SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 1–18

REMARKS ON THE DISCRETIZATION OF SOME NONCOERCIVE OPERATOR WITH APPLICATIONS TO HETEROGENEOUS MAXWELL EQUATIONS*

ANNALISA BUFFA†

Abstract. We aim to provide a framework for the analysis of convergence for the Galerkin approximation for a class of noncoercive problems. We provide a sufficient condition on the finite element space for the convergence and optimality of the Galerkin scheme. This theory is then applied to the study of the well-posedness and approximability of two problems in electromagnetism.

Key words. finite elements, boundary elements, Maxwell equations

AMS subject classifications. 65N30, 65N25, 65N38, 83N50

DOI. 10.1137/S003614290342385X

1. Introduction.

The aim of this paper is to provide a general framework for the analysis of convergence of Galerkin schemes for a class of linear continuous operators which are not definite (neither positive nor negative) and has, in general, no compact inverse. More precisely, letting X be a separable Hilbert space, we consider the class of operators verifying the following assumption.

Assumption 1. We assume that

(a) $A : X \to X'$ is a linear, continuous, and injective operator. We denote by M the continuity constant;

(b) there exists a stable splitting of the space X in $V \oplus W$ and denote by Θ the operator associated to the mapping: $u = v + w \mapsto v - w$;

(c) there exists a *compact* operator $T : X \to X'$ and a positive $\alpha \in \mathbb{R}_+$ such that

$$\mathrm{Re}\langle (A+T)u, \Theta\overline{u}\rangle_X \geq \alpha \, \|u\|_X^2. \tag{1.1}$$

Although the present theory could in principle be generalized to a more general expression of the operator Θ, we prefer to base the development of our main concepts on the above framework. This makes our exposition easier and, on the other hand, more general assumptions would be artificially complicated by the fact that we don't have a precise application in mind. Assumption 1 can be written in a different way, which might make it more clear. We rephrase it in section 2.

At the continuous level the invertibility of the operator A is an immediate consequence of Assumption 1. On the contrary, when we consider its Galerkin approximation, some care has to be devoted to the analysis of stability and convergence of the associated Galerkin scheme. The standard requirement [18] on the family of finite dimensional spaces $\{X_h\}_{h>0}$ is that $\overline{\bigcup_h X_h} = X$, but this turns out to be insufficient for ensuring the well-posedness of the associated Galerkin projection. Several papers exist on this subject and the most recent ones are devoted to the discretization of Maxwell equations; see, e.g., [20], [21], [5], [4], [17]. On the other hand, when concentrating on the edge elements approximation for the Maxwell problem with constant coefficients in a bounded domain (see section 4.1 for the definition of the Maxwell

*Received by the editors March 3, 2003; accepted for publication (in revised form) January 30, 2004; published electronically April 26, 2005.
http://www.siam.org/journals/sinum/43-1/42385.html

†MATI-CNR, Sede di Pavia, Via Ferrata 1, 27100 Pavia, Italy (annalisa@imati.cnr.it).

problem), the precise structure of the operators is used to write exhaustive, but dedicated, results which cannot be used for similar problems. For example, when trying to tackle the integral formulation of electromagnetic wave propagation, we find that the problem under consideration is mathematically more intricate (being of nonlocal, integrodifferential type) but reveals "almost" the same structure. This will be one of the applications we treat in this paper.

It is then natural to try to extend the known results to a general class of operators in order to fit all of them into the same framework. This is partially possible, and the aim of this paper is to collect some results in this direction. Thus, in section 3 we consider a condition on the finite element discretization (that we call the (GAP) property) and try to extract some consequences. The idea behind this condition goes back to Kato [28] and is the basis of the theory developed recently in the context of integral equations; see [14], [17], [8], [26]. We analyze further the consequences of this condition and attempt to give a comprehensive theory. We also approach the problem of deducing from (GAP) "spectral correctness" of the approximation in the sense of [20], but the results in this direction are incomplete. Finally, the question of whether, and/or when, (GAP) is a necessary condition for well-posedness is discussed but not answered.

Section 4 is devoted to applications. More precisely, in section 4.1, we show that the present theory easily permits the extension of the known results about the approximation of the Maxwell operator to the case of general bounded coefficients. In section 4.2, we consider a similar physical phenomenon, but one which has a more intricate mathematical formulation: the boundary integral formulation for electromagnetic wave propagation for piecewise homogeneous dielectric scatterers. Here, the effort to set up the problem is quite major, and it is a nontrivial application of the theory developed in this paper. We try our best to emphasize the structure of the problem. We dedicate a lot of room to this application because it shows the generality of the approach and, moreover, it is relevant in itself. Boundary integral discretization of electromagnetic problems is widely used in the engineering community, but mathematics had failed until now to prove the well-posedness of some related boundary element schemes.

Finally, in section 4 we inform the reader that we do not pretend that the section is self-contained; if it were, this would make the section far too long. Instead, we give detailed references for all results we use.

2. Setting of the problem. Let H , X be two complex separable Hilbert spaces such that $X \subset H$ with dense injection. We denote by X' the dual space of X when H plays the role of pivot space. We denote by $\|\cdot\|_H$ and $\|\cdot\|_X$ the associated norms and by $(\cdot,\cdot)_H$ and $(\cdot,\cdot)_X$ the inner product of H and X, respectively. Finally, $\langle\cdot,\cdot\rangle_X$ denotes the duality pairing in X.

First of all, we rephrase Assumption 1 in order to make more clear in which class of operators we are interested. Suppose that $A : X \to X'$ verifies Assumption 1. Since the splitting $X = V \oplus W$ is stable, there exists a projection $\Pi_V : X \to V$ such that Π is onto and $\ker\{\Pi_V\} = W$. We denote by Π_V' its adjoint with respect to the duality product $\langle\cdot,\cdot\rangle_X$. The operator A has a natural matrix representation associated with the splitting:

$$A \ \leftrightarrow \ \begin{pmatrix} A_{VV} & A_{VW} \\ A_{WV} & A_{WW} \end{pmatrix} \text{ with } \begin{array}{rl} A_{VV} = \Pi_V' A \Pi_V\,, & A_{WW} = (I-\Pi_V)'A(I-\Pi_V), \\ A_{VW} = \Pi_V' A(I-\Pi_V)\,, & A_{WV} = (I-\Pi_V)'A\Pi_V. \end{array}$$

Equation (1.1) can be expressed now by the following statement: The operator $A_{VV} -$

$A_{WW} + (A_{VW} - A_{WV})$ verifies the Gårding inequality. In the applications we have in mind, it will happen that $(A_{VW} - A_{WV})$ is either equal to zero (the self-adjoint case) or compact. In this case, Assumption 1 is equivalent to the requirement that $A_{VV} : V \to V'$ and $-A_{WW} : W \to W'$ verify a Gårding inequality.

Let $a : X \times X \to \mathbb{C}$ be the bilinear form associated with A, i.e., $a(u, u^t) = \langle Au, u^t \rangle_X$. We then solve the following.

PROBLEM 1 (continuous problem). *Given $f \in X'$, find*

$$u \in X \ : \ a(u, u^t) = \langle f, u^t \rangle_X \quad \forall\, u^t \in X.$$

It is easy to see by the following that Problem 1 is well-posed.

THEOREM 2.1. *For every $f \in X'$, there exists a unique solution $u \in X$ of the problem $Au = f$. Finally, there exists an isomorphism $\tilde{\Theta} : X \to X$ such that $\tilde{\Theta} - \Theta$ is compact and*

$$\mathrm{Re}\langle Au, \tilde{\Theta}\overline{u} \rangle \geq \alpha \|u\|_X^2.$$

Proof. The fact that A is injective and $A + T$ is invertible implies that A as well as its adjoint is invertible (see, e.g., [23]). It is immediate to construct $\tilde{\Theta}$ as $\tilde{\Theta} = (I + (A')^{-1}T')\Theta$, where A', T' are the adjoints of A, T. □

Finally, we are interested in operators enjoying some further properties and show how these have discrete counterparts for the associated Galerkin scheme.

To this aim we introduce two assumptions.

Assumption 2. Let $X = V \oplus W$ be the decomposition in Assumption 1; then $V \hookrightarrow H$ is compact.

Assumption 3. Let $X = V \oplus W$ be the decomposition in Assumption 1; then

$$\mathrm{Re}\, a(v, \overline{v}) \geq \beta \|v\|_X^2 \quad \forall\, v \in V.$$

3. Discretization. In this section we analyze the Galerkin discretization of Problem 1. Our aim is to provide sufficient conditions for the stability of the discrete problem and for the quasi optimality of the discretization scheme (in the sense of Ciarlet [18]).

The structure of this section is similar to the one chosen by Caorsi, Fernandes, and Raffetto in [16], while some of the results are a revision of the approach chosen in [17] and [8].

Let $\{X_h\}_{h \geq 0} \subset X$ be a family of finite dimensional subspace verifying the following.

Complete approximation space (CAS). $\lim_{h \downarrow 0} \inf_{u_h \in X_h} \|u - u_h\|_X = 0$.

When (CAS) is verified, we say that X_h is *approximating* in X. We denote by $I_h : X \to X_h$ the projection operator defined as $((u - I_h u, v))_X = 0 \ \forall\, v \in X_h$. The family $\{I_h\}_{h>0}$ is uniformly bounded with respect to h, and, if (CAS) holds, $I_h \to I$ pointwise in X.

The discrete variational problem reads as follows.

PROBLEM 2 (Galerkin projection). *Find $u_h \in X_h$ such that*

$$a(u_h, u_h^t) = \langle f, u_h^t \rangle_X \qquad \forall\, u_h^t \in X_h. \tag{3.1}$$

Gap property (GAP). We say that X_h verifies a gap property associated with Assumption 1 when there exist two subsets V_h, W_h of X_h such that

$$\delta_h = \max\{\delta(V_h, V), \delta(W_h, W)\} \to 0 \text{ when } h \to 0,$$

where

$$\delta(V_h, V) = \sup_{v_h \in V_h} \inf_{v \in V} \frac{\|v - v_h\|_X}{\|v_h\|_X}.$$

The (GAP) property has several consequences which are here written as theorems and lemmas. First of all, due to (GAP), there exists a continuous operator $\Pi : V_h \to V$ such that $\|v_h - \Pi v_h\| \leq 2\delta_h \|v_h\|_X$ (the same for W). This implies the following.

LEMMA 3.1. *The fact that $\delta(V_h, V) \leq \delta_h$, $\delta_h \to 0$ when $h \to 0$ implies that every continuous projector $P_0 : X \to V$, which is onto in V, verifies*

$$\|v_h - P_0 v_h\|_X \lesssim \delta_h \|v_h\|_X, \qquad v_h \in V_h.$$

Proof. It is enough to compute

$$\begin{aligned} \|v_h - P_0 v_h\|_X &= \|v_h - \Pi v_h + \Pi v_h - P_0 v_h\|_X \\ &= \|(I - P_0)(v_h - \Pi v_h)\|_X \leq 2\|I - P_0\|\delta_h \|v_h\|_X. \qquad \square \end{aligned}$$

We then can prove the following theorem.

THEOREM 3.2. *(GAP) implies that the splitting $X_h = V_h \oplus W_h$ is uniformly stable $\forall\ h < h_1$ for some $h_1 \in \mathbb{R}_+$.*

Proof (cf. [8]). We denote by P the projection with range V and kernel W. It commutes with conjugation. For any $(v_h, w_h) \in V_h \times W_h$ we have, with $u_h = v_h + w_h$,

$$\|v_h - Pv_h\|_X \leq 2\|I - P\|\delta_h \|v_h\|_X \lesssim \delta_h \|v_h\|_X, \tag{3.2}$$

and similarly,

$$\|Pw_h\|_X = \|w_h - (I - P)w_h\|_X \lesssim \delta_h \|w_h\|_X. \tag{3.3}$$

We use the identity

$$v_h = P(u_h) + ((I - P)w_h - w_h) + (v_h - Pv_h)$$

and, by triangle inequality, we obtain

$$\begin{aligned} \|v_h\|_X &\leq \|P(v_h + w_h)\|_X + \|(I - P)w_h - w_h\|_X + \|Pv_h - v_h\|_X \\ &\leq \|P\|\|u_h\|_X + 2\|P\|\delta_h \|w_h\|_X + 2\|I - P\|\delta_h \|v_h\|_X. \end{aligned}$$

Similarly,

$$\|w_h\|_X \leq \|I - P\|\|u_h\|_X + 2\|I - P\|\delta_h \|v_h\|_X + 2\|P\|\delta_h \|w_h\|_X.$$

Adding and rearranging we obtain for h small enough

$$\|v_h\|_X + \|w_h\|_X \leq 2(1 - \delta_h \max\{\|P\|, \|I - P\|\})^{-1} \max\{\|P\|, \|I - P\|\}\|u_h\|_X. \qquad \square \tag{3.4}$$

Since the splitting is stable for h sufficiently small, we denote by $P_h : X_h \to X_h$ the associated projection operator having V_h as range and W_h as kernel.

THEOREM 3.3. *(GAP) and (CAS) imply that V_h is approximating in V and that W_h is approximating in W, i.e.,*

$$\lim_{h \downarrow 0} \inf_{v_h \in V_h} \|v - v_h\|_X = 0,\ \lim_{h \downarrow 0} \inf_{w_h \in W_h} \|w - w_h\|_X = 0 \quad \forall\, v \in V,\ w \in W. \tag{3.5}$$

Proof. We estimate the best approximation error as follows:

$$\inf_{v_h \in V_h} \|v - v_h\|_X \leq \|v - P_h I_h v\|_X. \tag{3.6}$$

It is easy to see that $v - P_h I_h v = P(v - I_h v) + (P - P_h)(I_h v)$ and that, if we set $I_h v = v_h + w_h$ with $v_h \in V_h$, $w_h \in W_h$, we obtain

$$\|(P - P_h) I_h v\|_X \leq \|P v_h - v_h\|_X + \|P w_h\|_X \lesssim \delta_h \|I_h v\|_X.$$

As a consequence,

$$\begin{aligned} \inf_{v_h \in V_h} \|v - v_h\|_X &\lesssim \|v - P_h I_h v\|_X \\ &\lesssim \|P(v - v_h)\|_X + \delta_h \|I_h(v)\|_X \lesssim \|v - I_h v\|_X + \delta_h \|v\|_X. \end{aligned}$$

The statement is then proved since I_h is a linear and continuous operator, and (CAS) implies that $(I - I_h)$ is pointwise converging to 0. □

Remark 3.4. In the particular case in which $W_h \subset W$, we have of course that $\delta(W_h, W) = 0$. This means that the condition $\delta(V_h, V) \to 0$ for $h \to 0$ implies that W_h is approximating in W.

THEOREM 3.5. *(GAP) implies the following:*

(i) *If Assumption* 3 *holds, then there exists an* h_2 *such that* $\forall\ h < h_2$*, it holds that*

$$\operatorname{Re} a(v_h, v_h) \geq \frac{\beta}{2} \|v_h\|_X^2.$$

(ii) *If Assumption* 2 *holds, then we have the following: Let* $\{v_h\}_{h>0}$ *be a sequence such that* $v_h \in V_h\ \forall\ h > 0$*, and it verifies* $\|v_h\|_X \leq 1\ \forall\ h$*; there exists a subsequence* (*denoted again by* $\{v_h\}_{h>0}$) *and a* $v \in V$ *such that* $v_h \to v$ *strongly in* H.

Proof. (i) Let $v_h \in V_h$. Then

$$\begin{aligned} \operatorname{Re} a(v_h, \overline{v}_h) &= \operatorname{Re} a(\Pi v_h, \overline{\Pi v_h}) + \operatorname{Re}\{a(\Pi v_h - v_h, \overline{\Pi v_h}) - a(v_h, \overline{\Pi v_h - v_h})\} \\ &\geq \|\Pi v_h\|_X^2 - 2M \|v_h\|_X 2\delta_h \|v_h\|_X. \end{aligned}$$

We conclude by recalling the definition of Π and fixing h_2 such that $\beta(1 - 2\delta_h) - 4M\delta_h(1 + \delta_h) \geq \beta/2$.

(ii) Consider a sequence $\{v_{h_j}\}_{j \in \mathbb{N}}$, $h_j \to 0$, for $j \to \infty$ such that $\|v_{h_j}\|_X \leq 1$. We define $v_j = \Pi v_{h_j}$; by continuity of Π, we have $\|v_j\|_X \leq C$. This implies that there exists an increasing sequence j_k such that $v_{j_k} \to v$ strongly in H due to Assumption 3. On the other hand, (GAP) implies that $\|v_{h_{j_k}} - v_{j_k}\|_X \leq C\delta_{h_{j_k}}$, where $\delta_{h_{j_k}} \to 0$, when $k \to \infty$. Hence,

$$\|v_{h_{j_k}} - v\|_H \leq \|v_{j_k} - v\|_H + \|v_{h_{j_k}} - v_{j_k}\|_X \to 0, \quad k \to \infty, \tag{3.7}$$

which means $v_{h_{j_k}} \to v$ strongly in H. □

Remark 3.6. Property (ii) in Theorem 3.5 is commonly called the *discrete compactness property* for V_h and has been the object of several papers concerning edge elements approximation for Maxwell equations. See the very recent book [32] or papers [29], [30], [2], [3], [33]. Further comments are due, and we postpone them to section 4 (Remark 4.5).

We come now to the well-posedness of the discrete problem. Such a result has basically been proved in [14] (see also [8]). We report it here in its most general form.

THEOREM 3.7. *Let (CAS) and (GAP) hold. There exists an h_3 such that $\forall\, h < h_3$, Problem 2 is well-posed. Moreover, let $u \in X$, $u_h \in X_h$ be the solution of Problems 1 and 2. We have*

$$\|u - u_h\|_X \lesssim \inf_{\xi_h \in X_h} \|u - \xi_h\|_X. \tag{3.8}$$

Proof. As in the proof of Theorem 2.1, we have with $\tilde{\Theta} := (I + (A')^{-1}T')\Theta$ that

$$\forall u \in X : \quad \mathrm{Re}\langle Au, \tilde{\Theta}\overline{u}\rangle_X \geq \alpha \|u\|_X^2.$$

Moreover, $\tilde{\Theta} - \Theta = (A')^{-1}T'\Theta$ is compact.

Then

$$\left\|(I - I_h)\tilde{\Theta}u_h\right\|_X \leq \left\|(I - I_h)(\tilde{\Theta} - \Theta)u_h\right\|_X + \|(I - I_h)\Theta u_h\|_X\,.$$

Using (CAS) we have $\forall\ U \in X$ that $\|(I - I_h)U\|_X \to 0$ as $h \to 0$. Since $\tilde{\Theta} - \Theta$ is compact we obtain that

$$\epsilon_h := \left\|(I - I_h)(\tilde{\Theta} - \Theta)\right\|_{X \to X} \to 0 \quad \text{as } h \to 0.$$

Now let $u_h \in X_h$ be arbitrary. Then u_h has the decomposition $u_h = v + w$ with $v \in V$, $w \in W$, and we have $\Theta u_h = v - w$. There is also the decomposition $u_h = v_h + w_h$ with $v_h \in V_h$, $w_h \in W_h$. We have for $\Theta u_h = v - w$

$$\begin{aligned} \|(I - I_h)\Theta u_h\|_X &= \left\|(I - I_h)\Big(\Theta u_h - (v_h - w_h)\Big)\right\|_X \\ &\lesssim \|(v - w) - (v_h - w_h)\|_X \\ &\lesssim (\|v - v_h\|_X + \|w - w_h\|_X). \end{aligned} \tag{3.9}$$

Now, using the same argument as in the proof of Theorem 3.3, we have

$$\|v - v_h\|_X \lesssim (\|Pv_h - v_h\|_X + \|Pw_h\|_X) \lesssim \delta_h \|u_h\|_X.$$

As $w - w_h = -(v - v_h)$ we obtain

$$\|v - v_h\|_X + \|w - w_h\|_X \lesssim \delta_h\, \|u_h\|_X \tag{3.10}$$

and we obtain $\forall\ u_h \in X_h$

$$\left\|(I - I_h)\tilde{\Theta}u_h\right\|_X \leq (\epsilon_h + C\delta_h)\, \|u_h\|_X\,,$$

which implies that for sufficiently small h and $\forall\ u_h \in X_h$

$$\mathrm{Re}\langle Au_h, I_h\tilde{\Theta}u_h\rangle \geq \mathrm{Re}\langle Au_h, \tilde{\Theta}u_h\rangle - C(\varepsilon_h + C\delta_h)\, \|u_h\|_X^2 \geq \alpha/2\, \|u_h\|_X^2\,.$$

Since $I_h\tilde{\Theta} : X_h \to X_h$ is bounded independently of h, we have proved that there exist $\alpha > 0$ and $h_\star > 0$ such that $\forall h < h_\star$

$$\inf_{0 \neq u_h \in X_h} \sup_{0 \neq u_h^t \in X_h} \frac{\mathrm{Re}\langle Au_h, u_h^t\rangle}{\|u_h\|_X\, \|u_h^t\|_X} \geq \frac{\alpha}{2}. \tag{3.11}$$

It is well known that this discrete inf-sup condition implies that Problem 2 has a unique solution and that (3.8) holds. □

We end this section with a result concerning the correctness of the spectral approximation, i.e., we want to know if the Galerkin operator $A_h : X_h \to X'$ is a *correct spectral approximation* of A. To this aim, we need to recall some definitions and introduce some nomenclature.

First of all, we consider the solution operators $S : H \to X$, $S_h : X_h \to X_h$ defined as

$$a(Su, u^t) = (u, u^t)_H \;\; \forall\, u^t \in X \,, \quad a(S_h u_h, u_h^t) = (u_h, u_h^t)_H \;\; \forall\, u_h^t \in X_h. \tag{3.12}$$

By Theorems 2.1 and 3.7, we know that $S\,,\ S_h$ exist and are continuous, at least for h sufficiently small.

Let $\sigma(S)$ denote the spectrum of the operator S, and $\sigma(S_h)$ the one for the operator S_h.

Finally, following [21] (see also [20]), we define

$$\|S - S_h\|_h = \inf_{u_h \in X_h\,, \|u_h\|_X \le 1} \|(S - S)u_h\|_X.$$

It is known that if (CAS) holds and we also have that $\|S - S_h\|_h \to 0$ when $h \to 0$, then S_h provides a correct approximation of the spectrum in the sense expressed in [20]. We report the details of this definition only in the self-adjoint case and refer to [20] for the general case; when S and S_h are self-adjoint, we say that S_h is asymptotically spectrally correct if the following hold:

1. $\lim_{h \downarrow 0} \delta(\lambda, \sigma(S_h)) = 0 \,\forall\, \lambda \in \sigma(S_h)$.
2. If λ has multiplicity m, there are exactly m discrete eigenvalues converging to λ.
3. Let $\lambda \in \sigma(S)$ with multiplicity m and $E_\lambda(S)$ the associated eigenspace, and $\lambda_{h,i} \in \sigma(S_h)$, $i = 1, \dots, m$, the discrete approximation of λ with $E_{\lambda_{h,i}}(S_h)$ the corresponding eigenspace. Then
$$\delta(E_\lambda(S), \oplus_i E_{\lambda_{h,i}}(S_h))\,,\ \delta(\oplus_i E_{\lambda_{h,i}}(S_h), E_\lambda(S)) \to 0\,, \quad h \to 0.$$

THEOREM 3.8. *Let Assumption* 2 *hold and suppose, moreover, that*

$$\sup_{w_h \in W_h\,, \|w_h\|_X \le 1} \inf_{\lambda_h \in X_h} \|S w_h - \lambda_h\|_X \to 0 \textit{ when } h \to 0.$$

Then (CAS) and (GAP) imply that $\|S_h - S\|_h \to 0$ when $h \to 0$.

Proof. Let us fix $x_h \in X_h$, $\|x_h\|_X \le 1$. We estimate

$$\begin{aligned}
\|(S - S_h)x_h\|_X^2 &\lesssim a((S - S_h)x_h, \tilde{\Theta}(S - S_h)x_h) \\
&\lesssim a((S - S_h)x_h, \Theta(S - S_h)x_h) + \epsilon_h \|(S - S_h)x_h\|_X^2 \\
&\lesssim \inf_{\lambda_h \in X_h} a((S - S_h)x_h, \Theta(S - S_h)x_h - \Theta_h \lambda_h) \\
&\qquad + \epsilon_h \|(S - S_h)x_h\|_X^2 \\
&\lesssim \inf_{\lambda_h \in X_h} a((S - S_h)x_h, \Theta(Sx_h - \lambda_h)) + \epsilon_h \|(S - S_h)x_h\|_X^2 \\
&\qquad + \delta_h \|\lambda_h\|_X \|(S - S_h)x_h\|_X,
\end{aligned} \tag{3.13}$$

where we have used that $\tilde{\Theta} - \Theta$ is compact, the Galerkin orthogonality and, finally, that $\|(\Theta - \Theta_h)\lambda_h\|_X \lesssim \delta_h \|\lambda_h\|_X$, which is an immediate consequence of (GAP). From

(3.13), for h small enough, and using the continuity of the bilinear form a, we can deduce that

$$\|(S-S_h)x_h\|_X \lesssim \inf_{\lambda_h\in X_h} \|Sx_h-\lambda_h\|_X + \delta_h\|\lambda_h\|_X, \tag{3.14}$$

which is "almost" a pure approximation property. Since $\|Sx_h\| \lesssim 1$,

$$\inf_{\lambda_h\in X_h} \|Sx_h-\lambda_h\|_X = \|Sx_h - I_hSx_h\|_X = \inf_{\lambda_h\in X_h \|\lambda_h\|_X\lesssim 1} \|Sx_h-\lambda_h\|_X.$$

It is easy to see that, $\forall\ x_h$ such that $\|x_h\|_X \leq 1$, (3.14) implies

$$\|(S-S_h)x_h\|_X \lesssim \delta_h + \inf_{\lambda_h\in X_h} \|Sx_h-\lambda_h\|_X. \tag{3.15}$$

Now, select $x_h \in V_h$. Assumption 2 ensures that there exists a $\tilde{v}\in V$ such that, up to extractions, $\|x_h-\tilde{v}\|_H \to 0$ when $h\to 0$. Thus

$$\|(S-S_h)x_h\|_X \lesssim \delta_h + \|x_h-\tilde{v}\|_H + \inf_{\lambda_h\in X_h} \|S\tilde{v}-\lambda_h\|_X.$$

Hence, (CAS) allows us to conclude that

$$\sup_{x_h\in V_h\ ,\|V_h\|_X\leq 1} \|(S-S_h)x_h\| \to 0 \text{ when } h\to 0.$$

The statement is proved since

$$\begin{aligned}\sup_{u_h\in X_h\ ,\ \|u_h\|_X\leq 1} \|(S-S_h)u_h\|_X^2 \leq & \sup_{v_h\in V_h\ ,\ \|v_h\|_X\lesssim 1} \|(S-S_h)v_h\|_X^2 \\ & + \sup_{w_h\in w_h\ ,\ \|w_h\|_X\lesssim 1} \|(S-S_h)w_h\|_X^2,\end{aligned} \tag{3.16}$$

and the second term in the right-hand side is converging to 0 by assumption. □

Corollary 3.9. *If $S_{|W} : W \to X$ is either compact or a multipication by a sufficiently regular function, then S_h is asymptotically spectrally correct.*

Proof. It is a consequence of (3.15) and (3.16). □

Remark 3.10. This theorem is not completely satisfactory. We expect the statement to hold under much weaker conditions on $S_{|W}$. In the case of Maxwell equations in bounded domains, the discrete compactness property, which is a consequence of (GAP), turns out to be a sufficient condition for the associated Galerkin approximation to be spectrally correct [2] and also spurious free in the sense given in [16]. Finally, the following reasonable question remains open: "Letting (GAP) hold, under Assumptions 1, 2, can we prove that at least a part of the spectrum is well approximated for h sufficiently small?"

4. Applications. We will present here two applications of this theory. They concern two different problems in electromagnetism: (i) Compute solutions of Maxwell equations in a cavity characterized by variable magnetic and electric properties. (ii) Compute the electromagnetic diffraction due to a heterogeneous/piecewise homogeneous dielectric material.

Note that this section will not be self-contained in the sense that we will not recall (with precise statements) all the known properties about the finite elements we use; we provide instead a list of references and we try, when possible, to refer to the recent book [32] or to the review paper [24].

Let D (∂D, $\mathbf{n}_D$, resp.) denote a bounded connected Lipschitz domain (its boundary and the unit outer normal to D on ∂D, resp.) and D^c denote its complement. We set ϵ to be the electric permittivity and μ to be the magnetic permeability.

$\mathbf{E}$ and $\mathbf{H}$ denote the electric and magnetic fields, respectively, and we assume they satisfy the linear time-harmonic Maxwell equations

$$\mathbf{curl}\,\mathbf{E} - i\omega\mu\mathbf{H} = 0\,, \quad \mathbf{curl}\,\mathbf{H} + i\omega\epsilon\mathbf{E} = \mathbf{J} \quad \text{in} \quad D, \tag{4.1}$$

where $\omega \in \mathbb{R}_+$ is a fixed frequency and $\mathbf{J}$ is an imposed current density.

Let us introduce some Sobolev spaces and some operators which will be used throughout this section. We define

$$\begin{aligned}\mathbf{H}(\mathbf{curl}, D) &= \{\mathbf{u} \in L^2(D)^3 \ : \ \mathbf{curl}\,\mathbf{u} \in L^2(D)^3\};\\ \mathbf{H}_{\mathrm{loc}}(\mathbf{curl}, D^c) &= \{\mathbf{u} \in L^2_{\mathrm{loc}}(D^c)^3 \ : \ \mathbf{curl}\,\mathbf{u} \in L^2_{\mathrm{loc}}(D^c)^3\};\\ \mathbf{H}(\mathbf{curl}^2, D) &= \{\,\mathbf{u} \in \mathbf{H}(\mathbf{curl}, D) \mid \mathbf{curl}\,\mu^{-1}\,\mathbf{curl}\,\mathbf{u} \in L^2(D)^3\,\}.\end{aligned}$$

We denote by γ_D the tangential trace operator mapping $\mathbf{u}$ in $\mathbf{n}_D \times \mathbf{u}_{|\partial D}$, $\mathbf{u} \in C^\infty(\overline{D})$. Let $\mathbf{H}^{1/2}_\times(\partial D) = \gamma_D\{H^1(D)^3\}$ and $\mathbf{H}^{-1/2}_\times(\partial D)$ be its dual with respect to the natural duality pairing $b(\boldsymbol{\lambda}, \boldsymbol{\xi}) = \int_{\partial D} \boldsymbol{\lambda} \cdot (\boldsymbol{\xi} \times \mathbf{n})$. Note that the injection $\mathbf{H}^{1/2}_\times(\partial D) \hookrightarrow \mathbf{L}^2_t(\partial D) := \{\mathbf{u} \in L^2(\Gamma)^3 \ : \ \mathbf{u} \cdot \mathbf{n}_D = 0\}$ is compact.

We set $\mathbf{X}(\partial D) := \{\boldsymbol{\lambda} \in \mathbf{H}^{-1/2}_\times(\partial D) \ : \ \mathrm{div}_\Gamma\, \boldsymbol{\lambda} \in H^{-1/2}(\partial D)\}$. See [7], [9], [10], and also [14] for definitions and details. The idea to keep in mind is that vectors in $\mathbf{X}(\partial D)$ are tangential vector fields of Sobolev regularity $-1/2$, with surface divergence of Sobolev regularity $-1/2$, and that the related definitions for nonsmooth boundaries can be given as extensions of the same well-known definitions for regular manifolds.

It is known that $\gamma_D : \mathbf{H}(\mathbf{curl}, D) \to \mathbf{X}(\partial D)$ and $\gamma^c_D : \mathbf{H}_{\mathrm{loc}}(\mathbf{curl}, D^c) \to \mathbf{X}(\partial D)$ are linear continuous and admit a right inverse [12], [7]. Finally, we denote by γ_N the Neumann trace operator associated with the mapping $\mathbf{u} \mapsto \gamma_D(\mu^{-1}\,\mathbf{curl}\,\mathbf{u})$. It turns out [11], [14] that $\gamma_N : \mathbf{H}(\mathbf{curl}^2, D) \to \mathbf{X}(\partial D)$ is linear, continuous, and admits a right inverse. Finally, we set

$$\mathbf{H}_0(\mathbf{curl}, D) = \{\mathbf{u} \in L^2(D)^3 \ : \ \mathbf{curl}\,\mathbf{u} \in L^2(D)^3\,,\ \gamma_D(\mathbf{u}) = 0\}.$$

4.1. Maxwell interior problem.

Let Ω be a Lipschitz bounded polyhedron in $\mathbb{R}^3$.

In this section we consider (4.1) on Ω together with a perfect conductor boundary condition, i.e., $\gamma_D\mathbf{E} = 0$ on $\partial\Omega$. We suppose that $\epsilon\,, \mu \in L^\infty(\Omega)$, $0 < c_0 \leq \epsilon(\mathbf{x})\,, \mu(\mathbf{x}) \leq C_0$, for almost all $\mathbf{x} \in \Omega$.

Eliminating the field $\mathbf{H}$, defining $\mathbf{f} := i\omega\mathbf{J}$, and integrating by parts, we obtain the following (well-known) variational formulation.

PROBLEM 3. *Given* $\mathbf{f} \in L^2(\Omega)^3$, *find* $\mathbf{u} \in \mathbf{H}_0(\mathbf{curl}, \Omega)$ *such that* $\forall\, \mathbf{v} \in \mathbf{H}_0(\mathbf{curl}, \Omega)$

$$\int_\Omega \mu^{-1}\,\mathbf{curl}\,\mathbf{u}\,\mathbf{curl}\,\mathbf{v} - \omega^2 \int_\Omega \epsilon\,\mathbf{u} \cdot \mathbf{v} = \int_\Omega \mathbf{f} \cdot \mathbf{v}. \tag{4.2}$$

The following theorem is well known and has been proved, e.g., in [40] (see also [39] for the fundamental compactness result which is the basis of this theorem).

THEOREM 4.1. *Problem* 3 *admits a unique solution* $\mathbf{u} \in \mathbf{H}_0(\mathbf{curl}, \Omega)$ *except for* $\omega \subset \{0\} \cup \{\omega_j\}_{j>0}$, *where* $\{\omega_j\}$ *is a positive increasing sequence diverging to* $+\infty$. *If* $\mathrm{div}\,\mathbf{J} = 0$, *then* $\mathrm{div}(\epsilon\mathbf{u}) = 0$.

Now we decompose the space $\mathbf{X} := \mathbf{H}_0(\mathbf{curl}, \Omega)$ as $\mathbf{X} = \mathbf{V} \oplus \mathbf{W}$:

$$\mathbf{V} = \{\mathbf{u} \in \mathbf{H}_0(\mathbf{curl}, \Omega) \ : \ \operatorname{div} \mathbf{u} = 0\}\,, \ \mathbf{W} = \nabla H^1_0. \tag{4.3}$$

It is known (see, e.g., [1]) that there exists a positive σ such that $\mathbf{V} \hookrightarrow H^{1/2+\sigma}(\Omega)^3$. This means in particular that $\int_\Omega \epsilon \mathbf{v} \cdot \mathbf{w}$ is a compact bilinear form ($\mathbf{V} \hookrightarrow \mathbf{L}^2(\Omega)$ is compact). Calling Θ the mapping $\mathbf{u} = \mathbf{v} + \mathbf{w} \mapsto \mathbf{v} - \mathbf{w}$ associated with the decomposition (4.3), and $a(\cdot, \cdot)$ the bilinear form of the left-hand side of (4.2), we have

$$a(\mathbf{u}, \Theta\overline{\mathbf{u}}) \geq \alpha \|\mathbf{u}\|^2_{\mathbf{H}(\mathbf{curl},\Omega)} - c(\mathbf{u}, \Theta\overline{\mathbf{u}}),$$

where $c(\cdot, \cdot) : \mathbf{X} \times \mathbf{X} \to \mathbb{C}$ is a compact bilinear form. Thus, when ω is not an eigenvalue of Problem 3, i.e., $\omega \notin \{0\} \cup \{\omega_j\}_{j>0}$, Problem 3 fits exactly into Assumption 1 and also Assumption 2.

Now we pass to the discretization, and we consider the family of conforming finite dimensional spaces $\{\mathbf{X}_h\}_{h>0}$ generated by Nédélec finite elements of the first family of degree k fixed. The family $\{\mathbf{X}_h\}_{h>0}$ corresponds to a family of triangulations $\{\mathcal{T}_h(\Omega)\}_{h>0}$ of the domain Ω which, we assume for simplicity, to be made of tetrahedra. We defer the reader to [35], [22] (see also [31]), or again [24] for a precise definition. Here we list only the properties we need:

1. The space $\mathbf{X}_h$ constructed by Nédélec elements of degree k is approximating in $\mathbf{H}(\mathbf{curl}, \Omega)$, i.e., (CAS) is verified.
2. Let $\mathcal{P}_h$ be the H^1_0-conforming finite element space generated by piecewise polynomials of degree k on $\mathcal{T}_h(\Omega)$. Then, $\nabla \mathcal{P}_h = \{\mathbf{u}_h \in \mathbf{X}_h \ : \ \mathbf{curl}\, \mathbf{u}_h = 0\}$.
3. Denote by Π_k the Nédélec interpolant, we have that
   (a) Π_k is well defined on the space $\mathbf{V}^h = \{\mathbf{v} \in \mathbf{V} \ : \ \mathbf{curl}\, \mathbf{v} \subset \mathbf{curl}\, \mathbf{X}_h\}$ and continuous as an operator from $\mathbf{V}^h$ to $\mathbf{X}_h$;
   (b) $\forall\, \mathbf{v} \in \mathbf{V}^h$, $\mathbf{curl}\, \mathbf{v} = \mathbf{curl}\, \Pi_k \mathbf{v}$, and $\|\mathbf{v}\|_{\mathbf{H}(\mathbf{curl},\Omega)} \approx \|\Pi_k \mathbf{v}\|_{\mathbf{H}(\mathbf{curl},\Omega)}$.

Property 3 is a nontrivial property of Nédélec finite elements, which is basically due to V. Girault, and was first used in [19]. We refer to [24] for its proof and some comments. We set $\mathbf{V}_h = \Pi_k \mathbf{V}^h$ and $\mathbf{W}_h = \nabla \mathcal{P}_h$. We need to prove that $\mathbf{X}_h = \mathbf{V}_h + \mathbf{W}_h$. Letting $\mathbf{u}_h \in \mathbf{X}_h \subset \mathbf{X}$, it can be decomposed as $\mathbf{u}_h = \mathbf{v} + \nabla p$, $\mathbf{v} \in \mathbf{V}\,,\ \nabla p \in \mathbf{W}$. Since $\mathbf{curl}\, \mathbf{u}_h = \mathbf{curl}\, \mathbf{v}$, we deduce $\mathbf{v} \in \mathbf{V}^h$. On the other hand, $\mathbf{u}_h = \Pi_k \mathbf{u}_h = \Pi_k \mathbf{v} + \Pi_k \nabla p = \Pi_k \mathbf{v} + \nabla p_h$, $p_h \in \mathcal{P}_h$. Thus, $\nabla p_h \in \mathbf{W}_h$, $\Pi_k \mathbf{v} \in \mathbf{V}_h$.

It is immediate to see that $\mathbf{W}_h \subset \mathbf{W}$ and also $\mathbf{X}_h = \mathbf{V}_h \oplus \mathbf{W}_h$. We need only prove the following.

Theorem 4.2. *$\delta(\mathbf{V}_h, \mathbf{V})\,,\ \delta(\mathbf{W}_h, \mathbf{W}) \to 0$ when $h \to 0$, i.e., (GAP) holds.*

Proof. The building block of this proof basically exists in several papers; see, e.g., [24]. First of all, $\delta(\mathbf{W}_h, \mathbf{W}) = 0$.

Let $\mathbf{v}_h \in \mathbf{V}_h$. We know that $\mathbf{v}_h = \Pi_k \mathbf{v}$ for some $\mathbf{v} \in \mathbf{V}^h$. Moreover, $\mathbf{curl}\, \mathbf{v} = \mathbf{curl}\, \mathbf{v}_h$ and the regularity results proved in [1] ensure that there exists $\sigma > 0$ such that $\mathbf{v} \in H^{1/2+\sigma}(\Omega)$ with the continuity estimate

$$\|\mathbf{v}\|_{H^{1/2+\sigma}(\Omega)^3} \lesssim \|\,\mathbf{curl}\, \mathbf{v}_h\|_{L^2(\Omega)^3}.$$

Using the approximation properties of Π_k on $\mathbf{V}^h$, we have that $\forall\, \mathbf{v} \in \mathbf{V}^h$,

$$\begin{aligned} \|\mathbf{v} - \Pi_k \mathbf{v}\|_{L^2(\Omega)^3} &\lesssim h^{1/2+\sigma}(\|\mathbf{v}\|_{H^{1/2+\sigma}(\Omega)^3} + \|\,\mathbf{curl}\, \mathbf{v}_h\|_{L^2(\Omega)^3}) \\ &\lesssim h^{1/2+\sigma} \|\mathbf{v}_h\|_{\mathbf{H}(\mathbf{curl},\Omega)}. \qquad \square \end{aligned}$$

We are now ready to define the Galerkin approximation of Problem 3.

PROBLEM 4 (Galerkin). *Given $\mathbf{f} \in L^2(\Omega)^3$, let $\omega \in \mathbb{R}_+$ be such that Problem* 3 *is well-posed.*

Find $\mathbf{u}_h \in \mathbf{X}_h$ such that $\forall\, \mathbf{v} \in \mathbf{X}_h$,

$$\int_\Omega \mu^{-1}\, \mathbf{curl}\, \mathbf{u}_h \cdot \mathbf{curl}\, \mathbf{v}_h - \omega^2 \int_\Omega \epsilon\, \mathbf{u}_h \cdot \mathbf{v}_h = \int_\Omega \mathbf{f} \cdot \mathbf{v}_h. \tag{4.4}$$

The general setting developed in section 3 (and Theorem 3.7) provides the following statement as a corollary of Theorem 4.2.

COROLLARY 4.3. *There exists an $h_\star$ such that $\forall\, h < h_\star$ Problem* 4 *is wellposed. Moreover,*

$$\|\mathbf{u} - \mathbf{u}_h\|_{\mathbf{X}} \leq C \inf_{\mathbf{v}_h \in \mathbf{X}_h} \|\mathbf{u} - \mathbf{v}_h\|_{\mathbf{X}}.$$

Remark 4.4. Thanks to our general setting, the proof of well-posedness for Problem 4 is completely independent of the fact that we treat general coefficients (note that the numerical method is proved convergent under the most general assumption of $\epsilon\,,\ \mu \in L^\infty(\Omega)$, $0 < c_0 \leq \epsilon(\mathbf{x})\,,\ \mu(\mathbf{x}) < C_0$).

Remark 4.5 (spectral correctness). The relevant eigenvalue problem associated with Problem 3 is, of course, the one of computing the frequency ω for which the problem is not well-posed. This fits into the theory developed in section 3 when choosing as space H the space $L^2(\Omega)^3$ endowed with the inner product $(\mathbf{u}, \mathbf{u}^t)_H = \int_\Omega \epsilon\, \mathbf{u} \cdot \mathbf{u}^t$, $\mathbf{u}\,,\ \mathbf{u}^t \in L^2(\Omega)^3$. In this case, it is immediate to see that the associated solution operator S (see (3.12)) coincides with $-\frac{1}{\omega^2} I$ on $\mathbf{W}$, i.e., Theorem 3.8 and Corollary 3.9 can be applied and ensure asymptotic spectral correctness. This statement is not new for Maxwell equations and has been proved in [16]. Actually, there it was proved that the *discrete compactness property* (see Remark 3.6) together with the fact that $\mathbf{W}_h$ is approximating in $\mathbf{W}$ is a necessary and sufficient condition for the spurious free asymptotic spectral correctness (see [16] for definitions). Nonetheless, as before, the result expressed in our framework is completely independent of the fact that we treat general coefficients. Finally, it is easy to see that for this particular application, the *discrete compactness property* is equivalent to (GAP) and $\delta(\mathbf{W}_h, \mathbf{W}) = 0$.

4.2. Maxwell transmission problem. We suppose that the space is filled with different magnetic materials; i.e., the electric permittivity and the magnetic permeability ϵ and μ are piecewise positive constants on a fixed nonoverlapping polyhedral partition $\mathcal{P}$ of $\mathbb{R}^3$, $\mathbb{R}^3 = \bigcup_{j=1}^J \overline{\Omega}_j$. Moreover, we suppose that Ω_J is the only unbounded element of the partition and call $\Omega = \mathbb{R}^3 \setminus \overline{\Omega_J}$.

We define the piecewise constant function $k := k(\mathbf{x}) := \omega\sqrt{\epsilon\mu}$ and denote by Σ the set of interfaces, i.e., $\Sigma = \bigcup_{j=1}^J \partial\Omega_i$. Note that according to the notation introduced, $\mathbb{R}^3 = \bigcup_{j=1}^J \Omega_j \cup \Sigma$. The problem we want to solve is the following: Find $\mathbf{u} \in \mathbf{H}_{\mathrm{loc}}(\mathbf{curl}, \mathbb{R}^3 \setminus \Sigma)$ verifying (i)

$$\mathbf{curl}\, \mathbf{curl}\, \mathbf{u} - k^2 \mathbf{u} = 0 \quad \text{in } \mathbb{R}^3 \setminus \Sigma; \tag{4.5a}$$

(ii) Silver–Müller condition at infinity:

$$\left| \mathbf{curl}\, \mathbf{u}(\mathbf{r}) \times \frac{\mathbf{r}}{|\mathbf{r}|} - ik\mathbf{u} \right| = o\left(\frac{1}{|\mathbf{r}|}\right), \quad |\mathbf{r}| \to \infty; \tag{4.5b}$$

(iii) suitable transmission conditions on the set of interfaces Σ. In order to make precise the transmission conditions, we need to introduce suitable notation. First of all, for any vector $\mathbf{v}$ defined almost everywhere in $\mathbb{R}^3$, $\mathbf{v}_j$ always denotes its restriction to Ω_j.

Let $\Gamma_j = \partial\Omega_j$ and $\mathbf{n}_j$ be the unit outer normal to Ω_j, $j = 1, \dots, J$; we have at our disposal the space $\mathbf{X}(\Gamma_j)$ defined at the beginning of the section. We can then construct

$$\mathcal{X} := \mathbf{X}(\Gamma_1) \times \mathbf{X}(\Gamma_2) \times \cdots \times \mathbf{X}(\Gamma_J), \quad \underline{\boldsymbol{\xi}} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_J), \tag{4.6}$$

endowed with the product norm $\|\underline{\boldsymbol{\xi}}\|_{\mathcal{X}} = \sum_j \|\boldsymbol{\xi}_j\|_{\mathbf{X}(\Gamma_j)}$. On such a space, we define the *jump operator* $[\cdot]$ as the mapping

$$[\underline{\boldsymbol{\xi}}] \;:\; [\underline{\boldsymbol{\xi}}]_{|\Gamma_{ij}} = \boldsymbol{\xi}_i + \boldsymbol{\xi}_j \qquad \forall\, i, j = 1, \dots, J \text{ s.t. } \Gamma_j \cap \Gamma_i \neq \emptyset. \tag{4.7}$$

Now, if $\mathbf{u}$ solves (4.5a), we can construct the set of its Cauchy data as follows: Let $\boldsymbol{\xi}_j = \binom{\gamma_D \mathbf{u}_j}{\gamma_N \mathbf{u}_j}$ and $\underline{\boldsymbol{\xi}} = (\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_J)$.[1] Then, $\underline{\boldsymbol{\xi}} \in \mathcal{X} \times \mathcal{X}$. Applying the jump operator to each of the two lines of $\underline{\boldsymbol{\xi}}$, we impose the transmission condition as

$$[\underline{\boldsymbol{\xi}}] = \underline{\mathbf{f}} \qquad \text{for some fixed } \underline{\mathbf{f}},\ \underline{\mathbf{f}} \in [\mathcal{X}]^2. \tag{4.8}$$

Remark 4.6. We have to be careful in the definition of transmission conditions because the tangential trace operators depend on the orientation. Hence, the transmission condition cannot be defined directly on Σ, because Σ is never orientable.

A uniqueness result is available as follows.

THEOREM 4.7. *The problem* (4.5)–(4.8) *admits at most one solution.*

Proof. This is a direct consequence of Rellich's theorem (see Müller [34] for a proof). □

4.3. Statement of the problem. In order to show existence and to discretize the problem (4.5)–(4.8), we need to formulate it in terms of boundary integral equations on the interfaces Σ. We adopt a construction of the system of integral equations inspired by [14]. Other derivations are possible; see, e.g., [27].

We first introduce the 1-*dielectric* problem: Find $\mathbf{E} \in \mathbf{H}_{\text{loc}}(\mathbf{curl}, \Omega_j \cup \Omega_j^c)$ such that:

$$\begin{aligned} &\mathbf{curl}\,\mathbf{curl}\,\mathbf{E} - k_j^2 \mathbf{E} = 0 \qquad \text{in} \quad \Omega_j \cup \Omega_j^c, \\ &[\gamma_D]\mathbf{E} = \mathbf{m}; \qquad [\gamma_N]\mathbf{E} = \mathbf{j} \qquad \text{on} \quad \Gamma_j, \\ &\text{Silver–Müller radiation condition at } \infty, \end{aligned} \tag{4.9}$$

where $k_j \in \mathbb{R}_+$, $k_j = k(\mathbf{x})|_{\Omega_j}$, and $\mathbf{m}\,, \mathbf{j}$ are imposed transmission conditions. There exists an explicit representation for the solution $\mathbf{E}$ of (4.9), which reads for almost all $\mathbf{x} \in \Omega_j \cup \Omega_j^c$ as

$$\mathbf{E} = -\Psi_{\text{SL}}^j(\mathbf{j}) - \Psi_{\text{DL}}^j(\mathbf{m}),$$

[1] Note that the operators γ_D and γ_N are not indexed to keep the notation shorter. If applied to a field defined on Ω_j for some j, they represent the tangential trace operators on the boundary of Ω_j, i.e., Γ_j. Moreover, if they are applied to a field $\mathbf{u} \in \mathbf{H}_{\text{loc}}(\mathbf{curl}, \mathbb{R}^3)$, $\gamma_D \mathbf{u}$ stands for the vector $(\gamma_D \mathbf{u}_1, \dots, \gamma_D \mathbf{u}_J)$, γ_N stands for $(\gamma_N \mathbf{u}_1, \gamma_N \mathbf{u}_2, \dots, \gamma_N \mathbf{u}_J)$.

where Ψ^j_{SL} and Ψ^j_{DL} denote the single and double layer operators for the Maxwell problems as defined in [13]. More precisely, let $G(\mathbf{x},\mathbf{y}) = \frac{e^{ik|\mathbf{x}-\mathbf{y}|}}{4\pi|\mathbf{x}-\mathbf{y}|}$ be the standard Helmholtz kernel; then for $\mathbf{x} \in \Omega_j \cup \Omega_j^c$, Ψ^j_{SL} and Ψ^j_{DL} are given by

$$\Psi^j_{\mathrm{SL}}\mathbf{j}(\mathbf{x}) := \int_{\Gamma_j} G(\mathbf{x},\mathbf{y})\mathbf{j}(\mathbf{y})\,ds(y) + k^{-2}\nabla \int_{\Gamma_j} G(\mathbf{x},\mathbf{y})\,\mathrm{div}_\Gamma\,\mathbf{j}(\mathbf{y})\,ds(y),$$

$$\Psi_{\mathrm{DL}}\mathbf{m}(\mathbf{x}) := \mathbf{curl}\int_{\Gamma_j} G(\mathbf{x},\mathbf{y})\mathbf{m}(\mathbf{y})\,ds(y).$$

Let $\{\gamma_D\} := \frac{1}{2}(\gamma_D + \gamma_D^c)$, $\{\gamma_N\} := \frac{1}{2}(\gamma_N + \gamma_N^c)$. We construct the operator A_j associated with the domain Ω_j as

$$A_j \begin{pmatrix}\mathbf{m}\\ \mathbf{j}\end{pmatrix} := \begin{pmatrix}\{\gamma_D\}\\ \{\gamma_N\}\end{pmatrix}(-\Psi^j_{\mathrm{SL}}\mathbf{j} - \Psi^j_{\mathrm{DL}}\mathbf{m}). \tag{4.10}$$

Finally, for each Ω_j, we define the *antisymmetric* bilinear form $B_j \colon \mathbf{X}(\Gamma_j)^2 \times \mathbf{X}(\Gamma_j)^2 \to \mathbb{C}$ acting on sets of Cauchy data as

$$B_j\left(\begin{pmatrix}\mathbf{m}\\ \mathbf{j}\end{pmatrix}, \begin{pmatrix}\tilde{\mathbf{m}}\\ \tilde{\jmath}\end{pmatrix}\right) := -b_j(\mathbf{m},\tilde{\jmath}) + b_j(\tilde{\mathbf{m}},\mathbf{j}) \quad \forall \begin{pmatrix}\mathbf{m}\\ \mathbf{j}\end{pmatrix}, \begin{pmatrix}\tilde{\mathbf{m}}\\ \tilde{\jmath}\end{pmatrix} \in \mathbf{X}(\Gamma_j)^2, \tag{4.11}$$

where $b(\cdot,\cdot)$ is the duality pairing defined, as at the beginning of this section, as $b_j(\boldsymbol{\mu},\boldsymbol{\lambda}) = \int_{\Gamma_j} \boldsymbol{\mu}\cdot(\boldsymbol{\lambda}\times\mathbf{n}_j)$.

The following theorem has been proved in [14].

Theorem 4.8. *Let* $\mathbf{W}(\Gamma_j) = \{\boldsymbol{\lambda} \in \mathbf{X}(\Gamma_j) \ :\ \mathrm{div}_\Gamma\,\boldsymbol{\lambda} = 0\}$ *and let* $\mathbf{V}(\Gamma_j)$ *be any supplement of* $\mathbf{W}(\Gamma_j)$ *in* $\mathbf{X}(\Gamma_j)$ *such that* (i) $\mathbf{V}(\Gamma_j) \hookrightarrow \mathbf{L}^2_t(\Gamma_j)$ *is compact;* (ii) *the decomposition* $\mathbf{X}(\Gamma_j) = \mathbf{V}(\Gamma_j) \oplus \mathbf{W}(\Gamma_j)$ *is stable in* $\mathbf{X}(\Gamma_j)$. *Then call* $\Theta : \mathbf{X}(\Gamma_j)^2 \to \mathbf{X}(\Gamma_j)^2$ *the operator associated with the mapping* $\mathbf{u} = \mathbf{v}+\mathbf{w} \mapsto \mathbf{v}-\mathbf{w}$ *componentwise. The operator* A_j *is injective and verifies*

$$\mathrm{Re}\,B((A+T)\boldsymbol{\xi}, \Theta(\overline{\boldsymbol{\xi}})) \geq \|\boldsymbol{\xi}\|^2_X, \qquad T : \mathbf{X}(\Gamma_j)^2 \to \mathbf{X}(\Gamma_j)^2 \quad \textit{compact.}$$

In other words, we are within Assumptions 1 and 2. Note that the supplement $\mathbf{V}(\Gamma_j)$ fulfilling the assumptions of Theorem 4.8 can be constructed in several ways [14], [25].

We are now ready to give the integral formulation associated with (4.5)–(4.8).

Proposition 4.9. *Let* $\mathbb{A} = \mathrm{diag}\{A_1,\ldots,A_J\}$, *and*

$$\underline{B}(\underline{\boldsymbol{\xi}},\underline{\boldsymbol{\lambda}}) = \sum_{j=1}^{J} B_j(\boldsymbol{\xi}_j,\boldsymbol{\lambda}_j), \qquad \underline{\boldsymbol{\xi}},\ \underline{\boldsymbol{\lambda}} \in \mathcal{X}^2. \tag{4.12}$$

The vector field $\mathbf{u}$ *is a solution of* (4.5)–(4.8) if and only if *its Cauchy data verifies* $\underline{\boldsymbol{\xi}} = \underline{\boldsymbol{\xi}}^0 + \underline{\boldsymbol{\xi}}_{\boldsymbol{hom}}$, $[\underline{\boldsymbol{\xi}}^0] = \underline{\boldsymbol{f}}$, *and*

$$\underline{\boldsymbol{\xi}}_{\boldsymbol{hom}} \in \mathcal{X}^2_{\boldsymbol{hom}} \quad : \quad \underline{B}(\mathbb{A}\underline{\boldsymbol{\xi}}_{\boldsymbol{hom}}, \underline{\boldsymbol{\lambda}}_{\boldsymbol{hom}}) = \underline{B}((\tfrac{1}{2}\mathbb{I} - \mathbb{A})\underline{\boldsymbol{\xi}}^0, \underline{\boldsymbol{\lambda}}_{\boldsymbol{hom}}) \qquad \forall\, \underline{\boldsymbol{\lambda}}_{\boldsymbol{hom}} \in \mathcal{X}^2_{\boldsymbol{hom}}. \tag{4.13}$$

The proof of this result is completely equivalent to the one for the corresponding Helmholtz problem and can be found in [37] (see also [38] and [15]). The next section is devoted to showing that this problem fits into our framework.

4.3.1. Verification of Assumption 1. We want to prove that the operator $\mathbb{A} : \mathcal{X}^2_{\mathbf{hom}} \to (\mathcal{X}^2_{\mathbf{hom}})'$, $(\mathcal{X}^2_{\mathbf{hom}})'$ being the dual of $\mathcal{X}^2_{\mathbf{hom}}$ with respect to the duality product $\underline{B}(\cdot,\cdot)$ fits Assumption 1. We have the following theorem.

Theorem 4.10. *Define*

$$\mathbf{W}_{hom} = \{\underline{\boldsymbol{\xi}} \in \mathcal{X}_{hom}\,, \quad \mathrm{div}_\Gamma\, \boldsymbol{\xi}_j = 0 \ \forall\, j\}.$$

Then there exists a supplement $\mathbf{V}_{hom}$ of $\mathbf{W}_{hom}$ in $\mathcal{X}_{hom}$ such that

$$\mathcal{X}_{hom} = \mathbf{V}_{hom} \oplus \mathbf{W}_{hom} \qquad \textit{with} \quad \mathbf{V}_{hom} \hookrightarrow \underline{\mathbf{L}}^2_t(\Sigma) \ \textit{compact}, \tag{4.14}$$

where $\underline{\mathbf{L}}^2_t(\Sigma) = \otimes_{j=1}^J \mathbf{L}^2_t(\Gamma_j)$. Moreover, the splitting is stable, i.e., the following stability estimate holds: $\underline{\boldsymbol{\lambda}} \in \mathcal{X}$, $\underline{\boldsymbol{\lambda}} = \underline{\mathbf{v}} + \underline{\mathbf{w}}$, $\underline{\mathbf{v}} \in \underline{\mathbf{V}}$, $\underline{\mathbf{w}} \in \underline{\mathbf{W}}$, and

$$\|\underline{\mathbf{v}}\|_{\mathcal{X}} + \|\underline{\mathbf{w}}\|_{\mathcal{X}} \lesssim \|\underline{\boldsymbol{\lambda}}\|_{\mathcal{X}} \lesssim \|\underline{\mathbf{v}}\|_{\mathcal{X}} + \|\underline{\mathbf{w}}\|_{\mathcal{X}}. \tag{4.15}$$

Proof. We assume for simplicity that each Ω_j is connected and simply connected, we construct the space $\mathbf{V}_{\mathbf{hom}}$ by the definition of a projector $\Pi_{\mathbf{V}_{\mathbf{hom}}} : \mathcal{X}_{\mathbf{hom}} \to \mathcal{X}_{\mathbf{hom}}$ such that $\ker\{\Pi_{\mathbf{V}_{\mathbf{hom}}}\} = \mathbf{W}_{\mathbf{hom}}$, and choose $\mathbf{V}_{\mathbf{hom}} = \Pi_{\mathbf{V}_{\mathbf{hom}}}(\mathcal{X}_{\mathbf{hom}})$. Let $\underline{\boldsymbol{\xi}} \in \mathcal{X}_{\mathbf{hom}}$ and B_R be a ball of radius R sufficiently large to ensure that $\Sigma \subset B_R$. Solve the following problems in each Ω_j:

$$\begin{array}{ll} -\Delta p_j = 0 & \text{in } \Omega_j \cap B_R, \\ \nabla p_j \cdot \mathbf{n}_j = \mathrm{div}_\Gamma\, \boldsymbol{\xi}_j & \text{on } \Gamma_j \text{ (and, for } j = J\,, \nabla p_J \cdot \mathbf{n}_R = 0 \text{ on } \partial B_R). \end{array} \tag{4.16}$$

We denote by $\underline{\boldsymbol{\varphi}}$ the function defined as $\boldsymbol{\varphi}_j = \nabla p_j$. We deduce then that $\mathrm{div}\, \underline{\boldsymbol{\varphi}} = 0$ on B_R since $\mathrm{div}\, \boldsymbol{\varphi}_j = 0$ in Ω_j and $\boldsymbol{\varphi}_j \cdot \mathbf{n}_j + \boldsymbol{\varphi}_i \cdot \mathbf{n}_i = 0$ on Γ_{ij} $\forall\, i, j$, since $\underline{\boldsymbol{\xi}} \in \mathcal{X}_{\mathbf{hom}}$.

Using, e.g., [1], $\underline{\boldsymbol{\varphi}}$ admits a vector potential in B_R, $\underline{\mathbf{V}} \in \mathbf{H}^1(B_R) \cap \mathbf{H}_0(\mathbf{curl}, B_R)$ verifying

$$\mathbf{curl}\, \underline{\mathbf{V}} = \underline{\boldsymbol{\varphi}}, \quad \mathrm{div}\, \underline{\mathbf{V}} = 0.$$

By continuity we have

$$\|\underline{\mathbf{V}}\|_{\mathbf{H}^1(B_R)} \lesssim \|\underline{\boldsymbol{\varphi}}\|_{\mathbf{L}^2(B_R)} \lesssim \|\,\mathrm{div}_\Gamma\, \underline{\boldsymbol{\xi}}\|_{-1/2,\Sigma}.$$

Set $\Pi_{\mathbf{V}_{\mathbf{hom}}} \underline{\boldsymbol{\xi}} = \gamma_D \underline{\mathbf{V}}$. By construction $\Pi_{\mathbf{V}_{\mathbf{hom}}} : \mathcal{X} \to \mathbf{H}^{1/2}_\times(\Sigma)$ and $\ker\{\Pi_{\mathbf{V}_{\mathbf{hom}}}\} = \mathbf{W}_{\mathbf{hom}}$, and for $\mathbf{v}_j = (\gamma_D \underline{\mathbf{V}})_j$, $\mathrm{div}_\Gamma\, \mathbf{v}_j = (\mathbf{curl}\, \underline{\mathbf{V}})_{|\Omega_j} \cdot \mathbf{n}_j = \boldsymbol{\varphi}_j \cdot \mathbf{n}_j = \mathrm{div}_\Gamma\, \boldsymbol{\xi}_j$. Thus, $\Pi_{\mathbf{V}_{\mathbf{hom}}}$ is a projection and we conclude by observing that $\mathbf{H}^{1/2}_\times(\Sigma)$ is compactly embedded in $\underline{\mathbf{L}}^2_t(\Sigma)$. ◻

We consider an operator $\Theta : \mathcal{X}^2_{\mathbf{hom}} \to \mathcal{X}^2_{\mathbf{hom}}$ which maps $\underline{\boldsymbol{\xi}} \in \mathcal{X}^2_{\mathbf{hom}}$ with Hodge decomposition (Theorem 4.10 applied twice) $\underline{\boldsymbol{\xi}} = \underline{\mathbf{v}} + \underline{\mathbf{w}}$ into $\Theta(\underline{\boldsymbol{\xi}}) = \underline{\mathbf{v}} - \underline{\mathbf{w}}$. The next theorem follows then as an immediate consequence of Theorems 4.8 and 4.10.

Theorem 4.11. *There exists a compact operator $T : \mathcal{X}^2_{hom} \to \mathcal{X}^2_{hom}$ and a constant $\alpha > 0$ such that*

$$\mathrm{Re}\, \underline{B}((\mathbb{A} + T)\underline{\boldsymbol{\xi}}, \Theta(\overline{\underline{\boldsymbol{\xi}}})) \geq \alpha \left\|\underline{\boldsymbol{\xi}}\right\|^2_{\mathcal{X}^2}. \tag{4.17}$$

Proof. We use the definitions of $\underline{B}$ and $\mathbb{A}$:

$$\underline{B}(\mathbb{A}\underline{\boldsymbol{\xi}}, \Theta(\underline{\boldsymbol{\xi}})) = \sum_j B_j(A_j(\mathbf{v}_j + \mathbf{w}_j), \overline{\mathbf{v}}_j - \overline{\mathbf{w}}_j)$$

with $\underline{\boldsymbol{\xi}} = \underline{\mathbf{v}} + \underline{\mathbf{w}}$, $\underline{\mathbf{v}} = (\mathbf{v}_1, \dots, \mathbf{v}_J)$ and $\underline{\mathbf{w}} = (\mathbf{w}_1, \dots, \mathbf{w}_J)$. Applying Theorem 4.8, we know that, for each j, there exists a compact operator $T_j : \mathbf{X}(\Gamma_j)^2 \to \mathbf{X}(\Gamma_j)^2$ such that

$$\operatorname{Re} B_j(A_j + T_j(\mathbf{v}_j + \mathbf{w}_j), \overline{\mathbf{v}}_j - \overline{\mathbf{w}}_j) \geq \|\boldsymbol{\xi}_j\|_{\mathbf{X}(\Gamma_j)^2}. \tag{4.18}$$

Summing (4.18) over j, the statement is proved. □

Since we know that (4.13) admits at most one solution (as a consequence of Theorem 4.7 and Proposition 4.9), this theorem says that (4.13) is in our framework, i.e., it verifies Assumption 1. As a matter of fact, it verifies also Assumption 2.

4.3.2. Discretization and verification of the (GAP) property.

We concentrate now on the discretization of the problem (4.13). First we construct a compatible mesh on the interfaces Σ in the following way: Consider that there exists a triangulation $\mathcal{T}_h(B_R)$ of a ball B_R containing Σ and such that Σ is composed only of faces of the underlying triangulation $\mathcal{T}_h(B_R)$, i.e., Σ does not cut elements of the mesh. This automatically generates a compatible triangulation of Σ, $\mathcal{T}_h(\Sigma)$ and of course also triangulations of Γ_j's, that we denote by $\mathcal{T}_h(\Gamma_j)$.

Each single space $\mathbf{X}(\Gamma_j)$ is discretized by means of Raviart–Thomas finite elements of degree k defined on $\mathcal{T}_h(\Gamma_j)$, [36], [22]. We denote the discrete spaces by $\mathbf{X}_h(\Gamma_j)$.

Now, the discrete counterpart of $\mathcal{X}$ is $\mathcal{X}_h = \mathbf{X}_h(\Gamma_1) \times \cdots \times \mathbf{X}_h(\Gamma_J)$, and the discrete counterpart of $\mathcal{X}_{\mathbf{hom}}$ is

$$(\mathcal{X}_h)_{\mathbf{hom}} = \mathcal{X}_h \cap \mathcal{X}_{\mathbf{hom}}. \tag{4.19}$$

Note that, thanks to the fact that the nonempty intersections $\Gamma_{ij} = \Gamma_i \cap \Gamma_j$ are discretized by means of only one triangulation, namely, $\mathcal{T}_h(\Sigma)_{|\Gamma_{ij}}$, the space $(\mathcal{X}_h)_{\mathbf{hom}}$ is well defined as a constrained subspace of $\mathcal{X}_h$. We are then ready to state the theorem.

Theorem 4.12. *There exists a splitting $(\mathcal{X}_h)_{\boldsymbol{hom}} = (\mathbf{V}_h)_{\boldsymbol{hom}} \oplus (\mathbf{W}_h)_{\boldsymbol{hom}}$ such that*

$$\delta((\mathbf{V}_h)_{\boldsymbol{hom}}, \mathbf{V}_{\boldsymbol{hom}}) \to 0 \textit{ when } h \to 0 \qquad \textit{and} \qquad \delta((\mathbf{W}_h)_{\boldsymbol{hom}}, \mathbf{W}_{\boldsymbol{hom}}) = 0.$$

Proof. We assume for the sake of simplicity that each Ω_j is connected and simply connected.

We set $(\mathbf{W}_h)_{\mathbf{hom}} = \mathbf{W}_{\mathbf{hom}} \cap \mathcal{X}_h$. An alternate definition is the following: Set $\mathbf{W}_h(\Gamma_j) = \{\boldsymbol{\lambda}_h \in \mathbf{X}_h(\Gamma_j) \ : \ \operatorname{div}_\Gamma \boldsymbol{\lambda}_h = 0\}$. $\mathbf{W}_h(\Gamma_j)$ is the finite elements space of $\mathbf{curl}_\Gamma\, \mathcal{P}_h$, where $\mathcal{P}_h$ is the space of continuous piecewise polynomials of degree k [22], [11], i.e., characterized by degrees of freedom attached to vertices, edges, and triangles on Γ_j. Then we set $\mathbf{W}_h = \mathbf{W}_h(\Gamma_1) \times \cdots \times \mathbf{W}_h(\Gamma_J)$; this space has for each vertex (or edge, or triangle) belonging to an intersection Γ_{ij} two sets of independent degrees of freedom, one defining $\mathbf{W}_h(\Gamma_i)$ and the other $\mathbf{W}_h(\Gamma_j)$. Now, in $(\mathbf{W}_h)_{\mathbf{hom}} = \mathbf{W}_h \cap \mathcal{X}_{\mathbf{hom}}$, the degrees of freedom belonging to these two sets are constrained to be equal for each vertex, edge, or triangle.

Now we construct the supplement $(\mathbf{V}_h)_{\mathbf{hom}}$. We use two intermediate finite elements spaces: the Nédélec edge elements of degree k, $\mathbf{X}_h(B_R)$, as introduced in section 4.1, but on B_R (with vanishing tangential component on ∂B_R), and the Raviart–Thomas finite elements of degree k, $\mathbf{Y}_h(B_R)$, with vanishing normal component on ∂B_R. We construct local Nédélec and Raviart–Thomas elements as $\mathbf{X}_h(\Omega_j) =$

$\mathbf{X}_h(B_R)_{|\Omega_j}$ and $\mathbf{Y}_h(\Omega_j) := \mathbf{Y}_h(B_R)_{|\Omega_j}$ (and $\mathbf{X}_h(\Omega_J) = \mathbf{X}_h(B_R)_{\Omega_J \cap B_R}$, $\mathbf{Y}_h(\Omega_J) := \mathbf{Y}_h(B_R)_{|\Omega_J \cap B_R}$). We refer to [36], [6], [22] for suitable definitions and properties. We construct the space $(\mathbf{V}_h)_{\mathbf{hom}}$ as we did for $\mathbf{V}_{\mathbf{hom}}$ in the proof of Theorem 4.10, i.e., by constructing a projection operator $\Pi : (\mathcal{X}_h)_{\mathbf{hom}} \to (\mathcal{X}_h)_{\mathbf{hom}}$ with $\ker\{\Pi\} = (\mathbf{W}_h)_{\mathbf{hom}}$.

Then let $\underline{\boldsymbol{\lambda}}_h \in (\mathcal{X}_h)_{\mathbf{hom}}$, $\underline{\boldsymbol{\lambda}}_h = (\boldsymbol{\lambda}_{1,h}, \ldots, \boldsymbol{\lambda}_{J,h})$, $\boldsymbol{\lambda}_{j,h} \in \mathbf{X}_h(\Gamma_j)$. We solve the following:

$$\text{Find } \boldsymbol{\Xi}_{j,h} \in \mathbf{Y}_h(\Omega_j) \ : \begin{cases} \operatorname{div} \boldsymbol{\Xi}_{j,h} = 0 & \text{on } \Omega_j, \\ \int_{\Omega_j} \boldsymbol{\Xi}_{j,h} \, \mathbf{curl}\, \boldsymbol{\chi}_h = 0 & \forall\, \boldsymbol{\chi}_h \in \mathbf{X}_h(\Omega_j), \\ \boldsymbol{\Xi}_{j,h} \cdot \mathbf{n}_j = \operatorname{div}_\Gamma \boldsymbol{\lambda}_{j,h} & \text{on } \Gamma_j \,. \end{cases}$$

This problem is solvable since $\int_{\Gamma_j} \operatorname{div}_\Gamma \boldsymbol{\lambda}_{j,h} = 0$; moreover, it is uniquely solvable [6]. Now, since $\underline{\boldsymbol{\lambda}}_h \in (\mathcal{X}_h)_{\mathbf{hom}}$, by construction the vector $\boldsymbol{\Xi}_h$, $\boldsymbol{\Xi}_{h|\Omega_j} := \boldsymbol{\Xi}_{j,h}$, belongs to $\mathbf{Y}_h(B_R)$. Now we solve the discrete and continuous vector potential problem:

$$\text{Find } \boldsymbol{\Psi}_h \in \mathbf{X}_h(B_R) \ : \quad \begin{cases} \mathbf{curl}\, \boldsymbol{\Psi}_h = \boldsymbol{\Xi}_h & \text{on } B_R, \\ \int_{B_R} \boldsymbol{\Psi}_h \cdot \mathbf{w}_h = 0 & \forall\, \mathbf{w}_h \in \mathbf{X}_h(B_R)\,,\ \mathbf{curl}\, \mathbf{w}_h = 0, \end{cases}$$

$$\text{Find } \boldsymbol{\Psi} \in \mathbf{H}_0(\mathbf{curl}, B_R) \ : \ \begin{cases} \mathbf{curl}\, \boldsymbol{\Psi} = \boldsymbol{\Xi}_h & \text{on } B_R, \\ \operatorname{div} \boldsymbol{\Psi} = 0 & \text{on } B_R. \end{cases}$$

These problems are uniquely solvable [1]; moreover, the continuity estimate

$$\|\boldsymbol{\Psi}\|_{\mathbf{H}^1(B_R)} + \|\boldsymbol{\Psi}_h\|_{\mathbf{H}(\mathbf{curl}, B_R)} \lesssim \|\boldsymbol{\Xi}_h\|_{L^2(B_R)} \lesssim \sum_{j=1}^{J} \| \operatorname{div}_\Gamma \boldsymbol{\lambda}_j\|_{H^{-1/2}(\Gamma_j)} \tag{4.20}$$

holds.

Now the proof is basically finished. Denote by $\boldsymbol{\Psi}_{j,h}$ ($\boldsymbol{\Psi}_j$, resp.) the restriction of $\boldsymbol{\Psi}_h$ ($\boldsymbol{\Psi}$, resp.) to Ω_j (for $j = J$ to $\Omega_J \cap B_R$) $\forall\, j$.

It is enough to set $\Pi \underline{\boldsymbol{\lambda}}_h = \underline{\boldsymbol{\lambda}}_h^{\mathbf{V}} := (\gamma_D(\boldsymbol{\Psi}_{1,h}), \ldots, (\boldsymbol{\Psi}_{J,h}))$; by construction,

$$\operatorname{div}_\Gamma \boldsymbol{\lambda}_{j,h}^{\mathbf{V}} = \mathbf{curl}\, \boldsymbol{\Psi}_{j,h|\Gamma_j} \cdot \mathbf{n}_j = \boldsymbol{\Xi}_{j,h} \cdot \mathbf{n}_j = \operatorname{div}_\Gamma \boldsymbol{\lambda}_{j,h}.$$

Thus, $\ker\{\Pi\} = (\mathbf{W}_h)_{\mathbf{hom}}$. Set $\underline{\boldsymbol{\lambda}}^{\mathbf{V}} = (\gamma_D(\boldsymbol{\Psi}_1), \ldots, \gamma_D(\boldsymbol{\Psi}_J))$; by construction, $\underline{\boldsymbol{\lambda}}^{\mathbf{V}} \in \mathbf{V}_{\mathbf{hom}}$. It is the candidate in $\mathbf{V}_{\mathbf{hom}}$ that is needed to verify the (GAP) property:

$$\begin{aligned} \|\underline{\boldsymbol{\lambda}}^{\mathbf{V}} - \underline{\boldsymbol{\lambda}}_h^{\mathbf{V}}\|_{\mathcal{X}} &\lesssim \|\boldsymbol{\Psi} - \boldsymbol{\Psi}_h\|_{\mathbf{H}(\mathbf{curl}, B_R)} \lesssim \|\boldsymbol{\Psi} - \boldsymbol{\Psi}_h\|_{L^2(B_R)} \\ &\lesssim h \|\boldsymbol{\Psi}\|_{\mathbf{H}^1(B_R)} \lesssim h \|\boldsymbol{\Xi}_h\|_{L^2(B_R)} \\ &\lesssim h \sum_{j=1}^{J} \| \operatorname{div}_\Gamma \boldsymbol{\lambda}_{j,h}\|_{H^{-1/2}(\Gamma_j)} \lesssim h \|\underline{\boldsymbol{\lambda}}_h^{\mathbf{V}}\|_{\mathcal{X}}, \end{aligned} \tag{4.21}$$

which proves that $\delta((\mathbf{V}_h)_{\mathbf{hom}}, \mathbf{V}_{\mathbf{hom}}) \lesssim h$. Note that the forth estimate in (4.21) comes from the (GAP) property for Nédélec finite elements, which has been proved in section 4.1. □

COROLLARY 4.13. *The following Galerkin problem admits a unique solution when h is sufficiently small: Find* $\underline{\boldsymbol{\xi}}_h \in (\mathcal{X}_h)^2_{\mathit{hom}}$ *such that*

$$\underline{B}(\mathbb{A} \underline{\boldsymbol{\xi}}_h, \underline{\boldsymbol{\lambda}}_h) = \underline{B}((\tfrac{1}{2}\mathbb{I} - \mathbb{A})\underline{\boldsymbol{\xi}}^0, \underline{\boldsymbol{\lambda}}_h) \qquad \forall\, \underline{\boldsymbol{\lambda}}_h \in (\mathcal{X}_h)^2_{\mathit{hom}}, \tag{4.22}$$

Moreover, let $\underline{\boldsymbol{\xi}}_{hom} \in \mathcal{X}^2_{hom}$ be the solution of (4.13). *Then it holds that*

$$\|\underline{\boldsymbol{\xi}}_{hom} - \underline{\boldsymbol{\xi}}_h\|_{\mathcal{X}^2} \lesssim \inf_{\underline{\boldsymbol{\lambda}}_h \in (\mathcal{X}_h)^2_{hom}} \|\underline{\boldsymbol{\xi}}_{hom} - \underline{\boldsymbol{\lambda}}_h\|_{\mathcal{X}^2}.$$

REFERENCES

[1] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[2] D. Boffi, *Fortin operator and discrete compactness for edge elements*, Numer. Math., 87 (2000), pp. 229–246.

[3] D. Boffi, *A note on the de Rham complex and a discrete compactness property*, Appl. Math. Lett., 14 (2001), pp. 33–38.

[4] D. Boffi, F. Brezzi, and L. Gastaldi, *On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form*, Math. Comp., 69 (1999), pp. 121–140.

[5] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia, *Computational models of electromagnetic resonators: Analysis of edge element approximation*, SIAM J. Numer. Anal., 36 (1999), pp. 1264–1290.

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[7] A. Buffa, *Traces theorems for functional spaces related to Maxwell equations: An overview*, in Computational Electromagnetics, Lectures Notes in Comput. Sci. Engrg. 28, C. Carstensen et al., eds., Springer, Berlin, Germany, 2003, pp. 23–34.

[8] A. Buffa and S. Christiansen, *The electric field integral equation on Lipschitz screens: Definitions and numerical approximation*, Numer. Math., 94 (2003), pp. 229–267.

[9] A. Buffa and P. Ciarlet, Jr., *On traces for functional spaces related to Maxwell's equations. Part* I: *An integration by parts formula in Lipschitz polyhedra*, Math. Methods Appl. Sci., 21 (2001), pp. 9–30.

[10] A. Buffa and P. Ciarlet, Jr., *On traces for functional spaces related to Maxwell's equations. Part* II: *Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 21 (2001), pp. 31–48.

[11] A. Buffa, M. Costabel, and C. Schwab, *Boundary element methods for Maxwell equations in non-smooth domains*, Numer. Math., 92 (2002), pp. 679–710.

[12] A. Buffa, M. Costabel, and D. Sheen, *On traces for* $\mathbf{H}(\mathbf{curl}, \Omega)$ *for Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–876.

[13] A. Buffa and R. Hiptmair, *Galerkin boundary element methods for electromagnetic scattering*, in Topics in Computational Wave Propagation, Lecture Notes in Comput. Sci. Engrg. 31, Springer-Verlag, Berlin, 2003, pp. 83–124.

[14] A. Buffa, R. Hiptmair, T. von Petersdorff, and C. Schwab, *Boundary element methods for Maxwell transmission problems in Lipschitz domains*, Numer. Math., 95 (2003), pp. 459–485.

[15] A. Buffa and T. von Peterdorff, *Boundary Element Methods for Maxwell Equations in Complicated Domains*, Tech. report IMATI-CNR, 2003.

[16] S. Caorsi, P. Fernandes, and M. Raffetto, *On the convergence of Galerkin finite element approximations of electromagnetic eigenproblems*, SIAM J. Numer. Anal., 38 (2000), pp. 580–607.

[17] S. H. Christiansen, *Discrete Fredholm properties and convergence estimates for the electric field integral equation*, Math. Comp., 73 (2004), pp. 143–167.

[18] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[19] P. Ciarlet, Jr. and J. Zou, *Fully discrete finite element approaches for time-dependant Maxwell's equations*, Numer. Math., 82 (1999), pp. 193–219.

[20] J. Descloux, N. Nassif, and J. Rappaz, *On spectral approximation.* I. *The problem of convergence*, RAIRO Anal. Numér., 12 (1978), pp. 97–112, iii.

[21] J. Descloux, N. Nassif, and J. Rappaz, *On spectral approximation.* II. *Error estimates for the Galerkin method*, RAIRO Anal. Numér., 12 (1978), pp. 113–119, iii.

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

[23] S. Hildebrandt and E. Wienholtz, *Constructive proofs of representation theorems in separable Hilbert space*, Comm. Pure Appl. Math., 17 (1964), pp. 369–373.

[24] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.

[25] R. Hiptmair, *Coupling of finite elements and boundary elements in electromagnetic scattering*, SIAM J. Numer. Anal., 41 (2003), pp. 919–944.

[26] R. Hiptmair and C. Schwab, *Natural boundary element methods for the electric field integral equation on polyhedra*, SIAM J. Numer. Anal., 40 (2002), pp. 66–86.

[27] G. C. Hsiao, P. B. Monk, and N. Nigam, *Error analysis of a finite element–integral equation scheme for approximating the time-harmonic Maxwell system*, SIAM J. Numer. Anal., 40 (2002), pp. 198–219.

[28] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1976.

[29] F. Kikuchi, *On a discrete compactness property for the Nédélec finite elements*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 36 (1989), pp. 479–490.

[30] F. Kikuchi, *Numerical analysis of electromagnetic problems*, in Mathematical Modeling and Numerical Simulation in Continuum Mechanics (Yamaguchi, 2000), Lecture Notes Comput. Sci. Engrg. 19, Springer-Verlag, Berlin, 2002, pp. 109–124.

[31] P. Monk, *A finite element method for approximating the time-harmonic Maxwell equations*, Numer. Math., 63 (1992), pp. 243–261.

[32] P. Monk, *Finite Element Methods for Maxwell's Equations*, Numerical Mathematics and Scientific Computation, Oxford University Press, New York, 2003.

[33] P. Monk and L. Demkowicz, *Discrete compactness and the approximation of Maxwell's equations in* $\mathbb{R}^3$, Math. Comp., 70 (2001), pp. 507–523.

[34] C. Müller, *Foundations of the Mathematical Theory of Electromagnetic Waves*, Springer-Verlag, New York, Heidelberg, 1969.

[35] J. Nédélec, *Mixed finite element in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[36] P. Raviart and J. Thomas, *Primal hybrid finite element methods for second order elliptic problems*, Math. Comput., 31 (1977), pp. 391–413.

[37] T. von Petersdorff, *Boundary integral equations for mixed Dirichlet, Neumann and transmission problems*, Math. Methods Appl. Sci., 11 (1989), pp. 185–213.

[38] T. von Petersdorff, *Randwertprobleme der Elastizitätstheorie für Polyeder Singularitäten and Approximation mit Randelementmethoden*, Ph.D. thesis, Darmstadt, 1989.

[39] C. Weber, *A local compactness theorem for Maxwell's equations*, Math. Methods Appl. Sci., 2 (1980), pp. 12–25.

[40] N. Weck and K. J. Witsch, *Low-frequency asymptotics for dissipative Maxwell's equations in bounded domains*, Math. Methods Appl. Sci., 13 (1990), pp. 81–93.

SIAM J. Numer. Anal.
Vol. 43, No. 1, pp. 19–40

ON THE ERROR OF LINEAR INTERPOLATION AND THE ORIENTATION, ASPECT RATIO, AND INTERNAL ANGLES OF A TRIANGLE*

WEIMING CAO†

Abstract. In this paper, we attempt to reveal the precise relation between the error of linear interpolation on a general triangle and the geometric characters of the triangle. Taking the model problem of interpolating quadratic functions, we derive two exact formulas for the H^1-seminorm and L^2-norm of the interpolation error in terms of the area, aspect ratio, orientation, and internal angles of the triangle. These formulas indicate that (1) for highly anisotropic triangular meshes the H^1-seminorm of the interpolation error is almost a monotonically decreasing function of the angle between the orientations of the triangle and the function; (2) maximum angle condition is not essential if the mesh is aligned with the function and the aspect ratio is of magnitude $\sqrt{|\lambda_1/\lambda_2|}$ or less, where λ_1 and λ_2 are the eigenvalues of the Hessian matrix of the function. With these formulas we identify the optimal triangles, which produce the smallest H^1-seminorm of the interpolation error, to be the acute isosceles aligned with the solution and of an aspect ratio about $0.8|\frac{\lambda_1}{\lambda_2}|$. The L^2-norm of the interpolation error depends on the orientation and the aspect ratio of the triangle, but not directly on its maximum or minimum angles. The optimal triangles for the L^2-norm are those aligned with the solution and of an aspect ratio $\sqrt{|\lambda_1/\lambda_2|}$. These formulas can be used to formulate more accurate mesh quality measures and to derive tighter error bounds for interpolations.

Key words. anisotropic mesh, linear interpolation, aspect ratio, mesh alignment, maximum angle condition

AMS subject classifications. 65D05, 65L50, 65N15, 65N50

DOI. 10.1137/S0036142903433492

1. Introduction.

It is well known that on quasi-uniform meshes the accuracy of piecewise linear interpolation is first order in the H^1-norm. More precisely, denote by $H^m(\Omega)$ the usual Sobolev spaces of order m on a bounded domain $\Omega \in \mathcal{R}^2$. For any $u \in H^2(\Omega)$, the H^1-norm of the error between u and its piecewise linear interpolation u_I can be bounded by

$$\|u - u_I\|_{H^1(\Omega)} \leq ch|u|_{H^2(\Omega)},$$

where h is the diameter of the triangles in the mesh, and c is a constant independent of h and u.

If the mesh is not quasi-uniform, then the error contributed from each triangle K can be bounded by

$$\|u - u_I\|_{H^1(K)} \leq c\frac{h^2}{\rho}|u|_{H^2(K)},$$

where ρ is the diameter of the largest inscribed circle in K. This error bound guarantees the H^1-norm of the error converge to 0 as $h \to 0$, as long as $\frac{h}{\rho}$ remains bounded. This is equivalent to requiring that the minimum angle of K is bounded from 0 [5].

*Received by the editors August 19, 2003; accepted for publication (in revised form) May 7, 2004; published electronically April 26, 2005. This research was supported in part by the NSF under grant DMS-0209313.

http://www.siam.org/journals/sinum/43-1/43349.html

†Department of Applied Mathematics, University of Texas at San Antonio, San Antonio, TX 78249 (wcao@math.utsa.edu).

However, this error estimate is not tight when the triangle has only one small angle. Babuška and Aziz [2] showed that the minimum angle condition is actually not essential for the convergence of linear interpolation. They improved the error bound as

$$\|u - u_I\|_{H^1(K)} \le \Gamma(\alpha) h |u|_{H^2(K)},$$

where $\Gamma(\alpha)$ is an increasing function of the maximum angle α of K. Therefore, in order to guarantee the convergence it is only required that the maximum angle be bounded from π. This is the well-known maximum angle condition. This condition is necessary and sufficient for the convergence of the linear interpolation process over the class of functions of $H^2(K)$.

In adaptive computation, one often knows a priori or a posteriori some information of the functions to be approximated. This information can be used to restrict the set of functions considered and to avoid the divergence case although the triangle has a maximum angle close to π. Indeed, long and thin triangles violating the maximum angle condition have been used successfully in engineering, particularly in computational fluid dynamics [1, 15]. For problems with very different length scales in different spatial directions, long and thin triangles turn out to be better choices than shape regular ones if they are properly used. This motivated an intensive study on the error analysis for anisotropic meshes in the finite element method (FEM). For instance, Apel [1] described an error estimate in terms of the length scales h_1 and h_2 along the x and y directions, respectively,

$$|u - u_I|_{W^{m,p}(K)} \le c \sum_{\alpha_1+\alpha_2=\ell-m} h_1^{\alpha_1} h_2^{\alpha_2} |\frac{\partial^{\ell-m} u}{\partial x^{\alpha_1} \partial y^{\alpha_2}}|_{W^{m,p}(K)}, \qquad m = 0, 1,$$

where $W^{m,p}(K)$ is the Sobolev space of functions whose up to mth order derivatives are L^p-integratable. If u and K are aligned and the maximum angle condition is satisfied, this estimate is asymptotically accurate, i.e., both sides of the above inequality have the same h_i order. But when the maximum angle of K approaches π, the ratio of the error bound over the actual error norm goes to infinity [7]. Formaggia and Perotto [7] presented another type of estimate of the H^1-norm for the interpolation error based on the eigendecomposition of the affine mapping from a standard element to K. Their estimate is accurate when u and K are aligned and the maximum angle of K is close to π. However, the ratio of their estimate over the actual error norm goes to infinity when two angles of K approach $\frac{\pi}{2}$.

More recent error analyses have been based on the overall mesh properties and the behavior of the approximation functions. For instance, Berzins [4] developed a mesh quality indicator measuring the correlation between the anisotropic features of the mesh and those of the solutions. Kunert [11] proposed a so-called matching function to quantify how good overall a mesh is for a specific solution. Huang [8] introduced measures for three aspects of the mesh qualities—aspect ratio, alignment, and adaptation—and an overall quality mesh measure based on them. He formulated the error bounds in terms of these measures and proposed a variational formulation to optimize the overall mesh quality measure to control the interpolation error. A similar idea was used by Huang and Sun [9] in formulating the monitor function in variational mesh adaptation.

Needless to say, the interpolation error depends on the solution and the size and shape of the elements in the mesh. Understanding this relation is crucial for the generation of efficient and effective meshes for the FEM. However, in all the error estimates for anisotropic meshes, the relation between the error and the geometric characters of a triangle, such as the alignment, aspect ratio, and internal angles, has not been revealed explicitly. In the mesh generation community, this relation is studied more closely for the model problem of interpolating quadratic functions. This model is a reasonable simplification of the cases involving general functions, since quadratic functions are the leading terms in the local expansion of the linear interpolation errors. For instance, Nadler [12] derived an exact expression for the L^2-norm of the linear interpolation error in terms of the three sides $\boldsymbol{\ell}_1, \boldsymbol{\ell}_2$, and $\boldsymbol{\ell}_3$ of the triangle K,

$$\|u - u_I\|^2_{L^2(K)} = \frac{|K|}{180}[(d_1 + d_2 + d_3)^2 + d_1 d_2 + d_2 d_3 + d_1 d_3],$$

where $|K|$ is the area of the triangle, $d_i = \boldsymbol{\ell}_i \cdot H \boldsymbol{\ell}_i$ with H being the Hessian matrix of u. Bank and Smith [3] gave a similar formula for the H^1-seminorm of the linear interpolation error; see (8) in section 3. Assuming $u = \lambda_1 x^2 + \lambda_2 y^2$, D'Azevedo and Simpson [6] derived the exact formula for the maximum norm of the interpolation error

$$\|u - u_I\|^2_{L^\infty(K)} = \frac{D_{12} D_{23} D_{31}}{16 \lambda_1 \lambda_2 |K|^2},$$

where $D_{ij} = \boldsymbol{\ell}_i \cdot \mathrm{diag}(\lambda_1, \lambda_2) \boldsymbol{\ell}_i$. Based on the geometric interpretation of this formula, they proved that for a fixed area the optimal triangle, which produces the smallest maximum interpolation error, is the one obtained by compressing an equilateral triangle by factors $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$ along the two eigenvectors of the Hessian matrix of u. Furthermore, the optimal incidence for a given set of interpolation points is the Delaunay triangulation based on the stretching map (by factors $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$ along the two eigenvector directions) of the grid points. Rippa [13] showed that the mesh obtained this way is also optimal for the L^p-norm of the error for any $1 \le p \le \infty$.

Thought these formulas are exact, they do not describe explicitly the relation between the error and the geometric characters of the triangle. In this paper, we attempt to reveal this relation explicitly and precisely. Taking the model problem of linear interpolation of quadratic functions, we derive two exact expressions for the H^1-seminorm and L^2-norm of the interpolation error in terms of the area, aspect ratio, alignment direction, and internal angles of the triangle. From these formulas the effects of the geometric characters of a triangle can be clearly identified. They indicate that (1) for highly anisotropic triangular meshes the H^1-seminorm of the interpolation error is almost a monotonically decreasing function of the angle between the orientation of the triangle and the orientation of the function; (2) maximum angle condition is critical if the mesh is not aligned with the function or if the aspect ratio is larger than $\sqrt{|\lambda_1/\lambda_2|}$; (3) if the triangles are aligned with the function and the aspect ratio is about $\sqrt{|\lambda_1/\lambda_2|}$ or less, then the error is not sensitive to the maximum angle of the triangle. Also, we can easily identify the best triangles which produce the smallest H^1- or L^2-norm of the interpolation error. It turns out that in the sense of the H^1-seminorm, the optimal triangle with a given area is the acute isosceles aligned with the function and of the aspect ratio about $0.8|\frac{\lambda_1}{\lambda_2}|$. The L^2-norm of the interpolation error depends on the alignment and the aspect ratio of the triangle, but not directly on its maximum or minimum angles. The optimal triangles for the L^2-norm are those aligned with the solution and of an aspect ratio $\sqrt{|\lambda_1/\lambda_2|}$.

The organization of this paper is as follows. In section 2 we give a precise definition of the orientation and the aspect ratio of a triangle by using the singular value decomposition (SVD) of the affine mapping from a standard element to the triangle. In section 3 we derive the exact formulas for the H^1-seminorm and L^2-norm of the interpolation error for the quadratic functions. Then we identify in section 4 a number of special cases that are of interests to the optimal design of meshes. Optimal choices of the aspect ratio and the alignment direction are discussed here. In section 5 we elaborate in more detail the effects of various geometric characters of a triangle on the interpolation error. Finally we demonstrate by an example the accuracy of the linear interpolation errors on different meshes.

2. Mapping and the geometric characters of a triangle. Let K be a triangle in the xy-plane, and let $\boldsymbol{x}_1, \boldsymbol{x}_2$, and $\boldsymbol{x}_3$ be the vertices of K. Denote by $\boldsymbol{\ell}_i$ the vector of the side opposite to $\boldsymbol{x}_i$ (in counterclockwise direction). Let $\hat{K}$ be the equilateral triangle with the vertices

$$\boldsymbol{\xi}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \boldsymbol{\xi}_2 = \begin{bmatrix} -\frac{\sqrt{3}}{2} \\ -\frac{1}{2} \end{bmatrix}, \quad \boldsymbol{\xi}_3 = \begin{bmatrix} \frac{\sqrt{3}}{2} \\ -\frac{1}{2} \end{bmatrix}.$$

The three sides of $\hat{K}$ are $\boldsymbol{e}_i = \sqrt{3}[\cos(2(i-1)\pi/3), \sin(2(i-1)\pi/3)]^T, i = 1, 2, 3$. See Figure 1.

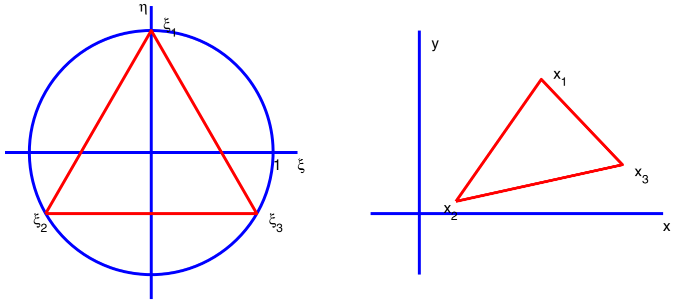


Fig. 1. *Standard element $\hat{K}$ and physical element K.*

Let $\boldsymbol{x}_c = \frac{1}{3}(\boldsymbol{x}_1 + \boldsymbol{x}_2 + \boldsymbol{x}_3)$ be the center of K. The affine mapping (which maps $\boldsymbol{\xi}_i$ to $\boldsymbol{x}_i$) from $\hat{K}$ to K can be expressed as $\boldsymbol{x} = M\boldsymbol{\xi} + \boldsymbol{x}_c$ with

$$M = \left(\frac{1}{\sqrt{3}}(\boldsymbol{x}_1 - \boldsymbol{x}_3), \boldsymbol{x}_2 - \boldsymbol{x}_c \right).$$

Denote by $M = U\Sigma V^*$ the SVD of the 2×2 matrix M. Without loss of generality, we assume $\Sigma = \text{diag}(\sigma_1, \sigma_2)$ with $\sigma_1 \geq \sigma_2 > 0$, and

$$U = R_{\phi_u} = \begin{bmatrix} \cos\phi_u & -\sin\phi_u \\ \sin\phi_u & \cos\phi_u \end{bmatrix}, \quad V = R_{\phi_v} = \begin{bmatrix} \cos\phi_v & -\sin\phi_v \\ \sin\phi_v & \cos\phi_v \end{bmatrix}.$$

R_{ϕ_u} and R_{ϕ_v} represent the linear transform of rotation (counterclockwise) by angles ϕ_u and ϕ_v, respectively.

The mapping $\boldsymbol{x} = M\boldsymbol{\xi}$ maps $\hat{K}$ into a triangle centered at the origin. Its effect can be understood as the composition of three operations (see, e.g., [16]): (1) rotation clockwise by angle ϕ_v; then (2) stretching by factors σ_1 and σ_2 in ξ and η directions, respectively; and finally (3) rotation counterclockwise by angle ϕ_u. A circle (centered at (0,0)) in the $\xi\eta$-plane will be mapped into an ellipse in the xy-plane with two axes σ_1 and σ_2, and the longer axis forms an angle ϕ_u with the x-axis. See Figure 2.

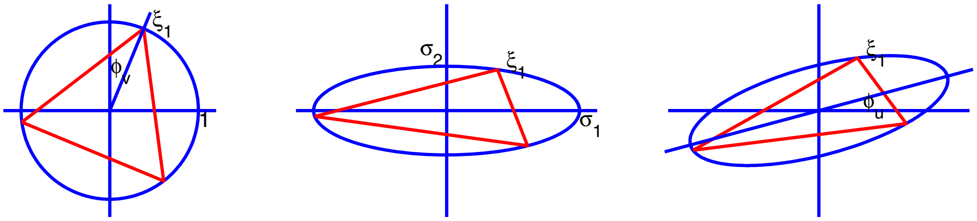


FIG. 2. *From $\hat{K}$ to K under the mapping M: $R^*_{\phi_v}\hat{K}$, $\Sigma R^*_{\phi_v}\hat{K}$, $R_{\phi_u}\Sigma R^*_{\phi_v}\hat{K}$.*

We define the aspect ratio of the triangle K as

$$r_{12} = \frac{\sigma_1}{\sigma_2} \tag{1}$$

and define the orientation of K as the direction of angle ϕ_u with the x-axis.

It can be seen that K is equilateral if and only if $\sigma_1 = \sigma_2$ and that K is isosceles if and only if $\phi_v = \frac{m\pi}{6}$ with integer m or $\sigma_1 = \sigma_2$. Together with the aspect ratio, ϕ_v determines the three internal angles of the triangle. For a fixed aspect ratio, the maximum internal angle of K is a periodic even function of ϕ_v with period $\frac{\pi}{3}$, and it is decreasing in $(0, \frac{\pi}{6})$. K is an obtuse isosceles triangle when $\phi_v = \frac{m\pi}{3}$ and an acute isosceles triangle when $\phi_v = \frac{(2m+1)\pi}{6}$. If K is an obtuse/acute isosceles, then its orientation is in parallel/perpendicular to its base. Furthermore, let b and h be the length of the base and the height over the base of an isosceles triangle; then

$$\left\{\begin{array}{lllll} \sigma_1 = \frac{\sqrt{3}}{3}b, & \sigma_2 = \frac{2}{3}h, & r_{12} = \frac{\sqrt{3}}{2}\frac{b}{h}, & \phi_v = 0 & \text{when } h \le \frac{\sqrt{3}}{2}b, \\ \sigma_1 = \frac{2}{3}h, & \sigma_2 = \frac{\sqrt{3}}{3}b, & r_{12} = \frac{2}{\sqrt{3}}\frac{h}{b}, & \phi_v = \frac{\pi}{6} & \text{when } h > \frac{\sqrt{3}}{2}b. \end{array}\right.$$

For general triangles, the aspect ratio r_{12} can be found as follows. Let

$$q(K) = \frac{1}{2}\left(\frac{\sigma_1}{\sigma_2} + \frac{\sigma_2}{\sigma_1}\right) \in [1, \infty).$$

This quantity can be calculated from the three sides and the area of K as (see formula (10) below)

$$q(K) = \frac{\sigma_1^2 + \sigma_2^2}{2\sigma_1\sigma_2} = \frac{|\boldsymbol{\ell}_1|^2 + |\boldsymbol{\ell}_2|^2 + |\boldsymbol{\ell}_3|^2}{4\sqrt{3}|K|}. \tag{2}$$

Therefore,

$$r_{12} = q + \sqrt{q^2 - 1}.$$

Clearly, the closer r_{12} is to 1, the closer K is to being equilateral. The same is true for the quantity $q(K)$. Therefore, both r_{12} and $q(K)$ can be used to measure the closeness of a triangle to being equilateral. Indeed, the reciprocal of the right-hand-side expression in (2) was used by Bank and Smith [3] as the "shape regularity quantity" of a triangle. $q(K)$ can also be expressed in terms of the matrix M directly. Note that σ_1 and σ_2 are eigenvalues of $(M^T M)^{1/2}$; it is easy to see that

$$q(K) = \frac{\|M\|_F^2}{2\ \det(M)},$$

where $\|\cdot\|_F$ stands for the Frobenius norm of a matrix. This formula was used by Knupp, Margolin, and Shashkov [10] in certain functionals characterizing the smoothness of the mesh.

There are several other ways to define the aspect ratio and the orientation of a triangular element in the finite element analysis. For instance, the aspect ratio is usually defined as the ratio of the length of the longest side over the perpendicular distance from it to the opposite vertex, or as the ratio of the diameter of the triangle over the diameter of the largest inscribed circle in the triangle. The orientation can be defined as the direction of the longest side. It is easy to see that these definitions are equivalent to (1) up to some bounded constants. However, they are not precise enough for describing accurately the behavior of the interpolation errors.

3. Formulas for H^1- and L^2- norms of the interpolation error. Without loss of generality, assume K is a triangle with its center at the origin and its orientation being the x-axis, i.e., $\boldsymbol{x}_c = \boldsymbol{0}, \phi_u = 0$. We study the H^1- and L^2-norms of the error for the linear interpolation of a quadratic function u over K. Since the first order terms of u have no contribution to the error of linear interpolation, we assume in particular that $u = \frac{1}{2}\boldsymbol{x} \cdot H\boldsymbol{x}$, where H is the Hessian of u. Since H is a 2×2 symmetric matrix, we may decompose it into

$$H = R_{\phi_h} \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix} R_{\phi_h}^T, \tag{3}$$

where R_{ϕ_h} is the matrix of rotation by an angle ϕ_h. We also assume that $\lambda_1 \geq |\lambda_2|$. Other cases can be covered by simply considering the function $-u$. It is easy to see that the contour lines of u are concentrical ellipses (when $\lambda_1\lambda_2 > 0$) or hyperbolas (when $\lambda_1\lambda_2 < 0$). Their axes are multiples of $1/\sqrt{\lambda_1}$ and $1/\sqrt{\lambda_2}$, and the longer axis (with λ_2) is of angle ϕ_h with the y-axis. We define this direction as the orientation of function u. When $\phi_h = \frac{\pi}{2}$, u is oriented along the x-axis, i.e., in the same direction as triangle K. In this case, we say K and u are aligned. When $\phi_h = 0$, u is oriented along the y-axis, i.e., perpendicular to the orientation of K. See Figure 3.

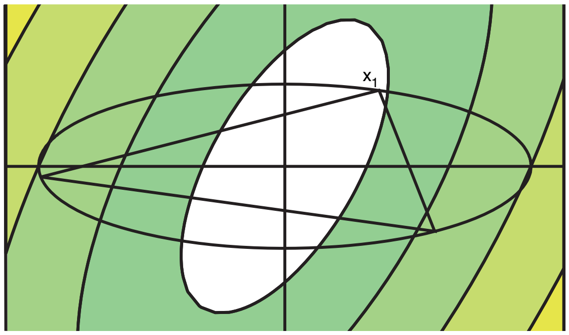


FIG. 3. *Triangle K and the contour lines of a function u. ϕ_h is about $-\frac{\pi}{6}$ in this graph.*

Denote by u_I the linear interpolation of u at the three vertices of K. Clearly the norm of the error $u - u_I$ depends on the size and the shape of triangle K, as well as the magnitude and the orientation of function u. We derive in this section the exact formulas for the H^1- and L^2-norms of the interpolation error in terms of these factors.

THEOREM 3.1. *Let K be a triangle oriented along the x-axis. r_{12} is the aspect ratio of K, $r_{21} = 1/r_{12}$. Let u be a quadratic function. λ_1 and λ_2 are the eigenvalues of the Hessian matrix of u, and the orientation of u is of angle ϕ_h with the y-axis; then*

$$\begin{aligned}\|\nabla(u-u_I)\|^2_{L^2(K)} = \frac{\sqrt{3}}{36}|K|^2\Big\{&[(r_{12}+r_{21})(\lambda_1^2+\lambda_2^2)+(r_{12}-r_{21})(\lambda_1^2-\lambda_2^2)\cos(2\phi_h)]\\ &+\frac{1}{2}\Big[-4\lambda_1\lambda_2+\frac{1}{4}(\ (\lambda_1+\lambda_2)(r_{12}+r_{21})+(\lambda_1-\lambda_2)(r_{12}-r_{21})\\ &\cdot\cos(2\phi_h)\ )^2\Big][(r_{12}+r_{21}+(r_{12}-r_{21})\cos(6\phi_v+4\theta)]\Big\},\end{aligned}\tag{4}$$

where ϕ_v is the rotation angle in the mapping from $\hat{K}$ to K, and

$$\theta = \frac{1}{2}\ \mathrm{atan}\left(\frac{2(\lambda_1-\lambda_2)\sin(2\phi_h)}{(\lambda_1+\lambda_2)(r_{12}-r_{21})+(\lambda_1-\lambda_2)(r_{12}+r_{21})\cos(2\phi_h)}\right).\tag{5}$$

Proof.

Step 1. We start with a result established by Bank and Smith in [3]. Let $c_i(x,y), i = 1,2,3$, be the barycentrical coordinates of a point (x,y) in K. The side basis functions (taking value $\frac{1}{4}$ at a midpoint) are

$$\begin{aligned}b_1(x,y) &= c_2(x,y)\ c_3(x,y),\\ b_2(x,y) &= c_3(x,y)\ c_1(x,y),\\ b_3(x,y) &= c_1(x,y)\ c_2(x,y).\end{aligned}$$

It is easy to see that

$$u - u_I = -\frac{1}{2}(v_1\ b_1(x,y) + v_2\ b_2(x,y) + v_3\ b_3(x,y)),\tag{6}$$

where

$$v_i = \boldsymbol{\ell}_i \cdot H\boldsymbol{\ell}_i, \qquad i = 1,2,3.\tag{7}$$

Let $\boldsymbol{v} = [v_1, v_2, v_3]^T$. It is further established in [3] that

$$\int_K |\nabla(u-u_I)|^2\ dxdy = \frac{1}{4}\boldsymbol{v}\cdot B\boldsymbol{v},\tag{8}$$

where

$$\begin{aligned}B &= \left(\int_K \nabla b_i \cdot \nabla b_j dxdy\right)\\ &= \frac{1}{48|K|}\begin{bmatrix} |\boldsymbol{\ell}_1|^2+|\boldsymbol{\ell}_2|^2+|\boldsymbol{\ell}_3|^2 & 2\boldsymbol{\ell}_1\cdot\boldsymbol{\ell}_2 & 2\boldsymbol{\ell}_1\cdot\boldsymbol{\ell}_3\\ & |\boldsymbol{\ell}_1|^2+|\boldsymbol{\ell}_2|^2+|\boldsymbol{\ell}_3|^2 & 2\boldsymbol{\ell}_2\cdot\boldsymbol{\ell}_3\\ \text{symm.} & & |\boldsymbol{\ell}_1|^2+|\boldsymbol{\ell}_2|^2+|\boldsymbol{\ell}_3|^2\end{bmatrix}.\end{aligned}$$

We first derive a formula for the matrix B in terms of the singular values and the rotation angle ϕ_v of the mapping M. Note that we assume $\phi_u = 0$; therefore

$$\boldsymbol{\ell}_i = M\boldsymbol{e}_i = \Sigma R^*_{\phi_v}\boldsymbol{e}_i = \sqrt{3}\begin{bmatrix} \sigma_1\cos\alpha_i \\ \sigma_2\sin\alpha_i \end{bmatrix},$$

where

$$\alpha_i = 2(i-1)\pi/3 - \phi_v, \qquad i = 1,2,3. \tag{9}$$

With formula (A.1) in the appendix it is easy to verify that

$$|\boldsymbol{\ell}_1|^2 + |\boldsymbol{\ell}_2|^2 + |\boldsymbol{\ell}_3|^2 = 3\sigma_1^2\sum_{i=1}^{3}\cos^2\alpha_i + 3\sigma_2^2\sum_{i=1}^{3}\sin^2\alpha_i = \frac{9}{2}(\sigma_1^2+\sigma_2^2). \tag{10}$$

Let

$$f_{ij} = \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}\cos\alpha_i\cos\alpha_j + \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}\sin\alpha_i\sin\alpha_j. \tag{11}$$

Then

$$\boldsymbol{\ell}_i\cdot\boldsymbol{\ell}_j = 3(\sigma_1^2\cos\alpha_i\cos\alpha_j + \sigma_2^2\sin\alpha_i\sin\alpha_j) = 3(\sigma_1^2+\sigma_2^2)f_{ij}, \qquad 1 \le i,j \le 3.$$

Note that $|K| = \frac{3\sqrt{3}}{4}\sigma_1\sigma_2$. We may express B as

$$B = \frac{\sqrt{3}}{24}\frac{\sigma_1^2+\sigma_2^2}{\sigma_1\sigma_2}\begin{bmatrix} 1 & \frac{4}{3}f_{12} & \frac{4}{3}f_{13} \\ & 1 & \frac{4}{3}f_{23} \\ & \text{symm.} & 1 \end{bmatrix},$$

and rewrite the norm of the error as

$$\|\nabla(u-u_I)\|^2_{L^2(K)} = \frac{\sqrt{3}}{96}\frac{\sigma_1^2+\sigma_2^2}{\sigma_1\sigma_2} \cdot \left[\sum_{i=1}^{3}(v_i)^2 + \frac{8}{3}(f_{12}\ v_1v_2 + f_{13}\ v_1v_3 + f_{23}\ v_2v_3)\right]. \tag{12}$$

Step 2. We next simplify the terms involving v_i. Assume the Hessian matrix $H = (h_{ij})$, and denote by

$$\tilde{H} = \begin{bmatrix} \frac{\sigma_1}{\sigma_2}h_{11} & h_{12} \\ h_{12} & \frac{\sigma_2}{\sigma_1}h_{22} \end{bmatrix}. \tag{13}$$

Since $\tilde{H}$ is a 2×2 symmetric matrix, we may decompose it into

$$\tilde{H} = R_\theta\begin{bmatrix} \mu_1 & \\ & \mu_2 \end{bmatrix}R_\theta^T, \tag{14}$$

where R_θ is the matrix for the counterclockwise rotation by an angle θ, and μ_1 and μ_2 are the eigenvalues of $\tilde{H}$. By the facts that

$$\begin{aligned} v_i &= \boldsymbol{\ell}_i\cdot H\boldsymbol{\ell}_i = (R^*_{\phi_v}\boldsymbol{e}_i)\cdot(\Sigma^T H\Sigma)\ (R^*_{\phi_v}\boldsymbol{e}_i) = \sigma_1\sigma_2(R^*_{\phi_v}\boldsymbol{e}_i)\cdot\tilde{H}\ (R^*_{\phi_v}\boldsymbol{e}_i) \\ &= \sigma_1\sigma_2(R^*_\theta R^*_{\phi_v}\boldsymbol{e}_i)\cdot\operatorname{diag}(\mu_1,\mu_2)\ (R^*_\theta R^*_{\phi_v}\boldsymbol{e}_i) \end{aligned}$$

and that

$$R_\theta^* R_{\phi_v}^* \boldsymbol{e}_i = \sqrt{3}[\cos(\alpha_i - \theta), \sin(\alpha_i - \theta)]^T,$$

we have

$$\begin{aligned} v_i &= 3\sigma_1\sigma_2[\mu_1 \cos^2(\alpha_i - \theta) + \mu_2 \sin^2(\alpha_i - \theta)] \\ &= \tfrac{3}{2}\sigma_1\sigma_2[(\mu_1 + \mu_2) + (\mu_1 - \mu_2)\cos 2(\alpha_i - \theta)]. \end{aligned}$$

Therefore it follows from formulas (A.2) and (A.3) that

$$\begin{aligned} \sum_{i=1}^{3}(v_i)^2 &= \frac{9}{4}(\sigma_1\sigma_2)^2 \left[3(\mu_1 + \mu_2)^2 + \frac{3}{2}(\mu_1 - \mu_2)^2\right] \\ &= \frac{81}{8}(\sigma_1\sigma_2)^2 \left[\mu_1^2 + \mu_2^2 + \frac{2}{3}\mu_1\mu_2\right]. \end{aligned} \tag{15}$$

Expand $v_i v_j$ as follows:

$$\begin{aligned} v_i v_j = \frac{9}{4}(\sigma_1\sigma_2)^2[(\mu_1 + \mu_2)^2 &+ (\mu_1^2 - \mu_2^2)(\cos 2(\alpha_i - \theta) + \cos 2(\alpha_j - \theta)) \\ &+ (\mu_1 - \mu_2)^2 \cos 2(\alpha_i - \theta)\cos 2(\alpha_j - \theta)]. \end{aligned} \tag{16}$$

Let $\beta = \phi_v + \theta$. Then $\alpha_i - \theta = 2(i-1)\pi/3 - \beta$, and

$$\cos 2(\alpha_i - \theta) + \cos 2(\alpha_j - \theta) = \begin{cases} \cos(2\beta + \frac{\pi}{3}) & \text{for } (i,j) = (1,2), \\ \cos(2\beta - \frac{\pi}{3}) & \text{for } (i,j) = (1,3), \\ \cos(2\beta + \pi) & \text{for } (i,j) = (2,3) \end{cases}$$

and

$$\cos 2(\alpha_i - \theta) \cdot \cos 2(\alpha_j - \theta) = \begin{cases} -\frac{1}{4} - \frac{1}{2}\cos(4\beta - \frac{\pi}{3}) & \text{for } (i,j) = (1,2), \\ -\frac{1}{4} - \frac{1}{2}\cos(4\beta + \frac{\pi}{3}) & \text{for } (i,j) = (1,3), \\ -\frac{1}{4} - \frac{1}{2}\cos(4\beta + \pi) & \text{for } (i,j) = (2,3). \end{cases}$$

Substituting the above formulas into (17), we have

$$\begin{aligned} & f_{12}\ v_1v_2 + f_{13}\ v_1v_3 + f_{23}\ v_2v_3 \\ & = \tfrac{9}{4}(\sigma_1\sigma_2)^2\ \{[(\mu_1 + \mu_2)^2 - \tfrac{1}{4}(\mu_1 - \mu_2)^2]\ (f_{12} + f_{13} + f_{23}) \\ & \quad + (\mu_1^2 - \mu_2^2)\ [f_{12}\cos(2\beta + \tfrac{\pi}{3}) + f_{13}\cos(2\beta - \tfrac{\pi}{3}) + f_{23}\cos(2\beta + \pi)] \\ & \quad - \tfrac{1}{2}(\mu_1 - \mu_2)^2\ [f_{12}\cos(4\beta - \tfrac{\pi}{3}) + f_{13}\cos(4\beta + \tfrac{\pi}{3}) + f_{23}\cos(4\beta + \pi)]\ \}. \end{aligned} \tag{17}$$

By using the definition (11) for f_{ij}, we have from formula (A.4) that

$$f_{12} + f_{13} + f_{23} = -\frac{3}{4}; \tag{18}$$

from formulas (A.5) and (A.6) that

$$\begin{aligned} & f_{12}\cos(2\beta + \tfrac{\pi}{3}) + f_{13}\cos(2\beta - \tfrac{\pi}{3}) + f_{23}\cos(2\beta + \pi) \\ & = -\tfrac{3}{4}\ \frac{\sigma_1^2 - \sigma_2^2}{\sigma_1^2 + \sigma_2^2}\cos(2\theta); \end{aligned} \tag{19}$$

and from formulas (A.7) and (A.8) that

$$\begin{aligned} &f_{12}\cos(4\beta-\tfrac{\pi}{3})+f_{13}\cos(4\beta+\tfrac{\pi}{3})+f_{23}\cos(4\beta+\pi)\\ &=-\tfrac{3}{4}\ \frac{\sigma_1^2-\sigma_2^2}{\sigma_1^2+\sigma_2^2}\cos(6\phi_v+4\theta). \end{aligned} \tag{20}$$

Put (17)–(20) into (13); we may express the norm of the error as

$$\begin{aligned} \|\nabla(u-u_I)\|^2_{L^2(K)}=\frac{3\sqrt{3}}{64}(\sigma_1^2+\sigma_2^2)\sigma_1\sigma_2\Bigg\{&\left[\mu_1^2+\mu_2^2-\frac{\sigma_1^2-\sigma_2^2}{\sigma_1^2+\sigma_2^2}(\mu_1^2-\mu_2^2)\cos(2\theta)\right]\\ &+\ \frac{1}{2}(\mu_1-\mu_2)^2\left[1+\frac{\sigma_1^2-\sigma_2^2}{\sigma_1^2+\sigma_2^2}\cos(6\phi_v+4\theta)\right]\Bigg\}. \end{aligned} \tag{21}$$

Step 3. Finally we express μ_1,μ_2, and θ in terms of the eigenvalues of the Hessian H and the aspect ratio r_{12} of K. According to the eigendecomposition (3) of H,

$$H=\begin{bmatrix} \lambda_1\cos^2\phi_h+\lambda_2\sin^2\phi_h & (\lambda_1-\lambda_2)\sin\phi_h\cos\phi_h\\ (\lambda_1-\lambda_2)\sin\phi_h\cos\phi_h & \lambda_1\sin^2\phi_h+\lambda_2\cos^2\phi_h \end{bmatrix}.$$

Therefore

$$\tilde{H}=(\tilde{h}_{ij})=\begin{bmatrix} r_{12}(A+D\cos2\phi_h) & D\sin2\phi_h\\ D\sin2\phi_h & r_{21}(A-D\cos2\phi_h) \end{bmatrix}$$

with

$$A=\frac{1}{2}(\lambda_1+\lambda_2),\qquad D=\frac{1}{2}(\lambda_1-\lambda_2).$$

Recall the eigendecomposition (14), it is not difficult to establish for the eigenvalues μ_1 and μ_2 of $\tilde{H}$ that

$$\begin{aligned} \mu_1^2+\mu_2^2 &=(\tilde{h}_{11})^2+(\tilde{h}_{22})^2+2(\tilde{h}_{12})^2\\ &=(A^2+D^2\cos^2(2\phi_h))(r_{12}^2+r_{21}^2)+2AD(r_{12}^2-r_{21}^2)\cos(2\phi_h)\\ &\quad+\ 2D^2\sin^2(2\phi_h),\\ (\mu_1^2-\mu_2^2)\cos2\theta &=(\tilde{h}_{11})^2-(\tilde{h}_{22})^2\\ &=(A^2+D^2\cos^2(2\phi_h))(r_{12}^2-r_{21}^2)+2AD(r_{12}^2+r_{21}^2)\cos(2\phi_h),\\ (\mu_1-\mu_2)^2 &=(\tilde{h}_{11}-\tilde{h}_{22})^2+4(\tilde{h}_{12})^2\\ &=4D^2+A^2(r_{12}-r_{21})^2+\ 2AD(r_{12}^2-r_{21}^2)\cos(2\phi_h)\\ &\quad+D^2(r_{12}-r_{21})^2\cos^2(2\phi_h)\\ &=4(D^2-A^2)+A^2(r_{12}+r_{21})^2+\ 2AD(r_{12}^2-r_{21}^2)\cos(2\phi_h)\\ &\quad+D^2(r_{12}-r_{21})^2\cos^2(2\phi_h)\\ &=-\ 4\lambda_1\lambda_2+[A(r_{12}+r_{21})+D(r_{12}-r_{21})\cos(2\phi_h)]^2 \end{aligned}$$

and that

$$\tan(2\theta)=\frac{2\tilde{h}_{12}}{\tilde{h}_{11}-\tilde{h}_{22}}=\frac{2D\sin2\phi_h}{A(r_{12}-r_{21})+D(r_{12}+r_{21})\cos2\phi_h}.$$

Substitute the above formulas into the right-hand side of (21) and simplify, and we obtain the formula (5). $\square$

We next derive the formula for the L^2-norm of the linear interpolation error.

THEOREM 3.2. *Let K be a triangle oriented along the x-axis. r_{12} is the aspect ratio of K, $r_{21} = 1/r_{12}$. Let u be a quadratic function. λ_1 and λ_2 are the eigenvalues of its Hessian, and the orientation of u is of angle ϕ_h with the y-axis. Then*

$$\|u - u_I\|^2_{L^2(K)} = \frac{|K|^3}{160}\left\{ -\frac{16}{9}\lambda_1\lambda_2 + [(\lambda_1+\lambda_2)(r_{12}+r_{21}) + (\lambda_1-\lambda_2)(r_{12}-r_{21})\cos(2\phi_h)]^2 \right\}. \tag{22}$$

Proof. It follows from (6) that

$$\int_K |u-u_I|^2 \; dxdy = \frac{1}{4}\boldsymbol{v}\cdot B_0\boldsymbol{v}, \tag{23}$$

where

$$B_0 = \left(\int_K b_i \cdot b_j dxdy\right).$$

An elementary calculation of the integrals leads to

$$B_0 = \frac{|K|}{180}\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}.$$

Therefore

$$\int_K |u-u_I|^2 \; dxdy = \frac{|K|}{360}\left(\sum_{i=1}^{3}(v_i)^2 + v_1v_2 + v_1v_3 + v_2v_3\right), \tag{24}$$

where v_1, v_2, and v_3 are defined as in (7). We have similar to (17) that

$$v_1v_2 + v_1v_3 + v_2v_3 = \frac{81}{16}(\sigma_1\sigma_2)^2\left(\mu_1^2 + \mu_2^2 + \frac{10}{3}\mu_1\mu_2\right),$$

which, together with (15), yields

$$\int_K |u-u_I|^2 \; dxdy = \frac{|K|^3}{40}\left((\mu_1+\mu_2)^2 - \frac{4}{9}\mu_1\mu_2\right). \tag{25}$$

Finally, by $\mu_1\mu_2 = \det(\tilde{H}) = \det(H) = \lambda_1\lambda_2$ and

$$\mu_1 + \mu_2 = \tilde{h}_{11} + \tilde{h}_{22} = A(r_{12}+r_{21}) + D(r_{12}-r_{21})\cos(2\phi_h),$$

we prove the conclusion of this theorem. □

4. Discussion about some special cases. Denote by T_1, T_2, and T_3 the terms in the square brackets on the right-hand side of (5), i.e.,

$$T_1 = (r_{12}+r_{21})(\lambda_1^2+\lambda_2^2) + (r_{12}-r_{21})(\lambda_1^2-\lambda_2^2)\cos(2\phi_h), \tag{26}$$

$$T_2 = -4\lambda_1\lambda_2 + \frac{1}{4}\left((\lambda_1+\lambda_2)(r_{12}+r_{21}) + (\lambda_1-\lambda_2)(r_{12}-r_{21})\cos(2\phi_h)\right)^2, \tag{27}$$

$$T_3 = r_{12} + r_{21} + (r_{12}-r_{21})\cos(6\phi_v + 4\theta). \tag{28}$$

Recall that we assume $r_{12} = \sigma_1/\sigma_2 \geq 1$ and $\lambda_1 \geq |\lambda_2|$; therefore all the three terms above are nonnegative, and we may write

$$\|\nabla(u - u_I)\|^2_{L^2(K)} = \frac{\sqrt{3}}{36}|K|^2 \left(T_1 + \frac{1}{2}\,T_2 \cdot T_3\right). \tag{29}$$

We may also restrict $0 \leq \phi_v < \frac{\pi}{3}$ and $0 \leq \phi_h < \pi$. Other cases are covered by the symmetry and periodicity of the error norms. For a fixed aspect ratio, the extremum values of T_1, T_2, and T_3 are

$$T_1 = \begin{cases} \min\,(T_1) = 2(r_{21}\lambda_1^2 + r_{12}\lambda_2^2) & \text{when } \phi_h = \frac{\pi}{2}, \\ \max\,(T_1) = 2(r_{12}\lambda_1^2 + r_{21}\lambda_2^2) & \text{when } \phi_h = 0, \end{cases} \tag{30}$$

$$T_2 = \begin{cases} \min\,(T_2) = (r_{21}\lambda_1 - r_{12}\lambda_2)^2 & \text{when } \phi_h = \frac{\pi}{2}, \\ \max\,(T_2) = (r_{12}\lambda_1 - r_{21}\lambda_2)^2 & \text{when } \phi_h = 0, \end{cases} \tag{31}$$

$$T_3 = \begin{cases} \min\,(T_3) = 2r_{21} & \text{when } 6\phi_v + 4\theta = (2m+1)\pi, \\ \max\,(T_3) = 2r_{12} & \text{when } 6\phi_v + 4\theta = 2m\pi. \end{cases} \tag{32}$$

We give more details about these cases.

Case 1. $\phi_h = \frac{\pi}{2}$ and $\phi_v = \frac{\pi}{6}$, i.e., K is an acute isosceles triangle aligned with u. In this case, $\theta = 0$ by (5). Therefore, T_1, T_2, and T_3 all take their minimum values, which yields

$$\|\nabla(u - u_I)\|^2_{L^2(K)} = \frac{\sqrt{3}}{36}|K|^2\ [\ 2(r_{21}\lambda_1^2 + r_{12}\lambda_2^2) + r_{21}(r_{21}\lambda_1 - r_{12}\lambda_2)^2\ ]. \tag{33}$$

This is the smallest H^1-seminorm of the interpolation error for all the triangles with a fixed aspect ratio and a fixed area.

We may consider the optimal choice of the aspect ratio in this case. Let

$$\begin{aligned} f(r) &= 2\Big(\tfrac{1}{r}\lambda_1^2 + r\lambda_2^2\Big) + \tfrac{1}{r}\Big(\tfrac{1}{r}\lambda_1 - r\lambda_2\Big)^2 \\ &= 3\lambda_2^2 r + 2(2\lambda_1^2 - 4\lambda_1\lambda_2)\tfrac{1}{r} + \lambda_1^2\frac{1}{r^3}. \end{aligned}$$

It can be seen that f is decreasing in $[1, r_*^{(1)})$ and increasing in $(r_*^{(1)}, \infty)$, where

$$r_*^{(1)} = \sqrt{\frac{\lambda_1(\lambda_1 - \lambda_2) + \lambda_1\sqrt{(\lambda_1 - \lambda_2)^2 + 9\lambda_2^2}}{3\lambda_2^2}}. \tag{34}$$

Therefore $r_{12} = r_*^{(1)}$ is the best aspect ratio. Moreover, since T_1, T_2, and T_3 all take minimum values in this case, we conclude that the acute isosceles triangle, which is aligned with u and of the aspect ratio $r_*^{(1)}$, is the best one that produces the smallest H^1-seminorm of the interpolation error among all the triangles with the same area!

It is noted that for u with $\lambda_1 = \lambda_2$ (isotropic u), the best aspect ratio is $r_*^{(1)} = 1$ (with isotropic K). For u with $\lambda_2 = -\lambda_1$, the best aspect ratio is $r_*^{(1)} = \sqrt{(2 + \sqrt{13})/3}$ ≈ 1.367. When $\lambda_1 >> |\lambda_2|$, we have

$$r_*^{(1)} \approx \sqrt{\frac{2}{3}\left|\frac{\lambda_1}{\lambda_2}\right|} \approx 0.816\left|\frac{\lambda_1}{\lambda_2}\right|,$$

with which the interpolation error is of the smallest possible magnitude

$$\|\nabla(u-u_I)\|^2_{L^2(K)} \approx \frac{\sqrt{2}}{6}|K|^2|\lambda_1\lambda_2| \approx 0.235|K|^2|\lambda_1\lambda_2|. \tag{35}$$

We may also compare the above optimal choice of the aspect ratio with some intuitive choices. For instance, if we choose $r_{12}=1$, then

$$\|\nabla(u-u_I)\|^2_{L^2(K)} \approx \frac{\sqrt{3}}{36}|K|^2(3(\lambda_1^2+\lambda_2^2)-2\lambda_1\lambda_2), \tag{36}$$

which is between $0.128|K|^2\lambda_1^2$ and $0.193|K|^2\lambda_1^2$. If we choose $r_{12}=|\frac{\lambda_1}{\lambda_2}|$, then in the case $\lambda_1 >> |\lambda_2|$,

$$\|\nabla(u-u_I)\|^2_{L^2(K)} \approx \frac{5\sqrt{3}}{36}|K|^2|\lambda_1\lambda_2| \approx 0.240|K|^2|\lambda_1\lambda_2|; \tag{37}$$

if we choose $r_{12}=\sqrt{|\lambda_1/\lambda_2|}$, then we have

$$\|\nabla(u-u_I)\|^2_{L^2(K)} = \frac{\sqrt{3}}{18}|K|^2\sqrt{|\lambda_1^3\lambda_2|} \approx 0.0962|K|^2\sqrt{|\lambda_1^3\lambda_2|}. \tag{38}$$

Note that when u and K are aligned, the inverse mapping $\boldsymbol{\xi} = M^{-1}\boldsymbol{x}$ with $r_{12} = \sqrt{|\lambda_1/\lambda_2|}$ transforms $u(x,y)$ into const.$(\xi^2 \pm \eta^2)$ and K into an equilateral triangle. It was shown by D'Azevedo and Simpson [6] and Rippa [13] that triangles with such an aspect ratio lead to the smallest maximum-norm and L^p-norm of the interpolation error among all triangles of the same area. However, this aspect ratio is not the optimal for the interpolation error in the sense of H^1-seminorm.

Case 2. $\phi_h = \frac{\pi}{2}$ and $\phi_v = 0$, i.e., K is an obtuse isosceles triangle aligned with u. We have in this case

$$\|\nabla(u-u_I)\|^2_{L^2(K)} = \frac{\sqrt{3}}{36}|K|^2\ [\ 2(r_{21}\lambda_1^2+r_{12}\lambda_2^2)+r_{12}(r_{21}\lambda_1-r_{12}\lambda_2)^2\ ]. \tag{39}$$

This error formula differs from that of Case 1 only in the coefficient of the second term.

To study the optimal choice of the aspect ratio in this case, let

$$\begin{aligned} g(r) &= 2\Big(\frac{1}{r}\lambda_1^2 + r\lambda_2^2\Big) + r\Big(\frac{1}{r}\lambda_1 - r\lambda_2\Big)^2 \\ &= 3\lambda_1^2\frac{1}{r} + 2(\lambda_2^2-\lambda_1\lambda_2)r + \lambda_2^2 r^3. \end{aligned}$$

It can be shown that the minimum of g is attained at

$$r_*^{(2)} = \sqrt{\frac{\lambda_2(\lambda_1-\lambda_2)+|\lambda_2|\sqrt{(\lambda_1-\lambda_2)^2+9\lambda_2^2}}{3\lambda_2^2}}. \tag{40}$$

When $\lambda_1=\lambda_2$, the best aspect ratio is $r_*^{(2)}=1$. When $\lambda_2=-\lambda_2$, the best aspect ratio is $r_*^{(1)}=\sqrt{(2+\sqrt{13})/3}\approx 1.367$. When $|\lambda_1| >> |\lambda_2|$, we have

$$r_*^{(2)} \approx \begin{cases} \sqrt{\frac{\sqrt{10}+1}{3}|\frac{\lambda_1}{\lambda_2}|} \approx 1.178\sqrt{|\frac{\lambda_1}{\lambda_2}|} & \text{when } \lambda_1\lambda_2>0, \\ \sqrt{\frac{\sqrt{10}-1}{3}|\frac{\lambda_1}{\lambda_2}|} \approx 0.849\sqrt{|\frac{\lambda_1}{\lambda_2}|} & \text{when } \lambda_1\lambda_2<0. \end{cases}$$

With this choice of aspect ratio, the interpolation error is of the magnitude

$$\|\nabla(u-u_I)\|^2_{L^2(K)} \approx \begin{cases} 0.0878|K|^2\sqrt{|\lambda_1^3\lambda_2|} & \text{when } \lambda_1\lambda_2 > 0, \\ 0.281|K|^2\sqrt{|\lambda_1^3\lambda_2|} & \text{when } \lambda_1\lambda_2 < 0. \end{cases} \tag{41}$$

We may compare this case (obtuse isosceles) with the previous one (acute isosceles). When the triangle is aligned with u and the aspect ratio $r_{12} \approx \sqrt{|\lambda_1/\lambda_2|}$, the interpolation error is nearly the same for both acute and obtuse triangles if $\lambda_1\lambda_2 > 0$. For u with $\lambda_1\lambda_2 < 0$, $\|\nabla(u-u_I)\|_{L^2(K)}$ is about 1.7 times smaller with the acute isosceles triangle than that with the obtuse one. Furthermore, because for general triangles with arbitrary ϕ_v, the error norm $\|\nabla(u-u_I)\|_{L^2(K)}$ is between those of Case 1 and Case 2, we may conclude that the H^1-seminorm of the interpolation error is not sensitive to the maximum angle of the triangle, as long as the triangle is aligned with the function u and the aspect ratio $r_{12} \approx \sqrt{|\lambda_1/\lambda_2|}$ is used.

Case 3. $\phi_h = 0$. In this case the orientation of u is perpendicular to the triangle. We have

$$\begin{aligned} \|\nabla(u-u_I)\|^2_{L^2(K)} = \frac{\sqrt{3}}{18}|K|^2\ [2(r_{12}\lambda_1^2 + r_{21}\lambda_2^2) + (r_{12}\lambda_1 - r_{21}\lambda_2)^2 \\ \cdot (r_{12} + r_{21} + (r_{12} - r_{21})\cos(6\phi_v))] \end{aligned} \tag{42}$$

For given $|K|$ and ϕ_v, it can be shown the error norm is an increasing function of r_{12}. Therefore, the best aspect ratio is $r_{12} = 1$. This is not surprising, since when the triangle is perpendicular to the orientation of u, increasing the aspect ratio leads to lower resolution in the needy direction and larger interpolation error, no matter what value ϕ_v takes.

We should emphasize that in the above discussion on the optimal aspect ratios the area and the orientation of the triangle are assumed fixed. In adaptive mesh generation there may be some other constraints on the triangles in a mesh, e.g., one side of a triangle or the vertices of the triangles are fixed. Optimal choices of the aspect ratio subject to those constraints can be quite different from the above values.

5. General discussion about orientation, aspect ratio, and angle condition.

In this section we present a general discussion about the effects of the orientation, aspect ratio, and the angle ϕ_v. We first study the H^1-seminorm of the interpolation error.

Mesh alignment. Given a triangle K (fixed $|K|$, r_{12}, and ϕ_v), it is commonly believed that the interpolation error is the smallest when the triangle is aligned with the function, i.e., when $\phi_h = \frac{\pi}{2}$. We present a justification of this viewpoint. We consider the case where both the solutions and the triangles are highly anisotropic, i.e., we assume $\lambda_1 >> |\lambda_2|$ and $r_{12} >> 1$. In this case we have for the three terms (26)–(28) in the error norm $\|\nabla(u-u_I)\|^2_{L^2(K)}$ that

$$\begin{aligned} T_1 &= r_{12}\lambda_1^2(1+\cos 2\phi_h) + \text{LOT}, \\ T_2 &= \tfrac{1}{4}(r_{12}\lambda_1)^2(1+\cos 2\phi_h)^2 + \text{LOT}, \\ T_3 &= r_{12}(1+\cos(6\phi_v + 4\theta)) + \text{LOT} \end{aligned}$$

and for the angle θ in (5) that

$$\tan(2\theta) \approx \frac{2\lambda_1 \sin 2\phi_h}{\lambda_1 r_{12}(1+\cos 2\phi_h)} = 2r_{21}\tan\phi_h,$$

where the lower order terms LOT include $r_{12}\lambda_2^2, r_{12}\lambda_1^2, r_{21}\lambda_2^2$, etc. In this case the error norm is dominated by

$$\|\nabla(u-u_I)\|^2_{L^2(K)} \approx \frac{\sqrt{3}}{36}|K|^2 \quad \begin{aligned}&[r_{12}\lambda_1^2 + \tfrac{1}{4}r_{12}^3\lambda_1^2(1+\cos 2\phi_h)(1+\cos(6\phi_v+4\theta))]\\ &\cdot\ (1+\cos 2\phi_h) + \text{LOT}.\end{aligned}$$

Note that by (5) θ is an increasing function of ϕ_h, and $\cos(2\phi_h)$ is a decreasing function of ϕ_h in $(0, \frac{\pi}{2})$. Therefore, we may conclude that the H^1-seminorm of the error is almost a monotonically decreasing function of the angle ϕ_h in $(0, \frac{\pi}{2})$.

We plot in Figure 4 the graph of $\|\nabla(u-u_I)\|^2_{L^2(K)}$ versus ϕ_h for various λ_1, λ_2, ϕ_v and r_{12}. It is observed that for all aspect ratios and ϕ_v, the error is smallest when ϕ_h is near $\frac{\pi}{2}$. Also seen in Figure 4 is a small peak of the error norm around $\phi_h = \frac{\pi}{2}$. This is because when ϕ_h is very close to $\frac{\pi}{2}$, the leading terms in T_1 and T_2 go to 0 and the lower order terms become the major components in the error.

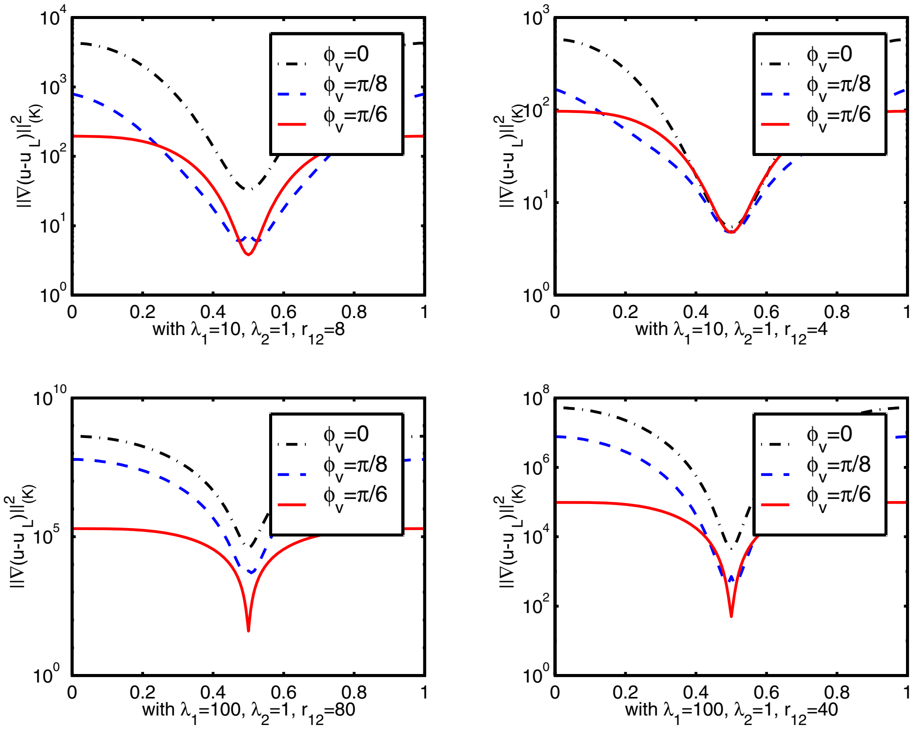


FIG. 4. $\|u-u_I\|^2_{L^2(K)}$ *versus the alignment angle* ϕ_h/π *with* $|K| = \frac{3\sqrt{3}}{4}$ *fixed.*

Aspect ratio. We are interested here in the question of, given a function u and a fixed triangle orientation ϕ_h, how one chooses the aspect ratio of the triangle. We plot in Figure 5 the graph of $\|\nabla(u-u_I)\|_{L^2(K)}$ versus r_{12} with various ϕ_h. It is observed that when $\phi_h = \frac{\pi}{2}$ (K aligned with u), the best aspect ratio r_{12} is in between $r_*^{(2)} = \sqrt{|\lambda_1/\lambda_2|}$ and $r_*^{(1)} \approx 0.8|\lambda_1/\lambda_2|$, while when $\phi_h \approx 0$ (K perpendicular to u), the best aspect ratio is 1. For other ϕ_v and ϕ_h, the optimal choice of r_{12} lies between 1 and $|\lambda_1/\lambda_2|$. Therefore, in practice, when a good mesh alignment is ensured, the proper aspect ratio should be between the magnitude of $\sqrt{|\lambda_1/\lambda_2|}$ and $0.8|\lambda_1/\lambda_2|$, with the lower end taken for the mesh with mostly obtuse triangles and the upper

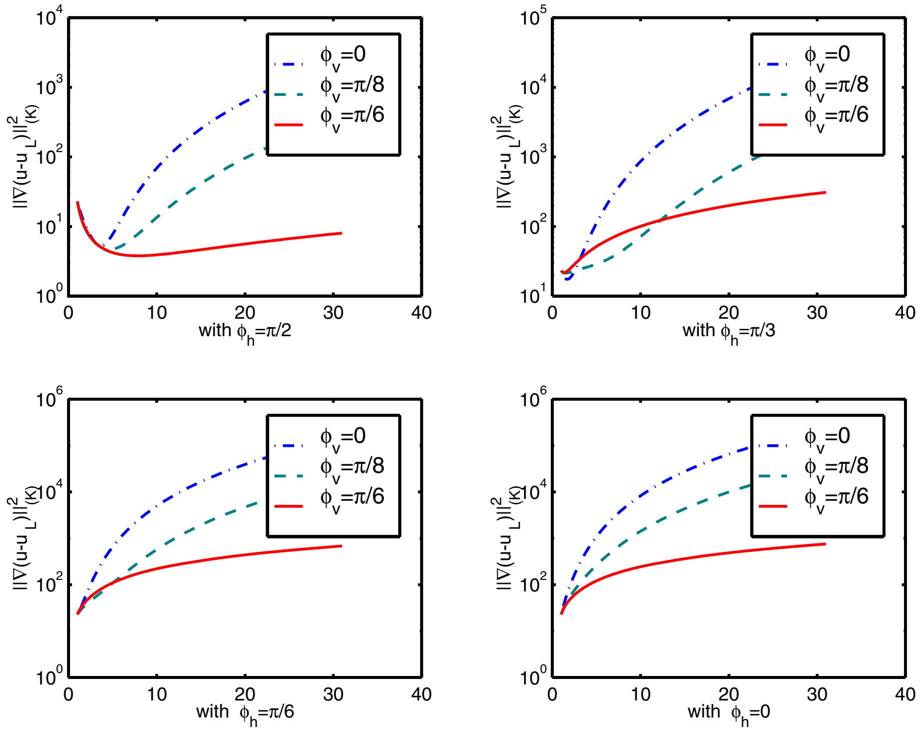


FIG. 5. $\|\nabla(u-u_I)\|^2_{L^2(K)}$ *versus the aspect ratio* r_{12} *with* $\lambda_1 = 10, \lambda_2 = 1$, *and* $|K| = \frac{3\sqrt{3}}{4}$.

end taken for the mesh with mostly acute triangles. The less the mesh alignment, the smaller r_{12} should be.

Maximum angle condition. It is well known that a good mesh for finite element approximation should respect the so-called maximum angle condition, i.e., the maximum angles of the triangular elements should be bounded from π. In [2], Babuška and Aziz gave an example showing that when the maximum angle condition is violated, the H^1-seminorm of the error of linear interpolation can go to infinity, although the element area approaches 0. Their example is in our notation the case with $\lambda_1 = 1, \lambda_2 = 0, |K| = \frac{\epsilon}{2}, r_{12} = \frac{\sqrt{3}}{2\epsilon}, \phi_h = 0, \phi_v = 0$, and thus $\|\nabla(u-u_I)\|^2_{L^2(K)} = \frac{9}{256}(\epsilon + \frac{3}{8\epsilon})$. We discuss here the impact of the maximum angle condition on the accuracy of linear interpolation. Taking as an example with $\lambda_1 = 100$ and $\lambda_2 = 1$, we plot in Figure 6 the graph of $\|\nabla(u-u_I)\|^2_{L^2(K)}$ versus ϕ_v with given aspect ratios and alignment directions. It is easy to see the following:

(a) When λ_1/λ_2 and r_{12} are high, triangles with $\phi_v \approx \frac{\pi}{6}$ or $\frac{\pi}{2}$ produce the smallest error, and triangles with $\phi_v \approx 0$ or $\frac{\pi}{3}$ produce the largest error, whether they are aligned with u or not. This can be justified by looking into formula (5) for the error norm. When $\lambda_1 >> |\lambda_2|$ and $r_{12} >> 1$, we have $\theta \approx 0$. Hence, the term T_3, and therefore $\|\nabla(u-u_I)\|_{L^2(K)}$, attains its minimum when $\phi_v \approx \frac{\pi}{6}$ or $\frac{\pi}{2}$.

Note that the maximum internal angle of K is an even periodic function of ϕ_v with the period $\pi/3$. It is decreasing in $(0, \pi/6)$. Hence $\phi_v = 0, \pi/3$ corresponds to the case where the maximum internal angle of K is the largest among all the triangles of the same aspect ratio, and $\phi_v = \pi/6, \pi/2$ corresponds to the case where the maximum internal angle is the smallest among all triangles. Therefore, in this case it is important to use triangles with smaller maximum internal angles, regardless of whether or not the mesh alignment is satisfied.

(b) When the triangle is aligned with u and the aspect ratio $r_{12} \approx \sqrt{|\lambda_1/\lambda_2|}$, the error is insensitive to ϕ_v and is therefore insensitive to the maximum internal angle of the triangle. In particular, if $\phi_h = \frac{\pi}{2}$, $\lambda_1\lambda_2 > 0$, and $r_{12} = \sqrt{\lambda_1/\lambda_2}$, then $T_2 = 0$, and by (29) $\|\nabla(u - u_I)\|_{L^2(K)}$ is independent of ϕ_v. See the lower left graph of Figure 6 and the analysis in Cases 1 and 2 in section 4.

(c) When $r_{12} \approx 1$, the best ϕ_v may be different from $\frac{\pi}{6}$ and $\frac{\pi}{2}$. However, in this case the difference between the maximum and the minimum of the error is not significant.

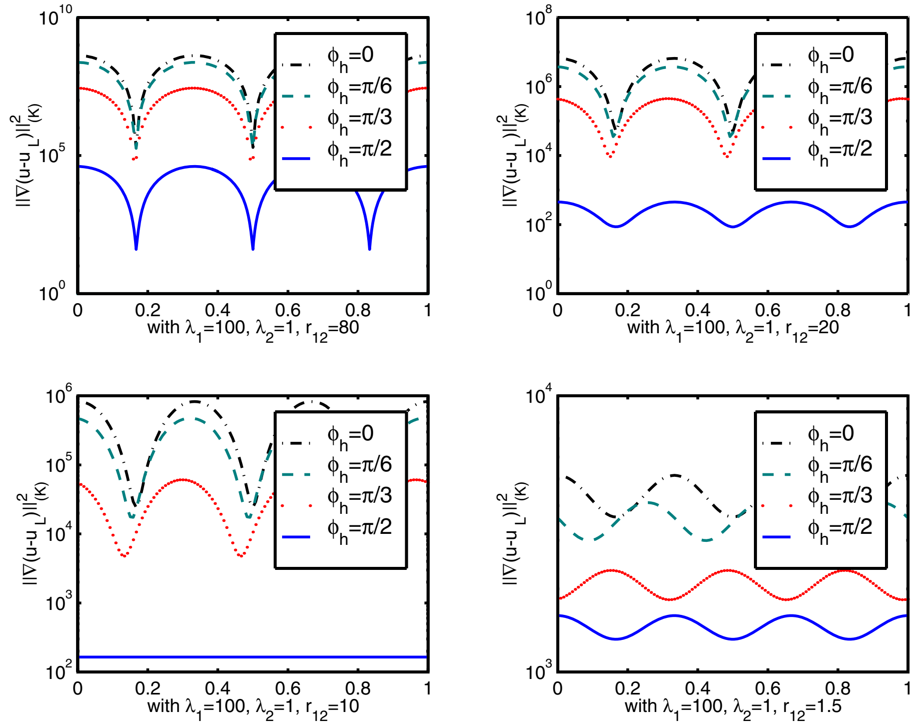


Fig. 6. $\|\nabla(u - u_I)\|^2_{L^2(K)}$ *versus the angle* ϕ_v/π *with* $|K| = \frac{3\sqrt{3}}{4}$ *fixed.*

In summary, we conclude that if the triangle is aligned with the function and the aspect ratio is of an magnitude $\sqrt{|\lambda_1/\lambda_2|}$ or smaller, then the maximum angle condition is not essential to the H^1-seminorm of the interpolation error. Otherwise, the maximum angle is critical to the magnitude of the error.

L^2-norm. From (23) we can conclude the following for the L^2-norm of the linear interpolation error:

(1) $\|u - u_I\|_{L^2(K)}$ does not depend on the angle ϕ_v. Therefore, for a given function, the L^2-norm of its linear interpolation error depends only on the area, aspect ratio, and orientation of the triangle. Maximum/minimum angle conditions are irrelevant to the L^2-norm of the interpolation error.

(2) $\|u - u_I\|_{L^2(K)}$ is a π-periodic even function of ϕ_h. It is decreasing in ϕ_h in

$(0, \frac{\pi}{2})$. When $\phi_h = \frac{\pi}{2}$ (K aligned with u),

$$\|u - u_I\|^2_{L^2(K)} = \frac{|K|^3}{40} \left[(\lambda_1 r_{21} + \lambda_2 r_{12})^2 - \frac{4}{9}\lambda_1\lambda_2 \right]$$

is the smallest, and when $\phi_h = 0$ (K perpendicular to u)

$$\|u - u_I\|^2_{L^2(K)} = \frac{|K|^3}{40} \left[(\lambda_1 r_{12} + \lambda_2 r_{21})^2 - \frac{4}{9}\lambda_1\lambda_2 \right]$$

is the largest.

(3) With the same orientation, it can be shown that when ϕ_h is in $(0, \frac{\pi}{4}) \cup (\frac{3\pi}{4}, \pi)$ (i.e., K is aligned more to the perpendicular direction of u), the best aspect ratio is $r_{12} = 1$; when $\frac{\pi}{4} \le \phi_h \le \frac{3\pi}{4}$ (i.e., K is aligned more to the orientation of u), the best aspect ratio is

$$r_{12} = \sqrt{\left| \frac{\lambda_1 + \lambda_2 - (\lambda_1 - \lambda_2)\cos(2\phi_h)}{\lambda_1 + \lambda_2 + (\lambda_1 - \lambda_2)\cos(2\phi_h)} \right|}. \tag{43}$$

We plot in Figure 7 the L^2-norm of the error versus the aspect ratio for the case of $\lambda_1 = 400, \lambda_2 = 1$, and $|K| = \frac{3\sqrt{3}}{4}$.

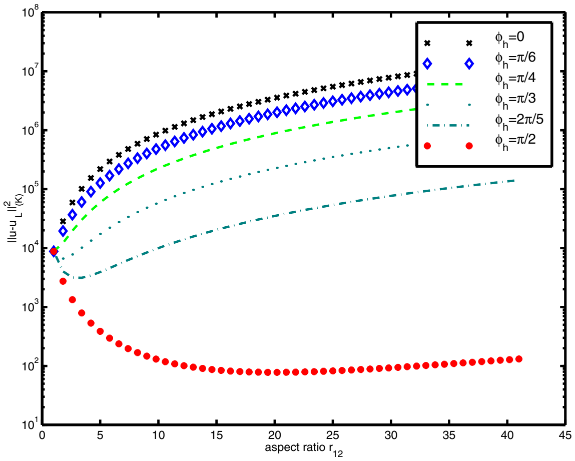


FIG. 7. $\|u - u_I\|^2_{L^2(K)}$ *versus the aspect ratio* r_{12} *with* $\lambda_1 = 400, \lambda_2 = 1$ *and* $|K| = \frac{3\sqrt{3}}{4}$.

(4) For a fixed triangle area, the minimum L^2-norm of the linear interpolation error is attained at $\phi_h = \frac{\pi}{2}$ and $r_{12} = \sqrt{|\lambda_1/\lambda_2|}$. The minimum value is

$$\|u - u_I\|^2_{L^2(K)} = \begin{cases} \frac{8}{90}|K|^3|\lambda_1\lambda_2| & \text{when } \lambda_1\lambda_2 > 0, \\ \frac{1}{90}|K|^3|\lambda_1\lambda_2| & \text{when } \lambda_1\lambda_2 < 0. \end{cases}$$

6. An example of linear interpolation on different meshes. In this section, we present the results of piecewise linear interpolation of a quadratic function on different types of mesh. We choose

$$u(x, y) = x^2 + 100y^2,$$

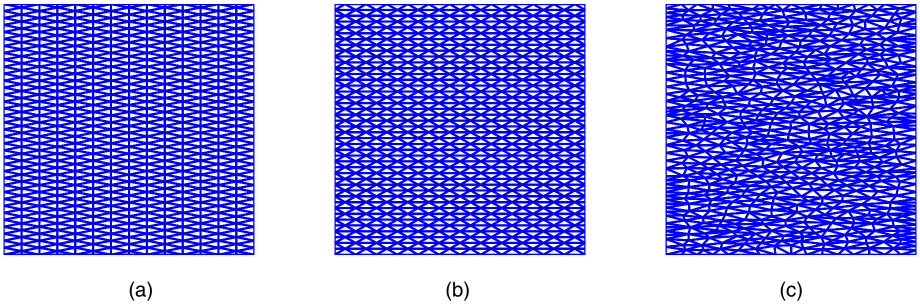


FIG. 8. *Three types of mesh used for piecewise linear interpolation.*

which corresponds to $\lambda_1 = 100, \lambda_2 = 1$, and the orientation along the x-axis. We consider the following three types of mesh on $[0,1] \times [0,1]$:

(a) The first type of mesh is as shown in Figure 8(a). There are M and N equal partitions in x and y directions, respectively. When $M/N < 2/\sqrt{3}$, all the triangles in this mesh, except those on the top and the bottom boundaries, are aligned exactly with the x-axis. Their aspect ratio is

$$r_{12} = \frac{2N}{\sqrt{3}M}$$

and their angle ϕ_v is $\frac{\pi}{6}$.

(b) The second type of mesh is as shown in Figure 8(b) with M and N equal partitions in x and y directions, respectively. When $M/N < \sqrt{3}/2$, all the triangles, except those on the left and the right boundaries, are aligned with the x-axis. Their aspect ratio is

$$r_{12} = \frac{\sqrt{3}N}{2M}$$

and their angle ϕ_v is 0.

(c) The third type is the unstructured meshes generated in the following way. First we create a Delaunay triangulation by using the package Triangle [14] on a rectangle $[0,1] \times [0, r_s]$, where $r_s \geq 1$ is a real number. Then we compress the rectangle in y direction by a factor r_s. The desired mesh on the unit square is the image of the Delaunay triangulation under the compression. See Figure 8(c) for a typical mesh of this type. By Delaunay triangulation the minimum internal angle of the triangles in the mesh is maximized, and all the triangles in the mesh over $[0,1] \times [0, r_s]$ are approximately of the same size and close to unilateral. Therefore, most triangles in the mesh over the unit square are roughly aligned with the x-axis (when $r_s > 1$) and are of the aspect ratio

$$r_{12} \approx r_s.$$

However, in this type of mesh the angle ϕ_v varies for different triangles (approximately uniformly distributed between 0 and $\pi/3$).

We are interested in the accuracy of the linear interpolation of u on different types of mesh and with various aspect ratios, in particular with the following three ratios: (1) $r_*^{(1)} = 81.6$, which is the best aspect ratio (for H^1-seminorm) calculated according to (34) for the case of acute isosceles triangles; (2) $r_*^{(2)} = 11.79$, which is the best aspect ratio (in H^1-seminorm) calculated according to (40) for the case of obtuse isosceles triangles; and (3) $r_*^{(\infty)} = \sqrt{|\frac{\lambda_1}{\lambda_2}|} = 10$, which is the best aspect ratio for the L^p-norm, $1 \leq p \leq \infty$. We also report the results with the aspect ratios $r_{12} = \lambda_1/\lambda_2 = 10$ and $r_{12} = 1$.

We list in Table 1 the various norms of the interpolation error with different choices of M, N, and r_s. Note that for all the meshes, the total number of elements is around 4000 and the total number of nodes is around 2200. The smallest H^1-seminorm for type (a) meshes is with $r_{12} = 92.37$; the smallest H^1-seminorm for type (b) meshes is with $r_{12} = 11.54$. For type (c) meshes, the smallest H^1-seminorm is achieved at $r_s = 11.78 \approx r_*^{(2)}$. This is because for this type of mesh, the error from the obtuse triangles (with $\phi_v \approx 0$) dominates in the global error. Therefore, the aspect ratio close to $r_*^{(2)}$ is the best choice for the global error norm.

For the L^2-, L^1-, and maximum norms, the smallest interpolation error is obtained with $r_{12} \approx r_*^{(\infty)}$ for all the three meshes.

In summary, we conclude that when the mesh is in good alignment with the solution and most of the triangles are acute isosceles, the aspect ratio should be chosen around $r_*^{(1)} \approx 0.8|\frac{\lambda_1}{\lambda_2}|$. If the mesh is in good alignment but with mostly obtuse triangles, or with varied maximum internal angles, then the aspect ratio should be chosen around $r_*^{(2)}$, which is approximately $1.178\sqrt{|\lambda_1/\lambda_2|}$ for the case $\lambda_1\lambda_1 > 0$, and $0.849\sqrt{|\lambda_1/\lambda_2|}$ for the case $\lambda_1\lambda_1 < 0$.

Table 1

The L^1- and L^2-norms, H^1-seminorm, and maximum norm of the interpolation error over the entire domain. (∗1) and (∗2) indicate the aspect ratio close to the best values for the H^1 seminorm in each case.

$M\times N$	Node #	Elem #	r_{12}	$\|u-u_I\|_{L^1}$	$\|u-u_I\|_{L^2}$	$\|u-u_I\|_{H^1}$	$\|u-u_I\|_\infty$
Type (a) meshes							
4×500	2507	4004	144.34	$5.23e{-3}$	$6.18e{-3}$	$8.28e{-2}$	$7.82e{-3}$
5×400	2409	4005	$92.37^{(*1)}$	$3.37e{-3}$	$3.97e{-3}$	$7.68e{-2}$	$5.01e{-3}$
6×333	2341	4002	64.08	$2.37e{-3}$	$2.78e{-3}$	$7.76e{-2}$	$3.50e{-3}$
14×143	2167	4018	11.79	$7.30e{-4}$	$8.01e{-4}$	$1.43e{-1}$	$9.74e{-4}$
15×133	2152	4005	10.24	$7.22e{-4}$	$7.92e{-4}$	$1.54e{-1}$	$9.64e{-4}$
48×42	2131	4080	1.01	$3.55e{-3}$	$4.20e{-3}$	$5.88e{-1}$	$7.08e{-3}$
Type (b) meshes							
3×572	2578	4004	165.11	$6.19e{-3}$	$7.60e{-3}$	$7.22e{+0}$	$1.38e{-2}$
4×445	2543	4005	96.34	$3.62e{-3}$	$4.39e{-3}$	$3.21e{+0}$	$7.81e{-3}$
5×364	2372	4004	63.04	$2.39e{-3}$	$2.87e{-3}$	$1.68e{+0}$	$5.00e{-3}$
12×160	2173	4000	$11.54^{(*2)}$	$7.47e{-4}$	$8.23e{-4}$	$1.53e{-1}$	$1.01e{-3}$
13×149	2175	4023	9.92	$7.35e{-4}$	$8.08e{-4}$	$1.60e{-1}$	$9.93e{-4}$
41×49	2125	4067	1.03	$3.50e{-3}$	$4.13e{-3}$	$5.87e{-1}$	$5.22e{-3}$
Type (c) meshes							
	2474	4101	100	$4.24e{-3}$	$5.34e{-3}$	$2.95e{+0}$	$1.91e{-2}$
	2426	4079	81.6	$3.55e{-3}$	$4.57e{-3}$	$2.48e{+0}$	$1.67e{-2}$
	2163	4036	11.78	$8.38e{-4}$	$9.62e{-4}$	$1.84e{-1}$	$2.42e{-3}$
	2127	4005	10	$8.34e{-4}$	$9.53e{-4}$	$1.87e{-1}$	$3.05e{-3}$
	2092	4036	1	$4.15e{-3}$	$5.21e{-3}$	$6.75e{-1}$	$2.01e{-2}$

Appendix. We list here some basic trigonometric formulas used to derive the results in the previous sections. Let α be any real number. Denote by $\alpha_i = 2(i-1)\pi/3 - \alpha, i = 1, 2, 3$.

$$\cos^2(\alpha_1) + \cos^2(\alpha_2) + \cos^2(\alpha_3) = \frac{3}{2}, \tag{A.1}$$

$$\cos(2\alpha_1) + \cos(2\alpha_2) + \cos(2\alpha_3) = 0, \tag{A.2}$$

$$\cos^2(2\alpha_1) + \cos^2(2\alpha_2) + \cos^2(2\alpha_3) = \frac{3}{2}, \tag{A.3}$$

$$\cos(\alpha_1)\cos(\alpha_2) + \cos(\alpha_1)\cos(\alpha_3) + \cos(\alpha_2)\cos(\alpha_3) = -\frac{3}{4}. \tag{A.4}$$

Similar relations hold for sine functions, too.

Let θ be any real number, and let $\beta = \alpha + \theta$. Then

$$\begin{aligned} &\cos(\alpha_1)\cos(\alpha_2)\cos\left(2\beta + \frac{\pi}{3}\right) + \cos(\alpha_1)\cos(\alpha_3)\cos\left(2\beta - \frac{\pi}{3}\right) \\ &+ \ \cos(\alpha_2)\cos(\alpha_3)\cos(2\beta + \pi) = -\frac{3}{4}\cos(2\theta), \end{aligned} \tag{A.5}$$

$$\begin{aligned} &\sin(\alpha_1)\sin(\alpha_2)\cos\left(2\beta + \frac{\pi}{3}\right) + \sin(\alpha_1)\sin(\alpha_3)\cos\left(2\beta - \frac{\pi}{3}\right) \\ &+ \ \sin(\alpha_2)\sin(\alpha_3)\cos(2\beta + \pi) = \frac{3}{4}\cos(2\theta), \end{aligned} \tag{A.6}$$

$$\begin{aligned} &\cos(\alpha_1)\cos(\alpha_2)\cos\left(4\beta - \frac{\pi}{3}\right) + \cos(\alpha_1)\cos(\alpha_3)\cos\left(4\beta + \frac{\pi}{3}\right) \\ &+ \ \cos(\alpha_2)\cos(\alpha_3)\cos(4\beta + \pi) = -\frac{3}{4}\cos(6\alpha + 4\theta), \end{aligned} \tag{A.7}$$

$$\begin{aligned} &\sin(\alpha_1)\sin(\alpha_2)\cos\left(4\beta - \frac{\pi}{3}\right) + \sin(\alpha_1)\sin(\alpha_3)\cos\left(4\beta + \frac{\pi}{3}\right) \\ &+ \ \sin(\alpha_2)\sin(\alpha_3)\cos(4\beta + \pi) = \frac{3}{4}\cos(6\alpha + 4\theta). \end{aligned} \tag{A.8}$$

REFERENCES

[1] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, 1999.

[2] I. Babuška and A. K. Aziz, *On the angle condition in the finite element method*, SIAM J. Numer. Anal., 13 (1976), pp. 214–226.

[3] R. E. Bank and R. K. Smith, *Mesh smoothing using a posteriori error estimates*, SIAM J. Numer. Anal., 34 (1997), pp. 979–997.

[4] M. Berzins, *A solution-based triangular and tetrahedral mesh quality indicator*, SIAM J. Sci. Comput., 19 (1998), pp. 2051–2060.

[5] J. Bramble and M. Zlámal, *Triangular elements in the finite element method*, Math. Comp., 24 (1970), pp. 809–820.

[6] E. F. D'Azevedo and R. B. Simpson, *On optimal interpolation triangle incidences*, SIAM J. Sci. Statist. Comput., 10 (1989), pp. 1063–1075.

[7] L. Formaggia and S. Perotto, *New anisotropic a priori error estimates*, Numer. Math., 89 (2001), pp. 641–667.

[8] W. Huang, *Measuring mesh qualities and application to variational mesh adaptation*, SIAM J. Sci. Comput., 26 (2005), pp. 1643–1666.

[9] W. Huang and W. Sun, *Variational mesh adaptation* II: *Error estimates and monitor functions*, J. Comput. Phys., 184 (2003), pp. 619–648.

[10] P. Knupp, L. G. Margolin, and M. Shashkov, *Reference Jacobian optimization-based rezone strategies for arbitrary Lagrange Eulerian methods*, J. Comput. Phys., 176 (2002), pp. 93–128.

[11] G. Kunert, *A Posteriori Error Estimation for Anisotropic Tetrahedral and Triangular Finite Element Meshes*, Ph.D. dissertation, TU Chemnitz, Chemnitz, Germany, 1999.

[12] E. J. Nadler, *Piecewise Linear Approximation on Triangulations of a Planar Region*, Ph.D. thesis, Division of Applied Mathematics, Brown University, Providence, RI, 1985.

[13] S. Rippa, *Long and thin triangles can be good for linear interpolation*, SIAM J. Numer. Anal., 29 (1992), pp. 257–270.

[14] J. R. Shewchuk, *Triangle: Engineering a 2D Quality Mesh Generator and Delaunay Triangulator*, Technical report, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, 1996.

[15] R. B. Simpson, *Anisotropic mesh transformations and optimal error control*, Appl. Numer. Math., 14 (1994), pp. 183–198.

[16] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

SIAM J. Numer. Anal.
Vol. 43, No. 1, pp. 41–57

INTERPOLATION AND APPROXIMATION OF PIECEWISE SMOOTH FUNCTIONS[*]

FRANCESC ARANDIGA[†], ALBERT COHEN[‡], ROSA DONAT[†], AND NIRA DYN[§]

Abstract. This paper provides approximation orders for a class of nonlinear interpolation procedures for uniformly sampled univariate data. The interpolation is based on essentially nonoscillatory (ENO) and subcell resolution (SR) reconstruction techniques. These nonlinear techniques aim at reducing significantly the approximation error for functions with isolated singularities and are therefore attractive for applications such as shock computations or image compression. We prove that in the presence of isolated singularities, the approximation order provided by the interpolation procedure is improved by a factor of h relative to the linear methods, where h is the sampling rate. Moreover, for h below a critical value, we recover the optimal approximation order as for uniformly smooth functions.

Key words. piecewise smooth functions, interpolation, ENO, subcell resolution, critical sampling rate

AMS subject classifications. 65D15, 65D05, 41A05, 41A10, 41A25

DOI. 10.1137/S0036142903426245

1. Introduction.

This paper is concerned with the analysis of a class of univariate high order interpolation and approximation techniques for piecewise smooth functions, introduced by Harten [10], namely, *essentially nonoscillatory* (ENO) and *subcell resolution* (SR) reconstructions. These methods automatically adapt near the singularities of the approximated function, and they are by essence data dependent and nonlinear.

While their initial motivation was in the context of finite volume methods for shock computations, ENO-SR methods have found natural applications in data compression algorithms, in particular through the development of multiscale decompositions, similar to wavelet expansions, which incorporate nonlinear reconstructions [11, 12, 4]. In such decompositions, the wavelet coefficients are interpreted as the errors between the sampled data and its reconstruction from a sampling at a twice coarser scale. When dealing with data sampled from a piecewise smooth function, the adaptive treatment of singularities results in more accurate reconstructions and therefore in sparser decompositions than when using standard wavelet basis. In recent years, ENO-SR techniques have been extended to two-dimensional (2D) image data, either by tensor product [1, 2, 6] or by intrinsically 2D reconstructions [14, 3]. Other related nonlinear multiscale representations have been introduced in [5] in the context of the lifting scheme.

From a theoretical point of view, the adaptive treatment of singularities allows

[*] Received by the editors April 17, 2003; accepted for publication (in revised form) June 17, 2004; published electronically April 26, 2005. This work was partially supported by EU-IHP project HPRN-CT-2002-00286.

http://www.siam.org/journals/sinum/43-1/42624.html

[†] Departamento de Matemàtica Aplicada, Universitat de València, C/ Dr. Moliner, 50, Burjasot, Valencia, Spain (arandiga@uv.es, http://gata.uv.es/~arandiga; donat@uv.es, http://gata.uv.es/~donat).

[‡] Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 175 Rue du Chevaleret, 75013 Paris, France (cohen@ann.jussieu.fr, http://www.ann.jussieu.fr/~cohen).

[§] School of Mathematical Sciences, Tel Aviv University, Ramat Aviv 69987, Israel (niradyn@math.tau.ac.il, http://www.math.tau.as.il/~niradyn).

us to expect strictly better approximation rates than with linear methods in the case of piecewise smooth functions and images. A rigorous analysis of this improvement in the one-dimensional (1D) case is the main objective of this paper. Our next perspective, which is the object of an ongoing work, is to study in a similar way the approximation properties of edge-adapted techniques for 2D functions and images such as those introduced in [14, 3]. It should well be understood that the 2D case is not a trivial generalization of the 1D case by tensor product technique: point discontinuities are then replaced by edges, which are not only characterized by their spatial locations but also by some geometrical features such as orientation and curvature. In turn, the development and analysis of 2D edge-adapted reconstruction strategies are significantly more involved, and the present paper can be viewed as an elementary yet instructive "starter" to this research program.

Consider at first the following situation: from a set of uniformly sampled data $(f(kh))_{k\in\mathbb{Z}}$, we are interested in building an interpolant $\mathcal{I}_h f$, i.e., a function such that $\mathcal{I}_h(kh) = f(kh)$ for all $k \in \mathbb{Z}$. There are many ways to build an interpolant $\mathcal{I}_h f$ of a prescribed order $m > 0$, i.e., such that if $f \in C^m$, one has

$$|\mathcal{I}_h f - f| \le C h^m \sup |f^{(m)}|. \tag{1}$$

Basically, one can do it with a linear operator $\mathcal{I}_h$ which is (i) local, (ii) exact for polynomials of degree $m-1$, and (iii) stable. We are interested in the interpolation of continuous functions f which are smooth everywhere except at isolated points. For such functions, we can only expect an error bound of order $\mathcal{O}(h)$ with a linear method, independently of its order.

In order to explain in a nutshell the principles of the ENO and SR techniques, first consider the following piecewise polynomial interpolation of the data $(f(kh))_{k\in\mathbb{Z}}$: to each interval

$$I_k := [kh, (k+1)h], \quad k \in \mathbb{Z}, \tag{2}$$

we attach the stencil S_k of size m around I_k, i.e.,

$$S_k := \{(k-m_1)h, \dots, (k+m_2)h\}, \tag{3}$$

where $m_1 \ge 0$ and $m_2 > 0$ are fixed integers such that $m_1 + m_2 = m - 1$. We define a unique polynomial $p_k \in \Pi_{m-1}$ which agrees with f on S_k. A linear interpolation operator is then defined by

$$\mathcal{I}_h f(x) = p_k(x), \quad x \in I_k. \tag{4}$$

This interpolant has accuracy of order m: if f is C^m on $[(k-m_1)h, (k+m_2)h]$, we have the estimate

$$\|f - \mathcal{I}_h f\|_{L^\infty(I_k)} \le C h^m \|f^{(m)}\|_{L^\infty([(k-m_1)h,(k+m_2)h])}. \tag{5}$$

Clearly, for a smooth function f with an isolated singularity of f' situated in the interval I_k, the order of accuracy is reduced to $\mathcal{O}(h)$ on all the intervals I_{k+l} for $l = -m_2+1, \dots, m_1$, due to the systematic use of a fixed stencil.

The principle of ENO interpolation is to allow for data-dependent stencils in order to reduce the influence of the singularity on the approximation. For this purpose, one typically introduces a measure of the oscillation of f on the stencil S_k. Since we are interested in detecting jump discontinuities in the first derivative, this measure is typically based on the evaluation of the second order differences,

$$\Delta_h^2 f(x) := f(x) - 2f(x+h) + f(x+2h) \tag{6}$$

for $x = (k-m_1)h, \dots, (k+m_2-2)h$. For each k, we select among all the stencils $\{S_{k-m_2+1}, \dots, S_{k+m_1}\}$ which contain I_k, the stencil $\tilde{S}_k$ which minimizes a chosen measure. The ENO interpolant is then given by

$$\mathcal{I}_h f(x) = \tilde{p}_k(x), \quad x \in I_k, \tag{7}$$

where $\tilde{p}_k$ is the polynomial which agrees with f on the stencil $\tilde{S}_k$. In comparison with the linear interpolation based on a fixed stencil, ENO interpolation has the same order of accuracy m and reduces the effect of an isolated singularity, since the selected stencil will tend to avoid it. We therefore expect that the precision only deteriorates on the interval which contains the singularity.

The goal of the SR technique is to improve the approximation properties of the interpolant even on this interval. It is based on a detection mechanism which labels as B (bad) an interval I_k which is suspected to contain a singularity, in the sense that the selected stencils for its immediate neighbors tend to avoid it. Thus I_k is B if $\tilde{S}_{k-1} = S_{k-m_2}$ and $\tilde{S}_{k+1} = S_{k+m_1+1}$. Other intervals are labeled as G (good). On a G interval I_k, we use the above-described ENO interpolation to define $\mathcal{I}_h f$. On a B interval I_k, we use the polynomials $\tilde{p}_{k-1}$ and $\tilde{p}_{k+1}$ to predict the location of the singularity: if these polynomials intersect at a single point a_k of I_k, we define for $x \in I_k$ the interpolant by

$$\mathcal{I}_h f(x) = \tilde{p}_{k-1}(x) \text{ if } x \le a_k, \;\; \tilde{p}_{k+1}(x) \text{ if } x \ge a_k. \tag{8}$$

In the case that these polynomials do not intersect at a single point of I_k, the interval is relabeled as G and the ENO interpolation is used.

An intuitive statement is that ENO-SR interpolation has accuracy of order $\mathcal{O}(h^m)$ for piecewise smooth functions. Some initial results suggesting the validity of this statement were given in [13] and in [10, 9]—for ENO and ENO-SR, respectively—in the context of building and analyzing high order schemes for conservation laws, and in [6] in the context of signal and image approximation and compression.

Our goal here is to investigate this statement in a rigorous way. For simplicity we consider functions which are smooth except at one unknown point a but are globally continuous. We also assume that $f \in C^m(\mathbb{R} \setminus \{a\})$ in the sense that its derivatives up to order m are uniformly bounded on $\mathbb{R} \setminus \{a\}$. Thus the derivatives of f have jumps $([f'], [f''], \dots)$ at the point a. Ideally we could hope for an estimate of the form

$$\|f - \mathcal{I}_h f\|_{L^\infty} \le C h^m \sup_{\mathbb{R}\setminus\{a\}} |f^{(m)}| \tag{9}$$

for all $h > 0$. Unfortunately, we shall see with a simple example that we cannot hope for such a result for $m > 2$. In fact (9) holds for h smaller than a fixed fraction of a critical scale h_c depending itself on the function f according to

$$h_c := \frac{[f']}{4 \sup_{x \in \mathbb{R}\setminus\{a\}} |f''(x)|}. \tag{10}$$

This critical scale h_c corresponds to the minimal level of resolution which ensures the detection of the singularity. To our knowledge, this notion was not used in previous works dealing with ENO and ENO-SR interpolation, particularly in the abovementioned references.

Therefore we can only achieve

$$\|f - \mathcal{I}_h f\|_{L^\infty} \le C h^m \sup_{\mathbb{R}\backslash\{a\}} |f^{(m)}|, \quad h \le K h_c(f). \tag{11}$$

We shall yet prove that we have for all $h > 0$ an estimate of the form

$$\|f - \mathcal{I}_h f\|_{L^\infty} \le C h^2 \sup_{\mathbb{R}\backslash\{a\}} |f''|, \tag{12}$$

i.e., at least a gain of one order of accuracy relative to any linear method. Note that since the interpolation process is local, our analysis applies also to the case of several isolated singularities which are sufficiently separated relative to the sampling scale (typically by mh).

When dealing with functions f which are piecewise smooth with an isolated jump discontinuity in the function itself, there is no more hope that a nonlinear reconstruction of f from its samples $f(kh)$ brings any improvement on the interval that contains the jump point, since the location of this point cannot be resolved at a finer resolution from these samples. Moreover these samples are not well defined a priori if f is not a continuous function. We should therefore replace the point value sampling by local averaging, in the sense that we are now given the cell averages $f_k^h := \frac{1}{h}\int_{kh}^{(k+1)h} f(t)dt$ for $k \in \mathbb{Z}$. We can build ENO-SR reconstruction procedures from such data in a way similar to that for point value data. In fact, reconstruction from cell averages can be derived by differentiating the point value interpolant obtained from the discrete primitive values $\sum_{l=0}^{k-1} f_l^h$. In turn, the results that we establish for piecewise smooth continuous functions in the point value setting can be used to establish similar results for piecewise smooth discontinuous functions in the cell average setting.

Note that most nonlinear approximation methods that deal with local singularities are based on either adaptive mesh refinement or wavelet thresholding (see, e.g., [8, 7] for surveys on such nonlinear approximation). A specific feature of the present approach is that it does not rely on any local refinement of the sampled data: the function is accurately reconstructed from a given uniform sampling, by a locally defined data dependent operator. A similar approach, yet based on different tools (in particular Fourier analysis), was developed in [15].

Our paper is organized as follows. We first show by an example in section 2 that one cannot hope for more than second order accuracy when a singularity occurs (still better than first order with linear methods). We introduce in section 3 a specific singularity detection mechanism together with an ENO-SR interpolation process, which slightly differs from the original ENO-SR, yet with the same basic principles, and we discuss the organization of the intervals which are detected by this mechanism. We prove in section 4 that detection always occurs for $h < h_c$ and that the position of the singularity is accurately estimated. We then use these results in section 5 to prove that our version of the ENO-SR interpolation technique has accuracy of order $\mathcal{O}(h^m)$ for h smaller than Kh_c, where $0 < K < 1$ is a fixed constant, and that it is second order accurate for all $h > 0$, which is the best that we can hope for according to the example of section 2. These findings are demonstrated in section 6 by numerical examples. Finally in section 7, we derive similar approximation results for piecewise smooth discontinuous functions in the cell average setting, measuring the error in the L^p norm as well as in the Hausdorff distance between graphs as a substitute to the L^∞ norm.

2. An instructive example. The following elementary example is meant to show that one cannot expect more than second order accuracy for a general class of continuous functions with jump discontinuities, as well as to illustrate the notion of critical scale.

Consider the functions f_+ and f_- which depend on $h_0 > 0$:

$$f_+(x) = 0 \text{ if } x < 0, \quad f_+(x) = x(x - h_0) \text{ if } x \geq 0, \tag{13}$$

and

$$f_-(x) = 0 \text{ if } x < h_0, \quad f_-(x) = x(x - h_0) \text{ if } x \geq h_0. \tag{14}$$

We notice that both functions agree on $h_0\mathbb{Z}$ so that if $\mathcal{I}_h$ is *any* interpolation operator on the grid $\mathbb{Z}h$, we have the following when $h = h_0$:

$$\mathcal{I}_h f_+ = \mathcal{I}_h f_-. \tag{15}$$

Since $\|f_+ - f_-\|_{L^\infty} = h^2/4$, by the triangle inequality we have either

$$\|f_+ - \mathcal{I}_h f_+\|_{L^\infty} \geq h^2/8 \geq \frac{h^2}{16} \sup_{x \in \mathbb{R} \setminus \{0\}} |f_+''| \tag{16}$$

or

$$\|f_- - \mathcal{I}_h f_-\|_{L^\infty} \geq h^2/8 \geq \frac{h^2}{16} \sup_{x \in \mathbb{R} \setminus \{h\}} |f_-''|. \tag{17}$$

Since we also have

$$\sup_{x \in \mathbb{R} \setminus \{h\}} |f_-^{(m)}| = \sup_{x \in \mathbb{R} \setminus \{0\}} |f_+^{(m)}| = 0, \quad m > 2, \tag{18}$$

this simple example shows us that (9) cannot be achieved with $m > 2$. Here h_0 plays the role of a *critical scale* above which singularities cannot be precisely detected. For $h \ll h_0$, our nonlinear interpolation method gives an exact reconstruction of f_+ and f_-. However, we certainly cannot ensure more than second order accuracy over all piecewise smooth functions and all $h > 0$.

3. A modified ENO-SR detection and interpolation mechanism. For a given approximation order m, our detection mechanism defines a set of intervals labeled B, which potentially contain the singularity, according to the following rules:

1. If

   $$|\Delta_h^2 f((k-1)h)| > |\Delta_h^2 f((k-1 \pm n)h)|, \quad n = 1, \dots, m. \tag{19}$$

   both I_{k-1} and I_k are labeled B. Notice that (19) indicates that the point kh lies at the center of the largest second divided difference (among those being compared). Hence either I_{k-1} or I_k could potentially contain the singularity.
2. If

   $$|\Delta_h^2 f(kh)| > |\Delta_h^2 f((k+n)h)|, \quad n = 1, \dots, m-1, \tag{20}$$

   and

   $$|\Delta_h^2 f((k-1)h)| > |\Delta_h^2 f((k-1-n)h)|, \quad n = 1, \dots, m-1, \tag{21}$$

   then I_k is labeled B. In this case the two largest divided differences involved in the comparison process include I_k, which is then a candidate to contain the singularity.

All other intervals are labeled G.

This detection mechanism is designed in such a way that for h sufficiently small, the interval I_k containing the singularity a is labeled B, while all intervals labeled G are in smooth regions of f. On the other hand it is also possible that an interval I_k might be labeled B in a smooth region at an arbitrarily small scale. In case of such false alarms, it is crucial that the polynomials which are used to construct the interpolation are built from stencils which only contain G intervals, i.e., from smooth regions. This is ensured by the following lemma, which describes the organization of the B and G intervals.

LEMMA 1. *The groups of adjacent B intervals are at most of size 2. They are separated by groups of adjacent G intervals which are at least of size $m-1$.*

Proof. Assume that I_0 and I_k are B with $1 < k < m$. We have three cases:

1. I_0 and I_k have been labeled B by the second rule. Then it follows that both $|\Delta_h^2 f(0)| > |\Delta_h^2 f((k-1)h)|$ and $|\Delta_h^2 f(0)| < |\Delta_h^2 f((k-1)h)|$, which is a contradiction.
2. I_0 has been labeled B by the second rule and I_k has been labeled B by the first rule. Then either I_{k-1} or I_{k+1} is also a B interval. Hence we obtain that both $|\Delta_h^2 f(0)| > |\Delta_h^2 f(qh)|$ and $|\Delta_h^2 f(0)| < |\Delta_h^2 f(qh)|$ for some $q \in \{k-1, k\}$, which is a contradiction. The case where I_0 has been labeled B by the first rule and I_k has been labeled B by the second rule is treated in a similar way.
3. I_0 and I_k have been labeled B by the first rule; hence each one is a member of a B-pair (two adjacent B intervals). Hence we obtain that both $|\Delta_h^2 f(ph)| > |\Delta_h^2 f(qh)|$ and $|\Delta_h^2 f(ph)| < |\Delta_h^2 f(qh)|$ for some $p \in \{-1, 0\}$ and $q \in \{k-1, k\}$, which is a contradiction.

We therefore obtain that no two B intervals can have a difference of indices strictly between 1 and m, which concludes the proof. □

Remark. Our detection mechanism is based only on comparing second order divided differences and is therefore different from the hierarchical mechanism originally proposed by Harten in [10]. The main motivation is that the final goal of our procedure is really to detect singularities and isolate them sufficiently from false alarms in order to derive our approximation results in an elementary way. Harten's detection mechanism needs to be modified to ensure Lemma 1; we need to compare a larger number of divided differences than that necessary for the stencil selection mechanism. The validity of Lemma 1 is crucial in order to obtain the desired approximation result, presented in Theorem 1. In fact, one can prove that this result does not hold when using the detection mechanism of [10], although the counterexamples seem to be of a pathological nature and are seldom observed numerically.

Based on the above-described detection mechanism, we propose the following interpolation procedure:

1. If I_k is a G interval, define $\mathcal{I}_h f$ on I_k as a polynomial p_k of degree $m-1$ obtained by interpolation of f on a stencil $\{ph, \dots, (p+m-1)h\}$ such that $p \le k < k+1 \le p+m-1$ and such that this stencil contains only G intervals. Such a stencil always exists, according to Lemma 1, yet is not unique. In practice, we may choose the stencil which is the most centered around the interval I_k or we may use the standard ENO procedure.
2. If I_k is an isolated B interval, we obtain polynomials p_k^- and p_k^+ of degree $m-1$ by interpolation of f on the stencils $\{(k-m+1)h, \dots, kh\}$ and $\{(k+1)h, \dots, (k+m)h\}$ and use them to predict the location of the singularity: if these polynomials intersect at a single point y of I_k, then for $x \in I_k$ we define

the interpolant by

$$\mathcal{I}_h f(x) = p_k^-(x) \text{ if } x \le y, \;\; p_k^+(x) \text{ if } x \ge y. \tag{22}$$

In the case where these polynomials do not intersect at a single point of I_k, the interval is relabeled G and we return to the previous case.

3. If (I_k, I_{k+1}) is a B-pair, we treat $I_k \cup I_{k+1}$ as I_k in the previous case; i.e., we obtain polynomials p_k^- and p_{k+1}^+ of degree $m-1$ by interpolation of f at stencils $\{(k-m+1)h, \ldots, kh\}$ and $\{(k+2)h, \ldots, (k+m+1)h\}$ and use them to predict the location of the singularity: if these polynomials intersect at a single point y of $I_k \cup I_{k+1}$, then for $x \in I_k \cup I_{k+1}$ we define the interpolant by

$$\mathcal{I}_h f(x) = p_k^-(x) \text{ if } x \le y, \;\; p_{k+1}^+(x) \text{ if } x \ge y. \tag{23}$$

In the case where these polynomials do not intersect at a single point of $I_k \cup I_{k+1}$, both intervals are relabeled G and we return to the first case.

Note that the interpolation operator $\mathcal{I}_h f$ described above does not make use of the data at midpoints of B-pairs. Hence $\mathcal{I}_h f$ does not interpolate f at these points. This is a specific feature of our modified ENO-SR interpolation which greatly facilitates the proof of our main approximation result in section 5.

4. Properties of the detection mechanism.

The goal of this section is to establish some properties of the detection mechanism which will be used in section 5 for proving the improved approximation order of $\mathcal{I}_h f$ announced in the introduction.

The properties are expressed by two lemmas. The first one ensures that the singularity is always detected under some critical scale.

LEMMA 2. *Let f be a globally continuous function with a bounded second derivative on $\mathbb{R} \setminus \{a\}$ and a discontinuity in the first derivative at a point a. Define the critical scale*

$$h_c := \frac{|[f']|}{4 \sup_{x \in \mathbb{R}\setminus\{a\}} |f''(x)|}, \tag{24}$$

where $[f']$ is the jump of the first derivative f' at the point a. Then for $h < h_c$, the interval I_k which contains a is labeled B. Moreover, if a is close to one endpoint of the interval I_k by at most a quarter of its size, then the interval adjacent to this endpoint is also labeled B.

Proof. Without loss of generality, we can assume that a is located on the first half of the interval I_0, i.e., $0 \le a \le h/2$. For $k > 0$ and $k < -1$, we find that

$$|\Delta_h^2 f(kh)| \le h^2 \sup_{x \in \mathbb{R}\setminus\{a\}} |f''(x)|. \tag{25}$$

For $k = -1$ and $k = 0$, the second order finite differences can be estimated by decomposing f into

$$f(x) = f_1(x) + f_2(x), \tag{26}$$

with $f_1(x) = [f'](x-a)_+$ and $f_2(x)$ a C^1 function with a bounded second derivative on $\mathbb{R} \setminus \{a\}$, such that

$$\sup_{x \in \mathbb{R}\setminus\{a\}} |f_2''(x)| = \sup_{x \in \mathbb{R}\setminus\{a\}} |f''(x)|. \tag{27}$$

We therefore have for all $k \in \mathbb{Z}$

$$|\Delta_h^2 f_2(kh)| \le h^2 \sup_{x \in \mathbb{R} \setminus \{a\}} |f''(x)|. \tag{28}$$

On the other hand, we have

$$|\Delta_h^2 f_1(-h)| = |(h-a)[f']|. \tag{29}$$

It follows that

$$|\Delta_h^2 f(-h)| \ge |(h-a)[f']| - h^2 \sup_{x \in \mathbb{R} \setminus \{a\}} |f''(x)|, \tag{30}$$

and therefore

$$|\Delta_h^2 f(-h)| \ge \frac{h}{2}|[f']| - h^2 \sup_{x \in \mathbb{R} \setminus \{a\}} |f''(x)|. \tag{31}$$

So if $h < h_c$, we get by (24)

$$|\Delta_h^2 f(-h)| > h^2 \sup_{x \in \mathbb{R} \setminus \{a\}} |f''(x)| \tag{32}$$

Combining this with (25), we find that if $h < h_c$

$$|\Delta_h^2 f(-h)| > |\Delta_h^2 f(kh)| \tag{33}$$

for $k < -1$ and $k > 0$. In the case where $|\Delta_h^2 f(-h)| > |\Delta_h^2 f(0)|$, we find that I_{-1} and I_0 are a B-pair according to the first detection rule. Otherwise, if $|\Delta_h^2 f(-h)| \le |\Delta_h^2 f(0)|$, we find that I_0 must be labeled B according to the second detection rule.

Finally, we notice that

$$|\Delta_h^2 f_1(0)| = |a[f']|, \tag{34}$$

so that

$$|\Delta_h^2 f(0)| \le |a[f']| + h^2 \sup_{x \in \mathbb{R} \setminus \{a\}} |f''(x)|. \tag{35}$$

Therefore, combining (35) and (30), we are always in the case of I_{-1} and I_0 constituting a B-pair whenever

$$2h^2 \sup_{x \in \mathbb{R} \setminus \{a\}} |f''(x)| < (h-2a)|[f']|, \tag{36}$$

which holds whenever $h < h_c$ and $a < h/4$. □

The next lemma expresses the fact that the location of the singularity is accurately estimated when h is less than a fixed fraction of the critical scale.

LEMMA 3. *There exist constants $C > 0$ and $0 < K < 1$ such that for all continuous f with uniformly bounded mth derivative on $\mathbb{R} \setminus \{a\}$ and for $h < Kh_c$ with h_c defined by* (24)*, the following holds:*

1. *The singularity a is contained in an isolated B interval I_k (Case 1) or in a B-pair (I_k, I_{k+1}) (Case 2).*

2. *The two polynomials* (p_k^-, p_k^+) *(Case* 1*) or* (p_k^-, p_{k+1}^+) *(Case* 2*) which are used in the definition of* $\mathcal{I}_h f$ *have only one intersection point* y *inside* I_k *(Case* 1*) or inside* $I_k \cup I_{k+1}$ *(Case* 2*).*
3. *The distance between* a *and* y *is bounded by*

$$|a-y| \leq C \frac{h^m \sup_{\mathbb{R}\setminus\{a\}} |f^{(m)}|}{|[f']|}. \tag{37}$$

Proof. Since $K < 1$, the first statement has already been proved in Lemma 2. Without loss of generality, we assume that $0 \leq a \leq h/2$. In this case we know by Lemma 2 that I_0 is B for $h < h_c$. For the sake of notational simplicity we denote by $I = [b, c]$ the interval where we perform the subcell resolution process, which is either I_0 (Case 1) or $I_{-1} \cup I_0$ (Case 2) or $I_0 \cup I_1$ (Case 2). By Lemma 2, we are ensured that $I = I_{-1} \cup I_0$ when $a < h/4$, and therefore

$$\min\{|a-b|, |a-c|\} \geq h/4. \tag{38}$$

We also denote by (p_-, p_+) the polynomials which are used in the subcell resolution of I. Finally we note that for any $2 \leq k \leq m$ we can write

$$f = f_- \chi_{]-\infty,a]} + f_+ \chi_{[a,+\infty[}, \tag{39}$$

where f^- and f^+ are functions which are globally C^k over $\mathbb{R}$ and such that

$$\sup_{x\in\mathbb{R}} |f_\pm^{(k)}(x)| \leq \sup_{x\in\mathbb{R}\setminus\{a\}} |f^{(k)}(x)|. \tag{40}$$

For example, we can define these functions by extension of f using its left or right Taylor expansion of order k at the point a. In order to prove the second statement of the lemma, we choose $k = 2$. We note that p_- and p_+ can also be viewed as Lagrange interpolation of f_- and f_+. It then follows from classical results on Lagrange interpolation that there exists a constant D independent of f such that for all $t \in I$,

$$|f_\pm(t) - p_\pm(t)| \leq Dh^2 \sup_{x\in\mathbb{R}} |f_\pm''(x)| = Dh^2 \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|, \tag{41}$$

and

$$|f_\pm'(t) - p_\pm'(t)| \leq Dh \sup_{x\in\mathbb{R}} |f_\pm''(x)| = Dh \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|. \tag{42}$$

Since $|t-a| \leq 2h$ when $t \in I$, we also have

$$|f_\pm'(t) - f_\pm'(a)| \leq 2h \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|, \tag{43}$$

and therefore we get from (42)

$$|f_\pm'(a) - p_\pm'(t)| \leq (D+2)h \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|, \quad t \in I. \tag{44}$$

It follows that for all $t \in I$,

$$|p_+'(t) - p_-'(t)| \geq |[f']| - 2(D+2)h \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|. \tag{45}$$

Thus, for $h < \frac{2}{D+2}h_c$ the function $p_+ - p_-$ is strictly monotone on I and has at most one root. Therefore, we are ensured that p_+ and p_- intersect at most at a single point inside I. In order to prove that this point y exists, we need to show that $p_+ - p_-$ has a sign change inside I. Without loss of generality, assume here that $[f'] > 0$. By second order Taylor expansion at the point a, we find that

$$(f_+ - f_-)(b) \leq -(a-b)[f'] + (a-b)^2 \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|, \tag{46}$$

and

$$(f_+ - f_-)(c) \geq (c-a)[f'] - (c-a)^2 \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|. \tag{47}$$

Combining with (41), we thus obtain

$$(p_+ - p_-)(b) \leq -(a-b)[f'] + ((a-b)^2 + 2Dh^2) \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|, \tag{48}$$

and

$$(p_+ - p_-)(c) \geq (c-a)[f'] - ((c-a)^2 + 2Dh^2) \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|. \tag{49}$$

Using (38), we therefore obtain

$$(p_+ - p_-)(b) \leq -\frac{h}{4}[f'] + (2D+4)h^2 \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|, \tag{50}$$

and

$$(p_+ - p_-)(c) \geq \frac{h}{4}[f'] - (2D+4)h^2 \sup_{x\in\mathbb{R}\setminus\{a\}} |f''(x)|. \tag{51}$$

It follows that for $h < \frac{1}{4D+8}h_c$, we have

$$(p_+ - p_-)(b) \leq -\frac{h}{8}[f'] < 0, \tag{52}$$

and

$$(p_+ - p_-)(c) \geq \frac{h}{8}[f'] > 0, \tag{53}$$

so that there exists a single intersection point $y \in I$. So defining $K := \frac{1}{4D+8}$, we have proved the two first statements of the lemma.

In order to prove the third statement (37), we now choose $k = m$ in the definition of the extensions f_+ and f_-. It again follows from classical results on Lagrange interpolation that there exists a constant $\tilde{D}$ such that for all $t \in I$,

$$|f_\pm(t) - p_\pm(t)| \leq \tilde{D}h^m \sup_{x\in\mathbb{R}\setminus\{a\}} |f^{(m)}(x)|, \tag{54}$$

and therefore, if we define $g = f_+ - f_-$ and $q = p_+ - p_-$, we obtain, for all $t \in I$,

$$|g(t) - q(t)| \leq 2\tilde{D}h^m \sup_{x\in\mathbb{R}\setminus\{a\}} |f^{(m)}(x)|, \tag{55}$$

so that for $t = a$,

$$|q(a)| \leq 2\tilde{D}h^m \sup_{x\in\mathbb{R}\setminus\{a\}} |f^{(m)}(x)|. \tag{56}$$

Now note that for $t \in I$ and $h < Kh_c$, by (45)

$$|q'(t)| = \frac{|[f']|}{2}, \tag{57}$$

and therefore

$$|q(a)| = |q(y) - q(a)| \geq |y - a|\frac{|[f']|}{2}. \tag{58}$$

Combining this with (56), we therefore obtain (37) with $C = 4\tilde{D}$. □

5. Approximation properties of $\mathcal{I}_h$. We are now ready to derive our main approximation result.

THEOREM 1. *For all continuous f with derivatives up to degree m uniformly bounded on $\mathbb{R} \setminus \{a\}$, the nonlinear interpolant $\mathcal{I}_h f$ satisfies*

$$\|f - \mathcal{I}_h f\|_{L^\infty} \leq Ch^2 \sup_{\mathbb{R}\setminus\{a\}} |f''| \tag{59}$$

for all $h > 0$, with $C > 0$ independent of f. Moreover there exists $0 < K < 1$ independent of f such that, for $h < Kh_c$ with h_c defined by (24), we have

$$\|f - \mathcal{I}_h f\|_{L^\infty} \leq Ch^m \sup_{\mathbb{R}\setminus\{a\}} |f^{(m)}|. \tag{60}$$

Proof. We choose for K the constant in Lemma 3. Note first that for $h < Kh_c$, according to Lemmas 1 and 2, all the polynomials which are used in the construction of $\mathcal{I}_h$ are built from stencils over which the function is smooth. It follows from classical results on Lagrange interpolation that the estimate

$$|f(x) - \mathcal{I}_h f(x)| \leq Ch^m \sup_{\mathbb{R}\setminus\{a\}} |f^{(m)}| \tag{61}$$

holds whenever x belongs to a G interval or to an isolated B interval or B-pair which does not contain a (i.e., false alarms do not deteriorate the convergence rate). Let us now assume that x belongs to the group of adjacent B intervals which contains a. Here, we shall assume, again without loss of generality, that $0 \leq a \leq h/2$ and use the notation $I = [b, c]$, p_+, p_-, f_+, f_- that were introduced in the proof of Lemma 3. We also assume that $a \leq y$, the case $y \leq a$ being treated in a similar way. For $x \in [b, a]$, we have the estimate

$$|f(x) - \mathcal{I}_h f(x)| = |f_-(x) - p_-(x)| \leq Ch^m \sup_{\mathbb{R}\setminus\{a\}} |f^{(m)}|, \tag{62}$$

and for $x \in [y, c]$, we have the estimate

$$|f(x) - \mathcal{I}_h f(x)| = |f_+(x) - p_+(x)| \leq Ch^m \sup_{\mathbb{R}\setminus\{a\}} |f^{(m)}|. \tag{63}$$

It remains to consider the case $a < x < y$. In this case, we have

$$|f(x) - \mathcal{I}_h f(x)| = |f_+(x) - p_-(x)| \leq |f_+(x) - f_-(x)| + |f_-(x) - p_-(x)|. \tag{64}$$

The second term is again bounded by $Ch^m \sup_{\mathbb{R}\backslash\{a\}} |f^{(m)}|$. For the first term, we use second order Taylor expansion to derive

$$\begin{aligned}|f_+(x) - f_-(x)| &\le |[f']|(y-a) + (y-a)^2 \sup_{\mathbb{R}\backslash\{a\}} |f''| \\ &\le (y-a)(|[f']| + h \sup_{\mathbb{R}\backslash\{a\}} |f''|).\end{aligned}$$

Since $h < h_c$, this gives

$$|f_+(x) - f_-(x)| \le \frac{5}{4}|[f']|](y-a). \tag{65}$$

Combining this with (37) of Lemma 3, we also obtain the bound $Ch^m \sup_{\mathbb{R}\backslash\{a\}} |f^{(m)}|$ for $|f_+(x) - f_-(x)|$, which concludes the proof in the case $h < Kh_c$.

In the case $h \ge Kh_c$, the estimate

$$|f(x) - \mathcal{I}_h f(x)| \le Ch^m \sup_{\mathbb{R}\backslash\{a\}} |f^{(m)}| \tag{66}$$

is guaranteed to hold only when x is at distance not less than $(m+1)h$ from a. We also have the lower order estimate

$$|f(x) - \mathcal{I}_h f(x)| \le Ch^2 \sup_{\mathbb{R}\backslash\{a\}} |f''|. \tag{67}$$

Let us now prove that this estimate remains valid if $|x-a| \le (m+1)h$. For this purpose we consider the decomposition $f = f_1 + f_2$ used in the proof of Lemma 2. The errors of polynomial interpolation of f_1 and f_2 are, respectively, bounded by $Ch|[f'|]$ and $Ch^2 \sup_{\mathbb{R}\backslash\{a\}} |f''|$. Since $h \ge Kh_c$ the second bound dominates the first one so that the above estimate is valid. The proof of the theorem is now complete. □

6. Numerical examples. We consider the functions

$$f_\varepsilon(x) = \begin{cases} (x - \pi/6)(x - \pi/6 - \varepsilon) + \sin(\pi x/8)/8, & x < \pi/6, \\ \sin(\pi x/8)/8 & \text{otherwise} \end{cases} \tag{68}$$

for four values of ε $(2^{-6}, 2^{-8}, 2^{-10}, 2^{-12})$. Each of these functions is globally continuous with a jump of ε in its first derivative at the point $\pi/6$, while its higher derivatives are uniformly bounded independently of ε. The f_ε are piecewise polynomial functions similar to the example in section 2, to which we add the smooth function $\sin(\pi x/8)/8$ in order to avoid a 0 approximation error for h less than a critical scale.

We apply our technique with $m = 4$. For different values of h we compute $\mathcal{I}_h f_\varepsilon$ for the four values of ε and plot log $\| f_\varepsilon - \mathcal{I}_h f_\varepsilon \|_\infty$ versus $-\log(h)$ in Figure 1. In Table 1 we give the corresponding values of $\| f_\varepsilon - \mathcal{I}_h f_\varepsilon \|_\infty$.

We observe that we always have accuracy of order $\mathcal{O}(h^2)$ and that for h smaller than some h_ε, we recover the optimal accuracy of order $\mathcal{O}(h^4)$. This is ensured by Theorem 1 for $h < Kh_c$, where

$$h_c := \frac{|[f'_\varepsilon]|}{4 \sup_{x\in\mathbb{R}-\{\pi/6\}} |f''_\varepsilon(x)|} = \varepsilon/8$$

and $0 < K < 1$. However, we observe that $h_\varepsilon > \varepsilon/8$, which means that the order of accuracy $\mathcal{O}(h^4)$ is attained even for some $h > Kh_c$.

Remark. We have performed the same test for the original SR technique of Harten [10], modified in order to ensure that Lemma 1 is satisfied. The behavior in terms of approximation errors and orders is absolutely similar to that shown in Figure 1 and Table 1.

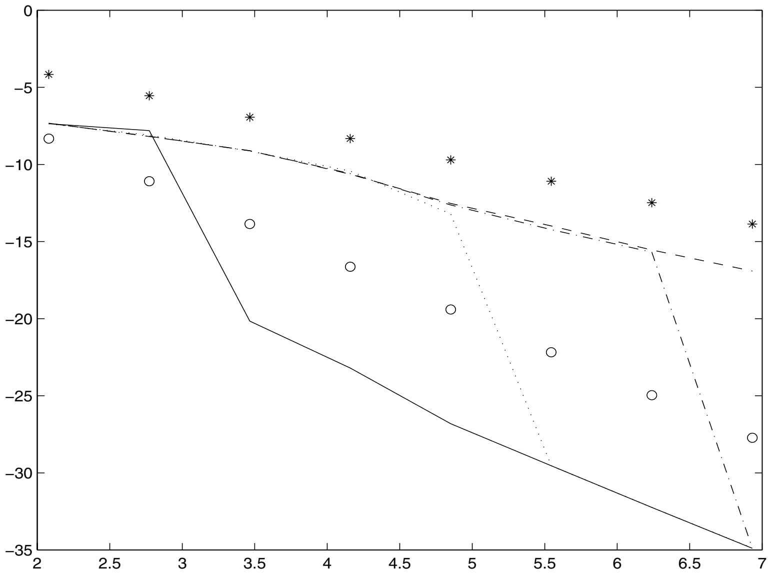


FIG. 1. *Solid, dotted, dash-dot, dashed lines: log* $\| f - I_h f_\varepsilon \|_\infty$ *versus* $-log(h)$ *for the different values of* ε *(*2^{-6}*,* 2^{-8}*,* 2^{-10}*, and* 2^{-12}*). Circles: log*(h^4) *versus* $-log(h)$*. Stars: log*(h^2) *versus* $-log(h)$*.*

TABLE 1
$\| f_\varepsilon - I_h f_\varepsilon \|_\infty$ *for different values of* ε *and different values of* h*.*

$\varepsilon \setminus h$	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-9}	2^{-10}
2^{-6}	6e-04	4e-04	2e-09	8e-11	2e-12	1e-13	1e-14	7e-16
2^{-8}	6e-04	3e-04	1e-04	3e-05	1e-06	1e-13	1e-14	7e-16
2^{-10}	7e-04	3e-04	1e-04	2e-05	3e-06	7e-07	2e-07	7e-16
2^{-12}	7e-04	3e-04	1e-04	2e-05	4e-06	8e-07	2e-07	4e-08

7. Approximation of discontinuous functions. In this last section, we derive similar approximation results for piecewise smooth functions with jump discontinuities, sampled by their cell averages

$$f_k^h := \frac{1}{h}\int_{kh}^{(k+1)h} f(t)dt. \tag{69}$$

For such data, the reconstruction $\mathcal{A}_h f$ is defined on each interval I_k by a polynomial or piecewise polynomial function such that its cell average on I_k coincides with f_k^h. In the case of a linear method, a polynomial q_k of degree $m-2$ is uniquely determined by the $f_{k-m_1}^h, \dots, f_{k+m_2-1}^h$, where $m_1 \geq 0$ and $m_2 > 0$ are fixed integers such that $m_1 + m_2 = m - 1$.

A simple connection can be established between cell average and point value reconstructions: the polynomial p_k of degree m which interpolates the point values of the primitive $F(x) = \int_0^x f(t)dt$ at the points $(k-m_1)h, \dots, (k+m_2)h$ satisfies

$$p_k' = q_k. \tag{70}$$

Since these point values are given by the discrete primitive

$$F(kh) = h\sum_{l=0}^{k-1} f_l^h,\ k>0,\ \ F(kh) = h\sum_{l=k}^{-1} f_l^h,\ k<0 \ \text{ and } \ F(0)=0, \tag{71}$$

the cell average–based reconstruction operator $\mathcal{A}_h$ by polynomials of degree $m-2$ can thus be interpreted as

$$\mathcal{A}_h = \mathcal{D}\mathcal{I}_h\mathcal{P}_d, \tag{72}$$

where $\mathcal{P}_d$ is the discrete primitive operator $\mathcal{P}_d(\{f_k^h\}) = \{F(kh)\}$, $\mathcal{I}_h$ is the point value interpolation operator of degree $m-1$, and $\mathcal{D}$ is the continuous differentiation $\mathcal{D}f = f'$ (see [4]). The choice of our ENO-SR point value interpolation of degree $m-1$ as $\mathcal{I}_h$ in the above formula leads to a natural definition of an ENO-SR cell average–based reconstruction operator $\mathcal{A}_h$.

We can use (72) in order to derive approximation results for $\mathcal{A}_h$ applied to piecewise smooth discontinuous functions from the approximation result of section 5. Here we shall analyze the error in L^p for $1 \le p < \infty$ on a fixed bounded interval I, since we necessarily have an $\mathcal{O}(1)$ error in the L^∞ norm near the discontinuity. We assume here that f is C^{m-1} on $\mathbb{R}\setminus\{a\}$ in the sense that its derivatives up to degree $m-1$ are uniformly bounded on $\mathbb{R}\setminus\{a\}$, with a jump $[f]$ at the point a. Therefore, its approximation $\mathcal{A}_h f$ by a linear cell average–based reconstruction will have $\mathcal{O}(1)$ accuracy on the intervals I_k which are such that a is contained in $[(k-m_1)h,(k+m_2)h]$, and $\mathcal{O}(h^{m-1})$ elsewhere, resulting in a global L^p error behaving like

$$\|f-\mathcal{A}_h f\|_{L^p(I)} \le C\max\{h^{m-1}, h^{1/p}\}. \tag{73}$$

If we now consider the nonlinear approximation $\mathcal{A}_h f$ obtained by our ENO-SR reconstruction, we know that

$$\mathcal{A}_h f = (\mathcal{I}_h F)', \tag{74}$$

where F is the primitive of f. Since F is C^m on $\mathbb{R}\setminus\{a\}$, defining the critical scale as

$$h_c := \frac{|[F']|}{4\sup_{x\in\mathbb{R}\setminus\{a\}}|F''(x)|} = \frac{|[f]|}{4\sup_{x\in\mathbb{R}\setminus\{a\}}|f'(x)|}, \tag{75}$$

we obtain the same conclusions as in Lemma 3 for $h < Kh_c$. We can then revisit the proof of Theorem 1 as follows.

For $h < Kh_c$, classical results on Lagrange interpolation show that the estimate

$$|F'(x) - (\mathcal{I}_h F)'(x)| \le Ch^{m-1}\sup_{\mathbb{R}\setminus\{a\}}|F^{(m)}|, \tag{76}$$

or, equivalently,

$$|f(x) - \mathcal{A}_h f(x)| \le Ch^{m-1}\sup_{\mathbb{R}\setminus\{a\}}|f^{(m-1)}|, \tag{77}$$

holds whenever x belongs to a G interval or to a group of adjacent B intervals which do not contain a (i.e., false alarms do not deteriorate the convergence rate). When x belongs to the group of adjacent B intervals which contains a, the same analysis as in the proof of Theorem 1 shows that this estimate remains valid when x is not located on the interval $[a, y]$ (or $[y, a]$) on which we cannot avoid an $\mathcal{O}(1)$ error. However, since this interval has its size estimated by

$$|a - y| \leq C \frac{h^m \sup_{\mathbb{R}\setminus\{a\}} |F^{(m)}|}{|[F']|} = C \frac{h^m \sup_{\mathbb{R}\setminus\{a\}} |f^{(m-1)}|}{|[f]|}, \tag{78}$$

we obtain a global L^p error behaving like

$$\|f - \mathcal{A}_h f\|_{L^p(I)} \leq C \max\{h^{m-1}, h^{m/p}\}. \tag{79}$$

For $h \geq K h_c$, the estimate

$$|f(x) - \mathcal{A}_h f(x)| \leq C h^{m-1} \sup_{\mathbb{R}\setminus\{a\}} |f^{(m-1)}| \tag{80}$$

is valid only when x is at distance at least $(m+1)h$ from a. For $|x - a| \leq (m+1)h$, we consider the decomposition $F = F_1 + F_2$ used in the proof of Lemma 2, for which we have

$$|F_1'(x) - (\mathcal{I}_h F_1)'(x)| \leq C|[F']| = C|[f]|, \tag{81}$$

and

$$|F_2'(x) - (\mathcal{I}_h F_2)'(x)| \leq C h \sup_{\mathbb{R}\setminus\{a\}} |F''| = C h \sup_{\mathbb{R}\setminus\{a\}} |f'|. \tag{82}$$

Since $h \geq K h_c$ the second bound dominates the first one. It follows that we obtain a global L^p error behaving like

$$\|f - \mathcal{A}_h f\|_{L^p(I)} \leq C \max\{h^{m-1}, h^{1+1/p}\}. \tag{83}$$

Note that both estimates (83) and (79) constitute an improvement of (73). Combining these estimates, we find that for all $h > 0$

$$\|f - \mathcal{A}_h f\|_{L^p(I)} \leq C \max\{h^{m-1}, h^{m/p}, h^{1+1/p}\}. \tag{84}$$

In the case $p = 1$ and $m > 1$, we find a statement very similar to that obtained in the point value setting: while approximation by a linear method behaves like $\mathcal{O}(h)$, our ENO-SR approximation behaves like $\mathcal{O}(h^{m-1})$ for $h < K h_c$ and like $\mathcal{O}(h^2)$ for all h.

The estimate (84) degenerates, however, to $\mathcal{O}(1)$ for $p = \infty$, which reflects the fact that we cannot hope to approximate a function with a jump in the L^∞ norm. For such discontinuous functions, a natural substitute to the L^∞ norm is given by the Hausdorff distance between graph, namely,

$$d(f, g) := d_{\mathcal{H}}(G_f, G_g), \tag{85}$$

where G_f and G_g are the completed graph of f and g and

$$d_{\mathcal{H}}(A, B) := \sup_{x \in A} \inf_{y \in B} |x - y| + \sup_{x \in B} \inf_{y \in A} |x - y| \tag{86}$$

is the Hausdorff distance between sets A and B of $\mathbb{R}^2$ (here $|x - y|$ is the Euclidean distance between x and y in $\mathbb{R}^2$). Recall that the completed graph of a discontinuous function f which admits left and right limits at its jumps consists of the points

$(x, f(x))$ when f is continuous at x and of the vertical segments $[f(x-0), f(x+0)]$ at the jump points x.

The approximation accuracy in the Hausdorff distance therefore measures the closeness in the L^∞ norm away from the jumps and the accuracy in resolving the exact location of the discontinuity. Note that this distance also penalizes the Gibbs phenomenon, which is the persistence of oscillations at fixed amplitude in the approximation near the discontinuity. In turn when applying a linear cell average–based reconstruction of degree $m-2>0$ to a piecewise smooth function with a jump discontinuity, we cannot expect any convergence in this distance and we thus have only $d(f, \mathcal{A}_h f) = \mathcal{O}(1)$.

Consider now the nonlinear approximation $\mathcal{A}_h f$ obtained by our ENO-SR reconstruction. For $h < Kh_c$, we have already seen that f is approximated from both sides with precision (77) and the jump point is approximated with accuracy (78). In this case, the Gibbs phenomenon is avoided and we obtain the asymptotical behavior

$$d(f, \mathcal{A}_h f) \leq Ch^{m-1}. \tag{87}$$

For $h \geq Kh_c$, the detection of the jump discontinuity is not ensured. In this case, we crudely estimate the Hausdorff distance by the L^∞ norm and obtain

$$d(f, \mathcal{A}_h f) \leq \|f - \mathcal{A}_h\|_{L^\infty} \leq Ch \sup_{x \in \mathbb{R} \setminus \{a\}} |f'(x)| + C|[f]|, \tag{88}$$

where the second term accounts for the Gibbs phenomenon generated by the singularity. Since $h \geq Kh_c$, the first term is dominant, and we therefore obtain for all $h > 0$ the estimate

$$d(f, \mathcal{A}_h f) \leq Ch. \tag{89}$$

The fact that we can obtain convergence rates in the Hausdorff distance is a very nice specific feature of ENO-SR nonlinear reconstruction, which reflects in particular its nonoscillatory nature.

We can summarize our findings on the cell average reconstruction of piecewise smooth functions with jump discontinuities in the following statement.

Theorem 2. *For all f with derivatives up to degree $m-1$ uniformly bounded on $\mathbb{R} \setminus \{a\}$, the nonlinear reconstruction $\mathcal{A}_h f$ satisfies*

$$\|f - \mathcal{A}_h f\|_{L^p(I)} \leq C \max\{h^{m-1}, h^{m/p}, h^{1+1/p}\}, \tag{90}$$

and

$$d(f, \mathcal{A}_h f) \leq Ch \tag{91}$$

for all $h > 0$. Moreover there exists $0 < K < 1$ independent of f such that for $h < Kh_c$ with h_c defined by (75)*, we have*

$$\|f - \mathcal{A}_h f\|_{L^p(I)} \leq C \max\{h^{m-1}, h^{m/p}\}, \tag{92}$$

and

$$d(f, \mathcal{A}_h f) \leq Ch^{m-1}. \tag{93}$$

Acknowledgment. The authors are grateful to the anonymous reviewers for their constructive criticism and for indicating to us some related works.

REFERENCES

[1] S. Amat, F. Arandiga, A. Cohen, and R. Donat, *Tensor product multiresolution with error control*, Signal Processing, 82 (2002), pp. 587–608.

[2] S. Amat, F. Arandiga, A. Cohen, R. Donat, G. Garcia, and M. Von Oehsen, *Data compression with ENO schemes—A case study*, Appl. Comput. Harmon. Anal., 11 (2001), pp. 273–288.

[3] F. Arandiga, A. Cohen, M. Doblas, and B. Matei, *Edge adapted nonlinear multiscale transforms for compact image representation*, in Proceedings of the IEEE International Conference of Image Processing, Vol. 1, Barcelona, 2003, pp. 701–704.

[4] F. Arandiga and R. Donat, *Nonlinear multiscale decompositions: The approach of A. Harten*, Numer. Algorithms, 23 (2000), pp. 175–216.

[5] R. Baraniuk, R. Claypoole, G. M. Davis, and W. Sweldens, *Nonlinear wavelet transforms for image coding via lifting*, IEEE Trans. Image Process., 12 (2003), pp. 1449–1459.

[6] T. Chan and H. Zhou, *ENO-Wavelet Transforms and Some Applications*, UCLA CAM report 02-50, 2002.

[7] A. Cohen, *Numerical Analysis of Wavelet Methods*, Elsevier, Amsterdam, 2003.

[8] R. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.

[9] R. Donat, *Studies on error propagation for certain nonlinear approximations to hyperbolic equations: Discontinuities in derivatives*, SIAM J. Numer. Anal., 31 (1994), pp. 655–679.

[10] A. Harten, *ENO schemes with subcell resolution*, J. Comput. Phys., 83 (1989), pp. 148–184.

[11] A. Harten, *Discrete multiresolution analysis and generalized wavelets*, J. Appl. Numer. Math., 12 (1993), pp. 153–192.

[12] A. Harten, *Multiresolution representation of data: A general framework*, SIAM J. Numer. Anal., 33 (1996), pp. 1205–1256.

[13] A. Harten, S. Osher, B. Engquist, and S. Chakravarthy, *Some results on uniformly high order accurate essentially non-oscillatory schemes*, Appl. Numer. Math., 2 (1986), pp. 347–377.

[14] B. Matei, *Méthodes Multirésolution Non-linéaires—Applications au traitement d'image*, Ph.D. dissertation, Université Paris, VI, 2002.

[15] M. Vetterli, P. Marziliano, and T. Blu, *Sampling signals with finite rate of innovation*, IEEE Trans. Signal Process., 50 (2002), pp. 1417–1428.

SIAM J. Numer. Anal.
Vol. 43, No. 1, pp. 58–75

HERMITE SPECTRAL METHODS WITH A TIME-DEPENDENT SCALING FOR PARABOLIC EQUATIONS IN UNBOUNDED DOMAINS[*]

HEPING MA[†], WEIWEI SUN[‡], AND TAO TANG[§]

Abstract. Hermite spectral methods are investigated for linear diffusion equations and nonlinear convection-diffusion equations in unbounded domains. When the solution domain is unbounded, the diffusion operator no longer has a compact resolvent, which makes the Hermite spectral methods unstable. To overcome this difficulty, a time-dependent scaling factor is employed in the Hermite expansions, which yields a positive bilinear form. As a consequence, stability and spectral convergence can be established for this approach. The present method plays a similar role in the stability of the similarity transformation technique proposed by Funaro and Kavian [*Math. Comp.*, 57 (1991), pp. 597–619]. However, since coordinate transformations are not required, the present approach is more efficient and is easier to implement. In fact, with the time-dependent scaling the resulting discretization system is of the same form as that associated with the classical (straightforward but unstable) Hermite spectral method. Numerical experiments are carried out to support the theoretical stability and convergence results.

Key words. Hermite spectral method, time-dependent scaling, stability, convergence

AMS subject classifications. 65N30, 76D99

DOI. 10.1137/S0036142903421278

1. Introduction.

Spectral methods for approximating solutions of differential equations in unbounded domains have received considerable attention, mainly due to their high accuracy and being free from using artificial boundary conditions. The spectral approaches employ orthogonal systems in unbounded domains, e.g., using the Laguerre spectral methods for problems in semibounded or exterior domains [2, 4, 8, 12, 17, 18, 21] and the Hermite spectral methods for the problems in unbounded domains [1, 5, 6, 7, 10, 20]. An alternative approximation for such problems is the rational spectral method which has also been studied by several authors [3, 9, 11, 13, 25].

When the Hermite method is applied to second-order differential equations directly, it is found in [7] that the nonsymmetric bilinear form is not of the desired coercity property. To see this, let us consider the following simple parabolic problem:

[*] Received by the editors January 13, 2003; accepted for publication (in revised form) April 28, 2004; published electronically April 26, 2005.
http://www.siam.org/journals/sinum/43-1/42127.html

[†] Department of Mathematics, Shanghai University, Shanghai 200436, China (hpma@mail.shu.edu.cn). The research of this author was supported by the Hong Kong Research Grants Council (project CityU 1084/02P), National Science Foundation of China (project NSFC 10471089), and Special Funds for Major Specialities of Shanghai Education Committee. The major portion of this research was carried out while this author was visiting the City University of Hong Kong.

[‡] Department of Mathematics, The City University of Hong Kong, Kowloon, Hong Kong (maweiw@math.cityu.edu.hk). The research of this author was supported by the Hong Kong Research Grants Council (project CityU 1084/02P).

[§] Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong, and Institute of Computational Mathematics, The Chinese Academy of Sciences, Beijing 100080, China (ttang@math.hkbu.edu.hk, ttang@lsec.cc.ac.cn). The research of this author was supported by the Hong Kong Research Grants Council (projects HKBU 2044/00P and HKBU2083/01P) and the International Research Team on Complex System of Chinese Academy of Sciences.

$$\begin{cases} \partial_t U - \nu \partial_x^2 U = f(x,t), & x \in \mathbb{R}, \quad t > 0, \\ U(x,0) = U_0(x), & x \in \mathbb{R}, \end{cases} \tag{1.1}$$

where the diffusion constant $\nu > 0$, and $\mathbb{R} = (-\infty, \infty)$. The solution U and its partial derivative $\partial_x U$ have to satisfy certain decay conditions as $|x| \to \infty$. Let $\mathbb{P}_N(\mathbb{R})$ be the space of polynomials of degree at most N and let

$$V_N = \{ v_N(x) = \omega_\beta \phi_N(x) \mid \phi_N(x) \in \mathbb{P}_N(\mathbb{R}) \}, \tag{1.2}$$

where $\omega_\beta = \mathrm{e}^{-(\beta x)^2}$ with β being a constant. The semidiscrete Hermite function method for (1.1) is to find $u_N(t) \in V_N$ such that for any $\varphi_N \in \mathbb{P}_N(\mathbb{R})$,

$$\begin{cases} (\partial_t u_N(t), \varphi_N) + \nu(\partial_x u_N(t), \partial_x \varphi_N) = (f(t), \varphi_N), & t > 0, \\ (u_N(0), \varphi_N) = (U_0, \varphi_N), \end{cases} \tag{1.3}$$

where $(\cdot,\cdot)$ is the conventional inner product in the $L^2(\mathbb{R})$ space.

We demonstrate that neither is the nonsymmetric bilinear form in (1.3) coercive nor can a corresponding Gårding's type inequality be established. To show this, we denote by $H_l(x)$ the Hermite polynomial of degree l orthogonal on $\mathbb{R}$ with respect to the weight $\omega_1(x) = e^{-x^2}$. Let $\beta > 0$ and let

$$\underline{H}_l(x) := (2^l l! \sqrt{\pi})^{-1/2} H_l(x), \qquad H_l^{(\beta)}(x) := \sqrt{\beta}\, \underline{H}_l(\beta x).$$

Note that $\|\underline{H}_l\|_{\omega_1} = 1$ and $\|H_l^{(\beta)}\|_{\omega_\beta} = 1$. Then, for $u_N = \omega_\beta \phi_N$ with $\phi_N := \sum_{l=0}^N \hat{u}_l H_l^{(\beta)}$, we have

$$\begin{aligned} (\partial_x u_N, \partial_x \phi_N) &= |\phi_N|^2_{1,\omega_\beta} + \beta^2 \|\phi_N\|^2_{\omega_\beta} - 2\beta^4 \|x\phi_N\|^2_{\omega_\beta} \\ &= -2\beta^2 \sum_{l=2}^N \sqrt{l(l-2)}\, \hat{u}_l \hat{u}_{l-2}, \end{aligned} \tag{1.4}$$

which cannot be controlled by $\|u_N\|^2_{\omega_\beta^{-1}} = \sum_{l=0}^N |\hat{u}_l|^2$. In other words, the stability for (1.3) cannot be established by using the classical energy method. On the other hand, the instability is observed numerically, as seen in section 6. To overcome this difficulty, a *similarity transformation* was introduced by Funaro and Kavian [6], which is defined by

$$s = \ln(1+t), \qquad y = x(1+t)^{-\frac{1}{2}}. \tag{1.5}$$

With this transformation, they were able to obtain the optimal error estimate of the Hermite function approximation for the linear problem (1.1). This similarity transformation technique has been extended recently to study the nonlinear convection-diffusion equations; see, e.g., [7, 10]. By using this transformation, the diffusion operator in (1.1) is changed into an operator whose eigenfunctions are the Hermite functions. This property can lead to a desired stability result. However, the transformation may make the underlying equations more complicated, which leads to difficulties in theoretical analysis and practical implementation. It is desirable to develop some simpler and more efficient Hermite spectral methods.

In this paper, we present a Petrov–Galerkin Hermite spectral method which uses a time-dependent weight function. On the one hand, the method keeps the advantage of the similarity transformation method, namely, it gives a positive definite bilinear form. On the other hand, the scheme can be easily formulated in the classical form of (1.3), without introducing any extra new terms. As a result, a priori explicit transformation is not needed. Moreover, the time-dependent weight function behaves like a spatial scaling. The importance of the scaling factor has been demonstrated by Tang [22] and Schumer and Holloway [20]. We will apply the proposed method to the analysis of the nonlinear convection-diffusion equations. Stability and optimal error estimates for the Hermite spectral methods, in both semidiscrete and fully discrete forms, are obtained for the nonlinear equation. It will be shown by numerical experiments that the time-dependent weight works well for solutions with time-dependent and time-independent decays.

An outline of the paper is as follows. In section 2 we briefly discuss the Hermite spectral methods with a time-dependent scaling. Section 3 presents some basic properties of the Hermite functions in weighted spaces, which will be useful in the stability and convergence analysis. In sections 4 and 5, stability and convergence analysis is carried out for the semidiscrete and fully discrete schemes, respectively. The analysis is devoted not only to the linear parabolic equation (1.1), but also to the nonlinear convection-diffusion problems. In section 6, numerical results will be presented.

2. Hermite method with time-dependent scaling. We present a Petrov–Galerkin Hermite spectral method with a time-dependent scaling for the simple model problem (1.1). Let $\alpha = \alpha(t) > 0$. We take

$$\alpha(t) = \frac{1}{2\sqrt{\nu\delta_0(\delta t + 1)}}, \tag{2.1}$$

where δ_0 and δ are some positive parameters. It can be verified that

$$\alpha'(t) = -2\nu\delta_0\delta\alpha^{3(t)}.$$

The motivation for this choice of α can be found in Remark 4.1 in section 4. The semidiscrete Hermite spectral method for (1.1) is to find $u_N(t) \in V_N(t)$ such that for any $\varphi_N \in \mathbb{P}_N(\mathbb{R})$,

$$\begin{cases} (\partial_t u_N(t), \varphi_N) + \nu(\partial_x u_N(t), \partial_x \varphi_N) = (f(t), \varphi_N), & t > 0, \\ (u_N(0), \varphi_N) = (U_0, \varphi_N), \end{cases} \tag{2.2}$$

where the trial space $V_N(t)$ is defined by

$$V_N(t) = \left\{ v_N(x) = \omega_{\alpha(t)}\phi_N(x) \mid \phi_N(x) \in \mathbb{P}_N(\mathbb{R}) \right\}. \tag{2.3}$$

The scheme (2.2) is almost the same as (1.3): the only difference is that here the weight function ω_α in the trial function space V_N varies with time. The scheme (2.2) can be rewritten as

$$\frac{d}{dt}(u_N(t), \varphi_N(t)) + (u_N(t), L^*\varphi_N(t)) = (f(t), \varphi_N(t)), \tag{2.4}$$

where $L^* := -\partial_t - \nu\partial_x^2$. To simplify the computation, let

(2.5)

$$u_N(x,t) = \frac{\omega_\alpha}{\sqrt{\pi}} \sum_{l=0}^{N} \hat{u}_l(t) H_l(\alpha x), \qquad \varphi_N(x,t) = \frac{\alpha(t)}{(2^m m!)} H_m(\alpha(t)x) \quad (0 \le m \le N).$$

In other words, we expand the unknown solution using the scaled Hermite functions with a time-dependent scaling factor. The test function φ_N is now also dependent on t. It can be verified that

$$\begin{aligned}
&(\omega_\alpha H_l(\alpha x), L^*(\alpha H_m(\alpha x)))\\
&\quad= -\alpha'\alpha^{-1}(\|H_m\|^2_{\omega_1}\delta_{lm} + (yH_l, H'_m)_{\omega_1}) - 2\nu\alpha^2(yH_l, H'_m)_{\omega_1} + \nu\alpha^2\|H'_m\|^2_{\omega_1}\delta_{lm}\\
&\quad= \delta_0\delta\nu\alpha^2\|H_m\|^2_{\omega_1}\delta_{lm} + (\delta_0\delta-1)\nu\alpha^2(H_{l+1}+2lH_{l-1}, 2mH_{m-1})_{\omega_1} + \nu\alpha^2|H_m|^2_{1,\omega_1}\delta_{lm}\\
&\quad= \nu\alpha^2 2^m m!\sqrt{\pi}(2\delta_0\delta\delta_{lm} + (\delta_0\delta-1)(\delta_{(l+2)m}+2m\delta_{lm}) + 2m\delta_{lm}).
\end{aligned}$$

Applying the above result to (2.2) gives

$$\tag{2.6}\begin{cases} \dfrac{d\mathbf{u}(t)}{dt} + \nu\alpha(t)^2\mathbf{A}\mathbf{u}(t) = \mathbf{f}(t), \quad t>0,\\ (\mathbf{u}(0))_m = \alpha(0)(2^m m!)^{-1}(U_0, H_m(\alpha(0)x)), \quad 0\le m\le N, \end{cases}$$

where $\alpha(0) = 1/2\sqrt{\nu\delta_0}$, $\mathbf{u} = (\hat{u}_0, \hat{u}_1, \dots, \hat{u}_N)^T$. The elements of the matrix $\mathbf{A}$ are given by

$$(\mathbf{A})_{ml} = \begin{cases} 2(m+1)\delta_0\delta, & l=m,\\ \delta_0\delta-1, & l=m-2,\\ 0 & \text{otherwise},\ 0\le l,m\le N, \end{cases}$$

and the entries f_m of $\mathbf{f}$ are given by

$$\begin{aligned}\hat{f}_m := (\mathbf{f})_m &= \alpha(2^m m!)^{-1}(f, H_m(\alpha x))\\ &= (2^m m!)^{-1}(\mathrm{e}^{y^2}f(\alpha^{-1}y), H_m(y))_{\omega_1}.\end{aligned}$$

Fully discrete methods can be designed by using (2.6) based on the method-of-lines approach. Here we consider the Crank–Nicolson scheme. Let τ be the time-step $t_k = k\tau$ $(k=0,1,\dots,n_T;\ T=n_T\tau)$, and let $\mathbf{v}^k = \mathbf{v}(t_k)$. The fully discrete Petrov–Galerkin method for (1.1) is to find

$$u_N^k = \frac{\omega_{\alpha(t_k)}}{\sqrt{\pi}}\sum_{l=0}^N \hat{u}_l^k H_l(\alpha(t_k)x)$$

such that

$$\tag{2.7}\begin{cases} \dfrac{\mathbf{u}^{k+1}-\mathbf{u}^k}{\tau} + \nu\alpha^2(t_k+\tau/2)\mathbf{A}\dfrac{\mathbf{u}^{k+1}+\mathbf{u}^k}{2} = \dfrac{\mathbf{f}^{k+1}+\mathbf{f}^k}{2}, & 0\le k\le n_T-1,\\ (\mathbf{u}^0)_m = (2^m m!)^{-1}(\mathrm{e}^{y^2}U_0(y/\alpha(0)), H_m(y))_{\omega_1}, & 0\le m\le N. \end{cases}$$

Since the matrix $\mathbf{A}$ is independent of time, the above scheme can be solved easily.

Remark 2.1. Note that the matrix $\mathbf{A}$ is an upper triangular matrix whose diagonal entries are $2(m+1)\delta_0\delta$. By the classical stability theory, both the semidiscrete scheme (2.6) and fully discrete scheme (2.7) are stable and convergent provided that $\delta_0\delta>0$. However, $\delta_0=0$ in the classical approach (1.3) yields numerical instability.

3. Approximation properties of Hermite functions. In this section, we present some basic approximation properties for the Hermite functions and the Hermite polynomials. Some of them are similar to those obtained in [5, 6, 7, 19, 23, 24] and we will only briefly outline the proofs.

Let $H^\sigma(\mathbb{R}) := W^{\sigma,2}(\mathbb{R})$ be the Sobolev spaces with the norm $\|\cdot\|_\sigma$ and seminorm $|\cdot|_\sigma$. For a nonnegative weight $\omega(x)$ on $\mathbb{R}$, the inner product and norm of $L^2_\omega(\mathbb{R})$ are denoted by $(\cdot,\cdot)_\omega$ and $\|\cdot\|_\omega$, respectively. The subscript ω will be dropped whenever $\omega(x) \equiv 1$. For a positive integer σ, the weighted Sobolev space $H^\sigma_\omega(\mathbb{R})$ is defined by

$$H^\sigma_\omega(\mathbb{R}) = \left\{ v \mid \partial^r_x v \in L^2_\omega(\mathbb{R}), \quad 0 \le r \le \sigma \right\}$$

with the seminorm and norm

$$|v|_{\sigma,\omega} = \|\partial^\sigma_x v\|_\omega, \qquad \|v\|_{\sigma,\omega} = \left(\sum_{r=0}^{\sigma} |v|^2_{r,\omega} \right)^{1/2}.$$

Denote by $H_l(x)$ the Hermite polynomial of degree l:

$$H_l(x) = (-1)^l \omega_1^{-1}(x) \partial^l_x(\omega_1(x)).$$

In theoretical analysis, it seems more convenient to use the normalized Hermite polynomials

$$\underline{H}_l(x) := (2^l l! \sqrt{\pi})^{-1/2} H_l(x).$$

We will work with the scaled Hermite polynomial $H^{(\beta)}_l(x) := \sqrt{\beta}\underline{H}_l(\beta x)$, where $\beta > 0$ is a constant. For nonnegative integers r and l, let

$$A^r_l = \begin{cases} l!/(l-r)!, & l \ge r \ge 1, \\ 1, & l \ge 0, r = 0, \\ 0, & l < r. \end{cases}$$

We have

$$(3.1) \qquad (\partial^r_x H^{(\beta)}_l, \partial^r_x H^{(\beta)}_m)_{\omega_\beta} = \beta^{2r}(\partial^r_x \underline{H}_l, \partial^r_x \underline{H}_m)_{\omega_1} = (2\beta^2)^r \sqrt{A^r_l A^r_m}\, \delta_{lm}$$

so that $\{\partial^r_x H^{(\beta)}_l\}$ are orthogonal on $\mathbb{R}$ with respect to the weight $\omega_\beta = e^{-(\beta x)^2}$. Let $P^\beta_N : L^2_{\omega_\beta}(\mathbb{R}) \to \mathbb{P}_N(\mathbb{R})$ be the $L^2_{\omega_\beta}$-orthogonal projection operator defined by

$$(3.2) \qquad (P^\beta_N v - v, \varphi_N)_{\omega_\beta} = 0 \qquad \forall\, \varphi_N \in \mathbb{P}_N(\mathbb{R}).$$

For $v \in H^r_{\omega_\beta}(\mathbb{R})$ $(r < N)$, we have $\partial^r_x P^\beta_N v = P^\beta_{N-r} \partial^r_x v$ and

$$(3.3) \qquad (\partial^r_x(P^\beta_N v - v), \varphi_{N-r})_{\omega_\beta} = 0 \qquad \forall\, \varphi_{N-r} \in \mathbb{P}_{N-r}(\mathbb{R}).$$

We consider the approximation by the Hermite functions; i.e., we approximate $v\omega_\beta^{-1}$ by using the Hermite polynomials. Let $\mathcal{P}^\beta_N : L^2_{\omega_\beta^{-1}}(\mathbb{R}) \to V_N$ be the $L^2_{\omega_\beta^{-1}}$-orthogonal projection operator defined by

$$(3.4) \qquad (\mathcal{P}^\beta_N v - v, \varphi_N)_{\omega_\beta^{-1}} = 0 \qquad \forall\, \varphi_N \in V_N.$$

It is easy to verify that $\mathcal{P}_N^\beta v = \omega_\beta P_N^\beta(v\omega_\beta^{-1})$. For $v \in H^r_{\omega_\beta^{-1}}(\mathbb{R})$ $(r < N)$, we have $\partial_x^r \mathcal{P}_N^\beta v = \mathcal{P}_{N+r}^\beta \partial_x^r v$ and

$$(\partial_x^r(\mathcal{P}_N^\beta v - v), \varphi_{N+r}) = 0 \qquad \forall\, \varphi_{N+r} \in \mathbb{P}_{N+r}(\mathbb{R}). \tag{3.5}$$

Lemma 3.1. *If r is a nonnegative integer, then $v \in H^r_{\omega_\beta^{-1}}(\mathbb{R})$ is equivalent to $v\omega_\beta^{-1} \in H^r_{\omega_\beta}(\mathbb{R})$. Moreover,*

$$\sum_{j=0}^{r}(2\beta^2)^{r-j}\|\partial_x^j[(I-P_m^\beta)(v\omega_\beta^{-1})]\|^2_{\omega_\beta} \le \|\partial_x^r[(I-\mathcal{P}_m^\beta)v]\|^2_{\omega_\beta^{-1}} \quad \forall m \ge 0, \tag{3.6}$$

$$\|\partial_x^r[(I-\mathcal{P}_N^\beta)v]\|_{\omega_\beta^{-1}} \le C(r)\|\partial_x^r[(I-P_N^\beta)(v\omega_\beta^{-1})]\|_{\omega_\beta} \quad \forall N > r, \tag{3.7}$$

where $P_0^\beta = \mathcal{P}_0^\beta = \mathbf{0}$ and $C(r)$ is a constant depending only on r.

Proof. By a direct calculation,

$$\partial_x^r(\omega_\beta H_l^{(\beta)}(x)) = (-\beta)^r 2^{r/2}\sqrt{A^r_{l+r}}\,\omega_\beta H^{(\beta)}_{l+r}(x). \tag{3.8}$$

Using this result we can verify that $\{\partial_x^r(\omega_\beta H_l^{(\beta)})\}$ are orthogonal with respect to the weight ω_β^{-1} on $\mathbb{R}$:

$$\begin{aligned}(\partial_x^r(\omega_\beta H_l^{(\beta)}), \partial_x^r(\omega_\beta H_m^{(\beta)}))_{\omega_\beta^{-1}} &= (2\beta^2)^r\sqrt{A^r_{l+r}A^r_{m+r}}(H^{(\beta)}_{l+r}, H^{(\beta)}_{m+r})_{\omega_\beta}\\ &= (2\beta^2)^r\sqrt{A^r_{l+r}A^r_{m+r}}\,\delta_{lm} \quad \forall l,\ m \ge r \ge 0.\end{aligned} \tag{3.9}$$

Let $v = \omega_\beta\sum_{l=0}^\infty \hat{v}_l H_l^{(\beta)}$. Then we have

$$(I-\mathcal{P}_m^\beta)v = \omega_\beta(I-P_m^\beta)(v\omega_\beta^{-1}) = \omega_\beta\sum_{l\ge m}\hat{v}_l H_l^{(\beta)}\,.$$

The above result, together with (3.9), gives

$$\begin{aligned}&\|\partial_x^r[(I-\mathcal{P}_m^\beta)v]\|^2_{\omega_\beta^{-1}} = (2\beta^2)^r\sum_{l\ge m}A^r_{l+r}|\hat{v}_l|^2\\ &\ge (2\beta^2)^r\sum_{l\ge m}\sum_{j=0}^{r}A_l^j|\hat{v}_l|^2 \ge (2\beta^2)^r\sum_{j=0}^{r}\sum_{l\ge\max\{m,j\}}A_l^j|\hat{v}_l|^2\\ &= \sum_{j=0}^{r}(2\beta^2)^{r-j}\|\partial_x^j[(I-P_m^\beta)(v\omega_\beta^{-1})]\|^2_{\omega_\beta}.\end{aligned}$$

This proves the result (3.6). The inequality (3.7) can be established similarly. □

Lemma 3.2. *If $0 \le r \le \sigma < N$, then*

$$\|\partial_x^r(v-P_N^\beta v)\|_{\omega_\beta} \le C(r,\sigma)(2\beta^2N)^{(r-\sigma)/2}\|\partial_x^\sigma v\|_{\omega_\beta} \qquad \forall\, v \in H^\sigma_{\omega_\beta}(\mathbb{R}), \tag{3.10}$$

$$\|\partial_x^r(v-\mathcal{P}_N^\beta v)\|_{\omega_\beta^{-1}} \le C(r,\sigma)(2\beta^2N)^{(r-\sigma)/2}\|\partial_x^\sigma v\|_{\omega_\beta^{-1}} \qquad \forall\, v \in H^\sigma_{\omega_\beta^{-1}}(\mathbb{R}), \tag{3.11}$$

where $C(r,\sigma)$ is a constant depending only on r and σ.

Proof. Let $v = \sum\limits_{l=0}^{\infty} \hat{v}_l H_l^{(\beta)}$. Then, it follows from (3.1) that

$$\begin{aligned}
\|\partial_x^r(v - P_N^\beta v)\|_{\omega_\beta}^2 &= \sum_{l>N} (2\beta^2)^r A_l^r |\hat{v}_l|^2 \\
&= (2\beta^2)^{r-\sigma} \sum_{l>N} (A_{l-r}^{\sigma-r})^{-1} (2\beta^2)^\sigma A_l^\sigma |\hat{v}_l|^2 \le (2\beta^2)^{r-\sigma} (A_{N+1-r}^{\sigma-r})^{-1} \|\partial_x^\sigma(v - P_N^\beta v)\|_{\omega_\beta}^2 \\
&= (2\beta^2)^{r-\sigma} \prod_{m=r-1}^{\sigma-2} \left(1 - \frac{m}{N}\right)^{-1} N^{r-\sigma} \|\partial_x^\sigma (v - P_N^\beta v)\|_{\omega_\beta}^2,
\end{aligned}$$

which gives (3.10). Using (3.7), (3.10), and (3.6) gives

$$\begin{aligned}
\|\partial_x^r(v - \mathcal{P}_N^\beta v)\|_{\omega_\beta^{-1}}^2 &\le C(r) \|\partial_x^r[(I - P_N^\beta)(v\omega_\beta^{-1})]\|_{\omega_\beta} \\
&\le C(r,\sigma)(2\beta^2 N)^{(r-\sigma)/2} \|\partial_x^\sigma[(I - P_N^\beta)(v\omega_\beta^{-1})]\|_{\omega_\beta} \\
&\le C(r,\sigma)(2\beta^2 N)^{(r-\sigma)/2} \|\partial_x^\sigma[(I - \mathcal{P}_N^\beta)v]\|_{\omega_\beta^{-1}},
\end{aligned}$$

which gives (3.11). □

LEMMA 3.3. *Let r, σ be nonnegative integers. We have*

$$\lim_{|x|\to\infty} x(\partial_x^r v)^2(x)\omega_\beta^{-1}(x) \to 0 \qquad \forall\, v \in H^\sigma_{\omega_\beta^{-1}}(\mathbb{R}),\ r \le \sigma - 1, \tag{3.12}$$

$$\|v^2\omega_\beta^{-1}\|_{L^\infty(\mathbb{R})} \le 2|v|_{1,\omega_\beta^{-1}} \|v\|_{\omega_\alpha^{-1}} \qquad \forall\, v \in H^1_{\omega_\beta^{-1}}(\mathbb{R}), \tag{3.13}$$

$$|\varphi_N|_{\sigma,\omega_\beta^{-1}} \le (4\beta^2 N)^{(\sigma-r)/2} |\varphi_N|_{r,\omega_\beta^{-1}} \qquad \forall\, \varphi_N \in V_N,\ r \le \sigma \le N, \tag{3.14}$$

$$\left\| \sqrt{\omega_\beta^{-1}}\varphi_N \right\|_{L^\infty(\mathbb{R})} \le 2(\beta^2 N)^{1/4} \|\varphi_N\|_{\omega_\beta^{-1}} \qquad \forall\, \varphi_N \in V_N,\ r \le \sigma \le N. \tag{3.15}$$

Proof. The first two results, (3.12) and (3.13), can be obtained by the arguments similar to those given in [5, 7]. Let $\varphi_N = \omega_\beta \sum_{l=0}^N \hat{\varphi}_l H_l^{(\beta)} \in V_N$. It follows from (3.9) that

$$\begin{aligned}
|\varphi_N|_{\sigma,\omega_\beta^{-1}}^2 &= (2\beta^2)^{\sigma-r} (2\beta^2)^r \sum_{l=0}^N A_{l+\sigma}^{\sigma-r} A_{l+r}^r |\hat{\varphi}_l|^2 \\
&\le (2\beta^2 N)^{\sigma-r} \prod_{j=r+1}^{\sigma} \left(1 + \frac{j}{N}\right) |\varphi_N|_{r,\omega_\beta^{-1}}^2,
\end{aligned}$$

which gives (3.14). Moreover, using (3.13) and (3.14) gives

$$\|\omega_\beta^{-1}\varphi_N^2\|_{L^\infty(\mathbb{R})} \le 2(4\beta^2 N)^{1/2} \|\varphi_N\|_{\omega_\beta^{-1}}^2 \le 4\beta N^{1/2} \|\varphi_N\|_{\omega_\beta^{-1}}^2.$$

This completes the proof of this lemma. □

4. Stability and convergence: Semidiscretization. To demonstrate the stability and convergence analysis for the proposed spectral method, we take the time-dependent weight

$$\omega_{\alpha(t)} = \mathrm{e}^{-(\alpha(t)x)^2}, \tag{4.1}$$

where $\alpha(t)$ is defined by (2.1). We expand

$$u_N(x,t) = \omega_{\alpha(t)} \sum_{l=0}^{N} \hat{u}_l(t) H_l^{(\alpha(t))}(x). \tag{4.2}$$

It can be verified that $\|u_N\|_{\omega_\alpha^{-1}} = \|\mathbf{u}\|$. The solution expansion (4.2) is slightly different from the one in (2.5) but is more suitable for theoretical analysis. With this expansion, the matrix form for the scheme (2.2) becomes

$$\frac{d\mathbf{u}}{dt} + \nu\alpha(t)^2 \mathbf{B}\mathbf{u} = \mathbf{f}, \tag{4.3}$$

where the elements of the matrices $\mathbf{B}$ and the vector $\mathbf{f}$ are given by

$$\begin{aligned} (\mathbf{B})_{ml} &= \begin{cases} \delta_0\delta(2m+1), & l = m, \\ (\delta_0\delta - 1)2\sqrt{m(m-1)}, & l = m-2, \\ 0 & \text{otherwise,} \end{cases} \\ \hat{f}_m &:= (\mathbf{f})_m = (f, H_m^{(\alpha)}), \qquad\qquad 0 \le l, m \le N. \end{aligned} \tag{4.4}$$

The stability and convergence properties can be established following the discussions in section 2. To be more precise, let

$$\underline{\delta} = \min\{1, 2\delta_0\delta - 1\} > 0, \quad \mathbf{D} = 2\mathrm{diag}(0, 1, \dots, N) \tag{4.5}$$

and let $\mathbf{I}$ be the identity matrix. Since

$$\mathbf{u}^T\mathbf{B}\mathbf{u} \ge \underline{\delta}\mathbf{u}^T(\mathbf{D}+\mathbf{I})\mathbf{u},$$

we obtain

$$\begin{aligned} &\|\mathbf{u}(t)\|^2 + \underline{\delta}\nu \int_0^t \alpha^2 \|(\mathbf{D}+\mathbf{I})^{1/2}\mathbf{u}(s)\|^2\, ds \\ &\quad\le \|\mathbf{u}(0)\|^2 + 4\delta_0\underline{\delta}^{-1} \int_0^t (\delta s + 1)\|(\mathbf{D}+\mathbf{I})^{-1/2}\mathbf{f}(s)\|^2\, ds, \quad t > 0, \end{aligned} \tag{4.6}$$

or, equivalently,

$$\begin{aligned} &\|u_N(t)\|^2_{\omega_\alpha^{-1}} + \underline{\delta}\nu \int_0^t |u_N(t)|^2_{1,\omega_\alpha^{-1}}\, ds \\ &\qquad \le \|u_N(0)\|^2_{\omega_\alpha^{-1}} + (\underline{\delta}\nu)^{-1} \int_0^t \|\partial_x^{-1} f(s)\|^2_{\omega_\alpha^{-1}}\, ds, \end{aligned} \tag{4.7}$$

where $\partial_x^{-1} v(x) = \int_{-\infty}^x v(y)\, dy$.

Remark 4.1. In the classical approach (1.3), we fail to obtain the stability due to the term $\|x\phi_N\|_{\omega_\beta}$ in (1.4). However, when α depends on time, an extra term is gained in the $\|xu_N\|_{\omega_\alpha}$ term:

$$\begin{aligned} &\frac{d}{dt}\|u_N(t)\|^2_{\omega_\alpha^{-1}} + 2\nu(|u_N|^2_{1,\omega_\alpha^{-1}} - \alpha^2\|u_N(t)\|^2_{\omega_\alpha^{-1}}) \\ &\qquad - 2\alpha(\alpha' + 2\nu\alpha^3)\|xu_N(t)\|^2_{\omega_\alpha^{-1}} = 2(f(t), u_N(t))_{\omega_\alpha^{-1}}. \end{aligned} \tag{4.8}$$

Stability can be obtained if α is chosen to satisfy $\alpha' + 2\nu\alpha^3 \le 0$.

We now briefly outline the convergence of the approximation (2.2). Our rigorous analysis will be carried out for the nonlinear convection-diffusion equations, which take (2.2) as a special case. It is interesting to note that the solutions of (1.3) and (2.2) are both of the same form: $u_N = \mathcal{P}_N^\alpha U$. In fact, assuming $U \in C(0,T; H^1_{\omega_\alpha}(\mathbb{R}))$, we have from (3.5) that for any $\varphi_N \in \mathbb{P}_N(\mathbb{R})$,

$$\tag{4.9} \begin{cases} (\partial_t \mathcal{P}_N^\alpha U(t), \varphi_N) + \nu(\partial_x \mathcal{P}_N^\alpha U(t), \partial_x \varphi_N) = (\partial_t U(t), \varphi_N) + \nu(\partial_x U(t), \partial_x \varphi_N) = (f(t), \varphi_N), \\ (\mathcal{P}_N^\alpha U(0), \varphi_N) = (U_0, \varphi_N). \end{cases}$$

However, the scheme (1.3) may not work since the bilinear form is not coercive. Since $u_N = \mathcal{P}_N^\alpha U$, it follows from (3.11) that if $U \in C(0,T; H^\sigma_{\omega_\alpha^{-1}}(\mathbb{R}))$ $(\sigma \geq 1)$, then

$$\tag{4.10} \|u_N(t) - U(t)\|_{r,\omega_\alpha^{-1}} \leq C N^{(r-\sigma)/2} \|U(t)\|_{\sigma,\omega_\alpha^{-1}} \quad \forall\, 0 \leq r \leq \sigma, \quad t \in (0,T),$$

which is analogous to the result obtained in [6] by using the similarity transformation.

The above method can be easily applied to some nonlinear equations. Consider the nonlinear convection-diffusion equation

$$\tag{4.11} \begin{cases} \partial_t U + \partial_x F(U) - \nu \partial_x^2 U = f(x,t), & (x,t) \in \mathbb{R} \times (0,T), \\ U(x,0) = U_0(x), & x \in \mathbb{R}, \end{cases}$$

where F is a smooth function, the constant $\nu > 0$, and U and $\partial_x U$ satisfy certain decay conditions at infinity. The semidiscrete Hermite function method for (4.11) is to find $u_N \in V_N$ such that for any $\varphi_N \in \mathbb{P}_N(\mathbb{R})$,

$$\tag{4.12} \begin{cases} (\partial_t u_N(t), \varphi_N) + (\partial_x F(u_N(t)), \varphi_N) + \nu(\partial_x u_N(t), \partial_x \varphi_N) = (f(t), \varphi_N), \quad t \in (0,T), \\ (u_N(0), \varphi_N) = (U_0, \varphi_N). \end{cases}$$

We investigate the stability property of the scheme (4.12). Suppose that u_N and the term on the right-hand side of (4.12) have the errors $\tilde{u}_N$ and $\tilde{f}$, respectively. Then, we have

$$\tag{4.13} (\partial_t \tilde{u}_N, \varphi_N) + (\partial_x \tilde{F}, \varphi_N) - \nu(\partial_x^2 \tilde{u}_N, \varphi_N) = (\tilde{f}, \varphi_N) \quad \forall \varphi_N \in \mathbb{P}_N(\mathbb{R}),\ t \in (0,T),$$

where $\tilde{F} := F(u_N + \tilde{u}_N) - F(u_N)$. Taking $\varphi_N = \omega_\alpha^{-1} \tilde{u}_N$ in (4.13), we obtain, similarly to (4.8),

$$\tag{4.14} \begin{aligned} &\frac{d}{dt}\|u_N(t)\|^2_{\omega_\alpha^{-1}} + \underline{\delta}\nu(|u_N(t)|^2_{1,\omega_\alpha^{-1}} + |\omega_\alpha^{-1} u_N(t)|^2_{1,\omega_\alpha}) \\ &\quad = 2(\tilde{f}(t) - \partial_x \tilde{F}(t), \tilde{u}_N(t))_{\omega_\alpha^{-1}} \\ &\quad \leq 2(\underline{\delta}\nu)^{-1}(\|\partial_x^{-1} \tilde{f}(t)\|^2_{\omega_\alpha^{-1}} + \|\tilde{F}\|^2_{\omega_\alpha^{-1}}) + \underline{\delta}\nu |\omega_\alpha^{-1} \tilde{u}_N(t)|^2_{1,\omega_\alpha}. \end{aligned}$$

Let $\tilde{M}$ be a positive constant and let

$$\tag{4.15} M(u) = \max_{0 \leq s \leq T} \|u_N(s)\|_{L^\infty(I)}, \qquad C_F = \max_{|z| \leq M(u) + \tilde{M}} |F'(z)|.$$

For any given $t \in (0,T)$, if

$$2(\alpha^2 N)^{1/4}\|\tilde{u}_N(s)\|_{\omega_\alpha^{-1}} \le \tilde{M} \qquad \forall\, s \in (0,t),$$

then by (3.15),

$$\begin{aligned}&\|\tilde{u}_N(s)\|_{L^\infty(I)} \le \tilde{M}\,,\\ &\|\tilde{F}(s)\|_{\omega_\alpha^{-1}} = \left\|\int_0^1 F'(u_N(s)+\theta\tilde{u}_N(s))\tilde{u}_N(s)\,d\theta\right\|_{\omega_\alpha^{-1}} \le C_F\|\tilde{u}_N(s)\|_{\omega_\alpha^{-1}} \quad \forall s \in (0,t).\end{aligned}$$

Substituting the above estimates into (4.14) gives

$$\frac{d}{dt}\|\tilde{u}_N(t)\|^2_{\omega_\alpha^{-1}} + \underline{\delta}\nu|\tilde{u}_N(t)|^2_{1,\omega_\alpha^{-1}} \le 2(\underline{\delta}\nu)^{-1}(C_F\|\tilde{u}_N(t)\|^2_{\omega_\alpha^{-1}} + \|\partial_x^{-1}\tilde{f}(t)\|^2_{\omega_\alpha^{-1}}). \tag{4.16}$$

Define

$$E(\tilde{u}_N,t) = \|\tilde{u}_N(t)\|^2_{\omega^{-1}_{\alpha(t)}} + \underline{\delta}\nu\int_0^t |\tilde{u}_N(s)|^2_{1,\omega^{-1}_{\alpha(s)}}\,ds, \tag{4.17}$$

$$\rho(\tilde{u}_N,\tilde{f},t) = \|\tilde{u}_N(0)\|^2_{\omega^{-1}_{\alpha(0)}} + 2(\underline{\delta}\nu)^{-1}\int_0^t \|\partial_x^{-1}\tilde{f}(s)\|^2_{\omega^{-1}_{\alpha(s)}}\,ds. \tag{4.18}$$

Integrating (4.16) with respect to t yields

$$E(\tilde{u}_N,t) \le \rho(\tilde{u}_N,\tilde{f},t) + C\int_0^t E(\tilde{u}_N,s)\,ds, \tag{4.19}$$

where C is a positive constant depending on $(\underline{\delta}\nu)^{-1}$ and C_F. Then, by a nonlinear Gronwall-like inequality [14],

$$E(\tilde{u}_N,t) \le \mathrm{e}^{Ct}\rho(\tilde{u}_N,\tilde{f},t) \qquad \forall\, 0 < t \le T, \tag{4.20}$$

provided that

$$4\alpha(t)N^{1/2}\mathrm{e}^{Ct}\rho(\tilde{u}_N,\tilde{f},t) \le \tilde{M}^2. \tag{4.21}$$

We now consider the convergence for the semidiscrete scheme (4.12). As we have shown for the linear problem (1.1), the projection $\mathcal{P}_N^\alpha U$ is a good comparison function. Let $u_* = \mathcal{P}_N^\alpha U$. Then, for any $\varphi_N \in \mathbb{P}_N(\mathbb{R})$,

(4.22)
$$\begin{cases}(\partial_t u_*(t),\varphi_N) + (\partial_x F(u_*(t)),\varphi_N) + \nu(\partial_x u_*(t),\partial_x\varphi_N) = (f(t),\varphi_N) - (\partial_x g(t),\varphi_N),\\ (u_*(0),\varphi_N) = (U_0,\varphi_N),\end{cases}$$

where $g(t) = F(U(t)) - F(u_*(t))$. Let $e_N = u_N - u_*$. We have

(4.23)
$$\begin{cases}(\partial_t e_N(t),\varphi_N) + (\partial_x G(t),\varphi_N) + \nu(\partial_x e_N(t),\partial_x\varphi_N) = (\partial_x g(t),\varphi_N), \quad t \in (0,T),\\ (e_N(0),\varphi_N) = 0,\end{cases}$$

where $G(t) = F(u_*(t) + e_N(t)) - F(u_*(t))$. Using the same argument as used in deriving the stability result (4.20), we can obtain

$$\begin{aligned}\|e_N(t)\|^2_{\omega^{-1}_{\alpha(t)}} &\le C\int_0^t \|g(s)\|^2_{\omega^{-1}_{\alpha(s)}}\,ds \le CC'_F\int_0^t \|(I-\mathcal{P}^\alpha_N)U(s)\|^2_{\omega^{-1}_{\alpha(s)}}\,ds\\ &\le CN^{-\sigma}\int_0^t \|\partial_x^\sigma U(s)\|^2_{\omega^{-1}_{\alpha(s)}}\,ds \le CN^{-\sigma}\|U\|^2_{L^2(0,T;H^\sigma_{\omega^{-1}_\alpha}(\mathbb{R}))}.\end{aligned}$$

THEOREM 4.1. *Let U and u_N be the solutions of* (4.11) *and* (4.12), *respectively. Assume that $U \in C(0,T;H^\sigma_{\omega^{-1}_\alpha}(\mathbb{R}))$ $(\sigma \ge 1)$, $F(z) \in C^1(\mathbb{R})$, the function $\alpha(t)$ is defined by* (2.1), *and $\underline{\delta}$ defined by* (4.5) *is positive. Then*

$$\|u_N(t) - U(t)\|_{\omega^{-1}_\alpha} \le CN^{-\sigma/2} \qquad \forall\, 0 < t < T,$$

where C is a constant depending on $(\underline{\delta}\nu)^{-1}, \delta_0, \delta, T$, and the regularity of $U$and F.

5. Stability and convergence: Fully discrete scheme. In this section, we further discretize the scheme (4.12) by using the method-of-lines approach. Without loss of generality, the analysis will be carried out for the nonlinear convection-diffusion equations. Noting that

$$(\partial_x F(u_N), H_m^{(\alpha)}) = -\sqrt{2m}\,\alpha(F(u_N), H_{m-1}^{(\alpha)}),$$

we can rewrite the scheme (4.12) in a matrix form as in (4.3):

$$\frac{d\mathbf{u}}{dt} - \alpha(t)\mathbf{D}^{1/2}\mathbf{F}(u_N) + \nu\alpha(t)^2\mathbf{B}\mathbf{u} = \mathbf{f}, \tag{5.1}$$

where $\mathbf{D}, \mathbf{B}, \mathbf{f}$ are the same as in (4.4) and (4.5), and the elements of the vector $\mathbf{F}$ are defined by

$$(\mathbf{F})_0 = 0\,, \quad (\mathbf{F})_m = (F(u_N), H_{m-1}^{(\alpha)}) \quad (1 \le m \le N).$$

For the time discretization, we use a second-order Crank–Nicolson/leapfrog scheme, which is implicit for the linear term and explicit for the nonlinear term [14, 15]. For the similarity transformation method (1.5), if the step size Δs for the transformed variable s is fixed, then the corresponding time-step in t is nonuniform. In our present approach, a uniform time-step is employed.

Let τ be the time-step size and let $t_k = k\tau$ $(k = 0, 1, \ldots, n_T;\ T = n_T\tau)$. We denote $v(x, t_k)$ by $v^k(x)$ or simply by v^k and $\mathbf{v}(t_k)$ by $\mathbf{v}^k$. Let

$$\mathbf{v}^k_{\hat{t}} = \frac{1}{2\tau}(\mathbf{v}^{k+1} - \mathbf{v}^{k-1}), \quad \mathbf{v}^{\hat{k}} = \frac{1}{2}(\mathbf{v}^{k+1} + \mathbf{v}^{k-1}).$$

For $v = \omega_{\alpha(t)}\sum_{l=0}^\infty \hat{v}_l(t)H_l^{(\alpha(t))}$, we define

$$D_t v = \omega_\alpha \sum_{l=0}^\infty \frac{d\hat{v}_l}{dt}H_l^{(\alpha)}.$$

The fully discrete Hermite spectral method to the nonlinear convection-diffusion equation (4.11) is to find

$$u_N^k = \omega_{\alpha(t)}\sum_{l=0}^N \hat{u}_l^k(t)H_l^{(\alpha(t))} \in V_N$$

satisfying

$$(5.2)\qquad \begin{cases} \mathbf{u}^k_{\hat t} - \alpha^k \mathbf{D}^{1/2}\mathbf{F}(u^k_N) + \nu(\alpha^k)^2 \mathbf{B}\mathbf{u}^{\hat k} = \mathbf{f}^k, & 1 \le k \le n_T - 1,\\ (\mathbf{u}^1)_m = (\mathbf{u}^0)_m + \tau(D_t U(0), H_m^{(\alpha(0))}), & 0 \le m \le N,\\ (\mathbf{u}^0)_m = (U_0, H_m^{(\alpha(0))}), & 0 \le m \le N, \end{cases}$$

where $(D_t U(0), H_m^{(\alpha(0))})$ can be computed from $\frac{d\mathbf{u}}{dt}(0)$ using the initial condition and (5.1).

We now present a stability analysis for the scheme (5.2). Assume that the solution and the term on the right-hand side of (5.2) have errors $\tilde{\mathbf{u}}^k := (\tilde{\hat{u}}^k_0, \tilde{\hat{u}}^k_1, \ldots, \tilde{\hat{u}}^k_N)^T$ and $\tilde{\mathbf{f}}^k$, respectively, with $\tilde u^k_N = \omega_\alpha \sum_{l=0}^N \tilde{\hat{u}}^k_l H_l^{(\alpha)}$. Then the errors satisfy

$$(5.3)\qquad \tilde{\mathbf{u}}^k_{\hat t} - \alpha^k \mathbf{D}^{1/2}\tilde{\mathbf{F}}^k + \nu(\alpha^k)^2 \mathbf{B}\tilde{\mathbf{u}}^{\hat k} = \tilde{\mathbf{f}}^k, \qquad 1 \le k \le n_T - 1,$$

where $\tilde{\mathbf{F}}^k = \mathbf{F}(u^k_N + \tilde u^k_N) - \mathbf{F}(u^k_N)$. Multiplying both sides of (5.3) with $2\tilde{\mathbf{u}}^{\hat k}$ and assuming that $\underline{\delta} = \min\{1, 2\delta_0\delta - 1\} > 0$, we obtain

$$(5.4)\qquad \begin{aligned} &(\|\tilde{\mathbf{u}}^k\|^2)_{\hat t} + 2\underline{\delta}\nu(\alpha^k)^2 \|(\mathbf{D}+\mathbf{I})^{1/2}\tilde{\mathbf{u}}^{\hat k}\|^2 \le 2(\tilde{\mathbf{f}}^{\hat k} + \alpha^k \mathbf{D}^{1/2}\tilde{\mathbf{F}}^k, \tilde{\mathbf{u}}^{\hat k})\\ &\quad \le 2(\underline{\delta}\nu)^{-1}((\alpha^k)^{-2}\|(\mathbf{D}+\mathbf{I})^{-1/2}\tilde{\mathbf{f}}^k\|^2 + \|\tilde{\mathbf{F}}^k\|^2) + \underline{\delta}\nu(\alpha^k)^2\|(\mathbf{D}+\mathbf{I})^{1/2}\tilde{\mathbf{u}}^{\hat k}\|^2. \end{aligned}$$

Let $\tilde M$ be a positive constant and let

$$(5.5)\qquad M(u) = \max_{0 \le k \le n_T} \|u^k_N\|_{L^\infty(I)}, \qquad C_F = \max_{|z| \le M(u) + \tilde M} |F'(z)|.$$

For a fixed $n \le n_T$, if

$$\|\tilde{\mathbf{u}}^k\| = \|\tilde u^k_N\|_{\omega_\alpha^{-1}} \le (4\alpha^k N^{1/2})^{-1/2}\tilde M \qquad \forall\, 1 \le k \le n-1,$$

then, by (3.15), we have $\|\tilde u^k_N\|_{L^\infty(I)} \le \tilde M$ and

$$\begin{aligned} \|\tilde{\mathbf{F}}^k\| &= \|\mathcal{P}^\alpha_{N-1}(F(u^k_N + \tilde u^k_N) - F(u^k_N))\|_{\omega_\alpha^{-1}}\\ &\le \|F(u^k_N + \tilde u^k_N) - F(u^k_N)\|_{\omega_\alpha^{-1}} \le C_F\|\tilde u^k_N\|^2_{\omega_\alpha^{-1}} = C_F\|\tilde{\mathbf{u}}^k\|. \end{aligned}$$

Define

$$(5.6)\qquad E^n(\mathbf{v}) = \|\mathbf{v}^n\|^2 + 2\underline{\delta}\nu\tau\sum_{k=1}^{n-1}(\alpha^k)^2\|(\mathbf{D}+\mathbf{I})^{1/2}\mathbf{v}^{\hat k}\|^2,$$

$$(5.7)\qquad \rho^n(\mathbf{v},\mathbf{g}) = \|\mathbf{v}^0\|^2 + \|\mathbf{v}^1\|^2 + 4(\underline{\delta}\nu)^{-1}\tau\sum_{k=0}^{n-1}(\alpha^k)^{-2}\|(\mathbf{D}+\mathbf{I})^{-1/2}\mathbf{g}^k\|^2.$$

Summing (5.4) for $1 \le k \le n-1$ gives

$$E^n(\tilde{\mathbf{u}}) \le \rho^n(\tilde{\mathbf{u}},\tilde{\mathbf{f}}) + 2(\underline{\delta}\nu)^{-1}C_F\tau\sum_{k=1}^{n-1}E^k(\tilde{\mathbf{u}}).$$

It follows from a discrete nonlinear Gronwall-like inequality [14] that

$$E^n(\tilde{\mathbf{u}}) \le \mathrm{e}^{Cn\tau}\rho^n(\tilde{\mathbf{u}},\tilde{\mathbf{f}}) \qquad \forall\, 0 < n \le n_T,$$

provided that $4\max_{0\le k\le n}\alpha^k N^{1/2}\mathrm{e}^{Ck\tau}\rho^k(\tilde{\mathbf{u}},\tilde{\mathbf{f}}) \le \tilde M^2$.

THEOREM 5.1. *Let u_N be the solution of* (4.12) *and let $\tilde M$ be a positive number. Assume that the function $\alpha(t)$ is defined by* (2.1) *and that $\underline{\delta}$ defined by* (4.5) *is positive. For $0 < n \le n_T$, if*

$$4\max_{0\le k\le n}\alpha^k N^{1/2}\mathrm{e}^{Ck\tau}\rho^k(\tilde{\mathbf{u}},\tilde{\mathbf{f}})\le \tilde{M}^2\,, \tag{5.8}$$

then

$$E^k(\tilde{\mathbf{u}})\le \mathrm{e}^{Ck\tau}\rho^k(\tilde{\mathbf{u}},\tilde{\mathbf{f}})\qquad \forall\, 0<k\le n,$$

where E^k and ρ^k are defined by (5.6) *and* (5.7)*, respectively, and C is a constant linearly proportional to $(\underline{\delta}\nu)^{-1}$ and C_F.*

We now analyze the convergence of the fully discrete scheme (5.2). Let $u_* = \mathcal{P}^{\alpha}_{N}U$ with the following Hermite expansion:

$$u_* = \omega_\alpha\sum_{l=0}^{N}\hat{u}_{*l}H^{(\alpha)}_l\,.$$

Denote the coefficients of the above expansion by $\mathbf{u}_* := (\hat{u}_{*0},\hat{u}_{*1},\dots,\hat{u}_{*N})^T$. It follows from (4.22) that

$$\mathbf{u}^k_{*\hat{t}} - \alpha^k\mathbf{D}^{1/2}\mathbf{F}(u^k_*) + \nu(\alpha^k)^2\mathbf{B}\mathbf{u}^{\hat{k}}_* = \mathbf{f}^k - \mathbf{g}^k, \tag{5.9}$$

where we split $\mathbf{g}^k$ into $\mathbf{g}^k_1$, $\mathbf{g}^k_2$, $\mathbf{g}^k_3$ as follows:

$$\begin{aligned}\mathbf{g}^k &= \left[\left(\frac{d\mathbf{u}_*}{dt}\right)^k - \mathbf{u}^k_{*\hat{t}}\right] + [\alpha^k\mathbf{D}^{1/2}(\mathbf{F}(u^k_*) - \mathbf{F}(U^k))] + [\nu(\alpha^k)^2\mathbf{B}(\mathbf{u}^k_* - \mathbf{u}^{\hat{k}}_*)]\\ &=: \mathbf{g}^k_1 + \mathbf{g}^k_2 + \mathbf{g}^k_3.\end{aligned} \tag{5.10}$$

Let $e^k_{N} = u^k_{N} - u^k_*$ and $\mathbf{e}^k = \mathbf{u}^k - \mathbf{u}^k_*$. Then it can be verified that

$$\begin{cases}\mathbf{e}^k_{\hat{t}} - \alpha^k\mathbf{D}^{1/2}\mathbf{G}^k + \nu(\alpha^k)^2\mathbf{B}\mathbf{e}^{\hat{k}} = \mathbf{g}^k, & 1\le k\le n_{T}-1,\\ \mathbf{e}^0 = 0, \qquad \mathbf{e}^1 = \mathbf{u}_*(0) + \tau\dfrac{d\mathbf{u}_*}{dt}(0) - \mathbf{u}_*(\tau), & \end{cases} \tag{5.11}$$

where $\mathbf{G}^k = \mathbf{F}(u^k_* + e^k_{N}) - \mathbf{F}(u^k_*)$. By the same arguments as in the stability analysis above, we can obtain

$$\|e^n_{N}\|^2_{\omega_\alpha^{-1}} = \|\mathbf{e}^n\|^2 \le C\left(\|\mathbf{e}^0\|^2 + \|\mathbf{e}^1\|^2 + (\underline{\delta}\nu)^{-1}\tau\sum_{k=0}^{n-1}(\alpha^k)^{-2}\|(\mathbf{D}+\mathbf{I})^{-1/2}\mathbf{g}^k\|^2\right).$$

The last term on the right-hand side can be bounded by using the facts below:

$$\tau\sum_{k=0}^{n-1}\|(\mathbf{D}+\mathbf{I})^{-1/2}\mathbf{g}^k_1\|^2 \le C\tau^4\|D^3_tU\|^2_{L^2(0,T;H^{-1}_{\omega_\alpha^{-1}}(\mathbb{R}))}, \tag{5.12}$$

$$\begin{aligned}\tau\sum_{k=0}^{n-1}\|(\mathbf{D}+\mathbf{I})^{-1/2}\mathbf{g}^k_2\|^2 &\le C\tau\sum_{k=0}^{n-1}\|F(u^k_*) - F(U^k)\|^2_{\omega_\alpha^{-1}}\\ &\le CC'_FN^{-\sigma}\|U\|^2_{C(0,T;H^{\sigma}_{\omega_\alpha^{-1}}(\mathbb{R}))},\end{aligned} \tag{5.13}$$

$$\tau\sum_{k=0}^{n-1}\|(\mathbf{D}+\mathbf{I})^{-1/2}\mathbf{g}^k_3\|^2 \le C\tau^4\|D^2_tU\|^2_{L^2(0,T;H^{1}_{\omega_\alpha^{-1}}(\mathbb{R}))}. \tag{5.14}$$

The initial errors can be bounded by using the Taylor expansion:

$$\begin{aligned} \|\mathbf{e}^1\| &= \left\| \int_0^\tau (\tau - s) \frac{d^2 \mathbf{u}_*}{dt^2}(s)\, ds \right\| \le \tau^2 \max_{0\le s\le\tau} \left\| \frac{d^2\mathbf{u}_*}{dt^2}(s) \right\| \\ &\le \tau^2 \|D_t^2 U(s)\|_{C(0,\tau;L^2_{\omega_\alpha^{-1}}(\mathbb{R}))}. \end{aligned} \tag{5.15}$$

Combining the above results, we arrive at the following optimal error estimate.

THEOREM 5.2. *Let U and u_{N} be the solutions of* (4.11) *and* (4.12)*, respectively. Assume that $U \in C(0,T;H^\sigma_{\omega_\alpha^{-1}}(\mathbb{R}))$ $(\sigma \ge 1)$, $D_t^2 U \in L^2(0,T;H^1_{\omega_\alpha^{-1}}(\mathbb{R})) \cap C(0,\tau;L^2_{\omega_\alpha^{-1}}(\mathbb{R}))$, $D_t^3 U \in L^2(0,T;H^{-1}_{\omega_\alpha^{-1}}(\mathbb{R}))$, and $F(z) \in C^1(\mathbb{R})$. Moreover, assume that the function $\alpha(t)$ is defined by* (2.1)*, $\underline{\delta}$ defined by* (4.5) *is positive, and $\tau N^{1/8} \le c_0$ is sufficiently small. Then, for $0 \le n \le n_{T}$,*

$$\|u_{N}^n - U^n\|_{\omega^{-1}_{\alpha(t_n)}} \le C(\tau^2 + N^{-\sigma/2}),$$

where C is a constant depending on $(\underline{\delta}\nu)^{-1}, \delta_0, \delta, T$, and the regularity of U and F.

Remark 5.1. If the underlying PDE solution does not satisfy the exponential decay property required by the Hermite function approximation, one may use the Hermite polynomial approximation directly. In this case, the Hermite polynomial approximation should be used together with a time-dependent scaling,

$$\alpha(t) = \frac{1}{2\sqrt{\nu\delta_0(\delta(T-t)+1)}}. \tag{5.16}$$

For the linear parabolic equation (1.1) and the nonlinear convection-diffusion equation (4.11), it can be verified that with the choice (5.16), the desired stability and convergence results can be established in some appropriate function space.

6. Numerical results. In this section, we present some numerical examples using the proposed method for both linear and nonlinear equations. The numerical results will be compared with those obtained by using the classical method (1.3) and by using the similarity transformation technique. In the following computations, the integrals involved are computed by the Hermite–Gauss quadrature rules with $N+1$ quadrature points. Let

$$E_{N}(t) = \|u_{N}(t) - U^{N}(t)\|_{\omega_\alpha^{-1}}, \qquad E_{N,\infty}(t) = \frac{\max_{0\le j\le N} |u_{N}(y_j,t) - U(y_j,t)|}{\max_{0\le j\le N} |U(y_j,t)|},$$

where $U^{N} \in V_{N}$ is the interpolation of U at the Hermite–Gauss points $\{y_j\}_{j=0}^N$. The examples used below are taken from [6] and [10], where the diffusion coefficient ν in (1.1) is chosen as 1. In the linear case our approach is appropriate for the general choice of $\nu > 0$ due to the use of the scaling factor (2.1). However, for nonlinear problems (such as Example 6.3 below) with sufficiently small values of ν, steep layers may be developed, and in this case some special techniques such as the spectral viscosity method [16] should be applied.

Example 6.1 (linear problem). Consider the parabolic problem (1.1) with $\nu = 1$ and the following source term:

$$f(x,t) = (x\cos x + (t+1)\sin x)(t+1)^{-3/2} e^{-x^2/4(t+1)}. \tag{6.1}$$

TABLE 6.1
Example 6.1*: Errors at $t = 1$ with $N = 20$ using different methods.*

Time step τ	Funaro and Kavian's scheme [6]	Classical scheme (1.3)	Proposed scheme (2.7)
250^{-1}	2.487E-03	1.948E-04	2.958E-06
1000^{-1}	6.203E-04	1.947E-04	1.189E-06
4000^{-1}	1.550E-04	1.947E-04	1.177E-06
16000^{-1}	3.886E-05	1.947E-04	1.177E-06

TABLE 6.2
Example 6.1*: Errors of the proposed scheme* (2.7) *with different τ and N.*

τ	N	$E_N(1)$	$E_{N,\infty}(1)$	Order
1E-1	30	1.697E-03	9.775E-04	
1E-2		1.697E-05	9.769E-06	$\tau^{2.00}$
1E-3		1.696E-07	9.769E-08	$\tau^{2.00}$
1E-4		1.696E-09	9.798E-10	$\tau^{2.00}$
1E-4	10	5.161E-03	1.192E-03	
	20	1.177E-06	1.246E-07	$N^{-12.10}$
	30	1.696E-09	9.798E-10	$N^{-16.14}$

This example was used by Funaro and Kavian [6]. Its exact solution is of the form

$$U(x,t) = \frac{\sin x}{\sqrt{t+1}} e^{-x^2/4(t+1)}. \tag{6.2}$$

We solve the above problem with $(\delta_0, \delta) = (1.5, 0)$, which corresponds to the classical approach (1.3), and with $(\delta_0, \delta) = (1, 1)$, which corresponds to the method proposed in this work. For ease of comparison, we use the same mesh size as used in [6]. Table 6.1 shows the error $E_{20}(t)$ at $t = 1$ with different time-steps. Note that the result in [6] is obtained by using (explicit) first-order forward difference in time.

Table 6.2 shows the order of accuracy for the scheme (2.7) with $\delta_0 = \delta = 1$. The numerical results are in good agreement with the theoretical prediction that the numerical scheme (2.7) is of second-order accuracy in time and spectral accuracy in space.

Example 6.2 (linear problem). Consider the parabolic problem (1.1) with $\nu = 1$ and the following source term:

$$f(x,t) = \left(k(1+4c^2x)\cos k(x+t) - (k^2 + 2c^2(1-2(cx)^2))\sin k(x+t)\right) \mathrm{e}^{-(cx)^2}, \tag{6.3}$$

where c is a constant. The exact solution of this example has a time-independent decay:

$$U(x,t) = \sin k(x+t)\mathrm{e}^{-c^2x^2}. \tag{6.4}$$

Table 6.3
Example 6.2: *Comparison of the classical approach and the present method.*

		Classical method (1.3)		Proposed method (2.7)	
τ	Steps	$E_{160}(1)$	$E_{160,\infty}(1)$	$E_{160}(1)$	$E_{160,\infty}(1)$
1E-3	250	5.66E-07	3.93E-07	4.30E-06	1.87E-06
	500	1.52E-04	8.50E-06	2.73E-06	2.03E-06
	750	3.72E+01	2.67E+00	2.08E-06	1.44E-06
	1000	8.20E+06	2.02E+05	1.75E-06	1.34E-06
1E-4	2500	5.66E-09	3.94E-09	4.30E-08	1.87E-08
	5000	1.45E-04	8.01E-06	2.73E-08	2.03E-08
	7500	4.30E+01	2.41E+00	2.08E-08	1.44E-08
	10000	8.95E+06	4.97E+04	1.73E-08	1.36E-08

The purpose for choosing this example is to demonstrate that the Hermite spectral method with a time-dependent scaling also works well for the solutions with time-independent decays. In our computations, the parameters k and c are taken as 5 and 0.5, respectively. We solve this problem by using a constant weight $\alpha(t) \equiv 0.5$, which not only corresponds to the classical method (1.3) but also matches the exponential solution-decay exactly. We also solve the problem by using the scheme (2.7) with $(\delta_0, \delta) = (0.6, 1)$. This choice of the parameters satisfies $\underline{\delta} = 0.2 > 0$, and therefore stability and convergence are expected. It is seen from Table 6.3 that although the classical method (1.3) matches the exponential decay exactly, the error is accumulated due to numerical instability. On the other hand, the Hermite spectral method with a time-dependent scaling produces highly accurate and stable numerical approximations.

Example 6.3 (nonlinear viscous Burgers equation). Consider the viscous Burgers equation

$$\partial_t U + U\partial_x U - \nu\partial_x^2 U = f(x,t), \quad x \in \mathbb{R}, \quad t > 0. \tag{6.5}$$

It was computed in [10] via the transformation

$$y = \frac{x}{2\sqrt{\nu(t+1)}}, \qquad s = \ln(t+1) \tag{6.6}$$

for a soliton-like solution

$$U(x,t) = \mathrm{e}^{-y^2}\operatorname{sech}^2(ay - bs - c). \tag{6.7}$$

We will recompute this problem with parameters $a = 0.3$, $b = 0.5$, $c = -3$, and $\nu = 1$.

We use the fully discrete scheme (5.2) to solve the problem with $(\delta_0, \delta) = (1, 1)$. The numerical errors at $t = \mathrm{e} - 1$ are presented in Table 6.4, where the comparison is made with those given in [10]. It is seen that the present method is more accurate than the similarity transformation solution.

To show the rate of convergence for (5.2), we list in Table 6.5 the numerical errors at $t = 1$ with various τ and N. The fully discrete scheme (5.2) is applied to the viscous Burgers problem with $(\delta_0, \delta) = (1, 1)$. It again confirms the theoretical prediction that the present method is of second-order accuracy in time and spectral accuracy in space.

Table 6.4
Example 6.3*: Errors at* $t = \mathrm{e} - 1$ *with* $\tau = 0.001 * t$.

N	Guo and Xu's result [10]	Proposed scheme (5.2)
8	1.381E-06	1.563E-05
16	1.381E-06	6.337E-07
32	1.381E-06	1.031E-07

Table 6.5
Example 6.3*: Errors of the proposed scheme* (5.2) *with different* τ *and* N.

τ	N	$E_N(1)$	$E_{N,\infty}(1)$	Order
1E-1		5.101E-04	4.677E-03	
1E-2		4.508E-06	4.548E-05	$\tau^{2.05}$
1E-3	40	4.454E-08	4.530E-07	$\tau^{2.01}$
1E-4		4.467E-10	4.372E-09	$\tau^{2.00}$
	8	6.685E-06	1.163E-04	
1E-4	16	2.684E-07	3.121E-06	$N^{-4.64}$
	32	7.888E-10	7.120E-09	$N^{-8.41}$

Acknowledgment. The third author thanks Prof. David Gottlieb for helpful discussions on numerical approximations in unbounded domains.

REFERENCES

[1] N. Adžić, *Modified Hermite polynomials in the spectral approximation for boundary layer problems*, Bull. Austral. Math. Soc., 45 (1992), pp. 267–276.

[2] C. Bernardi and Y. Maday, *Spectral methods*, in Handbook of Numerical Analysis, Vol. V, Handb. Numer. Anal. V, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1997, pp. 209–485.

[3] J. P. Boyd, *Spectral method using rational basis functions on an infinite interval*, J. Comput. Phys., 69 (1987), pp. 112–142.

[4] O. Coulaud, D. Funaro, and O. Kavian, *Laguerre spectral approximation of elliptic problems in exterior domains*, in Spectral and High Order Methods for Partial Differential Equations (Como, 1989), Comput. Methods Appl. Mech. Engrg., 80 (1990), pp. 451–458.

[5] J. C. M. Fok, B. Guo, and T. Tang, *Combined Hermite spectral-finite difference method for the Fokker-Planck equation*, Math. Comp., 71 (2002), pp. 1497–1528.

[6] D. Funaro and O. Kavian, *Approximation of some diffusion evolution equations in unbounded domains by Hermite functions*, Math. Comp., 57 (1991), pp. 597–619.

[7] B.-Y. Guo, *Error estimation of Hermite spectral method for nonlinear partial differential equations*, Math. Comp., 68 (1999), pp. 1067–1078.

[8] B.-Y. Guo and J. Shen, *Laguerre-Galerkin method for nonlinear partial differential equations on a semi-infinite interval*, Numer. Math., 86 (2000), pp. 635–654.

[9] B.-Y. Guo, J. Shen, and Z.-Q. Wang, *A rational approximation and its applications to differential equations on the half line*, J. Sci. Comput., 15 (2000), pp. 117–147.

[10] B.-Y. Guo and C.-L. Xu, *Hermite pseudospectral method for nonlinear partial differential equations*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 859–872.

[11] V. Iranzo and A. Falqués, *Some spectral approximations for differential equations in unbounded domains*, Comput. Methods Appl. Mech. Engrg., 98 (1992), pp. 105–126.

[12] I. K. Khabibrakhmanov and D. Summers, *The use of generalized Laguerre polynomials in spectral methods for nonlinear differential equations*, Comput. Math. Appl., 36 (1998), pp. 65–70.

[13] Y. Liu, L. Liu, and T. Tang, *The numerical computation of connecting orbits in dynamical systems: A rational spectral approach*, J. Comput. Phys., 111 (1994), pp. 373–380.

[14] H. Ma and W. Sun, *Optimal error estimates of the Legendre–Petrov–Galerkin method for the Korteweg–de Vries equation*, SIAM J. Numer. Anal., 39 (2001), pp. 1380–1394.

[15] H. Ma and W. Sun, *A Legendre–Petrov–Galerkin and Chebyshev collocation method for third-order differential equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1425–1438.

[16] Y. Maday and E. Tadmor, *Analysis of the spectral vanishing viscosity method for periodic conservation laws*, SIAM J. Numer. Anal., 26 (1989), pp. 854–870.

[17] Y. Maday, B. Pernaud-Thomas, and H. Vandeven, *Reappraisal of Laguerre type spectral methods*, Rech. Aérospat., 6 (1985), pp. 353–375.

[18] G. Mastroianni and G. Monegato, *Nyström interpolants based on zeros of Laguerre polynomials for some Weiner-Hopf equations*, IMA J. Numer. Anal., 17 (1997), pp. 621–642.

[19] B. Muckenhoupt, *Mean convergence of Hermite and Laguerre series*, II, Trans. Amer. Math. Soc., 147 (1970), pp. 433–460.

[20] J. W. Schumer and J. P. Holloway, *Vlasov simulations using velocity-scaled Hermite representations*, J. Comput. Phys., 144 (1998), pp. 626–661.

[21] J. Shen, *Stable and efficient spectral methods in unbounded domains using Laguerre functions*, SIAM J. Numer. Anal., 38 (2000), pp. 1113–1133.

[22] T. Tang, *The Hermite spectral method for Gaussian-type functions*, SIAM J. Sci. Comput., 14 (1993), pp. 594–606.

[23] S. Thangavelu, *Hermite expansions on* $\mathbf{R}^{2n}$ *for radial functions*, Rev. Mat. Iberoamericana, 6 (1990), pp. 61–73.

[24] S. Thangavelu, *Lectures on Hermite and Laguerre Expansions*, Princeton University Press, Princeton, NJ, 1993.

[25] J. A. C. Weideman, *The eigenvalues of Hermite and rational spectral differentiation matrices*, Numer. Math., 61 (1992), pp. 409–432.

SIAM J. Numer. Anal.
Vol. 43, No. 1, pp. 76–95

CONSTRUCTION ALGORITHMS FOR DIGITAL NETS WITH LOW WEIGHTED STAR DISCREPANCY[*]

JOSEF DICK[†], GUNTHER LEOBACHER[‡], AND FRIEDRICH PILLICHSHAMMER[‡]

Abstract. We introduce a new construction method for digital nets which yield point sets in the s-dimensional unit cube with low star discrepancy. The digital nets are constructed using polynomials over finite fields. It has long been known that there exist polynomials which yield point sets with low (unweighted) star discrepancy. This result was obtained by Niederreiter by the means of averaging over all polynomials. Hence concrete examples of good polynomials were not known in many cases. Here we show that good polynomials can be found by computer search. The search algorithm introduced in this paper is based on minimizing a quantity closely related to the star discrepancy.

It has been pointed out that many integration problems can be modeled by weighted function spaces and it has been shown that in this case point sets with low weighted discrepancy are required. Hence it is particularly useful to be able to adjust a point set to some given weights. We are able to extend our results from the unweighted case to show that this can be done using our construction algorithms. This way we can find point sets with low weighted star discrepancy, making such point sets especially useful for many applications.

Key words. digital net, weighted star discrepancy, component-by-component algorithm

AMS subject classifications. 11K38, 65D30, 65D32

DOI. 10.1137/040604662

1. Introduction.

In many applications one wants to approximate an s-dimensional integral over the unit cube,

$$I_s(F) := \int_{[0,1)^s} F(\boldsymbol{x}) d\boldsymbol{x},$$

by an N point quasi-Monte Carlo (QMC) rule,

$$Q_{N,s}(F) := \frac{1}{N} \sum_{n=0}^{N-1} F(\boldsymbol{x}_n).$$

For QMC rules the points $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1}$ are chosen deterministically, with the aim to obtain a small integration error. It has been shown that uniformly distributed point sets yield a small integration error for functions from certain function classes. Several quality measures of point sets in the unit cube are known. One popular way of measuring the distribution quality is based on the discrepancy function Δ. For a point set $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1}$ in the s-dimensional unit cube $[0,1)^s$ the discrepancy function Δ is defined as

$$\Delta(\alpha_1, \dots, \alpha_s) := \frac{A_N\left(\prod_{i=1}^s [0, \alpha_i)\right)}{N} - \alpha_1 \dots \alpha_s,$$

[*] Received by the editors March 2, 2004; accepted for publication (in revised form) September 20, 2004; published electronically April 26, 2005.

http://www.siam.org/journals/sinum/43-1/60466.html

[†] School of Mathematics, University of New South Wales, Sydney 2052, Australia (josi@maths.unsw.edu.au). The research of this author was supported by the Australian Research Council under its Center of Excellence Program.

[‡] Institut für Finanzmathematik, Universität Linz, Altenbergstraße 69, A-4040 Linz, Austria (gunther. leobacher@jku.at, friedrich.pillichshammer@jku.at). The research of the second author was supported by Erwin Schrödinger grant J 2215-N04. The research of the third author was supported by the Austrian Research Foundation, project S 8305 and project P17022-N12.

where $0 \le \alpha_1, \ldots, \alpha_s \le 1$. Here $A_N(E)$ denotes the number of indices n, $0 \le n \le N-1$, such that $\boldsymbol{x}_n$ is contained in the set E. By taking a norm of the discrepancy function we obtain a measure for the irregularity of distribution of the point set. We have the following definition (see, for example, Drmota and Tichy [4] or Kuipers and Niederreiter [10]).

Definition 1.1. *For a point set $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}$ in $[0,1)^s$ the star discrepancy D_N^* is defined as the supremums norm of the discrepancy function, i.e.,*

$$D_N^* = D_N^*(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}) := \sup_{\substack{0 \le \alpha_i \le 1 \\ 1 \le i \le s}} |\Delta(\alpha_1, \ldots, \alpha_s)|.$$

Informally we will say a point set has "low" or "small" star discrepancy if the star discrepancy is of order $O((\log N)^s/N)$; see [4, Chapter 3].

The star discrepancy of a finite point set is intimately related to the worst-case error of multivariate integration of functions with bounded variation in the sense of Hardy and Krause. Here the basic error estimate for the integration error is given by the Koksma–Hlawka inequality (see, for example, [10, Theorem 5.5] or [13, Theorem 2.11]), which states that

$$|I_s(F) - Q_{N,s}(F)| \le V(F) D_N^*,$$

where $V(F)$ is the variation of F in the sense of Hardy and Krause and D_N^* is the star discrepancy of the point set $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}$.

Having observed that different coordinates may have different influence on the quality of the approximation, Sloan and Woźniakowski [20] introduced a generalized ("weighted") version of the Koksma–Hlawka inequality. The star discrepancy for this case is then called weighted star discrepancy. We will give the definition subsequently, but first we introduce some notation used throughout the paper. Let $\boldsymbol{\gamma} = (\gamma_i)_{i \ge 1}$ denote a sequence of positive real numbers, the "weights," and let $E = \{1, 2, \ldots, s\}$ denote the set of coordinate indices. For $u \subseteq E$ let $\gamma_u = \prod_{i \in u} \gamma_i$, $\gamma_\emptyset = 1$, $|u|$ be the cardinality of u, and for a vector $\boldsymbol{z} \in [0,1)^s$ let $\boldsymbol{z}_u$ denote the vector $[0,1)^{|u|}$ containing only the components of $\boldsymbol{z}$ whose indices are in u. Moreover we write $(\boldsymbol{z}_u, \mathbf{1})$ for the vector that we obtain by replacing all the components of $\boldsymbol{z}$ not in u by 1. We have the following definition (see also [20]).

Definition 1.2. *For a point set $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}$ in $[0,1)^s$ and a sequence $\boldsymbol{\gamma} = (\gamma_i)_{i \ge 1}$ of weights the weighted star discrepancy $D_{N,\boldsymbol{\gamma}}^*$ is given by*

$$D_{N,\boldsymbol{\gamma}}^* = D_{N,\boldsymbol{\gamma}}^*(\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1}) := \sup_{\boldsymbol{z} \in [0,1)^s} \max_{\substack{u \subseteq E \\ u \neq \emptyset}} \gamma_u |\Delta(\boldsymbol{z}_u, \mathbf{1})|,$$

where $\gamma_u = \prod_{i \in u} \gamma_i$.

Note that for the choice $\boldsymbol{\gamma} = \mathbf{1}$, that is, $\gamma_i = 1$ for all $i \ge 1$, we have $D_{N,\mathbf{1}}^* = D_N^*$ from Definition 1.1. Henceforth we will refer to the unweighted star discrepancy as classical star discrepancy or simply star discrepancy.

Sloan and Woźniakowski [20] continued by showing that for all functions in the Sobolev space $W_2^{(1,\ldots,1)}([0,1)^s)$ we have

$$|I_s(F) - Q_{N,s}(F)| \le D_{N,\boldsymbol{\gamma}}^* \|F\|_{s,\boldsymbol{\gamma}},$$

where the norm is defined as

$$\|F\|_{s,\boldsymbol{\gamma}} := \sum_{u \subseteq E} \gamma_u^{-1} \int_{[0,1)^{|u|}} \left| \frac{\partial^{|u|}}{\partial \boldsymbol{x}_u} F(\boldsymbol{x}_u, \mathbf{1}) \right| d\boldsymbol{x}_u.$$

(Sloan and Woźniakowski [20] concentrated mainly on the Hilbertian case. See also [11] for a more specialized treatment of the L^q case for $q \neq 2$.) Therefore point sets with low weighted star discrepancy guarantee a small worst-case error for numerical integration and hence the need for point sets with low weighted star discrepancy.

Note that the best upper bounds on the classical star discrepancy are of the form $C_s(\log N)^s/N$ for some constant C_s independent of N. Because of the factor $(\log N)^s$ such bounds are useful only if N is exponentially large in the dimension s. Using a weighted star discrepancy that focuses mainly on lower dimensional projections, the effect of $(\log N)^s$ can be much reduced, yielding useful upper bounds on the star discrepancy even for large dimensions s.

Currently the most effective constructions of point sets with low star discrepancy are based on the concept of (t, m, s)-nets in a base b. For a definition of such nets see [13, Definition 4.1]. In [14] (see also [13, Chapter 4.4]) Niederreiter introduced a special construction of such nets. This construction is based on rational functions over finite fields.

Henceforth let p be a prime. Further let $\mathbb{F}_p((x^{-1}))$ be the field of formal Laurent series over the finite field $\mathbb{F}_p$ consisting of p elements. Thus elements of $\mathbb{F}_p((x^{-1}))$ are of the form

$$L = \sum_{l=w}^{\infty} t_l x^{-l},$$

where w is an arbitrary integer and all $t_l \in \mathbb{F}_p$. Note that $\mathbb{F}_p((x^{-1}))$ contains the field of rational functions over $\mathbb{F}_p$ as a subfield. Further let $\mathbb{F}_p[x]$ be the set of all polynomials over $\mathbb{F}_p$ and let $m \geq 1$ be an integer. For a given dimension $s \geq 2$, choose $f \in \mathbb{F}_p[x]$, with $\deg(f) = m$, and let $g_1, \dots, g_s \in \mathbb{F}_p[x]$. Let φ_m be the map from $\mathbb{F}_p((x^{-1}))$ to the interval $[0, 1)$ defined by

$$\varphi_m\left(\sum_{l=w}^{\infty} t_l x^{-l}\right) = \sum_{l=\max(1,w)}^{m} t_l p^{-l}.$$

For $0 \leq n < p^m$ let $n = n_0 + n_1 p + \cdots + n_{m-1}p^{m-1}$ be the p-adic expansion of n. With each such n we associate the polynomial

$$n(x) = \sum_{r=0}^{m-1} n_r x^r \in \mathbb{F}_p[x].$$

Then $P(\boldsymbol{g}, f)$ is the point set consisting of the p^m points

$$\boldsymbol{x}_n = \left(\varphi_m\left(\frac{n(x)g_1(x)}{f(x)}\right), \dots, \varphi_m\left(\frac{n(x)g_s(x)}{f(x)}\right)\right) \in [0, 1)^s$$

for $0 \leq n \leq p^m - 1$. Due to the construction principle, a QMC rule using the point set $P(\boldsymbol{g}, f)$ is often called a *polynomial lattice rule*. The vector $\boldsymbol{g}$ is called the generating vector of $P(\boldsymbol{g}, f)$ or the generating vector of the polynomial lattice rule, depending on the context.

It has been shown (see [13, Theorem 4.43]) that for a given polynomial f there always exists a vector of polynomials $\boldsymbol{g}$ such that $P(\boldsymbol{g}, f)$ has low star discrepancy. This result was obtained by averaging over all possible choices of $\boldsymbol{g}$. Hence good examples of generating vectors $\boldsymbol{g}$ were not known. Here we show that such vectors of polynomials $\boldsymbol{g}$ can be found using computers. First we consider the classical unweighted case. This is done in the following section, where we introduce a component-by-component and a Korobov construction algorithm for polynomial lattice rules based on the quantity $R(\boldsymbol{g}, f)$, which is intimately related to the star discrepancy. We show that the resulting point set $P(\boldsymbol{g}, f)$ has low star discrepancy. In section 3 we extend those results to the weighted case where we obtain similar results. (A similar approach was taken by Joe [8] for ordinary lattice rules.) Note that our algorithms allow us to adjust the digital net to given weights, thereby paying more attention to important projections. Construction algorithms which allow such an adjustment were first introduced for lattice rules; see [8]. Further note that our approach is different from the classical approach for (t, m, s)-nets or (t, s)-sequences where one aims at minimizing the quality parameter t. It has been shown that certain (t, s)-sequences have lower dimensional projections of poor quality (see [12]), hence the need for more flexible construction algorithms. Furthermore we note that in the weighted case we are able to obtain useful upper bounds on the star discrepancy even for large dimensions s and a relatively small number of points N (see Table 5.2 and Table 5.3). For the unweighted case useful upper bounds for such cases are known only from [5], but those results are only existence results, leaving no clue on how to construct such point sets in practice.

Our construction algorithms allow us to extend polynomial lattice rules in the dimension. In [15] Niederreiter showed the existence of polynomial lattice rules extensible in both the number of points N and the dimension. How to extend polynomial lattice rules in N remains an interesting open question.

Section 4 deals with the efficient calculation of $R(\boldsymbol{g}, f)$ and its weighted counterpart $\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g}, f)$. Concretely, we show that for a given $\boldsymbol{g}$ and f the quantities $R(\boldsymbol{g}, f)$ and $\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g}, f)$ can be computed in $O(p^m s)$ operations. In section 5 we present some numerical results and we explain how the construction cost can be further reduced by computing $R(\boldsymbol{g}, f)$ and $\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g}, f)$ recursively with an additional storage cost of $O(p^m)$.

We introduce some notation. For an arbitrary $\boldsymbol{k} = (k_1, \dots, k_s) \in \mathbb{F}_p[x]^s$, we define the "inner product"

$$\boldsymbol{k} \cdot \boldsymbol{g} = \sum_{i=1}^{s} k_i g_i$$

and we write $g \equiv 0 \,(\mathrm{mod}\, f)$ if f divides g in $\mathbb{F}_p[x]$. Further, as above, we often associate a nonnegative integer $k = \kappa_0 + \kappa_1 p + \cdots + \kappa_r p^r$ with the polynomial $k(x) = \kappa_0 + \kappa_1 x + \cdots + \kappa_r x^r \in \mathbb{F}_p[x]$ and vice versa.

Further let $G_{p,m} := \{h \in \mathbb{F}_p[x] : \deg(h) < m\}$ and let $|G_{p,m}|$ denote the number of elements of $G_{p,m}$.

2. The classical star discrepancy. In this section we deal with the classical star discrepancy of the digital net $P(\boldsymbol{g}, f)$, where the base p is restricted to prime numbers. We show that good polynomials $g_1, \dots, g_s$ (by good polynomials we mean polynomials such that $P(\boldsymbol{g}, f)$ has low star discrepancy) may be obtained by using a component-by-component or Korobov construction algorithm. Our algorithm depends on the quantity $R(\boldsymbol{g}, f)$, which we will introduce below.

For $h \in G_{p,m}$ we define

$$r_p(h) := \begin{cases} 1 & \text{if } h = 0, \\ \frac{1}{p^{g+1} \sin^2\left(\frac{\pi}{p}\kappa_g\right)} & \text{if } h = \kappa_0 + \kappa_1 x + \cdots + \kappa_g x^g,\ \kappa_g \neq 0. \end{cases}$$

For $f \in \mathbb{F}_p[x]$, with $\deg(f) = m$, and $\boldsymbol{g} = (g_1, \dots, g_s) \in G_{p,m}^s$ we define the quantity

$$R(\boldsymbol{g}, f) := \sum_{\substack{\boldsymbol{h} \in G_{p,m}^s \setminus \{\boldsymbol{0}\} \\ \boldsymbol{h} \cdot \boldsymbol{g} \equiv 0 \pmod{f}}} \prod_{i=1}^{s} r_p(h_i),$$

where for $\boldsymbol{h} \in G_{p,m}^s$ we write $\boldsymbol{h} = (h_1, \dots, h_s)$. With this definitions we obtain the following proposition.

PROPOSITION 2.1. *For the star discrepancy $D_N^*(\boldsymbol{g}, f)$ of the point set $P(\boldsymbol{g}, f)$ we have*

$$D_N^*(\boldsymbol{g}, f) \leq 1 - \left(1 - \frac{1}{N}\right)^s + R(\boldsymbol{g}, f) \leq \frac{s}{N} + R(\boldsymbol{g}, f), \tag{2.1}$$

where $N = p^m$.

Proof. From [6, Theorem 1(ii)] together with equality (2.2) in section 2.3 it follows that

$$D_N^*(\boldsymbol{g}, f) \leq 1 - \left(1 - \frac{1}{N}\right)^s + \sum_{\substack{\boldsymbol{h} \in G_{p,m}^s \setminus \{\boldsymbol{0}\} \\ \boldsymbol{h} \cdot \boldsymbol{g} \equiv 0 \pmod{f}}} \prod_{i=1}^{s} \rho_{\text{Walsh}}^*(h_i),$$

where

$$\rho_{\text{Walsh}}^*(h) := \begin{cases} 1 & \text{if } h = 0, \\ \frac{1}{p^{g+1} \sin\left(\frac{\pi}{p}\kappa_g\right)} & \text{if } h = \kappa_0 + \kappa_1 x + \cdots + \kappa_g x^g,\ \kappa_g \neq 0. \end{cases}$$

Now the result follows by observing that $\sin^{-1}(x) \leq \sin^{-2}(x)$ for $0 < x < \pi$. □

Observe that an analogue's result exists for lattice rules, which was used by Joe in [8] to obtain lattice rules with small (weighted) star discrepancy. Indeed, we use an analogues approach to [8].

We remark that a result similar to Proposition 2.1 holds for the so-called extreme discrepancy, which is an unanchored version of the star discrepancy; see [6]. This can be obtained by using [6, Theorem 1(i)] together with equality (2.2) in section 2.3 and hence the subsequent results can be modified to obtain construction algorithms yielding point sets with small extreme discrepancy.

As it is much easier to analyze $R(\boldsymbol{g}, f)$ than $D_N^*(\boldsymbol{g}, f)$, we will subsequently mainly deal with $R(\boldsymbol{g}, f)$ rather than the star discrepancy directly. The results on the star discrepancy are then obtained via inequality (2.1).

Note that our definition of $r_p(h)$ yields a slightly weaker bound on the star discrepancy than by using the original definition of ρ_{Walsh}^* from [6]. But this change makes it possible to compute $R(\boldsymbol{g}, f)$ at a cost of $O(p^m s)$ as shown in section 4. We will exploit this fact in our construction algorithms.

The following lemma will be useful for our subsequent investigations.

LEMMA 2.2. *For $p \in \mathbb{N}$, $p \geq 2$, we have*

$$\sum_{\boldsymbol{h} \in G_{p,m}^s} \prod_{i=1}^{s} r_p(h_i) = \left(1 + m\frac{p^2 - 1}{3p}\right)^s.$$

Proof. For $p = 2$ the result is proved in [13, Lemma 3.13]. We have

$$\sum_{\boldsymbol{h} \in G_{p,m}^s} \prod_{i=1}^{s} r_p(h_i) = \prod_{i=1}^{s} \sum_{h_i \in G_{p,m}} r_p(h_i) = \left(\sum_{h \in G_{p,m}} r_p(h)\right)^s.$$

Now for $h \in G_{p,m}$, with $\deg(h) = a-1$, we write $h = h_0 + h_1 x + \cdots + h_{a-1}x^{a-1}$, where $h_{a-1} \neq 0$. Then

$$\begin{aligned}\sum_{h \in G_{p,m}} r_p(h) &= 1 + \sum_{a=1}^{m} \sum_{\substack{h \in G_{p,m} \\ \deg(h)=a-1}} r_p(h) \\ &= 1 + \sum_{a=1}^{m} \frac{1}{p^a} \sum_{\substack{h \in G_{p,m} \\ \deg(h)=a-1}} \frac{1}{\sin^2\left(\frac{h_{a-1}\pi}{p}\right)} \\ &= 1 + \sum_{a=1}^{m} \frac{1}{p^a} \sum_{h_{a-1}=1}^{p-1} p^{a-1} \frac{1}{\sin^2\left(\frac{h_{a-1}\pi}{p}\right)} \\ &= 1 + \frac{1}{p} \sum_{a=1}^{m} \sum_{h_{a-1}=1}^{p-1} \frac{1}{\sin^2\left(\frac{h_{a-1}\pi}{p}\right)}.\end{aligned}$$

In [3, Appendix C] it was shown that $\sum_{h=1}^{p-1} \frac{1}{\sin^2\left(\frac{h\pi}{p}\right)} = \frac{p^2-1}{3}$. The result follows. □

As a benchmark we calculate the average of $R(\boldsymbol{g}, f)$ over all vectors $\boldsymbol{g} \in (G_{p,m} \setminus \{0\})^s$. A similar result was proved by Niederreiter [13, Remark 4.44]. We have the following theorem.

THEOREM 2.3. *Let $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$. *We have*

$$\frac{1}{|G_{p,m} \setminus \{0\}|^s} \sum_{\boldsymbol{g} \in (G_{p,m}\setminus\{0\})^s} R(\boldsymbol{g}, f) = \frac{1}{p^m - 1}\left(\left(1 + m\frac{p^2-1}{3p}\right)^s - 1 - sm\frac{p^2-1}{3p}\right).$$

Proof. First observe that $|G_{p,m} \setminus \{0\}| = p^m - 1$. We have

$$\begin{aligned}\frac{1}{|G_{p,m} \setminus \{0\}|^s} \sum_{\boldsymbol{g} \in (G_{p,m}\setminus\{0\})^s} R(\boldsymbol{g}, f) &= \frac{1}{(p^m-1)^s} \sum_{\boldsymbol{g} \in (G_{p,m}\setminus\{0\})^s} \sum_{\substack{\boldsymbol{h} \in G_{p,m}^s \setminus\{\boldsymbol{0}\} \\ \boldsymbol{h}\cdot\boldsymbol{g} \equiv 0 \,(\mathrm{mod}\, f)}} \prod_{i=1}^{s} r_p(h_i) \\ &= \frac{1}{(p^m-1)^s} \sum_{\boldsymbol{h} \in G_{p,m}^s \setminus\{\boldsymbol{0}\}} \prod_{i=1}^{s} r_p(h_i) \sum_{\substack{\boldsymbol{g} \in (G_{p,m}\setminus\{0\})^s \\ \boldsymbol{h}\cdot\boldsymbol{g} \equiv 0 \,(\mathrm{mod}\, f)}} 1.\end{aligned}$$

For a fixed $\boldsymbol{h} \in G_{p,m}^s \setminus \{\boldsymbol{0}\}$ we have

$$\sum_{\substack{\boldsymbol{g} \in (G_{p,m}\setminus\{0\})^s \\ \boldsymbol{h}\cdot\boldsymbol{g} \equiv 0 \,(\mathrm{mod}\, f)}} 1 = |\{\boldsymbol{g} \in (G_{p,m} \setminus \{0\})^s : \boldsymbol{h}\cdot\boldsymbol{g} \equiv 0 \,(\mathrm{mod}\, f)\}|.$$

If $\boldsymbol{h} = (0, \dots, 0, h_i, 0, \dots, 0)$, with $h_i \neq 0$, then there is no polynomial $\boldsymbol{g} \in (G_{p,m} \setminus \{0\})^s$ such that $\boldsymbol{h}\cdot\boldsymbol{g} = h_i g_i \equiv 0 \,(\mathrm{mod}\, f)$, as $g_i \neq 0$ and f is irreducible. Otherwise the number of polynomials $\boldsymbol{g} \in (G_{p,m} \setminus \{0\})^s$ is $(p^m-1)^{s-1}$. Therefore we have

$$\begin{aligned}&\frac{1}{|G_{p,m} \setminus \{0\}|^s} \sum_{\boldsymbol{g} \in (G_{p,m}\setminus\{0\})^s} R(\boldsymbol{g}, f) \\ &\quad= \frac{1}{p^m-1} \sum_{\boldsymbol{h} \in G_{p,m}^s \setminus\{\boldsymbol{0}\}} \prod_{i=1}^{s} r_p(h_i) - \frac{1}{p^m-1} \sum_{i=1}^{s} \sum_{h_i \in G_{p,m}\setminus\{0\}} r_p(h_i) \prod_{\substack{j=1 \\ j\neq i}}^{s} r_p(0).\end{aligned}$$

The result now follows from Lemma 2.2. □

Remark 2.4. Niederreiter [13, Theorem 4.43] also proved a similar result for arbitrary (not necessarily irreducible) polynomials $f \in \mathbb{F}_p[x]$.

2.1. A component-by-component construction of $P(g, f)$ based on $R(g, f)$. So far we know from Theorem 2.3 and Proposition 2.1 that there exist polynomials which yield point sets with low star discrepancy. In the following we show how good polynomials can be found by computer search.

Algorithm 2.5. *Given a prime p, $m \geq 1$ and a polynomial $f \in \mathbb{F}_p[x]$, with* $\deg(f) = m$,

```
1. set $g_1^* = 1$;
2. for $d = 2, 3, \ldots, s$ find $g_d^* \in G_{p,m} \setminus \{0\}$ by minimizing $R((g_1^*, \ldots, g_{d-1}^*, g_d), f)$.
```

Remark 2.6. In section 4 it is shown how the quantity $R(\boldsymbol{g}, f)$ can be calculated in $O(p^m s)$ operations. Hence the cost for Algorithm 2.5 is of $O(p^{2m} s^2)$ operations. This order is the same as for other component-by-component construction algorithms; see [2, 8, 19]. Further, in section 5 we explain how the construction cost can be reduced to $O(p^{2m} s)$ operations with an additional storage cost of $O(p^m)$ (see also [19]).

Theorem 2.7. *Let p be prime, $m \geq 1$, and $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$. *Suppose $\boldsymbol{g}^* = (g_1^*, \ldots, g_s^*) \in (G_{p,m} \setminus \{0\})^s$ is constructed according to Algorithm* 2.5. *Then for all $d = 1, 2, \ldots, s$ we have*

$$R((g_1^*, \ldots, g_d^*), f) \leq \frac{1}{p^m - 1} \left(1 + m\frac{p^2 - 1}{3p}\right)^d.$$

Proof. Since f is irreducible it follows that $R(1, f) = 0$ and the result follows for $d = 1$. Suppose now that for some $2 \leq d < s$ we have already constructed $\boldsymbol{g}^* \in (G_{p,m} \setminus \{0\})^d$ and

$$R(\boldsymbol{g}^*, f) \leq \frac{1}{p^m - 1} \left(1 + m\frac{p^2 - 1}{3p}\right)^d.$$

Now we consider $R((\boldsymbol{g}^*, g_{d+1}), f)$. We have

$$\begin{aligned} R((\boldsymbol{g}^*, g_{d+1}), f) &= \sum_{\substack{(\boldsymbol{h}, h_{d+1}) \in G_{p,m}^{d+1} \setminus \{\boldsymbol{0}\} \\ \boldsymbol{h} \cdot \boldsymbol{g}^* + h_{d+1} g_{d+1} \equiv 0 \pmod{f}}} \prod_{i=1}^{d+1} r_p(h_i) \\ &= \sum_{\substack{\boldsymbol{h} \in G_{p,m}^d \setminus \{\boldsymbol{0}\} \\ \boldsymbol{h} \cdot \boldsymbol{g}^* \equiv 0 \pmod{f}}} \prod_{i=1}^{d} r_p(h_i) + \theta(g_{d+1}) \\ &= R(\boldsymbol{g}^*, f) + \theta(g_{d+1}), \end{aligned}$$

where

$$\theta(g_{d+1}) = \sum_{h_{d+1} \in G_{p,m} \setminus \{0\}} \left(r_p(h_{d+1}) \sum_{\substack{\boldsymbol{h} \in G_{p,m}^d \\ \boldsymbol{h} \cdot \boldsymbol{g}^* \equiv -h_{d+1} g_{d+1} \pmod{f}}} \prod_{i=1}^{d} r_p(h_i) \right).$$

Since g_{d+1}^* is a minimizer of $R((\boldsymbol{g}^*, g_{d+1}), f)$ it follows that g_{d+1}^* is also a minimizer of $\theta(g_{d+1})$ and hence we obtain

$$\theta(g_{d+1}^*) \leq \frac{1}{p^m - 1} \sum_{g_{d+1} \in G_{p,m} \setminus \{0\}} \theta(g_{d+1}).$$

Now we have

$$\begin{aligned}
&\theta(g_{d+1}^*) \\
&\leq \frac{1}{p^m-1} \sum_{g_{d+1} \in G_{p,m} \setminus \{0\}} \sum_{h_{d+1} \in G_{p,m} \setminus \{0\}} \left(r_p(h_{d+1}) \sum_{\substack{\boldsymbol{h} \in G_{p,m}^d \\ \boldsymbol{h} \cdot \boldsymbol{g}^* \equiv -h_{d+1} g_{d+1} \pmod{f}}} \prod_{i=1}^{d} r_p(h_i) \right) \\
&= \frac{1}{p^m-1} \sum_{h_{d+1} \in G_{p,m} \setminus \{0\}} r_p(h_{d+1}) \sum_{\boldsymbol{h} \in G_{p,m}^d} \prod_{i=1}^{d} r_p(h_i) \sum_{\substack{g_{d+1} \in G_{p,m} \setminus \{0\} \\ g_{d+1} h_{d+1} \equiv -\boldsymbol{h} \cdot \boldsymbol{g}^* \pmod{f}}} 1.
\end{aligned}$$

Since $\gcd(h_{d+1}, f) = 1$ it follows that the congruence

$$g_{d+1} h_{d+1} \equiv -\boldsymbol{h} \cdot \boldsymbol{g}^* \pmod{f}$$

has exactly one solution $g_{d+1} \in G_{p,m} \setminus \{0\}$ if $-\boldsymbol{h} \cdot \boldsymbol{g}^* \not\equiv 0 \pmod{f}$ and no solution if $-\boldsymbol{h} \cdot \boldsymbol{g}^* \equiv 0 \pmod{f}$. Therefore we obtain

$$\begin{aligned}
\theta(g_{d+1}^*) &\leq \frac{1}{p^m-1} \sum_{h_{d+1} \in G_{p,m} \setminus \{0\}} r_p(h_{d+1}) \sum_{\boldsymbol{h} \in G_{p,m}^d} \prod_{i=1}^{d} r_p(h_i) \\
&= \frac{1}{p^m-1} \left(1 + m\frac{p^2-1}{3p}\right)^d \sum_{h_{d+1} \in G_{p,m} \setminus \{0\}} r_p(h_{d+1}).
\end{aligned}$$

Now we obtain

$$\begin{aligned}
R((\boldsymbol{g}^*, g_{d+1}^*), f) &\leq R(\boldsymbol{g}^*, f) + \frac{1}{p^m-1} \left(1 + m\frac{p^2-1}{3p}\right)^d \sum_{h_{d+1} \in G_{p,m} \setminus \{0\}} r_p(h_{d+1}) \\
&\leq \frac{1}{p^m-1} \left(1 + m\frac{p^2-1}{3p}\right)^d \sum_{h_{d+1} \in G_{p,m}} r_p(h_{d+1}) \\
&= \frac{1}{p^m-1} \left(1 + m\frac{p^2-1}{3p}\right)^{d+1}.
\end{aligned}$$

The result follows by induction. □

From inequality (2.1) and Theorem 2.7 we obtain the following corollary.

COROLLARY 2.8. *Let p be prime, $m \geq 1$, and $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$. *Suppose $\boldsymbol{g}^* = (g_1^*, \dots, g_s^*) \in G_{p,m}^s$ is constructed according to Algorithm* 2.5. *Then for all $d = 1, \dots, s$ we have*

$$D_{p^m}^*((g_1^*, \dots, g_d^*), f) \leq \frac{d}{p^m} + \frac{1}{p^m-1} \left(1 + m\frac{p^2-1}{3p}\right)^d.$$

2.2. A Korobov construction of $\boldsymbol{P(g, f)}$ based on $\boldsymbol{R(g, f)}$. In the method of good lattice points one often restricts the attention to lattice points whose coordinates are successive powers of a single integer. Such a choice was first proposed by Korobov (see [9]) and therefore such lattice points are often called *Korobov lattice points.* Here we consider now s-tuples $\boldsymbol{g} = (g_1, \dots, g_s)$ of polynomials that are obtained by taking a polynomial $g \in G_{p,m}$ and putting $g_i \equiv g^{i-1} \pmod{f}$ with $\deg g_i < m$ for $1 \leq i \leq s$. For such s-tuples we use the notation $\boldsymbol{v}_s(g) \equiv (1, g, g^2, \dots, g^{s-1}) \pmod{f}$.

Algorithm 2.9. *Given a prime p, a dimension $s \geq 2$, $m \geq 1$, and a polynomial $f \in \mathbb{F}_p[x]$, with* $\deg(f) = m$, *find $g^* \in G_{p,m} \setminus \{0\}$ by minimizing $R(\boldsymbol{v}_s(g), f)$.*

Remark 2.10. In section 4 it is shown how the quantity $R(\boldsymbol{g}, f)$ can be calculated in $O(p^m s)$ operations. Hence the cost for Algorithm 2.9 is of $O(p^{2m} s)$ operations. This order is the same as for other Korobov construction algorithms; see [22]. Note that compared to Algorithm 2.5 the search cost is reduced, or, if one uses the method explained in section 5, there is no additional storage cost.

Theorem 2.11. *Let p be prime, $s \geq 2$, $m \geq 1$, and $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$. *Suppose $g^* \in G_{p,m} \setminus \{0\}$ is constructed according to Algorithm* 2.9. *Then we have*

$$R(\boldsymbol{v}_s(g^*), f) \leq \frac{s-1}{p^m - 1} \left(1 + m\frac{p^2-1}{3p}\right)^s.$$

Proof. Define

$$M_s(f) := \frac{1}{p^m - 1} \sum_{g \in G_{p,m} \setminus \{0\}} R(\boldsymbol{v}_s(g), f).$$

It follows from Algorithm 2.9 that $R(\boldsymbol{v}_s(g^*), f) \leq M_s(f)$. Hence it suffices to show that $M_s(f)$ satisfies the bound. We have

$$\begin{aligned} M_s(f) &= \frac{1}{p^m - 1} \sum_{g \in G_{p,m} \setminus \{0\}} \sum_{\substack{\boldsymbol{h} \in G_{p,m}^s \setminus \{\boldsymbol{0}\} \\ \boldsymbol{v}_s(g) \cdot \boldsymbol{h} \equiv 0 \pmod{f}}} \prod_{i=1}^{s} r_p(h_i) \\ &= \frac{1}{p^m - 1} \sum_{\boldsymbol{h} \in G_{p,m}^s \setminus \{\boldsymbol{0}\}} \prod_{i=1}^{s} r_p(h_i) \sum_{\substack{g \in G_{p,m} \setminus \{0\} \\ \boldsymbol{v}_s(g) \cdot \boldsymbol{h} \equiv 0 \pmod{f}}} 1. \end{aligned}$$

Now we recall that for an irreducible polynomial $f \in \mathbb{F}_p[x]$, with $\deg(f) = m \geq 1$, and a nonzero $(h_1, \dots, h_s) \in \mathbb{F}_p[x]^s$ with $\deg(h_i) < m$, $i = 1, \dots, s$, the congruence

$$h_1 + h_2 g + \cdots + h_s g^{s-1} \equiv 0 \pmod{f}$$

has at most $s - 1$ solutions $g \in G_{p,m} \setminus \{0\}$. Thus we have

$$M_s(f) \leq \frac{s-1}{p^m - 1} \sum_{\boldsymbol{h} \in G_{p,m}^s} \prod_{i=1}^{s} r_p(h_i).$$

The result now follows from Lemma 2.2. □

From inequality (2.1) and Theorem 2.11 we obtain the following corollary.

Corollary 2.12. *Let p be prime, $s \geq 2$, $m \geq 1$, and $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$. *Suppose $g^* \in G_{p,m} \setminus \{0\}$ is constructed according to Algorithm* 2.9. *Then we have*

$$D_{p^m}^*(\boldsymbol{v}_s(g^*), f) \leq \frac{s}{p^m} + \frac{s-1}{p^m - 1} \left(1 + m\frac{p^2-1}{3p}\right)^s.$$

2.3. Walsh functions and a formula for $R(\boldsymbol{g}, f)$. Before we close this section we show that the quantity $R(\boldsymbol{g}, f)$ can be represented in terms of Walsh functions. For an integer $b \geq 2$ let $\omega_b = e^{2\pi \mathrm{i}/b}$. For a nonnegative integer h with base b representation $h = h_0 + h_1 b + \cdots + h_r b^r$ the function ${}_b\mathrm{wal}_h : \mathbb{R} \longrightarrow \mathbb{C}$, periodic with period one, is defined by

$$ {}_b\mathrm{wal}_h(x) = \omega_b^{h_0 x_1 + \cdots + h_r x_{r+1}}, $$

where $x \in [0, 1)$ has base b representation $x = x_1/b + x_2/b^2 + \cdots$ (unique in the sense that infinitely many of the x_i must be different from $b - 1$).

It is clear from the definition that Walsh functions are piecewise constant. Further it can be shown that for any $b \geq 2$ the system $\{{}_b\mathrm{wal}_h : h = 0, 1, \dots\}$ is a complete orthonormal system in $L_2([0, 1))$. More information on Walsh functions can be found, for example, in [1, 3, 16, 17, 21].

Subsequently we will make use of the following equality, which follows from [13, Lemma 2.20] and [3, Lemma 2]. For the point set $P(\boldsymbol{g}, f) = \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{p^m-1}\}$ with $\boldsymbol{x}_n = (x_n^{(1)}, \dots, x_n^{(s)})$ we have

$$ \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i=1}^{s} {}_p\mathrm{wal}_{h_i}(x_n^{(i)}) = \begin{cases} 1 & \text{if } \boldsymbol{g} \cdot \boldsymbol{h} \equiv 0 \,(\mathrm{mod}\, f), \\ 0 & \text{otherwise.} \end{cases} \tag{2.2} $$

As we always consider Walsh functions in base p we will often write wal_h instead of ${}_p\mathrm{wal}_h$.

Lemma 2.13. *Let $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{p^m-1}$ be the point set $P(\boldsymbol{g}, f)$, $\boldsymbol{x}_n = (x_n^{(1)}, \dots, x_n^{(s)})$, $0 \leq n \leq p^m - 1$. Then we have*

$$ R(\boldsymbol{g}, f) = -1 + \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i=1}^{s} \left(1 + \sum_{h=1}^{p^m-1} r_p(h) \mathrm{wal}_h(x_n^{(i)}) \right). $$

Proof. Note that here we use the identification of a polynomial $h_0 + h_1 x + \cdots + h_{m-1} x^{m-1} \in G_{p,m}$ with the integer with base p representation $h_0 + h_1 p + \cdots + h_{m-1} p^{m-1}$. We have

$$ \begin{aligned} -1 + \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i=1}^{s} \left(1 + \sum_{h=1}^{p^m-1} r_p(h) \mathrm{wal}_h(x_n^{(i)}) \right) & \\ &= -1 + \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i=1}^{s} \sum_{h=0}^{p^m-1} r_p(h) \mathrm{wal}_h(x_n^{(i)}) \\ &= -1 + \frac{1}{p^m} \sum_{n=0}^{p^m-1} \sum_{h_1, \dots, h_s = 0}^{p^m-1} \prod_{i=1}^{s} r_p(h_i) \mathrm{wal}_{h_i}(x_n^{(i)}) \\ &= -1 + \sum_{h_1, \dots, h_s = 0}^{p^m-1} \prod_{i=1}^{s} r_p(h_i) \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i=1}^{s} \mathrm{wal}_{h_i}(x_n^{(i)}). \end{aligned} $$

By using (2.2) it follows that the last sum equals

$$ -1 + \sum_{\substack{\boldsymbol{h} \in G_{p,m}^s \\ \boldsymbol{g} \cdot \boldsymbol{h} \equiv 0 \,(\mathrm{mod}\, f)}} \prod_{i=1}^{s} r_p(h_i) = \sum_{\substack{\boldsymbol{h} \in G_{p,m}^s \setminus \{\boldsymbol{0}\} \\ \boldsymbol{g} \cdot \boldsymbol{h} \equiv 0 \,(\mathrm{mod}\, f)}} \prod_{i=1}^{s} r_p(h_i) = R(\boldsymbol{g}, f) $$

and the result follows. □

3. The weighted star discrepancy. In this section we extend the results from the previous section to the weighted case. First we find an analogue to inequality (2.1) for the weighted star discrepancy.

For the weighted star discrepancy $D^*_{N,\boldsymbol{\gamma}}$ of a point set $\boldsymbol{x}_0,\dots,\boldsymbol{x}_{N-1}$ in $[0,1)^s$ we have

$$D^*_{N,\boldsymbol{\gamma}} = \sup_{\boldsymbol{z}\in[0,1)^s} \max_{\substack{u\subseteq E\\ u\neq\emptyset}} \gamma_u |\Delta(\boldsymbol{z}_u,\mathbf{1})| \leq \max_{\substack{u\subseteq E\\ u\neq\emptyset}} \gamma_u \sup_{\boldsymbol{z}_u\in[0,1)^{|u|}} |\Delta(\boldsymbol{z}_u)| \leq \sum_{\substack{u\subseteq E\\ u\neq\emptyset}} \gamma_u D^*_N(u),$$

where $D^*_N(u)$ denotes the star discrepancy of the projection of the point set $\boldsymbol{x}_0,\dots,\boldsymbol{x}_{N-1}$ to the coordinates given by u. If we consider the point set $P(\boldsymbol{g},f)$, then (2.1) yields

$$D^*_N(u) \leq 1-\left(1-\frac{1}{N}\right)^{|u|} + R(\boldsymbol{g}_u,f)$$

for $u\neq\emptyset$, where $\boldsymbol{g}_u=(g_j)_{j\in u}$ and $R(\boldsymbol{g}_u,f)$ is given by

$$R(\boldsymbol{g}_u,f) = \sum_{\substack{\boldsymbol{h}\in G^{|u|}_{p,m}\setminus\{\boldsymbol{0}\}\\ \boldsymbol{h}\cdot\boldsymbol{g}_u\equiv 0 \pmod f}} \prod_{i=1}^{|u|} r_p(h_i).$$

Hence for the weighted star discrepancy $D^*_{N,\boldsymbol{\gamma}}$ of the point set $P(\boldsymbol{g},f)$ we get

$$D^*_{N,\boldsymbol{\gamma}}(\boldsymbol{g},f) \leq \sum_{\substack{u\subseteq E\\ u\neq\emptyset}} \gamma_u \left(1-\left(1-\frac{1}{N}\right)^{|u|}\right) + \widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g},f), \tag{3.1}$$

where

$$\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g},f) := \sum_{\substack{u\subseteq E\\ u\neq\emptyset}} \gamma_u R(\boldsymbol{g}_u,f).$$

Remark 3.1. It was proved by Joe [8] that if the sequence of weights $(\gamma_i)_{i\geq 1}$ satisfies $\sum_{i=1}^\infty \gamma_i<\infty$, then we have

$$\sum_{\substack{u\subseteq E\\ u\neq\emptyset}} \gamma_u \left(1-\left(1-\frac{1}{N}\right)^{|u|}\right) \leq \frac{\max(1,\Gamma)e^{\sum_{i=1}^\infty \gamma_i}}{N} \quad \text{for all } s\geq 1,$$

where $\Gamma := \sum_{i=1}^\infty \frac{\gamma_i}{1+\gamma_i}$.

In the following proposition we obtain a formula for $\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g},f)$.

PROPOSITION 3.2. *We have*

$$\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g},f) = \sum_{\substack{\boldsymbol{h}\in G^{s}_{p,m}\setminus\{\boldsymbol{0}\}\\ \boldsymbol{g}\cdot\boldsymbol{h}\equiv 0 \pmod f}} \prod_{i=1}^{s} \widetilde{r}_p(h_i,\gamma_i),$$

where

$$\widetilde{r}_p(h,\gamma) := \begin{cases} 1+\gamma & \text{if } h=0,\\ \gamma r_p(h) & \text{if } h\neq 0.\end{cases} \tag{3.2}$$

Proof. Let $\boldsymbol{x}_0, \dots, \boldsymbol{x}_{p^m-1}$ be the point set $P(\boldsymbol{g}, f)$, $\boldsymbol{x}_n = (x_n^{(1)}, \dots, x_n^{(s)})$, $0 \le n \le p^m - 1$. From Lemma 2.13 it follows that for $u \neq \emptyset$ we have

$$R(\boldsymbol{g}_u, f) = -1 + \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i \in u} \left(1 + \sum_{h=1}^{p^m-1} r_p(h) \mathrm{wal}_h(x_n^{(i)})\right).$$

Now we have

$$\begin{aligned}
&\sum_{\substack{u \subseteq E \\ u \neq \emptyset}} \gamma_u R(\boldsymbol{g}_u, f) \\
&= -\sum_{\substack{u \subseteq E \\ u \neq \emptyset}} \gamma_u + \sum_{\substack{u \subseteq E \\ u \neq \emptyset}} \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i \in u} \gamma_i \left(1 + \sum_{h=1}^{p^m-1} r_p(h) \mathrm{wal}_h(x_n^{(i)})\right) \\
&= -\left(-1 + \prod_{i=1}^{s} (1+\gamma_i)\right) + \frac{1}{p^m} \sum_{n=0}^{p^m-1} \left(-1 + \prod_{i=1}^{s} \left(1 + \gamma_i + \gamma_i \sum_{h=1}^{p^m-1} r_p(h) \mathrm{wal}_h(x_n^{(i)})\right)\right) \\
&= -\prod_{i=1}^{s} (1+\gamma_i) + \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i=1}^{s} \left(\sum_{h=0}^{p^m-1} \widetilde{r}_p(h, \gamma_i) \mathrm{wal}_h(x_n^{(i)})\right) \\
&= -\prod_{i=1}^{s} (1+\gamma_i) + \sum_{\boldsymbol{h} \in G_{p,m}^s} \prod_{i=1}^{s} \widetilde{r}_p(h_i, \gamma_i) \frac{1}{p^m} \sum_{n=0}^{p^m-1} \prod_{i=1}^{s} \mathrm{wal}_{h_i}(x_n^{(i)}) \\
&= -\prod_{i=1}^{s} (1+\gamma_i) + \sum_{\substack{\boldsymbol{h} \in G_{p,m}^s \\ \boldsymbol{g} \cdot \boldsymbol{h} \equiv 0 \ (\mathrm{mod}\ f)}} \prod_{i=1}^{s} \widetilde{r}_p(h_i, \gamma_i) \\
&= \sum_{\substack{\boldsymbol{h} \in G_{p,m}^s \setminus \{\boldsymbol{0}\} \\ \boldsymbol{g} \cdot \boldsymbol{h} \equiv 0 \ (\mathrm{mod}\ f)}} \prod_{i=1}^{s} \widetilde{r}_p(h_i, \gamma_i),
\end{aligned}$$

where we used (2.2). □

Proposition 3.2 shows that R and $\widetilde{R}_{\boldsymbol{\gamma}}$ differ only by the definitions of r_p and $\widetilde{r}_p$. Hence the main part of the proofs of the theorems in section 2 apply also for the weighted case. Only Lemma 2.2 needs to be established using $\widetilde{r}_p$. This is done subsequently.

LEMMA 3.3. *Let $\widetilde{r}_p(h, \gamma)$ be given by (3.2). Then we have*

$$\sum_{\boldsymbol{h} \in G_{p,m}^s} \prod_{i=1}^{s} \widetilde{r}_p(h_i, \gamma_i) = \prod_{i=1}^{s} \left(1 + \gamma_i \left(1 + m \frac{p^2-1}{3p}\right)\right).$$

Proof. We have

$$\begin{aligned}
\sum_{\boldsymbol{h} \in G_{p,m}^s} \prod_{i=1}^{s} \widetilde{r}_p(h_i, \gamma_i) &= \prod_{i=1}^{s} \sum_{h \in G_{p,m}} \widetilde{r}_p(h, \gamma_i) = \prod_{i=1}^{s} \left(\widetilde{r}_p(0, \gamma_i) + \sum_{h \in G_{p,m} \setminus \{0\}} \widetilde{r}_p(h, \gamma_i)\right) \\
&= \prod_{i=1}^{s} \left(1 + \gamma_i + \sum_{h \in G_{p,m} \setminus \{0\}} \gamma_i r_p(h)\right) = \prod_{i=1}^{s} \left(1 + \gamma_i \sum_{h \in G_{p,m}} r_p(h)\right)
\end{aligned}$$

and hence the result follows from Lemma 2.2. □

Using Proposition 3.2 and Lemma 3.3 the proofs of the following results can be obtained from section 2.

As for the classical star discrepancy (Theorem 2.3), we can now, for a given irreducible polynomial $f \in \mathbb{F}_p[x]$ with $\deg(f) = m$, compute the average of $\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g}, f)$ over all vectors $\boldsymbol{g} \in (G_{p,m} \setminus \{0\})^s$.

Theorem 3.4. *Let $f \in \mathbb{F}_p[x]$ be irreducible with* $\deg(f) = m$. *We have*

$$\frac{1}{|G_{p,m} \setminus \{0\}|^s} \sum_{\boldsymbol{g} \in (G_{p,m} \setminus \{0\})^s} \widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g}, f) = \frac{1}{p^m - 1} \sum_{\substack{u \subseteq E \\ |u| \geq 2}} \prod_{i \in u} \left(\gamma_i \left(m \frac{p^2 - 1}{3p} \right) \right) \prod_{i \notin u} (1 + \gamma_i) \, .$$

Let $c_p > 0$ be some constant depending only on p and let $\sum_{i=1}^{\infty} \gamma_i < \infty$. Then it was shown in [7] that for every $\delta > 0$ there is some constant $C'_{\boldsymbol{\gamma},\delta} > 0$ such that

$$\text{(3.3)} \quad \frac{1}{p^m - 1} \sum_{\substack{u \subseteq E \\ |u| \geq 2}} \prod_{i \in u} (\gamma_i m c_p) \prod_{i \notin u} (1 + \gamma_i) \leq \frac{1}{p^m - 1} \prod_{i=1}^{s} (1 + \gamma_i (1 + m c_p)) \leq \frac{C'_{\boldsymbol{\gamma},\delta}}{p^{m(1-\delta)}}$$

for all $m > 0$.

Hence it follows from (3.1), Remark 3.1, and (3.3) that if $\sum_{i=1}^{\infty} \gamma_i < \infty$, then for every irreducible polynomial $f \in \mathbb{F}_p[x]$ there exists a constant $C_{\boldsymbol{\gamma},\delta}$, independent of s and m, and a sequence of polynomials $(g_i)_{i \geq 1}$, with $g_i \in G_{p,m} \setminus \{0\}$, such that the star discrepancy of $P((g_1, \ldots, g_s), f)$ satisfies

$$\text{(3.4)} \qquad D^*_{p^m,\boldsymbol{\gamma}}((g_1, \ldots, g_s), f) \leq \frac{C_{\boldsymbol{\gamma},\delta}}{p^{m(1-\delta)}} \quad \text{for all } m, s \geq 1, \text{ and } \delta > 0.$$

We emphasize that in this case the weighted star discrepancy is bounded independently of the dimension. (See [20] for a thorough discussion on (strong) tractability.)

In the following subsection we introduce an algorithm which shows how the polynomials $g_i \in G_{p,m} \setminus \{0\}$, which satisfy a bound of the form (3.4), can be found by computer search.

3.1. A component-by-component construction of $\boldsymbol{P(g, f)}$ based on $\boldsymbol{\widetilde{R}(g, f)}$.

We are now ready to formulate the weighted analogue to Algorithm 2.5 and Theorem 2.7.

Algorithm 3.5. *Given a prime p, $m \geq 1$, a polynomial $f \in \mathbb{F}_p[x]$, with* $\deg(f) = m$, *and a sequence of weights $\boldsymbol{\gamma} = (\gamma_i)_{i \geq 1}$,*

1. *set $g_1^* = 1$;*
2. *for $d = 2, 3, \ldots, s$ find $g_d^* \in G_{p,m} \setminus \{0\}$ by minimizing $\widetilde{R}_{\boldsymbol{\gamma}}((g_1^*, \ldots, g_{d-1}^*, g_d), f)$.*

Remark 3.6. In section 4 it is shown how the quantity $\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g}, f)$ can be calculated in $O(p^m s)$ operations. Hence the cost for Algorithm 3.5 is of $O(p^{2m} s^2)$ operations. This order is the same as for other component-by-component construction algorithms; see Algorithm 2.5 and [2, 8, 19]. Further, in section 5 we explain how the construction cost can be reduced to $O(p^{2m} s)$ operations with an additional storage cost of $O(p^m)$ (see also [19]).

Theorem 3.7. *Let p be prime, $m \geq 1$, and $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$. *Suppose $\boldsymbol{g}^* = (g_1^*, \ldots, g_s^*) \in G_{p,m}^s$ is constructed according to Algorithm* 3.5. *Then for all $d = 1, 2, \ldots, s$ we have*

$$\widetilde{R}_{\boldsymbol{\gamma}}((g_1^*, \ldots, g_d^*), f) \leq \frac{1}{p^m - 1} \prod_{i=1}^{s} \left(1 + \gamma_i \left(1 + m \frac{p^2 - 1}{3p} \right) \right) .$$

From inequality (3.1), Remark 3.1, and Theorem 3.7 we obtain the following corollary.

COROLLARY 3.8. *Let p be prime, $m \geq 1$, $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$, *and* $\gamma_u = \prod_{i \in u} \gamma_i$. *Suppose $\boldsymbol{g}^* \in G_{p,m}^s$ is constructed according to Algorithm* 3.5. *Then we have*

$$D^*_{p^m,\boldsymbol{\gamma}}(\boldsymbol{g}^*, f) \leq \sum_{\substack{u \subseteq E \\ u \neq \emptyset}} \gamma_u \left(1 - \left(1 - \frac{1}{p^m}\right)^{|u|}\right) + \frac{1}{p^m - 1} \prod_{i=1}^{s} \left(1 + \gamma_i \left(1 + m\frac{p^2 - 1}{3p}\right)\right).$$

Hence it follows from Remark 3.1 and (3.3) that if $\sum_{i=1}^{\infty} \gamma_i < \infty$, then there is a constant $\bar{C}_{\boldsymbol{\gamma},\delta}$, independent of s and m, such that

$$D^*_{p^m,\boldsymbol{\gamma}}(\boldsymbol{g}^*, f) \leq \frac{\bar{C}_{\boldsymbol{\gamma},\delta}}{p^{m(1-\delta)}} \quad \text{for all } \delta > 0.$$

Again we emphasize that this bound is independent of the dimension.

3.2. A Korobov construction of $P(g, f)$ based on $\widetilde{R}_\gamma(g, f)$.

As in subsection 2.2, we also have a Korobov construction algorithm for the weighted case.

ALGORITHM 3.9. *Given a prime p, a dimension $s \geq 2$, $m \geq 1$, and an irreducible polynomial $f \in \mathbb{F}_p[x]$, with* $\deg(f) = m$, *and a sequence of weights $\boldsymbol{\gamma} = (\gamma_i)_{i \geq 1}$, find $g^* \in G_{p,m} \setminus \{0\}$ by minimizing $\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{v}_s(g), f)$.*

We have the following result.

THEOREM 3.10. *Let p be a prime, $s \geq 2$, $m \geq 1$, and $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$. *A minimizer g^* obtained from Algorithm* 3.9 *satisfies*

$$\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{v}_s(g^*), f) \leq \frac{s-1}{p^m - 1} \prod_{i=1}^{s} \left(1 + \gamma_i \left(1 + m\frac{p^2 - 1}{3p}\right)\right).$$

We also obtain the following corollary.

COROLLARY 3.11. *Let p be prime, $s \geq 2$, $m \geq 1$, $f \in \mathbb{F}_p[x]$ be irreducible, with* $\deg(f) = m$, *and* $\gamma_u = \prod_{i \in u} \gamma_i$. *Suppose $g^* \in G_{p,m}$ is constructed according to Algorithm* 3.9. *Then we have*

$$D^*_{p^m,\boldsymbol{\gamma}}(\boldsymbol{v}_s(g^*), f) \leq \sum_{\substack{u \subseteq E \\ u \neq \emptyset}} \gamma_u \left(1 - \left(1 - \frac{1}{p^m}\right)^{|u|}\right) + \frac{s-1}{p^m - 1} \prod_{i=1}^{s} \left(1 + \gamma_i \left(1 + m\frac{p^2 - 1}{3p}\right)\right).$$

As in other Korobov-type construction algorithms (see [2, 22]), we obtain an upper bound which depends polynomially on the dimension. Whether an upper bound independent of the dimension can be obtained for Korobov type constructions is an open problem. (Note that this question is open not only for the star discrepancy of polynomial lattices but also for the L_2 discrepancy (see [2]) and for the L_2 discrepancy and the star discrepancy of lattices (see [8, 22]).)

4. Calculation of $R(g, f)$ and $\widetilde{R}_\gamma(g, f)$.

In this section we show how the quantities $R(\boldsymbol{g}, f)$ and $\widetilde{R}(\boldsymbol{g}, f)$ can be computed efficiently. We define

$$\phi_{p,m}(x) := \sum_{h=0}^{p^m - 1} r_p(h) \, {}_p\mathrm{wal}_h(x).$$

Let $\boldsymbol{x} = (x_1, \ldots, x_s)$, $f \in \mathbb{F}_p[x]$, with $\deg(f) = m$, and $\boldsymbol{g} \in G_{p,m}^s$. It follows from Lemma 2.13 that

$$R(\boldsymbol{g}, f) = -1 + \frac{1}{|P(\boldsymbol{g}, f)|} \sum_{\boldsymbol{x} \in P(\boldsymbol{g}, f)} \prod_{i=1}^{s} \phi_{p,m}(x_i). \tag{4.1}$$

In the following we show that the function $\phi_{p,m}$ can be simplified. Let $h = h_0 + h_1 p + \cdots + h_d p^d$ with $h_d \neq 0$. For $0 \leq d < m$ and $1 \leq h_d < p$ let

$$D_{d,h_d,p,m}(x) := \sum_{h=h_d p^d}^{h_d p^d + p^d - 1} {}_p\mathrm{wal}_h(x);$$

then it follows that

$$\phi_{p,m}(x) = 1 + \sum_{d=0}^{m-1} \sum_{h_d=1}^{p-1} r_p(h_d p^d) D_{d,h_d,p,m}(x).$$

Let $\omega_p = e^{2\pi \mathtt{i}/p}$ and $x = \frac{x_1}{p} + \frac{x_2}{p^2} + \cdots$. We have

$$D_{d,h_d,p,m}(x) = \sum_{h=h_d p^d}^{h_d p^d + p^d - 1} {}_p\mathrm{wal}_h(x) = \omega_p^{h_d x_{d+1}} \sum_{h_{d-1}=0}^{p-1} \omega_p^{h_{d-1} x_d} \cdots \sum_{h_0=0}^{p-1} \omega_p^{h_0 x_1}.$$

As $\sum_{h_i=0}^{p-1} \omega_p^{h_i x_{i+1}} = 0$ if $x_{i+1} \neq 0$ and $\sum_{h_i=0}^{p-1} \omega_p^{h_i x_{i+1}} = p$ if $x_{i+1} = 0$ we have

$$D_{d,h_d,p,m}(x) = \begin{cases} \omega_p^{h_d x_{d+1}} p^d & \text{if } x_1 = \cdots = x_d = 0 \text{ or if } d = 0, \\ 0 & \text{otherwise.} \end{cases}$$

We have $r_p(0) = 1$ and for $h > 0$ with $h = h_0 + h_1 p + \cdots + h_d p^d$ and $h_d \neq 0$ we have $r_p(h) = p^{-d-1} \sin^{-2}\left(\frac{h_d \pi}{p}\right)$.

We restate a result from [3, Appendix C] which will be used subsequently. For any $l \in \{0, \ldots, p-1\}$ we have

$$\sum_{h=1}^{p-1} \frac{\omega_p^{hl}}{\sin^2(h\pi/p)} = 2l(l-p) + \frac{p^2-1}{3}. \tag{4.2}$$

First let $x_1 = \cdots = x_m = 0$; then we have

$$\begin{aligned} \phi_{p,m}(x) &= 1 + \sum_{d=0}^{m-1} \sum_{h_d=1}^{p-1} r_p(h_d p^d) p^d = 1 + \sum_{d=0}^{m-1} p^d \sum_{h_d=1}^{p-1} \frac{1}{p^{d+1}} \frac{1}{\sin^2\left(\frac{h_d \pi}{p}\right)} \\ &= 1 + \frac{1}{p} \sum_{d=0}^{m-1} \frac{p^2-1}{3} = 1 + m\frac{p^2-1}{3p}, \end{aligned}$$

where we used (4.2) with $l = 0$.

Let $i_0 = i_0(x)$ be such that $x_1 = \cdots = x_{i_0-1} = 0$ and $x_{i_0} \neq 0$ with $1 \leq i_0 \leq m$. Then we have

$$\begin{aligned}\phi_{p,m}(x) &= 1 + \sum_{d=0}^{m-1} \sum_{h_d=1}^{p-1} r_p(h_d p^d) D_{d,h_d,p,m}(x)\\ &= 1 + \sum_{d=0}^{i_0-2} \sum_{h_d=1}^{p-1} r_p(h_d p^d) p^d + \sum_{h_{i_0-1}=1}^{p-1} r_p(h_{i_0-1} p^{i_0-1}) p^{i_0-1} \omega_p^{h_{i_0-1} x_{i_0}}.\end{aligned}$$

Now

$$\begin{aligned}\sum_{d=0}^{i_0-2} \sum_{h_d=1}^{p-1} r_p(h_d p^d) p^d &= \sum_{d=0}^{i_0-2} \sum_{h_d=1}^{p-1} p^d \frac{1}{p^{d+1}} \frac{1}{\sin^2\left(\frac{h_d \pi}{p}\right)} = \frac{1}{p} \sum_{d=0}^{i_0-2} \sum_{h_d=1}^{p-1} \frac{1}{\sin^2\left(\frac{h_d \pi}{p}\right)}\\ &= \frac{1}{p} \sum_{d=0}^{i_0-2} \frac{p^2-1}{3} = \frac{p^2-1}{3p}(i_0 - 1),\end{aligned}$$

where we used (4.2) with $l = 0$ again. Further we have

$$\begin{aligned}\sum_{h_{i_0-1}=1}^{p-1} r_p(h_{i_0-1} p^{i_0-1}) p^{i_0-1} \omega_p^{h_{i_0-1} x_{i_0}} &= \sum_{k=1}^{p-1} p^{i_0-1} \omega_p^{k x_{i_0}} \frac{1}{p^{i_0}} \frac{1}{\sin^2\left(\frac{k\pi}{p}\right)}\\ &= \frac{1}{p} \sum_{k=1}^{p-1} \frac{\omega_p^{k x_{i_0}}}{\sin^2\left(\frac{k\pi}{p}\right)}\\ &= \frac{1}{p}\left(2 x_{i_0}(x_{i_0} - p) + \frac{p^2-1}{3}\right),\end{aligned}$$

where we used (4.2) with $l = x_{i_0}$. It follows that

$$\phi_{p,m}(x) = 1 + i_0 \frac{p^2-1}{3p} + \frac{2}{p} x_{i_0}(x_{i_0} - p)\,.$$

Thus we have

$$\phi_{p,m}(x) = \begin{cases} 1 + i_0 \frac{p^2-1}{3p} + \frac{2}{p} x_{i_0}(x_{i_0} - p) & \text{if } x_1 = \cdots = x_{i_0-1} = 0 \text{ and } x_{i_0} \neq 0, \\ & \text{with } 1 \leq i_0 \leq m, \\ 1 + m \frac{p^2-1}{3p} & \text{otherwise.} \end{cases} \tag{4.3}$$

Thus using (4.1) and (4.3) we can compute $R(\boldsymbol{g}, f)$ in $O(p^m s)$ operations.

Now we turn to the weighted case. We have

$$\begin{aligned}\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g}, f) &= \sum_{\substack{u \subseteq E\\ u \neq \emptyset}} \gamma_u R(\boldsymbol{g}_u, f) = -\sum_{\substack{u \subseteq E\\ u \neq \emptyset}} \gamma_u + \frac{1}{|P(\boldsymbol{g}, f)|} \sum_{\boldsymbol{x} \in P(\boldsymbol{g}, f)} \sum_{\substack{u \subseteq E\\ u \neq \emptyset}} \gamma_u \prod_{i \in u} \phi_{p,m}(x_i)\\ &= -\prod_{i=1}^{s}(1 + \gamma_i) + \frac{1}{|P(\boldsymbol{g}, f)|} \sum_{\boldsymbol{x} \in P(\boldsymbol{g}, f)} \prod_{i=1}^{s}(1 + \gamma_i \phi_{p,m}(x_i)).\end{aligned}$$

Again this quantity can be computed in $O(p^m s)$ operations using (4.3).

5. Numerical results and discussion. In this section we show how the construction cost can be reduced to $O(p^{2m}s)$ with an additional cost of $O(p^m)$ storage (see also [19]). Further, we point out problems concerning the accuracy of the computation and how those problems can be avoided. Finally we present tables of numerical results, showing the values of $R(\boldsymbol{g}, f)$ and $\widetilde{R}_{\boldsymbol{\gamma}}(\boldsymbol{g}, f)$ of our construction algorithms, and compare them with the choices of $\boldsymbol{g}$ from the so-called Salzburg tables [18].

First we compute $\phi_{p,m}(\varphi_m(\frac{n(x)g(x)}{f(x)}))$ for all possible choices of polynomials n and g and store the results in some array. The computation of the values $\phi_{p,m}(x_{n,k})$ in the expression for R and $\widetilde{R}$ then reduces to picking the corresponding values from the array, performing some multiplication, and adding up the results. It is advantageous to save these tables because they do not depend on the weights and the dimension but only on m, and so they can be reused.

Next we compute R recursively with respect to the dimension in the component-by-component construction. Recall that for given polynomials $g_1, \ldots, g_d$ we compute N points where the kth coordinate depends on g_k. We want to minimize

$$R((g_1, \ldots, g_d, g_{d+1}), f) = -1 + \frac{1}{|P(g_{d+1}))|} \sum_{\boldsymbol{x} \in P(g_{d+1})} \prod_{i=1}^{d+1} \phi_{p,m}(x_i),$$

where we suppressed the dependence of the point set $P(g_{d+1})$ on $g_1, \ldots, g_d$. Disregarding constants this amounts to minimizing the inner product $\sum_{n=0}^{N-1} \Phi_{n,d}\phi_{p,m}(x_{n,d+1})$ where $\Phi_{n,d} = \prod_{i=1}^{d} \phi_{p,m}(x_{n,i})$. Obviously, the numbers $\Phi_{n,d}$ can be computed recursively by

$$\Phi_{n,1} = \phi_{p,m}(x_{n,1}),$$

where the $x_{n,1}$ are generated using the polynomial 1 and

$$\Phi_{n,d+1} = \Phi_{n,d}\phi_{p,m}(x_{n,d+1}),$$

where the $x_{n,d+1}$ are generated by the polynomial g_{d+1} which minimizes the inner product $\sum_{n=0}^{N-1} \Phi_{n,d}\phi_{p,m}(x_{n,d+1})$. In fact we can ignore the first addend since $x_{0,i} = 0$ for all i and therefore $\Phi_{0,d} = (1 + m\frac{p^2-1}{3p})^d$.

The weighted case is a bit more complicated. Recall that in each step we want to minimize

$$\begin{aligned}
&\widetilde{R}((g_1, \ldots, g_d, g_{d+1}), f) \\
&= -\prod_{i=1}^{d+1}(1 + \gamma_i) + \frac{1}{|P(g_{d+1}))|} \sum_{\boldsymbol{x} \in P(g_{d+1})} \prod_{i=1}^{d+1}(1 + \gamma_i \phi_{p,m}(x_i)).
\end{aligned}$$

In principle we could deal with this in the same way as before. But for small weights a computer does not distinguish between $(1 + \gamma_i \phi_{p,m}(x_i))$ and 1. So if we compute the product $\prod_{i=1}^{d+1}(1 + \gamma_i \phi_{p,m}(x_i))$ for small weights we simply get 1. This leads to the well-known effect in component-by-component construction algorithms that from some dimension onward one always gets the same optimizing polynomial.

So this is how we proceed: first disregard the additive constant $-\prod_{i=1}^{d+1}(1 + \gamma_i)$ and the multiplicative constant $\frac{1}{|P(g_{d+1}))|}$. Note that the first addend in the sum is again constant, so we may disregard it for minimization. Then write

$$\begin{aligned}\sum_{n=1}^{N-1}\prod_{i=1}^{d+1}(1+\gamma_i\phi_{p,m}(x_{n,i})) &= \sum_{n=1}^{N-1}\prod_{i=1}^{d}(1+\gamma_i\phi_{p,m}(x_{n,i})) \\ &\quad + \gamma_{d+1}\sum_{n=1}^{N-1}\left(\prod_{i=1}^{d}(1+\gamma_i\phi_{p,m}(x_{n,i}))-1\right)\phi_{p,m}(x_{n,d+1}) \\ &\quad + \gamma_{d+1}\sum_{n=1}^{N-1}\phi_{p,m}(x_{n,d+1})\,.\end{aligned}$$

Obviously the first term does not depend on g_{d+1} and is therefore irrelevant for our minimization problem. But the last term is independent of g_{d+1}, since for an arbitrary g_{d+1} the $x_{n,d+1}$ run through all nonzero m-bit numbers if n runs from 1 to $N-1$ and therefore always give the same sum.

It is therefore sufficient to minimize the inner product

$$\sum_{n=1}^{N-1}\Psi_{n,d}\,\phi_{p,m}(x_{n,d+1}),$$

where $\Psi_{n,d}=(\prod_{i=1}^{d}(1+\gamma_i\phi_{p,m}(x_{n,i}))-1)$. Note that the order of magnitude of $\Psi_{n,d}$ is not 1 (for small weights). We can compute the number $\Psi_{n,d}$ recursively by

$$\Psi_{n,1}=\gamma_1\phi_{p,m}(x_{n,1}),$$

where the $x_{n,1}$ are generated using the polynomial 1 and

$$\Psi_{n,d+1}=\Psi_{n,d}+(\Psi_{n,d}+1)\gamma_{d+1}\phi_{p,m}(x_{n,d+1}),$$

where the $x_{n,d+1}$ are generated by the polynomial g_{d+1} which minimizes the inner product $\sum_{n=1}^{N-1}\Psi_{n,d}\phi_{p,m}(x_{n,d+1})$. Finally, we start with the smallest weight first, i.e., we arrange the weights in increasing order. This seems to have the effect that we do not always get the same polynomial for small weights. Since our proof did not take into account the order of the weights we still get a point set with an $\widetilde{R}$ less than the average. However, as can be seen from Table 5.2, the resulting $\widetilde{R}$ is consistently bigger than for descending order.

In our tables we computed the best R's and $\widetilde{R}$'s for different values of $N=p^m$ and different weights, where we restrict ourselves to the case $p=2$. We write CBC for the component-by-component construction and RCBC (reversed) for the component-by-component construction with weights in ascending order. "Korobov Salztab" means using a Korobov rule with the defining polynomial taken from the Salzburg tables [18]. Since the latter were chosen to minimize the t-parameter of the net instead of R it is not a surprise that it gives a slightly inferior value for R. However, it can be seen that there are differences in the weighted cases, confirming the view that point sets which are good for the unweighted case need not be good for the weighted case. Hence there is a necessity to adjust a point set to some given weights (the weights are determined by the task at hand), and, as we have shown in this paper, this can be achieved using a component-by-component or a Korobov construction algorithm.

In our numerical examples we use point sets consisting of up to 2^{13} points. The same programs can also be used for 2^{15} points, unfortunately requiring a rather long

Table 5.1
Unweighted case for $s = 5$.

N	CBC plr	Korobov plr	Korobov Salztab
256	11.9666	11.96530	11.99270
512	9.65717	9.66028	9.69971
1024	7.47394	7.47476	7.49051
2048	5.58371	5.58585	5.58711
4096	4.04798	4.04970	4.06461
8192	2.86040	2.86191	2.86821

Table 5.2
Weighted case for $s = 50$, $\gamma_i = i^{-2}$.

N	CBC plr	RCBC plr	Korobov plr	Korobov Salztab
256	0.1862330	0.1929140	0.1907370	0.3237240
512	0.1309270	0.1341940	0.1351940	0.1737160
1024	0.0904281	0.0924048	0.0923820	0.1540320
2048	0.0612452	0.0626068	0.0627534	0.1313720
4096	0.0409122	0.0417111	0.0415957	0.1113260
8192	0.0270023	0.0274363	0.0274434	0.0348907

Table 5.3
Weighted case for $s = 50$, $\gamma_i = \frac{1}{50}$.

N	CBC plr	Korobov plr	Korobov Salztab
256	0.399518	0.398798	0.413967
512	0.325541	0.325211	0.329576
1024	0.261700	0.261230	0.268513
2048	0.207947	0.207750	0.215230
4096	0.163820	0.163738	0.172557
8192	0.128152	0.128195	0.129452

computation time, although more careful programming might reduce the computation time considerably. For an even greater number of points computing time and storage requirements become too high for a personal computer.

REFERENCES

[1] H. E. Chrestenson, *A class of generalized Walsh functions*, Pacific J. Math., 5 (1955), pp. 17–31.

[2] J. Dick, F. Y. Kuo, F. Pillichshammer, and I. H. Sloan, *Construction algorithms for polynomial lattice rules for multivariate integration*, in Math. Comp., to appear.

[3] J. Dick and F. Pillichshammer, *Multivariate integration in weighted Hilbert spaces based on Walsh functions and weighted Sobolev spaces*, J. Complexity, 21 (2005), pp. 149–195.

[4] M. Drmota and R. F. Tichy, *Sequences, Discrepancies, and Applications*, Lecture Notes in Math. 1651, Springer-Verlag, Berlin, 1997.

[5] S. Heinrich, E. Novak, G. Wasilkowski, and H. Woźniakowski, *The inverse of the star-discrepancy depends linearly on the dimension*, Acta Arith., 96 (2001), pp. 279–302.

[6] P. Hellekalek, *General discrepancy estimates: The Walsh function system*, Acta Arith., 67 (1994), pp. 209–218.

[7] F. J. Hickernell and H. Niederreiter, *The existence of good extensible rank-*1 *lattices*, J. Complexity, 19 (2003), pp. 286–300.

[8] S. Joe, *Construction of good rank-*1 *lattice rules based on the weighted star discrepancy*, submitted.

[9] N. M. Korobov, *Properties and calculation of optimal coefficients*, Dokl. Akad. Nauk SSSR, 132 (1960), pp. 1009–1012 (in Russian).

[10] L. Kuipers and H. Niederreiter, *Uniform Distribution of Sequences*, John Wiley, New York, 1974.

[11] G. Leobacher and F. Pillichshammer, *Bounds for the weighted Lp discrepancy and tractability of integration*, J. Complexity, 19 (2003), pp. 529–547.

[12] W. J. Morokoff and R. E. Caflisch, *Quasi-random sequences and their discrepancies*, SIAM J. Sci. Comput., 15 (1994), pp. 1251–1279.

[13] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conf. Ser. Appl. Math. 63, SIAM, Philadelphia, 1992.

[14] H. Niederreiter, *Low-discrepancy point sets obtained by digital constructions over finite fields*, Czechoslovak Math. J., 42 (1992), pp. 143–166.

[15] H. Niederreiter, *The existence of good extensible polynomial lattice rules*, Monatsh. Math., 139 (2003), pp. 295–307.

[16] G. Pirsic, *Schnell Konvergierende Walshreihen über Gruppen*, MS thesis, University of Salzburg, Austria, 1995; also available online from at http://www.ricam.oeaw.ac.at/people/page/pirsic/.

[17] T. J. Rivlin and E. B. Saff, *Joseph L. Walsh: Selected Papers*, Springer-Verlag, New York, 2000.

[18] W. Ch. Schmid, *Improvements and extensions of the "Salzburg Tables" by using irreducible polynomials*, in Monte Carlo and Quasi-Monte Carlo Methods, H. Niederreiter and J. Spanier, eds., Springer, Berlin, 1999, pp. 438–449.

[19] I. H. Sloan, F. Y. Kuo, and S. Joe, *Constructing randomly shifted lattice rules in weighted Sobolev spaces*, SIAM J. Numer. Anal., 40 (2002), pp. 1650–1665.

[20] I. H. Sloan and H. Woźniakowski, *When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?*, J. Complexity, 14 (1998), pp. 1–33.

[21] J. L. Walsh, *A closed set of normal orthogonal functions*, Amer. J. Math., 55 (1923), pp. 5–24.

[22] X. Wang, I. H. Sloan, and J. Dick, *On Korobov lattice rules in weighted Korobov spaces*, submitted.

SIAM J. Numer. Anal.
Vol. 43, No. 1, pp. 96–115

INEXACT SEMIMONOTONIC AUGMENTED LAGRANGIANS WITH OPTIMAL FEASIBILITY CONVERGENCE FOR CONVEX BOUND AND EQUALITY CONSTRAINED QUADRATIC PROGRAMMING[*]

Z. DOSTÁL[†]

Abstract. A variant of the augmented Lagrangian-type algorithm for strictly convex quadratic programming problems with bounds and equality constraints is considered. The algorithm exploits the adaptive precision control in the solution of auxiliary bound constraint problems in the inner loop while the Lagrange multipliers for the equality constraints are updated in the outer loop. The update rule for the penalty parameter is introduced that depends on the increase of the augmented Lagrangian. Global convergence is proved and an explicit bound on the penalty parameter is given. A qualitatively new feature of our algorithm is a bound on the feasibility error that is independent of conditioning of the constraints. When applied to the class of problems with the spectrum of the Hessian matrix in a given interval, the algorithm returns the solution in $O(1)$ matrix-vector multiplications. The results are valid even for linearly dependent constraints. Theoretical results are illustrated by numerical experiments including the solution of an elliptic variational inequality.

Key words. quadratic programming, box and equality constraints, augmented Lagrangians, adaptive precision control

AMS subject classifications. Primary, 65K05; Secondary, 90C20

DOI. 10.1137/S0036142903436393

1. Introduction.

We shall be concerned with the problem of finding a minimizer of a strictly convex quadratic function subject to simple bounds and linear equality constraints, that is

$$\text{minimize} \quad q(x) \quad \text{subject to (s.t.)} \quad x \in \Omega \tag{1.1}$$

with $\Omega = \{x \in \mathbb{R}^n : x \geq \ell \text{ and } Cx = 0\}$, $q(x) = \frac{1}{2}x^T Ax - b^T x$, $b \in \mathbb{R}^n$, $\ell \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ symmetric positive definite, and $C \in \mathbb{R}^{m\times n}$. We do not require that C is a full row rank matrix, as for some large problems it may be not trivial to verify such assumption, but we shall assume that Ω is not empty. Let us point out that confining ourselves to the homogeneous equality constraints does not mean any loss of generality, as we can use a simple transform to reduce any nonhomogeneous equality constraints to our case. Moreover, our results may be useful also in the more general case when A is positive definite only on the kernel of C, i.e., when only $R^T AR$ is positive definite with the matrix R formed by the basis of the kernel of C. The modification of our algorithm can be based on observation that in the latter case there is $\widehat{\rho} > 0$ such that $A + \widehat{\rho} C^T C$ is positive definite and that the problem (1.1) has the same solution as the problem

$$\text{minimize} \quad \widehat{q}(x) \quad \text{s.t.} \quad x \in \Omega, \quad \widehat{q}(x) = \frac{1}{2}x^T(A + \widehat{\rho} C^T C)x - b^T x.$$

[*] Received by the editors October 22, 2003; accepted for publication (in revised form) July 21, 2004; published electronically April 26, 2005. This research was supported by grants GAČR 101/01/0538 and 101/04/1145 and by the grant ME641 of the Ministry of Education of the Czech Republic.

http://www.siam.org/journals/sinum/43-1/43639.html

[†] Department of Applied Mathematics, FEI VŠB-Technical University Ostrava, 17. listopadu 15, CZ-70833 Ostrava, Czech Republic (zdenek.dostal@vsb.cz).

We shall be especially interested in problems with the matrix A reasonably conditioned and either large and sparse or defined as a product of large and sparse matrices, so that the conjugate gradient–based methods are more efficient for the unconstrained minimization of q than the direct solvers. We shall also assume that the matrix C is sparse. Such problems arise, for example, from the discretization of semicoercive elliptic variational inequalities (e.g., Dostál [7]) or from application of the duality-based domain decomposition to the contact problems of elasticity (e.g., Dostál, Friedlander, and Santos [12], Dostál, Gomes Neto, and Santos [16] or Avery et al. [1]).

We restrict our attention to the algorithms that reduce problem (1.1) to a sequence of bound constrained quadratic programming problems. Our approach has been motivated by an effort to exploit recent progress in the solution of the latter problems, namely, effective application of projections and adaptive precision control of the solution of auxiliary problems (e.g., [23, 4, 8]). Most recently, these results were combined with the results on the gradient projections [29] to get algorithms with the R-linear rate of convergence in terms of the bounds on the spectrum of A [10, 21]. It simply follows that such algorithms can solve any class of problems with the spectrum of the Hessian matrix in a given interval $[a, b]$, $a > 0$, at the cost of $O(1)$ matrix-vector multiplications. These results were applied to the development of some optimal (i.e., with linear complexity) algorithms for the solution of elliptic boundary variational inequalities [18, 19, 20].

Our development is based on the algorithm proposed by Conn, Gould and Toint [5], who adapted the augmented Lagrangian method [2, 24] of Powell [28] and Hestenes [26] to the solution of problems with a general cost function subject to general equality constraints and simple bounds. When applied to (1.1), their algorithm reduces the solution to a sequence of simple bound constrained problems of the form

$$\text{minimize} \quad L(x, \mu^k, \rho_k) \quad \text{s.t.} \quad x \geq \ell, \tag{1.2}$$

where

$$L(x, \mu^k, \rho_k) = q(x) + (\mu^k)^T Cx + \frac{\rho_k}{2}\|Cx\|^2 \tag{1.3}$$

is known as the augmented Lagrangian function, $\mu^k = (\mu_1^k, \dots, \mu_m^k)^T$ is the vector of Lagrange multipliers for the equality constraints, ρ_k is the penalty parameter, and $\|\cdot\|$ denotes the Euclidean norm. In [5] the authors developed basic methods of analysis, proved the convergence results that cover also the possibility of solving inexactly the auxiliary problems (1.2), and established that a potentially troublesome penalty parameter is bounded. They implemented successfully their algorithm using the well-known package LANCELOT [6]. Later, Friedlander and Santos with the present author [15] proposed an adaptive precision control of the solution of the auxiliary problems (1.2) based on simple observation that the precision of the solution x^k of the auxiliary problems (1.2) should be related to the feasibility of x^k, i.e., $\|Cx^k\|$, since it does not seem reasonable to solve (1.2) to the high precision when μ^k is still far from the Lagrange multiplier corresponding to the solution of (1.1). Due to the choice introduced for the precision control, we obtained an estimate of the rate of convergence that does not have any term accounting for the inexact minimization. Moreover, it was proved that the penalty parameter generated by our algorithm remains bounded even with the inexact minimization of the auxiliary problems. It was also demonstrated that large penalty parameters need not slow down the rate of convergence in the inner loop [9]. The latter algorithm was then extensively used in development of parallel algorithms for the solution of variational inequalities [12, 16, 17]. For example, in combination with duality-based domain decomposition methods, the algorithm required as few as 65 conjugate gradient iterations and 3 outer iterations to solve an elliptic variational inequality discretized by 8,454,272 nodes with 2049 nodes on the contact interface [18], much cheaper then one direct solve of the related linear problem. Let us point out that the precision control that we use was proposed by Hager [25] for the solution of equality constrained problems by the augmented Lagrangian method with the least squares update of Lagrange multipliers.

Despite the improved results on the rate of convergence, the analysis in [15] indicates fast deterioration of the rate of convergence of the Lagrange multipliers with the diminishing smallest eigenvalue of $\widehat{C}\widehat{A}^{-1}\widehat{C}^T$, where $\widehat{C}$ is formed by the columns of C with indices of the constraints that are not active at the solution and $\widehat{A}$ is the corresponding diagonal block of A. This is not surprising, because when the matrix $\widehat{C}$ is not a full row rank matrix, then the multipliers of the solution are not uniquely determined and there is no reason to expect fast convergence when the rows of $\widehat{C}$ are nearly dependent. In fact, the analysis of [6, 15] requires that the solution is regular, i.e., that $\widehat{C}$ is a full column rank matrix. Moreover, the estimate assumes large values of the penalty parameter that are also related to the conditioning of $\widehat{C}\widehat{A}^{-1}\widehat{C}^T$. The results turned out to be insufficient to prove numerical scalability of our algorithms for solution of variational inequalities, despite the experimental evidence [17].

In this paper we show that it is possible to get results on convergence of the feasibility error that are independent of the conditioning of the equality constraints. Our main tool is a simple explicit bound on the penalty parameter [11] which guarantees an increase of the augmented Lagrangian provided the precision of the solution x^k of the auxiliary bound constrained problems is proportional to the feasibility error $\|Cx^k\|$. We exploit this observation to propose a new, less aggressive update rule for the penalty parameter which enables one to get an upper bound on the rate of convergence of $\|Cx^k\|$ that is independent of the eigenvalues of $\widehat{C}\widehat{A}^{-1}\widehat{C}^T$ as well as an explicit bound on the penalty parameter that is fully independent of the constraints data. When implemented with one of the algorithms mentioned above in the inner loop and applied to the class of problems with the spectrum of the Hessian matrix in a given interval of positive numbers, our new algorithm returns the solution of (1.1) in $O(1)$ matrix-vector multiplications. Our results on the rate of convergence of the algorithm remain valid even for the constraint matrix with dependent rows.

In section 2 we present the algorithm and prove that it is well defined. In section 3 we prove simple inequalities that will be exploited in the analysis of the algorithm in section 4. In section 5 we give new results on the convergence of the feasibility error that are independent of the form of the constraints and give a result on the "optimality" of the feasibility error estimate. The convergence results are presented in section 6. The results are illustrated on numerical solution of model problems in section 7. Finally, some conclusions are discussed in section 8.

2. Notation and preliminaries. Given nonempty sets of indices $\mathcal{I}, \mathcal{J} \subseteq \mathcal{N} \equiv \{1, \dots, n\}$, a matrix B, and a vector x, we define the submatrix $B_{\mathcal{I}\mathcal{J}}$ and the subvector $x_{\mathcal{I}}$ that comprise rows and columns determined by the sets $\mathcal{I}, \mathcal{J}$. For the matrix C of problem (1.1), we shall denote by $C_{\mathcal{J}}$ the submatrix of C formed by the same rows as C and columns in $\mathcal{J}$.

The first order update of the vector of Lagrange multipliers of the problem (1.1) and the gradients of the augmented Lagrangians (1.3) will be denoted, respectively, by

$$\widetilde{\mu} = \mu + \rho C x \tag{2.1}$$

and

$$g(x, \mu, \rho) = \nabla_x L(x, \mu, \rho) = \nabla q(x) + C^T \mu + \rho C^T C x. \tag{2.2}$$

For each vector $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$, we shall denote by

$$\|v\| = (v_1^2 + v_2^2 + \cdots + v_p^2)^{1/2} \quad \text{and} \quad \|\mathrm{v}\|_1 = |\mathrm{v}_1| + |\mathrm{v}_2| + \cdots + |\mathrm{v}_\mathrm{p}|$$

its Euclidean and ℓ_1-norms, respectively.

Since problem (1.1) comprises minimization of a strictly convex cost function on a convex set, it is well known (e.g., [3]) that its solution x^* exists and is necessarily unique. It satisfies the Karush–Kuhn–Tucker (KKT) conditions for problems that may be conveniently described by the *projected gradient* g^P that is defined by

$$\begin{array}{lcllll} g_i^P(x, \mu, \rho) & = & g_i(x, \mu, \rho) & \text{if} \;\; x_i > \ell_i \;\; \text{or} & x_i = \ell_i & \text{and} \;\; g_i(x, \mu, \rho) < 0, \\ g_i^P(x, \mu, \rho) & = & 0 & \text{otherwise, i.e.,} & x_i = \ell_i & \text{and} \;\; g_i(x, \mu, \rho) \geq 0, \end{array} \tag{2.3}$$

where $g(x, \mu, \rho) = (g_1, \ldots, g_n)^T$. Thus the solution x^* is the only feasible (for problem (1.1)) vector for which there is μ such that $g^P(x, \mu, 0) = 0$, and the KKT conditions for problem (1.2) are satisfied at $x \geq \ell$ if and only if there is μ such that $g^P(x, \mu, \rho) = 0$.

For each vector $x \in \mathbb{R}^n$, we shall denote by $\mathcal{A}(x)$ and $\mathcal{F}(x)$ the *active* and *free* set of x, respectively, so that

$$\mathcal{A}(x) = \{i \in \mathcal{N} : x_i = \ell_i\} \;\; \text{and} \;\; \mathcal{F}(\mathrm{x}) = \{\mathrm{i} \in \mathcal{N} : \mathrm{x}_\mathrm{i} \neq \ell_\mathrm{i}\}.$$

We shall also define the *binding* set

$$\mathcal{B}(x, \mu, \rho) = \{i \in \mathcal{A}(x) : g_i(x, \mu, \rho) \geq 0\}.$$

Thus $\mathcal{B}(x, \mu, \rho)$ comprises the indices of the active constraints that satisfy the KKT conditions for problem (1.2).

As mentioned, the earlier papers required that the solution $x^* \in \Omega$ of problem (1.1) is *regular*, i.e., that the gradients of all the active constraints (equalities and inequalities) at x^* are linearly independent. We will not need this assumption here, although we will be able to prove some stronger results if the solution is regular.

3. Semimonotonic algorithm for bound and equality constraints.

The following algorithm is a modification of the classical augmented Lagrangian method for the solution of strictly convex quadratic programming problems with equality constraints that enables adaptive precision control of the solution of auxiliary problems.

ALGORITHM 3.1. Given $\eta > 0,\; \beta > 1,\; M > 0,\; \rho_0 > 0,$ and $\mu^0 \in \mathbb{R}^m$, set $k = 0$.

Step 1. {Inner iteration with adaptive precision control.}

Find x^k such that

$$\|g^P(x^k, \mu^k, \rho_k)\| \leq \min\{M\|Cx^k\|, \eta\}. \tag{3.1}$$

Step 2. {Update μ.}

$$\mu^{k+1} = \mu^k + \rho_k C x^k. \tag{3.2}$$

Step 3. {Update ρ provided the increase of the Lagrangian is not sufficient.}
If $k > 0$ and

$$L(x^k, \mu^k, \rho^k) < L(x^{k-1}, \mu^{k-1}, \rho_{k-1}) + \frac{\rho_k}{2}\|Cx^k\|^2, \tag{3.3}$$

then

$$\rho_{k+1} = \beta\rho_k; \tag{3.4}$$

else

$$\rho_{k+1} = \rho_k. \tag{3.5}$$

Step 4. Set $k = k + 1$ and return to Step 1.

In Step 1 we can use any convergent algorithm for minimizing the strictly convex quadratic function subject to the bound constraints, such as [23, 4, 10, 21, 8]. Let us point out that Algorithm 3.1 differs from those considered by Conn, Gould, and Toint [5] or Dostál, Friedlander, and Santos [15] by the condition on the update of the penalization parameter in Step 3.

The next lemma shows that Algorithm 3.1 is well defined, that is, any convergent algorithm for the solution of the auxiliary problem required in Step 1 will generate either x^k that satisfies (3.1) in a finite number of steps or a sequence of approximations that converges to the solution of (1.1). It is also clear that there is no hidden enforcement of exact solution in (3.1) and consequently typically inexact solutions of the auxiliary unconstrained problems are obtained in Step 1.

Lemma 3.1. *Let $M > 0, \mu \in \mathbb{R}^m$, and $\rho \geq 0$ be given and let $\{y^k\}$ denote any sequence that converges to the unique solution $\overline{y}$ of the problem*

$$\textit{minimize} \quad L(y, \mu, \rho) \quad \textit{s.t.} \quad y \geq \ell. \tag{3.6}$$

Then either $\{y^k\}$ converges to the solution $\widehat{x}$ of problem (1.1) *or there is an index k such that*

$$\|g^P(y^k, \mu, \rho)\| \leq \min\{M\|Cy^k\|, \eta\}. \tag{3.7}$$

Proof. See [15]. □

4. Inequalities involving the augmented Lagrangian. In this section we shall establish basic inequalities that relate the bound on the norm of the gradient g of the augmented Lagrangian L to the values of the augmented Lagrangian L. These inequalities will be the key ingredients in the proof of convergence of Algorithm 3.1.

Lemma 4.1. *Let $x, y, \ell \in \mathbb{R}^n,\ x \geq \ell,\ y \geq \ell,\ \mu \in \mathbb{R}^m,\ \rho > 0,\ \eta > 0,$ and $M > 0$. Let α denote the least eigenvalue of A and $\widetilde{\mu} = \mu + \rho Cx$.*

(i) *If*

$$\|g^P(x, \mu, \rho)\| \leq M\|Cx\|, \tag{4.1}$$

then

$$L(y, \widetilde{\mu}, \rho) \geq L(x, \mu, \rho) + \frac{1}{2}\left(\rho - \frac{M^2}{\alpha}\right)\|Cx\|^2 + \frac{\rho}{2}\|Cy\|^2. \tag{4.2}$$

(ii) *If*

$$\|g^P(x, \mu, \rho)\| \leq \eta, \tag{4.3}$$

then

$$L(y,\widetilde{\mu},\rho) \geq L(x,\mu,\rho) + \frac{\rho}{2}\|Cx\|^2 + \frac{\rho}{2}\|Cy\|^2 - \frac{\eta^2}{2\alpha}. \tag{4.4}$$

(iii) *If* (4.3) *holds and* $z_0 \in \Omega$, *then*

$$L(x,\mu,\rho) \leq q(z_0) + \frac{\eta^2}{2\alpha}. \tag{4.5}$$

Proof. Let us denote $\delta = y - x$, $H = A + \rho C^T C$ and recall that we assume $x \geq \ell$ and $y \geq \ell$, so that it may be easily verified that

$$\delta^T g^P(x,\mu,\rho) \leq \delta^T g(x,\mu,\rho).$$

Using the latter inequality with

$$L(x,\widetilde{\mu},\rho) = L(x,\mu,\rho) + \rho\|Cx\|^2 \quad \text{and} \quad g(x,\widetilde{\mu},\rho) = g(x,\mu,\rho) + \rho C^T Cx,$$

we get

$$\begin{aligned}
L(y,\widetilde{\mu},\rho) &= L(x,\widetilde{\mu},\rho) + \delta^T g(x,\widetilde{\mu},\rho) + \frac{1}{2}\delta^T H\delta \\
&= L(x,\mu,\rho) + \delta^T g(x,\mu,\rho) + \rho\delta^T C^T Cx + \frac{1}{2}\delta^T H\delta + \rho\|Cx\|^2 \\
&\geq L(x,\mu,\rho) + \delta^T g^P(x,\mu,\rho) + \rho\delta^T C^T Cx + \frac{1}{2}\delta^T H\delta + \rho\|Cx\|^2 \\
&\geq L(x,\mu,\rho) + \delta^T g^P(x,\mu,\rho) + \rho\delta^T C^T Cx + \frac{\alpha}{2}\|\delta\|^2 + \frac{\rho}{2}\|C\delta\|^2 + \rho\|Cx\|^2.
\end{aligned}$$

Noticing that

$$\frac{\rho}{2}\|Cy\|^2 = \frac{\rho}{2}\|C(\delta + x)\|^2 = \rho\delta^T C^T Cx + \frac{\rho}{2}\|C\delta\|^2 + \frac{\rho}{2}\|Cx\|^2,$$

we obtain

$$L(y,\widetilde{\mu},\rho) \geq L(x,\mu,\rho) + \delta^T g^P(x,\mu,\rho) + \frac{\alpha}{2}\|\delta\|^2 + \frac{\rho}{2}\|Cx\|^2 + \frac{\rho}{2}\|Cy\|^2. \tag{4.6}$$

Using (4.1) and simple manipulations we obtain

$$\begin{aligned}
L(y,\widetilde{\mu},\rho) &\geq L(x,\mu,\rho) - M\|\delta\|\|Cx\| + \frac{\alpha}{2}\|\delta\|^2 + \frac{\rho}{2}\|Cx\|^2 + \frac{\rho}{2}\|Cy\|^2 \\
&= L(x,\mu,\rho) + \left(\frac{\alpha}{2}\|\delta\|^2 - M\|\delta\|\|Cx\| + \frac{M^2\|Cx\|^2}{2\alpha}\right) \\
&\quad - \frac{M^2\|Cx\|^2}{2\alpha} + \frac{\rho}{2}\|Cx\|^2 + \frac{\rho}{2}\|Cy\|^2 \\
&\geq L(x,\mu,\rho) + \frac{1}{2}\left(\rho - \frac{M^2}{\alpha}\right)\|Cx\|^2 + \frac{\rho}{2}\|Cy\|^2,
\end{aligned}$$

which proves (i).

If we assume that (4.3) holds, then by (4.6)

$$\begin{aligned}
L(y,\widetilde{\mu},\rho) &\geq L(x,\mu,\rho) - \|\delta\|\eta + \frac{\alpha}{2}\|\delta\|^2 + \frac{\rho}{2}\|Cx\|^2 + \frac{\rho}{2}\|Cy\|^2 \\
&\geq L(x,\mu,\rho) + \frac{\rho}{2}\|Cx\|^2 + \frac{\rho}{2}\|Cy\|^2 - \frac{\eta^2}{2\alpha},
\end{aligned}$$

which proves (ii).

Finally, let $\overline{x}$ denote the solution of the auxiliary problem

$$\text{minimize } \ L(z,\mu,\rho) \ \text{ s.t. } \ z \geq \ell, \tag{4.7}$$

let $z_0 \in \Omega$ so that $Cz_0 = 0$, and let $\overline{\delta} = \overline{x} - x$. If (4.3) holds, then

$$0 \geq L(\overline{x},\mu,\rho) - L(x,\mu,\rho) = \overline{\delta}^T g(x,\mu,\rho) + \frac{1}{2}\overline{\delta}^T H\overline{\delta} \geq \overline{\delta}^T g^P(x,\mu,\rho) + \frac{1}{2}\overline{\delta}^T H\overline{\delta}$$
$$\geq -\|\overline{\delta}\|\eta + \frac{1}{2}\alpha\|\overline{\delta}\|^2 \geq -\frac{\eta^2}{2\alpha}.$$

Since $L(\overline{x},\mu,\rho) \leq L(z_0,\mu,\rho) = q(z_0)$, we conclude that

$$L(x,\mu,\rho) \leq L(x,\mu,\rho) - L(\overline{x},\mu,\rho) + q(z_0) \leq q(z_0) + \frac{\eta^2}{2\alpha}. \qquad \square$$

5. Monotonicity and feasibility.

LEMMA 5.1. *Let $\{x^k\}, \{\mu^k\}$, and $\{\rho_k\}$ be generated by Algorithm* 3.1 *with $\eta > 0$, $\beta > 1$, $M > 0$, $\rho_0 > 0$, and $\mu^0 \in \mathbb{R}^m$. Let α denote the least eigenvalue of the Hessian A of the quadratic q.*

(i) *If $k \geq 0$ and*

$$\rho_k \geq M^2/\alpha, \tag{5.1}$$

then

$$L(x^{k+1},\mu^{k+1},\rho_{k+1}) \geq L(x^k,\mu^k,\rho_k) + \frac{\rho_{k+1}}{2}\|Cx^{k+1}\|^2. \tag{5.2}$$

(ii) *For any $k \geq 0$*

$$L(x^{k+1},\mu^{k+1},\rho_{k+1}) \geq L(x^k,\mu^k,\rho_k) + \frac{\rho_k}{2}\|Cx^k\|^2 + \frac{\rho_{k+1}}{2}\|Cx^{k+1}\|^2 - \frac{\eta^2}{2\alpha}. \tag{5.3}$$

(iii) *For any $k \geq 0$ and $z_0 \in \Omega$*

$$L(x^k,\mu^k,\rho_k) \leq q(z_0) + \frac{\eta^2}{2\alpha}. \tag{5.4}$$

Proof. Let us substitute in Lemma 4.1 $x = x^k, \mu = \mu^k, \rho = \rho_k, y = x^{k+1}$, so that by (3.1) the inequality (4.1) holds and by (3.2) $\widetilde{\mu} = \mu^{k+1}$.

If (5.1) holds, we shall get by (4.2)

$$L(x^{k+1},\mu^{k+1},\rho_k) \geq L(x^k,\mu^k,\rho_k) + \frac{\rho_k}{2}\|Cx^{k+1}\|^2. \tag{5.5}$$

To prove (5.2), it is enough to add

$$\frac{\rho_{k+1}-\rho_k}{2}\|Cx^{k+1}\|^2 \tag{5.6}$$

to both sides of (5.5) and to use that

$$L(x^{k+1},\mu^{k+1},\rho_{k+1}) = L(x^{k+1},\mu^{k+1},\rho_k) + \frac{\rho_{k+1}-\rho_k}{2}\|Cx^{k+1}\|^2. \tag{5.7}$$

If we notice that, by the definition of Step 1 of Algorithm 3.1,

$$\|g(x^k, \mu^k, \rho_k)\| \leq \eta$$

and apply the same substitution as above to Lemma 3.1(ii), we shall get

$$L(x^{k+1}, \mu^{k+1}, \rho_k) \geq L(x^k, \mu^k, \rho_k) + \frac{\rho_k}{2}\|Cx^k\|^2 + \frac{\rho_k}{2}\|Cx^{k+1}\|^2 - \frac{\eta^2}{2\alpha}, \tag{5.8}$$

so that, after adding the nonnegative expression (5.6) to both sides of (5.8) and using (5.7), we get (5.3). Similarly, the definition of Algorithm 2.1 and application of the substitution to Lemma 4.1(iii) implies the inequality (5.4). □

THEOREM 5.2. *Let $\{x^k\}, \{\mu^k\}$, and $\{\rho_k\}$ be generated by Algorithm* 3.1 *with $\eta > 0,\ \beta > 1,\ M > 0,\ \rho_0 > 0,\ \ \mu^0 \in \mathbb{R}^m$. Let α denote the least eigenvalue of the Hessian A of the quadratic q and let $s \geq 0$ denote the smallest integer such that $\beta^s \rho_0 \geq M^2/\alpha$.*

(i) *The sequence $\{\rho_k\}$ is bounded and*

$$\rho_k \leq \beta^s \rho_0. \tag{5.9}$$

(ii) *If $z_0 \in \Omega$, then*

$$\sum_{k=1}^{\infty} \frac{\rho_k}{2}\|Cx^k\|^2 \leq q(z_0) - L(x^0, \mu^0, \rho_0) + (1+s)\frac{\eta^2}{2\alpha}. \tag{5.10}$$

(iii) *$\|Cx^k\|$ converges to* 0.

Proof. Let $s \geq 0$ denote the smallest integer such that $\beta^s \rho_0 \geq M^2/\alpha$ and let $\mathcal{I}$ denote the set of all indices k_i such that $\{\rho_{k_i}\}$ are generated in Step 3 of Algorithm 3.1 by (3.4). Using Lemma 5.1(i), $\rho_{k_i} = \beta\rho_{k_{i-1}} = \beta^i \rho_0$ for $k_i \in \mathcal{I}$, and $\beta^s \rho_0 \geq M^2/\alpha$, we conclude that there is no k such that $\rho_k > \beta^s \rho_0$. Thus $\mathcal{I}$ has at most s elements and (5.9) holds.

By the definition of Step 3, for $k > 0$ either $k+1 \notin \mathcal{I}$ and

$$\frac{\rho_k}{2}\|Cx^k\|^2 \leq L(x^k, \mu^k, \rho_k) - L(x^{k-1}, \mu^{k-1}, \rho_{k-1})$$

or $k+1 \in \mathcal{I}$ and by (5.3)

$$\begin{aligned}\frac{\rho_k}{2}\|Cx^k\|^2 &\leq \frac{\rho_{k-1}}{2}\|Cx^{k-1}\|^2 + \frac{\rho_k}{2}\|Cx^k\|^2 \\ &\leq L(x^k, \mu^k, \rho_k) - L(x^{k-1}, \mu^{k-1}, \rho_{k-1}) + \frac{\eta^2}{2\alpha}.\end{aligned}$$

Summing up appropriate cases of the last two inequalities for $k = 1, \ldots, n$ and taking into account that $\mathcal{I}$ has at most s elements, we get

$$\sum_{k=1}^{n} \frac{\rho_k}{2}\|Cx^k\|^2 \leq L(x^n, \mu^n, \rho_n) - L(x^0, \mu^0, \rho_0) + s\frac{\eta^2}{2\alpha}. \tag{5.11}$$

To get (5.10), it is enough to replace $L(x^n, \mu^n, \rho_n)$ by the upper bound (5.4).

Statement (iii) is an immediate consequence of (ii). □

Theorem 5.2 suggests that it is possible to give a uniform upper bound on the number of the outer iterations of Algorithm 3.1 that is necessary to achieve a prescribed feasibility error for any problem of (5.12). To present explicitly this qualitatively new feature of Algorithm 3.1, at least as compared to the related algorithms [15], let $\mathcal{T}$ denote any set of indices and let for any $t \in \mathcal{T}$ be defined a problem

$$\text{minimize } \; q_t(x) \;\text{ s.t. }\; x \in \Omega_t \tag{5.12}$$

with $\Omega_t = \{x \in \mathbb{R}^{n_t} : C_t x = 0 \;\text{ and }\; x \geq \ell_t\}$, $q_t(x) = \frac{1}{2}x^T A_t x - b_t^T x$, $A_t \in \mathbb{R}^{n_t \times n_t}$ symmetric positive definite, $C_t \in \mathbb{R}^{m_t \times n_t}$, and $b_t, \ell_t \in \mathbb{R}^{n_t}$. To simplify our exposition, we shall assume that the bound constraints ℓ_t are not positive so that $0 \in \Omega_t$. Our optimality result reads as follows.

Theorem 5.3. *Let $\{x_t^k\}, \{\mu_t^k\}$, and $\{\rho_{t,k}\}$ be generated by Algorithm 3.1 for (5.12) with $\|b_t\| \geq \eta_t > 0$, $\beta > 1$, $M > 0$, $\rho_{t,0} = \rho_0 > 0$, $\mu_t^0 = 0$. Let there be an $\alpha > 0$ such that the least eigenvalue α_t of the Hessian A_t of the quadratics q_t satisfies $\alpha_t \geq \alpha$, and let $s \geq 0$ denote the smallest integer such that $\beta^s \rho_0 \geq M^2/\alpha$. Then for each $\epsilon > 0$ and*

$$j \geq \frac{2+s}{\epsilon^2 \alpha \rho_0} \tag{5.13}$$

there are indices k_t such that

$$k_t \leq j \;\; \text{and} \;\; M^{-1}\|g^P(x_t^{k_t}, \mu_t^{k_t}, \rho_{t,k_t})\| \leq \|C_t x_t^{k_t}\| \leq \epsilon \|b_t\|. \tag{5.14}$$

Proof. First notice that for any index k

$$\frac{\rho_0 k}{2} \min\{\|C_t x_t^i\|^2 : i = 1, \dots, k\} \leq \sum_{i=1}^{k} \frac{\rho_{t,i}}{2}\|C_t x_t^i\|^2 \leq \sum_{i=1}^{\infty} \frac{\rho_{t,i}}{2}\|C_t x_t^i\|^2. \tag{5.15}$$

Denoting by $L_t(x, \mu, \rho)$ the augmented Lagrangian for the problem (5.12), we get for any $x \in \mathbb{R}^p$ and $\rho \geq 0$

$$L_t(x, 0, \rho) = \frac{1}{2}x^T(A_t + \rho C_t^T C_t)x - b_t^T x \geq \frac{1}{2}\alpha\|x\|^2 - \|b_t\|\|x\| \geq -\frac{\|b_t\|^2}{2\alpha}.$$

If we substitute this inequality and $z_0 = 0$ to (5.10) and use the assumption $\|b_t\| \geq \eta_t$, we get

$$\sum_{i=1}^{\infty} \frac{\rho_i}{2}\|C_t x_t^i\|^2 \leq \frac{\|b_t\|^2}{2\alpha} + (1+s)\frac{\eta^2}{2\alpha} \leq \frac{(2+s)\|b_t\|^2}{2\alpha}. \tag{5.16}$$

Taking for j any integer that satisfies (5.13) and denoting for any $t \in \mathcal{T}$ by $k_t \in \{1, \dots, j\}$ the index which minimizes $\{\|C_t x_t^i\| : i = 1, \dots, j\}$, we can use (5.15) and (5.16) with simple manipulations to obtain

$$\|C_t x_t^{k_t}\|^2 = \min\{\|C_t x_t^i\|^2 : i = 1, \dots, k\} \leq \frac{(2+s)\|b_t\|^2}{j\alpha\rho_0} \leq \epsilon^2\|b_t\|^2.$$

The inequality

$$M^{-1}\|g^P(x_t^{k_t}, \mu_t^{k_t}, \rho_{t,k_t})\| \leq \|C_t x_t^{k_t}\|$$

results easily from the definition of Step 1 of Algorithm 3.1. □

6. Convergence.

Lemma 6.1. *Let $\{x^k\}, \{\mu^k\}$, and $\{\rho_k\}$ be generated by Algorithm* 3.1 *with $\eta > 0,\ \beta > 1,\ M > 0,\ \rho_0 > 0,\ \ \mu^0 \in \mathbb{R}^m$. Then the sequence $\{x^k\}$ is bounded.*

Proof. Since there is only a finite number of different subsets $\mathcal{F}$ of $\mathcal{N}$ and $\{x^k\}$ is bounded if and only if $\{x^k_{\mathcal{F}(x^k)}\}$ is bounded, we can restrict our attention to analysis of infinite subsequences $\{x^k_{\mathcal{F}} : \mathcal{F}(x^k) = \mathcal{F}\}$ that are defined by nonempty subsets $\mathcal{F}$ of $\mathcal{N}$.

Let $\mathcal{F} \subseteq \mathcal{N},\ \mathcal{F} \neq \emptyset$ be such that $\{x^k : \mathcal{F}(x^k) = \mathcal{F}\}$ is infinite and denote $\mathcal{A} = \mathcal{N} \setminus \mathcal{F}$. Using Theorem 5.2(i), we get that there is an integer k_0 such that $\rho_k = \rho_{k_0}$ for $k \geq k_0$. Thus, for $k \geq k_0$, we can denote $H = A + \rho_k C^T C$ and

$$g^k = g(x^k, \mu^k, \rho_k) = Hx^k + C^T\mu^k - b,$$

so that

$$\begin{pmatrix} H_{\mathcal{F}\mathcal{F}} & C_{\mathcal{F}}^T \\ C_{\mathcal{F}} & 0 \end{pmatrix} \begin{pmatrix} x^k_{\mathcal{F}} \\ \mu^k \end{pmatrix} = \begin{pmatrix} g^k_{\mathcal{F}} + b_{\mathcal{F}} - H_{\mathcal{F}\mathcal{A}}\ell_{\mathcal{A}} \\ C_{\mathcal{F}}x^k_{\mathcal{F}} \end{pmatrix}. \tag{6.1}$$

Since $C_{\mathcal{F}}x^k_{\mathcal{F}} = Cx^k - C_{\mathcal{A}}\ell_{\mathcal{A}},\ \|g^k_{\mathcal{F}}\| = \|g_{\mathcal{F}}(x^k, \mu^k, \rho_k)\| \leq \|g^P(x^k\mu^k, \rho_k)\|$, and both $\|g^P(x^k, \mu^k, \rho_k)\|$ and $\|Cx^k\|$ converge to zero, the right-hand side of (6.1) is bounded. Thus both x^k and μ^k are bounded provided the matrix of the system (6.1) is regular. This happens when $C_{\mathcal{F}}$ is a full row rank matrix (e.g., [13]).

If $C_{\mathcal{F}} \in \mathbb{R}^{ps}$ is not a full row rank matrix, then its rank r satisfies $r < p$ and by the singular value decomposition theorem (see Theorem 7.3.5 in [27]) there are orthogonal matrices $U = (u_1, \dots, u_p)^T$, $V = (v_1, \dots, v_s)^T$ and a matrix $\Sigma \in \mathbb{R}^{ps}$ defined by a diagonal matrix $D = \operatorname{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0) \in \mathbb{R}^{t \times t},\ t = \min\{p, s\}$ padded with zeros so that $C_{\mathcal{F}} = U^T\Sigma V$. Thus, taking $\widehat{U} = (u_1, \dots, u_r)^T$, $\widehat{D} = \operatorname{diag}(\sigma_1, \dots, \sigma_r)$ and $\widehat{V} = (v_1, \dots, v_r)^T$, we have $C_{\mathcal{F}} = \widehat{U}^T\widehat{D}\widehat{V}$ and we can define a full row rank matrix

$$\widehat{C}_{\mathcal{F}} = \ \widehat{D}\widehat{V} = \widehat{U}C_{\mathcal{F}}$$

that satisfies $\widehat{C}_{\mathcal{F}}^T\widehat{C}_{\mathcal{F}} = C_{\mathcal{F}}^TC_{\mathcal{F}}$ and $\|\widehat{C}_{\mathcal{F}}x_{\mathcal{F}}\| = \|C_{\mathcal{F}}x_{\mathcal{F}}\|$ for any vector x. We shall assign to any $\mu \in \mathbb{R}^m$ the vector

$$\widehat{\mu} = \widehat{U}\mu$$

so that $\widehat{C}_{\mathcal{F}}^T\widehat{\mu} = C_{\mathcal{F}}^T\mu$. Substituting the latter identity to (6.1), we get

$$\begin{pmatrix} H_{\mathcal{F}\mathcal{F}} & \widehat{C}_{\mathcal{F}}^T \\ C_{\mathcal{F}} & 0 \end{pmatrix} \begin{pmatrix} x^k_{\mathcal{F}} \\ \widehat{\mu}^k \end{pmatrix} = \begin{pmatrix} g^k_{\mathcal{F}} + b_{\mathcal{F}} - H_{\mathcal{F}\mathcal{A}}\ell_{\mathcal{A}} \\ C_{\mathcal{F}}x^k_{\mathcal{F}} \end{pmatrix}. \tag{6.2}$$

Since $C_{\mathcal{F}} = \widehat{U}^T\widehat{D}\widehat{V} = \widehat{U}^T\widehat{C}_{\mathcal{F}}$ and $\widehat{U}^T$ is a full column rank matrix, the latter system is equivalent to the system

$$\begin{pmatrix} H_{\mathcal{F}\mathcal{F}} & \widehat{C}_{\mathcal{F}}^T \\ \widehat{C}_{\mathcal{F}} & 0 \end{pmatrix} \begin{pmatrix} x^k_{\mathcal{F}} \\ \widehat{\mu}^k \end{pmatrix} = \begin{pmatrix} g^k_{\mathcal{F}} + b_{\mathcal{F}} - H_{\mathcal{F}\mathcal{A}}\ell_{\mathcal{A}} \\ \widehat{C}_{\mathcal{F}}x^k_{\mathcal{F}} \end{pmatrix} \tag{6.3}$$

with a regular matrix. The right-hand side of (6.3) being bounded due to $\|\widehat{C}_{\mathcal{F}}x^k_{\mathcal{F}}\| = \|C_{\mathcal{F}}x^k_{\mathcal{F}}\|$, we conclude that $\{x^k_{\mathcal{F}} : \mathcal{F}(x^k) = \mathcal{F}\}$ is bounded. □

Lemma 6.2. *Let $\{z^k\}$ denote a bounded sequence, let B denote a full column rank matrix, and let there be a sequence $\{\tau^k\}$ such that $B\tau^k \geq z^k$. Then there is a bounded sequence $\{\widehat{\tau}^k\}$ such that $B\widehat{\tau}^k \geq z^k$.*

Proof. Let us denote $e = (1, 1, \dots, 1)^T$ and consider for a given integer k a linear programming problem of the form

$$\min\{e^T B\xi : B\xi \geq z^k\} \tag{6.4}$$

with B and z_k of the lemma. Since τ^k satisfies $B\tau^k \geq z^k$, it follows that the problem (6.4) is feasible. Moreover, observing that for any feasible ξ

$$e^T B\xi = e^T(B\xi - z^k) + e^T z^k \geq e^T z^k,$$

we conclude that the problem (6.4) is also bounded from below, so that it has a solution ξ^k. Using the well known duality theory of linear programming (e.g., [3]), it follows that the dual problem

$$\max\{\eta^T z^k : \eta \geq 0 \ \text{ and } \ \mathrm{B^T}\eta = \mathrm{e}\} \tag{6.5}$$

is feasible and bounded from above, so that it attains its solution η^k at a vertex of the convex boundary of the feasible set of the dual problem (6.5) and

$$(\eta^k)^T z^k = e^T B\xi^k.$$

Since the number of the vertices is finite, it follows that there is only a finite number of different η^k, so that, as $\{z^k\}$ is bounded, there is a constant c such that $e^T B\xi^k = (\eta^k)^T z^k \leq c$ for any integer k. Thus

$$\begin{aligned}\|B\xi^k\|_1 &\leq \|B\xi^k - z^k\|_1 + \|z^k\|_1 = e^T(B\xi^k - z^k) + \|z^k\|_1\\ &\leq e^T B\xi^k + 2\|z^k\|_1 \leq c + 2\|z^k\|_1.\end{aligned}$$

Since $\{z^k\}$ is bounded and B is a full column rank matrix, also the vectors ξ^k are bounded and $\widehat{\tau}^k = \xi^k$ satisfies the statement of the lemma. □

LEMMA 6.3. *Let $\{x^k\}, \{\mu^k\}$, and $\{\rho_k\}$ be generated by Algorithm* 3.1 *with $\eta > 0,\ \beta > 1,\ M > 0,\ \rho_0 > 0,\ \ \mu^0 \in \mathbb{R}^m$. Then there is a bounded sequence $\widehat{\mu}^k$ such that*

$$g^P(x^k, \widehat{\mu}^k, \rho_k) = g^P(x^k, \mu^k, \rho_k). \tag{6.6}$$

Proof. Let $\mathcal{B} \subset \mathcal{N},\ \mathcal{B} \neq \emptyset,\ \mathcal{B} \neq \mathcal{N}$ be such that $\{x^k : \mathcal{B}(x^k, \mu^k, \rho^k) = \mathcal{B}\}$ is infinite and denote $\mathcal{C} = \mathcal{N} \setminus \mathcal{B}$. Using a variant of the Gramm–Schmidt orthogonalization process, we can find a regular matrices R such that

$$\begin{pmatrix} C_{\mathcal{C}}^T \\ C_{\mathcal{B}}^T \end{pmatrix} R = \begin{pmatrix} P & 0 & 0 \\ Q & T & 0 \end{pmatrix},$$

where P and T are full column rank matrices. Thus decomposing properly $R^{-1}\mu^k$ into the blocks $R^{-1}\mu^k = (\xi^k, \tau^k, \nu^k)^T$, we get

$$\begin{pmatrix} C_{\mathcal{C}}^T \\ C_{\mathcal{B}}^T \end{pmatrix} \mu^k = \begin{pmatrix} P & 0 & 0 \\ Q & T & 0 \end{pmatrix} \begin{pmatrix} \xi^k \\ \tau^k \\ \nu^k \end{pmatrix} = \begin{pmatrix} P & 0 & 0 \\ Q & T & 0 \end{pmatrix} \begin{pmatrix} \xi^k \\ \tau^k \\ 0 \end{pmatrix}. \tag{6.7}$$

Using Theorem 5.2(i), we get that there is an integer k_0 such that $\rho_k = \rho_{k_0}$ for $k \geq k_0$. Let us denote $H = A + \rho_{k_0} C^T C$ and $g^k = g(x^k, \mu^k, \rho_k)$, so that for $k \geq k_0$

$$C^T \mu^k = b + g^k - Hx^k. \tag{6.8}$$

Substituting into (6.7), we get for $k \geq k_0$

$$C_{\mathcal{C}}^T \mu^k = P\xi^k = b_{\mathcal{C}} + g_{\mathcal{C}}^k - H_{\mathcal{C}\mathcal{N}} x^k.$$

Since P is full column rank matrix, $\|g_{\mathcal{C}}^k\| = \|g^P(x^k, \mu^k, \rho_k)\|$ and both x^k and $g^P(x^k, \mu^k, \rho_k)$ are bounded, it follows that ξ^k is bounded. Similarly, for $k \geq k_0$ and $\mathcal{B} = \mathcal{B}(x^k, \mu^k, \rho^k)$

$$C_{\mathcal{B}}^T \mu^k = Q\xi^k + T\tau^k = b_{\mathcal{B}} + g_{\mathcal{B}}^k - H_{\mathcal{B}\mathcal{N}} x^k \geq b_{\mathcal{B}} - H_{\mathcal{B}\mathcal{N}} x^k.$$

The vectors x^k being bounded due to Lemma 6.1, we can apply Lemma 6.2 to get bounded sequence $\widehat{\tau}^k$ such that

$$T\widehat{\tau}^k \geq b_{\mathcal{B}} - H_{\mathcal{B}\mathcal{N}} x^k - Q\xi^k. \tag{6.9}$$

Let us now define for $k \geq k_0$ a bounded sequence

$$\widehat{\mu}^k = R \begin{pmatrix} \xi^k \\ \widehat{\tau}^k \\ 0 \end{pmatrix}$$

so that by (6.7)

$$C_{\mathcal{C}}^T \widehat{\mu}^k = C_{\mathcal{C}}^T R \begin{pmatrix} \xi^k \\ \widehat{\tau}^k \\ 0 \end{pmatrix} = P\xi^k = C_{\mathcal{C}}^T \mu^k$$

and

$$g_{\mathcal{C}}^k = g_{\mathcal{C}}(x^k, \mu^k, \rho_k) = g_{\mathcal{C}}^P(x^k, \mu^k, \rho_k). \tag{6.10}$$

Similarly,

$$C_{\mathcal{B}}^T \widehat{\mu}^k = C_{\mathcal{B}}^T R \begin{pmatrix} \xi^k \\ \widehat{\tau}^k \\ 0 \end{pmatrix} = Q\xi^k + T\widehat{\tau}^k \tag{6.11}$$

and by (6.9)

$$g_{\mathcal{B}}(x^k, \widehat{\mu}^k, \rho_k) = H_{\mathcal{B}\mathcal{N}} x^k - b_{\mathcal{B}} + C_{\mathcal{B}}^T \widehat{\mu}^k = H_{\mathcal{B}\mathcal{N}} x^k - b_{\mathcal{B}} + Q\xi^k + T\widehat{\tau}^k \geq 0.$$

Recalling that we assume that $\mathcal{B}(x^k, \mu^k, \rho^k) = \mathcal{B}$, the last equation together with (6.10) yields

$$g^P(x^k, \mu^k, \rho_k) = g^P(x^k, \widehat{\mu}^k, \rho_k).$$

If $\mathcal{B} = \emptyset$ or $\mathcal{B} = \mathcal{N}$ are such that $\{x^k : \mathcal{B}(x^k, \mu^k, \rho^k) = \mathcal{B}\}$ is infinite, we can find the multipliers $\widehat{\mu}^k$ that satisfy the statement of our lemma by specializing the above arguments. Since there is only a finite number of different subsets $\mathcal{B}$ of $\mathcal{N}$, we have shown that there are the multipliers $\widehat{\mu}^k$ that satisfy the statement of our lemma for all k except possibly finite number of indices for which we shall define $\widehat{\mu}^k = \mu^k$. This completes the proof. □

THEOREM 6.4. *Let* $\{x^k\}, \{\mu^k\}$, *and* $\{\rho_k\}$ *be generated by Algorithm* 3.1 *with* $\eta > 0,\ \beta > 1,\ M > 0,\ \rho_0 > 0,\ \ \mu^0 \in \mathbb{R}^m$.

(i) *The sequence* $\{x^k\}$ *converges to the solution* x^* *of* (1.1).

(ii) *If the solution* x^* *of* (1.1) *is regular, then* $\{x^k\}$ *and* $\{\mu^k\}$ *converge to the solution* x^* *and the vector* μ^* *of the Lagrange multipliers of* (1.1), *respectively.*

Proof. Let $\widehat{\mu}^k$ denote the sequence of Lemma 6.3 so that it satisfies $g^P(x^k, \mu^k, \rho_k) = g^P(x^k, \widehat{\mu}^k, \rho_k)$. Since both x^k and $\widehat{\mu}^k$ are bounded, it follows that there is a cluster point $(\bar{x}, \bar{\mu})$ of the sequence $(x^k, \widehat{\mu}^k)$. Using Theorem 5.2(i), we get that there is k_0 such that $\rho_k = \rho_{k_0}$ for $k \geq k_0$. Moreover, by Theorem 5.2(iii) and the definition of Step 1 of Algorithm 3.1, $C\bar{x} = 0$ and $g^P(\bar{x}, \bar{\mu}, \rho_{k_0}) = g^P(\bar{x}, \bar{\mu}, 0) = 0$. Since $\bar{x} \geq 0$, $\bar{x}$ is the solution of (1.1). The solution x^* of (1.1) being unique, it follows that x^k converges to $\bar{x} = x^*$.

Let k_0 be as above and assume that the solution x^* of (1.1) is regular. Since we have just proved that $\{x^k\}$ converges to x^*, it follows that there is $k_1 \geq k_0$ such that $\{x^k\}$ is regular for $k \geq k_1$. Denoting $\mathcal{F} = \mathcal{F}(x^*)$ and $H = A + \rho_{k_0} C_{\mathcal{F}}^T C_{\mathcal{F}}$, it follows that

$$g_{\mathcal{F}}^P(x^k, \mu^k, \rho_k) = H_{\mathcal{F}\mathcal{N}} x^k - b_{\mathcal{F}} + C_{\mathcal{F}}^T \mu^k$$

converges to zero, so that the sequence

$$C_{\mathcal{F}}^T \mu^k = b_{\mathcal{F}} - H_{\mathcal{F}\mathcal{N}} x^k + g_{\mathcal{F}}^P(x^k, \mu^k, \rho_k)$$

is bounded. Using that x^* is the regular solution of (1.1) so that $C_{\mathcal{F}}^T$ is a full column rank matrix, we conclude that μ^k is bounded and there is a cluster point $(x^*, \bar{\mu})$ of the sequence (x^k, μ^k) that satisfies $g^P(x^*, \bar{\mu}, 0) = 0$. Thus $(x^*, \bar{\mu})$ is the KKT couple for (1.1). Since the KKT couple (x^*, μ^*) of (1.1) is unique when the solution x^* of (1.1) is regular, it follows that μ^k converges to $\bar{\mu} = \mu^*$. □

7. Numerical experiments.

We have implemented Algorithm 3.1 in Matlab and solved two benchmarks with an aim to illustrate both its efficiency and the main theoretical results. The inner loop (Step 1) was realized by the reduced gradient projection algorithm with proportioning [21].

Problem 7.1. The first problem was designed to demonstrate the efficiency of our algorithm on the solution of a class of well conditioned sparse problems. The Hesssian matrix $A = A_P, P > 2$, of the quadratic function $q = q_P$ is the symmetric Toeplitz matrix of the order $2 * P^2$ that is fully determined by the entries $a_{11} = 12, a_{12} = -1$ and $a_{1,P-1} = -1$. The other vectors $b = b_P$ and $l = l_P$ are defined by the entries $b_i = -1,\ i = 1, \ldots, 2 * P^2$, and $l_i = -0.125 + 0.1 * \cos(2 * \pi * i/P^2),\ i = 1, \ldots, P^2$. The remaining entries of the vector l are set to $-\infty$. Finally the matrix $C = C_P$ has P rows with $2 * P^2$ entries which are zeros except that $c_{i,P^2-i+1} = 1$ and $c_{i,P^2+i} = -1,\ i = 1, \ldots, P$. The matrix C is designed to enforce the relations $x_{P^2-i+1} = x_{P^2+i},\ i = 1, \ldots, P$. Using the Gersghorin theorem [27], it is easy to see that the eigenvalues λ_i of any A_P satisfy $8 \leq \lambda_i \leq 16$. The initial approximation for x in the first run of the inner loop is zero, so that no bound constraints are active.

We solved the problem with $\eta = \|b_P\|,\ \beta = 10$ and $\mu^0 = 0$ using the stopping criteria $\left\|g_P^P(x, \mu, 0)\right\| \leq 10^{-5} \|b_P\|$ and $\|C_P x\| \leq 10^{-5} \|b_P\|$, where we denoted by g_P the gradient of the augmented Lagrangian

$$L_P(x, \mu, \rho) = q_P + \frac{1}{2}\rho \|C_P x\|^2 + \mu^T C_P x.$$

We have not observed any update of the penalty parameter, which confirms that our update rule is less aggressive than that introduced in [5] and used in our previous papers.

Table 7.1
Optimality of the semimonotonic algorithm with $M = 1$ and $\rho = 20$.

P	Dimension	Active bounds	Equality constraints	Outer iterations	cg iterations
10	200	47	10	9	38
50	5000	1239	50	9	41
100	20000	4997	100	8	39
250	125000	31193	250	8	44
500	500000	124887	500	8	44

Table 7.2
Effect of the penalty parameter ρ for $P = 100$ and $M = 1$.

ρ	Dimension	Outer iterations	cg iterations
1	20000	89	108
10	20000	13	42
100	20000	4	82
1000	20000	2	397

The results of numerical experiments with $M = 1$, $P \in \{10, 50, 100, 250, 500\}$, and $\rho = 20$ are in Table 7.1. For each value of P, the table includes also the number of the bound constraints that are active at the solution and the number of the inequality constraints in the columns labeled "Active bounds" and "Equality constraints," respectively. Observe that the value of ρ is an upper bound on the spectrum of A_P. This choice may be easily implemented in a more realistic situation and seems to work for well conditioned constraints. We can see that the number of outer iterations does not increase with the dimension of the problem which confirms a kind of optimality predicted by Theorem 5.3.

We examined also the effect of the choice of the penalty parameter ρ. In Table 7.2 there are the results of computations for $\rho \in \{1, 10, 100, 1000\}$ with $P = 100$ and $M = 1$, so that the dimension of the problem and the number of constraints was 20,000 and 100, respectively. We can observe the large number of iterations for $\rho = 1000$. This is caused by the short gradient projection step (in the algorithm for Step 1) which is inversely proportional to the penalty parameter and by the large number of active constraints of the solution that are not active at the initial approximation. Notice that the algorithm had to identify 4997 active bound constraints starting from the empty active set. The number of outer iterations decreases with the increasing penalty parameter.

To see the effect of the conditioning of the constraints, we modified the matrix C by adding to each row the sum of the rows of the original matrix that are above the modified row, normalized the modified rows with respect to the ℓ_2-norm, and then normalized the whole constraint matrix with respect to the ℓ_∞-norm. We used the penalty parameter $\rho = 200$. The results are in Table 7.3, which includes the conditioning of the constraints. We can see that the number of the outer iterations decreases although there is deteriorating conditioning of the constraints. The number of the inner iterations also surprisingly decreases with the increasing condition number of the constraint matrix.

To see what happens when the constraints are dependent, we first formed an auxiliary matrix $\widehat{C}$ by appending the first $P/2$ rows of C to C so that $\widehat{C}$ had $1.5 * P$ rows, and then we modified the matrix $\widehat{C}$ by summing and normalizing as above.

TABLE 7.3
Effect of conditioning of the equality constraints for $\rho = 200$ and $M = 1$.

P	Dimension	Equality constraints	Outer iterations	cg iterations	Conditioning
10	200	10	123	263	15.0816
50	5000	50	68	171	79.5278
100	20000	100	43	143	161.4914
250	125000	2500	22	118	408.9432

TABLE 7.4
Experiments with dependent rows with $\rho = 200$ and $M = 1$.

P	Dimension	Equality constraints	Outer iterations	cg iterations	Conditioning
10	200	15	95	233	19.5937
50	5000	75	26	135	87.1887
100	20000	150	16	139	170.7854
250	125000	375	10	145	539.1429

Thus the resulting constraint matrix corresponding to P has the same first P rows as the constraint matrix of the previous experiments and $P/2$ additional rows that are linear combinations of the first ones. We can observe that the results in Table 7.4 are even better that those in Table 7.3. By "Conditioning" we understand here the ratio of the largest singular value of the constraint matrix to the smallest nonzero one. The constraint matrix corresponding to P has in these experiments P nonzero singular values of $1.5 * P$.

Problem 7.2. The second problem illustrates the performance of the algorithm on the solution of a model elliptic boundary variational inequality introduced in [14] modifying the experimental code produced by Horák [17]. We shall describe it here only briefly, referring the interested reader to [14] or [17].

Let us start from the following continuous problem:

$$\text{(7.1)} \qquad \begin{aligned} &\text{Minimize } q(u_1, u_2) = \sum_{i=1}^{2} \left(\int_{\Omega^i} |\nabla u_i|^2 d\Omega - \int_{\Omega^i} f u_i d\Omega \right) \\ &\text{s.t. } u_1(0, y) \equiv 0 \text{ and } u_1(1, y) \leq u_2(1, y) \text{ for } y \in [0, 1], \end{aligned}$$

where $\Omega^1 = (0,1)\times(0,1)$, $\Omega^2 = (1,2)\times(0,1)$, $f(x,y) = -5$ for $(x,y) \in (0,1)\times[0.75,1)$, $f(x,y) = 0$ for $(x,y) \in (0,1) \times (0,0.75)$, $f(x,y) = -1$ for $(x,y) \in (1,2) \times (0,0.25)$, and $f(x,y) = 0$ for $(x,y) \in (1,2) \times (0.25,1)$. This problem is semicoercive due to the lack of Dirichlet data on the boundary of Ω^2.

The solution of the model problem may be interpreted as the displacement of two membranes under the traction f. The left membrane is fixed on the left and the left edge of the right membrane is not allowed to penetrate below the right edge of the left membrane. The solution is unique because the right membrane is pressed down. More details about this model problem, including some other results, may be found in [14].

To solve the problem (7.1), we used the well established FETI domain decomposition method with the natural coarse grid preconditioning that was introduced for linear problems by Farhat, Mandel, and Roux (e.g., [22]) and adapted to variational inequalities by Friedlander, Gomes, Santos, and the present author [12, 14]. Each domain Ω^i, $i = 1, 2$, was first decomposed into identical squares Ω^{ij} with sides of the length H. The squares were discretized by regular grids defined by the stepsize h. We fixed the ratio $H/h = 4$ and considered $H = 1/2, 1/4$, and $1/8$. After having introduced the equality constraints and the inequalities to enforce continuity across the artificial interfaces between the squares and nonpenetration of the edges of the membranes, respectively, indexing contiguously the nodes and entries of corresponding vectors in the subdomains, and using the finite element discretization, we get the discretized version of problem (7.1) with the auxiliary domain decomposition that reads

$$\min \frac{1}{2} u^T K u - f^T u \quad \text{s.t.} \quad B^I u \leq 0 \quad \text{and} \quad B^E u = 0. \tag{7.2}$$

In (7.2), $K = \operatorname{diag}(K_1, \ldots, K_s),\ s = 1, \ldots, 2/H^2$ denotes a positive semidefinite block-diagonal stiffness matrix, the full rank matrices B^I and B^E describe the discretized nonpenetration and gluing conditions, respectively, and f represents the discrete analogue of the linear term.

Introducing the notation

$$\lambda = \begin{bmatrix} \lambda^I \\ \lambda^E \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B^I \\ B^E \end{bmatrix},$$

we can write the Lagrangian associated with problem (7.2) briefly as

$$L(u, \lambda) = \frac{1}{2} u^T K u - f^T u + \lambda^T B u.$$

It is well known [3] that (7.2) is equivalent to the saddle point problem

$$\text{Find} \quad (\overline{u}, \overline{\lambda}) \quad \text{s.t.} \quad L(\overline{u}, \overline{\lambda}) = \sup_{\lambda_I \geq 0} \inf_{u} L(u, \lambda). \tag{7.3}$$

After eliminating the primal variables u from (7.3), we shall obtain the minimization problem

$$\min \Theta(\lambda) \quad \text{s.t.} \quad \lambda_I \geq 0 \quad \text{and} \quad R^T(f - B^T \lambda) = 0, \tag{7.4}$$

where

$$\Theta(\lambda) = \frac{1}{2} \lambda^T B A^\dagger B^T \lambda - \lambda^T B A^\dagger f, \tag{7.5}$$

$A^\dagger$ denotes a generalized inverse that satisfies $AA^\dagger A = A$, and R denotes the full rank matrix whose columns span the kernel of A. Let us recall that the multiplication by $A^\dagger$ may be efficiently evaluated by means of the triangular decomposition $A = LL^T$ [14]. We shall choose R so that its entries belong to $\{0, 1\}$ and each column corresponds to some floating auxiliary subdomain Ω^{ij} with the nonzero entries in the positions corresponding to the indices of nodes belonging to Ω^{ij}.

Although problem (7.4) is much more suitable for computations than (7.2), further improvement may be achieved by adapting some simple observations and the results of Farhat, Mandel, and Roux [22]. Let us denote

$$F = BA^\dagger B^T, \quad G = R^T B^T, \quad \widetilde{e} = R^T f, \quad \widetilde{d} = BA^\dagger f,$$

and let $\widetilde{\lambda}$ solve $G\widetilde{\lambda} = \widetilde{e}$. We can now transform the problem (7.4) to minimization on the subset of the vector space by looking for the solution in the form $\lambda = \mu + \widetilde{\lambda}$. Since

$$\frac{1}{2} \lambda^\top F \lambda - \lambda^\top \widetilde{d} = \frac{1}{2} \mu^\top F \mu - \mu^\top (\widetilde{d} - F\widetilde{\lambda}) + \frac{1}{2} \widetilde{\lambda}^\top F \widetilde{\lambda} - \widetilde{\lambda}^\top \widetilde{d},$$

problem (7.4) is, after returning to the old notation, equivalent to

$$\min \quad \frac{1}{2}\lambda^\top F\lambda - \lambda^\top d \quad \text{s.t} \quad G\lambda = 0 \quad \text{and} \quad \lambda^I \geq -\widetilde{\lambda}^I \tag{7.6}$$

with $d = \widetilde{d} - F\widetilde{\lambda}$. It is possible [19] to find $\widetilde{\lambda}$ such that $\widetilde{\lambda}^I \geq 0$.

Our final step is based on observation that the problem (7.6) is equivalent to

$$\min q(\lambda) \qquad \text{s.t} \quad G\lambda = 0 \quad \text{and} \quad \lambda^I \geq -\widetilde{\lambda}^I \tag{7.7}$$

where

$$q(\lambda) \ = \ \frac{1}{2}\lambda^T(PFP + \widehat{\rho}Q)\lambda - \lambda^T Pd$$

and

$$Q = G^T(GG^T)^{-1}G, \qquad\qquad P = I - Q,$$

denote the orthogonal projectors on the image space of $G^\top$ and on the kernel of G, respectively, and $\widehat{\rho}$ is a (small) positive number. Let us point out that $\widehat{\rho}$ is important rather from the point of view of analysis than in computations. In what follows we shall denote by ρ_i the sum of $\widehat{\rho}$ and of the penalty parameter in the augmented Lagrangian. Since we implemented Step 1 of Algorithm 3.1 with the conjugate gradient method, it is important to note that image spaces of the projectors P and Q are the invariant subspaces of the Hessian $PFP + \rho Q$ of the augmented Lagrangian for the problem (7.7), so that the convergence of the conjugate gradients may be fast even for large penalty parameters [9].

Each discretization of the model problem (7.1) is determined by a couple $D = (H, h)$ of the decomposition and the discretization parameter H and h, respectively. Thus using the domain decomposition and the discretization of the model problem (7.1), we get the set of quadratic programming problems

$$\min\ q_D(\lambda) \ \text{ s.t. } \ C_D\lambda = 0 \ \text{ and } \ \lambda \geq e_D, \qquad q_D = \lambda^T A_D\lambda - b_D^T\lambda, \tag{7.8}$$

that are defined for each

$$D \in \mathcal{T} = \{D = (H, h) : 0 < h < H \leq 1, h^{-1} \in \mathbb{N}, H^{-1} \in \mathbb{N}, H/h \in \mathbb{N}\}.$$

Let us recall that the Hessian A_D is in this case the product of eight large and sparse matrices. More details concerning the application of FETI may be found in [14, 17].

We first solved the problems by Algorithm 3.1 with $\rho_0 = 100, \eta = \|b_D\|, \beta = 10$, and $\mu^0 = 0$ using the stopping criteria $\left\|g_D^P(\lambda, \mu, 0)\right\| \leq 10^{-4}\|b_D\|$ and $\|C_D\lambda\| \leq 10^{-4}\|b_D\|$, where we denoted by g_D the gradient of the augmented Lagrangian for the problem (7.7) with the discretization D. We kept H/h fixed so that the assumptions of Theorem 5.3 were satisfied [19]. The results are given in Table 7.5. We can see that the number of the outer iterations is only 3 regardless the value of the parameter H, the dimension of the problem, and even the number of constraints. An interesting feature of the experiments is also the small number of the conjugate gradient iterations due to efficiency of both FETI and our algorithms. We shall discuss this point elsewhere.

Theorem 5.2 indicates that the rate of convergence of the feasibility error depends on the initial penalty parameter. To illustrate this feature, we used varying initial penalty parameter to solve our model problem with $H/h = 8$ and $H = 1/4$. The

Table 7.5
Performance of the semimonotonic algorithm for $H/h = 4$.

H	Dual dimension	Number of constraints	Outer iterations	Primal dimension	cg iterations
1/2	47	6	3	200	23
1/4	239	28	3	800	24
1/8	1055	120	3	3200	38

Table 7.6
Development of the feasibility error $\|Cx^k\|$.

Penalty	cg iterations	Outer iterations	$\|Cx^0\|$	$\|Cx^1\|$	$\|Cx^2\|$	$\|Cx^3\|$
1	23	4	$6.19e-2$	$1.82e-3$	$8.90e-5$	$4.23e-6$
10	28	3	$3.96e-4$	$1.97e-5$	$1.04e-6$	
100	28	2	$4.14e-5$	$2.23e-6$		
1000	24	1	$4.16e-6$			

dimension of the dual and the primal problem was 447 and 2592, respectively, with 28 equality constraints. The record of the development of the feasibility error is given in Table 7.6. We observed that the penalty parameter was updated only once, after the first iteration with $\rho_0 = 1$, which was the only case when we recorded the decrease of the augmented Lagrangian.

The values of the augmented Lagrangian are given in Table 7.7. Observe that the algorithm can generate for the small penalty parameters values of the augmented Lagrangian that are greater than its value at the solution. We conclude that the results are in agreement with the theory.

8. Comments and conclusions. We have introduced a new update rule for the penalty parameter that enforces at a certain stage monotonic increase of the augmented Lagrangian. We implemented this rule using our earlier algorithm and proved global convergence results for the augmented Lagrangian method. This method uses adaptive precision control in the solution of the auxiliary problems for quadratic programming problems with equality constraints. The precision is controlled by the feasibility of the current iteration. The new feature of the algorithm is that it does not take any special measures to treat the iterations that are not regular and the theory supports convergence even in the case when the equality constraints are dependent.

The algorithm is a variant of the well established algorithm [15] which has already proved to be useful in the development of scalable algorithms for the numerical solution of elliptic variational inequalities [17] and for the solution of contact problems of elasticity [12]. In fact, if ρ_0 is chosen in the algorithm [15] to be sufficiently large so that it is not updated, and if all the iterations are regular, then the performance of both algorithms is identical. This paper may be considered a complement of [15] in the sense that it explains what happens when the penalty parameter in this type of algorithm is relatively small without any reference to convergence of the Lagrange multipliers.

The main results of the paper are the *bounds on the feasibility error and the penalty parameters that do not depend on the form of constraints*. When applied to a class of problems with the spectrum of the Hessian matrix in a given interval, the algorithm returns the solution in $O(1)$ matrix-vector multiplications. In combination with the recent results on the solution of the bound constrained quadratic programming problems [10, 21], it follows that if the cost of the matrix-vector multiplication

TABLE 7.7
Values of the augmented Lagrangian $L_k = L(x^k, \mu^k, \rho^k)$.

Penalty	L_0	L_1	L_2	L_3
1	−.2043688	−.2090037	−.2076275	−.2075860
10	−.2084112	−.2075881	−.2075860	
100	−.2076722	−.2075859		
1000	−.2075946			

by the Hessian matrix is proportional to its dimension, *the algorithm finds the approximate solution to the prescribed relative precision at the optimal (i.e., asymptotically proportional to the dimension of the problem) cost.* The results were also confirmed numerically on the solution of two model problems and were shown to be valid even for linearly dependent constraints. The result of the paper is an important ingredient in the development of scalable algorithms for variational inequalities. We shall describe this application in more detail elsewhere.

Let us recall that there are algorithms with even superlinear convergence (e.g., variants of the primal-dual interior point methods [30]), but their typical step requires the solution of an auxiliary ill conditioned linear problem which may be, for sufficiently large and sparse problems, much more expensive than the solution of the whole problem by the algorithm presented here.

Acknowledgments. I wish to express here my thanks to D. Horák, who assisted with the numerical experiments and reading the paper. My special thanks go to A. R. Conn and the anonymous referee for their valuable comments that resulted in considerable improvement of the paper.

REFERENCES

[1] PH. AVERY, G. REBEL, M. LESOINNE, AND C. FARHAT, *A numerically scalable dual-primal substructuring method for the solution of contact problems—part* I: *The frictionless case*, Computer Methods in Applied Mechanics and Engineering, 193 (2004), pp. 2403–2426.

[2] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, London, 1982.

[3] D. P. BERTSEKAS, *Nonlinear Optimization*, Athena Scientific, Belmont, 1999.

[4] E. G. BIRGIN AND J. M. MARTINEZ, *Large-Scale Active-Set Box-Constrained Optimization Method with Spectral Projected Gradients*, Comput. Optim. Appl., 23 (2002), pp. 101–125.

[5] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds*, SIAM J. Numer. Anal., 28 (1991), pp. 545–572.

[6] A. R. CONN, N. I. M. GOULD, AND PH. L. TOINT, *LANCELOT: A Fortran Package for Large Scale Nonlinear Optimization*, Springer-Verlag, Berlin, 1992.

[7] Z. DOSTÁL, *Duality based domain decomposition with proportioning for the solution of free boundary problems*, J. Comput. Appl. Math., 63 (1995), pp. 203–208.

[8] Z. DOSTÁL, *Box constrained quadratic programming with proportioning and projections*, SIAM J. Optim., 7 (1997), pp. 871–887.

[9] Z. DOSTÁL, *On preconditioning and penalized matrices*, Numer. Linear Algebra Appl., 6 (1999), pp. 109–114.

[10] Z. DOSTÁL, *A proportioning based algorithm for bound constrained quadratic programming with the rate of convergence*, Numer. Algorithms, 34 (2003), pp. 293–302.

[11] Z. DOSTÁL, *Semi-monotonic inexact augmented Lagrangians for quadratic programming with equality constraints*, Optim. Methods Softw., to appear.

[12] Z. DOSTÁL, A. FRIEDLANDER, AND S. A. SANTOS, *Solution of Coercive and Semicoercive Contact Problems by FETI Domain Decomposition*, Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 83–93.

[13] Z. Dostál, A. Friedlander, and S. A. Santos, *Augmented Lagrangians with adaptive precision control for quadratic programming with equality constraints*, Comput. Optim. Appl., 14 (1999), pp. 37–53.

[14] Z. Dostál, F. A. M. Gomes, and S. A. Santos, *Duality based domain decomposition with natural coarse space for variational inequalities*, J. Comput. Appl. Math., 126 (2000), pp. 397–415.

[15] Z. Dostál, A. Friedlander, and S. A. Santos, *Augmented Lagrangians with adaptive precision control for quadratic programming with simple bounds and equality constraints*, SIAM J. Optim., 13 (2003), pp. 1120–1140.

[16] Z. Dostál, F. A. M. Gomes Neto, and S. A. Santos, *Solution of contact problems by FETI domain decomposition with natural coarse space projection*, Comput. Meth. Appl. Mech. Eng., 190 (2000), pp. 1611–1627.

[17] Z. Dostál and D. Horák, *Scalability and FETI based algorithm for large discretized variational inequalities*, Math. Comput. Simulation, 61 (2003), pp. 347–357.

[18] Z. Dostál and D. Horák, *Scalable FETI with Optimal Dual Penalty for a Variational Inequality*, Numer. Linear Algebra Appl., 11 (2004), pp. 455–472.

[19] Z. Dostál and D. Horák, *Scalable FETI with Optimal Dual Penalty for Semicoercive Variational Inequalities*, Contemp. Math. 329, AMS, Providence, RI, 2003, pp. 79–88.

[20] Z. Dostál, D. Horák, and D. Stefanica, *A scalable FETI–DP algorithm for coercive variational inequalities*, IMACS J. Appl. Numer. Anal., to appear.

[21] Z. Dostál and J. Schöberl, *Minimizing quadratic functions subject to bound constraints with the rate of convergence and finite termination*, Comput. Optim. Appl., 30 (2005), pp. 23–44.

[22] C. Farhat, J. Mandel, and F.-X. Roux, *Optimal convergence properties of the FETI domain decomposition method*, Comput. Methods Appl. Mech. Eng., 115 (1994), pp. 365–385.

[23] A. Friedlander, J. M. Martínez, and M. Raydan, *A new method for large scale box constrained quadratic minimization problems*, Optim. Methods Softw., 5 (1995), pp. 57–74.

[24] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics*, SIAM Stud. Appl. Math. 9, SIAM, Philadelphia, 1989.

[25] W. W. Hager, *Analysis and implementation of a dual algorithm for constraint optimization*, J. Optim. Theory Appl., 79 (1993), pp. 37–71.

[26] M. R. Hestenes, *Multiplier and gradient methods*, J. Optim. Theory Appl., 4 (1969), pp. 303–320.

[27] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1985.

[28] M. J. D. Powell, *A Method for Nonlinear Constraints in Minimization Problems*, R. Fletcher, ed., Academic Press, New York, 1969, pp. 283–298.

[29] J. Schöberl, *Solving the Signorini problem on the basis of domain decomposition techniques*, Computing, 60 (1998), pp. 323–344.

[30] S. J. Wright, *Primal-Dual Interior-Point Methods*, SIAM, Philadelphia, 1997.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 116–126

A DIV-CURL LEMMA FOR EDGE ELEMENTS[*]

SNORRE H. CHRISTIANSEN[†]

Abstract. A variant of Murat and Tartar's div-curl lemma is stated and proved for Nédélec's edge elements. Given two sequences of vector fields of this type, converging weakly in L^2 as the mesh width tends to 0, we prove that their scalar product converges in the sense of distributions when one of the sequences consists of so-called discrete divergence-free fields whereas the other has relatively compact curl in H^{-1}. The proof uses a uniform norm equivalence related to discrete compactness properties of vector finite element spaces and a super-approximation property of scalar finite element spaces.

Key words. div-curl lemma, compensated compactness, edge elements, discrete compactness

AMS subject classifications. 65N12, 65N30

DOI. 10.1137/S0036142903433807

1. Introduction.

Given two sequences of vector fields converging weakly in L^2, the div-curl lemma of Murat [29] and Tartar [37], recalled in Theorem 1.1 below, gives sufficient conditions under which their scalar product converges to the right scalar field in the weak-star sense of distributions. Following Schwartz, the space of compactly supported smooth functions is denoted $\mathcal{D}$, and the space of distributions, which contains L^1 and is dual to $\mathcal{D}$ for a certain topology, is denoted $\mathcal{D}'$.

THEOREM 1.1. *Suppose (u_h) and (u'_h) are sequences of vector fields converging weakly in L^2 to u and u'. Suppose, furthermore, that $(\operatorname{div} u_h)$ is relatively compact in H^{-1} and that $(\operatorname{curl} u'_h)$ is relatively compact in H^{-1}. Then for each $\phi \in \mathcal{D}$ we have*

$$\lim_h \int (u_h \cdot u'_h)\phi = \int (u \cdot u')\phi. \tag{1.1}$$

Following Folland [20] we shall refer to weak-star as vague convergence, since this leads to adjectives more easily. Thus (1.1) expresses that the sequence $(u_h \cdot u'_h)$ converges vaguely to $u \cdot u'$ in $\mathcal{D}'$.

Over the years the curl conforming finite element (FE) spaces constructed by Nédélec [31, 32], also referred to as edge elements since the principal degrees of freedom are attached to the edges, have emerged as the spaces of choice for discretizing Maxwell's equations (Cessenat [12], Jackson [24], Nédélec [33]). This success is based on numerous numerical and theoretical studies; for surveys and references, see Monk [28], Hiptmair [23], and Joly [25]. For the purposes of this paper we notice in particular that a discrete compactness property of Kikuchi [27] has been used to show that edge element spaces do not yield spurious eigenvalues when used to compute eigenvalues of the curl curl operator, contrary to some tempting so-called nodal FE spaces (see Boffi, Brezzi, and Gastaldi [4], Boffi, Fernandes, and Gastaldi [5], and Boffi [6]).

In this paper we prove another good property of edge elements, this time related to nonlinear convergence theory. Most numerical schemes of electromagnetics do not offer direct control of the divergence of the discrete vector fields. However, it is widely acknowledged that the success of edge elements is related to weak divergence control.

[*] Received by the editors August 26, 2003; accepted for publication (in revised form) July 6, 2004; published electronically April 26, 2005.
http://www.siam.org/journals/sinum/43-1/43380.html

[†] CMA, Universitet i Oslo, PB 1053 Blindern, NO-0316 Oslo, Norway (snorrec@math.uio.no).

In the simplest case the mechanism is as follows: Let X_h denote the space of lowest order Nédélec edge elements on a simplicial triangulation with mesh width h, and let Y_h denote the continuous piecewise affine functions on the same triangulation. The first (by now standard) remark is that the cohomology group of the sequence

$$Y_h \xrightarrow{\text{grad}} X_h \xrightarrow{\text{curl}} \mathrm{L}^2, \tag{1.2}$$

has the dimension of its continuous counterpart, essentially because simplicial cohomology is equivalent to De Rham cohomology. This accounts for the naturality of considering Y_h and X_h together and has attracted much interest in the numerical analysis community; see Arnold [2]. We denote by $\widetilde{Y}_h$ the intersection $Y_h \cap \mathrm{H}^1_0$. In the Galerkin formulation of electromagnetic problems the only information available a priori on the divergence of an electric field $u_h \in X_h$ is through the graph

$$\left\{ \left(p_h, \int u_h \cdot \operatorname{grad} p_h \right) : \; p_h \in \widetilde{Y}_h \right\}. \tag{1.3}$$

How does this discrete divergence information translate into H^{-1} estimates on $\operatorname{div} u_h$? For instance a basic question is, suppose that we have a sequence $(\mathcal{T}_h)$ of meshes with mesh width $h \to 0$, yielding sequences of spaces (X_h) and (Y_h) and suppose that (u_h) is a sequence of vector fields $u_h \in X_h$, bounded in L^2 and discrete divergence-free in the sense that

$$\forall p_h \in \widetilde{Y}_h, \qquad \int u_h \cdot \operatorname{grad} p_h = 0; \tag{1.4}$$

is $(\operatorname{div} u_h)$ relatively compact in H^{-1}?

I do not know the answer to this question but if affirmative, then a direct consequence of Theorem 1.1 would be Corollary 4.2. One of the main goals of this paper is nevertheless to prove this corollary. It is a special case of Theorem 4.1, which is a div-curl lemma for edge elements paralleling Theorem 1.1. Thus, loosely speaking, we prove that with respect to the convergence of edge element vector fields under div-curl control, everything happens as if the answer to the above question were affirmative.

The paper is organized as follows. In section 2 we introduce the general setting we shall use in this paper and check that the edge element spaces fit into it. In section 3 we introduce some tools used to prove the proposed div-curl lemma. This proof is provided in section 4. Finally in section 5 we establish some links between the estimates proved in this paper, other estimates on discrete Helmholtz decompositions, and the previously mentioned discrete compactness property of Kikuchi.

2. Setting.

2.1. Generalities.

Let Ω be a connected bounded domain in $\mathbb{R}^3$ with a smooth boundary. For simplicity we suppose that Ω has the topological property that the kernel of the curl operator on $\mathrm{L}^2 = \mathrm{L}^2(\Omega)$ is exactly the range of the gradient on $\mathrm{H}^1 = \mathrm{H}^1(\Omega)$. The extension of the following results to general topology is straightforward using the fact that the edge elements approximate the L^2 realizations of cohomology groups (harmonic vector fields) extremely well (see, e.g., section 2.3 in [15]). The standard norms and seminorms on H^k are denoted $\| \cdot \|_k$ and $| \cdot |_k$, respectively.

Let X denote the space L^2, W the kernel of the curl : $X \to \mathrm{H}^{-1}$, and V its L^2 orthogonal in X. Thus we have

$$W = \{\operatorname{grad} p\colon \; p \in \mathrm{H}^1\}, \tag{2.1}$$

and, denoting by n the exterior normal on $\partial\Omega$,

$$V = \{v \in \mathrm{L}^2\colon\ \operatorname{div} v = 0 \text{ and } v \cdot n|_{\partial\Omega} = 0\}. \tag{2.2}$$

Let P_V denote the projection with range V and kernel W. It is nothing but the L^2 projection onto V and it preserves the curl; i.e., for all vector fields $u \in \mathrm{L}^2$,

$$\operatorname{curl} P_V u = \operatorname{curl} u. \tag{2.3}$$

A less straightforward property of V is the following.

Proposition 2.1. *On V, $\|\operatorname{curl}(\cdot)\|_{-1}$ is a norm equivalent to the* L^2 *norm.*

Proof. The map curl : $V \to \mathrm{H}^{-1}$ is continuous and injective, so it remains to prove that it has closed range. By the closed range theorems it suffices to prove that the range of curl : $\mathrm{H}^1_0 \to \mathrm{L}^2$ is closed. However, this latter space is a subspace of V, and Theorem 3.20 (which requires $C^{1,1}$ regularity of the boundary) in [1] provides a closed subspace U of H^1_0 such that curl : $U \to V$ is an isomorphism. In particular the range of curl : $\mathrm{H}^1_0 \to \mathrm{L}^2$ must be V, which is closed. This completes the proof. □

Remark. M. Costabel has indicated to me that by slightly adapting techniques from Girault–Raviart [22] one can show that the above result remains true for domains whose boundary is locally the graph of Lipschitz functions.

We will use the following setting. In accordance with widespread notational conventions for FE discretizations, we consider families indexed by a given countable set of positive reals accumulating only at 0. The dummy variable is denoted h and we are interested in the limit $h \to 0$. Let (X_h) and (Y_h) denote two sequences of finite-dimensional spaces such that

$$Y_h \subset \mathrm{H}^1 \quad \text{and} \quad X_h \subset \{u \in \mathrm{L}^2\colon\ \operatorname{curl} u \in \mathrm{L}^2\}. \tag{2.4}$$

We suppose that Y_h contains the constant scalar fields and that the operator grad maps Y_h *onto* the kernel W_h of the curl operator restricted to X_h. We denote by V_h the L^2-orthogonal of W_h in X_h. Thus we have an exact sequence

$$\mathbb{R} \to Y_h \to X_h \to \mathrm{L}^2, \tag{2.5}$$

and L^2-orthogonal splittings

$$X_h = V_h \oplus W_h. \tag{2.6}$$

We also put $\widetilde{Y}_h = Y_h \cap \mathrm{H}^1_0$, define $\widetilde{W}_h = \operatorname{grad} \widetilde{Y}_h$, and denote by $\widetilde{V}_h$ the L^2-orthogonal of $\widetilde{W}_h$ in X_h.

Some approximation properties are also assumed for (X_h) and (Y_h) throughout the rest of this paper without further notice:

$$\forall u \in \mathrm{L}^2, \quad \lim_{h\to 0} \inf_{u_h \in X_h} \|u - u_h\|_0 = 0, \tag{2.7}$$

$$\forall p \in \mathrm{H}^1, \quad \lim_{h\to 0} \inf_{p_h \in Y_h} \|p - p_h\|_1 = 0, \tag{2.8}$$

$$\forall p \in \mathrm{H}^1_0, \quad \lim_{h\to 0} \inf_{p_h \in \tilde{Y}_h} \|p - p_h\|_1 = 0. \tag{2.9}$$

Of course all FE spaces satisfy them, and they can be interpreted as the pointwise convergence of the corresponding L^2 and H^1 projectors.

Unless specified otherwise, the above hypotheses are the only ones we assume in the next sections of the paper. However, for most of our results one or two of the following assumptions will play a crucial role.

UNE (uniform norm equivalence). There is $C > 0$ such that

$$\forall h \; \forall v_h \in V_h, \quad \|v_h\|_0 \leq C \| \operatorname{curl} v_h \|_{-1}. \tag{2.10}$$

SA (super-approximation). For any function $\phi \in \mathcal{D}$ (smooth and compactly supported) we have

$$\lim_{h \to 0} \sup_{p_h \in Y_h} \inf_{\widetilde{p}_h \in \widetilde{Y}_h} \|\phi p_h - \widetilde{p}_h\|_1 / \|p_h\|_1 = 0. \tag{2.11}$$

When they are needed, we will specify them as part of the hypotheses of the statements.

Remark. Since Y_h is assumed to contain the constant functions and $\mathcal{D}$ is dense in H^1_0, (SA) implies (2.9).

2.2. Edge elements. Let $(\mathcal{T}_h)$ be a family of simplicial meshes on Ω such that (in accordance with widespread notational practice) the mesh width of $\mathcal{T}_h$ is h. We suppose furthermore that $(\mathcal{T}_h)$ is shape-regular and uniform in the standard senses (see, e.g., Braess [8, p. 61]), so that in particular inverse inequalities can be used.

Given an integer $k \geq 1$ we consider (for each h) on $\mathcal{T}_h$ the edge element space X_h of Nédélec of the first or second family and constructed with (incomplete vector) polynomials of maximum degree k, as well as the associated space Y_h of continuous piecewise polynomial functions. We suppose that these spaces have been adequately fitted near the curved boundary as in Dubois [19]. These spaces fit into the above setting, and in the next two propositions we check that they also satisfy (UNE) and (SA).

PROPOSITION 2.2. *The family* (X_h) *satisfies* (UNE).

As we shall see this is a very strong statement. For instance it is a considerable strengthening of Proposition 4.6 in [1], but has a more restrictive boundary regularity hypothesis. It is unknown to me if the present proposition remains true for Lipschitz domains or without the uniformity hypothesis on the meshes.

Proof. Let Π_h denote the standard edge element interpolator. It maps curl-free fields to curl-free fields, due to the commuting diagram property of standard interpolation operators. For any v_h in V_h, $P_V v_h$ is in H^1, due to our regularity assumptions on the boundary, and has a piecewise smooth curl. Therefore Π_h is well defined on $P_V v_h$ and we may write

$$\operatorname{curl}(\Pi_h P_V v_h - v_h) = \operatorname{curl} \Pi_h (P_V v_h - v_h) = 0. \tag{2.12}$$

Hence v_h, $P_V v_h$, and $\Pi_h P_V v_h$ have the same curl. Next we use a trick due to V. Girault for which I refer to the proof of Lemma 4.1 in Ciarlet–Zou [16]. A Bramble–Hilbert type error estimate for Π_h, using the additional information available on $\operatorname{curl} v_h$ (when transported from the cells of $\mathcal{T}_h$ to the reference cell, it lies in a fixed finite-dimensional space), yields

$$\|\Pi_h P_V v_h - P_V v_h\|_0 \leq Ch |P_V v_h|_1. \tag{2.13}$$

Then Lemma 2.11 in [1] gives

$$\|\Pi_h P_V v_h - P_V v_h\|_0 \leq Ch \| \operatorname{curl} v_h \|_0, \tag{2.14}$$

and we proceed as in the proof of Theorem 3.5 in [13].

Putting $w_h = \Pi_h P_V v_h - v_h \in W_h$ we remark that w_h is L^2-orthogonal to $v_h \in V_h$ and (being a gradient) also to $P_V v_h \in V$. Hence we have

$$\|w_h\|_0^2 = \int w_h \cdot (\Pi_h P_V v_h - P_V v_h) \leq \|w_h\|_0 \|\Pi_h P_V v_h - P_V v_h\|_0. \tag{2.15}$$

It follows that

$$\|w_h\|_0 \leq Ch \|\operatorname{curl} v_h\|_0. \tag{2.16}$$

Now, using an inverse inequality, we obtain

$$\begin{aligned} \|P_V v_h - v_h\|_0 &\leq \|\Pi_h P_V v_h - P_V v_h\|_0 + \|\Pi_h P_V v_h - v_h\|_0 && (2.17)\\ &\leq Ch \|\operatorname{curl} v_h\|_0 && (2.18)\\ &\leq C \|\operatorname{curl} v_h\|_{-1}. && (2.19) \end{aligned}$$

Using Proposition 2.1 we deduce

$$\begin{aligned} \|v_h\|_0 &\leq \|P_V v_h\|_0 + \|P_V v_h - v_h\|_0 && (2.20)\\ &\leq C \|\operatorname{curl} v_h\|_{-1}, && (2.21) \end{aligned}$$

which is the desired result. □

PROPOSITION 2.3. *The family* (Y_h) *satisfies* (SA).

Remark. In fact much more elaborate estimates are known and used in proofs of super-convergence properties; see Wahlbin [39, pp. 36–37] for a discussion of the techniques involved.

Remark. I thank one of the referees for showing me how to get rid of the uniformity hypothesis on the meshes in the proof of this proposition.

Proof. Denote by k the maximum degree of the polynomials used to construct the spaces Y_h. Let Π_h be the standard nodal interpolator. For any tetrahedron T we denote by h_T its diameter, and by $\|\cdot\|_{T,i}$ and $|\cdot|_{T,i}$ the standard $\mathrm{H}^i(T)$ norms and seminorms, respectively.

Pick $\phi \in \mathcal{D}$. For each h, each $p_h \in Y_h$, and each tetrahedron T of $\mathcal{T}_h$, we may write

$$\begin{aligned} \|\phi p_h - \Pi_h(\phi p_h)\|_{T,1} &\leq C h_T^k |\phi p_h|_{T,k+1} && (2.22)\\ &\leq C h_T^k \sum_{i=0}^{k} |p_h|_{T,i}, && (2.23) \end{aligned}$$

where we used first the standard Bramble–Hilbert lemma and second the Leibniz rule together with the fact that the derivatives of p_h of order $k+1$ vanish. Next, for $2 \leq i \leq k$, we use an inverse estimate (local to each T) to obtain

$$|p_h|_{T,i} \leq C h_T^{1-i} \|p_h\|_{T,1}. \tag{2.24}$$

Combining the above estimates and adding over all tetrahedra of $\mathcal{T}_h$ yields

$$\|\phi p_h - \Pi_h(\phi p_h)\|_1 \leq Ch \|p_h\|_1. \tag{2.25}$$

This completes the proof. □

3. Three tools. The proof of the proposed div-curl lemma is based on the following three results.

Lemma 3.1. *Suppose (u_h) is a sequence of vector fields $u_h \in X_h$ converging weakly in* L^2 *to $u = v + w$ with $v \in V$ and $w \in W$. Then with the decomposition $u_h = v_h + w_h$, with $v_h \in V_h$ and $w_h \in W_h$, the sequences (v_h) and (w_h) converge weakly to v and w in* L^2. *A similar result holds for the splitting $X_h = \widetilde{V}_h \oplus \widetilde{W}_h$.*

Proof. Put $w_h = \operatorname{grad} p_h$ with $p_h \in Y_h$ with vanishing integral. Let P_h denote the Galerkin projection onto the subspace of elements of Y_h with vanishing integral, with respect to the bilinear form

$$(p, p') \to \int \operatorname{grad} p \cdot \operatorname{grad} p'. \tag{3.1}$$

Then we have, for all $p' \in \mathrm{H}^1(\Omega)$ with vanishing integral,

$$\int \operatorname{grad} p_h \cdot \operatorname{grad} p' = \int \operatorname{grad} p_h \cdot \operatorname{grad} P_h p' = \int u_h \cdot \operatorname{grad} P_h p'. \tag{3.2}$$

Since $(P_h p')$ converges strongly to p' in H^1 by the approximation property of (2.8), the above quantity converges to

$$\int u \cdot \operatorname{grad} p' = \int w \cdot \operatorname{grad} p'. \tag{3.3}$$

It follows that (w_h) converges weakly in L^2 to w. Therefore (v_h) also converges weakly to v.

To prove the similar result for the splitting $X_h = \widetilde{V}_h \oplus \widetilde{W}_h$ one can use a similar technique, relying on estimate (2.9) instead of estimate (2.8). □

Proposition 3.2. *Suppose* (UNE) *holds. Let (v_h) be a sequence of vector fields such that $v_h \in V_h$. If* $(\operatorname{curl} v_h)$ *converges to* $\operatorname{curl} v$ *in* H^{-1} *for some $v \in V$, then (v_h) converges to v in* L^2.

Proof. By Proposition 2.1, $(P_V v_h)$ converges to v in L^2. Also, if we let P_h denote the L^2 projection onto X_h, $(P_h v)$ converges to v in L^2 by the approximation property of (2.7). Moreover P_h maps V into V_h. We can now write

$$\|P_h v - v_h\|_0 \leq C\|\operatorname{curl} P_h v - \operatorname{curl} v_h\|_{-1} \tag{3.4}$$

$$\leq C\|\operatorname{curl} P_h v - \operatorname{curl} P_V v_h\|_{-1} \tag{3.5}$$

$$\leq C\|P_h v - P_V v_h\|_0. \tag{3.6}$$

It follows that (v_h) converges to v. □

Corollary 3.3. *Suppose* (UNE) *holds. Let (v_h) be a sequence of vector fields such that $v_h \in V_h$. If* $(\operatorname{curl} v_h)$ *is relatively compact in* H^{-1}, *then (v_h) is relatively compact in* L^2.

Proposition 3.4. *Suppose* (SA) *holds. Suppose (v_h) is a sequence of vector fields $v_h \in \widetilde{V}_h$ converging weakly in* L^2 *to v, and (p_h) is a sequence of scalar fields $p_h \in Y_h$ converging weakly in* H^1 *to p. Then $(v_h \cdot \operatorname{grad} p_h)$ converges vaguely to $v \cdot \operatorname{grad} p$ in $\mathcal{D}'$.*

Proof. Pick $\phi \in \mathcal{D}$. We have

$$\int (v_h \cdot \operatorname{grad} p_h)\phi = \int v_h \cdot \operatorname{grad}(\phi p_h) - \int (v_h \cdot \operatorname{grad} \phi) p_h. \tag{3.7}$$

Since $v_h \in \widetilde{V}_h$, the first term can be replaced by

$$\int v_h \cdot \operatorname{grad}(\phi p_h - \widetilde{p}_h) \tag{3.8}$$

for any $\widetilde{p}_h \in \widetilde{Y}_h$. It converges to 0 by (SA).

In the second term we remark that $(v_h \cdot \operatorname{grad} \phi)$ converges weakly to $v \cdot \operatorname{grad} \phi$ in L^2, whereas (p_h) converges strongly in L^2 by the compactness of the injection $\mathrm{H}^1 \to \mathrm{L}^2$.

Therefore,

$$\int (v_h \cdot \operatorname{grad} p_h)\phi \;\to\; -\int (v \cdot \operatorname{grad} \phi) p = \int (v \cdot \operatorname{grad} p)\phi. \tag{3.9}$$

This completes the proof. □

4. Div-curl lemma for edge elements. We are now ready to prove our main result.

THEOREM 4.1. *Suppose both* (UNE) *and* (SA) *hold. Suppose* (u_h) *and* (u'_h) *are sequences of vector fields* $u_h, u'_h \in X_h$ *converging weakly in* L^2 *to* u *and* u'. *Suppose, furthermore, that with the decomposition* $u_h = v_h + \operatorname{grad} p_h$ *for* $v_h \in \widetilde{V}_h$ *and* $p_h \in \widetilde{Y}_h$, (p_h) *is relatively compact in* H^1_0, *and* $(\operatorname{curl} u'_h)$ *is relatively compact in* H^{-1}.

Then $(u_h \cdot u'_h)$ *converges vaguely to* $u \cdot u'$ *in* $\mathcal{D}'$.

Proof. Decompose also $u'_h = v'_h + \operatorname{grad} p'_h$ with $v'_h \in V_h$ and $p'_h \in Y_h$. Recall that by Lemma 3.1, all terms of the decompositions converge weakly. The limits will be denoted by skipping the subscript h.

Pick $\phi \in \mathcal{D}$. Each $\int (u_h \cdot u'_h)\phi$ can be decomposed into several terms studied separately.

First, we remark that $(\operatorname{grad} p_h)$ converges strongly in L^2 by the hypothesis. Therefore we have

$$\int (\operatorname{grad} p_h \cdot u'_h)\ \phi \;\to \int (\operatorname{grad} p \cdot u')\phi. \tag{4.1}$$

Next, we remark that by Proposition 3.2, (v'_h) converges strongly in L^2 to v. Hence we have

$$\int (v_h \cdot v'_h)\phi \;\to\; \int (v \cdot v')\phi. \tag{4.2}$$

Finally, we remark that by Proposition 3.4 we have

$$\int (v_h \cdot \operatorname{grad} p'_h)\ \phi \;\to \int (v \cdot \operatorname{grad} p')\phi. \tag{4.3}$$

This completes the proof. □

A useful special case of Theorem 4.1 is the following.

COROLLARY 4.2. *Suppose both* (UNE) *and* (SA) *hold. Suppose* (u_h) *and* (u'_h) *are sequences of vector fields* $u_h, u'_h \in X_h$ *converging weakly in* L^2 *to* u *and* u'. *Suppose furthermore that* u_h *is discrete divergence-free (i.e., is in* $\widetilde{V}_h$*) and that* $(\operatorname{curl} u'_h)$ *is bounded in* L^2.

Then $(u_h \cdot u'_h)$ *converges vaguely to* $u \cdot u'$ *in* $\mathcal{D}'$.

The next remark shows that on one point the hypothesis of Theorem 4.1 is not strictly stronger than that of Theorem 1.1.

Remark. Suppose (u_h) is a sequence of vector fields $u_h \in X_h$ such that $(\operatorname{div} u_h)$ is relatively compact in H^{-1}. Then with the decomposition $u_h = v_h + \operatorname{grad} p_h$, for $v_h \in \widetilde{V}_h$ and $p_h \in \widetilde{Y}_h$, (p_h) is relatively compact in H^1_0.

Proof. This follows from the uniform continuity and pointwise convergence in H^1_0 of the Galerkin projectors onto $\widetilde{Y}_h$ with respect to the bilinear form with expression (3.1). □

5. On discrete compactness. In this section we provide results on the relationship between the estimates of this paper and some more well-known ones.

First we show that (UNE) implies a *gap property* of discrete Helmholtz or Hodge decompositions, which in turn implies a *discrete compactness property* of edge elements in the sense of Kikuchi [27] (with natural rather than essential boundary conditions for the curl operator). The gap property proved useful in analyzing some integral equations describing electromagnetic phenomena (Christiansen [13] and Buffa and Christiansen [11]) whereas the discrete compactness property is crucial in the analysis of discrete eigenvalue problems of electromagnetics (see in particular Boffi, Brezzi, and Gastaldi [4] and Boffi, Fernandes, and Gastaldi [5]).

Then we show that a weakened (local) version of discrete compactness can be deduced from the proposed discrete div-curl lemma (without reference to (UNE) or (SA)).

Suppose G is a Hilbert space such that $\mathrm{L}^2 \subset G \subset \mathrm{H}^{-1}$ with continuous inclusions. Suppose furthermore that the injection $G \to \mathrm{H}^{-1}$ is compact.

PROPOSITION 5.1. *Suppose* (X_h) *satisfies* (UNE). *Then we have*

$$\lim_{h \to 0} \sup_{v_h \in V_h} \|v_h - P_V v_h\|_0 / \|\operatorname{curl} v_h\|_G = 0. \tag{5.1}$$

Proof. Indeed if this were not true we would have a subsequence (v_h) with $v_h \in V_h$ such that $(\|\operatorname{curl} v_h\|_G)$ is bounded and $(\|v_h - P_V v_h\|_0)$ does not converge to 0. From it we can extract a subsequence such that $(\operatorname{curl} v_h)$ converges weakly in G. But then $(\operatorname{curl} v_h)$ converges strongly in H^{-1}. One obtains a contradiction with Proposition 3.2. □

Consider the Hilbert space X^G of vector fields in L^2 with curl in G equipped with its natural norm $\|\cdot\|_{0G}$. In X^G let V^G be the L^2-orthogonal of (the closed subspace) $W = \operatorname{grad} \mathrm{H}^1$ and let $\delta_G(\cdot,\cdot)$ denote the gap between subspaces in the sense of Kato [26]. Thus for nonzero subspaces U, U' we have

$$\delta_G(U, U') = \sup_{u \in U, \|u\|=1} \inf_{u' \in U'} \|u - u'\|_{0G}. \tag{5.2}$$

COROLLARY 5.2. *Suppose* (X_h) *satisfies estimate* (5.1). *Then we have*

$$\lim_{h \to 0} \delta_G(V_h, V^G) = 0. \tag{5.3}$$

Proof. Just remark that

$$\|v_h - P_V v_h\|_{0G} = \|v_h - P_V v_h\|_0 \quad \text{and} \quad \|\operatorname{curl} v_h\|_G \le \|v_h\|_{0G}, \tag{5.4}$$

and conclude using estimate (5.1). □

From this gap property in the case $G = \mathrm{H}^0 = \mathrm{L}^2$ one easily deduces the following discrete compactness property. Let X^0 denote the space of vector fields u in L^2 such that $\operatorname{curl} u$ is in L^2. Let V^0 denote the L^2-orthogonal of the closed subspace

$W = \operatorname{grad} \mathrm{H}^1$ of X^0. The elements of V^0 thus have vanishing divergence and vanishing normal component on $\partial\Omega$. By a result of Weber, for which we refer to Theorem 2.8 in [1], the space V^0, when equipped with the norm inherited from X^0, is compactly embedded in L^2. This result has the following discrete analogue.

Corollary 5.3. *Suppose* (X_h) *satisfies the gap property* (5.3) *and that* (v_h) *is a sequence of vector fields* $v_h \in V_h$*, which is bounded in* X^0*. Then there is a subsequence which converges in* L^2*.*

Proof. If (v_h) is such a sequence, then $(P_V v_h)$ is bounded in V^0 and $(\|v_h - P_V v_h\|_0)$ tends to 0. □

Concerning directly the relationship between Theorem 4.1 and the discrete compactness, we make the following remark.

Proposition 5.4. *Suppose that* (X_h) *is a family of subspaces (not necessarily satisfying* (UNE) *or* (SA) *a priori) for which the conclusions of Theorem* 4.1 *(resp., Corollary* 4.2*) are true. Then for each sequence* (u_h) *of vector fields* $u_h \in \tilde{V}_h$ *which is bounded in* L^2 *and such that* $(\operatorname{curl} u_h)$ *is relatively compact in* H^{-1} *(resp., bounded in* L^2*), there is a subsequence converging in* L^2_{loc}*.*

Proof. Indeed we can extract a subsequence converging weakly in L^2 to some u. Then we have, for all $\phi \in \mathcal{D}$,

$$\int |u_h|^2 \phi \to \int |u|^2 \phi. \tag{5.5}$$

Following standard procedure we deduce

$$\int |u_h - u|^2 \phi = \int |u_h|^2 \phi - 2 \int (u_h \cdot u)\phi + \int |u|^2 \phi \to 0, \tag{5.6}$$

which is a characterization of L^2_{loc} convergence. □

In other words, the above local discrete compactness property can be viewed as the special case of the discrete div-curl lemma where the sequences (u_h) and (u'_h) are equal. Furthermore, we remark that in the absence of boundary conditions on u_h one should not expect to be able to strengthen convergence from L^2_{loc} to L^2 in this result; see Proposition 2.7 in [1].

6. Discussion.

We close this paper with some remarks.

Of course, if an entire physical field is to be approximated numerically, a scheme which is only vaguely converging is of little practical use. Indeed the consensus seems to be to aim for rapid convergence in the energy norm. However, a method which satisfies a given property (vague convergence, for instance) under *the weakest of assumptions* shows a sign of *robustness*, in the sense that under extreme conditions it performs as well as one can reasonably hope for. Therefore the above result can be interpreted as a robustness property of edge elements with respect to a class of nonlinearities.

It is sometimes useful to state properties of discretized electromagnetic fields in terms of (nonsmooth) differentiable forms, a point of view developed by Bossavit [7] (see Hiptmair [23] for more on this). Likewise the div-curl lemma of Murat and Tartar can be stated in terms of differential forms (see, e.g., Taylor [38, pp. 358–359]). The translation of the results of this paper into this framework is unproblematic. A case of particular interest is when considering div conforming face elements [34, 10]. One obtains a similar div-curl lemma for face elements based this time on the discrete information on the curl (obtained by integrating against the curl of the naturally associated edge element space).

The question whether L^2-bounded sequences in $\widetilde{V}_h$ (discrete vector fields which are discrete divergence-free) have relatively compact divergence in H^{-1} is as of today unsettled for the standard FE spaces and, as mentioned in the introduction, the present paper can be seen as a circumvention of this question. The affirmative would imply a slightly stronger version of Theorem 4.1 (the condition $u'_h \in X_h$ can then be dropped).

The div-curl lemma is closely related to Hardy space estimates (see, e.g., Coifman et al. [17]) and a translation of the above results into uniform Hardy space estimates for edge elements could be an important step forward. More generally, this lemma is the prototype for so-called bilinear estimates which play a prominent role in the analysis of many nonlinear PDE. The original motivation for the present work was to develop some tools for the numerical analysis of the discretization of Yang–Mills equations [14], where consequences of weak divergence control are investigated.

REFERENCES

[1] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[2] D. N. Arnold, *Differential complexes and numerical stability*, in Proceedings of the International Congress of Mathematicians, Vol. 1, Higher Ed. Press, Beijing, 2002, pp. 137–157.

[3] I. Babuska, *Error bounds for the finite element method*, Numer. Math., 16 (1971), pp. 322–333.

[4] D. Boffi, F. Brezzi, and L. Gastaldi, *On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form*, Math. Comp., 69 (1999), pp. 121–140.

[5] D. Boffi, P. Fernandes, L. Gastaldi, and I. Perugia, *Computational models of electromagnetic resonators: Analysis of edge element approximation*, SIAM J. Numer. Anal., 36 (1999), pp. 1264–1290.

[6] D. Boffi, *A note on the DeRham complex and a discrete compactness property*, Appl. Math. Lett., 14 (2001), pp. 33–38.

[7] A. Bossavit, *Mixed finite elements and the complex of Whitney forms*, in The Mathematics of Finite Elements and Applications, VI, Uxbridge, 1987, Academic Press, London, 1988, pp. 137–144.

[8] D. Braess, *Finite Elements. Theory, Fast Solvers and Applications in Solid Mechanics*, 2nd ed., Cambridge University Press, Cambridge, 2001.

[9] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers*, RAIRO Anal. Numér., 8 (1974), pp. 129–151.

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

[11] A. Buffa and S. H. Christiansen, *The electric field integral equation on Lipschitz screens: Definitions and numerical approximation*, Numer. Math., 94 (2003), pp. 229–267.

[12] M. Cessenat, *Mathematical Methods in Electromagnetism, Linear Theory and Applications*, World Scientific, River Edge, NJ, 1996.

[13] S. H. Christiansen, *Discrete Fredholm properties and convergence estimates for the electric field integral equation*, Math. Comp., 73 (2004), pp. 143–167.

[14] S. H. Christiansen and R. Winther, *On constraint preservation in numerical simulations of Yang–Mills equations*, preprint, Pure Mathematics, University of Oslo, 33, 2004, http://www.math.uio.no/div/eprint/pure_math/2004/33-04.html.

[15] S. H. Christiansen and J.-C. Nédélec, *A preconditioner for the electric field integral equation based on Calderon formulas*, SIAM J. Numer. Anal., 40 (2002), pp. 1100–1135.

[16] P. Ciarlet Jr. and J. Zou, *Fully discrete finite element approaches for time-dependent Maxwell's equations*, Numer. Math., 82 (1999), pp. 193–219.

[17] R. Coifman, P.-L. Lions, Y. Meyer, and S. Semmes, *Compensated compactness and Hardy spaces*, J. Math. Pures Appl. (9), 72 (1993), pp. 247–286.

[18] M. Costabel and M. Dauge, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276.

[19] F. Dubois, *Discrete vector potential representation of a divergence-free vector field in three-dimensional domains: Numerical analysis of a model problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1103–1141.

[20] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*, 2nd ed., John Wiley & Sons, New York, 1999.

[21] V. Girault, *Incompressible finite element methods for Navier-Stokes equations with nonstandard boundary conditions in* $\mathbb{R}^3$, Math. Comp., 51 (1988), pp. 55–74.

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

[23] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.

[24] J. D. Jackson, *Classical Electrodynamics*, 2nd ed. John Wiley & Sons, New York, 1975.

[25] P. Joly, *Variational methods for time-dependent wave propagation problems*, in Topics in Computational Wave Propagation, Lect. Notes Comput. Sci. Eng. 31, Springer-Verlag, Berlin, 2003, pp. 201–264.

[26] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer-Verlag, Berlin, 1976.

[27] F. Kikuchi, *On a discrete compactness property for the Nédélec finite elements*, J. Fac. Sci. Univ. Tokyo, Sect. IA Math., 36 (1989), pp. 479–490.

[28] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, New York, 2003.

[29] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa. Cl. Sci. (4), 5 (1978), pp. 485–507.

[30] J. Nečas, *Les Méthodes Directes en Théorie Des Équations Elliptiques*, Masson, Paris, 1967.

[31] J.-C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[32] J.-C. Nédélec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 50 (1986), pp. 57–81.

[33] J.-C. Nédélec, *Acoustic and Electromagnetic Equations, Integral Representations for Harmonic Problems*, Springer-Verlag, New York, 2001.

[34] P. A. Raviart and J.-M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, Berlin, 1977, pp. 292–315.

[35] G. De Rham, *Variétés Différentiables. Formes, Courants, Formes Harmoniques*, 3rd ed., Hermann, Paris, 1973.

[36] W. Rudin, *Functional Analysis*, 2nd ed., McGraw–Hill, New York, 1991.

[37] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, Res. Notes in Math. 39, Pitman, Boston, 1979, pp. 136–212.

[38] M. Taylor, *Partial Differential Equations, Vol.* I*, Basic Theory*, Springer-Verlag, New York, 1996.

[39] L. B. Wahlbin, *Superconvergence in Galerkin Finite Element Methods*, Lecture Notes in Math. 1605, Springer-Verlag, Berlin, 1995.

[40] H. Whitney, *Geometric Integration Theory*, Princeton University Press, Princeton, NJ, 1957.

SIAM J. Numer. Anal.
Vol. 43, No. 1, pp. 127–155

hp-INTERPOLATION OF NONSMOOTH FUNCTIONS AND AN APPLICATION TO *hp*-A POSTERIORI ERROR ESTIMATION[*]

J. M. MELENK[†]

Abstract. The quasi-interpolation operators of Clément and Scott–Zhang type are generalized to the setting of the *hp*-FEM (finite element method). New polynomial lifting and inverse estimates are presented. The classical residual based a posteriori error estimator is generalized to the *hp*-FEM.

Key words. Clément interpolant, quasi interpolation, *hp*-FEM, polynomial inverse estimates, a posteriori error estimation

AMS subject classifications. 65N30, 65N35, 65N50

DOI. 10.1137/S0036142903432930

1. Introduction.

Quasi-interpolation operators, that is, operators achieving optimal rates of convergence also for classes of functions of low regularity have a long history, for example, in spline theory (see, e.g., [24] for an overview). In connection with the finite element method (FEM) such an operator was constructed by Clément in [22], where he showed how H^1-functions can be approximated by piecewise linear functions. Subsequent refinements and variations include [5,13,16,20,21,27,36,38] to account for higher order polynomials of fixed degree p, preservation of piecewise polynomial boundary conditions, curvilinear elements, Hermite elements, and anisotropic elements. Several of these refinements were done with a view to an application in residual-based finite element error estimation as discussed in the monographs [4,7,40].

While quasi interpolation in the context of the h-version FEM is well documented in the literature, the situation is less favorable for the p-version and particularly the hp-version of the FEM, where the approximation properties of spaces of piecewise polynomials are quantified in terms of both the local mesh size and the local polynomial degree. The one-dimensional situation of polynomial approximation on an interval has been thoroughly studied, and we refer the reader to [24] for an excellent exposition of pertinent results. In higher dimensions, the situation is somewhat less developed: Approximation results suitable for the application to the p-version FEM/spectral method in higher dimensions can be found in the survey article [17] (for L^2-based weighted and unweighted Sobolev spaces on tensor product domains) and [8,37] (likewise for L^2-based Sobolev spaces), where, however, extra regularity of the function to be approximated is assumed, namely, that it be in the Sobolev spaces H^s for some $s > d/2$, where $d \in \mathbb{N}$ is the spatial dimension. Quasi-interpolation operators making minimal regularity assumption have been constructed in [2,3] (for the Sobolev spaces $W^{k,q}$); their approximation results include, however, logarithmic factors if $q \neq 2$.

In the present paper, we develop optimal quasi-interpolation operators suitable for an application in the framework of the hp-version of the FEM. We exhibit two kinds of closely related operators: Clément-type operators (see Theorem 3.1) defined on the

[*] Received by the editors August 6, 2003; accepted for publication (in revised form) October 19, 2004; published electronically April 26, 2005.

http://www.siam.org/journals/sinum/43-1/43293.html

[†] MPI für Mathematik in den Naturwissenschaften, Inselstr. 22-26, D-04103 Leipzig, Germany. Current address: Department of Mathematics, The University of Reading, PO Box 220, Reading RG6 6AX, United Kingdom (j.m.melenk@reading.ac.uk).

space L^1 and Scott/Zhang-type operators (see Theorems 3.3, 3.4) defined on $W^{1,q}$ (so that traces on the boundary are defined) that preserve piecewise polynomial boundary conditions. Both operators achieve optimal rates of convergence. The paper restricts itself to problems in $\mathbb{R}^2$. This restriction is largely due to the way the operators that preserve piecewise polynomial boundary conditions are constructed, that is, to the method of the proof of Theorems 3.3 and 3.4. Theorem 3.1 can readily be extended to higher dimensions.

A particular application of the operators developed in the present paper is that they permit the extension of the h-FEM residual-based error estimation to the hp-FEM [34]. We illustrate the salient features in section 4.

This paper is organized as follows: In section 2, we introduce the necessary notation, in particular γ-shape regular triangulations of two-dimensional domains and the hp-FEM spaces of piecewise mapped polynomials. We emphasize that the element maps need not be affine, which is an important aspect in hp-FEM, and that variable approximation order is considered. Section 3, which is the heart of our paper, presents the quasi-interpolation operators. Section 4 illustrates how the quasi-interpolation operators of section 3 can be employed for reliable a posteriori error estimation. Finally, section 5 is devoted to the proof of the approximation properties of our quasi-interpolation operators. Since polynomial approximation of $W^{k,q}$-functions on reference configurations and polynomial liftings rely on fairly technical constructions, several of these constructions are relegated to the appendix.

2. Notation and assumptions.

We will denote by $\mathbb{N} = \{1, 2, \dots\}$ the positive integers; $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ denotes the nonnegative integers.

2.1. Triangulations.

We start with the standard definitions of meshes and triangulations for two-dimensional domains.

A triangulation $\mathcal{T}$ of a set $\Omega \subset \mathbb{R}^2$ is a collection of elements $K \in \mathcal{T}$; associated with each element K is an element map $F_K : \widehat{K} \to K$, where the reference element $\widehat{K}$ corresponding to K is either the reference square $S = (0,1)^2$ or the reference triangle

$$T = \{(x,y) \in \mathbb{R}^2 \,|\, 0 < x < 1,\ 0 < y < \min(x, 1-x)\}. \tag{2.1}$$

We consider triangulations that satisfy the following standard conditions:

(M1) The element maps $F_K : \widehat{K} \to K = F_K(\widehat{K})$ are C^1-diffeomorphisms between $\overline{\widehat{K}}$ and $\overline{K}$; i.e., there exist domains $\widehat{K}'$ and K' with $\overline{\widehat{K}} \subset \widehat{K}'$, $\overline{K} \subset K'$ such that F_K is in fact a C^1-diffeomorphism between $\widehat{K}'$ and K'.

(M2) For two elements K, K' the intersection $\Gamma := \overline{K} \cap \overline{K'}$ falls into exactly one of the following categories: Γ is empty, or a vertex, or a whole edge, or K and K' coincide (i.e., $F_K^{-1}(\Gamma)$ and $F_{K'}^{-1}(\Gamma)$ are edges, or vertices of the corresponding reference elements $\widehat{K}$, $\widehat{K}'$). Additionally, we require the map

$$Q : F_K^{-1}(\Gamma) \to F_{K'}^{-1}(\Gamma) : \quad x \mapsto (F_{K'}^{-1} \circ F_K)(x)$$

to be an affine homeomorphism.

(M3) $\Omega \setminus \cup_{K \in \mathcal{T}}$ is a set of Lebesgue measure zero.

A triangulation $\mathcal{T}$ is called γ-shape regular if additionally

$$h_K^{-1} \|F_K'\|_{L^\infty(\widehat{K})} + h_K \| \left(F_K'\right)^{-1} \|_{L^\infty(\widehat{K})} \le \gamma, \tag{2.2}$$

where $h_K = \operatorname{diam} K$. We say that the triangulation is *affine* if all element maps F_K are affine maps. The restriction $\mathcal{T}|_\omega$ denotes the subset of $\mathcal{T}$ that represents the triangulation of $\omega \subset \Omega$ satisfying (M1)–(M3).

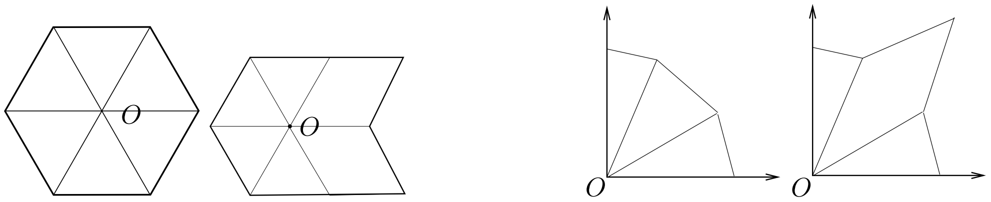


FIG. 2.1. *Left: interior reference patches $\omega_{int,6,1}$ and $\omega_{int,6,j}$ for some $j \in \{2,\dots,2^6\}$. Right: boundary reference patches $\omega_{bdy,3,1}$ and $\omega_{bdy,3,j}$ for some $j \in \{2,\dots,2^3\}$.*

For each element $K \in \mathcal{T}$ we denote by $\mathcal{E}(K)$ the set of edges of K and by $\mathcal{N}(K)$ the set of vertices of K. Similarly, $\mathcal{N}(\mathcal{T})$ denotes the set of all vertices of $\mathcal{T}$ and $\mathcal{E}(\mathcal{T})$ the set of all edges. Setting

$$\hat{I} = (0,1)$$

the assumption (M2) implies that we can define *edge maps* $F_e : \hat{I} \to e$ for each $e \in \mathcal{E}(\mathcal{T})$ by taking an element K such that e is an edge of K, then identifying the edge $F_K^{-1}(e)$ of $\hat{K}$ with $\hat{I}$ via an affine map, and finally taking F_e as the restriction of F_K to $F_K^{-1}(e)$; the assumption (M2) guarantees that the map F_e obtained in this way is independent of the choice of K. Additionally, we introduce the notion of the *patch* ω_V associated with a node $V \in \mathcal{N}(\mathcal{T})$ by

$$\omega_V := \{x \in \Omega \,|\, x \in \overline{K} \quad \text{for some } K \text{ with } V \in \overline{K}\}^\circ, \tag{2.3}$$

where A° denotes the interior of the set A. We note that the patches ω_V are open subsets of Ω. Of importance will be the connectivity of the patches. Our tool for classifying patches according to their connectivity will be the notion of *reference patches* that we make precise in the following definition.

DEFINITION 2.1 (reference patch). *Reference patches are Lipschitz domains that are either labeled* interior *or* boundary *patches. They are characterized as follows:*

1. Interior patches: *For each $M \in \mathbb{N}$, $M \geq 3$, we define 2^M* interior reference patches *$\omega_{int,M,j}$, $j = 1,\dots,2^M$, as follows: $\omega_{int,M,1}$ is defined to be the regular polygon with M edges of length 1 that is centered at the origin $0 \in \mathbb{R}^2$ and is triangulated with M triangles all sharing the vertex 0. The remaining $2^M - 1$ reference patches are obtained from this one by replacing one or several of these isosceles triangles by parallelograms (see Figure 2.1).*
2. Boundary patches: *For each $M \in \mathbb{N}$ we define 2^M* boundary reference patches *$\omega_{bdy,M,j}$, $j = 1,\dots,2^M$, in the following way: $\omega_{bdy,M,1} \subset \{(x,y) \,|\, x > 0, y > 0\}$ is the polygon that consists of M isosceles triangles all sharing the vertex $0 \in \mathbb{R}^2$ and having angle $\pi/(2M)$ at 0. The remaining $2^M - 1$ patches are obtained from this one by replacing one or several of these isosceles triangles by parallelograms (see Figure 2.1).*

We will only consider triangulations whose patches can be related to these reference patches:

(M4) For each vertex $V \in \mathcal{N}(\mathcal{T})$ there exists a reference patch $\widehat{\omega}_V$ of the form given in Definition 2.1 together with a homeomorphism $F_V : \widehat{\omega}_V \to \omega_V$ with $F_V(0) = V$, which has the form

$$F_V^{-1}|_K = A_{K,V} \circ F_K^{-1} \qquad \forall K \in \mathcal{T}|_{\omega_V},$$

where the maps $A_{K,V} : \mathbb{R}^2 \to \mathbb{R}^2$ are *affine.*

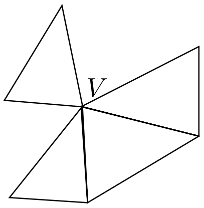


FIG. 2.2. *Example of a mesh excluded by* (M4).

Remark 2.2. It is worth pointing out that slit domains are not excluded by (M4). However, other kinds of domains that fail to be Lipschitz domain are not covered by the present results: For example, domains such as the one depicted in Figure 2.2 are not admitted since the vertex cannot be mapped to a boundary reference patch in the way condition (M4) requires.

We finish this subsection by noting that the γ-shape regularity of the element maps implies that only a finite number of elements can meet at a vertex.

LEMMA 2.3. *Let $\mathcal{T}$ be a γ-shape regular triangulation satisfying* (M1)–(M3). *Then there exists a constant $M \in \mathbb{N}$, which depends only on γ, such that*

1. *no more than M elements share a common vertex;*
2. *for any two elements K, K' with $\overline{K} \cap \overline{K'} \neq \emptyset$ there holds $M^{-1}h_K \leq h_{K'} \leq Mh_K$.*

If the triangulation satisfies additionally (M4), *then the maps $A_{K,V}$ appearing in condition* (M4) *satisfy*

$$\|A'_{K,V}\|_{L^\infty(\hat{K})} + \|A'^{-1}_{K,V}\|_{L^\infty(\hat{K})} \leq C$$

for some $C > 0$ that depends only on γ. Additionally, $F_V \in W^{1,\infty}(\widehat{\omega}_V)$ and $F_V^{-1} \in W^{1,\infty}(\omega_V)$, and we have the bound

$$h_V^{-1}\|F_V'\|_{L^\infty(\widehat{\omega}_V)} + h_V\left\|\left(F_V'\right)^{-1}\right\|_{L^\infty(\widehat{\omega}_V)} \leq C, \qquad h_V = \min_{K:V\in\mathcal{N}(K)} h_K$$

for some $C > 0$ depending solely on γ.

Proof.

Step 1. The element maps F_K are C^1 up to the boundary of the reference elements. The fact that the interior angles of the reference elements are nondegenerate and the γ-shape regularity assumption (2.2) then imply that the interior angles of elements $K \in \mathcal{T}$ are within $(\varepsilon, \pi - \varepsilon)$ for an $\varepsilon > 0$, which depends solely on γ. The first claim of the lemma then follows if we choose $M \in \mathbb{N}$ such that $M \geq 2\pi/\varepsilon$.

Step 2. The γ-shape regularity assumption (2.2) also implies the existence of $C > 0$ depending solely on γ such that

$$C^{-1}h_K \leq |e| \leq Ch_K \qquad \forall e \in \mathcal{E}(K) \qquad \forall K \in \mathcal{T}.$$

This fact together with the observation of the first step easily implies the second claim after appropriately adjusting the constant M.

Step 3. We will only show that $F_V^{-1} \in W^{1,\infty}(\omega_V)$ with the corresponding bound for the derivative. By assumption $F_V^{-1}|_K \in C^1(\overline{K})$ for each element $K \in \mathcal{T}|_{\omega_V}$. Then an elementwise integration by parts together with the observation $F_V^{-1} \in C(\omega_V)$ implies that the weak derivative is elementwise given by $(F_V^{-1})'|_K = A_{K,V} \cdot (F_K^{-1})'$. From this representation, we readily infer $F_V^{-1} \in W^{1,\infty}(\omega_V)$ and the desired bound. □

2.2. Polynomial spaces. The finite element spaces that we consider are the variable order piecewise mapped polynomials, an early implementation of which is discussed in [23]: For each element $K \in \mathcal{T}$, we choose a polynomial degree $p_K \in \mathbb{N}$ and collect these numbers in the polynomial degree vector $\mathbf{p} = (p_K)_{K\in\mathcal{T}}$. We then define the space $S^{\mathbf{p}}(\mathcal{T}) \subset W^{1,\infty}(\Omega)$ by

$$S^{\mathbf{p}}(\mathcal{T}) = \{u \in C(\Omega) \,|\, u|_K \circ F_K \in \Pi_{p_K}(\widehat{K})\}, \tag{2.4}$$

where we set

$$\Pi_p(\widehat{K}) = \begin{cases} \mathcal{P}_p := \operatorname{span}\{x^i y^j \,|\, 0 \le i+j \le p\} & \text{if } \widehat{K} = T, \\ \mathcal{Q}_p := \operatorname{span}\{x^i y^j \,|\, 0 \le i,j \le p\} & \text{if } \widehat{K} = S. \end{cases} \tag{2.5}$$

We will write $S^p(\mathcal{T})$ if the degree vector $\mathbf{p}$ satisfies $p_K = p$ for all $K \in \mathcal{T}$. In this case, we will also permit the choice $p = 0$, where $S^0(\mathcal{T})$ reduces to a one-dimensional space.

A key property of the spaces $S^{\mathbf{p}}(\mathcal{T})$ is that we can identify "nodal shape functions" that form a partition of unity; i.e., for each vertex $V \in \mathcal{N}(\mathcal{T})$, we can find a function $\varphi_V \in S^1(\mathcal{T})$ such that

$$\varphi_V|_{\Omega\setminus\omega_V} \equiv 0 \qquad \text{and} \qquad \sum_{V\in\mathcal{N}(\mathcal{T})} \varphi_V \equiv 1 \quad \text{on } \Omega. \tag{2.6}$$

A well-known consequence of the γ-shape regularity of the triangulation is that these nodal shape functions satisfy for some constant $C > 0$, which depends solely on γ,

$$\|\varphi_V\|_{L^\infty(\Omega)} \le 1, \qquad \|\nabla\varphi_V\|_{L^\infty(\Omega)} \le C h_K^{-1} \qquad \forall K \in \mathcal{T}|_{\omega_V}. \tag{2.7}$$

In the present paper we consider only γ-shape regular triangulations. Such triangulations have the property that neighboring elements are comparable in size (cf. Lemma 2.3). We impose a similar condition on the polynomial degree distribution:

$$\gamma^{-1} p_K \le p_{K'} \le \gamma p_K \qquad \forall K, K' \in \mathcal{T} \quad \text{such that (s.t.) } \overline{K} \cap \overline{K'} \ne \emptyset. \tag{2.8}$$

We will also employ the notation

$$p_V := \min\{p_K \,|\; V \in \mathcal{N}(K)\}, \qquad p_e := \min\{p_K \,|\; e \in \mathcal{E}(K)\}. \tag{2.9}$$

2.3. Notation for Sobolev spaces. For domains $\Omega \subset \mathbb{R}^2$ and $k \in \mathbb{N}_0$, $q \in [1,\infty]$ we employ standard Sobolev spaces $W^{k,q}(\Omega)$ as described in, e.g., [1]. For the reference interval $\hat{I} = (0,1)$, $\kappa \in (0,1)$, and $q \in [1,\infty)$, we equip the space $W^{\kappa,q}(\hat{I})$ with the Slobodeckij norm

$$\|u\|^q_{W^{\kappa,q}(\hat{I})} = \|u\|^q_{L^q(\hat{I})} + \int_{\hat{I}}\int_{\hat{I}} \frac{|u(x) - y(y)|^q}{|x-y|^{1+q\kappa}}\,dx\,dy. \tag{2.10}$$

We will also require the spaces $\widetilde{W}^{\kappa,q}(\hat{I})$, which consist of the functions $u \in W^{\kappa,q}(\hat{I})$ such that their trivial extension (i.e., by zero) to $\mathbb{R}$ is an element of $W^{\kappa,q}(\mathbb{R})$. This space is equipped with the norm

$$\|u\|^q_{\widetilde{W}^{\kappa,q}(\hat{I})} = \|u\|^q_{W^{\kappa,q}(\hat{I})} + \int_0^1 \frac{|u(x)|^q}{x^{\kappa q}}\,dx + \int_0^1 \frac{|u(x)|^q}{(1-x)^{\kappa q}}\,dx. \tag{2.11}$$

In analogy with the spaces $\widetilde{W}^{\kappa,q}(\hat{I})$ we can define the spaces $\widetilde{W}_l^{\kappa,q}(\hat{I})$ if the trivial extension to $I' = \{x \in \mathbb{R} \,|\, x < 1\}$ is in $W^{\kappa,q}(I')$. This space is equipped with the norm

$$\|u\|^q_{\widetilde{W}_l^{\kappa,q}(\hat{I})} = \|u\|^q_{W^{\kappa,q}(\hat{I})} + \int_0^1 \frac{|u(x)|^q}{x^{\kappa q}}\,dx. \tag{2.12}$$

We also note how functions transform under concatenation with the patch maps F_V.

Lemma 2.4. *Let $\mathcal{T}$ be a γ-shape regular triangulation satisfying* (M1)–(M4). *Let $q \in [1,\infty]$. Then for every patch ω_V, $V \in \mathcal{N}(\mathcal{T})$, and every $u \in W^{1,q}(\omega_V)$ we have that $\widehat{u} := u \circ F_V \in W^{1,q}(\widehat{\omega}_V)$ and*

$$\|\widehat{u}\|_{L^q(\widehat{\omega}_V)} \sim h_V^{-2/q}\|u\|_{L^q(\omega_V)}, \qquad \|\nabla\widehat{u}\|_{L^q(\widehat{\omega}_V)} \sim h_V^{1-2/q}\|\nabla u\|_{L^q(\omega_V)}, \tag{2.13}$$

where $h_V = \min_{K:K\subset\omega_V} h_K$. The constants hidden in the $\sim$-notation depend solely on γ and q.

Proof. We claim that the pull-back $\widehat{u}$ is in $W^{1,q}(\widehat{\omega}_V)$. To see this, we first consider the case $q < \infty$. For each element K of the patch ω_V and its corresponding element $K' := F_V^{-1}(K) \subset \widehat{\omega}_V$, the assumption (M1) guarantees that $F_V|_{K'} \in C^1(\overline{K'})$ and likewise $F_V^{-1} \in C^1(\overline{K})$. Hence by standard properties of Sobolev space (see, e.g., [1, Chap. III, Thm. 3.35]) we have for each element K that $u\circ F_V|_{K'} \in W^{1,q}(K')$ and the derivative satisfies $(\nabla(u\circ F_V))|_{K'} = (\nabla u \circ F_V)F_V'$. In order to see that $u\circ F_V$ is in $W^{1,q}(\widehat{\omega}_V)$ we have to check that the traces on the edges shared by two elements K_1, K_2 of $\widehat{\omega}_V$ coincide. This follows easily from the assumption (M3). The case $q = \infty$ is obtained by inspection: Since the weak derivative has been identified as $(\nabla u \circ F_V)F_V'$, one merely has to check that it is in $L^\infty(\widehat{\omega}_V)$, which is indeed the case. The bounds (2.13) now follow from (2.2). □

3. Quasi interpolation of nonsmooth functions.

We present two types of quasi-interpolation results for $W^{1,q}$-functions: In Theorem 3.1 we exhibit a quasi-interpolation operator of Clément type; in Theorem 3.3 we present an operator that additionally preserves homogeneous boundary conditions that may be imposed on parts of the boundary. This latter operator is generalized in Theorem 3.4 to an operator that preserves arbitrary piecewise polynomial Dirichlet boundary conditions.

In order to formulate these results, we introduce the following additional notation: For $e \in \mathcal{E}(\mathcal{T})$ we denote by $\mathcal{N}(e)$ the two endpoints of e, i.e., $\mathcal{N}(e) = \{V \in \mathcal{N}(\mathcal{T}) \,|\, V \in \overline{e}\}$. Patches of order $j \in \mathbb{N}$ associated with an element $K \in \mathcal{T}$ or an edge $e \in \mathcal{E}(\mathcal{T})$ are defined thus:

$$\omega_e^1 := \bigcup_{V\in\mathcal{N}(e)} \omega_V, \qquad \omega_e^{j+1} := \bigcup_{V\in\mathcal{N}(\mathcal{T}):V\in\overline{\omega_e^j}} \omega_V, \quad j = 1,2,\ldots, \tag{3.1}$$

$$\omega_K^1 := \bigcup_{V\in\mathcal{N}(K)} \omega_V, \qquad \omega_K^{j+1} := \bigcup_{V\in\mathcal{N}(\mathcal{T}):V\in\overline{\omega_K^j}} \omega_V, \quad j = 1,2,\ldots. \tag{3.2}$$

3.1. Clément-type approximation.

Quasi interpolation of Clément type takes the following form.

Theorem 3.1 (Clément-type quasi interpolation). *Let $\mathcal{T}$ be a γ-shape regular triangulation of a domain $\Omega \subset \mathbb{R}^2$ satisfying* (M1)–(M4) *and let $\mathbf{p}$ be a polynomial degree distribution satisfying* (2.8). *Then there exists a bounded linear operator I^{hp} : $L^1(\Omega) \to S^{\mathbf{p}}(\mathcal{T}) \subset L^1(\Omega)$, and there exists a constant $C > 0$, which depends solely on*

$q \in [1, \infty]$ and γ, such that for every $u \in W^{1,q}(\Omega)$ and all elements $K \in \mathcal{T}$ and all edges $e \in \mathcal{E}(\mathcal{T})$

$$\|u - I^{hp}u\|_{L^q(K)} + \frac{h_K}{p_K}\|\nabla(u - I^{hp}u)\|_{L^q(K)} \le C\frac{h_K}{p_K}\|\nabla u\|_{L^q(\omega_K^1)}, \tag{3.3}$$

$$\|u - I^{hp}u\|_{L^q(e)} \le C\left(\frac{h_e}{p_e}\right)^{1-1/q}\|\nabla u\|_{L^q(\omega_e^1)}. \tag{3.4}$$

Proof. The proof can be found in section 5.2. □

3.2. Scott–Zhang-type approximation. The operator I^{hp} of Theorem 3.1 does not preserve piecewise polynomial boundary conditions if applied to functions of $W^{1,q}(\Omega)$. The operators of Theorem 3.1 can, however, be modified to have this property.

Let a set $\mathcal{B} \subset \mathcal{E}(\mathcal{T})$ of boundary edges of the triangulation $\mathcal{T}$ be given, i.e.,

$$\mathcal{B} \subset \mathcal{E}(\mathcal{T}) \quad \text{and} \quad b \subset \partial\Omega \qquad \forall b \in \mathcal{B}. \tag{3.5}$$

Next, we define for $q \in (1, \infty)$ the spaces

$$W^{1,q}_{\mathcal{B},0} := \{u \in W^{1,q}(\Omega) \,|\, u|_b = 0 \text{ for all } b \in \mathcal{B}\}, \tag{3.6}$$

$$W^{1,q}_{\mathcal{B},\mathbf{p}} := \{u \in W^{1,q}(\Omega) \,|\, u|_b \circ F_b \in \mathcal{P}_{p_b} \text{ for all } b \in \mathcal{B} \text{ and (3.8) holds}\}, \tag{3.7}$$

where the continuity condition (3.8) is

$$\text{for all } b, b' \in \mathcal{B} \text{ and } V \in \mathcal{N}(b) \cap \mathcal{N}(b') \text{ there holds } \lim_{\substack{x\to V\\ x\in b}} u(x) = \lim_{\substack{x\to V\\ x\in b'}} u(x). \tag{3.8}$$

Remark 3.2. Since the edges of $\mathcal{B}$ are part of the boundary of $\partial\Omega$, the function values are understood in the sense of traces. In the case of slit domains appropriate limits have to be taken.

We then have the following approximation results.

THEOREM 3.3 (homogeneous boundary conditions). *Let $\mathcal{T}$ be a γ-shape regular triangulation of a domain $\Omega \subset \mathbb{R}^2$ satisfying* (M1)–(M4)*. Let $\mathbf{p}$ be a polynomial degree distribution satisfying* (2.8)*. Let $q \in (1, \infty)$ and a set $\mathcal{B} \subset \mathcal{E}(\mathcal{T})$ of boundary edges be given. Then there exists a linear operator $I^{hp}_{hom} : W^{1,q}_{\mathcal{B},0}(\Omega) \to S^{\mathbf{p}}(\mathcal{T}) \cap W^{1,q}_{\mathcal{B},0}(\Omega)$, and there exists a constant $C > 0$ depending solely on γ and q such that*

$$\|u - I^{hp}_{hom}u\|_{L^q(K)} + \frac{h_K}{p_K}\|\nabla(u - I^{hp}_{hom}u)\|_{L^q(K)} \le C\frac{h_K}{p_K}\|\nabla u\|_{L^q(\omega_K^1)}, \tag{3.9}$$

$$\|u - I^{hp}_{hom}u\|_{L^q(e)} \le C\left(\frac{h_e}{p_e}\right)^{1-1/q}\|\nabla u\|_{L^q(\omega_e^1)}. \tag{3.10}$$

Proof. The proof can be found in section 5.3. □

A slightly different situation arises if nonhomogeneous piecewise polynomial boundary conditions are to be preserved: The domain of influence in the local bounds is enlarged, and we impose a restriction on the variation in polynomial degree distribution for elements near the Dirichlet part of the boundary.

THEOREM 3.4 (Scott–Zhang-type quasi interpolation). *Let $q \in (1, \infty)$ and let $\mathcal{T}$ be a γ-shape regular triangulation of a domain $\Omega \subset \mathbb{R}^2$ satisfying* (M1)–(M4)*. Let $\mathbf{p}$*

be a polynomial degree distribution satisfying (2.8)*. Let* $\mathcal{B} \subset \mathcal{E}(\mathcal{T})$ *be a collection of boundary edges. If* $q \neq 2$*, assume additionally that*

$$|p_K - p_{K'}| \leq \gamma \qquad \forall K, K' \text{ s.t. } \overline{K} \cap \overline{K'} \cap \bar{b} \neq \emptyset \text{ for some } b \in \mathcal{B}. \tag{3.11}$$

Then there exists a linear operator $I^{hp}_{inhom} : W^{1,q}_{\mathcal{B},\mathbf{p}}(\Omega) \to S^{\mathbf{p}}(\mathcal{T})$ *such that*

$$(I^{hp}_{inhom} u)|_b = u|_b \qquad \forall b \in \mathcal{B}.$$

Furthermore, there exists a constant $C > 0$ *depending only on* γ *and* q *such that for all elements* $K \in \mathcal{T}$ *and all edges* $e \in \mathcal{E}(\mathcal{T})$

$$\|u - I^{hp}_{inhom} u\|_{L^q(K)} + \frac{h_K}{p_K} \|\nabla(u - I^{hp}_{inhom} u)\|_{L^q(K)} \leq C \frac{h_K}{p_K} \|\nabla u\|_{L^q(\omega^4_K)},$$

$$\|u - I^{hp}_{inhom} u\|_{L^q(e)} \leq C \left(\frac{h_K}{p_K}\right)^{1-1/q} \|\nabla u\|_{L^q(\omega^4_e)}.$$

Proof. The proof can be found in section 5.4. □

Remark 3.5. The dependence on the domains ω^4_K, ω^4_e is not optimal. A careful inspection of the proof allows slightly sharper bounds. For example, for elements K such that $\omega^4_K \subset\subset \Omega$ we can replace ω^4_K with ω^1_K.

4. Residual-based a posteriori error estimation.

The Clément-type interpolation operators can be utilized for a posteriori error estimation as discussed in [34]. The following proposition illustrates the type of results that can be obtained for a simple model problem.

PROPOSITION 4.1. *Let* $\mathcal{T}$ *be a triangulation of a domain* $\Omega \subset \mathbb{R}^2$ *that satisfies* (M1)–(M4)*. Assume that a polynomial degree distribution* $\mathbf{p}$ *satisfies* (2.8)*. For* $f \in L^2(\Omega)$ *let* $u \in H^1_0(\Omega)$ *be the weak solution of*

$$-\Delta u = f \qquad \text{on } \Omega, \qquad u|_{\partial\Omega} = 0. \tag{4.1}$$

Let $u_{FE} \in S^{\mathbf{p}}(\mathcal{T}) \cap H^1_0(\Omega)$ *be the finite element approximation to* u*. Then*

$$\|\nabla(u - u_{FE})\|^2_{L^2(\Omega)} \leq C \sum_{K \in \mathcal{T}} \eta^2_K, \tag{4.2}$$

where the local error indicators η_K *are defined by*

$$\eta^2_K := \left(\frac{h_K}{p_K}\right)^2 \|f + \Delta u_{FE}\|^2_{L^2(K)} + \sum_{\substack{e \in \mathcal{E}(K) \\ e \not\subset \partial\Omega}} \frac{h_e}{p_e} \|[\partial_n u_{FE}]\|^2_{L^2(e)}, \tag{4.3}$$

and $[\partial_n u_{FE}]$ *denotes the jump of the normal derivative of* u_{FE} *across the edge* e*. The constant* $C > 0$ *in* (4.2) *depends solely on* γ*.*

Proof. The proof follows along standard lines of residual-based a posteriori error estimation as outlined, for example, in [40]. We have the characterization

$$\|\nabla(u - u_{FE})\|_{L^2(\Omega)} = \sup_{v \in H^1_0(\Omega)} \frac{R(v)}{\|\nabla v\|_{L^2(\Omega)}}, \qquad R(v) := \int_\Omega \nabla(u - u_{FE}) \cdot \nabla v \, dx. \tag{4.4}$$

From the standard Galerkin orthogonality satisfied by the finite element solution u_{FE}, we have $R(v) = 0$ for all $v \in S^{\mathbf{p}}(\mathcal{T}) \cap H_0^1(\Omega)$. In particular, therefore $R(v) = R(v - I_{hom}^{hp} v)$ for all $v \in H_0^1(\Omega)$. Breaking up the integral over Ω into a sum of integrals over elements and an elementwise integration by parts yields for arbitrary $v \in H_0^1(\Omega)$

$$R(v) = R(v - I_{hom}^{hp} v) = \sum_{K \in \mathcal{T}} \int_K (f + \Delta u_{FE})(v - I_{hom}^{hp} v) + \int_{\partial K} \partial_n u_{FE}\, (v - I_{hom}^{hp} v)\, ds.$$

Hence using the abbreviations $r_K := f + \Delta u_{FE}$ and $r_e := [\partial_n u_{FE}]$ for the element residuals and edge residuals, the Cauchy–Schwarz inequality and Theorem 3.3 yield

$$\begin{aligned} |R(v)| &\leq \sum_{K \in \mathcal{T}} \|r_K\|_{L^2(K)} \|v - I_{hom}^{hp} v\|_{L^2(K)} + \sum_{\substack{e \in \mathcal{E}(\mathcal{T}) \\ e \not\subset \partial\Omega}} \|r_e\|_{L^2(e)} \|v - I_{hom}^{hp} v\|_{L^2(e)} \\ &\leq C \left\{ \sum_{K \in \mathcal{T}} \eta_K^2 \right\}^{1/2} \left\{ \sum_{K \in \mathcal{T}} \|\nabla v\|^2_{L^2(\omega_K^1)} \right\}^{1/2} \leq C \left\{ \sum_{K \in \mathcal{T}} \eta_K^2 \right\}^{1/2} \|\nabla v\|_{L^2(\Omega)}, \end{aligned}$$

where, in the last step, we employed a variation of Lemma 2.3 to conclude the existence of a constant $M > 0$ depending solely on γ such that each element $K \in \mathcal{T}$ is contained in not more than M sets of the form $\omega_{K'}^1$, $K' \in \mathcal{T}$. This concludes the argument. $\square$

Proposition 4.1 shows that the error estimator defined as the sum of the error indicators (4.1) is reliable. A corresponding lower bound, also known as efficiency estimate, is well known in the *h*-FEM. The difficulty in the present case of the *hp*-version is that polynomial inverse estimates have to be employed. The resulting lower bound then turns out to be independent of the local mesh size but dependent on the local polynomial degree. We refer to [34] for more details.

5. Proofs of the approximation results of section 3. We prove Theorems 3.1, 3.3, and 3.4 in turn, since Theorem 3.3 depends on Theorem 3.1, and Theorem 3.4 depends on both Theorem 3.1 and Theorem 3.3. For the proof of Theorem 3.1, we require polynomial approximation results on a reference configuration, which we choose to be a hyper cube, in section 5.1.1. For the proof of Theorems 3.3 and 3.4, we additionally require polynomial lifting results that are provided in section 5.1.2.

5.1. Polynomial approximation and lifting.

5.1.1. Polynomial approximation. For polynomial approximation on hyper cubes we have the following result.

THEOREM 5.1. *Let $d \in \mathbb{N}$ and I_i, $i = 1, \ldots, d$, be bounded intervals. Set $I = I_1 \times \cdots \times I_d$. Let $R \in \mathbb{N}$. Then for each $N \in \mathbb{N}_0$ there exists a bounded linear operator $J_{R,N} : L^1(I) \to \mathcal{Q}_N(I)$ with the following properties: For each $q \in [1, \infty]$ there exists a constant $C > 0$, which depends only on R, q, and I, such that for all $N \geq R - 1$ and all $0 \leq r \leq R$*

$$J_{R,N} u = u \qquad \forall u \in \mathcal{Q}_{R-1}, \tag{5.1}$$

$$\|u - J_{R,N} u\|_{W^{l,q}(I)} \leq C (N+1)^{-(r-l)} |u|_{W^{r,q}(I)}, \qquad l = 0, \ldots, r. \tag{5.2}$$

Proof. The operator $J_{R,N}$ is constructed as the tensor product of one-dimensional operators that are often employed in approximation theory. The detailed arguments can be found in Appendix A. $\square$

5.1.2. Polynomial liftings and extensions. In order to enforce polynomial boundary conditions, polynomial lifting results are required.

THEOREM 5.2. *Let K be the reference triangle or the reference square. Let $\Gamma = \overline{\Gamma} \subset \partial K$ be the union of closed edges of K. Let $q \in (1,\infty)$. Then there exists a constant $C > 0$ with the following property: For each $f \in C(\Gamma)$ such that f is a polynomial of degree p on each edge contained in Γ, there exists a polynomial $F \in \mathcal{P}_p$ if K is the triangle or $F \in \mathcal{Q}_p$ if K is the square such that $F|_\Gamma = f$ and*

$$\|F\|_{L^q(T)} \leq C\|f\|_{L^q(\Gamma)},$$
$$\|F\|_{W^{1,q}(T)} \leq C\|f\|_{W^{1-1/q,q}(\Gamma)}.$$

Moreover, the mapping $f \mapsto F$ is linear, and it reproduces constant functions.

Proof. The proof can be found in Appendix B.2.1. □

The lifting result Theorem 5.2 allows us to construct $W^{1,q}$-stable liftings from the boundary of the reference element. Our constructions will require an additional lifting where the L^q-norm of the lifting is smaller than that of Theorem 5.2. This is the merit of the next proposition.

PROPOSITION 5.3. *Let K be the reference triangle or the reference square. Let $\Gamma = \overline{\Gamma} \subset \partial K$ be a union of edges of K. Let $q \in (1,\infty)$. Then there exists $C > 0$ such that for every $f \in C(\Gamma)$, which is a polynomial of degree $p \in \mathbb{N}$ on each edge of K, there exists a polynomial F (if K is the reference triangle, then $F \in \mathcal{P}_{3p}$; otherwise $F \in \mathcal{Q}_{4p}$) such that $F|_{\partial\Gamma} = f$ and*

$$p\|F\|_{L^q(K)} + \|F\|_{W^{1,q}(K)} \leq C\|f\|_{W^{1-1/q,q}(\Gamma)} + Cp^{1-1/q}\|f\|_{L^q(\Gamma)}.$$

Furthermore, the mapping $f \mapsto F$ is linear.

Proof. The proof can be found in Appendix B.2.2. We mention that the proof of Proposition 5.3 can be modified so that $F \in \mathcal{P}_{\lceil \lambda p \rceil}$ or $F \in \mathcal{Q}_{\lceil \lambda p \rceil}$ for arbitrarily chosen $\lambda > 1$. The constant $C > 0$ does depends on λ, however. □

We will also employ the following one-dimensional extension result.

LEMMA 5.4. *Let $\hat{I} = (0,1)$, $q \in (1,\infty)$. Let $k \in \mathbb{N}_0$. Then there exists $C > 0$ such that for every $p \in \mathbb{N}_0$ with $p \geq k$ there exists a linear operator $Z_{p,p-k} : \mathcal{P}_p \to \mathcal{P}_{p-k}$ with the following properties:*

$$Z_{p,p-k}1 = 1, \qquad (Z_{p,p-k}u)(0) = u(0),$$
$$\|Z_{p,p-k}u\|_{L^q(\hat{I})} \leq C\|u\|_{L^q(\hat{I})}, \qquad \|Z_{p,p-k}u\|_{W^{1-1/q,q}(\hat{I})} \leq C\|u\|_{W^{1-1/q,q}(\hat{I})},$$
$$\|Z_{p,p-k}u - u\|_{\widetilde{W}_l^{1-1/q,q}(\hat{I})} \leq C\|u\|_{W^{1-1/q,q}(\hat{I})}.$$

Proof. The proof can be found in Appendix C. □

For $q = 2$ a sharper construction is possible, namely, to take the Gauß–Lobatto interpolant.

LEMMA 5.5. *Let $q = 2$ and $\lambda \in (0,1)$. Denote by $\mathcal{I}_{\lfloor \lambda p \rfloor} : \mathcal{P}_p \to \mathcal{P}_{\lfloor \lambda p \rfloor}$, the Gauß–Lobatto interpolation operator scaled to the interval $\hat{I}$. Then $i_{\lfloor \lambda p \rfloor}1 = 1$, $(\mathcal{I}_{\lfloor \lambda p \rfloor}u)(0) = u(0)$ for all $u \in \mathcal{P}_p$, and there exists $C > 0$ depending only on λ such that*

$$\|\mathcal{I}_{\lfloor \lambda p \rfloor}u\|_{L^2(\hat{I})} \leq C\|u\|_{L^2(\hat{I})}, \quad \|\mathcal{I}_{\lfloor \lambda p \rfloor}u\|_{W^{1/2,2}(\hat{I})} \leq C\|u\|_{W^{1/2,2}(\hat{I})}, \tag{5.3}$$

$$\|u - \mathcal{I}_{\lfloor \lambda p \rfloor}u\|_{\widetilde{W}^{1/2,2}(\hat{I})} \leq C\|u\|_{W^{1/2,2}(\hat{I})}. \tag{5.4}$$

Proof. The key step of the proof consists in stability estimates for the Gauß–Lobatto interpolation and can be obtained by combining the results of [17, Rem. 13.5]

(L^2-stability of the Gauß–Lobatto interpolation when applied to polynomials), [17, eq. (13.27)] (stability of the Gauß–Lobatto interpolation operator in H^1), and a bound in a weighted L^2-space [17, Thm. 13.4]. For a detailed version of the proof, we refer to [25, Lem. 4.1]. □

5.2. Proof of Theorem 3.1. The Clément-type interpolation operator I^{hp} is constructed in two steps: For each patch ω_V a local approximation I_V operator is constructed; in a second step, these local approximations are combined into a global one using the ideas of the partition of unity method [33]. The local approximation operator I_V is defined on the corresponding reference patch $\widehat{\omega}_V$ by extending to a suitable square containing $\widehat{\omega}_V$ and then employing the polynomial approximation result (Theorem 5.1).

We start by recalling a result from [33].

Lemma 5.6. *Let $\mathcal{T}$ be a γ-shape regular triangulation triangulation of a domain $\Omega \subset \mathbb{R}^2$ satisfying* (M1)–(M3). *Let $q \in [1, \infty]$, and let $\mathbf{p}$ be an arbitrary polynomial degree distribution. Let $u \in W^{1,q}(\Omega)$ and let, for each $V \in \mathcal{N}(\mathcal{T})$, a function $u_V \in S^{p_V - 1}(\mathcal{T}|_{\omega_V})$ be given, where p_V is defined in* (2.9). *Then there exists $C > 0$ depending solely on γ such that the function $\widetilde{u} := \sum_{V \in \mathcal{N}(\mathcal{T})} \varphi_V u_V \in S^{\mathbf{p}}(\mathcal{T})$ and*

$$\|u - \tilde{u}\|_{L^q(K)} \leq C \sum_{V \in \mathcal{N}(K)} \|u - u_V\|_{L^q(K)},$$

$$\|\nabla(u - \tilde{u})\|_{L^q(K)} \leq C \sum_{V \in \mathcal{N}(K)} \left[\|\nabla(u - u_V)\|_{L^q(K)} + \frac{1}{h_K} \|u - u_V\|_{L^q(K)} \right],$$

$$\|u - \tilde{u}\|_{L^q(e)} \leq C \sum_{V \in \mathcal{N}(e)} \|u - u_V\|_{L^q(e)}.$$

Proof. We start by ascertaining $\tilde{u} \in C(\Omega)$. This follows easily from the support properties of the functions $\varphi_V \in C(\Omega)$, namely, $\varphi|_{\Omega \setminus \omega_V} \equiv 0$, together with $u_V \in C(\omega_V) \cap L^\infty(\omega_V)$. In order to see $\tilde{u} \in S^{\mathbf{p}}(\mathcal{T})$ we have to make sure that $(\varphi_V u_V) \circ F_K \in \Pi_{p_K}(\hat{K})$ for all $K \in \mathcal{T}|_{\omega_V}$ for all $V \in \mathcal{N}(\mathcal{T})$. This follows easily from $\varphi_V \in S^1(\mathcal{T})$ and $u_V \in S^{p_V - 1}(\mathcal{T}|_{\omega_V})$. The essential ingredient for proving the estimates is the observation that $\sum_{V \in \mathcal{N}(K)} \varphi_V \equiv 1$ on K for every $K \in \mathcal{T}$ and $\sum_{V \in \mathcal{N}(e)} \varphi_V \equiv 1$ on e for every $e \in \mathcal{E}(\mathcal{T})$. The bounds on $(u - \tilde{u})|_K$ then follow from the observation that $(u - \tilde{u})|_K = \sum_{V \in \mathcal{N}(K)} \varphi_V (u - u_V)$, where the sum extends over at most four terms, and from the bounds (2.6) on the functions φ_V. □

Lemma 5.7. *Let $\mathcal{T}$ a γ-shape regular triangulation of a domain $\Omega \subset \mathbb{R}^2$ satisfying* (M1)–(M4). *Assume that the polynomial degree distribution $\mathbf{p}$ satisfies* (2.8). *Then for each vertex V there exists a bounded linear operator $I_V : L^1(\omega_V) \to S^{p_V - 1}(\mathcal{T}|_{\omega_V})$, and there exists a constant $C > 0$, which depends solely on γ, such that for each $u \in W^{1,q}(\omega_V)$, each $K \in \mathcal{T}|_{\omega_V}$, and each edge $e \in \mathcal{E}(\mathcal{T}|_{\omega_V})$*

$$\|u - I_V u\|_{L^q(K)} + \frac{h_K}{p_K} \|\nabla(u - I_V u)\|_{L^q(K)} \leq C \frac{h_V}{p_V} \|\nabla u\|_{L^q(\omega_V)},$$

$$\|u - I_V u\|_{L^q(e)} \leq C \left(\frac{h_V}{p_V} \right)^{1 - 1/q} \|\nabla u\|_{L^q(\omega_V)}.$$

Proof. Consider a patch ω_V. Condition (M4) provides the patch map $F_V : \widehat{\omega}_V \to \omega_V$ and Lemma 2.4 gives $\widehat{u}_V = u|_K \circ F_V \in W^{1,q}(\widehat{\omega}_V)$ together with

$$\|\widehat{u}_V\|_{L^q(\widehat{\omega}_V)} \leq C h_V^{2/q} \|u\|_{L^q(\omega_V)}, \qquad \|\nabla \widehat{u}_V\|_{L^q(\widehat{\omega}_V)} \leq C h_V^{1 - 2/q} \|\nabla u\|_{L^q(\omega_V)}.$$

Let $\widehat{S}$ be a square such that $\widehat{\omega}_V \subset\subset \widehat{S}$ and denote by $E : L^1(\widehat{\omega}_V) \to L^1(\widehat{S})$ the universal linear extension operator of [39]. We then have the existence of a constant $C_q > 0$, which depends solely on $q \in [1,\infty]$ and $\widehat{\omega}_V$, such that

$$\|E\widehat{u}_V\|_{L^q(\widehat{S})} \le C_q \|\widehat{u}_V\|_{L^q(\widehat{\omega}_V)}, \qquad \|E\widehat{u}_V\|_{W^{1,q}(\widehat{S})} \le C\|\widehat{u}_V\|_{W^{1,q}(\widehat{\omega}_V)}.$$

Choosing $N = \lfloor (p_V - 1)/2 \rfloor$ in the approximation result (Theorem 5.1), we obtain a bounded linear operator $J_{1,N} : L^1(\widehat{S}) \to \mathcal{Q}_N \subset \mathcal{P}_{p_V-1}$ that reproduces constant functions and satisfies

$$(p_V+1)\|v - J_{1,N}v\|_{L^q(\widehat{S})} + \|\nabla(v - J_{1,N}v)\|_{L^q(\widehat{S})} \le C\|v\|_{W^{1,q}(\widehat{S})} \qquad \forall v \in W^{1,q}(\widehat{S}).$$

We next define the operator $J_{p_V} : L^1(\widehat{\omega}_V) \to \mathcal{P}_{p_V-1}$ by

$$J_{p_V} v := \overline{v} + J_{1,N} \circ E(v - \overline{v}), \qquad \overline{v} := \frac{1}{|\widehat{\omega}_V|}\int_{\widehat{\omega}_V} v(x)\,dx.$$

J_{p_V} is a bounded linear operator on $L^1(\widehat{\omega}_V)$, and we obtain for $W^{1,q}$-functions

$$(p_V+1)\|v - J_{p_V}v\|_{L^q(\widehat{\omega}_V)} + \|\nabla(v - J_{p_V}v)\|_{L^q(\widehat{\omega}_V)} \le C\|v - \overline{v}\|_{W^{1,q}(\widehat{\omega}_V)} \le C\|\nabla v\|_{L^q(\widehat{\omega}_V)},$$

where in the last estimate we employed the second Poincaré inequality. Applying this operator to the pull-back $\widehat{u}_V$, we obtain

$$\begin{aligned}(p_V+1)\|\widehat{u}_V - J_{p_V}\widehat{u}_V\|_{L^q(\widehat{\omega}_V)} + \|\nabla(\widehat{u}_V - J_{p_V}\widehat{u}_V)\|_{L^q(\widehat{\omega}_V)} &\le C\|\nabla \widehat{u}_V\|_{L^q(\widehat{\omega}_V)}\\ &\le Ch_V^{1-2/q}\|u_V\|_{L^q(\omega_V)}.\end{aligned}$$

Returning to the patch ω_V, we observe that the function u_{p_V} defined on ω_V by $u_{p_V} = (J_{p_V}\widehat{u}_V) \circ F_V^{-1}$ is an element of $S^{p_V-1}(\mathcal{T}|_{\omega_V})$ (this is due to the fact that elementwise F_V is the composition of an *affine* map and the element map) and

$$(p_V+1)h_V^{-2/q}\|u_V - u_{p_V}\|_{L^q(\omega_V)} + h_V^{1-2/q}\|\nabla(u_V - u_{p_V})\|_{L^q(\omega_V)} \le Ch_V^{1-2/q}\|\widehat{u}_V\|_{L^q(\omega_V)}.$$

This leads to the desired bound on elements $K \in \mathcal{T}|_{\omega_V}$. For the bound on an edge $e \in \mathcal{E}(\mathcal{T}|_{\omega_V})$, we employ a trace theorem on $\widehat{\omega}_V$ before transforming back to ω_V.

Checking the steps of the construction, we see that the map $u_V \mapsto u_{p_V}$ is linear and that it is at the same time a bounded linear map $L^1(\omega_V) \to \mathcal{P}_{p_V-1}$.

The constant in the last estimate does depend on the reference patch $\widehat{\omega}_V$. We observe, however, that for a given (upper bound on) γ, only finitely many reference patches have to be considered since only finitely many elements can abut on a vertex (cf. Lemma 2.3). This concludes the argument. □

Proof of Theorem 3.1. Theorem 3.1 now follows from combining Lemmata 5.6 and 5.7. For each vertex V, we construct the local approximation $I_V u \in S^{p_V-1}(\mathcal{T}|_{\omega_V})$ with the aid of Lemma 5.7. The operator $I^{hp} : L^1(\Omega) \to S^{\mathbf{p}}(\mathcal{T})$ is then defined as

$$I^{hp}u = \sum_{V \in \mathcal{N}(\mathcal{T})} \varphi_V I_V u,$$

where the vertex shape functions $\varphi_V \in S^1(\mathcal{T})$ have the support properties of (2.6). The operator I^{hp} maps indeed into $S^{\mathbf{p}}(\mathcal{T})$ since $I_V u \in s^{p_V-1}(\mathcal{T}|_{\omega_V})$. By inspection, we observe that $I^{hp} : L^1(\Omega) \to S^{\mathbf{p}}(\mathcal{T})$ is a bounded linear operator. Its approximation properties, when applied to $W^{1,q}$-functions, follow from Lemmata 5.6 and 5.7. □

5.3. Proof of Theorem 3.3. We modify the approximation operator of Theorem 3.1 so as to enforce homogeneous Dirichlet boundary conditions. Since we need the trace theorem to hold, the operator is now defined on $W^{1,q}(\Omega)$ instead of $L^1(\Omega)$. The construction of this operator is again patch oriented. The difference is that we will change the definition of the linear maps I_V for $V \in \mathcal{N}(\mathcal{B})$. Here, we defined

$$\mathcal{N}(\mathcal{B}) := \bigcup_{b\in\mathcal{B}} \mathcal{N}(b). \tag{5.5}$$

We first analyze the prototypical situation on a boundary reference patch.

LEMMA 5.8. *Let $q \in (1,\infty)$. Let $\widehat{\omega} = \omega_{bdy,M,j}$ for some $M \in \mathbb{N}$ and $j \in \{1,\dots,2^M\}$ and denote by $\widehat{\mathcal{T}}$ the triangulation of $\widehat{\omega}$. Denote by Γ_0 the edge of $\widehat{\omega}$ lying on the x-axis and by Γ_M the edge lying on the y-axis (cf. Figure* 5.1*). Let Γ_D be either Γ_0, Γ_M, or $\Gamma_0 \cup \Gamma_M \cup \{0\}$. Denote $W^{1,q}_{\Gamma_D,0}(\widehat{\omega}) = \{u \in W^{1,q}(\widehat{\omega}) \,|\, u|_{\Gamma_D} = 0\}$. Then for every $p \in \mathbb{N}_0$ there exists a bounded linear map $I_p : W^{1,q}_{\Gamma_D,0}(\widehat{\omega}) \to S^p(\widehat{\mathcal{T}}) \cap W^{1,q}_{\Gamma_D,0}(\widehat{\omega})$ such that*

$$(p+1)\|u - I_p u\|_{L^q(\widehat{\omega})} + \|\nabla(u - I_p u)\|_{L^q(\widehat{\omega})} \le C\|\nabla u\|_{L^q(\widehat{\omega})}, \tag{5.6}$$

where the constant $C > 0$ is independent of p and $u \in W^{1,q}_{\Gamma_D,0}(\widehat{\omega})$.

Proof. We will demonstrate the result for the case $\Gamma_D = \Gamma_0 \cup \Gamma_M \cup \{0\}$, the other two cases being handled similarly. The construction of I_p is done in two steps: First, we let $J_p : L^1(\widehat{\omega}) \to \mathcal{P}_p$ be the linear operator of the proof of Lemma 5.7. It satisfies for $u \in W^{1,q}(\widehat{\omega})$

$$(p+1)\|u - J_p u\|_{L^q(\widehat{\omega})} + \|\nabla(u - J_p u)\|_{L^q(\widehat{\omega})} \le C\|\nabla u\|_{L^q(\widehat{\omega})}.$$

In particular, from the multiplicative trace inequality (see, e.g., [19, Thm. 1.6.6]) and the fact that $u|_{\Gamma_D} = 0$, we get

$$\begin{aligned}\|J_p u\|_{L^q(\Gamma_D)} &= \|u - J_p u\|_{L^q(\Gamma_D)} \le C(p+1)^{-(1-1/q)}\|\nabla u\|_{L^q(\widehat{\omega})},\\ \|J_p u\|_{W^{1-1/q,q}(\Gamma_D)} &= \|u - J_p u\|_{W^{1-1/q,q}(\Gamma_D)} \le C\|\nabla u\|_{L^q(\widehat{\omega})}.\end{aligned}$$

The function $J_p u$ does not, however, satisfy homogeneous boundary conditions on Γ_D. This is corrected in a second step by an element-by-element construction using appropriate polynomial liftings. To that end, we enumerate the edges of $\widehat{\mathcal{T}}$ emanating from the origin in a counterclockwise fashion as depicted in Figure 5.1. Likewise, the elements are labeled K_i, $i = 0,\dots,M-1$. Next, we observe that the functions $u_0 := (u - J_p u)|_{\Gamma_0} = (J_p u)|_{\Gamma_0}$ and $u_M := (u - J_p u)|_{\Gamma_M} = (J_p u)|_{\Gamma_M}$ are polynomials of degree p. Additionally, we note that each edge Γ_i, $i = 0,\dots,M$, is homeomorphic to the reference interval $\hat{I} = (0,1)$ by means of an affine map $\gamma_i : \hat{I} \to \Gamma_i$, which we may choose to satisfy $\gamma_i(0) = 0$ for $i \in \{0,\dots,M\}$. We then define a function $z \in C(\cup_{i=0}^M \overline{\Gamma_i})$ by

$$\begin{aligned} z \circ \gamma_i(x) &:= u_0 \circ \gamma_0(x), \qquad x \in \overline{\hat{I}}, \quad i = 0,\dots,M-1,\\ z \circ \gamma_M(x) &:= u_M \circ \gamma_M(x).\end{aligned}$$

Clearly, for $i \in \{0,\dots,M\}$ we have

$$\|z\|_{L^q(\Gamma_i)} \le C\left[\|u_0\|_{L^q(\Gamma_0)} + \|u_M\|_{L^q(\Gamma_M)}\right] \le C(p+1)^{-1+1/q}\|\nabla u\|_{L^q(\widehat{\omega})}.$$

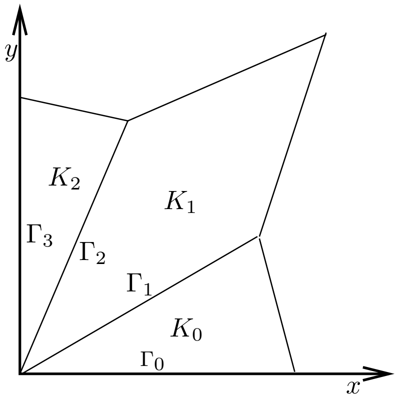


FIG. 5.1. *Numbering of elements and edges of a boundary reference patch for $M = 3$.*

We next define for each element K_i the set

$$\Gamma_{i,i+1} := \Gamma_i \cup \Gamma_{i+1} \cup \{0\} \subset \partial K_i, \qquad i = 0, \dots, M-1. \tag{5.7}$$

We then ascertain

$$\|z\|_{W^{1-1/q,q}(\Gamma_{i,i+1})} \leq C\|J_p u\|_{W^{1-1/q,q}(\Gamma_D)} \leq C\|u\|_{W^{1,q}(\widehat{\omega})}.$$

This follows easily from the fact that for two edges Γ_i, Γ_j with $i \neq j$ we have the characterization (see, e.g., [28, Thm. 1.5.2.3])

$$\|z\|^q_{W^{1-1/q,q}(\Gamma_i \cup \Gamma_j \cup \{0\})} \sim \|\hat{z}_i\|^q_{W^{1-1/q,q}(\hat{I})} + \|\hat{z}_j\|^q_{W^{1-1/q,q}(\hat{I})} + \int_0^1 \frac{|\hat{z}_i(x) - \hat{z}_j(x)|^q}{x^{q-1}}\,dx,$$

where we wrote $\hat{z}_i = z \circ \gamma_i$, $\hat{z}_j = z \circ \gamma_j$. Here, the constants hidden in the $\sim$-notation depend solely on Γ_i, Γ_j, and q. We finally construct with the aid of Proposition 5.3 a function $Z \in S^{4p}(\widehat{\mathcal{T}})$ such that $Z|_{\Gamma_i} = z|_{\Gamma_i}$ for all $i \in \{0, \dots, M\}$ and

$$\begin{aligned}(p+1)\|Z\|_{L^q(K_i)} + \|\nabla Z\|_{L^q(K_i)} &\leq C\left[\|z\|_{W^{1-1/q,q}(\Gamma_{i,i+1})} + (p+1)^{1-1/q}\|z\|_{L^q(\Gamma_{i,i+1})}\right] \\ &\leq C\|\nabla u\|_{L^q(\widehat{\omega})}, \qquad i = 0, \dots, M-1.\end{aligned}$$

We conclude the argument by noting that the map $u \mapsto J_p u + Z$ is linear and bounded. Since $J_p u + Z \in S^{4p}(\widehat{\mathcal{T}})$, replacing p with $\lfloor p/4 \rfloor$ gives the desired result. ∎

Proof of Theorem 3.3. The proof of Theorem 3.3 now follows by the same arguments as that of Theorem 3.1. Merely for the patches ω_V with $V \in \mathcal{N}(\mathcal{B})$ we replace the local approximation $I_V u$ of Lemma 5.7 with the pushforward $(I_{p_V} \widehat{u}_V) \circ F_V^{-1}$ of $I_{p_V} \widehat{u}_V$, where $I_{p_V} \widehat{u}_V$ with $\widehat{u}_V = u|_{\omega_V} \circ F_V$ is defined in Lemma 5.8. ∎

5.4. Proof of Theorem 3.4.

5.4.1. Lifting from $\mathcal{B}$. The proof of Theorem 3.4 follows along the same lines as that of Theorem 3.3. The key difference is that additionally appropriate (polynomial) liftings are required. Providing these is the purpose of the present subsection.

We start with a "vertex lifting" result on boundary reference patches that yields the correct value at a boundary vertex. Given a collection of boundary edges $\widehat{\mathcal{B}}$ of the reference patch $\widehat{\omega}$, the spaces $W^{1,q}_{\widehat{\mathcal{B}},\mathbf{p}}(\widehat{\omega})$ on $\widehat{\omega}$ are defined analogously to the way the spaces $W^{1,q}_{\mathcal{B},\mathbf{p}}(\Omega)$ are defined in (3.6). We then have the following lemma.

LEMMA 5.9. *Let $\widehat{\omega} = \widehat{\omega}_{bdy,M,j}$ for some $M \in \mathbb{N}$, $j \in \{1,\ldots,2^M\}$, and denote by $\widehat{\mathcal{T}}$ the triangulation of $\widehat{\omega}$. Let $\mathbf{p}$ be a polynomial degree distribution on $\widehat{\mathcal{T}}$ that satisfies* (2.8)*. If $q \neq 2$, assume additionally*

$$|p_K - p_{K'}| \le k \qquad \forall K, K' \in \widehat{\mathcal{T}}.$$

Define $p' := \min\{p_K - 1 \,|\, K \in \widehat{\mathcal{T}}\} \in \mathbb{N}_0$.

Let $\widehat{\mathcal{B}} = \{\Gamma_0\}$ or $\widehat{\mathcal{B}} = \{\Gamma_M\}$ or $\widehat{\mathcal{B}} = \{\Gamma_0, \Gamma_M\}$ (cf. Figure 5.1*). Then there exists a constant $C > 0$ that depends solely on $\widehat{\omega}$ (i.e., on M, j) and γ of* (2.8)*, k (if $q \neq 2$), q, and there exists a bounded linear operator $L : W^{1,q}_{\widehat{\mathcal{B}},\mathbf{p}}(\widehat{\omega}) \to S^{p'}(\widehat{\mathcal{T}})$ such that*

$$(Lu - u)(0) = 0, \tag{5.8}$$

$$\|Lu - u\|_{W^{1,q}(\widehat{\omega})} \le C\|\nabla u\|_{L^q(\widehat{\omega})}, \tag{5.9}$$

$$\|(Lu - u)\circ\gamma_b\|_{\widetilde{W}_l^{1-1/q,q}(\hat{I})} \le C\|\nabla u\|_{L^q(\widehat{\omega})} \qquad \forall b \in \widehat{\mathcal{B}}, \tag{5.10}$$

where $\gamma_b : \hat{I} \to b$ is the affine parametrization of $b \in \widehat{\mathcal{B}}$ satisfying $\gamma_b(0) = 0$.

Proof. We employ ideas similar to those of the proof of Theorem 3.3. For simplicity of notation, we consider the case $\widehat{\mathcal{B}} = \{\Gamma_0, \Gamma_M\}$; the other two cases are treated in a similar fashion. We denote by $\gamma_i : \hat{I} \to \Gamma_i$, $i = 0,\ldots,M$, the affine parametrizations of the edges Γ_i, which are assumed without loss of generality to satisfy $\gamma_i(0) = 0$. We will construct Lu first on the edges Γ_i and in a second step define Lu on the elements via appropriate liftings. We will only consider the case $q \neq 2$—the reader may check that in the case $q = 2$, the operator $Z_{p,p'}$ of Lemma 5.4 can be replaced with the Gauß–Lobatto interpolation operator $\mathcal{I}_{p'}$ discussed in Lemma 5.5.

We write $p = \max\{p_K \,|\, K \in \widehat{\mathcal{T}}\} \in \mathbb{N}$. Choose $b \in \widehat{\mathcal{B}}$. Without loss of generality, we assume that $b = \Gamma_0$. By assumption $u \circ \gamma_0 \in \mathcal{P}_p$, so that we may define $l_0 := Z_{p,p'}(u\circ\gamma_0)$, where the linear operator $Z_{p,p'} : \mathcal{P}_p \to \mathcal{P}_{p'}$ is the polynomial extension operator of Lemma 5.4. We then have $l_0(0) = u(0)$ and additionally by properties of $Z_{p,p'}$ and the trace theorem

$$\|l_0 - u|_{\Gamma_0}\circ\gamma_0\|_{\widetilde{W}_l^{1-1/q,q}(\hat{I})} \le C\|u\|_{W^{1-1/q,q}(\Gamma_0)} \le C\|u\|_{W^{1,q}(\widehat{\omega})}, \tag{5.11}$$

$$\|l_0 - u|_{\Gamma_M}\circ\gamma_M\|_{\widetilde{W}_l^{1-1/q,q}(\hat{I})} \le C\|u\|_{W^{1,q}(\widehat{\omega})}. \tag{5.12}$$

Next, we define

$$(Lu)|_{\Gamma_i} = l_0 \circ \gamma_i^{-1}, \qquad i = 0,\ldots,M.$$

This gives $(Lu)(0) = u(0)$. Furthermore, this definition of $(Lu)|_{\Gamma_i}$ in conjunction with the bounds (5.11), (5.12) implies

$$\|Lu\|_{W^{1-1/q,q}(\Gamma_{i,i+1})} \le C\|u\|_{W^{1,q}(\widehat{\omega})}, \qquad i = 0,\ldots,M-1,$$

where we abbreviate $\Gamma_{i,i+1} = \Gamma_i \cup \Gamma_{i+1} \cup \{0\}$ as in (5.7). From the lifting result (Theorem 5.2), there exists then a function $Lu \in S^{p'}(\widehat{\mathcal{T}})$ with

$$\|Lu\|_{W^{1,q}(\widehat{\omega})} \le C\|u\|_{W^{1,q}(\widehat{\omega})}. \tag{5.13}$$

Furthermore, inspection of the construction of Lu reveals that $u \mapsto Lu$ is linear. Since the operator $Z_{p,p'}$ of Lemma 5.4 satisfies $Z_{p,p'}1 = 1$ and the lifting of Theorem 5.2 likewise ensures that constant functions are reproduced, we conclude $L1 = 1$. By a standard argument, we can therefore strengthen (5.13) to yield (5.9). The estimates (5.10) are ensured by the way we defined $(Lu)|_{\Gamma_i}$ for $i \in \{0, \ldots, M\}$. □

Lemma 5.7 allows us to construct a lifting operator as follows.

PROPOSITION 5.10. *Let $\mathcal{T}$ be a γ-shape regular triangulation of a domain $\Omega \subset \mathbb{R}^2$ satisfying* (M1)–(M4). *Let $\mathcal{B} \subset \mathcal{E}(\mathcal{T})$ be a collection of boundary edges. Let $q \in (1, \infty)$ be given. Assume that the polynomial degree distribution $\mathbf{p}$ satisfies* (2.8) *and additionally* (3.11) *if $q \neq 2$. Then there exists a constant $C > 0$, which depends solely on γ and q, and there exists a bounded linear operator $I^{hp}_{lift} : W^{1,q}_{\mathcal{B},\mathbf{p}}(\Omega) \to S^{\mathbf{p}}(\mathcal{T})$ such that*

$$
\begin{aligned}
(I^{hp}_{lift}u)|_b &= u|_b \qquad \forall b \in \mathcal{B},\\
(I^{hp}_{lift}u)|_K &= 0 \qquad \textit{if } \omega_{K,\mathcal{B}} = \emptyset,
\end{aligned}
$$

$$\|I^{hp}_{lift}u\|_{L^q(K)} + h_K\|\nabla I^{hp}_{lift}u\|_{L^q(K)} \le C\big[\|u\|_{L^q(\omega_{K,\mathcal{B}})} + h_K\|\nabla u\|_{L^q(\omega_{K,\mathcal{B}})}\big] \quad \textit{if } \omega_{K,\mathcal{B}} \neq \emptyset,$$

where for an element $K \in \mathcal{T}$ we define

$$\omega_{K,\mathcal{B}} := \bigcup_{V \in \mathcal{N}(K) \cap \mathcal{N}(\mathcal{B})} \omega_V. \tag{5.14}$$

Proof. The lifting $I^{hp}_{lift}u$ is constructed as the sum of u_1 and u_2. The term u_1 is constructed such that the correct behavior at the vertices of the triangulation is ensured. In this way, the construction of the lifting is then reduced to an edgewise construction, which defines u_2.

Given $u \in W^{1,q}_{\mathcal{B},\mathbf{p}}(\Omega)$, we construct $u_1 \in S^{\mathbf{p}}(\mathcal{T})$ patchwise as

$$u_1 = \sum_{V \in \mathcal{N}(\mathcal{T})} \varphi_V L_V u,$$

where the patch operators L_V are defined with the aid of Lemma 5.9 according to the following rules:

(a) if $V \notin \mathcal{N}(\mathcal{B})$, then $L_V u = 0$;

(b) if $V \in \mathcal{N}(\mathcal{B})$, then $L_V u$ is defined on the corresponding reference patch $\widehat{\omega}_V$ as $(L_V u) \circ F_V = L\widehat{u}$, where L is the operator of Lemma 5.9. Here, $\widehat{u} = u \circ F_V$ and the polynomial degrees p and p' are defined as $p = \max\{p_K \mid K \in \mathcal{T}|_{\omega_V}\}$ and $p' = \min\{p_K \mid K \in \mathcal{T}|_{\omega_V}\} - 1$.

By the choice of the polynomial degrees p', we get $u_1 \in S^{\mathbf{p}}(\mathcal{T})$. Additionally, the function $L_V u$ satisfies for $V \in \mathcal{N}(\mathcal{B})$

$$
\begin{aligned}
(L_V u)(V) &= u(V),\\
\|L_V u\|_{L^q(\omega_V)} &\le C\left[\|u\|_{L^q(\omega_V)} + h_V\|\nabla u\|_{L^q(\omega_V)}\right],\\
\|\nabla L_V u\|_{L^q(\omega_V)} &\le C\|\nabla u\|_{L^q(\omega_V)}.
\end{aligned}
$$

Moreover, for edges $b \in \mathcal{B}$ and vertices $V \in \mathcal{N}(b)$ we have upon denoting by $\gamma_{b,V}$ the map $\gamma_{b,V} : \hat{I} \to b$ that is determined by the element maps and the condition $\gamma_{b,V}(0) = V$, the following bound:

$$\|(u - L_V u) \circ \gamma_{b,V}\|_{\widetilde{W}^{1-1/q,q}_l(\hat{I})} \le C h_V^{1-2/q}\|\nabla u\|_{L^q(\omega_V)} \qquad \forall b \in \mathcal{B}, \quad V \in \mathcal{N}(b).$$

For elements K with $\omega_{K,\mathcal{B}} = \emptyset$, our construction implies $(u_1)|_K = 0$. For elements K with $\omega_{K,\mathcal{B}} \neq \emptyset$ we get

$$\begin{aligned} u_1(V) &= u(V) \qquad \forall V \in \mathcal{N}(\mathcal{B}), \\ \|u_1\|_{L^q(K)} + h_K\|\nabla u_1\|_{L^q(K)} &\leq C\left[\|u\|_{L^q(\omega_{K,\mathcal{B}})} + h_K\|\nabla u\|_{L^q(\omega_{K,\mathcal{B}})}\right], \\ \|(u-u_1)\circ\gamma_b\|_{\widetilde{W}^{1-1/q,q}(\hat{I})} &\leq C h_b^{1-2/q}\|\nabla u\|_{L^q(\omega_b^1)} \qquad \forall b \in \mathcal{B}. \end{aligned}$$

For the last estimate, we employed additionally Lemma C.2.

We now turn to the construction of u_2. Since u_1 and u coincide in the vertices that lie on the Dirichlet boundary, we can proceed in an element-by-element fashion. For elements K with $\mathcal{E}(K)\cap\mathcal{B} = \emptyset$, we set $u_2|_K = 0$. For elements K with $\mathcal{E}(K)\cap\mathcal{B} \neq \emptyset$, we construct $u_2|_K$ using the following considerations: We set $\mathcal{B}_K = \mathcal{E}(K)\cap\mathcal{B}$, denote by $\widehat{b} := F_K^{-1}(b)$ the pull-back of an edge $b \in \mathcal{B}_K$, and construct with the aid of the lifting result (Theorem 5.2) on the reference element $\hat{K}$ the polynomial $\widehat{u}_{2,K} \in \mathcal{P}_{p_K}$ such that

$$\begin{aligned} \widehat{u}_{2,K}|_{\widehat{b}} &= ((u-u_1)\circ F_K)|_{\widehat{b}} \qquad \forall b \in \mathcal{B}_K, \\ \widehat{u}_{2,K}|_{F_K^{-1}(e)} &= 0 \qquad \forall e \in \mathcal{E}(K)\setminus\mathcal{B}_K, \\ \|\widehat{u}_2\|_{W^{1,q}(\widehat{K})} &\leq C\sum_{b\in\mathcal{B}_K}\|(u-u_1)\circ F_b\|_{\widetilde{W}^{1-1/q,q}(\hat{I})} \leq C h_K^{1-2/q}\|\nabla u\|_{L^q(\omega_{K,\mathcal{B}})}, \end{aligned}$$

where $F_b : \hat{I} \to b$ denotes the parametrization of b determined by the element maps. Pushing forward these estimates to the element K, the function $u_2|_K := \widehat{u}_{2,K}\circ F_K^{-1}$ then satisfies

$$\begin{aligned} u_2|_b &= (u-u_1)|_b \qquad \forall b \in \mathcal{B}_K, \\ u_2|_e &= 0 \qquad \forall e \in \mathcal{E}(K)\setminus\mathcal{B}_K, \\ \|u_2\|_{L^q(K)} &\leq C h_K^{2/q}\|\widehat{u}_2\|_{L^q(\widehat{K})} \leq C h_K\|\nabla u\|_{L^q(\omega_{K,\mathcal{B}})}, \\ \|\nabla u_2\|_{L^q(K)} &\leq C h_K^{2/q-1}\|\nabla\widehat{u}_2\|_{L^q(\widehat{K})} \leq C\|\nabla u\|_{L^q(\omega_{K,\mathcal{B}})}. \end{aligned}$$

The sum $u_1 + u_2$ is an element of $S^{\mathbf{p}}(\mathcal{T})$; it satisfies $(u_1+u_2)|_b = u|_b$ for all $b \in \mathcal{B}$, and we have the estimates

$$\|u_1+u_2\|_{L^q(K)} + h_K\|\nabla(u_1+u_2)\|_{L^q(K)} \leq C\left[\|u\|_{L^q(\omega_{K,\mathcal{B}})} + h_K\|\nabla u\|_{L^q(\omega_{K,\mathcal{B}})}\right].$$

Inspection of the construction shows that the map $u \mapsto u_1 + u_2$ is linear. □

5.4.2. Proof of Theorem 3.4. We are now in a position to prove Theorem 3.4.

Proof of Theorem 3.4. We employ the lifting operator I^{hp}_{lift} of Proposition 5.10 and the approximation operators I^{hp}_{hom}, I^{hp} of Theorems 3.3 and 3.1. We define

$$\begin{aligned} L &:= I^{hp}_{lift}\circ(\mathrm{Id} - I^{hp}) + I^{hp}, \\ I^{hp}_{inhom} &:= L + I^{hp}_{hom}\circ(\mathrm{Id} - L). \end{aligned}$$

I^{hp}_{inhom} is a linear operator mapping into $S^{\mathbf{p}}(\mathcal{T})$. We easily check that for $u \in W^{1,q}_{\mathcal{B},\mathbf{p}}(\Omega)$

$$(I^{hp}_{inhom}u)|_b = u|_b \qquad \forall b \in \mathcal{B}.$$

It remains to check the approximation properties. Upon writing

$$(\operatorname{Id} - I^{hp}_{inhom}) = (\operatorname{Id} - I^{hp}_{hom}) \circ (\operatorname{Id} - L),$$

we see that the desired approximation follow from the approximation properties of I^{hp}_{hom} together with stability properties of $\operatorname{Id} - L$. These stability properties can be inferred by writing

$$\operatorname{Id} - L = (\operatorname{Id} - I^{hp}_{lift}) \circ (\operatorname{Id} - I^{hp})$$

and then observing that Proposition 5.10 implies

$$\|\nabla(u - Lu)\|_{L^q(K)} \leq C \left[\frac{1}{h_K} \|u - I^{hp} u\|_{L^q(\omega^2_K)} + \|\nabla(u - I^{hp} u)\|_{L^q(\omega^2_K)} \right] \leq C \|\nabla u\|_{L^q(\omega^3_K)}.$$

Theorem 3.3 then implies

$$\|u - I^{hp}_{inhom} u\|_{L^q(K)} + \frac{h_K}{p_K} \|\nabla(u - I^{hp}_{inhom} u)\|_{L^q(K)} \leq C \frac{h_K}{p_K} \|\nabla u\|_{L^q(\omega^4_K)}.$$

From this, the desired estimate for the edges follows. □

Appendix A. Approximation results.

A.1. Polynomial approximation results on hyper cubes. We establish polynomial approximation results for the approximation of functions of Sobolev spaces $W^{r,q}(I)$, where I is a hyper cube. Similar results have been obtained in [2,3]. Our exposition here ignores effects related to the behavior of polynomials near the endpoints of an interval. While in the one-dimensional situation a characterization of the functions that can be approximated at a certain rate can be done using weighted spaces, these results do not easily extend to higher dimensions. We refer to [24] for an exposition of the one-dimensional results (so-called direct and inverse estimates) and mention also [9,10] where related results for the two-dimensional case are proved.

We recall a one-dimensional result on simultaneous trigonometric approximation.

LEMMA A.1. *Let $\mathbb{T}$ be the one-dimensional torus and denote for $r \in \mathbb{N}_0$, $q \in [1, \infty]$ by $W^{r,q}(\mathbb{T})$ the set of functions with r weak derivatives whose derivatives are in $L^q(\mathbb{T})$. Denote by T_N the set of trigonometric polynomials of degree $N \in \mathbb{N}$. Then for each $R \in \mathbb{N}$ and each N there exists a linear operator $J_{R,N} : L^1(\mathbb{T}) \to T_N$ and a constant $C_R > 0$ (which depends solely on R) such that for all $r \in \mathbb{N}_0$ with $0 \leq r \leq R$, all $q \in [1, \infty]$, and all $u \in W^{r,q}(\mathbb{T})$*

$$\|\,(u - J_{R,N}\, u)^{(j)}\,\|_{L^q(\mathbb{T})} \leq C_r (N+1)^{-(r-j)} \|u^{(r)}\|_{L^q(\mathbb{T})}, \qquad j = 0, \ldots, r. \tag{A.1}$$

Proof. Jackson-type results of this form are well known in approximation theory. The linear operators $J_{R,N}$, whose existence is ascertained in Lemma A.1, can be chosen as in [24, Chap. 7, eq. (2.8)]. The results concerning simultaneous approximation then follow from combining Theorems 2.3, 2.7, and 2.8 of [24, Chap. 7] and a check that the case $q = \infty$ is included in the form stated in Lemma A.1. The details can be found in [32, Prop. E.1]. □

As is well known, trigonometric approximation result implies polynomial approximation results by means of the transformation $x = \cos\theta$. For future reference, we formulate this in the following proposition.

PROPOSITION A.2. *Let $I \subset \mathbb{R}$ be a bounded interval. Let $R \in \mathbb{N}$ and $q \in [1, \infty]$. Then for each $N \in \mathbb{N}_0$ there exists a linear operator $J_{R,N} : L^1(I) \to \mathcal{P}_N$ and a constant $C > 0$, which depends only on R, q, I, such that for each $0 \leq r \leq R$*

$$\|u - J_{R,N} u\|_{W^{j,q}(I)} \leq C(N+1)^{-(r-j)} \|u\|_{W^{r,q}(I)}, \qquad j = 0, \ldots, r. \tag{A.2}$$

Furthermore, the linear operator $J_{R,N}$ may be constructed such that for $0 \leq r \leq R$ and $N \geq R - 1$

$$J_{R,N} u = u \qquad \forall u \in \mathcal{P}_{R-1} \tag{A.3}$$

$$\|u - J_{R,N} u\|_{W^{j,q}(I)} \leq C(N+1)^{-(r-j)} |u|_{W^{r,q}(I)}, \qquad j = 0, \ldots, r. \tag{A.4}$$

Proof. The proof for $N = 0$ is trivial; we will therefore assume $N \in \mathbb{N}$. We will obtain the results for polynomial approximation from those for trigonometric approximation. We construct for given $u \in W^{r,q}(I)$ the approximant $J_N u \in \mathcal{P}_N$; tracing the steps of the construction then reveals that $u \mapsto J_N u$ is in fact a linear operator. Without loss of generality, we may assume that I is such that the closed interval $\overline{I}$ satisfies $\overline{I} = [-\cos\varepsilon, \cos\varepsilon]$ for some chosen $\varepsilon \in (0, \pi/2)$.

Step 1. Define the interval $\Theta = (\varepsilon, \pi - \varepsilon)$. For every function v (defined on I) we define a function v_θ on Θ by $v_\theta(\theta) = v(\cos\theta)$. Then for every $j \in \mathbb{N}_0$, $q \in [1, \infty]$ there exists a constant $C > 0$, which depends only on j, q, and ε, such that

$$C^{-1} \|v\|_{W^{j,q}(I)} \leq \|v_\theta\|_{W^{j,q}(\Theta)} \leq C \|v\|_{W^{j,q}(I)}. \tag{A.5}$$

Step 2. We construct a function $\widetilde{u}_\theta$ on the torus $\mathbb{T}$ with the properties that (a) $\widetilde{u}_\theta = u_\theta$ on Θ, (b) $\widetilde{u}_\theta$ is symmetric with respect to $\theta = 0$, and (c) $\|\widetilde{u}_\theta\|_{W^{r,q}(\mathbb{T})} \leq C\|u\|_{W^{r,q}(I)}$. To that end, we extend u_θ to a function in $W^{r,q}(\mathbb{R})$ such that the extended function (again denoted by u_θ) satisfies

$$\|u_\theta\|_{W^{r,q}(\mathbb{R})} \leq C \|u_\theta\|_{W^{r,q}(\Theta)};$$

such an extension is constructed, for example, in [39]. Furthermore, using smooth cut-off functions, we may assume that this extension satisfies $\operatorname{supp} u_\theta \subset [\varepsilon/2, \pi - \varepsilon/2]$. We then define on the interval $(-\pi, \pi)$ the symmetric extension of u_θ by

$$\widetilde{u}_\theta(x) := \begin{cases} u_\theta(x) & \text{if } x \in (0, \pi), \\ u_\theta(-x) & \text{if } x \in (-\pi, 0). \end{cases}$$

By the support properties of u_θ we then conclude

$$\|\widetilde{u}_\theta\|_{W^{r,q}(\mathbb{T})} \leq C \|u\|_{W^{r,q}(I)}.$$

Step 3. From Lemma A.1, we get for the trigonometric polynomial $J_N := J_{R,N}\, \widetilde{u}_\theta$

$$\|\widetilde{u}_\theta - J_N\|_{W^{j,q}(\mathbb{T})} \leq C N^{-(r-j)} \|u\|_{W^{r,q}(I)}. \tag{A.6}$$

We wish to approximate $\widetilde{u}_\theta$ by a symmetric trigonometric polynomial. Since $\widetilde{u}_\theta$ is symmetric with respect to $\theta = 0$, we get that the trigonometric polynomial $\widetilde{J}_N$ defined by $\widetilde{J}_N(x) = J_N(-x)$ also satisfies

$$\|\widetilde{u}_\theta - \widetilde{J}_N\|_{W^{j,q}(\mathbb{T})} \leq C N^{-(r-j)} \|u\|_{W^{r,q}(I)}. \tag{A.7}$$

Combining (A.6), (A.7), we conclude that the symmetric trigonometric polynomial

$$\widehat{J}_N := \frac{1}{2}\left(J_N + \widetilde{J}_N\right)$$

satisfies

$$\|\widetilde{u}_\theta - \widehat{J}_N\|_{W^{j,q}(\mathbb{T})} \leq CN^{-(r-j)}\|u\|_{W^{r,q}(I)}. \tag{A.8}$$

Step 4. Since the symmetric trigonometric polynomial $\widehat{J}_N$ can be written in the form $\widehat{J}_N(\theta) = P_N(\cos(\theta))$ for a polynomial $P_N \in \mathcal{P}_N$, we get the desired operator $J_{R,N}$ and the bound (A.2) from (A.8) and (A.5).

Step 5. As a preparation to the final step, we recall the Bramble–Hilbert lemma [19, Lem. 4.3.8]: For $r \geq 1$, we have

$$\inf_{v \in \mathcal{P}_{r-1}} \|u - v\|_{W^{r,q}(I)} \leq C\|u^{(r)}\|_{L^q(I)}. \tag{A.9}$$

Next, we choose a bounded linear operator $Q : L^1(I) \to \mathcal{P}_{R-1}$ such that $Qv = v$ for all $v \in \mathcal{P}_{R-1}$. Such a projector is constructed, for example, in [19, sect. 4.1], where it is shown in [19, Prop. 4.3.8] that

$$\|u - Qu\|_{W^{r,q}(I)} \leq \|u^{(r)}\|_{L^q(I)}, \qquad 0 \leq r \leq R.$$

We claim that the operator $u \mapsto J_{R,N}(u - Qu) + Qu$ has all the desired properties. By exploiting the stability properties (A.2) of $J_{R,N}$ constructed so far we get

$$\begin{aligned}\|u - (J_{R,N}(u - Qu) + Qu)\|_{W^{j,q}(I)} &\leq C(N+1)^{-(r-j)}\|u - Qu\|_{W^{r,q}(I)}\\ &\leq C(N+1)^{-(r-j)}\|u^{(r)}\|_{L^q(I)}.\end{aligned}$$

This concludes the argument, since $J_{R,N}(u - Qu) + Qu = u$ for $u \in \mathcal{P}_{R-1}$. □

The one-dimensional operator $J_{R,N}$ of Proposition A.2 can be tensorized to yield the polynomial approximation result (Theorem 5.1) for functions defined on hyper cubes. This result is proved as follows.

Proof of Theorem 5.1. The operator $J_{R,N}$ is taken as the tensor product of the one-dimensional ones given by Proposition A.2. To simplify the notation, we will drop the indices R, N and write $J_1, \ldots, J_d$ to denote these one-dimensional operators and J to denote the tensor product. From Proposition A.2 we obtain the following stability and approximation results.

$$|J_i u|_{W^{l,q}(I_i)} \leq C|u|_{W^{l,q}(I_i)}, \qquad l = 0, \ldots, r, \tag{A.10}$$

$$|u - J_i u|_{W^{l,q}(I_i)} \leq C(N+1)^{-(r-l)}|u|_{W^{l,q}(I_i)}, \qquad 0 \leq l \leq r, \tag{A.11}$$

where $i = 1, \ldots, d$. These stability estimates then allow us to obtain approximation results in the standard way. We illustrate the procedure for the case $d = 2$. Let α, $\beta \geq 0$ with $\alpha + \beta = l \leq r$. Since the operators ∂_i and J_j commute if $i \neq j$, we get

$$\begin{aligned}\|\partial_1^\alpha \partial_2^\beta (u - J_1 \otimes J_2 u)\|_{L^q(I)} &\leq \|\partial_1^\alpha \partial_2^\beta (u - J_1 u)\|_{L^q(I)} + \|\partial_1^\alpha \partial_2^\beta J_1(u - J_2 u)\|_{L^q(I)}\\ &\leq \|\partial_1^\alpha \left((\partial_2^\beta u) - J_1(\partial_2^\beta u)\right)\|_{L^q(I)} + \|\partial_1^\alpha J_1 \partial_2^\beta (u - J_2 u)\|_{L^q(I)}.\end{aligned}$$

We consider the first term. The function $v(\cdot, x_2) = \partial_2^\beta u(\cdot, x_2)$ is defined for a.e. $x_2 \in I_2$ and $v(\cdot, x_2) \in W^{r-\beta,q}(I_1)$. Hence, we obtain from (A.11) for a.e. $x_2 \in I_2$

$$\|\partial_1^\alpha (v(\cdot, x_2) - J_1 v(\cdot, x_2))\|_{L^q(I_1)} \leq C(N+1)^{-(r-\beta-\alpha)}|v(\cdot, x_2)|_{W^{r-\beta,q}(I_1)}.$$

Substituting again the definition of v and integrating over I_2 yields

$$\|\partial_1^\alpha \left((\partial_2^\beta u) - J_1(\partial_2^\beta u)\right) \|_{L^q(I)} \leq C(N+1)^{-(r-\beta-\alpha)}|u|_{W^{r,q}(I)}.$$

By similar arguments (here, we additionally employ the stability result (A.10)) we can bound

$$\|\partial_1^\alpha J_1 \partial_2^\beta (u - J_2 u)\|_{L^q(I)} \leq C(N+1)^{-(r-\alpha-\beta)}|u|_{W^{r,q}(I)}.$$

Combining these last two estimates and summing over all combinations of α, β with $\alpha + \beta = l$ yields the desired bound. The fact (5.1) follows readily from the property (A.3) of the one-dimensional operators. □

Appendix B. Polynomial liftings. A general trace lifting operator of the form (B.1) was studied, for example, in [26, 30]. The subsequent observation that it also maps polynomials to polynomials (cf. Proposition B.1) was the basis for polynomial liftings from $H^{1/2}(\partial\hat{K})$ to $H^1(\hat{K})$, where $\hat{K}$ is the reference square or triangle [6, 14, 31]. We generalize these results to the L^q-setting. In principle, the techniques employed here are applicable to three-dimensional problems as well, although they are technically more involved. Polynomial lifting results for hexahedra, prisms, and tetrahedra are available (in Hilbert space settings) in [11, 12, 15, 35].

B.1. The operator $F^{[f]}$. We recall the definition of the reference triangle T in (2.1) and denote its bottom side by

$$\Gamma = \{(x, 0) \mid 0 < x < 1\}.$$

We will view Γ as embedded in $\mathbb{R}$ in the natural way. We choose $\alpha \in (0, 1)$ and define for a function $f \in L^q(\mathbb{R})$ the extension operator by

$$f \mapsto F^{[f]}(x, y) = \frac{1}{2\alpha y} \int_{x-\alpha y}^{x+\alpha y} f(t)\, dt. \tag{B.1}$$

Proposition B.1. *Let the extension operator be given by* (B.1). *Then $f \mapsto F^{[f]}$ is linear and $F^{[f]} \in \mathcal{P}_p$ if $f \in \mathcal{P}_p$. Furthermore, $F^{[f]}|_T$ depends only on the values of f on Γ, and for each $q \in (1, \infty)$ there exists a constant $C > 0$ such that for functions f defined on Γ the following bounds hold (provided that the right-hand side is finite):*

$$\|F^{[f]}\|_{L^q(T)} \leq C\|(x(1-x))^{1/q} f\|_{L^q(\Gamma)}, \tag{B.2}$$

$$\|F^{[f]}\|_{W^{1,q}(T)} \leq C\|f\|_{W^{1-1/q,q}(\Gamma)}, \tag{B.3}$$

$$\|(x-y)F^{[f/t]}\|_{L^q(T)} \leq C\|f\|_{L^q(\Gamma)}, \tag{B.4}$$

$$\|(x-y)(1-x-y)F^{[f/(t(1-t))]}\|_{L^q(T)} \leq C\|f\|_{L^q(\Gamma)}, \tag{B.5}$$

$$\|(x-y)F^{[f/t]}\|_{W^{1,q}(T)} \leq C\Bigg[\|f\|_{W^{1-1/q,q}(\Gamma)} + \left\|\frac{f(x)}{x^{1-1/q}}\right\|_{L^q(\Gamma)}\Bigg], \tag{B.6}$$

$$\|(x-y)(1-x-y)F^{[f/(t(1-t))]}\|_{W^{1,q}(T)} \leq C\|f\|_{\widetilde{W}^{1-1/q,q}(\Gamma)}. \tag{B.7}$$

Here, we employed the shorthand f/t to indicate the function $t \mapsto f(t)/t$ and $(x - y)F^{[f/t]}$ to denote the function $(x, y) \mapsto (x-y)F^{[f/t]}(x, y)$. Additionally, we have

$$\|F^{[f]}\|_{L^q(\partial T)} \leq C\|f\|_{L^q(\Gamma)}.$$

Proof. We first show (B.2). From (B.11) below we bound for each fixed $x \in (0,1)$

$$\int_{y=0}^{\min(x,1-x)} |F^{[f]}(x,y)|^q \, dy \le C \int_{x-\alpha\min(x,1-x)}^{x+\alpha\min(x,1-x)} |f(y)|^q \, dy.$$

Integrating over $x \in (0,1)$ we get from Lemma B.3

$$\|F^{[f]}\|^q_{L^q(T)} \le C \int_{y=0}^{1} y(1-y)|f(y)|^q \, dy = C\|(x(1-x))^{1/q} f\|^q_{L^q(\Gamma)}.$$

The estimate (B.4) follows immediately from (B.13) of Lemma B.2. The symmetry of T with respect to the line $x = 1/2$ together with (B.4) implies

$$\|(1-x-y)F^{[f/(1-t)]}\|_{L^q(T)} \le C\|f\|_{L^q(\Gamma)}.$$

From this and (B.4) we can easily obtain (B.5) if we observe

$$\frac{f(t)}{t(1-t)} = \frac{f(t)}{t} + \frac{f(t)}{1-t}.$$

It remains to obtain the bounds (B.3), (B.6), and (B.7). We compute

$$\begin{aligned}
\partial_x F^{[f]}(x,y) &= \frac{1}{2\alpha y}\left[f(x-\alpha y) - f(x+\alpha y)\right],\\
\partial_y F^{[f]}(x,y) &= \frac{1}{2\alpha}\left[-\frac{1}{y^2}\int_{x-\alpha y}^{x+\alpha y} f(t)\,dt + \frac{\alpha}{y}\left(f(x+\alpha y) + f(x-\alpha y)\right)\right]\\
&= -\frac{1}{2\alpha y^2}\int_{x-\alpha y}^{x+\alpha y} f(t) - f(x)\,dt - \frac{f(x) - f(x-\alpha y)}{2y} - \frac{f(x) - f(x+\alpha y)}{2y}.
\end{aligned}$$

From the definition of the $W^{1-1/q,q}$-norm and the bound (B.10) of Lemma B.2 we get

$$\|\nabla F^{[f]}\|_{W^{1,q}(T)} \le C\|f\|_{W^{1-1/q,q}(\Gamma)},$$

which shows (B.3). For the bound (B.6), we compute

$$\begin{aligned}
\partial_x\left((x-y)F^{[f/t]}\right) &= F^{[f/t]} + (x-y)\partial_x F^{[f/t]},\\
\partial_y\left((x-y)F^{[f/t]}\right) &= -F^{[f/t]} + (x-y)\partial_y F^{[f/t]},
\end{aligned}$$

and

$$\begin{aligned}
\partial_x F^{[f/t]} &= \frac{1}{2\alpha y}\left[\frac{f(x+\alpha y)}{x+\alpha y} - \frac{f(x-\alpha y)}{x-\alpha y}\right],\\
\partial_y F^{[f/t]} &= -\frac{1}{2\alpha y^2}\int_{x-\alpha y}^{x+\alpha y}\frac{f(t)}{t}\,dt + \frac{1}{2y}\left[\frac{f(x+\alpha y)}{x+\alpha y} + \frac{f(x-\alpha y)}{x-\alpha y}\right]\\
&= \frac{-1}{2\alpha y^2}\int_{x-\alpha y}^{x+\alpha y}\frac{f(t)}{t} - \frac{f(x)}{x}\,dt + \frac{1}{2y}\left[\frac{f(x+\alpha y)}{x+\alpha y} - \frac{f(x)}{x} + \frac{f(x-\alpha y)}{x-\alpha y} - \frac{f(x)}{x}\right].
\end{aligned}$$

With (B.9) and Lemma B.3 we can bound

$$\begin{aligned}
\|F^{[f/t]}\|^q_{L^q(T)} &\le C\int_{x=0}^{1}\int_{x-\alpha\min(x,1-x)}^{x+\alpha\min(x,1-x)}\left|\frac{f(t)}{t}\right|^q dt\,dx \le C\int_{x=0}^{1} x(1-x)\left|\frac{f(x)}{x}\right|^q dx\\
&\le C\int_{x=0}^{1}\left|\frac{f(x)}{x^{1-1/q}}\right|^q dx. \qquad \text{(B.8)}
\end{aligned}$$

The estimate (B.15) of Lemma B.2 implies

$$\int_T \left| \frac{(x-y)}{y^2} \int_{x-\alpha y}^{x+\alpha y} \frac{f(t)}{t} - \frac{f(x)}{x}\, dt \right|^q dx\, dy \le C \left[\|f\|^q_{W^{1-1/q,q}(\Gamma)} + \|f(x)/x\|_{L^q(\Gamma)} \right].$$

It remains to bound terms of the form

$$\left\| \frac{x-y}{y} \left(\frac{f(x \pm \alpha y)}{x \pm \alpha y} - \frac{f(x)}{x} \right) \right\|_{L^q(T)}.$$

Rearranging terms, we arrive at

$$\left\| \frac{x-y}{y} \left(\frac{f(x \pm \alpha y)}{x \pm \alpha y} - \frac{f(x)}{x} \right) \right\|_{L^q(T)} =$$
$$\left\| \frac{x-y}{x \pm \alpha y} \frac{f(x \pm \alpha y) - f(x)}{y} - \frac{\pm\alpha(x-y)}{x \pm \alpha y} \frac{f(x)}{x} \right\|_{L^q(T)} \le |f|_{W^{1-1/q,q}(\Gamma)} + \left\| \frac{f(x)}{x} \right\|_{L^q(T)},$$

where we employed the observation $|\frac{x-y}{x \pm \alpha y}| \le 1$ for $0 < y < x$. The term $\|\frac{f(x)}{x}\|_{L^q(T)}$ is now controlled in the desired fashion as in (B.8).

It remains to show (B.7). By symmetry considerations we obtain, analogous to (B.6),

$$\|(1-x-y)F^{[f/(1-t)]}\|_{W^{1,q}(T)} \le C \left[\|f\|_{W^{1-1/q,q}(\Gamma)} + \left\| \frac{f(x)}{(1-x)^{1-1/q}} \right\|_{L^q(\Gamma)} \right].$$

Since $\frac{f(t)}{t(1-t)} = \frac{f(t)}{t} + \frac{f(t)}{1-t}$, the desired bound (B.7) now follows easily. □

The following lemma contains estimates of Hardy type.

LEMMA B.2. *Let $a \le b$, $\alpha \in (0,1)$, T be the reference triangle. Then for $q \in (1,\infty)$*

$$\int_a^b \left| \frac{1}{x-a} \int_a^x |g(\xi)|\, d\xi \right|^q \le \left(\frac{q}{q-1} \right)^q \int_a^b |g(\xi)|^q\, d\xi, \tag{B.9}$$

$$\int_a^b \left| \frac{1}{(x-a)^2} \int_a^x g(\xi) - g(a)\, d\xi \right|^q dx \le \left(\frac{q}{q-1} \right)^q \int_a^b \left| \frac{g(\xi) - g(a)}{\xi - a} \right|^q d\xi. \tag{B.10}$$

Furthermore, for each $x \in (0,1)$ we have upon setting $m := \min(x, 1-x)$

$$\int_{y=0}^m \left| \frac{1}{y} \int_{x-\alpha y}^{x+\alpha y} g(t)\, dt \right|^q \le \left(\frac{q}{q-1} \right)^q \alpha^{q-1} \int_{x-\alpha m}^{x+\alpha m} |g(y)|^q\, dy, \tag{B.11}$$

$$\int_{y=0}^m \frac{1}{y^2} \int_{x-\alpha y}^{x+\alpha y} |g(t) - g(x)|\, dt\, dy \le \alpha^{2q-1} \left(\frac{q}{q-1} \right)^q \times \int_{x-\alpha m}^{x+\alpha m} \left| \frac{g(t) - g(x)}{t-x} \right|^q dt. \tag{B.12}$$

Finally, we have for some constant $C > 0$ that depends only on q and α,

$$\int_T \left| (x-y) \frac{1}{y} \int_{x-\alpha y}^{x+\alpha y} \frac{g(t)}{t}\, dt \right|^q dx\, dy \le C \|g\|^q_{L^q(\Gamma)}, \tag{B.13}$$

(B.14)

$$\int_T \left| \frac{x-y}{y^2} \int_{x-\alpha y}^{x+\alpha y} \frac{g(t)}{t} - \frac{g(x)}{x}\,dt \right|^q dxdy \le C \Big[\|g\|^q_{W^{1-1/q,q}(\Gamma)} + \left\| \frac{g(x)}{x} \right\|_{L^q(\Gamma)} \Big].$$

In all the above estimates, it is implicitly assumed that the right-hand side is finite.

Proof. The first estimate is the well-known Hardy inequality [29, Thm. 327]. For the second estimate, we note

$$\begin{aligned} \int_a^b \left| \frac{1}{|x-a|^2} \int_a^x |g(\xi)-g(a)| \right|^q dx &= \int_a^b \left| \frac{1}{|x-a|^2} \int_a^x \frac{|g(\xi)-g(a)|}{|\xi-a|} |\xi - a|\,d\xi\,dx \right|^q \\ &\le \int_a^b \left| \frac{1}{|x-a|} \int_a^x \frac{|g(\xi)-g(a)|}{|\xi-a|}\,d\xi \right|^q dx. \end{aligned}$$

The result (B.10) now follows from (B.9). The bounds (B.11), (B.12) follow from (B.9) and (B.10), respectively. To proceed further, we note that for $x \in (0,1)$ we have

(B.15) $$(1-\alpha)\,x \le x - \alpha y \le x + \alpha y \le (1+\alpha)x, \qquad 0 \le y \le \min(x, 1-x).$$

We are now in a position to prove (B.13). From (B.15) and (B.11) we get (again with the abbreviation $m = \min(x, 1-x)$)

$$\int_T \left| \frac{(x-y)}{y} \int_{x-\alpha y}^{x+\alpha y} \frac{g(t)}{t}\,dt \right|^q dx \le C \int_T \left| \frac{1}{y} \int_{x-\alpha y}^{x+\alpha y} g(t)\,dt \right|^q \le C \|g\|^q_{L^q(\Gamma)}.$$

We now turn to the proof of the last inequality, (B.15). We employ (B.12) and obtain

$$\int_T \left| \frac{x-y}{y^2} \int_{x-\alpha y}^{x+\alpha y} \frac{g(t)}{t} - \frac{g(x)}{x}\,dt \right|^q dy\,dx \le C \int_{x=0}^1 x^q \int_{t=x-\alpha m}^{x+\alpha m} \left| \frac{\frac{g(t)}{t} - \frac{g(x)}{x}}{t-x} \right|^q dt\,dx.$$

We next rewrite the integrand as

$$\frac{g(t)}{t} - \frac{g(x)}{x} = \frac{g(t)-g(x)}{t} - \frac{t-x}{tx} g(x)$$

and arrive at

$$\begin{aligned} &\int_T \left| \frac{x-y}{y^2} \int_{x-\alpha y}^{x+\alpha y} \frac{g(t)}{t} - \frac{g(x)}{x}\,dt \right|^q dy\,dx \\ &\le C \int_{x=0}^1 x^q \int_{t=x-\alpha m}^{x+\alpha m} t^{-q} \left| \frac{g(t)-g(x)}{t-x} \right|^q + C \int_{x=0}^1 x^q \int_{t=x-\alpha m}^{x+\alpha m} |g(x)|^q \frac{1}{|tx|^q}\,dt\,dx \\ &\le C \int_{\Gamma\times\Gamma} \left| \frac{g(t)-g(x)}{t-x} \right|^q dt\,dx + C \int_\Gamma \left| \frac{g(x)}{x^{1-1/q}} \right|^q dx. \end{aligned}$$

This concludes the proof of the lemma. □

LEMMA B.3. *Let $\alpha \in (0,1)$. Then for some $C > 0$ depending solely on α,*

$$\int_{x=0}^1 \int_{y=x-\alpha\min(x,1-x)}^{x+\alpha\min(x,1-x)} |g(y)|\,dy\,dx \le C \int_{y=0}^1 y(1-y)|g(y)|\,dy.$$

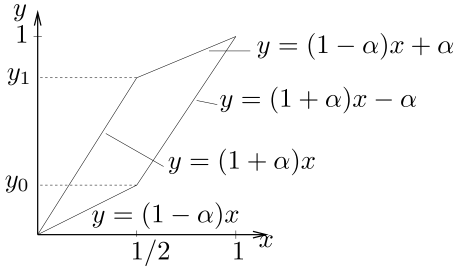


FIG. B.1. *Integration domain in Lemma* B.3.

Proof. The integration domain is sketched in Figure B.1. Interchanging the order of integration, we get

$$\int_{x=0}^{1} \int_{y=x-\alpha \min(x,1-x)}^{x+\alpha \min(x,1-x)} |g(y)|\, dy\, dx = \int_{y=0}^{y_0} \int_{x=y/(1+\alpha)}^{y/(1-\alpha)} |g(y)|\, dx\, dy$$
$$+ \int_{y=y_0}^{y_1} \int_{x=y/(1+\alpha)}^{(y+\alpha)/(1+\alpha)} |g(y)|\, dx\, dy + \int_{y=y_1}^{1} \int_{x=(y-\alpha)/(1-\alpha)}^{(y+\alpha)/(1+\alpha)} |g(y)|\, dx\, dy,$$

where $y_0 = \frac{1-\alpha}{2}$ and $y_1 = \frac{1+\alpha}{2}$. The result follows by elementary calculations. □

B.2. Polynomial lifting from the boundary.

B.2.1. $W^{1,q}$-stable liftings. The operator $f \mapsto F^{[f]}$ is the basic building block for the polynomial trace liftings of Theorem 5.2, which we now prove.

Proof of Theorem 5.2. We consider the case $\hat{K} = T$. Three cases may occur:

Γ *is a single edge:* We may assume $\Gamma = \{(x,0)\,|\, 0 < x < 1\}$ and choose F as $F^{[f]}$.

Γ *consists of two edges* Γ_1, Γ_2*:* The function F is defined in two steps. First, the lifting operator F of section B.1 is employed to construct a function $F_1 \in \mathcal{P}_p$ with $F_1|_{\Gamma_1} = f|_{\Gamma_1}$ and

$$\|F_1\|_{W^{1,q}(T)} \leq C\|f\|_{W^{1-1/q,q}(\Gamma_1)}, \quad \|F_1\|_{L^q(\Gamma_2)} + \|F_1\|_{L^q(T)} \leq C\|f\|_{L^q(\Gamma_1)}.$$

We have thus reduced the problem to one where f vanishes on one of the sides Γ_1, Γ_2. Without loss of generality, we assume that $\Gamma_1 = \{(x,0)\,|\, 0 < x < 1\}$ and $f|_{\Gamma_2} = 0$ with $\Gamma_2 = \{(x,x)\,|\, 0 < x < 1/2\}$. The mapping $f \mapsto (x-y)F^{[f/t]}(x,y)$ of section B.1 then has the desired properties.

$\Gamma = \partial K$*:* After having constructed a lifting from two adjacent edges as in the above construction, we may assume that f vanishes on two sides of T. Without loss of generality, we may therefore assume that $f|_{\Gamma_i} = 0$ for $i \in \{2,3\}$, where Γ_1 is the third side of ∂T given by $\Gamma_1 = \{(x,0)\,|\, 0 < x < 1\}$. The construction of a polynomial F with the property $F|_{\Gamma_1} = f$ and $F|_{\Gamma_i} = 0$ for $i \in \{2,3\}$ is then achieved with the operator $f \mapsto (x-y)(1-x-y)F^{[f/(t(1-t))]}(x,y)$ of section B.1.

The case of a square is proved similarly using the ideas of [6]. Note that in the case of a square, the set Γ may be disconnected, i.e., it consists of two parallel edges of $\hat{K}$. In this event, we easily reduce the construction to the case where f vanishes on one of the two edges and construct $F \in \mathcal{Q}_p$ in the same way as the function U in the proof of Lemma C.1.

We finally turn to the statement that constant functions are reproduced. This follows from the observation that the lifting operator constructed above is independent of the polynomial degree p. Since for $p = 0$ the constant function is reproduced, this operator reproduces constants for any $p \in \mathbb{N}_0$. □

B.2.2. Polynomial liftings with improved L^q-bounds. The basis of the results of section B.2.1 is the operator $f \mapsto F^{[f]}$ of section B.1. We introduce a new operator $\widetilde{F}^{[f]}$ by

$$\widetilde{F}^{[f]}(x,y) = (1-y)^p F^{[f]}(x,y). \tag{B.16}$$

We note that, if $f \in \mathcal{P}_p$, then $\widetilde{F} \in \mathcal{P}_{2p}$. Furthermore, $\widetilde{F}^{[f]}|_\Gamma = f$, where $\Gamma = \{(x,0)\,|\,0<x<1\}$. We also have the following lemma.

Lemma B.4. *Let T be the reference triangle. Then there exists $C>0$ such that for every $p \in \mathbb{N}$ the functions $\widetilde{F}^{[f]}$, $F_1 := (x-y)\widetilde{F}^{[f/t]}$, $F_2 := (x-y)(1-x-y)\widetilde{F}^{[f/(t(1-t))]}$ satisfy*

$$\begin{aligned}
p\|\widetilde{F}^{[f]}\|_{L^q(T)} + \|\widetilde{F}^{[f]}\|_{W^{1,q}(T)} &\le C\|f\|_{W^{1-1/q,q}(\Gamma)} + Cp^{1-1/q}\|f\|_{L^q(\Gamma)},\\
p\|F_1\|_{L^q(T)} + \|F_1\|_{W^{1,q}(T)} &\le C\|f\|_{W^{1-1/q,q}(\Gamma)} + C\|\frac{f(x)}{x^{1-1/q}}\|_{L^q(\Gamma)} + Cp^{1-1/q}\|f\|_{L^q(\Gamma)},\\
p\|F_2\|_{L^q(T)} + \|F_2\|_{W^{1,q}(T)} &\le C\|f\|_{\widetilde{W}^{1-1/q,q}(\Gamma)} + Cp^{1-1/q}\|f\|_{L^q(\Gamma)}.
\end{aligned}$$

Furthermore,

$$\|\widetilde{F}^{[f]}\|_{L^q(\partial T)} \le C\|f\|_{L^q(\Gamma)}.$$

Proof. The lemma follows from Lemma B.5 below and the properties of the operator $F^{[f]}$ of section B.1. □

This lemma allows us now to prove Proposition 5.3.

Proof of Proposition 5.3. The proof is similar to that of Theorem 5.2. The appeals to Proposition B.1 are replaced with those to Lemma B.4. □

Lemma B.5. *K be the reference triangle or the reference square. Set $\Gamma = \{(x,0)\,|\,0<x<1\}$ and let $q \in (1,\infty)$. Then there exists $C>0$ such that for every $p \in \mathbb{N}$ and every function $g \in W^{1,q}(K)$*

$$p\|(1-y)^p g\|_{L^q(K)} + \|(1-y)^p g\|_{W^{1,q}(K)} \le C|g|_{W^{1,q}(K)} + p^{1-1/q}\|g\|_{L^q(\Gamma)}.$$

Proof. We express the function g for $y>0$ as $g(x,y) = g(x,0) + \int_{t=0}^y g_y(x,t)\,dt$. Then

$$(1-y)^p g(x,y) = [(1-y)^p y]\,\frac{1}{y}\int_0^y g_y(x,t)\,dt + (1-y)^p g(x,0).$$

Since $\sup_{y\in(0,1)}(1-y)^p y \le \frac{C}{p}$, we conclude with the Hardy inequality (B.9)

$$\|(1-y)^p g\|_{L^q(K)} \le \frac{C}{p}|g|_{W^{1,q}(K)} + p^{-1/q}\|g\|_{L^q(\Gamma)}.$$

For the bound on the derivative of $(1-y)^p g$, we write

$$\nabla((1-y)^p g) = p(1-y)^{p-1}g + (1-y)^p\nabla g;$$

we treat the first term as above and for the second term we use

$$|1-y| \le 1 \text{ on } K. \qquad \square$$

Appendix C. One-dimensional extension operators.

Proof of Lemma 5.4. The key to this result is the following approximation result of [18, Cor. 3.7]:

$$\inf_{v \in \mathcal{P}_{p-k}} \|u - v\|_{L^\infty(J)} \leq 12(4p)^{k-1} |J|^{p-k+1} \|u\|_{L^\infty(\hat{I})} \qquad \forall u \in \mathcal{P}_p, \tag{C.1}$$

where $J \subset \hat{I}$ is an arbitrary subinterval of $\hat{I}$ and $|J|$ denotes the length of J. We (arbitrarily) choose $J = (0, 1/2)$, denote by $\mathcal{I}_{p-k} : C(\overline{J}) \to \mathcal{P}_{p-k}$ the Gauß–Lobatto interpolation operator and set

$$(Z_{p,p-k} u)(x) = (\mathcal{I}_{p-k} u)(2x).$$

By construction $(Z_{p,p-k} u) \in \mathcal{P}_{p-k}$. Additionally, the fact that the endpoint 0 of J is an interpolation point implies $(Z_{p,p-k} u)(0) = u(0)$. In order to see the remaining estimates, we see that (C.1) together with standard inverse estimates (see, e.g., [24, Chap. 4, Thms. 1.4 and 2.6]) implies

$$\|u - \mathcal{I}_{p-k} u\|_{W^{1,\infty}(J)} \leq C p^2 \|u - \mathcal{I}_{p-k} u\|_{L^\infty(J)} \leq C \rho^{p-k} \|u\|_{L^\infty(\hat{I})} \leq C \tilde{\rho}^{p-k} \|u\|_{L^q(\hat{I})}$$

for some suitable $C > 0$ and $\tilde{\rho} \in (0, 1)$ that are both independent of p and u. In particular, since $u(0) = (\mathcal{I}_{p-k} u)(0)$, we get

$$\max_{x \in J} \frac{|u(x) - (\mathcal{I}_{p-k} u)(x)|}{|x|} \leq \|u' - (\mathcal{I}_{p-k} u)'\|_{L^\infty(J)} \leq C \tilde{\rho}^{p-k} \|u\|_{L^q(\hat{I})}.$$

From this and the triangle inequality, we can easily infer the estimates

$$\|\mathcal{I}_{p-k} u\|_{L^q(J)} \leq C \|u\|_{L^q(\hat{I})},$$
$$\|\mathcal{I}_{p-k} u\|_{W^{1-1/q,q}(J)} \leq C \|u\|_{W^{1-1/q,q}(\hat{I})},$$
$$\int_0^{1/2} \frac{|u(x) - \mathcal{I}_{p-k} u(x)|^q}{x^{q-1}} \, dx \leq C \|u\|^q_{L^q(\hat{I})}.$$

This and the change of variables $x \mapsto 2x$ imply the desired bounds for $Z_{p,p-k}$. □

A closely related extension result is the following.

Lemma C.1. *Let $\hat{I} = (0, 1)$ and $q \in (1, \infty)$. Then there exists a bounded linear operator $Z : W^{1-1/q,q}(\hat{I}) \to W^{1-1/q,q}(\hat{I})$ with the following properties:*

1. $\int_{x=0}^1 \frac{|u(x) - Zu(x)|^q}{x^{q-1}} \, dx \leq C \|u\|^q_{W^{1-1/q,q}(\hat{I})}$;
2. *if $u \in \mathcal{P}_p$, then $Zu \in \mathcal{P}_p$;*
3. $Zu(1) = 0$ *and* $\int_{-1}^0 \frac{|Zu(x)|^q}{(1-x)^{q-1}} \, dx \leq C \|u\|^q_{W^{1-1/q,q}(\hat{I})}$;
4. $\|Zu\|_{L^q(I)} \leq C \|u\|_{L^q(\hat{I})}$.

Proof. Let T be the reference triangle and identify $\hat{I}$ with the edge of $\hat{K}$ lying on the x-axis. Consider the trapezoid $\tilde{T} := \{(x, y) \in T \,|\, 0 < y < 1/4\}$ and define $\Gamma = \{(x, x) \,|\, 0 < x < 1/4\}$. An elementary calculation reveals

$$\|F^{[u]}(\cdot, 1/4)\|_{W^{1,\infty}((1/4,3/4))} \leq C \|u\|_{L^q(\hat{I})}$$

for some appropriate $C > 0$. Hence, the function

$$U(x, y) = F^{[u]}(x, y) - 4y F^{[u]}(x, 1/4)$$

satisfies

$$\|U\|_{W^{1,q}(\tilde{T})} \leq C\|u\|_{W^{1-1/q,q}(\hat{I})}, \quad \|U\|_{L^q(\Gamma)} \leq C\|u\|_{L^q(\hat{I})}, \quad U|_{y=1/4} = 0.$$

Additionally, if $u \in \mathcal{P}_p$, then $U \in \mathcal{Q}_p$. Defining $Zu(x)$ for $x \in (0,1)$ by $(Zu)(x) = U(x/4, x/4)$ and appealing to the trace theorem concludes the proof. □

Lemma C.2. *Let $\hat{I} = (0,1)$ and $q \in (1,\infty)$. Then $u \in W^{1-1/q,q}(\hat{I})$ implies that the function $\tilde{u} : x \mapsto xu(x)$ is in $W^{1-1/q,q}(\hat{I})$ and satisfies*

$$\int_0^1 \frac{|\tilde{u}(x)|^q}{x^{q-1}}\, dx + \|\tilde{u}\|^q_{W^{1-1/q,q}(\hat{I})} \leq C\|u\|^q_{W^{1-1/q,q}(\hat{I})}$$

for some $C > 0$ that is independent of u.

Proof. The estimate $\|\tilde{u}\|_{W^{1-1/q,q}(\hat{I})} \leq C\|u\|_{W^{1-1/q,q}(\hat{I})}$ follows from the smoothness of the function $x \mapsto x$. The remaining estimate follows by inspection. □

Acknowledgments. The author is grateful to the Isaac Newton Institute for Mathematical Sciences for their hospitality. This research was carried out in conjunction with the author's participation in the institute's program on Computational Challenges in Partial Differential Equations.

REFERENCES

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

[2] M. Ainsworth and D. Kay, *The approximation theory for the p-version finite element method and application to non-linear elliptic PDEs*, Numer. Math., 82 (1999), pp. 351–388.

[3] M. Ainsworth and D. Kay, *Approximation theory for the hp-version finite element method and application to the non-linear Laplacian*, Appl. Numer. Math., 34 (2000), pp. 329–344.

[4] M. Ainsworth and T.J. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley, New York, 2000.

[5] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Leipzig, 1999.

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

[7] I. Babuška and T. Strouboulis, *The Finite Element Method and its Reliability*, Oxford University Press, Oxford, UK, 2001.

[8] I. Babuška and M. Suri, *The optimal convergence rate of the p-version of the finite element method*, SIAM J. Numer. Anal., 24 (1987), pp. 750–776.

[9] Ivo Babuska and Benqi Guo, *Optimal estimates for lower and upper bounds of approximation errors in the p-version of the finite element method in two dimensions*, Numer. Math., 85 (2000), pp. 219–255.

[10] Ivo Babuska and Benqi Guo, *Direct and inverse approximation theorems for the p-version of the finite element method in the framework of weighted Besov spaces.* I: *Approximability of functions in the weighted Besov spaces*, SIAM J. Numer. Anal., 39 (2001), pp. 1512–1538.

[11] F. Ben Belgacem, *Relèvement polynômiaux sur un cube*, Technical report HI-72/7780, Electricité de France, Clamart, France, 1992.

[12] F. Ben Belgacem, *Polynomial extensions of compatible polynomial traces in three dimensions*, Comput. Meth. Appl. Mech. Engrg., 116 (1994), pp. 235–241.

[13] C. Bernardi, *Optimal finite element interpolation*, SIAM J. Numer. Anal., 26 (1989), pp. 430–454.

[14] C. Bernardi, M. Dauge, and Y. Maday, *Trace liftings which preserve polynomials*, C. R. Acad. Sci. Paris Sér. I, 315 (1992), pp. 333–338.

[15] C. Bernardi, M. Dauge, and Y. Maday, *Interpolation of nullspaces for polynomial approximation of divergence-free functions in a cube*, in Boundary value Problems and Integral Equations in Nonsmooth Domains, Lectures Notes in Pure and Appl. Math., 167, M. Costabel, M. Dauge, and S. Nicaise, eds., Dekker, New York, 1995, pp. 27–46.

[16] C. Bernardi and V. Girault, *A local regularization operator for triangular and quadrilateral finite elements*, SIAM J. Numer. Anal., 35 (1998), pp. 1893–1916.

[17] C. Bernardi and Y. Maday, *Spectral methods*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., Vol. 5, North-Holland, Amsterdam, 1997, pp. 209–485.

[18] S. Börm, M. Löhndorf, and J. M. Melenk, *Approximation of integral operators by variable-order interpolation*, Numer. Math., 99 (2005), pp. 605–643.

[19] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1994.

[20] C. Carstensen, *Quasi interpolation and a posteriori error analysis in finite element methods*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1187–1202.

[21] C. Carstensen and R. Verfürth, *Edge residuals dominate a posteriori error estimates for low order finite element methods*, SIAM J. Numer. Anal., 36 (1999), pp. 1571–1587.

[22] Ph. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numérique, 9 (1975), pp. 77–84.

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

[24] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.

[25] T. Eibner and J. M. Melenk, *Local error analysis of the boundary concentrated fem*, IMA J. Numer. Anal., to appear.

[26] E. Gagliardo, *Caratterizzazione delle tracce sulla frontiera relative ad alcune classi di funzioni in n variabili*, Rend. Sem. Mat. Univ. Padova, 27 (1957), pp. 284–305.

[27] V. Girault and L. R. Scott, *Hermite interpolation of nonsmooth functions preserving boundary conditions*, Math. Comput., 73 (2002), pp. 1043–1074.

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

[29] G. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, in Cambridge Mathematical Library, Cambridge University Press, Cambridge 1991.

[30] J. L. Lions, *Théorèmes de trace et d'interpolation (*IV*)*, Math. Annalen, 152 (1963), pp. 42–56.

[31] Y. Maday, *Relèvement de traces polyômiales et interpolations hilbertiennes entres espaces de polynômes*, C. R. Acad. Sci. Paris, Série I, 309 (1989), pp. 463–468.

[32] J. M. Melenk, *hp-interpolation of nonsmooth functions*, Technical report NI03050, Isaac Newton Institute for Mathematics Sciences, Cambridge, UK, 2003.

[33] J. M. Melenk and I. Babuška, *The partition of unity finite element method: basic theory and applications*, Comput. Meth. Appl. Mech. Engrg., 139 (1996), pp. 289–314.

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

[35] R. Muñoz-Sola, *Polynomial liftings on a tetrahedron and applications to the hp-version of the finite element method in three dimensions*, SIAM J. Numer. Anal., 34 (1997), pp. 282–314.

[36] N. Neuss, *Homogenisierung und Mehrgitter*, Ph.D. thesis, University of Heidelberg, 1996.

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

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

[39] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

[40] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Teubner-Wiley, Leipzig, 1996.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 156–173

AN OPTIMAL A PRIORI ERROR ESTIMATE FOR NONLINEAR MULTIBODY CONTACT PROBLEMS*

S. HÜEBER[†] AND B. I. WOHLMUTH[†]

Abstract. Nonconforming domain decomposition methods provide a powerful tool for the numerical approximation of partial differential equations. For the discretization of a nonlinear multibody contact problem, we use linear mortar finite elements based on dual Lagrange multipliers. Under some regularity assumptions on the solution, an optimal convergence order of $h^{0.5+\nu}$, $0 < \nu \leq 0.5$, can be established in two dimensions (2D) and three dimensions (3D). Compared with a standard linear saddle point formulation, two additional terms which provide a measure for the nonconformity and the nonlinearity of the approach have to be taken in account. Numerical examples illustrating the performance of the nonconforming method and confirming our theoretical result are presented.

Key words. multibody contact problems, a priori error estimates, mortar finite element methods, dual Lagrange multipliers, linear elasticity

AMS subject classifications. 65N15, 65N30, 74B10, 74M15

DOI. 10.1137/S0036142903436678

1. Introduction.

The numerical approximation of nonlinear multibody contact problems is a challenging task. Modern discretization techniques are very often based on flexible nonconforming approaches. Here, we provide a new optimal a priori bound for the discretization error of the mortar method. We focus on piecewise linear dual Lagrange multipliers which have been introduced for the lowest-order case in [Woh00]. Dual Lagrange multiplier spaces are based on a biorthogonal basis resulting in a diagonal mass matrix. Thus we obtain for each node on the slave side a local nonpenetration condition. The main advantage of dual Lagrange multipliers is that fast and efficient monotone multigrid methods [KK01] can be generalized to multibody contact problems [WK03]. In case of standard Lagrange multipliers, local monotone multigrid methods cannot be applied.

In this paper, we give a new optimal a priori error estimate for the discretization error in the H^1-norm for the displacements and in the $H^{-1/2}$-norm for the Lagrange multiplier. The interest in contact problems and variational inequalities has led to an increased research activity in this field. Abstract error estimates for variational inequalities can be found, e.g., in [Fal74, BHR77, Glo84] and a priori bounds for the discretization error of unilateral contact problems are given, e.g., in [Hc81, Hcc96]. Recently a lot of work has been done to analyze mortar formulations based on standard Lagrange multipliers. A priori error estimates for the displacements in the H^1-norm and for the Lagrange multiplier in the $H^{-1/2}$-norm of order $h^{0.75}$ have been established; see, e.g., [BR03, LS99, BHL99, CHLS01], under H^2-regularity assumption. Using additional regularity assumptions on the Lagrange multiplier, order h has been shown; see, e.g., [Hil00, CHLS01]. Although the order h is optimal, the regularity assumptions are quite strong and restrictive. These first a priori results have been considerably improved during the last couple of years. In [Ben00], order

*Received by the editors October 20, 2003; accepted for publication (in revised form) July 15, 2004; published electronically April 26, 2005. This work was supported in part by the Deutsche Forschungsgemeinschaft, SFB 404, C12.

http://www.siam.org/journals/sinum/43-1/43667.html

[†]Institute of Applied Analysis and Numerical Simulations (IANS), University of Stuttgart, Germany (hueeber@ians.uni-stuttgart.de, wohlmuth@ians.uni-stuttgart.de).

$h^{0.5+\nu}$, $0 < \nu < 0.5$, and order $h\sqrt{|\log h|}$ a priori results are given for the H^1-norm if the solution is $H^{3/2+\nu}$- and H^2-regular, respectively. A new and stronger result is given in [BR03], where order $h\sqrt[4]{|\log h|}$ can be found. In [HL02], quadratic finite elements are discussed and a priori estimates for the discretization error are given. Most of the theoretical results are obtained for standard Lagrange multipliers and in the two-dimensional setting.

In this paper, we consider the case of linear mortar finite elements based on dual Lagrange multipliers in two dimensions (2D) and three dimensions (3D). Moreover, we show order $h^{0.5+\nu}$ a priori estimates for the displacements and the Lagrange multiplier if the solution is $H^{3/2+\nu}$-regular, $0 < \nu \leq 0.5$. The techniques are based on introducing locally defined truncation operators measuring the nonconformity of the discretization and on Sobolev–Slobodeckij norms.

The rest of this paper is organized as follows. In section 2, we present the mathematical formulation of the nonlinear multibody contact problem. In section 3, we then consider the mortar discretization technique and establish the optimal a priori estimate. Finally in section 4, we provide several numerical examples illustrating the performance of the approach and confirming our theoretical results.

2. Unilateral Signorini contact problem. In this section, we give the formulation of a unilateral contact problem in linear elasticity. Let Ω_k, $k = m, s$, denote two elastic bodies, where $\Omega_k \subset \mathbb{R}^d$, $d = 2, 3$, are two bounded polyhedral domains. The nonmortar side is associated with the subdomain Ω_s and the mortar side with the subdomain Ω_m. We recall that in the mortar setting, the Lagrange multiplier is defined on the nonmortar side and that the displacements on the nonmortar side depend on the ones on the mortar side. Mortar and nonmortar sides are also called master and slave sides, respectively. This motivates the subscript s for the nonmortar side and the subscript m for the mortar side. Let $\Gamma_k := \partial\Omega_k$ be $\Gamma_k = \overline{\Gamma_{k,D}} \cup \overline{\Gamma_{k,N}} \cup \overline{\Gamma_{k,C}}$ with disjoint open subsets $\Gamma_{k,i}$, $i = D, N, C$. We assume homogeneous Dirichlet data on $\Gamma_D := \Gamma_{s,D} \cup \Gamma_{m,D}$ and we assume that $\Gamma_{k,D}$, $k = m, s$, has a nonzero measure. On $\Gamma_N := \Gamma_{s,N} \cup \Gamma_{m,N}$ Neumann data are given, and $\Gamma_{k,C}$ denotes the possible contact zone between the two bodies. Furthermore, we assume that the two bodies in the initial configuration are in contact on their common boundary part, i.e., $\Gamma_C := \Gamma_{m,C} = \Gamma_{s,C}$ and that $\overline{\Gamma_C}$ is a compact subset of $\partial\Omega_s \backslash \overline{\Gamma_{s,D}}$. The normal vector $\mathbf{n} := \mathbf{n}_s = -\mathbf{n}_m$ is assumed to be constant on Γ_C. Let $\mathbf{f}_k \in [L^2(\Omega_k)]^d$ be the volume forces on Ω_k. On each domain Ω_k, $k = m, s$, we have to consider the boundary value problem

$$
\begin{aligned}
-\operatorname{div} \boldsymbol{\sigma}_k(\mathbf{u}_k) &= \mathbf{f}_k \quad \text{in} \quad \Omega_k,\\
\mathbf{u}_k &= \mathbf{0} \quad \text{on} \quad \Gamma_{k,D},\\
\boldsymbol{\sigma}_k(\mathbf{u}_k)\,\mathbf{n}_k &= \mathbf{p}_k \quad \text{on} \quad \Gamma_{k,N}
\end{aligned}
\tag{2.1}
$$

with given boundary stresses $\mathbf{p}_k$ on $\Gamma_{k,N}$. The stress tensor $\boldsymbol{\sigma}_k$ on Ω_k is given by Hooke's law

$$\boldsymbol{\sigma}_{k,ij} := \mathbf{C}_{k,ijml}\boldsymbol{\epsilon}_{k,ml},$$

where the components of the symmetric and positive definite tensor $\mathbf{C}_k$ are

$$\mathbf{C}_{k,ijml} := \lambda_k\,\delta_{ij}\delta_{ml} + \mu_k\,(\delta_{im}\delta_{jl} + \delta_{il}\delta_{jm})$$

with the constant Lamé parameters λ_k and μ_k on Ω_k. Here $\boldsymbol{\epsilon}_k$ denotes the linearized strain tensor

$$\boldsymbol{\epsilon}_k(\mathbf{u}_k) := \frac{1}{2}\Big(\nabla\mathbf{u}_k + \big(\nabla\mathbf{u}_k\big)^\top\Big).$$

We refer the reader to [KO88, Wri01, Lau02] for an introduction to linear elasticity and contact problems. To specify the conditions modeling unilateral contact on the possible contact part Γ_C, we define for $k = m, s$ the normal stress $\sigma_{k,n}(\mathbf{u}_k)$ and the tangential stress $\boldsymbol{\sigma}_{k,T}$ on Γ_C by

$$\sigma_{k,n}(\mathbf{u}_k) := \big(\boldsymbol{\sigma}_k\left(\mathbf{u}_k\right)\mathbf{n}_k\big)\cdot\mathbf{n}_k \quad \text{and} \quad \boldsymbol{\sigma}_{k,T}\left(\mathbf{u}\right) := \boldsymbol{\sigma}_k\left(\mathbf{u}_k\right)\mathbf{n}_k - \sigma_{k,n}(\mathbf{u}_k)\,\mathbf{n}_k,$$

respectively. We consider frictionless unilateral contact problems, thus the contact conditions on the possible contact boundary Γ_C for $k = m, s$ are given by

$$\begin{aligned} -\boldsymbol{\sigma}_{s,T}\left(\mathbf{u}_s\right) = \boldsymbol{\sigma}_{m,T}\left(\mathbf{u}_m\right) &= 0, \\ \sigma_{m,n}(\mathbf{u}_m) = \sigma_{s,n}(\mathbf{u}_s) &\leq 0, \\ [\mathbf{u}\cdot\mathbf{n}] &\leq 0, \\ \sigma_n(\mathbf{u})[\mathbf{u}\cdot\mathbf{n}] &= 0, \end{aligned} \tag{2.2}$$

where $[\mathbf{u}\cdot\mathbf{n}]$ stands for the jump of the normal displacements $\mathbf{u}$ on the possible contact boundary Γ_C. It is defined as $[\mathbf{u}\cdot\mathbf{n}] := \mathbf{u}_m\cdot\mathbf{n}_m + \mathbf{u}_s\cdot\mathbf{n}_s$, and the normal part of the stresses on Γ_C is given by $\sigma_n(\mathbf{u}) := \sigma_{m,n}(\mathbf{u}_m) = \sigma_{s,n}(\mathbf{u}_s)$. The third condition is called the nonpenetration condition of the two bodies. Together with (2.1), the contact condition (2.2) formulate the problem of frictionless unilateral contact. The last condition in (2.2) implies that at the possible contact area Γ_C, we have either zero boundary stresses, i.e., $\sigma_n(\mathbf{u}) = 0$, or contact between the two bodies, i.e., $[\mathbf{u}\cdot\mathbf{n}] = 0$. We remark that we do not need any function modeling the distance between the two bodies, since we assume that they are in contact in the reference configuration. In the more general case, we have to replace the third condition in (2.2) by $[\mathbf{u}\cdot\mathbf{n}] \leq g$, where the gap function $g : \Gamma_{s,C} \to \mathbb{R}$ models the distance between the two bodies, and the jump has to be defined in terms of a suitable parametrization. To give the weak formulation of problem (2.1) with the contact conditions (2.2), we introduce the product space $\mathbf{V} := [H^1_*(\Omega_m)]^d \times [H^1_*(\Omega_s)]^d$, equipped with the broken H^1-norm $\|\mathbf{v}\|^2_{1,\Omega} := \sum_{k=m,s}\|\mathbf{v}\|^2_{1,\Omega_k}$, where the spaces $[H^1_*(\Omega_k)]^d$ are defined by

$$[H^1_*(\Omega_k)]^d := \left\{\mathbf{v}_k \in \left[H^1\left(\Omega_k\right)\right]^d \,:\, \mathbf{v}_{k|_{\Gamma_{k,D}}} = \mathbf{0}\right\}.$$

For $\mathbf{u} := (\mathbf{u}_m, \mathbf{u}_s) \in \mathbf{V}$ and $\mathbf{v} := (\mathbf{v}_m, \mathbf{v}_s) \in \mathbf{V}$, we define the bilinear form $a(\mathbf{u},\mathbf{v})$ and the linear form $f(\mathbf{v})$ by

$$a(\mathbf{u},\mathbf{v}) := \sum_{k=m,s}\int_{\Omega_k}\boldsymbol{\sigma}_k\left(\mathbf{u}_k\right) : \nabla\mathbf{v}_k\,dx,$$

$$f(\mathbf{v}) := \sum_{k=m,s}\left(\int_{\Omega_k}\mathbf{f}_k\cdot\mathbf{v}_k\,dx + \int_{\Gamma_{k,N}}\mathbf{p}_k\cdot\mathbf{v}_k\,ds\right).$$

Then the weak solution of our nonlinear contact problem (2.1) and (2.2) can be obtained as a solution of a minimization problem on the convex subset $\mathbf{K} := \{\mathbf{v} \in \mathbf{V} \,:\, [\mathbf{v}\cdot\mathbf{n}] \leq 0 \text{ on } \Gamma_C\}$

$$J(\mathbf{u}) = \min_{\mathbf{v}\in\mathbf{K}} J(\mathbf{v}) := \frac{1}{2}a(\mathbf{v},\mathbf{v}) - f(\mathbf{v}); \tag{2.3}$$

see, e.g., [Hc80, BGK87]. Furthermore, the minimization problem (2.3) is equivalent to a variational inequality on the convex subset $\mathbf{K} \subset \mathbf{V}$: find $\mathbf{u} \in \mathbf{K}$ such that

$$a(\mathbf{u}, \mathbf{v} - \mathbf{u}) \geq f(\mathbf{v} - \mathbf{u}), \qquad \mathbf{v} \in \mathbf{K}. \tag{2.4}$$

To give the saddle point formulation, we introduce a Lagrange multiplier space $\mathbf{M}$, being the dual space of the trace space $\mathbf{W}$ of $[H^1_*(\Omega_s)]^d$ restricted to Γ_C. By assumption $\overline{\Gamma_C}$ is a compact subset of $\partial\Omega_s \backslash \overline{\Gamma_{s,D}}$, and thus we have $\mathbf{W} = [H^{1/2}(\Gamma_C)]^d$. In the case $\overline{\Gamma_C} = \partial\Omega_s \backslash \Gamma_{s,D}$, we have to work with $[H^{1/2}_{00}(\Gamma_C)]^d$ instead of $[H^{1/2}(\Gamma_C)]^d$. We now define the following convex cone of Lagrange multipliers:

$$\mathbf{M}^+ := \left\{ \boldsymbol{\mu} \in \mathbf{M} \, : \, \langle \boldsymbol{\mu}, \mathbf{v} \rangle_{\Gamma_C} \geq 0, \quad \mathbf{v} \in \mathbf{W}^+ \right\}, \tag{2.5}$$

where $\langle \cdot, \cdot \rangle_{\Gamma_C}$ denotes the duality pairing between $\mathbf{M}$ and $\mathbf{W}$, and $\mathbf{W}^+ := \{\mathbf{v} \in \mathbf{W} : \mathbf{v} \cdot \mathbf{n} \geq 0\}$. Defining the bilinear form $b(\cdot, \cdot)$ on the product space $\mathbf{V} \times \mathbf{M}$ by

$$b(\mathbf{v}, \boldsymbol{\mu}) := \langle \boldsymbol{\mu} \cdot \mathbf{n}, [\mathbf{v} \cdot \mathbf{n}] \rangle_{\Gamma_C}, \qquad \boldsymbol{\mu} \in \mathbf{M}, \quad \mathbf{v} \in \mathbf{V},$$

we get the saddle point formulation of the unilateral contact problem without friction; see, e.g., [Hcc96, Chapter 1.3]: find $\mathbf{u} \in \mathbf{V}$ and $\boldsymbol{\lambda} \in \mathbf{M}^+$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, \boldsymbol{\lambda}) &= f(\mathbf{v}), & \mathbf{v} &\in \mathbf{V}, \\ b(\mathbf{u}, \boldsymbol{\mu} - \boldsymbol{\lambda}) &\leq 0, & \boldsymbol{\mu} &\in \mathbf{M}^+. \end{aligned} \tag{2.6}$$

The existence and uniqueness of $(\mathbf{u}, \boldsymbol{\lambda}) \in \mathbf{V} \times \mathbf{M}^+$ of (2.6) has been stated, e.g., in [Hcc96, Theorem 3.11 and Remark 3.10]. Moreover, $\mathbf{u}$ is also the unique solution of the minimization problem (2.3) and the variational inequality (2.4), and we find that $\boldsymbol{\lambda} = -\sigma_s(\mathbf{u}_s)\mathbf{n}$ and thus $\boldsymbol{\lambda} \cdot \mathbf{n} = -\sigma_n(\mathbf{u})$ and $\boldsymbol{\lambda} \times \mathbf{n} = 0$.

In the following, we denote by $\gamma_a \subset \Gamma_C$ the actual contact set, i.e., $[\mathbf{u} \cdot \mathbf{n}] = 0$ on γ_a and $[\mathbf{u} \cdot \mathbf{n}] < 0$ on $\gamma_c := \Gamma_C \backslash \gamma_a$. If the displacement $\mathbf{u}$ is a continuous function, then the actual contact zone γ_a is a well-defined and closed subset of Γ_C.

3. Discretization techniques and a priori error estimates. In this section, we consider a discrete formulation of the saddle point formulation (2.6). Mortar techniques with standard Lagrange multiplier spaces for contact problems have been considered and analyzed, e.g., in [BHL99, Hil00, CHLS01, HL02, BR03]. Here, we apply these techniques to dual Lagrange multiplier spaces; see, e.g., [Woh00]. New optimal a priori error estimates for the discretization error in the H^1-norm and for the Lagrange multiplier in the $H^{-1/2}$-norm will be given. To approximate $\mathbf{V}$, we use standard conforming finite elements of lowest order on quasi-uniform simplicial, quadrilateral, or hexahedral triangulations. The finite element space associated with the shape regular triangulation $\mathcal{T}_{h,\Omega_k}$ is denoted by $S_1(\Omega_k, \mathcal{T}_{h,\Omega_k})$. The meshsize h is defined by the maximal diameter of the elements in $\mathcal{T}_{h,\Omega_m}$ and $\mathcal{T}_{h,\Omega_s}$. We define the discrete product space $\mathbf{V}_h$ by

$$\mathbf{V}_h := \left\{ \mathbf{v}_h \in \prod_{k \in \{s,m\}} \left[S_1\left(\Omega_k, \, \mathcal{T}_{h,\Omega_k}\right)\right]^d \, : \, \mathbf{v}_{h|_{\Gamma_D}} = \mathbf{0} \right\} \subset \mathbf{V}.$$

As it is standard in the mortar context, the Lagrange multiplier space inherits its $(d-1)$-dimensional mesh from the d-dimensional triangulation on the slave side. We assume that Γ_C can be written as union of edges in 2D and faces in 3D on the slave side. Here, we use discontinuous piecewise linear or bilinear nodal basis functions for the dual Lagrange multiplier. The discrete Lagrange multiplier space is denoted by $\mathbf{M}_h$ and can be spanned by $\{\psi_i \mathbf{e}_l,\ i = 1, \ldots, N_{\mathbf{M}_h},\ l = 1, \ldots, d\}$, where $\mathbf{e}_l$ denotes the lth unit vector, ψ_i the ith scalar dual basis function, and $N_{\mathbf{M}_h}$ is the number of vertices on the slave side of $\overline{\Gamma_C}$. In contrast to the general mortar setting, we do not have to remove the degrees of freedom of $\mathbf{M}_h$ on $\partial\Gamma_C$. We note that this has to be done if $\overline{\Gamma_{s,D}} \cap \overline{\Gamma_{s,C}} \neq \emptyset$ but not in our case. We now define the discrete space of $\mathbf{M}^+$ (see (2.5)) by

$$\mathbf{M}_h^+ := \Big\{\boldsymbol{\mu}_h \in \mathbf{M}_h \,:\, \langle \boldsymbol{\mu}_h, \mathbf{v}_h \rangle_{\Gamma_C} \geq 0,\, \mathbf{v}_h \in \mathbf{W}_h^+ \Big\}, \tag{3.1}$$

where $\mathbf{W}_h^+ := \{\mathbf{v}_h \in \mathbf{W}_h \,:\, \mathbf{v}_h \cdot \mathbf{n} \geq 0\}$ and $\mathbf{W}_h$ is the vector-valued trace space of $[S_1(\Omega_s, \mathcal{T}_{h,\Omega_s})]^d$ restricted to Γ_C. The following lemma shows that the space $\mathbf{M}_h^+$ can be equivalently defined by

$$\widehat{\mathbf{M}}_h^+ := \left\{ \boldsymbol{\mu}_h = \sum_{i=1}^{N_{\mathbf{M}_h}} \boldsymbol{\alpha}_i \psi_i \,:\, \boldsymbol{\alpha}_i \in \mathbb{R}^d,\, \boldsymbol{\alpha}_i = \alpha_i^n \mathbf{n},\, \alpha_i^n \in \mathbb{R},\, \alpha_i^n \geq 0,\, i = 1, \ldots, N_{\mathbf{M}_h} \right\}.$$

LEMMA 3.1. $\mathbf{M}_h^+ = \widehat{\mathbf{M}}_h^+$.

Proof. To show the equivalence, we write $\mathbf{v}_h \in \mathbf{W}_h^+$ as $\mathbf{v}_h = \sum_{i=1}^{N_{\mathbf{M}_h}} (\beta_i^n \mathbf{n} + \boldsymbol{\beta}_i^t)\varphi_i$ with $\beta_i^n \in \mathbb{R}$, $\beta_i^n \geq 0$, and $\boldsymbol{\beta}_i^t \in \mathbb{R}^d$, $\boldsymbol{\beta}_i^t \cdot \mathbf{n} = 0$, where φ_j are the scalar nodal basis function of $S_1(\Omega_s, \mathcal{T}_{h,\Omega_s})$ restricted to Γ_C. Each $\boldsymbol{\mu}_h \in \mathbf{M}_h$ can be written in terms of the basis functions ψ_i as $\boldsymbol{\mu}_h = \sum_{i=1}^{N_{\mathbf{M}_h}} (\alpha_i^n \mathbf{n} + \boldsymbol{\alpha}_i^t)\psi_i$ with $\alpha_i^n \in \mathbb{R}$ and $\boldsymbol{\alpha}_i^t \in \mathbb{R}^d$, $\boldsymbol{\alpha}_i^t \cdot \mathbf{n} = 0$. Due to the biorthogonality of the basis functions

$$\int_{\Gamma_C} \psi_i \varphi_j \, ds = \delta_{ij} c_j, \qquad c_j := \int_{\Gamma_C} \varphi_j \, ds > 0,$$

we have

$$\langle \boldsymbol{\mu}_h, \mathbf{v}_h \rangle_{\Gamma_C} = \sum_{i=1}^{N_{\mathbf{M}_h}} \left(\alpha_i^n \beta_i^n + \boldsymbol{\alpha}_i^t \cdot \boldsymbol{\beta}_i^t \right) c_i. \tag{3.2}$$

Letting $\boldsymbol{\mu}_h \in \mathbf{M}_h^+$ and using the definition (3.1), the choice $\boldsymbol{\beta}_i^t = \mathbf{0}$ and $\beta_i^n = \delta_{ij}$ yields $\alpha_j^n \geq 0$ and the choice $\beta_i^n = 0$ and $\boldsymbol{\beta}_i^t = \pm\delta_{ij}\boldsymbol{\alpha}_j^t$ yields $\boldsymbol{\alpha}_j^t = \mathbf{0}$ and thus $\boldsymbol{\mu}_h \in \widehat{\mathbf{M}}_h^+$. Letting $\boldsymbol{\mu}_h \in \widehat{\mathbf{M}}_h^+$ and using (3.2) and $\beta_i^n \geq 0$, we find that $\boldsymbol{\mu}_h \in \mathbf{M}_h^+$. □

We remark that this is not a conforming approach for the Lagrange multiplier space, i.e., $\mathbf{M}_h^+ \not\subset \mathbf{M}^+$.

Remark 3.2. Using standard Lagrange multipliers, the spaces $\mathbf{M}_h^+$ and $\widehat{\mathbf{M}}_h^+$ are not the same. In that case, the convex cone $\widehat{\mathbf{M}}_h^+$ is a subset of the convex cone $\mathbf{M}_h^+$.

The discrete mortar formulation of the saddle point problem (2.6) is now given by the following: find $\mathbf{u}_h \in \mathbf{V}_h$ and $\boldsymbol{\lambda}_h \in \mathbf{M}_h^+$ such that

$$\begin{aligned} a(\mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, \boldsymbol{\lambda}_h) &= f(\mathbf{v}_h), & \mathbf{v}_h &\in \mathbf{V}_h, \\ b(\mathbf{u}_h, \boldsymbol{\mu}_h - \boldsymbol{\lambda}_h) &\leq 0, & \boldsymbol{\mu}_h &\in \mathbf{M}_h^+. \end{aligned} \tag{3.3}$$

Existence and uniqueness of a solution follows from a discrete inf-sup condition (see, e.g., [Woh01, Chapter 1.2.3]) for the spaces $\mathbf{M}_h^n$ and $\mathbf{V}_h$ with $\mathbf{M}_h^+ \subset \mathbf{M}_h^n := \{\boldsymbol{\mu}_h \in \mathbf{M}_h : \boldsymbol{\mu}_h \times \mathbf{n} = 0\}$ (see, e.g., [Hcc96, Chapter 2.4.2] or [HL02]).

In the rest of this section, we consider optimal a priori estimates for the discretization errors in the primal and dual variables. The starting point is the following abstract lemma; see, e.g., [HL02].

LEMMA 3.3. *Let* $(\mathbf{u}, \boldsymbol{\lambda}) \in \mathbf{V} \times \mathbf{M}^+$ *be the solution of* (2.6) *and let* $(\mathbf{u}_h, \boldsymbol{\lambda}_h) \in \mathbf{V}_h \times \mathbf{M}_h^+$ *be the solution of the discrete formulation* (3.3). *Then there exists a constant* $0 < C < \infty$ *independent of the meshsize* h, *such that*

$$\|\mathbf{u}-\mathbf{u}_h\|_{1,\Omega} + \|\boldsymbol{\lambda}-\boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C} \le C\Big\{ \inf_{\mathbf{v}_h\in\mathbf{V}_h} \|\mathbf{u}-\mathbf{v}_h\|_{1,\Omega} + \inf_{\boldsymbol{\mu}_h\in\mathbf{M}_h^n} \|\boldsymbol{\lambda}-\boldsymbol{\mu}_h\|_{-\frac{1}{2},\Gamma_C} \\ + \max\big(b(\mathbf{u},\boldsymbol{\lambda}_h),\, 0\big)^{\frac{1}{2}} + \max(b(\mathbf{u}_h,\boldsymbol{\lambda}),\, 0)^{\frac{1}{2}} \Big\}.$$

Proof. The proof can be found in [HL02] for scalar-valued standard Lagrange multipliers and quadratic finite elements. It applies also to our situation. For convenience of the reader, we briefly recall the basic steps. Starting with $a(\mathbf{u}-\mathbf{u}_h, \mathbf{u}-\mathbf{u}_h)$, we find for $\mathbf{v}_h \in \mathbf{V}_h$, using (2.6) and (3.3),

$$\begin{aligned} a(\mathbf{u}-\mathbf{u}_h, \mathbf{u}-\mathbf{u}_h) &= a(\mathbf{u}-\mathbf{u}_h, \mathbf{u}-\mathbf{v}_h) + a(\mathbf{u}-\mathbf{u}_h, \mathbf{v}_h-\mathbf{u}_h) \\ &= a(\mathbf{u}-\mathbf{u}_h, \mathbf{u}-\mathbf{v}_h) - b(\mathbf{v}_h-\mathbf{u}_h, \boldsymbol{\lambda}) + b(\mathbf{v}_h-\mathbf{u}_h, \boldsymbol{\lambda}_h) \\ &= a(\mathbf{u}-\mathbf{u}_h, \mathbf{u}-\mathbf{v}_h) - b(\mathbf{v}_h-\mathbf{u}, \boldsymbol{\lambda}-\boldsymbol{\lambda}_h) - b(\mathbf{u}-\mathbf{u}_h, \boldsymbol{\lambda}-\boldsymbol{\lambda}_h). \end{aligned}$$

Using the continuity of the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, the trace theorem, and Korn's inequality, we obtain the upper bound

$$\|\mathbf{u}-\mathbf{u}_h\|^2_{1,\Omega} \le C(\|\mathbf{u}-\mathbf{u}_h\|_{1,\Omega} + \|\boldsymbol{\lambda}-\boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C})\|\mathbf{u}-\mathbf{v}_h\|_{1,\Omega} - b(\mathbf{u}-\mathbf{u}_h, \boldsymbol{\lambda}-\boldsymbol{\lambda}_h). \tag{3.4}$$

In terms of the discrete inf-sup condition (see, e.g., [Woh01, Chapter 1.2.3]) and observing that $\|\boldsymbol{\mu}_h\|_{-\frac{1}{2},\Gamma_C} = \|\boldsymbol{\mu}_h \cdot \mathbf{n}\|_{-\frac{1}{2},\Gamma_C}$, $\boldsymbol{\mu}_h \in \mathbf{M}_h^n$, we find the upper bound

$$\begin{aligned} \|\boldsymbol{\mu}_h - \boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C} &\le C \sup_{\mathbf{w}_h\in\mathbf{v}_h} \frac{b(\mathbf{w}_h, \boldsymbol{\mu}_h-\boldsymbol{\lambda}_h)}{\|\mathbf{w}_h\|_{1,\Omega}} \\ &= C \sup_{\mathbf{w}_h\in\mathbf{v}_h} \frac{b(\mathbf{w}_h, \boldsymbol{\mu}_h-\boldsymbol{\lambda}) + a(\mathbf{u}_h-\mathbf{u}, \mathbf{w}_h)}{\|\mathbf{w}_h\|_{1,\Omega}} \\ &\le C(\|\boldsymbol{\mu}_h-\boldsymbol{\lambda}\|_{-\frac{1}{2},\Gamma_C} + \|\mathbf{u}_h-\mathbf{u}\|_{1,\Omega}). \end{aligned}$$

Setting $\boldsymbol{\mu} = \mathbf{0}$ in (2.6) and $\boldsymbol{\mu}_h = \mathbf{0}$ in (3.3), we find that $b(\mathbf{u}, \boldsymbol{\lambda}) \ge 0$ and $b(\mathbf{u}_h, \boldsymbol{\lambda}_h) \ge 0$. We note that $\mathbf{M}^+$ and $\mathbf{M}_h^+$ are convex cones. Using $\boldsymbol{\mu} = 2\boldsymbol{\lambda}$ and $\boldsymbol{\mu}_h = 2\boldsymbol{\lambda}_h$, we get $b(\mathbf{u}, \boldsymbol{\lambda}) \le 0$ and $b(\mathbf{u}_h, \boldsymbol{\lambda}_h) \le 0$, respectively. In terms of $b(\mathbf{u}, \boldsymbol{\lambda}) = b(\mathbf{u}_h, \boldsymbol{\lambda}_h) = 0$, the third term on the right side of (3.4) can be written as

$$-b(\mathbf{u}-\mathbf{u}_h, \boldsymbol{\lambda}-\boldsymbol{\lambda}_h) = b(\mathbf{u}, \boldsymbol{\lambda}_h) + b(\mathbf{u}_h, \boldsymbol{\lambda}) \le \max(b(\mathbf{u}, \boldsymbol{\lambda}_h),\, 0) + \max(b(\mathbf{u}_h, \boldsymbol{\lambda}),\, 0). \qquad \square$$

We remark that the term $\max(b(\mathbf{u}_h, \boldsymbol{\lambda}), 0)$ takes into account the discrete penetration of the two bodies on the actual contact set γ_a. The term $\max\big(b(\mathbf{u}, \boldsymbol{\lambda}_h), 0\big)$ can be greater than zero if the discrete Lagrange multiplier $\boldsymbol{\lambda}_h \cdot \mathbf{n}$ is negative on a part of γ_c. We recall that $\mathbf{M}_h^+$ is not a subspace of $\mathbf{M}^+$, and thus $\boldsymbol{\lambda}_h \cdot \mathbf{n}$, $\boldsymbol{\lambda}_h \in \mathbf{M}_h^+$, can be smaller than zero. The first two terms in the upper bound of Lemma 3.3 are the best approximation errors. They reflect the quality of the approximation of the spaces $\mathbf{v}_h$ and $\mathbf{M}_h^n$. The third and the fourth terms measure the nonconformity of the approach.

To prove optimal a priori error estimates under the H^s-regularity assumption for the displacements u with $\frac{3}{2} < s \leq 2$, we have to consider the two last terms for the bilinear form $b(\cdot,\cdot)$ in the upper bound of Lemma 3.3. In a first step, we briefly recall the definition of the mortar projection $\mathbf{\Pi}_h : [L^2(\Gamma_C)]^d \to \mathbf{w}_h$ and its dual operator $\mathbf{\Pi}_h^* : [L^2(\Gamma_C)]^d \to \mathbf{M}_h$:

$$\begin{aligned}\int_{\Gamma_C} (\mathbf{\Pi}_h \mathbf{w}) \cdot \boldsymbol{\mu}_h \, ds &:= \int_{\Gamma_C} \mathbf{w} \cdot \boldsymbol{\mu}_h \, ds, \qquad \boldsymbol{\mu}_h \in \mathbf{M}_h,\\ \int_{\Gamma_C} \mathbf{v}_h \cdot (\mathbf{\Pi}_h^* \boldsymbol{\mu}) \, ds &:= \int_{\Gamma_C} \mathbf{v}_h \cdot \boldsymbol{\mu} \, ds, \qquad \mathbf{v}_h \in \mathbf{w}_h.\end{aligned}$$

In terms of the stability of $\mathbf{\Pi}_h$ and $\mathbf{\Pi}_h^*$ (see the general mortar setting with dual Lagrange multipliers in [Woh01, Chapters 1.2.1 and 1.2.2]), both operators satisfy an approximation property

$$\|\mathbf{w} - \mathbf{\Pi}_h \mathbf{w}\|_{0,\Gamma_C} \leq C h^\tau |\mathbf{w}|_{\tau,\Gamma_C}, \qquad \|\boldsymbol{\mu} - \mathbf{\Pi}_h^* \boldsymbol{\mu}\|_{0,\Gamma_C} \leq C h^\nu |\boldsymbol{\mu}|_{\nu,\Gamma_C} \tag{3.5}$$

for $\mathbf{w} \in [H^\tau(\Gamma_C)]^d$, $0 \leq \tau \leq 2$, and $\boldsymbol{\mu} \in [H^\nu(\Gamma_C)]^d$, $0 \leq \nu \leq 1$.

The proof of the upper bound of $b(\mathbf{u}, \boldsymbol{\lambda}_h)$ is based on a regularity assumption for the actual contact zone.

Assumption 3.1. Regularity assumption on γ_a.

- In 2D, we assume that the number of points in $\overline{\mathring{\gamma}_a} \cap \overline{\gamma_c}$ is finite.
- In 3D, the regularity assumption will be specified in the proof of Lemma 3.5.

We now give two lemmas providing upper bounds for the consistency errors.

LEMMA 3.4. *Let* $(\mathbf{u}, \boldsymbol{\lambda}) \in \mathbf{v} \times \mathbf{M}^+$ *be the solution of* (2.6) *and let* $(\mathbf{u}_h, \boldsymbol{\lambda}_h) \in \mathbf{v}_h \times \mathbf{M}_h^+$ *be the solution of the discrete formulation* (3.3). *Under the regularity assumption* $\mathbf{u} \in \left[H^{\frac{3}{2}+\nu}(\Omega)\right]^d$, $0 < \nu \leq \frac{1}{2}$, *we then have the a priori error estimate*

$$b(\mathbf{u}_h, \boldsymbol{\lambda}) \leq C \left(h^{1+2\nu} |\mathbf{u}|^2_{\frac{3}{2}+\nu,\Omega} + h^{\frac{1}{2}+\nu} |\mathbf{u}|_{\frac{3}{2}+\nu,\Omega} \|\mathbf{u} - \mathbf{u}_h\|_{1,\Omega} \right)$$

for a positive constant $C < \infty$ *independent of* h.

Proof. For standard Lagrange multipliers, we refer the reader to [HL02]. Although our dual basis functions of $\mathbf{M}_h$ are not positive, we can apply the same techniques. Using the discrete saddle point formulation (3.3) and the definition of the mortar projection, we find that

$$b(\mathbf{u}_h, \boldsymbol{\mu}_h - \boldsymbol{\lambda}_h) = \int_{\Gamma_C} ((\boldsymbol{\mu}_h - \boldsymbol{\lambda}_h) \cdot \mathbf{n}) \, (\mathbf{\Pi}_h [\mathbf{u}_h] \cdot \mathbf{n}) \, ds \leq 0, \qquad \boldsymbol{\mu}_h \in \mathbf{M}_h^+.$$

The normal component of the mortar projection of $[\mathbf{u}_h]$ can be written as $\mathbf{\Pi}_h[\mathbf{u}_h] \cdot \mathbf{n} = \sum_{i=1}^{N_{\mathbf{M}_h}} \alpha_i^n \varphi_i$. Writing the normal components of $\boldsymbol{\mu}_h, \boldsymbol{\lambda}_h \in \mathbf{M}_h^+$ as linear combination of the dual basis functions yields $\boldsymbol{\mu}_h \cdot \mathbf{n} = \sum_{i=1}^{N_{\mathbf{M}_h}} \beta_i^n \psi_i$ and $\boldsymbol{\lambda}_h \cdot \mathbf{n} = \sum_{i=1}^{N_{\mathbf{M}_h}} \gamma_i^n \psi_i$, $\beta_i^n, \gamma_i^n \geq 0$. Setting $\beta_i^n := \gamma_i^n + \delta_{ij}$, we obtain

$$0 \geq b(\mathbf{u}_h, \boldsymbol{\mu}_h - \boldsymbol{\lambda}_h) = \int_{\Gamma_C} \psi_j \, (\mathbf{\Pi}_h [\mathbf{u}_h] \cdot \mathbf{n}) \, ds = \alpha_j^n \int_{\Gamma_C} \varphi_j \, ds.$$

We recall that $\varphi_j \geq 0$ and thus $\alpha_j^n \leq 0$, $1 \leq j \leq N_{\mathbf{M}_h}$. This results, in combination with $\boldsymbol{\lambda} \in \mathbf{M}^+$, in $\int_{\Gamma_C} (\boldsymbol{\lambda} \cdot \mathbf{n}) \, (\mathbf{\Pi}_h [\mathbf{u}_h] \cdot \mathbf{n}) \, ds \leq 0$. In contrast to the general mortar setting with crosspoints, $\mathbf{\Pi}_h$ restricted to the finite element trace space on the slave side is the identity. The approximation properties (3.5) yield the upper bound

$$\begin{aligned}
&b(\mathbf{u}_h, \boldsymbol{\lambda})\\
&= \int_{\Gamma_C} (\boldsymbol{\lambda} \cdot \mathbf{n})\big([\mathbf{u}_h \cdot \mathbf{n}] - \mathbf{\Pi}_h[\mathbf{u}_h] \cdot \mathbf{n} + \mathbf{\Pi}_h[\mathbf{u}_h] \cdot \mathbf{n}\big)\, ds\\
&\le \int_{\Gamma_C} \left(\boldsymbol{\lambda} - \mathbf{\Pi}_h^* \boldsymbol{\lambda}\right) \cdot \mathbf{n} \left([\mathbf{u}_h] - \mathbf{\Pi}_h[\mathbf{u}_h]\right) \cdot \mathbf{n}\, ds\\
&= \int_{\Gamma_C} \left(\boldsymbol{\lambda} - \mathbf{\Pi}_h^* \boldsymbol{\lambda}\right) \cdot \mathbf{n} \left(\mathbf{\Pi}_h \mathbf{u}_h^m - \mathbf{u}_h^m\right) \cdot \mathbf{n}\, ds\\
&\le \|\boldsymbol{\lambda} - \mathbf{\Pi}_h^* \boldsymbol{\lambda}\|_{0,\Gamma_C} \big(\|(\mathbf{u}_m - \mathbf{u}_h^m) - \mathbf{\Pi}_h(\mathbf{u}_m - \mathbf{u}_h^m)\|_{0,\Gamma_C} + \|\mathbf{u}_m - \mathbf{\Pi}_h \mathbf{u}_m\|_{0,\Gamma_C}\big)\\
&\le h^\nu |\boldsymbol{\lambda}|_{\nu,\Gamma_C} \big(h^{\frac{1}{2}} |\mathbf{u}_m - \mathbf{u}_h^m|_{\frac{1}{2},\Gamma_C} + h^{1+\nu} |\mathbf{u}_m|_{1+\nu,\Gamma_C}\big)\\
&\le C\big(h^{\frac{1}{2}+\nu} |\mathbf{u}|_{\frac{3}{2}+\nu,\Omega} \|\mathbf{u} - \mathbf{u}_h\|_{1,\Omega} + h^{1+2\nu} |\mathbf{u}|^2_{\frac{3}{2}+\nu,\Omega}\big). \qquad \square
\end{aligned}$$

A similar bound for the term $b(\mathbf{u}, \boldsymbol{\lambda}_h)$ can be established.

LEMMA 3.5. *Let* $(\mathbf{u}, \boldsymbol{\lambda}) \in \mathbf{v} \times \mathbf{M}^+$ *be the solution of* (2.6) *and let* $(\mathbf{u}_h, \boldsymbol{\lambda}_h) \in \mathbf{v}_h \times \mathbf{M}_h^+$ *be the solution of the discrete formulation* (3.3)*. Under Assumption* 3.1 *and the regularity assumption* $\mathbf{u} \in \left[H^{\frac{3}{2}+\nu}(\Omega)\right]^d$, $0 < \nu \le \frac{1}{2}$, *we then have the a priori error estimate*

$$b(\mathbf{u}, \boldsymbol{\lambda}_h) \le C h^{\frac{1}{2}+\nu} |\mathbf{u}|_{\frac{3}{2}+\nu,\Omega} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C}$$

for a positive constant $C < \infty$ *independent of* $h < h_0$*.*

Proof. The regularity assumptions guarantee that $[\mathbf{u} \cdot \mathbf{n}]$ restricted to Γ_C is continuous. In a first step, we consider the two-dimensional case. We denote by I_h the Lagrange interpolation operator with respect to the mesh on the slave side. Based on I_h, we consider a modified interpolation operator $\tilde{I}_h$. Let $\mathcal{P}_C := \{p_i \,:\, 1 \le i \le N_{\mathbf{M}_h}\}$ be the set of vertices on the slave side of $\overline{\Gamma_C}$ and $\mathcal{W}_C := \{w_j \,:\, 1 \le j \le N_w\}$ be the set of points in $\overline{\mathring{\gamma}_a} \cap \overline{\gamma_c}$. The minimum distance between the elements in $\mathcal{W}_C$ is denoted by a, i.e., $a = \inf\{|w_j - w_k| \,:\, 1 \le j \ne k \le N_w\}$, where $|\cdot|$ denotes the Euclidean norm. By Assumption 3.1, $N_w < \infty$ and thus $a > 0$. For $h < \frac{a}{2} =: h_0$, we find between two neighbor points in $\mathcal{W}_C$ at least two vertices in $\mathcal{P}_C$.

We now define the modified Lagrange interpolation operator $\tilde{I}_h$ as

$$(\tilde{I}_h\, [\mathbf{u} \cdot \mathbf{n}])(p_i) := \begin{cases} [\mathbf{u} \cdot \mathbf{n}](p_i) & \text{if supp } \varphi_i \subset \overline{\gamma_c}, \\ \mathbf{0} & \text{else}, \end{cases} \tag{3.6}$$

where φ_i is the standard nodal basis function associated with the vertex $p_i \in \mathcal{P}_C$; see Figure 3.1. Using $b(\mathbf{u}, \boldsymbol{\lambda}) = 0$, we find

$$b\,(\mathbf{u}, \boldsymbol{\lambda}_h) = \int_{\Gamma_C} ([\mathbf{u} \cdot \mathbf{n}] - \tilde{I}_h[\mathbf{u} \cdot \mathbf{n}])\big((\boldsymbol{\lambda}_h - \boldsymbol{\lambda}) \cdot \mathbf{n}\big)\, ds + \int_{\Gamma_C} \tilde{I}_h[\mathbf{u} \cdot \mathbf{n}]\big((\boldsymbol{\lambda}_h - \boldsymbol{\lambda}) \cdot \mathbf{n}\big)\, ds. \tag{3.7}$$

By construction $\tilde{I}_h[\mathbf{u} \cdot \mathbf{n}] \le 0$ and supp $\tilde{I}_h[\mathbf{u} \cdot \mathbf{n}] \subset$ supp $[\mathbf{u} \cdot \mathbf{n}] = \overline{\gamma_c}$, we then find in terms of (2.2) and (3.1) that

$$\int_{\Gamma_C} \tilde{I}_h[\mathbf{u} \cdot \mathbf{n}]\, (\boldsymbol{\lambda}_h \cdot \mathbf{n})\, ds \le 0, \qquad \int_{\Gamma_C} \tilde{I}_h[\mathbf{u} \cdot \mathbf{n}]\, (\boldsymbol{\lambda} \cdot \mathbf{n})\, ds = 0. \tag{3.8}$$

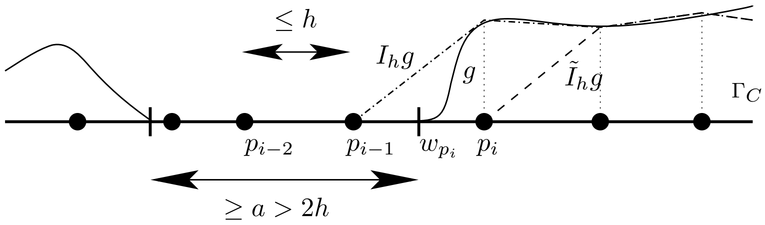


FIG. 3.1. *Functions g, $I_h g$, and $\tilde{I}_h g$.*

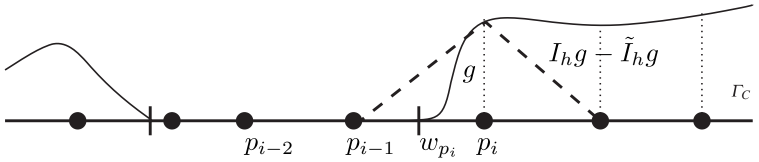


FIG. 3.2. *Difference of the interpolated functions $I_h g - \tilde{I}_h g$.*

Now we get from (3.7) the estimate

$$b\left(\mathbf{u}, \boldsymbol{\lambda}_h\right) \leq \|[\mathbf{u}\cdot\mathbf{n}] - \tilde{I}_h[\mathbf{u}\cdot\mathbf{n}]\|_{\frac{1}{2},\Gamma_C} \|\boldsymbol{\lambda}_h - \boldsymbol{\lambda}\|_{-\frac{1}{2},\Gamma_C}. \tag{3.9}$$

To estimate the term $\|[\mathbf{u}\cdot\mathbf{n}] - \tilde{I}_h[\mathbf{u}\cdot\mathbf{n}]\|_{\frac{1}{2},\Gamma_C}$, we define $g := -[\mathbf{u}\cdot\mathbf{n}]$ and consider the difference $I_h g - \tilde{I}_h g$ of the interpolated functions; see Figures 3.1 and 3.2.

In terms of the inverse inequality (see, e.g., [Bra97, Chapter 6]), we get

$$\|I_h g - \tilde{I}_h g\|^2_{\frac{1}{2},\Gamma_C} \leq \frac{C}{h}\|I_h g - \tilde{I}_h g\|^2_{0,\Gamma_C} \leq C \sum_{p_i \in \mathcal{M}_C} \left(g(p_i)\right)^2, \tag{3.10}$$

where the set of points $\mathcal{M}_C$ on the contact boundary Γ_C is defined by $\mathcal{M}_C := \{p_i \in \mathcal{P}_C \,:\, \tilde{I}_h g(p_i) \neq I_h g(p_i)\}$. Let $w_{p_i} \in \mathcal{W}_C$ be the unique closest point to $p_i \in \mathcal{M}_C$. Without loss of generality, we consider a lexicographically ordering of the indices and the case that $g = 0$ in the left neighborhood of w_{p_i} and $g > 0$ in the right neighborhood of w_{p_i}.

For $h < \frac{a}{2}$, we have $I_h g - \tilde{I}_h g = 0$ on $[p_{i-2}, p_{i-1}]$; see Figure 3.2. The regularity assumption on $\mathbf{u}$ yields $g \in H^{1+\nu}(\Gamma_C)$ and moreover $g = 0$ on $[p_{i-2}, w_{p_i}]$. Now the Cauchy–Schwarz inequality gives for each point $p_i \in \mathcal{M}_C$ the estimate

$$\begin{aligned}
\left(g(p_i)\right)^2 &= \left(\int_{w_{p_i}}^{p_i} g'(s)\,ds\right)^2 \\
&= \frac{1}{|w_{p_i} - p_{i-2}|^2}\left(\int_{w_{p_i}}^{p_i}\int_{p_{i-2}}^{w_{p_i}} \frac{g'(s) - g'(t)}{|s-t|^{\frac{1+2\nu}{2}}} |s-t|^{\frac{1+2\nu}{2}}\,dt\,ds\right)^2 \\
&\leq \frac{1}{|w_{p_i} - p_{i-2}|^2} \int_{p_{i-2}}^{p_i}\int_{p_{i-2}}^{p_i} \frac{\left(g'(s) - g'(t)\right)^2}{|s-t|^{1+2\nu}}\,dt\,ds \int_{w_{p_i}}^{p_i}\int_{p_{i-2}}^{w_{p_i}} |s-t|^{1+2\nu}\,dt\,ds \\
&\leq C\frac{1}{|w_{p_i} - p_{i-2}|^2} |g'|^2_{\nu,[p_{i-2},\,p_i]}\, h^{1+2\nu}\, |p_i - w_{p_i}|\, |w_{p_i} - p_{i-2}| \\
&= C|g'|^2_{\nu,[p_{i-2},\,p_i]}\, h^{1+2\nu}\, \frac{|p_i - w_{p_i}|}{|w_{p_i} - p_{i-2}|} \leq C|g'|^2_{\nu,[p_{i-2},\,p_i]}\, h^{1+2\nu},
\end{aligned}$$

where we used the shape regularity of the triangulation. To obtain an upper bound for $\|I_h g - \tilde{I}_h g\|_{\frac{1}{2},\Gamma_C}$, we have to sum over all $p_i \in \mathcal{M}_C$. We observe that by Assumption 3.1, the number of elements in $\mathcal{M}_C$ is finite. In terms of (3.9) and (3.10), we find now the bound

$$\begin{aligned} b(\mathbf{u}, \boldsymbol{\lambda}_h) &\leq \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C} (\|g - I_h g\|_{\frac{1}{2},\Gamma_C} + \|I_h g - \tilde{I}_h g\|_{\frac{1}{2},\Gamma_C}) \\ &\leq C h^{\frac{1}{2}+\nu} |g|_{1+\nu,\Gamma_C} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C} \leq C h^{\frac{1}{2}+\nu} |\mathbf{u}|_{\frac{3}{2}+\nu,\Omega} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C}. \end{aligned}$$

If not stated otherwise, we use the same definitions and symbols in the three-dimensional setting as for the two-dimensional situation. The regularity assumption on $\mathbf{u}$ guarantees that g is continuous. In the three-dimensional case, we replace the Lagrange interpolation operator by a quasi-interpolation operator which is defined locally by its values at the nodes $p \in \mathcal{P}_C$

$$I_h g(p) := \frac{1}{2c_o h} \int_{S_p} g(s)\ ds,$$

where S_p is a one-dimensional interval of length $2c_0 h$, direction z, and p is the midpoint of S_p. The direction z is arbitrary but has to be fixed for each node $p \in \mathcal{P}_C$. The constant $c_0 > 0$ depends on the shape regularity of the mesh, but not on h. It is taken such that $B_p(c_0 h) \subset \operatorname{supp}(\varphi_p)$, where $B_p(c_0 h)$ is the circle with radius $c_0 h$ and center p. We note that I_h restricted to $\mathbf{n} \cdot \mathbf{w}_h$ is not the identity, but it reproduces a linear function. This operator is similar to the ones of Clément and Scott–Zhang, and thus it is easy to verify the estimate

$$\|g - I_h g\|_{\frac{1}{2},\Gamma_C} \leq C h^{\frac{1}{2}+\nu} |g|_{1+\nu,\Gamma_C}.$$

We also define analogously to (3.6) the modified operator $\tilde{I}_h$

$$(\tilde{I}_h g)(p_i) := \begin{cases} I_h g(p_i) & \text{if } \operatorname{supp} \varphi_i \subset \overline{\gamma_c}, \\ 0 & \text{else.} \end{cases}$$

We note that we cannot use the standard Scott–Zhang operator because it does not necessarily preserves the sign. The same reasoning as in the two-dimensional case yields (3.8) and thus (3.9). Furthermore in the three-dimensional case, the estimate (3.10) has the form

$$\|I_h g - \tilde{I}_h g\|^2_{\frac{1}{2},\Gamma_C} \leq \frac{C}{h} \|I_h g - \tilde{I}_h g\|^2_{0,\Gamma_C} \leq C h \sum_{p_i \in \mathcal{M}_C} (I_h g(p_i))^2\,, \tag{3.11}$$

and we have to estimate the terms $I_h g(p_i)$ for all $p_i \in \mathcal{M}_C$. The vertices $p_i \in \mathcal{M}_C$ are marked with empty circles and γ_c is the shaded domain, in the left picture of Figure 3.3.

We assume that for h small enough, there exists for each $p \in \mathcal{M}_C$ a shape regular quadrilateral R_p, such that S_p is one edge, and $R_p \cap \gamma_a$ contains a shape regular quadrilateral including the opposite edge of S_p, where all regularity constants depend only on the regularity of the triangulation and of $\overline{\gamma_c}$ but neither on the node p nor on h. This quadrilateral R_p can now be mapped by F_p to the reference square $\hat{R} := [0,h]^2$ of length h such that $\hat{g}(\hat{s}) := g(F_p^{-1}(\hat{s}))$, restricted to the lower half $\hat{R}_2 := [0,h] \times [0, 1/2h]$ of the square, is zero, and such that the Jacobian of F_p and its

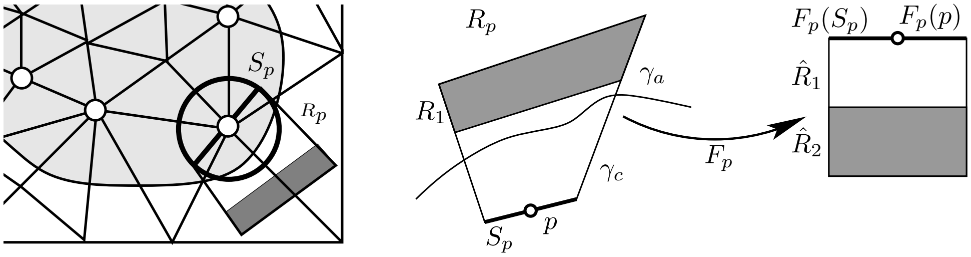


FIG. 3.3. *Vertices in $\mathcal{M}_C$ (left) and edge S_p; quadrilateral R_p and reference square $\hat{R}$ (right).*

inverse are bounded independently of h. We further assume that F_p restricted to S_p is affine and $F_p(S_p) = [0,h] \times \{1\}$. The upper half of the square is denoted by $\hat{R}_1$; see the right picture of Figure 3.3. Now we can proceed as in the two-dimensional case and find an upper bound for $I_h g(p)$

$$\begin{aligned}(I_h g(p))^2 &= \frac{1}{h^2}\left(\int_0^h \hat{g}(s_1,h)ds_1\right)^2 = \frac{1}{h^2}\left(\int_{\hat{R}_1} \hat{g}_{s_2}(s)ds\right)^2 \\ &= \frac{4}{h^6}\left(\int_{\hat{R}_2}\left(\int_{\hat{R}_1} \frac{\hat{g}_{s_2}(s)-\hat{g}_{s_2}(\tilde{s})}{|s-\tilde{s}|^{1+\nu}}|s-\tilde{s}|^{1+\nu}ds\right)d\tilde{s}\right)^2 \\ &\le \frac{4}{h^6}\int_{\hat{R}_2}\int_{\hat{R}_1}\frac{\left(\hat{g}_{s_2}(s)-\hat{g}_{s_2}(\tilde{s})\right)^2}{|s-\tilde{s}|^{2+2\nu}}dsd\tilde{s}\int_{\hat{R}_2}\int_{\hat{R}_1}|s-\tilde{s}|^{2+2\nu}ds\,d\tilde{s} \\ &\le \frac{C}{h^6}h^{2+2\nu}h^4\int_{\hat{R}}\int_{\hat{R}}\frac{\left(\hat{g}_{s_2}(s)-\hat{g}_{s_2}(\tilde{s})\right)^2}{|s-\tilde{s}|^{2+2\nu}}ds\,d\tilde{s} \\ &\le Ch^{2\nu}|\hat{g}_{s_2}|^2_{\nu,\hat{R}} \le Ch^{2\nu}|g|^2_{1+\nu,R_p}.\end{aligned}$$

In the last step, we have used the properties of the Jacobian of F_p. Summing over all $p \in \mathcal{M}_C$ and using a coloring argument, (3.11) yields finally as in the two-dimensional case the upper bound

$$\|I_h g - \tilde{I}_h g\|_{\frac{1}{2},\Gamma_C} \le Ch^{\frac{1}{2}+\nu}|g|_{1+\nu,\Gamma_C},$$

where the constant does not depend on the meshsize. □

Now we combine the previous results and formulate the optimal a priori error estimate for multibody contact problems.

THEOREM 3.6. *Let $(\mathbf{u},\boldsymbol{\lambda}) \in \mathbf{v} \times \mathbf{M}^+$ be the solution of* (2.6) *and let $(\mathbf{u}_h,\boldsymbol{\lambda}_h) \in \mathbf{v}_h \times \mathbf{M}_h^+$ be the solution of the discrete formulation* (3.3)*. Under Assumption* 3.1 *and the regularity assumption $\mathbf{u} \in \left[H^{\frac{3}{2}+\nu}(\Omega)\right]^d$, $0 < \nu \le \frac{1}{2}$, we then have the a priori error estimate*

$$\|\mathbf{u}-\mathbf{u}_h\|_{1,\Omega} + \|\boldsymbol{\lambda}-\boldsymbol{\lambda}_h\|_{-\frac{1}{2},\Gamma_C} \le Ch^{\frac{1}{2}+\nu}|\mathbf{u}|_{\frac{3}{2}+\nu,\Omega}$$

for a positive constant C.

Proof. Using the well-known approximation property for the spaces $\mathbf{v}_h$ and $\mathbf{M}_h$, the proof is a direct consequence of Lemmas 3.3, 3.4, and 3.5 and the application of Young's inequality. □

Remark 3.7. It is possible to establish an optimal a priori estimate for the Lagrange multiplier in a weighted L^2-norm, i.e., $\sqrt{h}\|\boldsymbol{\lambda}-\boldsymbol{\lambda}_h\|_{0,\Gamma_C}$. The proof follows the same lines as for the $H^{-1/2}$-norm and uses an inverse estimate.

Remark 3.8. Replacing $\mathbf{v}_h$ by quadratic finite elements, we obtain a higher-order a priori estimate. Quadratic finite elements and linear Lagrange multipliers yield an order $h^{\frac{1}{2}+\nu}$, $0 < \nu < 1$, upper bound for the discretization error if the solution is $H^{\frac{3}{2}+\nu}$-regular. Replacing the linear Lagrange multiplier space by quadratic Lagrange multipliers does not give a higher order; see, e.g., [HL02]. For a proof and numerical results comparing quadratic finite elements with linear Lagrange multipliers and with quadratic Lagrange multipliers, we refer the reader to [HMW04]; see also Table 4.4.

Remark 3.9. The same theoretical results can be obtained if a different dual Lagrange multiplier space is used, e.g., piecewise constant dual Lagrange multipliers or continuous piecewise cubic dual Lagrange multipliers as proposed in [Woh02]

4. Numerical results. In this section, we consider different examples for multibody contact problems. We focus on the discretization errors of the displacement in the L^2- and the H^1-norm and of the Lagrange multiplier in a weighted L^2-norm on the contact boundary. In general, multibody contact problems do not admit an analytical solution. To evaluate the discretization errors, we compute a reference solution denoted by $\mathbf{u}_{\text{ref}}$ for the displacements and a reference solution $\boldsymbol{\lambda}_{\text{ref}}$ for the Lagrange multiplier corresponding to a finer mesh. We note that in all our examples, the meshsize h_{ref} for the reference solution satisfies $h_{\text{ref}} \leq 1/4h$. A reference meshsize of $h_{\text{ref}} = 1/2h$ does not guarantee reliable numbers for the discretization errors. For the iterative solution of the nonlinear system, we use an inexact primal dual active set strategy; see [HW03]. In each outer iteration step, we apply one linear multigrid cycle. Alternatively, nonlinear Dirichlet–Neumann solvers [KW02, BSS02, EW03], monotone multigrid methods [KK01, Kra01, WK03], or finite element tearing and interconnecting (FETI) approaches [DFS98, DGS00, DH03] can be applied. The implementation is based on the finite element toolbox UG; see [BBJ$^+$97]. We start with a coarse triangulation and use uniform refinement techniques. Each element is decomposed into four subelements within the next refinement step.

Example 1. In our first example, we consider the problem depicted in Figure 4.1. The two bodies in their reference configuration are scaled squares, $\Omega_m := (0, 0.01) \times (0, 0.01)$ and $\Omega_s := (0.0, 0.01) \times (0.01, 0.02)$. On Ω_m, we set $E = 15 \times 10^9$ and $\nu = 0.2$ and on Ω_s, we use a different material with $E = 20 \times 10^9$ and $\nu = 0.4$. The lower boundary of Ω_m is clamped. The applied load $\mathbf{p}$ is given by $(10^5, -10^6)^\top$ on the left side and by $(-10^5, -10^6)^\top$ on the right side of Ω_m. Homogeneous Neumann boundary conditions are applied to both sides of Ω_s, and inhomogeneous Dirichlet data are given on the top of Ω_s. The displacements are set to be $(0.0, -5 \times 10^{-7})^\top$.

The initial triangulation on level 0 is shown in the left picture of Figure 4.1. We recall that the Lagrange multiplier $-\boldsymbol{\lambda}_h \cdot n$ of the mortar discretization approximates the contact pressure. In the right picture of Figure 4.1, we give the normal and tangential component of $\boldsymbol{\lambda}_h$ on level 6. The normal component is nonzero only in the part of Γ_C where the two bodies are actually in contact. As expected, the tangential component is equal to zero because no friction is taken into account. The distorted domains and stress components $\boldsymbol{\sigma}_{11}$ and $\boldsymbol{\sigma}_{22}$ are shown in the middle picture of Figure 4.1.

We use the finite element solution on level 8 as reference solution $(\mathbf{u}_{\text{ref}}, \boldsymbol{\lambda}_{\text{ref}})$. On level 8, we have 262,144 elements on Ω_s and 589,824 elements on Ω_m. In Table 4.1, we give the discretization errors of the displacements in the relative $L^2(\Omega)$-norm and in the relative $H^1(\Omega)$-norm. The error in the Lagrange multiplier is measured in a weighted $L^2(\Gamma_C)$-norm. The convergence rates in the L^2-norm for the displacements are approximately 1.8. Asymptotically, the convergence rates tend to 1 in the

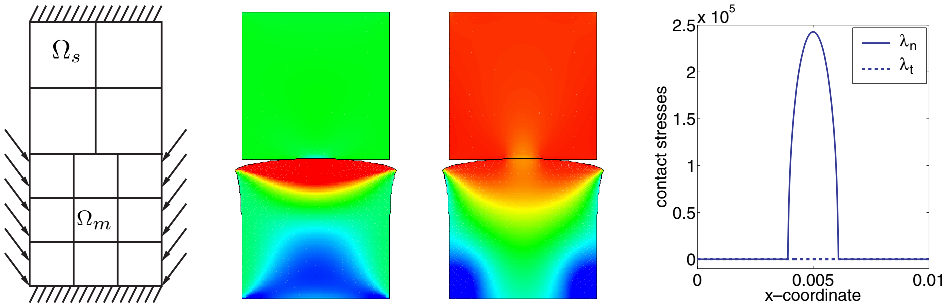


FIG. 4.1. *Example* 1: *Problem definition and initial triangulation (left), stress components $\boldsymbol{\sigma}_{11}$ and $\boldsymbol{\sigma}_{22}$ on distorted domains (distortion scaled by factor* 1000*) (middle), and contact stresses in normal and tangential direction (right).*

TABLE 4.1
Example 1: *Relative $L^2(\Omega)$-error and relative $H^1(\Omega)$-error of $\mathbf{u}_h$ with respect to $\mathbf{u}_{ref}$, weighted $L^2(\Gamma_C)$-error of $\boldsymbol{\lambda}_h$ with respect to $\boldsymbol{\lambda}_{ref}$, and the numerical convergence orders.*

Level	$\Vert\mathbf{u}_h - \mathbf{u}_{\mathrm{ref}}\Vert_{0,\Omega}/\Vert\mathbf{u}_{\mathrm{ref}}\Vert_{0,\Omega}$		$\vert\mathbf{u}_h - \mathbf{u}_{\mathrm{ref}}\vert_{1,\Omega}/\vert\mathbf{u}_{\mathrm{ref}}\vert_{1,\Omega}$		$\Vert\boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{\mathrm{ref}}\Vert_{-\frac{1}{2},h,\Gamma_C}$	
0	$6.447659e-02$	–	$4.447386e-01$	–	$9.507368e+02$	–
1	$1.947047e-02$	1.73	$2.422984e-01$	0.87	$3.709632e+02$	1.36
2	$5.719462e-03$	1.77	$1.323250e-01$	0.87	$1.937997e+02$	0.94
3	$1.798368e-03$	1.67	$7.137496e-02$	0.89	$8.890783e+01$	1.12
4	$5.181605e-04$	1.80	$3.815553e-02$	0.90	$3.743692e+01$	1.25
5	$1.551852e-04$	1.74	$2.012575e-02$	0.92	$1.503053e+01$	1.32
6	$4.544415e-05$	1.77	$1.027931e-02$	0.97	$5.607133e+00$	1.42

$H^1(\Omega)$-norm. As in the linear case, we observe better convergence rates for the Lagrange multiplier. The best approximation error of the Lagrange multiplier in the weighted $L^2(\Gamma_C)$-norm is of order $h^{1.5}$.

Example 2. In our second example, we use the same geometry and the same material parameters as in Example 1; see [HL02]. The upper part of Ω_s and the lower part of Ω_m are clamped. Inhomogeneous Neumann data are applied to the left part of Ω_s and to the right part of Ω_m. The scaled boundary forces are given by $\mathbf{p} = (10^5, -10^6)^\top$ and $\mathbf{p} = (-10^5, -10^6)^\top$; see Figure 4.2.

In the middle picture of Figure 4.2, the stress component $\boldsymbol{\sigma}_{11}$ is shown on the distorted domains. The normal component of the Lagrange multiplier is given in the right picture. In contrast to Example 1, we observe a singularity at the left endpoint of the contact zone. Here, we use bilinear finite elements and a conforming triangulation on the possible contact boundary Γ_C. The finite element solution on level 8 is taken as reference solution. Due to the lower regularity of the solution, we cannot expect a convergence rate of order h. The discretization errors for the relative $L^2(\Omega)$- and $H^1(\Omega)$-norm of the displacement are given in columns three and four of Table 4.2, respectively. The error of the Lagrange multiplier in the weighted $L^2(\Gamma_C)$-norm can be found in the last column. Asymptotically, the convergence rates tend to 0.9 in the $H^1(\Omega)$-norm. As in Example 1, we observe a higher convergence rate for the Lagrange multiplier. The convergence rate for the L^2-norm is approximately 1.5. The low regularity of the problem is reflected in smaller numerical convergence rates.

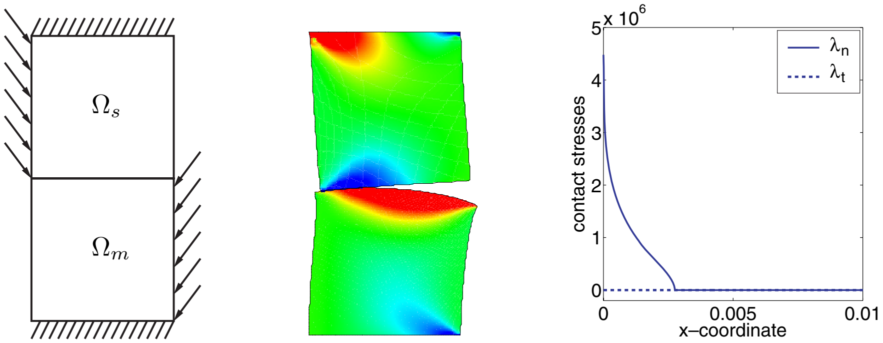


FIG. 4.2. *Example* 2: *Problem definition (left), stress component $\boldsymbol{\sigma}_{11}$ on distorted domains (distortion scaled by factor* 1000*) (middle), and contact stresses in normal and tangential direction (right).*

TABLE 4.2
Example 2: *Relative $L^2(\Omega)$-error and relative $H^1(\Omega)$-error of $\mathbf{u}_h$ with respect to $\mathbf{u}_{ref}$, weighted $L^2(\Gamma_C)$-error of $\boldsymbol{\lambda}_h$ with respect to $\boldsymbol{\lambda}_{ref}$, and the numerical convergence orders.*

Level	$\Vert\mathbf{u}_h - \mathbf{u}_{\text{ref}}\Vert_{0,\Omega}/\Vert\mathbf{u}_{\text{ref}}\Vert_{0,\Omega}$		$\vert\mathbf{u}_h - \mathbf{u}_{\text{ref}}\vert_{1,\Omega}/\vert\mathbf{u}_{\text{ref}}\vert_{1,\Omega}$		$\Vert\boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{\text{ref}}\Vert_{-\frac{1}{2},h,\Gamma_C}$	
0	$6.314604e-01$	–	$8.113448e-01$	–	$4.648877e+03$	–
1	$3.513681e-01$	0.85	$5.138377e-01$	0.66	$2.210844e+03$	1.07
2	$1.414198e-01$	1.31	$2.786256e-01$	0.88	$1.668986e+03$	0.41
3	$5.196291e-02$	1.44	$1.502248e-01$	0.89	$9.087467e+02$	0.88
4	$1.885312e-02$	1.46	$8.271511e-02$	0.86	$4.278806e+02$	1.09
5	$6.983378e-03$	1.43	$4.614857e-02$	0.84	$1.994283e+02$	1.10
6	$2.611373e-03$	1.42	$2.576322e-02$	0.84	$9.358784e+01$	1.09
7	$9.293584e-04$	1.49	$1.399176e-02$	0.88	$4.454642e+01$	1.07

Example 3. Our third example is the Hertzian contact problem of a linear elastic circle with a plane. In this example, the contact stresses can be computed analytically; see [Her82, KO88]. So we are able to compare the numerically computed boundary stresses to the analytical ones. If an elastic circle with radius r and material parameters E and ν is pressed by a single point load $f = (0, -f)^\top$ on the top to a rigid plane, the analytical contact pressure is given by (see, e.g., [KO88, Chapter 6.6])

$$p(x) = \frac{2f}{\pi b^2}\sqrt{\left(b^2 - x^2\right)}, \qquad x \le b, \qquad b := 2\sqrt{\frac{fr(1-\nu^2)}{E\pi}}, \tag{4.1}$$

where b is the half-width of the contact surface and x is the distance to the center of the contact surface. In our setting, we replace the rigid plane by a linear elastic rectangle. Young's modulus of the rectangle is set to be larger than the one of the circle. We apply homogeneous Dirichlet boundary conditions at the bottom and at the two sides of the rectangle. On the circle with radius $r = 1$, we set the material parameters to be $E = 7000$ and $\nu = 0.3$. For the rectangle with height $w = 1$, we use $E = 10^6$ and $\nu = 0.45$. The circle, assumed to be the slave side, is pressed by a point load $f = (0, -100)^\top$ at the top of the rectangle. The problem definition and the geometry are shown in Figure 4.3. As done in [CSW99, KW02], we replace the point load by a surface load to avoid a strong singularity on the upper part of the circle.

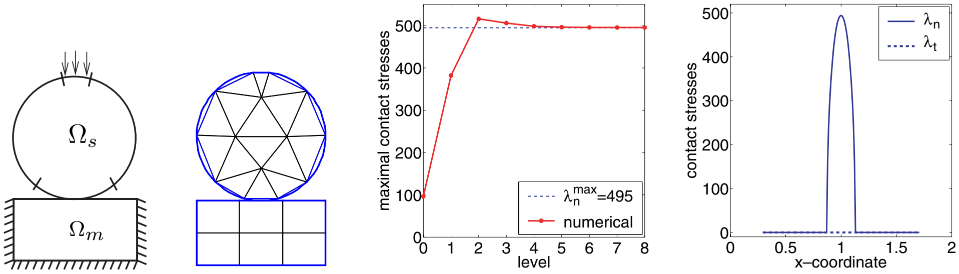


FIG. 4.3. *Example* 3: *Problem definition and initial triangulation (left), maximal contact stresses (middle), and contact stresses (right).*

TABLE 4.3
Example 3: *Relative* $L^2(\Omega)$*-error and relative* $H^1(\Omega)$*-error of* $\mathbf{u}_h$ *with respect to* $\mathbf{u}_{ref}$*, weighted* $L^2(\Gamma_C)$*-error of* $\boldsymbol{\lambda}_h$ *with respect to* $\boldsymbol{\lambda}_{ref}$*, and the numerical convergence orders.*

Level	$\Vert\mathbf{u}_h - \mathbf{u}_{\mathrm{ref}}\Vert_{0,\Omega}/\Vert\mathbf{u}_{\mathrm{ref}}\Vert_{0,\Omega}$		$\vert\mathbf{u}_h - \mathbf{u}_{\mathrm{ref}}\vert_{1,\Omega}/\vert\mathbf{u}_{\mathrm{ref}}\vert_{1,\Omega}$		$\Vert\boldsymbol{\lambda}_h - \boldsymbol{\lambda}_{\mathrm{ref}}\Vert_{-\frac{1}{2},h,\Gamma_C}$	
1	$5.227370e-02$	–	$4.663629e-01$	–	$5.845408+01$	–
2	$1.381026e-02$	1.92	$3.214708e-01$	0.54	$4.998951+01$	0.23
3	$4.844067e-03$	1.51	$1.807113e-01$	0.83	$2.120954+01$	1.24
4	$1.350194e-03$	1.84	$9.735769e-02$	0.89	$8.377665+00$	1.34
5	$4.175901e-04$	1.69	$5.111927e-02$	0.93	$3.269378+00$	1.36
6	$1.102450e-04$	1.92	$2.584385e-02$	0.98	$1.168388+00$	1.48

In order to obtain a unique discrete solution, we eliminate the horizontal degrees of freedom of the two inner nodes on the vertical symmetry axis; see Figure 4.3.

Using (4.1), the maximal normal contact stresses of the Hertzian contact problem is then given by $\boldsymbol{\lambda}_n^{\max} = 494.8$, and for the half-width of the contact zone we find $b = 0.129$. The pictures on the right of Figure 4.3 illustrate the numerical approximation of the contact stress. From level 4 on, the discrete maximal normal stress is a very good approximation of the analytical one. In the right picture, the normal and tangential contribution of the Lagrange multiplier are given.

To compute the discretization errors, we use the mortar finite element solution on level 8 as reference solution. The errors are given in Table 4.3. Asymptotically, we observe optimal convergence rates. In the L^2-norm, the convergence rate tends to 2 with increasing number of refinement steps, whereas the convergence rate in the $H^1(\Omega)$-norm tends to 1. We observe asymptotically a convergence rate of 1.5 for the Lagrange multiplier in the weighted L^2-norm on the contact zone. This results from the fact that the error in the energy norm restricted to a strip of width h can be bounded by $C\sqrt{h}\Vert\mathbf{u}\Vert_{1,\Omega}$.

In a last test, we consider the influence of the choice of the dual Lagrange multiplier on the displacement and the maximal contact stress. To do so, we compare three different dual Lagrange multipliers. In addition to the linear Lagrange multiplier, we use discontinuous piecewise constant and continuous piecewise cubic Lagrange multipliers; see [Woh02]. Table 4.4 shows in the first two columns the relative H^1-error for the dual constant and the dual cubic Lagrange multipliers. If we compare the results of Table 4.4 with those of Table 4.3, we find that all three of our dual Lagrange multipliers yield almost the same results. The differences in the H^1-norm of the error is negligible. This is also true for the maximal contact stress. The last three columns of

TABLE 4.4
Example 3: *Relative $H^1(\Omega)$-error of $\mathbf{u}_h$ with respect to $\mathbf{u}_{ref}$ and the maximal contact stress $(\boldsymbol{\lambda}_h \cdot n)_{max}$ for dual linear, dual constant, and dual cubic Lagrange multipliers.*

Level	$\lvert\mathbf{u}_h - \mathbf{u}_{\text{ref}}\rvert_{1,\Omega}/\lvert\mathbf{u}_{\text{ref}}\rvert_{1,\Omega}$		$(\boldsymbol{\lambda}_h \cdot n)_{\max}$		
	Constant	Cubic	Linear	Constant	Cubic
1	$4.663629e-01$	$4.663629e-01$	382.057	382.057	382.057
2	$3.214712e-01$	$3.214712e-01$	514.166	514.172	514.172
3	$1.807105e-01$	$1.807104e-01$	504.190	504.229	504.231
4	$9.735749e-02$	$9.735742e-02$	496.765	496.775	496.775
5	$5.111913e-02$	$5.111915e-02$	494.805	494.809	494.809
6	$2.584384e-02$	$2.584384e-02$	494.264	494.266	494.266
7			494.174	494.175	494.175
8			494.202	494.202	494.202

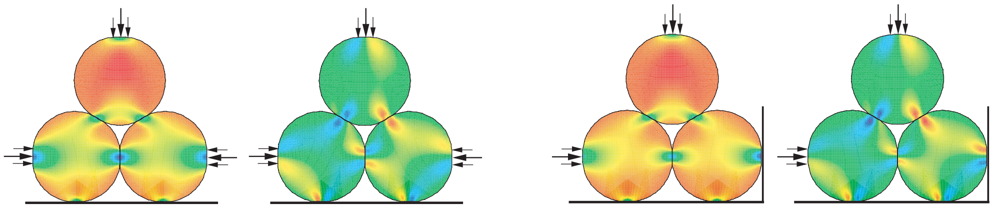


FIG. 4.4. *Example* 4: *Stress components $\boldsymbol{\sigma}_{11}$ and $\boldsymbol{\sigma}_{12}$ for a symmetric (left) and nonsymmetric (right) settings.*

Table 4.4 show $(\boldsymbol{\lambda}_h \cdot n)_{\max}$ for the three different discretizations.

Example 4. Finally, we consider an example with more than two subdomains. We use three circles and a rigid obstacle (see Figure 4.4) and two different sets of boundary data. The possible contact zone can be decomposed into two different types: the contact between two elastic circles and the contact between one elastic circle and the rigid obstacle. For this example, one circle has to have a master and a slave interface. The contact between the rigid obstacle and the elastic body is also discretized in terms of dual Lagrange multipliers defined on the elastic body which is the slave side. The two pictures in the left of Figure 4.4 show the stress components $\boldsymbol{\sigma}_{11}$ and $\boldsymbol{\sigma}_{12}$ of a symmetric boundary data. In the two pictures on the right, the rigid obstacle forms an L-shape, and the solution is nonsymmetric.

REFERENCES

[BBJ+97] P. BASTIAN, K. BIRKEN, K. JOHANNSEN, S. LANG, N. NEUSS, H. RENTZ-REICHERT, AND C. WIENERS, *UG—a flexible software toolbox for solving partial differential equations*, Comput. Vis. Sci., 1 (1997), pp. 27–40.

[Ben00] F. BEN BELGACEM, *Numerical simulation of some variational inequalities arisen from unilateral contact problems by the finite element methods*, SIAM J. Numer. Anal., 37 (2000), pp. 1198–1216.

[BGK87] P. BOIERI, F. GASTALDI, AND D. KINDERLEHRER, *Existence, uniqueness, and regularity results for the two-body contact problem*, Appl. Math. Optim., 15 (1987), pp. 251–277.

[BHL99] F. BEN BELGACEM, P. HILD, AND P. LABORDE, *Extension of the mortar finite element method to a variational inequality modeling unilateral contact*, Math. Models Methods Appl. Sci., 9 (1999), pp. 287–303.

[BHR77] F. BREZZI, W. HAGER, AND P. RAVIART, *Error estimates for the finite element solution of variational inequalities*, Numer. Math., 28 (1977), pp. 431–443.

[BR03] F. Ben Belgacem and Y. Renard, *Hybrid finite element methods for the Signorini problem*, Math. Comp., 72 (2003), pp. 1117–1145.

[Bra97] D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, 1997.

[BSS02] G. Bayada, J. Sabil, and T. Sassi, *Algorithme de Neumann-Dirichlet pour des problèmes de contact unilatéral: Résultat de convergence,* C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 381–386.

[CHLS01] P. Coorevits, P. Hild, K. Lhalouani, and T. Sassi, *Mixed finite element methods for unilateral problems: Convergence analysis and numerical studies*, Math. Comp., 71 (2001), pp. 1–25.

[CSW99] C. Carstensen, O. Scherf, and P. Wriggers, *Adaptive finite elements for elastic bodies in contact*, SIAM J. Sci. Comput., 20 (1999), pp. 1605–1626.

[DFS98] Z. Dostál, A. Friedlander, and S. Santos, *Solution of coercive and semicoercive contact problems by FETI domain decomposition*, in Domain Decomposition Methods, Boulder, CO, 1997, Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 82–93.

[DGS00] Z. Dostál, F. Gomes Neto, and S. Santos, *Solution of contact problems by FETI domain decomposition with natural coarse space projections*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 1611–1627.

[DH03] Z. Dostál and D. Horák, *Scalability and FETI based algorithm for large discretized variational inequalities*, Math. Comput. Simulation, 61 (2003), pp. 347–357.

[EW03] C. Eck and B. Wohlmuth, *Convergence of a contact-Neumann iteration for the solution of two-body contact problems*, Math. Models Methods Appl. Sci., 13 (2003), pp. 1103–1118.

[Fal74] R. Falk, *Error estimates for the approximation of a class of variational inequalities*, Math. Comp., 28 (1974), pp. 963–971.

[Glo84] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems*, Springer, New York, 1984.

[Hc80] J. Haslinger and I. Hlaváček, *Contact between two elastic bodies.* I. *Continuous problems*, Apl. Mat., 25 (1980), pp. 324–347.

[Hc81] J. Haslinger and I. Hlaváček, *Contact between two elastic bodies.* II. *Finite element analysis*, Apl. Mat., 26 (1981), pp. 263–290.

[Hcc96] J. Haslinger, I. Hlaváček, and J. Nečas, *Numerical methods for unilateral problems in solid mechanics*, in Handbook of Numerical Analysis, Vol. IV, North–Holland, Amsterdam, 1996, pp. 313–485.

[Her82] H. Hertz, *Über die Berührung fester elastischer Körper*, J. Reine Angew. Math., 92 (1882) pp. 156–171.

[Hil00] P. Hild, *Numerical implementation of two nonconforming finite element methods for unilateral contact*, Comput. Methods Appl. Mech. Engrg., 184 (2000), pp. 99–123.

[HL02] P. Hild and P. Laborde, *Quadratic finite element methods for unilateral contact problems*, Appl. Numer. Math., 41 (2002), pp. 410–421.

[HMW04] S. Hüeber, M. Mair, and B. Wohlmuth, *A priori error estimates and an inexact primal-dual active set strategy for linear and quadratic finite elements applied to multibody contact problems*, Appl. Numer. Math., to appear.

[HW03] S. Hüeber and B. Wohlmuth, *A Primal-Dual Active Strategy for Non-Linear Multibody Contact Problems*, Technical report 40, Universität Stuttgart, SFB 404, 2003. Comput. Methods Appl. Mech. Engrg., to appear.

[KK01] R. Kornhuber and R. Krause, *Adaptive multigrid methods for Signorini's problem in linear elasticity*, Comput. Vis. Sci., 4 (2001), pp. 9–20.

[KO88] N. Kikuchi and J. T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM Stud. Appl. Math. 8, SIAM, Philadelphia, 1988.

[Kra01] R. Krause, *Monotone Multigrid Methods for Signorini's Problem with Friction*, Ph.D. thesis, Freie Universität Berlin, Berlin, 2001.

[KW02] R. Krause and B. Wohlmuth, *A Dirichlet-Neumann type algorithm for contact problems with friction*, Comput. Vis. Sci., 5 (2002), pp. 139–148.

[Lau02] T. Laursen, *Computational Contact and Impact Mechanics*, Springer, Berlin, 2002.

[LS99] K. Lhalouani and T. Sassi, *Nonconfirming mixed variational inequalities and domain decomposition for unilateral problems*, East-West J. Numer. Math., 7 (1999), pp. 23–30.

[WK03] B. I. Wohlmuth and R. H. Krause, *Monotone multigrid methods on nonmatching grids for nonlinear multibody contact problems*, SIAM J. Sci. Comput., 25 (2003), pp. 324–347.

[Woh00] B. I. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

[Woh01] B. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Springer, Berlin, 2001.

[Woh02] B. Wohlmuth, *A comparison of dual Lagrange multiplier spaces for mortar finite element discretizations*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 995–1012.

[Wri01] P. Wriggers, *Nichtlineare Finite-Element-Methoden*, Springer, Berlin, 2001.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 174–194

PRESSURE CORRECTION ALGEBRAIC SPLITTING METHODS FOR THE INCOMPRESSIBLE NAVIER–STOKES EQUATIONS*

F. SALERI† AND A. VENEZIANI†

Abstract. In this paper we present a new family of methods for the effective numerical solution of the incompressible unsteady Navier–Stokes equations. These methods resort to an algebraic splitting of the discretized problem based on inexact LU block factorizations of the corresponding matrix (following [A. Quarteroni, F. Saleri, and A. Veneziani, *Comput. Methods Appl. Mech. Engrg.*, 188 (2000), pp. 505–526]. In particular, we will start from inexact algebraic factorizations of *algebraic Chorin–Temam* and *Yosida* type and introduce a *pressure correction step* aimed at improving the time accuracy. One of the schemes obtained in this way (the algebraic Chorin–Temam pressure correction method) resembles a method previously introduced in the framework of differential projection schemes (see [L. Timmermans, P. Minev, and F. V. de Vosse, *Internat. J. Numer. Methods Fluids*, 22 (1996), pp. 673–688], [A. Prohl, *Projection and Quasi-Compressibility Methods for Solving the Incompressible Navier–Stokes Equations*, Teubner, Stuttgart, 1997]. The stability and the dependence of splitting error on the time step of the new methods is investigated and tested on several numerical cases.

Key words. incompressible Navier–Stokes equations, fractional steps schemes

AMS subject classifications. 76M25, 76D10, 35Q30, 65M12

DOI. 10.1137/S0036142903435429

1. Introduction.

The numerical computation of the unsteady Navier–Stokes equations for incompressible flows in real applications requires the solution of linear systems of large dimensions. These systems are typically not definite nor well conditioned, and therefore the setup of efficient methods is mandatory. Perhaps one of the most successful approaches is provided by the class of the projection methods at the differential (see, e.g., [3] and, more recently, [14]) and the algebraic level (see [20] and [22], [23]). These methods typically compute the velocity and pressure fields separately, (i) by computing an *auxiliary* (or *intermediate*) velocity; (ii) by solving a suitable problem for the pressure; (iii) by correcting the velocity (*end-of-step velocity*), to enforce the incompressibility constraint. In [23] we have introduced a general class of algebraic splitting methods that can be regarded in this framework. We recall, in particular, the *algebraic Chorin–Temam* scheme (see [20]) and the *Yosida* scheme (see [22]). These methods are based on a splitting that reduces the computational effort, without affecting the time accuracy of the solution driven by the time discretization. This is true for first order time discretizations, while for higher order accuracy the setup of suitable splittings is still an open problem (see, e.g., in the framework of differential schemes, [10]).

In the present paper, we aim at investigating splitting methods that arise whenever, in addition to the velocity, the pressure is obtained after a suitable *correction*

*Received by the editors September 30, 2003; accepted for publication (in revised form) September 21, 2004; published electronically April 26, 2005. This work was supported by the Italian Scientific Research Ministry grant Cofin 2001 "Advanced Numerical Methods for Applications of Partial Differential Equations."

http://www.siam.org/journals/sinum/43-1/43542.html

†Department of Mathematics "F. Brioschi," MOX (Modeling and Scientific Computing), Politecnico di Milano, Piazza L. da Vinci 32, I-20133 Milan, Italy (fausto.saleri@polimi.it, alessandro.veneziani@mate.polimi.it). The research of the second author was partially supported by the EU RTN Project "Haemodel."

step. This step is set up in order to reduce the error associated with the splitting and obtain definitively a higher order of accuracy in time.

After a brief introduction to algebraic splitting methods (section 2), we will provide a general approach for setting up such *pressure correction* schemes (section 3). Then we will analyze, in particular, the schemes arising from the pressure correction of both the algebraic Chorin–Temam and the Yosida methods. The former is investigated in section 4. We analyze the splitting error reduction induced by the pressure correction step and prove that in the Stokes (linear) case the scheme features unconditional stability when starting from a backward difference (implicit) time discretization. Then (section 4.2) we establish some formal analogies between our scheme and other pressure correction schemes proposed at a differential level (see [26]).

The pressure correction formulation of the Yosida method is investigated in section 5. In particular, it is possible to prove that the consistency error induced by the splitting on the matrix to be solved at each time step depends on the cube of the time step, which is an improvement of the original Yosida scheme introduced by the pressure correction. On the other hand, we prove that the stability of this scheme is in general conditional, even if applied to the Stokes problem discretized with a backward difference (implicit) scheme.

In section 6 we provide several numerical results, testing the properties of pressure correction schemes. We analyze, in particular, the improvements induced by the pressure correction when applied to high order time discretization schemes. Some conclusions are drawn in section 7.

2. Inexact algebraic factorizations for the Navier–Stokes equations. Consider an open and bounded domain $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$ with boundary $\partial\Omega$ for a time $t \geq 0$. The Navier–Stokes equations for incompressible flows in terms of the velocity, $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$, and the pressure, $p = p(\mathbf{x}, t)$, read as

$$\left\{ \begin{array}{l} \dfrac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} - \nu \triangle \mathbf{u} + \nabla p = \mathbf{f}, \\ \nabla \cdot \mathbf{u} = 0 \end{array} \right. \tag{2.1}$$

for any $(\mathbf{x}, t) \in \Omega \times (0, T]$, with $T > 0$. This system must be completed with the initial condition $\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}^0(\mathbf{x})$ (where $\mathbf{u}^0(\mathbf{x})$ is a given function) and suitable boundary conditions on $\partial\Omega$. Since in the framework of algebraic splitting methods there is a complete independence of the numerical methods of the boundary conditions, we do not specify a specific boundary set. It is, however, understood that some (reasonable) boundary conditions are prescribed on $\partial\Omega$.

In order to have a quantitative evaluation of the flow field in real applications, a numerical approximation has to be carried out. To this aim, the problem has to be discretized with respect to time and space variables. Concerning the *space discretization* issue, we will basically refer to the *Galerkin method* and, in particular, to the *finite element method* (FEM). The most part of what follows can, however, be applied to other space discretization methods as well. For any details concerning the FEM discretization of the Navier–Stokes problem, we refer the reader, e.g., to [24]. In fact, we choose functional spaces for the approximate velocity and pressure fields which satisfy the *inf-sup* or *LBB condition* (see, e.g., [1]). We will denote by $N_{\mathbf{u}}$ the number of velocity degrees of freedom and by N_p the number of pressure degrees of freedom.

For what concerns the *time discretization*, we will refer to classic backward differences methods. Namely, we consider a decomposition of the time interval into

N subintervals (t^n, t^{n+1}) with $t^n = n\Delta t$, where $\Delta t = T/N$ is the uniform positive time step, and collocate the equation in the instants $t^n = n\Delta t$. For the treatment of the nonlinear convective term, we resort to the usual (semi-implicit) linearization $(\mathbf{u}(t^{n+1}) \cdot \nabla)\mathbf{u}(t^{n+1}) \simeq (\mathbf{u}(t^n) \cdot \nabla)\mathbf{u}(t^{n+1})$ or similar ones featuring higher order of time accuracy (see, e.g., [10]). We point out, however, that the most part of the methods investigated here relies on the pattern of the matrix associated with the discrete Navier–Stokes problem. Since the pattern is also the same for the inner iteration of a Newton method (solution of the Jacobian), the methods introduced here can also be applied to a fully implicit treatment of the nonlinear term.

The fully discretized and linearized incompressible Navier–Stokes equations at the time t^{n+1} therefore read

$$\mathcal{A}\mathbf{w}^{n+1} = \mathbf{b}^{n+1}, \tag{2.2}$$

where the vector $\mathbf{b}^k = \left[b_1^k, b_2^k\right]^T$ contains forcing terms and contributions of the boundary conditions, $\mathbf{w}^{n+1} = (\mathbf{u}^{n+1}, \mathbf{p}^{n+1})^T$ denotes the vector of the nodal values of the discrete velocity and pressure, and

$$\mathcal{A} = \begin{bmatrix} \mathrm{C} & \mathrm{D}^T \\ \mathrm{D} & 0 \end{bmatrix}. \tag{2.3}$$

Here D is the discrete divergence operator (i.e., $d_{ij} = -\int_\Omega \nabla \cdot \mathbf{v}_j q_i$, where $\{\mathbf{v}_k\}$ band $\{q_k\}$ are the basis functions for the velocity and the pressure, respectively). D^T denotes the discrete gradient operator, and C collects contributions from the time derivative and the advection and diffusion operators. More specifically, we denote

$$\mathrm{C} = \frac{\alpha}{\Delta t}\mathrm{M} + \mathrm{K},$$

where M is the velocity mass matrix, α is the coefficient of the backward difference formula (BDF) scheme at hand for the velocity field at time t^{n+1}, and K corresponds to the discretization of the diffusive and of the convective terms. In the case of the Stokes problem, K corresponds just to the Laplacian of the velocity, and it is therefore symmetric and positive definite (s.p.d.).

System (2.2) typically features large dimensions and bad conditioning properties. The splitting between the computation of the velocity field from that of the pressure is almost mandatory when large three-dimensional problems are faced. This can be obtained through inexact block LU decompositions. These strategies stem from the following "exact" LU-block factorization of $\mathcal{A}$:

$$\mathcal{A} = \begin{bmatrix} \mathrm{C} & 0 \\ \mathrm{D} & -\mathrm{D}\mathrm{C}^{-1}\mathrm{D}^T \end{bmatrix} \begin{bmatrix} \mathrm{I} & \mathrm{C}^{-1}\mathrm{D}^T \\ 0 & \mathrm{I} \end{bmatrix}.$$

Since the inverse C^{-1} is seldom available, we can set up different schemes, achieving a reduction in the computational cost by suitably approximating C^{-1} with a matrix H_1 in the L-block and H_2 in the U-block. This leads to the following *inexact block LU factorization* (see [23] and also [5] and [6]):

$$\widehat{\mathcal{A}} = \begin{bmatrix} \mathrm{C} & 0 \\ \mathrm{D} & -\mathrm{D}\mathrm{H}_1\mathrm{D}^T \end{bmatrix} \begin{bmatrix} \mathrm{I} & \mathrm{H}_2\mathrm{D}^T \\ 0 & \mathrm{I} \end{bmatrix} = \begin{bmatrix} \mathrm{C} & \mathrm{C}\mathrm{H}_2\mathrm{D}^T \\ \mathrm{D} & \mathrm{D}(\mathrm{H}_2 - \mathrm{H}_1)\mathrm{D}^T \end{bmatrix}. \tag{2.4}$$

The corresponding algebraic fractional-step methods require at the generic time level t^{n+1} the solution of the following systems:

$$\text{L-step}\quad \begin{cases} \mathrm{C}\widetilde{\mathbf{u}}^{n+1} = \mathbf{b}_1^{n+1}, \\ \mathrm{D}\widetilde{\mathbf{u}}^{n+1} - \mathrm{D}\mathrm{H}_1\mathrm{D}^T\tilde{\mathbf{p}}^{n+1} = \mathbf{b}_2^{n+1}, \end{cases} \tag{2.5}$$

$$\text{U-step}\quad \begin{cases} \mathbf{p}^{n+1} = \widetilde{\mathbf{p}}^{n+1}, \\ \mathbf{u}^{n+1} + \mathrm{H}_2\mathrm{D}^T\mathbf{p}^{n+1} = \widetilde{\mathbf{u}}^{n+1}. \end{cases} \tag{2.6}$$

Different choices can be pursued for the approximating matrices H_1 and H_2. In particular, we could take

$$\mathrm{H}_1 = \mathrm{H}_2 = \frac{\Delta t}{\alpha}\mathrm{M}^{-1}$$

or

$$\mathrm{H}_1 = \frac{\Delta t}{\alpha}\mathrm{M}^{-1}, \quad \mathrm{H}_2 = \mathrm{C}^{-1}.$$

The former choice yields a scheme (see [20]) that can be considered the *algebraic* counterpart of the *Chorin–Temam scheme* (briefly, ACT), because of the formal analogy with the original differential splitting method. The latter choice yields the so-called *Yosida method*, introduced in [23] and analyzed in [22]. See also [5] and [6].

The main difference between the two possibilities is that in the Chorin–Temam scheme only the discretized momentum equation is perturbed, while in the Yosida scheme only the mass conservation equation is perturbed (see [27]). In what follows, we will say that an algebraic fractional-step method is of *Yosida type* if $\mathrm{H}_1 = \mathrm{H}$ and $\mathrm{H}_2 = \mathrm{C}^{-1}$, and of *Chorin–Temam* type if $\mathrm{H}_1 = \mathrm{H}_2 = \mathrm{H}$, with H being any convenient approximation of C^{-1}. Note that the approximation of C^{-1} with $\Delta t/\alpha\mathrm{M}^{-1}$ stems from a truncation to the first term of the following well-known Neumann expansion:

$$\mathrm{C}^{-1} = \frac{\Delta t}{\alpha}\left(\mathrm{I}_{N_{\mathbf{u}}} + \frac{\Delta t}{\alpha}\mathrm{M}^{-1}K\right)^{-1}, \quad \mathrm{M}^{-1} = \frac{\Delta t}{\alpha}\sum_{i=0}^{\infty}\left(-\Delta t\mathrm{M}^{-1}K\right)^{i}\mathrm{M}^{-1}.$$

Here $\mathrm{I}_{N_{\mathbf{u}}}$ (I_{N_p}) denotes the identity matrix of dimension $N_{\mathbf{u}}$ (N_p). α is the coefficient of the term evaluated at t^{n+1} in the time discretization scheme adopted. For the implicit Euler scheme, $\alpha = 1$; for a BDF scheme of order two, $\alpha = 3/2$; and for a BDF scheme of order three, $\alpha = 11/6$. In order to improve the accuracy of the inexact factorization, one could choose H by taking more terms in the Neumann expansion. This strategy has been analyzed in [27], and it can lead to some relevant instabilities, even for the Stokes problem. In what follows, we will set

$$\mathrm{H} = \frac{\Delta t}{\alpha}\mathrm{M}^{-1}, \tag{2.7}$$

so that $\mathrm{C}^{-1} = \mathrm{H} + \mathcal{O}(\Delta t)$.

As previously pointed out, note that for both approaches the pressure is only predicted (in the L-step), while the velocity is first predicted in the L-step (with the so-called *intermediate velocity*, $\widetilde{\mathbf{u}}^{n+1}$) and next corrected in the U-step, computing the *end-step velocity.* One could expect some improvements in the accuracy of computation by resorting to a correction also for the pressure field, leading to an *end-step pressure.* In the next section we investigate such schemes.

Remark. A popular modification of the schemes presented above is the so-called *incremental approach.* This approach can be applied to differential projection methods and to algebraic splittings as well. It basically consists of a reformulation of the time-discrete Navier–Stokes problem in such a way that the pressure field p^{n+1} is computed as the sum of an *extrapolation* $\sigma_{n+1}p$ (which is a linear combination of the pressure at the previous time steps) and an *increment* $\delta_{n+1}p$. This reformulation for the algebraic (fully discretized) problem reads

$$\mathcal{A}\begin{bmatrix}\mathbf{U}^{n+1}\\ \mathbf{P}^{n+1}\end{bmatrix}=\begin{bmatrix}\mathbf{b}_1\\ \mathbf{b}_2\end{bmatrix}\Rightarrow\mathcal{A}\begin{bmatrix}\mathbf{U}^{n+1}\\ \delta_{n+1}\mathbf{P}\end{bmatrix}=\begin{bmatrix}\mathbf{b}_1-\sigma_{n+1}\mathbf{P}\\ \mathbf{b}_2\end{bmatrix}, \tag{2.8}$$

having set $\mathbf{P}^{n+1}=\sigma_{n+1}\mathbf{P}+\delta_{n+1}\mathbf{P}$. In this way, possible errors introduced by the splitting affect the pressure increment rather than the pressure itself. This modification improves the accuracy of the solution. For instance, in the van Kan scheme (see [16], [25]), we have

$$\sigma_{n+1}p=p^n,\quad \delta_{n+1}p=p^{n+1}-\sigma_{n+1}=p^{n+1}-p^n, \tag{2.9}$$

which coupled with a Crank–Nicolson time discretization yields a second order scheme (see also [10]). Recent results (see [15]) show that the incremental approach can improve the accuracy of the solution also for higher order time discretizations, e.g., coupling a BDF of order three with an incremental approach by setting

$$\sigma_{n+1}p=2p^n-p^{n-1},\quad \delta_{n+1}p=p^{n+1}-\sigma_{n+1}=p^{n+1}-2p^n+p^{n-1}. \tag{2.10}$$

In our analysis we do not consider the incremental approach, in order to put in evidence the role of the pressure corrections proposed in the next sections. However, in the numerical results section, we will present and comment on the impact of the incremental modification on pressure corrected schemes.

Remark. In the literature, there is an open discussion about the pros and cons of algebraic or differential approaches in the splitting (see [11]). Here, we investigate essentially an algebraic approach, even if we do not claim that this necessarily leads to more accurate results. In fact, as previously pointed out, the algebraic approach has, for sure, the advantage of including all the possible boundary conditions (not only Dirichlet conditions) without taking care of the setup of special (approximate) pressure conditions, which is conversely needed in differential splitting schemes. This makes the algebraic approach appealing in many real problems.

3. Pressure correction algebraic schemes. Let us consider the following *modified* inexact LU factorization of the matrix $\mathcal{A}$, defined in (2.3),

$$\widehat{\mathcal{A}}=\begin{bmatrix}\mathrm{C} & 0\\ \mathrm{D} & -\mathrm{DH}_1\mathrm{D}^T\end{bmatrix}\begin{bmatrix}\mathrm{I} & \mathrm{H}_2\mathrm{D}^T\mathrm{R}\\ 0 & \mathrm{Q}\end{bmatrix}=\begin{bmatrix}\mathrm{C} & \mathrm{CH}_2\mathrm{D}^T\mathrm{R}\\ \mathrm{D} & \mathrm{D}(\mathrm{H}_2\mathrm{D}^T\mathrm{R}-\mathrm{H}_1\mathrm{D}^T\mathrm{Q})\end{bmatrix}, \tag{3.1}$$

where R and Q are square $N_p\times N_p$ matrices that we choose in order to minimize the difference $\mathcal{A}-\widehat{\mathcal{A}}$ in some sense. This (generic) factorization leads to new algebraic fractional step methods, where the L-step is still given by (2.5), while the U-step becomes

$$\text{U-step}\quad \left\{ \begin{array}{l} \mathrm{Q}\mathbf{p}^{n+1} = \widetilde{\mathbf{p}}^{n+1}, \\ \mathbf{u}^{n+1} + \mathrm{H}_2\mathrm{D}^T\mathrm{R}\mathbf{p}^{n+1} = \widetilde{\mathbf{u}}^{n+1}. \end{array} \right. \tag{3.2}$$

Since we have now a pressure correction $(3.2)_1$, we give to schemes in the form (2.5), (3.2) the name *pressure correction methods.*

We still distinguish two approaches: the Chorin–Temam approach, in which we set $\mathrm{H}_1 = \mathrm{H}_2 = \mathrm{H}$, and the Yosida one, where $\mathrm{H}_2 = \mathrm{C}^{-1}$. In the first case, we obtain

$$\widehat{\mathcal{A}} = \left[\begin{array}{cc} \mathrm{C} & \mathrm{CHD}^T\mathrm{R} \\ \mathrm{D} & \mathrm{D}(\mathrm{HD}^T\mathrm{R} - \mathrm{HD}^T\mathrm{Q}) \end{array} \right]. \tag{3.3}$$

If we select $\mathrm{R} = \mathrm{Q}$, the (discrete) mass conservation equation is satisfied without any approximation. This choice will be investigated in the next section.

In the second case, we have

$$\widehat{\mathcal{A}} = \left[\begin{array}{cc} \mathrm{C} & \mathrm{D}^T\mathrm{R} \\ \mathrm{D} & \Sigma\mathrm{R} - \mathrm{SQ} \end{array} \right], \tag{3.4}$$

where

$$\Sigma = \mathrm{DC}^{-1}\mathrm{D}^T, \quad \mathrm{S} = \mathrm{DHD}^T. \tag{3.5}$$

Matrix $-\Sigma$ is the so-called pressure *Schur complement* of $\mathcal{A}$. Observe that thanks to (2.7), S is $\mathcal{O}(\Delta t)$. In this case, if $\mathrm{R} = \mathrm{I}_{N_p}$, the (discrete) momentum equation is fulfilled exactly. We investigate this choice in section 5.

4. The algebraic Chorin–Temam pressure correction scheme. Let us investigate the choice $\mathrm{H}_1 = \mathrm{H}_2$, with $\mathrm{R} = \mathrm{Q}$. Since we are introducing a pressure correction in the ACT method, we will call this choice the *algebraic Chorin–Temam pressure correction* (CTPC) scheme.

The splitting error matrix is

$$\widehat{\mathcal{E}} = \mathcal{A} - \widehat{\mathcal{A}} = \left[\begin{array}{cc} 0 & \mathrm{D}^T - \widehat{\mathrm{D}}^T \\ 0 & 0 \end{array} \right], \qquad \widehat{\mathrm{D}}^T = \mathrm{CHD}^T\mathrm{Q}; \tag{4.1}$$

thus the splitting error vanishes if

$$\mathrm{CHD}^T\mathrm{Q} = \mathrm{D}^T. \tag{4.2}$$

Matrix equation (4.2) is an overdetermined problem. In order to obtain a solution, multiply both the sides of (4.2) by the matrix DH, yielding $\mathrm{BQ} = \mathrm{S}$, where

$$\mathrm{B} = \mathrm{DHCHD}^T. \tag{4.3}$$

This implies that the matrix equation (4.2) is solved up to a nonzero matrix Z such that $\mathrm{DHZ} = 0$. Observe that if the inf-sup condition is fulfilled, matrix B is nonsingular, and then we can compute

$$\mathrm{Q} = \mathrm{B}^{-1}\mathrm{S}. \tag{4.4}$$

Remark. In the case of the Stokes problem, C is s.p.d., and the matrix Q corresponds to solving (4.2) in the least square sense, where the solution yields the minimal error in the norm $\| \cdot \|_{\mathrm{C}}$.

At each time step, the CTPC scheme reads (we neglect the time index $n+1$ for the sake of simplicity) as follows:

1. *Intermediate velocity computation*: $\mathrm{C}\widetilde{\mathbf{U}} = \mathbf{b}_1$.
2. *Intermediate pressure computation.* $\mathrm{S}\widetilde{\mathbf{P}} = \mathrm{D}\widetilde{\mathbf{U}} - \mathbf{b}_2$.
3. *End-of-step pressure computation*: $\mathrm{S}\mathbf{P} = \mathrm{B}\widetilde{\mathbf{P}}$.
4. *End-of-step velocity computation*: $\mathbf{U} = \widetilde{\mathbf{U}} - \mathrm{HD}^T\widetilde{\mathbf{P}}$.

Observe that the two systems for the pressure computation share the same matrix S. This is useful if S can be factorized all at once, allowing an effective direct strategy for the solution of the related systems (see section 6).

Remark. In order to solve (4.2), another possibility resorts to manipulate it in the form $\mathrm{HD}^T\mathrm{Q} = \mathrm{C}^{-1}\mathrm{D}^T$ and then solve it in the least square sense with respect to the matrix norm $\| \cdot \|_{\mathrm{H}^{1/2}}$. This strategy yields $\mathrm{Q} = \mathrm{S}^{-1}\Sigma$. This means that the pressure correction step reads $\mathrm{Q}\mathbf{P} = \widetilde{\mathbf{P}} \Rightarrow \Sigma\mathbf{P} = \mathrm{S}\widetilde{\mathbf{P}}$. This step involves the pressure Schur complement Σ, and it is, in fact, an (ineffective) reformulation of the pressure matrix method (see [24]), corresponding to exploiting the exact factorization (2.3). Therefore it is not feasible.

4.1. Stability and splitting error analysis of CTPC.

4.1.1. Preliminary results. We start with some preliminary notation and lemmas. Starting from the identity

$$\mathrm{CH} = \mathrm{I}_{N_{\mathbf{u}}} + \mathrm{E}_1, \qquad \mathrm{E}_1 = \mathrm{KH}, \tag{4.5}$$

we have

$$\mathrm{B} = \mathrm{DH}\left(\mathrm{I}_{N_{\mathbf{u}}} + \mathrm{E}_1\right)\mathrm{D}^T = \mathrm{S} + \mathrm{W} = \mathrm{S}\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right), \tag{4.6}$$

where

$$\mathrm{W} = \mathrm{DHKHD}^T, \qquad \mathrm{E}_2 = \mathrm{S}^{-1}\mathrm{W}. \tag{4.7}$$

Observe that W is $\mathcal{O}(\Delta t^2)$, thanks to (2.7). Consequently, E_2 is $\mathcal{O}(\Delta t)$. From (4.6) we have

$$\mathrm{B}^{-1} = \left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1}\mathrm{S}^{-1} \Rightarrow \mathrm{B}^{-1}\mathrm{S} = \left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1}. \tag{4.8}$$

Therefore, matrix $\widehat{\mathrm{D}}^T$ introduced in (4.1) admits the following factorization:

$$\hat{\mathrm{D}}^T = \mathrm{CHD}^T\mathrm{B}^{-1}\mathrm{S} = \left(\mathrm{I}_{N_{\mathbf{u}}} + \mathrm{E}_1\right)\mathrm{D}^T\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1}. \tag{4.9}$$

Observe that if H is proportional to $\mathrm{I}_{N_{\mathbf{u}}}$, as it happens in a finite difference discretization, it is possible to verify that the matrix DH is the Moore–Penrose pseudoinverse of $\mathrm{D}^T\mathrm{S}^{-1}$. In general (for a finite element or a spectral discretization) this is not true. In fact, it is possible to verify that $\mathrm{H}^{1/2}\mathrm{D}^T\mathrm{S}^{-1}$ is the Moore–Penrose pseudoinverse of $\mathrm{DH}^{1/2}$. However, for the purpose of the present work, it is useful in the following lemma.

Lemma 4.1. *Matrix* $\mathrm{I}_{N_{\mathbf{u}}} - \mathrm{D}^T\mathrm{S}^{-1}\mathrm{DH}$ *is similar to the matrix*

$$\begin{bmatrix} 0 & 0 \\ 0 & \mathrm{I}_{N_{\mathbf{u}}-N_p} \end{bmatrix}.$$

Proof. The singular value decomposition of the matrix $\mathrm{H}^{1/2}\mathrm{D}^T$ reads

$$\mathrm{H}^{1/2}\mathrm{D}^T = \mathrm{U}^T\Pi\mathrm{V}, \tag{4.10}$$

where U is an orthogonal $N_{\mathbf{u}} \times N_{\mathbf{u}}$ matrix, V is an orthogonal $N_p \times N_p$ matrix, and Π is the $N_{\mathbf{u}} \times N_p$ matrix such that $\Pi_{ij} = \sigma_i\delta_{ij}$, where $\{\sigma_i\}$ are the singular values and δ_{ij} is the Kronecker symbol. Observe that, thanks to the inf-sup condition, $\sigma_i \neq 0$ for any $i = 1, \dots, N_p$. Now, we have

$$\mathrm{I}_{N_{\mathbf{u}}} - \mathrm{D}^T\mathrm{S}^{-1}\mathrm{DH} = \mathrm{H}^{-1/2}\left(\mathrm{I}_{N_{\mathbf{u}}} - \mathrm{H}^{1/2}\mathrm{D}^T\left(\mathrm{DH}^{1/2}\mathrm{H}^{1/2}\mathrm{D}^T\right)^{-1}\mathrm{DH}^{1/2}\right)\mathrm{H}^{1/2}$$

and, thanks to (4.10),

$$\mathrm{I}_{N_{\mathbf{u}}} - \mathrm{H}^{1/2}\mathrm{D}^T\left(\mathrm{DH}^{1/2}\mathrm{H}^{1/2}\mathrm{D}^T\right)^{-1}\mathrm{DH}^{1/2} = \mathrm{I}_{N_{\mathbf{u}}} - \mathrm{U}^T\Pi\mathrm{V}\left(\mathrm{V}^T\Pi^T\mathrm{U}^T\mathrm{U}\Pi\mathrm{V}\right)^{-1}\mathrm{V}^T\Pi^T\mathrm{U}.$$

Observe that $\Pi^T\Pi$ is the $N_p \times N_p$ diagonal matrix with the square of the singular values on the diagonal. In what follows we will set $\Pi_0^2 = \Pi^T\Pi$. The thesis is a consequence of the fact that U and V are orthogonal and that

$$\Pi\Pi_0^{-2}\Pi^T = \begin{bmatrix} \mathrm{I}_{N_p} & 0 \\ 0 & 0 \end{bmatrix}. \qquad \square$$

4.1.2. Splitting error analysis. We are now in position to investigate the splitting error matrix $\mathcal{A} - \hat{\mathcal{A}}$ associated with the CTPC scheme, that is,

$$\widehat{\mathcal{E}}_{CTPC} = \begin{bmatrix} 0 & \mathrm{D}^T - \widehat{\mathrm{D}}^T \\ 0 & 0 \end{bmatrix}.$$

Setting

$$\mathrm{E} = \mathrm{D}^T - \widehat{\mathrm{D}}^T, \tag{4.11}$$

from the definition of $\widehat{\mathrm{D}}^T$ we can straightforwardly verify that

$$\mathrm{DHE} = 0. \tag{4.12}$$

This was to be expected, since we have solved the overdetermined problem (4.2), by projecting it into the subspace image of DH. From (4.9), it follows that

$$\mathrm{E} = \mathrm{D}^T - \left(\mathrm{I}_{N_{\mathbf{u}}} + \mathrm{E}_1\right)\mathrm{D}^T\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1}. \tag{4.13}$$

Assuming that Δt is small enough to exploit the Neumann expansion, which makes sense since E_2 is $\mathcal{O}(\Delta t)$,

$$\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1} = \sum_{k=0}^{\infty}(-\mathrm{E}_2)^k,$$

from (4.6), (4.7), (4.8), and (4.5) we have

$$\begin{aligned} &\mathrm{D}^T - \left(\mathrm{I}_{N_{\mathbf{u}}} + \mathrm{E}_1\right)\mathrm{D}^T\left(\mathrm{I}_{N_p} - \mathrm{E}_2 + \mathrm{E}_2^2 - \dots\right) \\ &= \mathrm{D}^T - \left(\mathrm{I}_{N_{\mathbf{u}}} + \mathrm{E}_1\right)\mathrm{D}^T\left(\mathrm{I}_{N_p} - \mathrm{E}_2 + \mathcal{O}(\Delta t^2)\right), \end{aligned} \tag{4.14}$$

yielding

$$\mathrm{E} = -\mathrm{E}_1\mathrm{D}^T \boxed{+\mathrm{D}^T\mathrm{E}_2} + \mathcal{O}(\Delta t^2).$$

In the boxed term we put in evidence the specific contribution on the error given by the pressure correction. Since both the matrices KHD^T and $\mathrm{D}^T\mathrm{S}^{-1}\mathrm{W}$ are $\mathcal{O}(\Delta t)$, we conclude that the splitting error is at least first order in time. Unfortunately, this conclusion does not give significant improvements in terms of order of accuracy with respect to the original ACT scheme. The main difference is that in the latter scheme the splitting error is dominated by the term KHD^T, while in the CTPC scheme the matrix error is dominated by

$$-\left(\mathrm{I}_{N_{\mathbf{u}}} - \mathrm{D}^T\mathrm{S}^{-1}\mathrm{DH}\right)\mathrm{E}_1\mathrm{D}^T. \tag{4.15}$$

Thanks to Lemma 4.1, it is to be expected that the matrix in brackets, having N_p null eigenvalues, will reduce the error associated with the scheme, in comparison with the ACT scheme, even if it is not possible to prove that the order of accuracy of the scheme is improved. Numerical results confirm that the scheme is in general only first order accurate in time (see section 6).

However, for some special space discretization, it is interesting to point out the following circumstance.

PROPOSITION 4.2. *If* $\mathrm{KHD}^T = \nu\mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{DHD}^T$*, the splitting error matrix* E *vanishes.*

Proof. From (4.13) it follows that

$$\mathrm{E}\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right) = \mathrm{D}^T + \mathrm{D}^T\mathrm{E}_2 - \mathrm{D}^T - \mathrm{E}_1\mathrm{D}^T = \mathrm{D}^T\mathrm{E}_2 - \mathrm{E}_2\mathrm{D}^T. \tag{4.16}$$

If $\mathrm{E}_1\mathrm{D}^T = \mathrm{KHD}^T = \nu\mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{DHD}^T$, then we have $\mathrm{E}_2 = \nu\mathrm{S}^{-1}\mathrm{DHD}^T\mathrm{M}_p^{-1}\mathrm{DHD}^T = \nu\mathrm{M}_p^{-1}\mathrm{S}$. Recalling that $\mathrm{S} = \mathrm{DHD}^T$, from (4.16) it follows that $\mathrm{E}\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right) = \nu\mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{S} - \nu\mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{DHD}^T = 0$. Thus, since $\mathrm{I}_{N_p} + \mathrm{E}_2$ is invertible, the thesis follows. □

Observe that the hypothesis in Proposition 4.2 reinterpreted at the continuous level yields the following identity:

$$-\nu\triangle\nabla = -\nu\nabla\cdot(\nabla)\nabla.$$

Moreover, in the context of finite difference space discretization the same hypothesis has been advocated in [18] as a *compatibility condition* holding for a special set of velocity boundary conditions in rectangular domains.

4.1.3. Stability analysis. The main result is the following proposition.

PROPOSITION 4.3. *Consider the Stokes problem discretized with an implicit Euler time discretization scheme ($\alpha = 1$). Then the CTPC scheme is* unconditionally stable.

Proof. We consider a problem where the forcing term and the boundary conditions are null as well, since they do not influence the stability properties of the scheme.

The CTPC scheme actually amounts to solving the following problem:

$$\begin{bmatrix} \mathrm{C} & \hat{\mathrm{D}}^T \\ \mathrm{D} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{U}^{n+1} \\ \mathbf{P}^{n+1} \end{bmatrix} = \begin{bmatrix} \dfrac{1}{\Delta t}\mathrm{M}\mathbf{U}^n \\ \mathbf{0} \end{bmatrix}, \tag{4.17}$$

where $\mathrm{C} = \frac{1}{\Delta t}\mathrm{M}+\mathrm{K}$ (with K s.p.d.) and $\hat{\mathrm{D}}^T$ has been introduced in (4.1). Eliminating the pressure unknowns in (4.17), we obtain

$$\mathbf{U}^{n+1} = \frac{1}{\Delta t}\mathrm{C}^{-1}\mathrm{M}\mathbf{U}^n - \frac{1}{\Delta t}\mathrm{C}^{-1}\hat{\mathrm{D}}^T\left(\mathrm{D}\mathrm{C}^{-1}\hat{\mathrm{D}}^T\right)^{-1}\mathrm{D}\mathrm{C}^{-1}\mathrm{M}\mathbf{U}^n. \tag{4.18}$$

Observe that

$$\begin{aligned}\hat{\mathrm{D}}^T \quad & \left(\mathrm{D}\mathrm{C}^{-1}\hat{\mathrm{D}}^T\right)^{-1}\\ &= \left(\mathrm{I}_{N_\mathbf{u}}+\mathrm{E}_1\right)\mathrm{D}^T\left(\mathrm{I}_{N_p}+\mathrm{E}_2\right)^{-1}\left(\mathrm{D}\mathrm{C}^{-1}\left(\mathrm{I}_{N_\mathbf{u}}+\mathrm{E}_1\right)\mathrm{D}^T\left(\mathrm{I}_{N_p}+\mathrm{E}_2\right)^{-1}\right)^{-1}\\ &= \left(\mathrm{I}_{N_\mathbf{u}}+\mathrm{E}_1\right)\mathrm{D}^T\left(\mathrm{D}\mathrm{C}^{-1}\left(\mathrm{I}_{N_\mathbf{u}}+\mathrm{E}_1\right)\mathrm{D}^T\right)^{-1}.\end{aligned}$$

Moreover, from the definition of E_1 and C, we have

$$\mathrm{C}^{-1}\left(\mathrm{I}_{N_\mathbf{u}}+\mathrm{E}_1\right) = \Delta t\mathrm{M}^{-1}\left(\mathrm{I}_{N_\mathbf{u}}+\Delta t\mathrm{K}\mathrm{M}^{-1}\right)^{-1}\left(\mathrm{I}_{N_\mathbf{u}}+\mathrm{E}_1\right) = \Delta t\mathrm{M}^{-1},$$

so that $\hat{\mathrm{D}}^T(\mathrm{D}\mathrm{C}^{-1}\hat{\mathrm{D}}^T)^{-1}$ reduces to $\left(\mathrm{I}_{N_\mathbf{u}}+\Delta t\mathrm{K}\mathrm{M}^{-1}\right)\mathrm{D}^T\mathrm{S}^{-1}$. We have therefore that the second matrix on the right-hand side of (4.18) becomes

$$\frac{1}{\Delta t}\mathrm{C}^{-1}\hat{\mathrm{D}}^T\left(\mathrm{D}\mathrm{C}^{-1}\hat{\mathrm{D}}^T\right)^{-1}\mathrm{D}\mathrm{C}^{-1}\mathrm{M} = \mathrm{M}^{-1}\mathrm{D}^T\mathrm{S}^{-1}\mathrm{D}\mathrm{C}^{-1}\mathrm{M},$$

yielding

$$\mathbf{U}^{n+1} = \frac{1}{\Delta t}\left(\mathrm{M}^{-1}-\mathrm{M}^{-1}\mathrm{D}^T\left(\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T\right)^{-1}\mathrm{D}\mathrm{M}^{-1}\right)\mathrm{M}\mathrm{C}^{-1}\mathrm{M}\mathbf{U}^n, \tag{4.19}$$

or, equivalently,

$$\mathbf{U}^{n+1} = \mathrm{M}^{-1}\left(\mathrm{I}_{N_\mathbf{u}}-\mathrm{D}^T\left(\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T\right)^{-1}\mathrm{D}\mathrm{M}^{-1}\right)\mathrm{M}\left(\mathrm{I}_{N_\mathbf{u}}+\Delta t\mathrm{M}^{-1}\mathrm{K}\right)\mathbf{U}^n. \tag{4.20}$$

We prove that for any $\Delta t > 0$

$$\left\|\mathrm{M}^{-1}\left(\mathrm{I}_{N_\mathbf{u}}-\mathrm{D}^T\left(\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T\right)^{-1}\mathrm{D}\mathrm{M}^{-1}\right)\mathrm{M}\left(\mathrm{I}_{N_\mathbf{u}}+\Delta t\mathrm{M}^{-1}\mathrm{K}\right)^{-1}\right\|_2 < 1. \tag{4.21}$$

For a generic matrix X, we will denote by ρ_X its spectral radius. First of all, observe that

$$||\mathrm{I}_{N_\mathbf{u}}-\mathrm{D}^T\left(\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T\right)^{-1}\mathrm{D}\mathrm{M}^{-1}||_2 = 1. \tag{4.22}$$

This is, in fact, a consequence of Lemma 4.1, since the matrix in (4.22) corresponds exactly to the matrix $\mathrm{I}_{N_\mathbf{u}} - \mathrm{D}^T\mathrm{S}^{-1}\mathrm{D}\mathrm{H}$ considered in the lemma. If follows that the matrix $\mathrm{M}^{-1}(\mathrm{I}_{N_\mathbf{u}} - \mathrm{D}^T\left(\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T\right)^{-1}\mathrm{D}\mathrm{M}^{-1})\mathrm{M}$ still has a unit spectral radius, being similar to the one in (4.22). Now, due to the positiveness of $\Delta t\mathrm{M}^{-1}\mathrm{K}$ with respect to the scalar product weighted by the s.p.d. mass matrix M, we have that $\rho_{(\mathrm{I}_{N_\mathbf{u}}+\Delta t\mathrm{M}^{-1}\mathrm{K})^{-1}} < 1$. Therefore, we have

$$\begin{aligned}&\left\|\mathrm{M}^{-1}\left(\mathrm{I}_{N_\mathbf{u}}-\mathrm{D}^T\left(\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T\right)^{-1}\mathrm{D}\mathrm{M}^{-1}\right)\mathrm{M}\left(\mathrm{I}_{N_\mathbf{u}}+\Delta t\mathrm{M}^{-1}\mathrm{K}\right)^{-1}\right\|_2\\ &\le \left\|\mathrm{M}^{-1}\left(\mathrm{I}_{N_\mathbf{u}}-\mathrm{D}^T\left(\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T\right)^{-1}\mathrm{D}\mathrm{M}^{-1}\right)\right\|_2\left\|\left(\mathrm{I}_{N_\mathbf{u}}+\Delta t\mathrm{M}^{-1}\mathrm{K}\right)^{-1}\right\|_2 < 1.\end{aligned} \tag{4.23}$$

The first inequality is due to (4.22) and the fact that each matrix of the product is s.p.d., so its norm corresponds to its spectral radius. Equations (4.18) and (4.23) yield

$$||\mathbf{U}^{n+1}||_2 < ||\mathbf{U}^n||_2$$

for any $\Delta t > 0$, giving the unconditional stability result. □

Remark. With a similar approach, it can be proved that the unconditional stability holds also for any implicit (unconditionally stable) BDF time discretization.

Remark. It is worthwhile to point out the fact that on the stability of the scheme the error matrix E_2 does not play any role. Actually, since the error matrix E_1 of the present scheme is the same of the original ACT scheme, the unconditional stability of CTPC could be directly inferred from the unconditional stability of ACT. E_2 is therefore set up in such a way that it should reduce the splitting error without affecting the stability of the scheme. Other schemes featuring a high order splitting error do not share the same stability properties. As we will see, also the pressure correction of the Yosida scheme is affected by instabilities.

4.2. CTPC and differential pressure correction schemes. The introduction of a pressure correction for improving the numerical solution is not new in the field of projection methods for the Navier–Stokes equations. In Timmermans, Minev, and de Vosse [26] a variant of the second order van Kan scheme (see [16]) is proposed for improving the pressure computation in the framework of splitting (differential) schemes. Following Prohl [21], who has extensively analyzed this scheme, we present the corresponding formulation starting from the Chorin–Temam method. Suppose we have homogeneous Dirichlet conditions on the whole boundary of the computational domain Ω. First, compute the intermediate velocity $\widetilde{\mathbf{u}}^{n+1}$ as the solution of the following advection-diffusion semidiscrete problem:

$$\frac{1}{\Delta t}\left(\widetilde{\mathbf{u}}^{n+1} - \mathbf{u}^n\right) - \nu\triangle\widetilde{\mathbf{u}}^{n+1} + \left(\mathbf{u}^n \cdot \nabla\right)\widetilde{\mathbf{u}}^{n+1} = \mathbf{f}, \tag{4.24}$$

with $\widetilde{\mathbf{u}}^{n+1} = \mathbf{0}$ on $\partial\Omega$. Then compute an "intermediate" pressure as the solution of the following Poisson problem:

$$\Delta t\triangle\widetilde{p}^{n+1} = \nabla \cdot \widetilde{\mathbf{u}}^{n+1}, \tag{4.25}$$

with $\partial_{\mathbf{n}}\widetilde{p}^{n+1} = 0$ on $\partial\Omega$, where $\mathbf{n}$ is the outward normal unit vector to $\partial\Omega$. The end-of-step velocity is now given by

$$\mathbf{u}^{n+1} = \widetilde{\mathbf{u}}^{n+1} - \Delta t\nabla\widetilde{p}^{n+1}. \tag{4.26}$$

If we take $\widetilde{p}^{n+1}$ as the end-of-step pressure, we have actually the classical Chorin–Temam scheme. In the Timmermans proposal, we take

$$p^{n+1} = \widetilde{p}^{n+1} - \nu\nabla \cdot \widetilde{\mathbf{u}}^{n+1}. \tag{4.27}$$

At the semidiscrete level (time-discrete and space-continuous) it is possible to verify that this scheme is *strongly consistent* with the Stokes problem, i.e., that solving the Chorin–Temam method with the correction (4.27) amounts exactly to solve the Stokes equations, without any splitting error. This strong consistency, however, fails to be verified in the case of the Navier–Stokes problem. Actually, in this case, a Lagrangian treatment of the time derivative needs to be pursued [26]. Prohl has, moreover, proved that this pressure correction introduces a "smoothing effect" only on the pressure error in the interior domain, and, however, the pressure correction step does not improve the order of accuracy of the method.

Here we want to establish some relations between the Timmermans method and our CTPC scheme. Indeed, exploiting (4.25) we can eliminate $\widetilde{\mathbf{u}}^{n+1}$ in (4.27), yielding

$$p^{n+1} = \widetilde{p}^{n+1} - \nu\Delta t\Delta\widetilde{p}^{n+1} = (\mathcal{I} - \nu\Delta t\Delta)\,\widetilde{p}^{n+1}, \tag{4.28}$$

where $\mathcal{I}$ denotes the identity operator.

On the other hand, concerning the CTPC scheme applied to the Stokes problem, we assume that matrix $\Delta t\mathrm{K}\mathrm{M}^{-1}\mathrm{D}^T$ can be factorized as $\nu\Delta t\mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T$ (as we have assumed in Proposition 4.2). The end-of-step pressure of the CTPC schemes becomes

$$\begin{aligned}\mathbf{P}^{n+1} &= \mathrm{S}^{-1}\mathrm{B}\widetilde{\mathbf{P}}^{n+1} = \left(\mathrm{I}_{N_p} + \mathrm{S}^{-1}\mathrm{W}\right)\widetilde{\mathbf{P}}^{n+1}\\ &= \left(\mathrm{I}_{N_p} + \Delta t^2\mathrm{S}^{-1}\mathrm{D}\mathrm{M}^{-1}\mathrm{K}\mathrm{M}^{-1}\mathrm{D}^T\right)\widetilde{\mathbf{P}}^{n+1} = \left(\mathrm{I}_{N_p} + \nu\Delta t\mathrm{M}_p^{-1}\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T\right)\widetilde{\mathbf{P}}^{n+1}.\end{aligned}$$

A formal analogy with (4.28) can be drawn if we read matrix $\nu\mathrm{D}\mathrm{M}^{-1}\mathrm{D}^T$ as a discrete counterpart of the pressure operator $-\nu\Delta$.

The algebraic reformulation, actually, has some advantages. As already pointed out (also in [26]), in the differential framework there is the problem of determining boundary conditions for the pressure problem, which in the algebraic approach are not required. Moreover, we point out that the CTPC scheme naturally embodies the presence of a convective term, and it does not necessarily need a Lagrangian treatment of the time derivative or to an explicit treatment of the convective term, as required in the Timmermans work.

Finally, we point out that the CTPC scheme provides *discrete-divergence* solutions, which is not true for the Timmermans method (whose divergence is null at the continuous level).

5. The algebraic Yosida pressure correction scheme. Let us consider now the pressure correction approach applied following the Yosida strategy, i.e., $\mathrm{H}_2 = \mathrm{C}^{-1}$ and $\mathrm{R} = \mathrm{I}_{N_p}$. The splitting error matrix in this case is

$$\widehat{\mathcal{E}}_{YPC} = \mathcal{A} - \widehat{\mathcal{A}}_{YPC} = \begin{bmatrix} 0 & 0 \\ 0 & \mathrm{SQ} - \Sigma \end{bmatrix}. \tag{5.1}$$

The problem of finding out a matrix such that the splitting error vanishes this time is clearly well posed, and the solution is

$$\Sigma - \mathrm{SQ} = \mathbf{0} \Rightarrow \mathrm{Q} = \mathrm{S}^{-1}\Sigma, \tag{5.2}$$

corresponding to the pressure correction step $\Sigma\mathbf{P}^{n+1} = \mathrm{S}\widetilde{\mathbf{P}}^{n+1}$. This is another formulation of the pressure matrix method (see [24]) and therefore is not interesting in the perspective of the present work.

However, we would like to introduce an approximate computation of Q which is computationally affordable. It corresponds to solve again an overdetermined problem related to (5.2). Multiply the two sides of the first equation in (5.2) by $\mathrm{D}^T\mathrm{M}_p^{-1}$, yielding $\mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{D}\mathrm{C}^{-1}\mathrm{D}^T = \mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{D}\mathrm{H}\mathrm{D}^T\mathrm{Q}$. Since $\mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{D}$ is nonsingular if the inf-sup condition is fulfilled, we can write

$$\mathrm{C}^{-1}\mathrm{D}^T = \mathrm{H}\mathrm{D}^T\mathrm{Q} \Rightarrow \mathrm{D}\mathrm{H}\mathrm{D}^T = \mathrm{D}\mathrm{H}\mathrm{C}\mathrm{H}\mathrm{D}^T\mathrm{Q},$$

corresponding to the choice $\mathrm{Q} = \mathrm{B}^{-1}\mathrm{S}$, which is exactly the same matrix setup for the CTPC scheme (see (4.4), (4.3)). Dropping again the time index $n+1$, the Yosida pressure correction scheme therefore reads as follows:

1. *Intermediate velocity computation*: $\mathrm{C}\widetilde{\mathbf{U}} = \mathbf{b}_1$.
2. *Intermediate pressure computation*: $\mathrm{S}\widetilde{\mathbf{P}} = \mathrm{D}\widetilde{\mathbf{U}} - \mathbf{b}_2$.
3. *End-of-step pressure computation*: $\mathrm{S}\mathbf{P} = \mathrm{B}\widetilde{\mathbf{P}}$.
4. *End-of-step velocity computation*: $\mathrm{C}\mathbf{U} = \mathbf{b}_1 - \mathrm{D}^T\mathbf{P}$.

Observe that the two problems for both the intermediate and end-of-step pressures are still solved by solving the same matrix S.

5.1. Stability and splitting error analysis.

5.1.1. Splitting error analysis. Starting from (5.1), the splitting error is given by the block (2,2) of $\mathcal{E}_{YPC}$, where,

$$\mathrm{E}_{YPC} = \mathrm{S}\mathrm{B}^{-1}\mathrm{S} - \mathrm{D}\mathrm{C}^{-1}\mathrm{D}^T.$$

Exploiting (4.6), (4.7), (4.8), and (4.11), the previous matrix can be reduced to

$$\begin{aligned}
\mathrm{E}_{YPC} &= \mathrm{S}\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1} - \mathrm{DH}\left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)^{-1}\mathrm{D}^T \\
&= \mathrm{DHD}^T\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1} - \mathrm{DH}\left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)^{-1}\mathrm{D}^T \\
&= \mathrm{DH}\left(\mathrm{D}^T\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1} - \left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)^{-1}\mathrm{D}^T\right) \\
&= \mathrm{DH}\left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)^{-1}\left(\left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)\mathrm{D}^T\left(\mathrm{I}_{N_p} + \mathrm{S}^{-1}\mathrm{W}\right)^{-1} - \mathrm{D}^T\right) \\
&= -\mathrm{DH}\left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)^{-1}\mathrm{E}.
\end{aligned} \tag{5.3}$$

The splitting error associated with the Yosida pressure correction (YPC) scheme can now be estimated, by assuming that Δt is small enough to exploit the following Neumann expansion:

$$\left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)^{-1} = \sum_{i=0}^{\infty}(-\mathrm{E}_1)^i. \tag{5.4}$$

This makes sense, since, as we have pointed out, $\mathrm{E}_1 = \mathcal{O}(\Delta t)$.

PROPOSITION 5.1. *The splitting error matrix associated with the YPC scheme is* $\mathcal{O}(\Delta t^3)$

Proof. By exploiting the Neumann expansion (5.4), we have from (5.3) and (4.12)

$$\mathrm{DH}\left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)^{-1}\mathrm{E} = \mathrm{DHE} - \mathrm{DHE}_1\mathrm{E} + \mathrm{DHE}_1^2\mathrm{E} - \cdots = -\mathrm{DHE}_1\mathrm{E} + \text{high order terms.} \tag{5.5}$$

On the other hand, we have that $\mathrm{H} = \mathcal{O}(\Delta t)$, $\mathrm{E}_1 = \mathcal{O}(\Delta t)$, and (from the analysis carried out in section 4.1) $\mathrm{E} = \mathcal{O}(\Delta t)$, so that the thesis is proven. □

In order to have better insight to the benefits introduced by the pressure correction, let us suppose that, besides the Neumann expansion (5.4), it is possible to also expand $(\mathrm{I}_{N_p} + \mathrm{E}_2)^{-1}$, where $\mathrm{E}_2 = \mathrm{S}^{-1}\mathrm{W}$: $(\mathrm{I}_{N_p} + \mathrm{E}_2)^{-1} = \sum_{i=0}^{\infty}(-\mathrm{E}_2)^i$. This still makes sense since also E_2 is $\mathcal{O}(\Delta t)$. Exploiting the Neumann expansions and recalling (4.8), we have

$$\begin{aligned}
&\mathrm{E}_{YPC} = \mathrm{S}\left(\mathrm{I}_{N_p} + \mathrm{E}_2\right)^{-1} - \mathrm{DH}\left(\mathrm{I}_{N_\mathbf{u}} + \mathrm{E}_1\right)^{-1}\mathrm{D}^T \\
&= \mathrm{S}\left(\mathrm{I}_{N_p} - \boxed{\mathrm{E}_2 - \mathrm{E}_2^2 + \mathcal{O}(\Delta t^3)}\right) - \mathrm{DH}\left(\mathrm{I}_{N_\mathbf{u}} - \mathrm{E}_1 - E_1^2 + \mathcal{O}(\Delta t^3)\right)\mathrm{D}^T.
\end{aligned}$$

The boxed terms are the contribution of the error due to the pressure correction; i.e., they were absent in the original Yosida scheme. This is to outline how the pressure correction acts. Indeed, by recalling the definition of E_1 and E_2 in (3.5) and (4.7), the first two terms of the two Neumann expansions cancel themselves (in the original Yosida scheme only the first ones were canceled), yielding

$$\mathrm{E}_{YPC} = -\mathrm{DHKH}\left(\mathrm{I}_{N_{\mathbf{u}}} - \mathrm{D}^T\mathrm{S}^{-1}\mathrm{DH}\right)\mathrm{KHD}^T + \mathcal{O}(\Delta t^4),$$

the first term being $\mathcal{O}(\Delta t^3)$, as already proved. Observe that the matrix in brackets in this term is the same one investigated in Lemma 4.1, and its spectral radius is therefore 1.

Remark. From the expression of the splitting error (5.3) and Proposition 4.2 it follows that also for the YPC scheme *the splitting error vanishes* whenever $\mathrm{KHD}^T = \nu\mathrm{D}^T\mathrm{M}_p^{-1}\mathrm{DHD}^T$.

5.1.2. Stability analysis. A different formulation of the matrix E_{YPC} is needed in order to carry out a stability analysis. Let us introduce the QR factorization (remember that H is s.p.d.) $\mathrm{H}^{1/2}\mathrm{D}^T = \mathrm{UR}$, where U is an orthogonal square ($N_{\mathbf{u}}\times N_{\mathbf{u}}$) matrix, and, if the inf-sup condition holds, R is a triangular full-rank $N_{\mathbf{u}} \times N_p$ matrix such that $\mathrm{R} = \left[\begin{smallmatrix}\mathrm{R}_0\\ \mathbf{0}\end{smallmatrix}\right]$, where R_0 is nonsingular and square ($N_p \times N_p$). In this way, we have the Cholesky factorization $\mathrm{S} = \mathrm{R}_0^T\mathrm{R}_0$. Matrix $\mathrm{SB}^{-1}\mathrm{S}$ therefore reads

$$\begin{aligned}\mathrm{SB}^{-1}\mathrm{S} &= \mathrm{S}\left(\mathrm{DHCHD}^T\right)^{-1}\mathrm{S}\\ &= \mathrm{R}_0^T\mathrm{R}_0\left(\mathrm{R}^T\mathrm{U}^T\mathrm{H}^{1/2}\left(\mathrm{H}^{-1}+\mathrm{K}\right)\mathrm{H}^{1/2}\mathrm{UR}\right)^{-1}\mathrm{R}_0^T\mathrm{R}_0\\ &= \mathrm{R}_0^T\left(\mathrm{R}_0^{-T}\mathrm{R}^T\mathrm{U}^T\mathrm{H}^{1/2}\left(\mathrm{H}^{-1}+\mathrm{K}\right)\mathrm{H}^{1/2}\mathrm{URR}_0^{-1}\right)^{-1}\mathrm{R}_0\\ &= \mathrm{R}_0^T\left(\begin{bmatrix}\mathrm{I}_{N_p} & \mathbf{0}\end{bmatrix}\left(\mathrm{I}_{N_{\mathbf{u}}}+\mathrm{U}^T\mathrm{H}^{1/2}\mathrm{KH}^{1/2}\mathrm{U}\right)\begin{bmatrix}\mathrm{I}_{N_p} & \mathbf{0}\end{bmatrix}^T\right)^{-1}\mathrm{R}_0.\end{aligned} \tag{5.6}$$

Matrix Σ, on the other hand, can be resorted as follows:

$$\begin{aligned}&\Sigma = \mathrm{D}\left(\mathrm{H}^{-1}+\mathrm{K}\right)^{-1}\mathrm{D}^T = \mathrm{DH}^{1/2}\left(\mathrm{I}_{N_{\mathbf{u}}}+\mathrm{H}^{1/2}\mathrm{KH}^{1/2}\right)^{-1}\mathrm{H}^{1/2}\mathrm{D}^T\\ &= \mathrm{R}^T\mathrm{U}^T\left(\mathrm{I}_{N_{\mathbf{u}}}+\mathrm{H}^{1/2}\mathrm{KH}^{1/2}\right)^{-1}\mathrm{UR} = \begin{bmatrix}\mathrm{R}_0^T & \mathbf{0}\end{bmatrix}\left(\mathrm{I}_{N_{\mathbf{u}}}+\mathrm{U}^T\mathrm{H}^{1/2}\mathrm{KH}^{1/2}\mathrm{U}\right)^{-1}\begin{bmatrix}\mathrm{R}_0^T & \mathbf{0}\end{bmatrix}^T.\end{aligned} \tag{5.7}$$

Now set $\mathrm{N} = \mathrm{I}_{N_{\mathbf{u}}} + \mathrm{U}^T\mathrm{H}^{1/2}\mathrm{KH}^{1/2}\mathrm{U}$, and since N is s.p.d., we denote

$$\mathrm{N} = \begin{bmatrix}\mathrm{N}_{11} & \mathrm{N}_{12}\\ \mathrm{N}_{12}^T & \mathrm{N}_{22}\end{bmatrix},$$

where N_{11} is $N_p \times N_p$ and N_{22} is $(N_{\mathbf{u}} - N_p) \times (N_{\mathbf{u}} - N_p)$. Observe that, in the case of the Stokes problem (i.e., K is s.p.d.), from the Sylvester criterion both the diagonal blocks are s.p.d. With these positions, we have that

$$\mathrm{SQ} - \Sigma = \mathrm{R}_0^T\left(\mathrm{N}_{11}^{-1} - \left(\mathrm{N}_{11} - \mathrm{N}_{12}\mathrm{N}_{22}^{-1}\mathrm{N}_{12}^T\right)^{-1}\right)\mathrm{R}_0. \tag{5.8}$$

Observe that also matrix $\mathrm{N}_{11} - \mathrm{N}_{12}\mathrm{N}_{22}^{-1}\mathrm{N}_{12}^T$ is s.p.d., being the first $N_p \times N_p$ block component of the inverse of N.

Since N_{11} is s.p.d., we can rearrange the previous matrix in the following way:

$$\mathrm{SQ} - \Sigma = \mathrm{R}_0^T\mathrm{N}_{11}^{-1/2}\left(\mathrm{I}_{N_p} - \left(\mathrm{I}_{N_p} - \mathrm{N}_{11}^{-1/2}\mathrm{N}_{12}\mathrm{N}_{22}^{-1}\mathrm{N}_{12}^T\mathrm{N}_{11}^{-1/2}\right)^{-1}\right)\mathrm{N}_{11}^{-1/2}\mathrm{R}_0. \tag{5.9}$$

Starting from this reformulation of the error matrix, unfortunately we are not able to prove an unconditional stability result, even for the Stokes problem with an implicit Euler time discretization. Indeed, we have the following results.

Proposition 5.2. *In the Stokes case, discretized with the implicit Euler scheme, error matrix* $\mathrm{SQ} - \Sigma$ *is symmetric and negative semidefinite.*

Proof. Resorting to (5.9), consider the following scalar product:

$$s = \mathbf{x}^T \mathrm{R}_0^T \mathrm{N}_{11}^{-1/2} \left(\mathrm{I}_{N_p} - \left(\mathrm{I}_{N_p} - \mathrm{N}_{11}^{-1/2} \mathrm{N}_{12} \mathrm{N}_{22}^{-1} \mathrm{N}_{12}^T \mathrm{N}_{11}^{-1/2} \right)^{-1} \right) \mathrm{N}_{11}^{-1/2} \mathrm{R}_0 \mathbf{x}.$$

We prove that $s \leq 0$ for each $\mathbf{x} \in \mathbb{R}^{N_p}$. Set $\mathrm{V} = \mathrm{N}_{11}^{-1/2} \mathrm{N}_{12} \mathrm{N}_{22}^{-1} \mathrm{N}_{12}^T \mathrm{N}_{11}^{-1/2}$ and $\mathrm{J} = \left(\mathrm{I}_{N_p} - \mathrm{V} \right)^{-1}$. As previously pointed out, J is s.d.p. The thesis amounts therefore to prove that $\mathrm{I}_{N_p} - \mathrm{J}$ is negative semidefinite. First of all, observe that if we set $\mathbf{v} = \mathrm{J}\mathbf{y}$, we obtain by definition $\mathbf{v} - \mathrm{V}\mathbf{v} = \mathbf{y} \Rightarrow \mathrm{J}\mathbf{y} - \mathrm{VJ}\mathbf{y} = \mathbf{y} \Rightarrow \left(\mathrm{I}_{N_p} - \mathrm{J} \right) \mathbf{y} = -\mathrm{VJ}\mathbf{y}$. Therefore, since J is s.d.p. we get, with obvious notation, $s = -\mathbf{y}^T \mathrm{VJ} \mathbf{y} = -\mathbf{y}^T \mathrm{J}^{1/2} \mathrm{J}^{-1/2} \mathrm{V} \mathrm{J}^{1/2} \mathrm{J}^{1/2} \mathbf{y} = -\mathbf{z}^T \mathrm{J}^{-1/2} \mathrm{V} \mathrm{J}^{1/2} \mathbf{z}$, where $\mathbf{y} = \mathrm{N}_{11}^{-1/2} \mathrm{R}_0 \mathbf{x}$ and $\mathbf{z} = \mathrm{J}^{1/2} \mathbf{y}$. Since V and J are both symmetric and, by construction, share the same set of orthogonal eigenvectors, it is possible to verify that $\mathbf{z}^T \mathrm{J}^{-1/2} \mathrm{V} \mathrm{J}^{1/2} \mathbf{z} \geq 0$, yielding the thesis. □

The previous result is negative in view of the stability analysis of the scheme. Actually, it implies that the scheme introduces a mass source in the fluid. This clearly reflects negatively on the stability of the scheme.

Proposition 5.3. *In the case of the Stokes problem discretized in time with the implicit Euler method, the YPC scheme is conditionally stable.*

Proof. The YPC scheme associated with a backward Euler time discretization, in a homogeneous Dirichlet case with null forcing terms, resorts to solve at each time step the following system:

$$\left[\begin{array}{cc} \mathrm{C} & \mathrm{D}^T \\ -\mathrm{D} & \mathrm{SQ} - \Sigma \end{array} \right] \left[\begin{array}{c} \mathbf{U}^{n+1} \\ \mathbf{P}^{n+1} \end{array} \right] = \left[\begin{array}{c} \dfrac{1}{\Delta t} \mathrm{M} \mathbf{U}^n \\ \mathbf{0} \end{array} \right].$$

The conditional stability is proved by multiplying the two sides by $\left[\mathbf{U}^{n+1} \mathbf{P}^{n+1} \right]^T$ and applying the Young inequality. We obtain, indeed,

$$\frac{1}{2\Delta t} \mathbf{U}^{n+1} \mathrm{M} \mathbf{U}^{n+1} + \mathbf{U}^{n+1} \mathrm{K} \mathbf{U}^{n+1} - \left| \mathbf{P}^{n+1} \left(\mathrm{SQ} - \Sigma \right) \mathbf{P}^{n+1} \right| \leq \frac{1}{2\Delta t} \mathbf{U}^n \mathrm{M} \mathbf{U}^n.$$

Since $\mathrm{SQ} - \Sigma$ vanishes when Δt tends to zero, it is possible to select a time step Δt_{max} such that for each $\Delta t \leq \Delta t_{max}$ we have $\mathbf{U}^{n+1} \mathrm{K} \mathbf{U}^{n+1} - \left| \mathbf{P}^{n+1} \left(\mathrm{SQ} - \Sigma \right) \mathbf{P}^{n+1} \right| \geq 0$, yielding the (conditional) stability of the scheme. □

The actual impact of this conditional stability on numerical results will be discussed in section 6.

Remark. In [27] the use of inexact algebraic factorizations (Yosida and ACT without pressure corrections) as preconditioners for the Navier–Stokes problem has been extensively investigated. Following the same idea, the inexact factorizations with pressure corrections can be used in the same fashion. The outcome is a fast preconditioner which seems to be an effective generalization of the well-known Cahouet–Chabard preconditioner for the Stokes problem. Preliminary numerical results about this preconditioner can be found in [9].

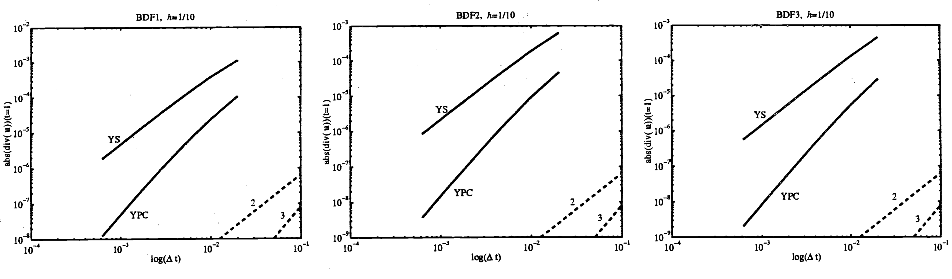


FIG. 6.1. *Divergence of the computed velocity field at $T_f = 1$ for different values of Δt. Comparison between the YS and YPC schemes with a BDF1 (left), a BDF2 (center), and a BDF3 (right) time discretization scheme.*

6. Numerical results. In this section we present some numerical results[1] that confirm the analysis carried out above and give a deeper insight into the real accuracy and stability properties of the methods presented. In particular, we consider BDF schemes of order 1, 2, and 3, both in the *nonincremental* and *incremental* formulations. In particular, in the incremental case, we refer to the approach proposed in [15], with a pressure increment as in (2.9) for a BDF of order 2 and as in (2.10) for a BDF of order 3 (see the remark above (2.8)).

We refer to the two-dimensional nonlinear Navier–Stokes problem on the unit domain $(0,1)^2$ in the time interval $(0,1)$ where (time-dependent) Dirichlet boundary conditions for the velocity, initial conditions, and the forcing term are prescribed in such a way that the analytical solution is (see [11])

$$\mathbf{u}(x,y,t) = (\sin(x)\sin(y+t), \cos(x)\cos(y+t))^T, \ p(x,y,t) = \cos(x)\sin(y+t).$$

Similar results have been obtained also for other test cases, such as the Kim and Moin (see [17]) and the Timmermans (see [26]) cases.

For what concerns the space-discretization, we have adopted an *inf-sup* compatible couple of finite element spaces. In particular, for the numerical results of the present section, we resorted to a piecewise linear functions space $\mathbf{P}^1$ for the pressure fields, and we have used $\tilde{\mathbf{P}}^2 = \mathbf{P}^2 \oplus b$ finite elements for each component of the velocity, where b is a cubic bubble function. Following [4], the role of the bubble function is to give nonsingular (velocity) mass lumped matrices, which is useful in solving systems for matrix S.

Mass conservation and pressure errors (Yosida and YPC schemes). We start focusing our attention on the Yosida (denoted by YS) and YPC schemes. In particular, we consider the divergence of the velocity field computed by the two schemes at the final time ($T_f = 1$). From the analysis of section 5.1, the first effect of the pressure correction is to modify the dependence on Δt of the residual of the mass equation, which changes from $\mathcal{O}(\Delta t^2)$ to $\mathcal{O}(\Delta t^3)$, independently of the time discretization adopted. This is clearly confirmed by Figure 6.1, where for the three different orders of BDF schemes the divergence of the velocity field is computed for several time step sizes. The effect of the pressure correction is evident. The circumstance that the divergence of the computed velocity (independently of the time discretization scheme) is $\mathcal{O}(\Delta t^3)$

[1] Numerical results of the present section have been obtained with a MATLAB code developed by the authors.

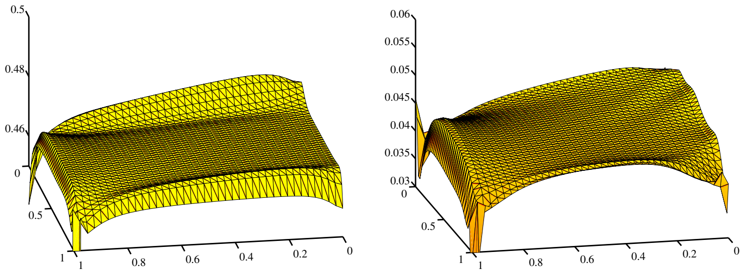


Fig. 6.2. *Space distribution of the pressure error for the YS (left) and the YPC (right) schemes.*

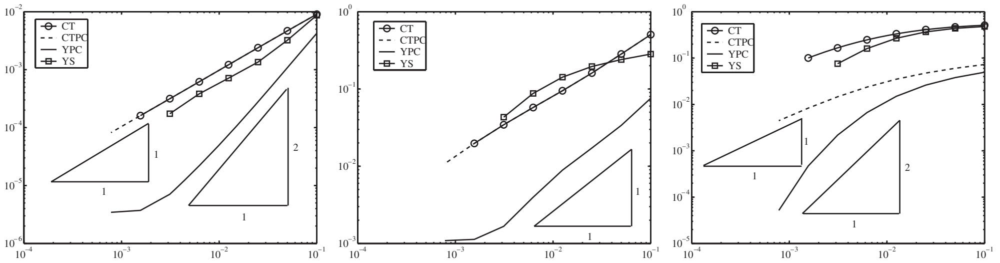


Fig. 6.3. *Errors in the* $L^\infty(\mathbf{L}^2)$*(left) and* $L^2(\mathbf{H}^1)$*(center) norms for the velocity and in the* $L^2(L^2)$*norm for the pressure (right) in the case of* BDF2 nonincremental *time discretization. Note that in all the figures presented in this section the error reduction with the time step stops whenever the error is completely due to the space discretization.*

is shared also by other schemes proposed in the differential splitting framework, which are an evolution of the Timmermans scheme (see [13] and also [12]).

Another way of investigating the effect of the pressure correction is to check the space distribution of the pressure error (see Figure 6.2). From this picture we infer that in the algebraic approach we actually do not have significant boundary layers for the pressure error (which is the case of "standard" differential splitting techniques), and this is particularly true for the pressure corrected scheme, whose associated error is significantly smaller than in the uncorrected case.

*Accuracy tests (BDF*2 *and BDF*3*).* We compare the numerical results obtained for $h = 1/40$ for different sizes of the time step. We consider, in particular, BDF time discretization schemes of order 2 and 3. The errors have been computed with respect to the norms $L^\infty(0,T,L^2(\Omega)\times L^2(\Omega)$ and $L^2(0,T,H^1(\Omega)\times H^1(\Omega))$ for the velocity and $L^2(0,T,L^2(\Omega))$ for the pressure. (In what follows, these norms will be denoted $L^\infty(\mathbf{L}^2)$, $L^2(\mathbf{H}^1)$, and $L^2(L^2)$, respectively.)

In Figure 6.3 we illustrate the results for the BDF2 *nonincremental* schemes. Results suggest that the pressure correction has a relevant effect for the YS, both for the velocity and the pressure. For the (algebraic) Chorin–Temam scheme (denoted by CT in the figures), the pressure correction gives a significant improvement only on the pressure. In particular, in the $L^\infty(\mathbf{L}^2)$norm, YPC exhibits a second order of accuracy which is not shared by the other schemes (in particular CTPC). In the $L^2(\mathbf{H}^1)$norm all the schemes are first order accurate, even if YPC features an error significantly lower than the others. For the pressure, results suggest that YPC is asymptotically second order accurate, while CTPC seems to be first order, even if there is an evident

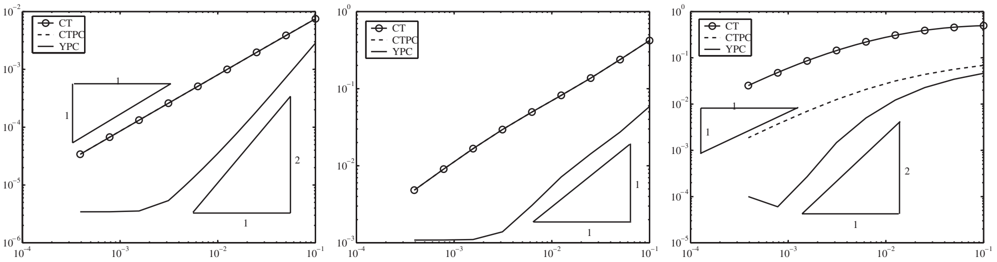


FIG. 6.4. *Errors in the* $L^\infty(\mathbf{L}^2)$*(left) and* $L^2(\mathbf{H}^1)$*(center) norms for the velocity and in the* $L^2(L^2)$*(right) norm for the pressure in the case of* BDF3 nonincremental *time discretization.*

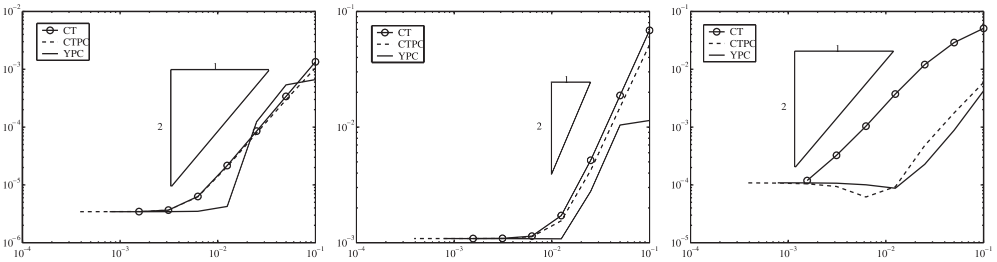


FIG. 6.5. *Errors in the* $L^\infty(\mathbf{L}^2)$*(left) and* $L^2(\mathbf{H}^1)$*(center) norms for the velocity and in the* $L^2(L^2)$*norm for the pressure (right) in the case of* BDF2 incremental *time discretization.*

error reduction with respect to the uncorrected CT scheme.

In Figure 6.4 numerical results for the BDF3 *nonincremental* schemes are reported. For what concerns the convergence order, we observe that it is substantially unchanged with respect to the case of a BDF2 time discretization. In particular, CTPC is first order accurate with respect to all the monitored norms (even if the pressure is by far more accurate with respect to the CT scheme), while YPC is second order accurate in the $L^\infty(\mathbf{L}^2)$(velocity) and $L^2(L^2)$(pressure) norms. It is first order accurate in the $L^2(\mathbf{H}^1)$norm of the velocity error. This means that the pressure correction by itself yields a splitting error $\mathcal{O}(\Delta t^2)$ which therefore does not affect the accuracy of a BDF2 time discretization, while it reduces the accuracy of the BDF3 one. It is, however, worthwhile to point out that the errors in the BDF3 case are slightly lower than the corresponding ones of the BDF2 case.

Now, let us consider the *incremental* case (see the remark above (2.8)). In Figure 6.5 we present the results of an BDF2 incremental scheme with a first order pressure extrapolation. Numerical results suggest that the method is second order accurate with respect to all the norms considered here. Actually, the improvements given by the pressure correction are minimal on the velocity for the CTPC scheme and more significant, if Δt is sufficiently small, for the YPC scheme. Pressure correction yields a relevant improvement of the solution on the pressure solution for both CTPC and YPC. For large values of Δt, YPC exhibits some strange behavior which is probably due to the poor stability properties of the method.

Specific considerations have to be deserved to the case of BDF3 incremental version. From the numerical results presented in Figure 6.6 we observe the following:

1. The uncorrected YS method in fact is second order accurate for the velocity and third order accurate for the pressure.

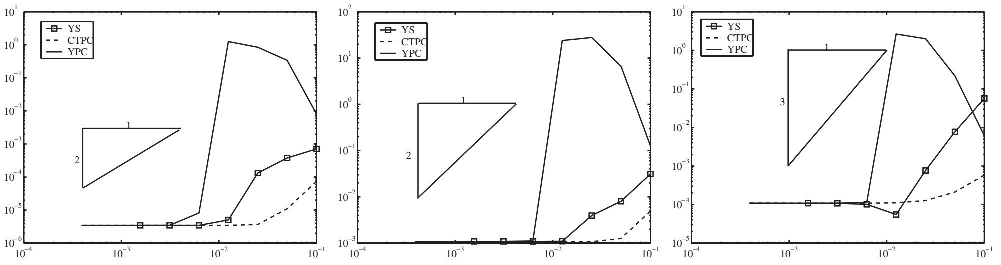


FIG. 6.6. *Errors in the* $L^\infty(\mathbf{L}^2)$*(left) and* $L^2(\mathbf{H}^1)$*(center) norms for the velocity and in the* $L^2(L^2)$*norm for the pressure (right) in the case of* BDF3 incremental *time discretization.*

2. The CTPC scheme features very good results also for large values of the time step, in such a way that it is difficult to draw an order of accuracy.
3. The YPC method has a strange behavior when Δt is large, which is probably induced by numerical instability. Surprisingly enough, under a threshold on the time step (which in our simulations is about 10^{-2}) the scheme is extremely accurate, reducing immediately the error to the contribution of the space discretization solely. We guess that in this case YPC is affected by some instabilities possibly due to the combination of the conditional stability of the BDF3 and of the YPC scheme. Further numerical investigations will be carried out for a deeper analysis of this circumstance.

Remark. The pressure correction yields improvements in the accuracy of the solution in particular on the pressure field. However, the computational cost is clearly increasing, since we need to solve two linear systems for S, rather than one. This increment can be reduced (at least in two-dimensional problems) by resorting to a direct method of solution. Since H is a s.p.d. matrix, we exploit the QR factorization $\mathrm{H}^{1/2}\mathrm{D}^T = \mathrm{QR}$, where Q is an orthogonal square ($N_{\mathbf{u}} \times N_{\mathbf{u}}$) matrix and, if the inf-sup condition holds, R is a triangular full-rank $N_{\mathbf{u}} \times N_p$ matrix such that $\mathrm{R} = \left[\mathrm{R}_0^T \; \mathbf{0}\right]^T$, where R_0 is a nonsingular $N_p \times N_p$ triangular matrix. In this way, $\mathrm{S} = \mathrm{R}_0^T\mathrm{R}_0$, yielding the Cholesky factorization of S. In this way, the solution for the system in S reduces to the solution of two triangular systems, whose computational cost is significantly lower. (For more details, see [27] and [19].) Another possibility for three-dimensional computations (see [2], [8]) is based on the idea of extracting at a given time step relevant information on the solution from the previous systems solved that share the same matrix S.

7. Conclusions and future developments. In this paper we introduce a new family of methods for the Navier–Stokes equations, based on a pressure correction step. The idea of pressure correction has already been introduced in the framework of differential schemes (Timmermans scheme), but it is new in the field of algebraic splitting. We give a mathematical basis to this approach and numerically verify that it actually improves solutions, yielding a reduction of the errors or even, in some cases, an increment of the accuracy order. The latter conclusion holds true in particular for the YPC scheme. The accuracy improvement, however, seems limited to the second order, at least in the nonincremental approach.

The pressure correction with a BDF3 incremental time advancing gives interesting results in the case of the CTPC scheme. While in all the other cases this method was usually worse than the YPC, in this case it exhibits good results that need to be investigated further. YPC features very good results in the nonincremental approach. On the other hand, whenever it is coupled with the incremental approach it can be unstable. In fact, we proved that YPC is only conditional stable, but with the nonincremental formulation we actually never observed numerical instabilities in a reasonable range for the time step sizes. Actually, it seems that the incremental approach can be somehow less stable. Moreover, it is worthwhile to mention that from preliminary numerical results the stability bound on Δt in the BDF3 incremental time advancing is proportional to the inverse of the viscosity. For low values of the viscosity (which means for high Reynolds numbers) our conjecture is that the stability bound becomes less restrictive. This observation is in agreement with the circumstance that the use of YPC as a preconditioner of the Navier–Stokes solver is well suited in particular for low viscosity problems (see [9] and also [7]). A more specific stability analysis for this scheme and, in particular, the role of the incremental formulation will be, however, the subject of a future development of this work.

REFERENCES

[1] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser., Comput. Math. 15, Springer-Verlag, New York, 1991.

[2] T. F. Chan and W. L. Wan, *Analysis of projection methods for solving linear systems with multiple right-hand sides*, SIAM J. Sci. Comput., 18 (1997), pp. 1698–1721.

[3] A. Chorin, *Numerical solution of the Navier–Stokes equations*, Math. Comp., 22 (1968), pp. 745–762.

[4] G. Cohen, P. Joly, J. E. Roberts, and N. Tordjman, *Higher order triangular finite elements with mass lumping for the wave equation*, SIAM J. Numer. Anal., 38 (2001), pp. 2047–2078.

[5] M. O. Deville, P. F. Fischer, and E. H. Mund, *High Order Methods for Incompressible Fluid Flow*, Cambridge University Press, Cambridge, UK, 2002.

[6] J. Donea and A. Huerta, *Finite Element Methods for Flow Problems*, John Wiley and Sons, New York, 2003.

[7] H. C. Elman, *Preconditioning for the steady-state Navier–Stokes equations with low viscosity*, SIAM J. Sci. Comput., 20 (1999), pp. 1299–1316.

[8] P. Fischer, *Projection techniques for iterative solution of ax=b with successive right-hand sides*, Comput. Methods Appl. Mech. Eng., 163 (1998), pp. 193–204.

[9] A. Gauthier, F. Saleri, and A. Veneziani, *A fast preconditioner for the incompressible Navier-Stokes equations*, Comput. Vis. Sci., 6 (2004), pp. 105–112.

[10] J. Guermond, *Un resultat de convergence d'ordre deux en temps pour l'approximation des equations de Navier–Stokes par une technique de projection incrementale*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 169–189.

[11] J. Guermond, P. Minev, and J. Shen, *Error Analysis of Pressure-Correction Schemes for the Navier-Stokes Equations with Open Boundary Conditions*, TICAM report 03-08, Institute for Computational Engineering and Science, Austin, TX, 2003.

[12] J. Guermond and J. Shen, *A new class of truly consistent splitting schemes for incompressible flows*, J. Comput. Phys., 192 (2003), pp. 262–276.

[13] J. L. Guermond and J. Shen, *Velocity-correction projection methods for incompressible flows*, SIAM J. Numer. Anal., 41 (2003), pp. 112–134.

[14] J.-L. Guermond and L. Quartapelle, *On the approximation of the unsteady Navier–Stokes equations by finite element projection methods*, Numer. Math., 80 (1998), pp. 207–238.

[15] M. Henriksen and J. Holmen, *Algebraic splitting for incompressible Navier–Stokes equations*, J. Comput. Phys., 175 (2002), pp. 438–453.

[16] J. van Kan, *A second-order accurate pressure-correction scheme for viscous incompressible flow*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 870–891.

[17] J. Kim and P. Moin, *Application of a fractional step method to incompressible Navier–Stokes equations*, J. Comput. Phys., 59 (1985), pp. 308–323.

[18] G. M. Kobelkow and M. Olshanskii, *Effective preconditioning of Uzawa type schemes for a generalized Stokes problem*, Numer. Math., 86 (2000), pp. 443–470.

[19] P. Matstoms, *Sparse QR Factorization with Application to Linear Least Squares Problems*, Ph.D. thesis, Linköping University, Sweden, 1994.

[20] B. Perot, *An analysis of the fractional step method*, J. Comput. Phys., 108 (1993), pp. 51–58.

[21] A. Prohl, *Projection and Quasi-Compressibility Methods for Solving the Incompressible Navier–Stokes Equations*, Springer-Verlag, 1997.

[22] A. Quarteroni, F. Saleri, and A. Veneziani, *Analysis of the Yosida method for the incompressible Navier-Stokes equations*, J. Math. Pures Appl. (9), 78 (1999), pp. 473–503.

[23] A. Quarteroni, F. Saleri, and A. Veneziani, *Factorization methods for the numerical approximation of Navier–Stokes equations*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 505–526.

[24] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.

[25] J. Shen, *On error estimates of the projection methods for Navier–Stokes equations: Second order schemes*, Math. Comp., 65 (1996), pp. 1039–1065.

[26] L. Timmermans, P. Minev, and F. V. de Vosse, *An approximate projection scheme for incompressible flow using spectral methods*, Internat. J. Numer. Methods Fluids, 22 (1996), pp. 673–688.

[27] A. Veneziani, *Block factorized preconditioners for high-order accurate in time approximation of the Navier–Stokes equations*, Num. Methods Partial Differential Equations, 19 (2003), pp. 487–510.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 195–219

SYMMETRIC AND NONSYMMETRIC DISCONTINUOUS GALERKIN METHODS FOR REACTIVE TRANSPORT IN POROUS MEDIA[*]

SHUYU SUN[†] AND MARY F. WHEELER[‡]

Abstract. For solving reactive transport problems in porous media, we analyze three primal discontinuous Galerkin (DG) methods with penalty, namely, symmetric interior penalty Galerkin (SIPG), nonsymmetric interior penalty Galerkin (NIPG), and incomplete interior penalty Galerkin (IIPG). A cut-off operator is introduced in DG to treat general kinetic chemistry. Error estimates in $L^2(H^1)$ are established, which are optimal in h and nearly optimal in p. We develop a parabolic lift technique for SIPG, which leads to h-optimal and nearly p-optimal error estimates in the $L^2(L^2)$ and negative norms. Numerical results validate these estimates. We also discuss implementation issues including penalty parameters and the choice of physical versus reference polynomial spaces.

Key words. error estimates, discontinuous Galerkin methods, reactive transport, porous media, parabolic partial differential equations, SIPG, NIPG, IIPG

AMS subject classifications. 65M12, 65M15, 65M60, 35K57

DOI. 10.1137/S003614290241708X

1. Introduction.

Discontinuous Galerkin (DG) methods employ discontinuous piecewise polynomials to approximate the solutions of differential equations, with boundary conditions and interelement continuity weakly imposed through bilinear forms. Even though they often have larger numbers of degrees of freedom than conforming approaches, DG methods have recently gained popularity for a number of attractive features [19, 3, 4, 23, 27, 11, 25, 26, 9, 18, 15]: (1) they are element-wise conservative; (2) they support general nonconforming spaces including unstructured meshes, nonmatching grids and variable degrees of local approximations, thus allowing efficient h-, p-, and hp-adaptivities; (3) they tend to have localized errors, allowing sharp *a posteriori* error indicators and effective adaptivities; (4) they have less numerical diffusion; (5) they treat rough coefficient problems and effectively capture discontinuities in solutions; (6) they are robust and nonoscillatory in the presence of high gradients; (7) with appropriate meshing, they are capable of delivering exponential rates of convergences; (8) they have excellent parallel efficiency since data communications are relatively local; (9) for time-dependent problems in particular, their mass matrices are block diagonal, providing substantial computational advantages if explicit time integrations are used. In addition, by a simple extension from the average of the fluxes on element faces, DG can provide a continuous flux field defined over the entire domain, allowing efficient coupling with conforming methods.

Numerical modeling of reactive transport in porous media has important applications in hydrology, earth sciences, environmental protection, oil recovery, chemical

[*] Received by the editors November 2, 2002; accepted for publication (in revised form) February 25, 2005; published electronically May 27, 2005. This research was partially supported by National Science Foundation grant DMS-0411413.

http://www.siam.org/journals/sinum/43-1/41708.html

[†] The Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, 201 E. 24th St. ACE 5.316, Austin, TX 78712 (shuyu@ices.utexas.edu).

[‡] ICES, Department of Aerospace Engineering and Engineering Mechanics, Department of Petroleum and Geosystems Engineering, and Department of Mathematics, The University of Texas at Austin, 201 E. 24th St. ACE 5.324, Austin, TX 78712 (mfw@ices.utexas.edu).

industry, and biomedical engineering. Realistic simulations for simultaneous advection, diffusion, and chemical reactions present significant computational challenges [2, 40, 10, 14, 24, 37, 41, 28, 7, 16, 8]. Recently, it has been shown that adaptive DG can effectively capture moving concentration fronts in reactive transport [31, 33, 36, 32, 29]. A posteriori error estimates of DG for reactive transport problems have been derived in the $L^2(L^2)$ [32] and $L^2(H^1)$ norms [35]. In addition, DG has been applied to coupled flow and transport problems in porous media [34, 39, 30]. However, to the best of our knowledge, optimal a priori *hp*-estimates in the $L^2(L^2)$ and negative norms have not been established.

The primal DG methods include four members: Oden–Babuška–Baumann DG (OBB-DG) formulation [19], symmetric interior penalty Galerkin (SIPG) [38], nonsymmetric interior penalty Galerkin (NIPG) [23, 21], and incomplete interior penalty Galerkin (IIPG) [12, 29]. In this paper, we analyze the three primal DG methods with penalty, i.e., SIPG, NIPG, and IIPG, for solving reactive transport problems in porous media. The primal DG method without penalty, i.e., the OBB-DG scheme, has been analyzed for reactive transport problems elsewhere [22]. In the following section, we describe the modeling equations. The DG schemes are introduced in section 3. Section 4 contains the $L^2(H^1)$ error analysis for SIPG, NIPG, and IIPG. In section 5, a parabolic lift technique is developed, and an $L^2(L^2)$ error analysis for SIPG is conducted. Optimal negative norm estimates are derived in section 6. In section 7, we present numerical studies of *h*- and *p*-convergences for the four primal DG schemes. In section 8, we discuss choices of penalty parameters as well as DG implementations using reference versus physical polynomial spaces. Conclusions are given in the last section.

2. Governing equations. For convenience of presentation, we consider reactive transport problems of only one species in a single flowing phase in porous media. Results for systems of multiple species with kinetic reactions can be derived by similar arguments. We assume that a Darcy velocity field $\mathbf{u}$ is given and time-independent, and satisfies $\nabla \cdot \mathbf{u} = q$, where q is the imposed external total flow rate. In addition, we assume that Ω is a polygonal and bounded domain in $\mathbb{R}^d$ ($d = 1$, 2, or 3) with boundary $\partial\Omega = \overline{\Gamma}_{\text{in}} \cup \overline{\Gamma}_{\text{out}}$. Here we denote by Γ_{in} the inflow boundary and by Γ_{out} the outflow/no-flow boundary, i.e.,

$$\Gamma_{\text{in}} := \{x \in \partial\Omega : \ \mathbf{u} \cdot \mathbf{n} < 0\},$$
$$\Gamma_{\text{out}} := \{x \in \partial\Omega : \ \mathbf{u} \cdot \mathbf{n} \geq 0\},$$

where $\mathbf{n}$ denotes the unit outward normal vector to $\partial\Omega$. Let T be the final simulation time. The classical advection-diffusion-reaction equation in porous media is given by

$$(2.1) \qquad \frac{\partial \phi c}{\partial t} + \nabla \cdot (\mathbf{u}c - \mathbf{D}(\mathbf{u})\nabla c) = qc^* + r(c), \quad (x,t) \in \Omega \times (0,T],$$

where the unknown variable c is the concentration of a species (amount per volume). Here ϕ is the effective porosity and is assumed to be time-independent, uniformly bounded above and below by positive numbers; $\mathbf{D}(\mathbf{u})$ is the dispersion-diffusion tensor and is assumed to be uniformly symmetric positive definite and bounded from above; $r(c)$ is the reaction term; qc^* is the source term, where the imposed external total flow rate q is a sum of sources (injection) and sinks (extraction); c^* is the injected concentration c_w if $q \geq 0$ and is the resident concentration c if $q < 0$.

We consider the following boundary conditions for this problem:

$$(\mathbf{u}c - \mathbf{D}(\mathbf{u})\nabla c)\cdot \mathbf{n} = c_B \mathbf{u}\cdot\mathbf{n}, \qquad (x,t) \in \Gamma_{\text{in}} \times (0,T], \tag{2.2}$$

$$(-\mathbf{D}(\mathbf{u})\nabla c)\cdot \mathbf{n} = 0, \qquad (x,t) \in \Gamma_{\text{out}} \times (0,T], \tag{2.3}$$

where c_B is the inflow concentration. The initial concentration is specified by

$$c(x,0) = c_0(x), \qquad x \in \Omega. \tag{2.4}$$

3. Discontinuous Galerkin schemes.

3.1. Notation. Let $\mathcal{E}_h$ be a family of nondegenerate, quasi-uniform and possibly nonconforming partitions of Ω composed of triangles or quadrilaterals if $d = 2$, or tetrahedra, prisms, or hexahedra if $d = 3$. The nondegeneracy requirement (also called regularity) is that the element is convex, and that there exists $\rho > 0$ such that if h_j is the diameter of $E_j \in \mathcal{E}_h$, then each of the subtriangles (for $d = 2$) or subtetrahedra (for $d = 3$) of element E_j contains a ball of radius ρh_j in its interior. The quasi-uniformity requirement is that there is $\tau > 0$ such that $(h/h_j) \le \tau$ for all $E_j \in \mathcal{E}_h$, where h is the maximum diameter of all elements. We assume that no element crosses the boundaries of Γ_{in} or Γ_{out}. The set of all interior edges (for $d = 2$) or faces (for $d = 3$) for $\mathcal{E}_h$ is denoted by Γ_h. On each edge or face $\gamma \in \Gamma_h$, a unit normal vector $\mathbf{n}_\gamma$ is chosen. The sets of all edges or faces on Γ_{out} and on Γ_{in} for $\mathcal{E}_h$ are denoted by $\Gamma_{h,\text{out}}$ and $\Gamma_{h,\text{in}}$, respectively, for which the normal vector $\mathbf{n}_\gamma$ coincides with the outward unit normal vector.

We now define the average and jump for $\phi \in H^s(\mathcal{E}_h)$, $s > 1/2$. Let $E_i, E_j \in \mathcal{E}_h$ and $\gamma = \partial E_i \cap \partial E_j \in \Gamma_h$ with $\mathbf{n}_\gamma$ exterior to E_i. We denote

$$\{\phi\} := \frac{1}{2}((\phi|_{E_i})|_\gamma + (\phi|_{E_j})|_\gamma), \qquad [\phi] := (\phi|_{E_i})|_\gamma - (\phi|_{E_j})|_\gamma.$$

The upwind value of a concentration $c^*|_\gamma$ is defined as

$$c^*|_\gamma := \begin{cases} c|_{E_i} & \text{if } \mathbf{u}\cdot\mathbf{n}_\gamma \ge 0, \\ c|_{E_j} & \text{if } \mathbf{u}\cdot\mathbf{n}_\gamma < 0. \end{cases}$$

We denote by $\|\cdot\|_{m,R}$ the usual Sobolev norm over a domain R [1]. The Sobolev norm $\|\cdot\|_{m,\Omega}$ over the entire domain Ω is also denoted simply by $\|\cdot\|_m$. For $s \ge 0$, we define the broken Sobolev space

$$H^s(\mathcal{E}_h) := \{\phi \in L^2(\Omega) : \ \phi|_E \in H^s(E), \ E \in \mathcal{E}_h\}.$$

One can show that $H^s(\mathcal{E}_h)$ is a normed linear space with its norm defined by

$$\|\phi\|_{H^s(\mathcal{E}_h)} := \left(\sum_{E\in\mathcal{E}_h} \|\phi\|^2_{s,E} \right)^{1/2}.$$

Following the tradition, we also use the notation $|\!|\!| \cdot |\!|\!|_s$ to denote the broken norm $\|\cdot\|_{H^s(\mathcal{E}_h)}$. For a given normed space X and a number $p \ge 1$, we define

$$L^p(0,T;X) := \{\phi : \ \phi(t) \in X, \|\phi\|_X \in L^p(0,T)\}.$$

The space $L^p(0,T;X)$ is also a normed linear space with its norm given by

$$\|\phi\|_{L^p(0,T;X)} := \|(\|\phi\|_X)\|_{L^p(0,T)}.$$

The broken norm $\|\cdot\|_{L^p(0,T;H^s(\mathcal{E}_h))}$ is also written as $|\!|\!|\cdot|\!|\!|_{L^p(0,T;H^s)}$ in the triple bar notation. We denote by $(\cdot,\cdot)_R$ the inner product in $(L^2(R))^d$ or $L^2(R)$ over a domain R. The inner product $(\cdot,\cdot)_\Omega$ over the entire domain Ω is also denoted simply by $(\cdot,\cdot)$. We also need the space $W_\infty^{r,s}$ and its norm:

$$W_\infty^{r,s}((0,T)\times\Omega) := \{f \in L^2((0,T)\times\Omega) : \ \|f\|_{W_\infty^{r,s}} < \infty\},$$
$$\|f\|_{W_\infty^{r,s}} := \sum_{|\alpha|\le r,\, \beta\le s} \operatorname{ess\,sup}_{(0,T)\times\Omega}(|D_x^\alpha f| + |D_t^\beta f|).$$

The discontinuous finite element space is taken to be

$$\mathcal{D}_r(\mathcal{E}_h) := \{\phi \in L^2(\Omega) : \ \phi|_E \in \mathbb{P}_r(E), \ E \in \mathcal{E}_h\}, \tag{3.1}$$

where $\mathbb{P}_r(E)$ denotes the space of polynomials of (total) degree less than or equal to r on E. Note that we present *hp*-results in this paper for the local space $\mathbb{P}_r$, but the results also apply to the local space $\mathbb{Q}_r$ because $\mathbb{P}_r(E) \subset \mathbb{Q}_r(E)$.

We define a cut-off operator as

$$\mathcal{M}(c)(x) := \min(c(x), M), \tag{3.2}$$

where M is a large positive constant. By a straightforward algebraic argument, we can show that the cut-off operator $\mathcal{M}$ is uniformly Lipschitz continuous.

Lemma 1 (property of operator $\mathcal{M}$). *The cut-off operator $\mathcal{M}$ defined in* (3.2) *is uniformly Lipschitz continuous with a Lipschitz constant of one; that is,*

$$\|\mathcal{M}(c) - \mathcal{M}(w)\|_{L^\infty(\Omega)} \le \|c - w\|_{L^\infty(\Omega)}. \tag{3.3}$$

We use the following *hp*-approximation results, which can be proved using the techniques in [6, 5]. Let $E \in \mathcal{E}_h$ and $\phi \in H^s(E)$. Then there exists a constant K, independent of ϕ, r, and h_E, and a sequence of $z_r^h \in \mathbb{P}_r(E)$, $r = 1, 2, \ldots$, such that

$$\begin{cases} \left\|\phi - z_r^h\right\|_{q,E} \le K \dfrac{h_E^{\mu-q}}{r^{s-q}} \|\phi\|_{s,E}, & 0 \le q < \mu, \\ \left\|\phi - z_r^h\right\|_{q,\partial E} \le K \dfrac{h_E^{\mu-q-\frac{1}{2}}}{r^{s-q-\frac{1}{2}}} \|\phi\|_{s,E}, & 0 \le q < \mu - \frac{1}{2}, \end{cases} \tag{3.4}$$

where $\mu = \min(r+1, s)$ and h_E denotes the diameter of E.

We shall also use the following inverse inequalities, which can be derived using the method in [27]. Let $E \in \mathcal{E}_h$ and $v \in \mathbb{P}_r(E)$. Then there exists a constant K, independent of v, r, and h_E, such that

$$\begin{cases} \|D^q v\|_{0,\partial E} \le K \dfrac{r}{h_E^{1/2}} \|D^q v\|_E, & q \ge 0, \\ \|D^{q+1} v\|_{0,E} \le K \dfrac{r^2}{h_E} \|D^q v\|_{0,E}, & q \ge 0. \end{cases} \tag{3.5}$$

3.2. Continuous-in-time DG schemes. We introduce a bilinear form:

$$\begin{aligned} B(c, w; \mathbf{u}) := {} & \sum_{E\in\mathcal{E}_h} \int_E (\mathbf{D}(\mathbf{u})\nabla c - c\mathbf{u}) \cdot \nabla w - \int_\Omega c q^- w \\ & - \sum_{\gamma\in\Gamma_h} \int_\gamma \{\mathbf{D}(\mathbf{u})\nabla c \cdot \mathbf{n}_\gamma\}[w] - s_{\text{form}} \sum_{\gamma\in\Gamma_h} \int_\gamma \{\mathbf{D}(\mathbf{u})\nabla w \cdot \mathbf{n}_\gamma\}[c] \\ & + \sum_{\gamma\in\Gamma_h} \int_\gamma c^* \mathbf{u} \cdot \mathbf{n}_\gamma [w] + \sum_{\gamma\in\Gamma_{h,\text{out}}} \int_\gamma c\mathbf{u} \cdot \mathbf{n}_\gamma w + J_0^\sigma(c, w). \end{aligned}$$

Here $s_{\text{form}} = 1$ for SIPG; $s_{\text{form}} = -1$ for OBB-DG or NIPG; and $s_{\text{form}} = 0$ for IIPG. For convenience of presentation, we denote the bilinear form as $B_S(c, w; \mathbf{u})$ when it is symmetric, i.e., $s_{\text{form}} = 1$. We denote by q^+ the injection source term and by q^- the extraction source term, i.e., $q^+ = \max(q, 0)$ and $q^- = \min(q, 0)$. By definition, we have $q = q^+ + q^-$. To impose interelement continuity weakly, an interior penalty term $J_0^\sigma(c, w)$ is formulated:

$$J_0^\sigma(c, w) := \sum_{\gamma \in \Gamma_h} \frac{r^2 \sigma_\gamma}{h_\gamma} \int_\gamma [c][w], \tag{3.6}$$

where σ is a discrete positive function that takes the constant value σ_γ on the edge or face γ. There is no penalty term, i.e., $\sigma = 0$, for OBB-DG. In the analysis of SIPG, NIPG, and IIPG in this paper, we assume $0 < \sigma_0 \leq \sigma_\gamma \leq \sigma_m$. In addition we define a linear functional:

$$L(w; \mathbf{u}, c) := \int_\Omega r(\mathcal{M}(c)) w + \int_\Omega c_w q^+ w - \sum_{\gamma \in \Gamma_{h,\text{in}}} \int_\gamma c_B \mathbf{u} \cdot \mathbf{n}_\gamma w. \tag{3.7}$$

The reactive transport problem can be stated in the following equivalent weak formulation.

Lemma 2 (weak formulation). *If c is a solution of* (2.1)–(2.3) *and c is essentially bounded, then c satisfies*

$$\left(\frac{\partial \phi c}{\partial t}, w \right) + B(c, w; \mathbf{u}) = L(w; \mathbf{u}, c) \qquad \forall w \in H^s(\mathcal{E}_h),\ s > \frac{3}{2} \quad \forall t \in (0, T], \tag{3.8}$$

provided that the constant M for the cut-off operator is sufficiently large.

Proof. Let $w \in H^s(\mathcal{E}_h)$, $s > 3/2$ and $E \in \mathcal{E}_h$. Multiplying (2.1) by w, integrating over E, and then integrating by parts, we observe

$$\left(\frac{\partial \phi c}{\partial t}, w \right)_E - \int_E (\mathbf{u}c - \mathbf{D}(\mathbf{u}) \nabla c) \cdot \nabla w + \int_{\partial E} (\mathbf{u}c - \mathbf{D}(\mathbf{u}) \nabla c) \cdot \mathbf{n}_{\partial E} w = \int_E q c^* w + r(c) w.$$

Summing it over all elements in $\mathcal{E}_h$, noting the fact that the traces of the concentration and its normal flux are continuous across element faces, and applying the boundary conditions, we obtain the desired result. □

The continuous-in-time DG approximation $C^{DG}(\cdot, t) \in \mathcal{D}_r(\mathcal{E}_h)$ to the solution of (2.1)–(2.4) is defined by

$$\left(\frac{\partial \phi C^{DG}}{\partial t}, w \right) + B(C^{DG}, w; \mathbf{u}) = L(w; \mathbf{u}, C^{DG}) \qquad \forall w \in \mathcal{D}_r(\mathcal{E}_h) \quad \forall t \in (0, T], \tag{3.9}$$

$$(\phi C^{DG}, w) = (\phi c_0, w) \qquad \forall w \in \mathcal{D}_r(\mathcal{E}_h), \quad t = 0. \tag{3.10}$$

As a valuable property, DG schemes possess element-wise mass conservation. OBB-DG satisfies local conservation strictly, whereas SIPG, NIPG, and IIPG are locally conservative if the concentration jump term is considered as part of the computed diffusive flux:

Lemma 3 (local mass balance). *The approximation of the concentration satisfies on each element E the following local mass balance equation:*

$$\begin{aligned}(3.11)\qquad &\int_E \frac{\partial \phi C^{DG}}{\partial t} - \int_{\partial E\setminus\partial\Omega} \{\mathbf{D}(\mathbf{u})\nabla C^{DG}\cdot \mathbf{n}_{\partial E}\} + \int_{\partial E} C^{DG*}\mathbf{u}\cdot\mathbf{n}_{\partial E}\\ &+ \sum_{\gamma\subset\partial E\setminus\partial\Omega} \frac{r^2\sigma_\gamma}{h_\gamma}\int_\gamma (C^{DG}|_E - C^{DG}|_{\Omega\setminus\overline{E}})\\ &= \int_E C^{DG*}q + \int_E r(\mathcal{M}(C^{DG})).\end{aligned}$$

Proof. The relationship (3.11) follows immediately from the DG schemes by fixing an element E and letting $w \in \mathcal{D}_r(\mathcal{E}_h)$ with $w|_E = 1$, $w|_{\Omega\setminus E} = 0$. □

It is also important to know that a DG scheme has a solution.

Lemma 4 (existence of a solution). *Assume that the reaction rate is a locally Lipschitz continuous function of the concentration. Then the discontinuous Galerkin scheme* (3.9) *and* (3.10) *has a unique solution for $t > 0$.*

Proof. We let $\{v_i\}_{i=1}^M$ be a basis of $\mathcal{D}_r(\mathcal{E}_h)$ and write $C^{DG} = \sum_{i=1}^M \zeta_i(t)v_i(x)$. Then (3.9) and (3.10) reduce to the following initial value problem:

$$\begin{cases} A\dfrac{d\zeta}{dt} = -B\zeta + R(\zeta),\\ A\zeta(0) = b,\end{cases}$$

where the mass matrix A is block-diagonal, symmetric, and positive definite. From the properties of the cut-off operator $\mathcal{M}$ and the reaction function, we observe that $R(\zeta)$ is (globally) Lipschitz continuous. It follows from the theory of ordinary differential equations that $\zeta(t)$ exists and is unique for $t > 0$. □

4. $L^2(H^1)$ and $L^\infty(L^2)$ error estimates.

Throughout the paper, we denote by K a generic positive constant independent of h and r, and by ϵ a fixed positive constant that may be chosen arbitrarily small.

Theorem 1 ($L^2(H^1)$ and $L^\infty(L^2)$ error estimates). *Let c be the solution to* (2.1)–(2.4), *and assume $c \in L^2(0,T;H^s(\mathcal{E}_h))$, $\partial c/\partial t \in L^2(0,T;H^{s-1}(\mathcal{E}_h))$, and $c_0 \in H^{s-1}(\mathcal{E}_h)$. We further assume that c, $\mathbf{u}$ and q are essentially bounded, that the reaction rate is a locally Lipschitz continuous function of c, and that the cut-off constant M and the penalty parameter σ_0 are sufficiently large. Then there exists a constant K, independent of h and r, such that*

$$\begin{aligned}&\|C^{DG} - c\|_{L^\infty(0,T;L^2)} + \|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla(C^{DG} - c)\|_{L^2(0,T;L^2)}\\ &+\left(\int_0^T J_0^\sigma(C^{DG}-c, C^{DG}-c)\right)^{\frac{1}{2}}\\ &\le K\frac{h^{\mu-1}}{r^{s-1-\delta}}\|c\|_{L^2(0,T;H^s)} + K\frac{h^{\mu-1}}{r^{s-1}}(\|\partial c/\partial t\|_{L^2(0,T;H^{s-1})} + \|c_0\|_{s-1}),\end{aligned}$$

where $\mu = \min(r+1, s)$, $r \ge 1$, $s \ge 2$, $\delta = 0$ for conforming meshes with triangles or tetrahedra, and $\delta = 1/2$ in general.

Proof. We let $\widehat{c} \in \mathcal{D}_r(\mathcal{E}_h)$ be an interpolant of concentration c such that the *hp*-results (3.4) hold, and define

$$(4.1)\qquad \xi = C^{DG} - c,$$

$$\xi^I = c - \widehat{c}, \tag{4.2}$$

$$\xi^A = C^{DG} - \widehat{c} = \xi + \xi^I. \tag{4.3}$$

Subtracting the weak formulation (3.8) from the DG scheme (3.9), choosing $w = \xi^A$, we obtain

$$\begin{aligned}(4.4)\qquad &\left(\frac{\partial \phi \xi^A}{\partial t}, \xi^A\right) + B(\xi^A, \xi^A; \mathbf{u}) \\ &\quad = L(\xi^A; \mathbf{u}, C^{DG}) - L(\xi^A; \mathbf{u}, c) + \left(\frac{\partial \phi \xi^I}{\partial t}, \xi^A\right) + B(\xi^I, \xi^A; \mathbf{u}).\end{aligned}$$

The first term of the error equation (4.4) may be written in a time derivative of an L^2 norm:

$$\left(\frac{\partial \phi \xi^A}{\partial t}, \xi^A\right) = \frac{1}{2}\frac{d}{dt}\left\|\sqrt{\phi}\xi^A\right\|_{0,\Omega}^2.$$

We expand the second term of (4.4) as

$$\begin{aligned} B(\xi^A, \xi^A; \mathbf{u}) = &\sum_{E\in\mathcal{E}_h}\int_E (\mathbf{D}(\mathbf{u})\nabla\xi^A - \xi^A\mathbf{u})\cdot\nabla\xi^A - \int_\Omega q^-(\xi^A)^2 \\ &-(1+s_{\text{form}})\sum_{\gamma\in\Gamma_h}\int_\gamma \{\mathbf{D}(\mathbf{u})\nabla\xi^A\cdot\mathbf{n}_\gamma\}[\xi^A] \\ &+\sum_{\gamma\in\Gamma_h}\int_\gamma \xi^{A*}\mathbf{u}\cdot\mathbf{n}_\gamma[\xi^A] + \sum_{\gamma\in\Gamma_{h,\text{out}}}\int_\gamma \mathbf{u}\cdot\mathbf{n}_\gamma(\xi^A)^2 + J_0^\sigma(\xi^A,\xi^A).\end{aligned}$$

Integrating the advection term by parts, we observe

$$\begin{aligned} &-\sum_{E\in\mathcal{E}_h}\int_E \xi^A\mathbf{u}\cdot\nabla\xi^A \\ &= -\frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_E \mathbf{u}\cdot\nabla(\xi^A)^2 = -\frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_{\partial E}\mathbf{u}\cdot\mathbf{n}_{\partial E}(\xi^A)^2 + \frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_E q(\xi^A)^2 \\ &= -\frac{1}{2}\sum_{\gamma\in\Gamma_h}\int_\gamma \mathbf{u}\cdot\mathbf{n}_\gamma[(\xi^A)^2] - \frac{1}{2}\sum_{\gamma\in\Gamma_{h,\text{in}}\cup\Gamma_{h,\text{out}}}\int_\gamma \mathbf{u}\cdot\mathbf{n}_\gamma(\xi^A)^2 + \frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_E q(\xi^A)^2.\end{aligned}$$

In addition, noting that $[c^2] = 2\{c\}[c]$ and $(c^* - \{c\})\text{sign}(\mathbf{u}\cdot\mathbf{n}) = [c]/2$, we have

$$\begin{aligned} B(\xi^A, \xi^A; \mathbf{u}) = &\|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A\|_0^2 + \frac{1}{2}\int_\Omega |q|(\xi^A)^2 - T_0 + J_0^\sigma(\xi^A,\xi^A) \\ &+\frac{1}{2}\sum_{\gamma\in\Gamma_h}\int_\gamma |\mathbf{u}\cdot\mathbf{n}_\gamma|[\xi^A]^2 + \frac{1}{2}\sum_{\gamma\in\Gamma_{h,\text{in}}\cup\Gamma_{h,\text{out}}}\int_\gamma |\mathbf{u}\cdot\mathbf{n}_\gamma|(\xi^A)^2,\end{aligned}$$

where T_0 is defined by

$$T_0 := (1+s_{\text{form}})\sum_{\gamma\in\Gamma_h}\int_\gamma \{\mathbf{D}(\mathbf{u})\nabla\xi^A\cdot\mathbf{n}_\gamma\}[\xi^A].$$

If the penalty parameter σ_0 is chosen to be sufficiently large, we may bound T_0 by applying the Cauchy–Schwarz and inverse inequalities:

$$\begin{aligned}(4.5)\qquad T_0 &\le \frac{h}{Kr^2}\sum_{E\in\mathcal{E}_h}\left\|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A\cdot\mathbf{n}_{\partial E}\right\|^2_{0,\partial E} + \frac{Kr^2}{h}\sum_{\gamma\in\Gamma_h}\|[\xi^A]\|^2_{0,\gamma}\\ &\le \frac{1}{2}|\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|\!|\!|^2_0 + \frac{1}{2}J_0^\sigma(\xi^A,\xi^A).\end{aligned}$$

The first two terms on the right-hand side of (4.4) may be estimated, by using Lemma 1, as

$$\begin{aligned}L(\xi^A;\mathbf{u},C^{DG}) - L(\xi^A;\mathbf{u},c) &= \int_\Omega (r(\mathcal{M}(C^{DG})) - r(\mathcal{M}(c)))\xi^A\\ \le K\|\sqrt{\phi}\xi^A\|^2_0 + K\|\xi^I\|^2_0 &\le K\|\sqrt{\phi}\xi^A\|^2_0 + K\frac{h^{2\mu}}{r^{2s}}|\!|\!|c|\!|\!|^2_s.\end{aligned}$$

We have a similar result for the third term:

$$\begin{aligned}&\left(\frac{\partial\phi\xi^I}{\partial t},\xi^A\right) \le K\left\|\frac{\partial\xi^I}{\partial t}\right\|_0\left\|\sqrt{\phi}\xi^A\right\|_0\\ &\le K\left\|\sqrt{\phi}\xi^A\right\|^2_0 + K\left\|\frac{\partial\xi^I}{\partial t}\right\|^2_0 \le K\left\|\sqrt{\phi}\xi^A\right\|^2_0 + K\frac{h^{2\mu-2}}{r^{2s-2}}|\!|\!|c_t|\!|\!|^2_{s-1}.\end{aligned}$$

The fourth term on the right-hand side of (4.4) consists of eight pieces:

$$\begin{aligned}&B(\xi^I,\xi^A;\mathbf{u})\\ &= \sum_{E\in\mathcal{E}_h}\int_E \mathbf{D}(\mathbf{u})\nabla\xi^I\cdot\nabla\xi^A - \sum_{E\in\mathcal{E}_h}\int_E \xi^I\mathbf{u}\cdot\nabla\xi^A - \int_\Omega q^-\xi^I\xi^A\\ &\quad - \sum_{\gamma\in\Gamma_h}\int_\gamma\{\mathbf{D}(\mathbf{u})\nabla\xi^I\cdot\mathbf{n}_\gamma\}[\xi^A] - s_{\mathrm{form}}\sum_{\gamma\in\Gamma_h}\int_\gamma\{\mathbf{D}(\mathbf{u})\nabla\xi^A\cdot\mathbf{n}_\gamma\}[\xi^I]\\ &\quad + \sum_{\gamma\in\Gamma_h}\int_\gamma \xi^{I*}\mathbf{u}\cdot\mathbf{n}_\gamma[\xi^A] + \sum_{\gamma\in\Gamma_{h,\mathrm{out}}}\int_\gamma \mathbf{u}\cdot\mathbf{n}_\gamma\xi^I\xi^A + J_0^\sigma(\xi^I,\xi^A)\\ &=: \sum_{i=1}^{8} T_i.\end{aligned}$$

The Cauchy–Schwarz inequality and approximation results yield

$$\begin{aligned}T_1 &\le \epsilon|\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|\!|\!|^2_0 + K\frac{h^{2\mu-2}}{r^{2s-2}}|\!|\!|c|\!|\!|^2_s,\\ T_2 &\le \epsilon|\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|\!|\!|^2_0 + K\frac{h^{2\mu}}{r^{2s}}|\!|\!|c|\!|\!|^2_s,\\ T_3 &\le \epsilon\int_\Omega |q^-|(\xi^A)^2 + K\frac{h^{2\mu}}{r^{2s}}|\!|\!|c|\!|\!|^2_s.\end{aligned}$$

We bound the terms T_4 and T_5 by hiding a large constant in the penalty term and by using the inverse inequality, respectively,

$$\begin{aligned}T_4 &\leq \epsilon\frac{\sigma_0 r^2}{h}\sum_{\gamma\in\Gamma_h}\|[\xi^A]\|_{0,\gamma}^2 + \frac{Kh}{r^2}\sum_{E\in\mathcal{E}_h}\|\nabla\xi^I\cdot\mathbf{n}_{\partial E}\|_{0,\partial E}^2\\ &\leq \epsilon J_0^\sigma(\xi^A,\xi^A) + K\frac{h^{2\mu-2}}{r^{2s-1}}|||c|||_s^2,\\ T_5 &\leq \frac{\epsilon h}{Kr^2}\sum_{E\in\mathcal{E}_h}\|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A\cdot\mathbf{n}_{\partial E}\|_{0,\partial E}^2 + \frac{Kr^2}{h}\sum_{E\in\mathcal{E}_h}\|\xi^I\|_{0,\partial E}^2\\ &\leq \epsilon|||\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|||_0^2 + K\frac{h^{2\mu-2}}{r^{2s-3}}|||c|||_s^2.\end{aligned}$$

Similar applications of the Cauchy–Schwarz inequality and approximation results give

$$\begin{aligned}T_6 &\leq \epsilon\sum_{\gamma\in\Gamma_h}\int_\gamma|\mathbf{u}\cdot\mathbf{n}_\gamma|[\xi^A]^2 + K\frac{h^{2\mu-1}}{r^{2s-1}}|||c|||_s^2,\\ T_7 &\leq \epsilon\sum_{\gamma\in\Gamma_{h,\mathrm{out}}}\int_\gamma|\mathbf{u}\cdot\mathbf{n}_\gamma|(\xi^A)^2 + K\frac{h^{2\mu-1}}{r^{2s-1}}|||c|||_s^2,\\ T_8 &\leq \epsilon J_0^\sigma(\xi^A,\xi^A) + K\frac{h^{2\mu-2}}{r^{2s-3}}|||c|||_s^2.\end{aligned}$$

For conforming meshes with triangles or tetrahedra, we can choose a continuous approximation $\widehat{c}$ to make the two terms T_5 and T_8 vanish. Substituting all the estimates into (4.4), we see that

$$\begin{aligned}&\frac{d}{dt}\|\sqrt{\phi}\xi^A\|_0^2 + |||\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|||_0^2 + J_0^\sigma(\xi^A,\xi^A)\\ &\quad\leq K\|\sqrt{\phi}\xi^A\|_0^2 + K\frac{h^{2\mu-2}}{r^{2s-2-2\delta}}|||c|||_s^2 + K\frac{h^{2\mu-2}}{r^{2s-2}}|||c_t|||_{s-1}^2,\end{aligned}\tag{4.6}$$

where $\delta = 0$ for conforming meshes with triangles or tetrahedra, and $\delta = 1/2$ in general. Integrating (4.6) with respect to the time t, noting that

$$\|\sqrt{\phi}E^A\|_0(0) \leq K\frac{h^{\mu-1}}{r^{s-1}}|||c_0|||_{s-1},$$

and applying Gronwall's inequality, we conclude that

$$\begin{aligned}&\|\sqrt{\phi}\xi^A\|_{L^\infty(0,T;L^2)} + |||\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|||_{L^2(0,T;L^2)} + (\int_0^T J_0^\sigma(\xi^A,\xi^A))^{\frac{1}{2}}\\ &\quad\leq K\frac{h^{\mu-1}}{r^{s-1-\delta}}|||c|||_{L^2(0,T;H^s)} + K\frac{h^{\mu-1}}{r^{s-1}}(|||\partial c/\partial t|||_{L^2(0,T;H^{s-1})} + |||c_0|||_{s-1}).\end{aligned}$$

The theorem follows by applying the triangle inequality, the approximation results and the fact that

$$|||c|||_{L^\infty(0,T;H^{s-1})} \leq K|||c_t|||_{L^2(0,T;H^{s-1})} + |||c_0|||_{s-1}. \qquad \square \tag{4.7}$$

We remark that, in [22], $L^\infty(L^2) + L^2(H^1)$ error estimates for the OBB-DG diffusion scheme applied to the transport problem established optimality in h and suboptimality in p by $3/2$. Here for SIPG, NIPG, and IIPG, we obtain optimality in h and p for conforming meshes with triangles and tetrahedra and a loss of $1/2$ in p for general grids. Obviously, penalty terms improve the provable p-optimality of DGs.

5. Optimal $\boldsymbol{L^2(L^2)}$ error estimates for the symmetric scheme. In this and following sections, we restrict our attention to SIPG. The derivation in this section is motivated by the h-optimal L^2 result for SIPG applied to an elliptic problem by Wheeler [38] and the h-optimal $L^2(L^2)$ result for continuous Galerkin methods applied to a parabolic problem by Palmer [20]. See also the h-optimal $L^2(L^2)$ result for continuous finite element modified methods of characteristics applied to a coupled system of partial differential equations (PDEs) by Dawson, Russell, and Wheeler [13] and the h-optimal $L^\infty(L^2)$ result for SIPG applied to a parabolic equation with diffusion term by Arnold [4, 3]. We first recall a theorem proved in [20, 17].

Theorem 2. *Consider the parabolic equation:*

$$\begin{aligned} \frac{\partial \phi \Phi}{\partial t} + \nabla \cdot (\mathbf{u}\Phi - \mathbf{D}\nabla\Phi) + a\Phi = f, &\qquad x \in \Omega,\ t \in (0,T], \\ \mathbf{D}\nabla\Phi \cdot \mathbf{n} = 0, &\qquad x \in \partial\Omega,\ t \in (0,T], \\ \Phi = 0, &\qquad x \in \Omega,\ t = 0. \end{aligned}$$

Assume that $0 < \phi_0 \le \phi(t,x) \le \phi_m$, $\mathbf{D}$ is uniformly symmetric positive definite and bounded from above, $\phi \in W^{2;1}_\infty((0,T)\times\Omega)$, $\mathbf{D}_{ij} \in W^{1;0}_\infty((0,T)\times\Omega)$, $\mathbf{u}_i \in L^\infty(\Omega)$ ($\mathbf{u}$ being independent of time), $a \in L^2(0,T;L^\infty(\Omega))$ and $f \in L^2(0,T;L^2(\Omega))$. Then there exists a unique solution Φ satisfying the above equation and the regularity bounds given by

$$\|\Phi\|_{L^\infty(0,T;H^1)} + \|\Phi\|_{L^2(0,T;H^2)} \le K\|f\|_{L^2(0,T;L^2)},$$

where K is a constant independent of the input data f.

For simplicity of presentation, we consider problems with no-flow boundary conditions, though the result can be generalized. We make additional assumptions: $\phi \in W^{2;1}_\infty((0,T)\times\Omega)$, $\mathbf{D}_{ij} \in W^{1;0}_\infty((0,T)\times\Omega)$, and $q^+ \in L^2(0,T;L^\infty(\Omega))$.

5.1. Parabolic lift for SIPG.

Lemma 5 (parabolic lift). *Let $a \in L^2(0,T;L^\infty(\Omega))$ and $e \in L^2(0,T;H^1(\mathcal{E}_h))$ satisfy*

$$\left(\frac{\partial \phi e}{\partial t}, w\right) + B_S(e,w;\mathbf{u}) + (ae,w) = 0 \qquad \forall w \in \mathcal{D}_r(\mathcal{E}_h) \quad \forall t \in (0,T], \tag{5.1}$$

$$(\phi e, w) = 0 \qquad \forall w \in \mathcal{D}_r(\mathcal{E}_h), \quad t = 0. \tag{5.2}$$

In addition we let the assumptions in Theorem 1 hold. Then there exists a constant K, independent of h, r, and e, such that

$$\begin{aligned} &\|e\|_{L^2(0,T;L^2)} \\ &\quad \le K\frac{h}{r}\|e\|_{L^\infty(0,T;L^2)} + K\frac{h^2}{r^2}\|e_t\|_{L^2(0,T;L^2)} \\ &\quad +K\frac{h}{r}\|\mathbf{D}^{\frac12}(\mathbf{u})\nabla e\|_{L^2(0,T;L^2)} + K\frac{h}{r^{\frac32-2\delta}}\left(\int_0^T J_0^\sigma(e,e)\right)^{\frac12} \\ &\quad +K\delta\frac{h^{\frac32}}{r^{\frac32}}\left(\sum_{E\in\mathcal{E}_h}\left(\|e\|^2_{L^2(0,T;L^2(\partial E))} + \|\nabla e \cdot \mathbf{n}_{\partial E}\|^2_{L^2(0,T;L^2(\partial E))}\right)\right)^{\frac12}, \end{aligned}$$

where $\delta = 0$ for conforming meshes with triangles or tetrahedra, and $\delta = 1/2$ in general.

Proof. Consider the backward or adjoint parabolic equation:

$$-\frac{\partial \phi\Phi}{\partial t} + \nabla\cdot(-\mathbf{u}\Phi - \mathbf{D}(\mathbf{u})\nabla\Phi) + (a+q^+)\Phi = e, \quad x\in\Omega,\, t\in[0,T), \tag{5.3}$$

$$\mathbf{D}(\mathbf{u})\nabla\Phi\cdot\mathbf{n}_{\partial\Omega} = 0, \quad x\in\partial\Omega,\, t\in[0,T), \tag{5.4}$$

$$\Phi = 0, \quad x\in\Omega,\, t=T. \tag{5.5}$$

Theorem 2 suggests a unique solution Φ for (5.3)–(5.5) satisfying

$$\|\Phi\|_{L^\infty(0,T;H^1)} + \|\Phi\|_{L^2(0,T;H^2)} \le K\|e\|_{L^2(0,T;L^2)}. \tag{5.6}$$

Observing that $\mathbf{D}(\mathbf{u})\nabla\Phi\cdot\mathbf{n}_{\partial\Omega}=0$ on $\partial\Omega$, $\nabla\cdot\mathbf{u}=q$, and $[\mathbf{D}(\mathbf{u})\nabla\Phi\cdot\mathbf{n}_\gamma]=[\Phi]=0$, we multiply both sides of the adjoint equation (5.3) by e, integrate it over the domain Ω, and then apply integration by parts to conclude that

$$\begin{aligned}\|e\|_0^2 &= -\frac{d}{dt}\sum_{E\in\mathcal{E}_h}(e,\phi\Phi)_E + \sum_{E\in\mathcal{E}_h}\left(\phi\frac{\partial e}{\partial t},\Phi\right)_E + \sum_{E\in\mathcal{E}_h}((a-q^-)e,\Phi)_E \\ &\quad + \sum_{E\in\mathcal{E}_h}(\nabla e,\mathbf{D}(\mathbf{u})\nabla\Phi)_E - \sum_{\gamma\in\Gamma_h}\int_\gamma\{\mathbf{D}(\mathbf{u})\nabla\Phi\cdot\mathbf{n}_\gamma\}[e] - \sum_{E\in\mathcal{E}_h}(e,\mathbf{u}\nabla\cdot\Phi)_E \\ &= -\frac{d}{dt}(e,\phi\Phi) + \left(\phi\frac{\partial e}{\partial t},\Phi\right) + (ae,\Phi) + B_S(e,\Phi;\mathbf{u}).\end{aligned}$$

Applying the orthogonality condition (5.1), we obtain

$$\|e\|_0^2 = -\frac{d}{dt}(e,\phi\Phi) + \left(\phi\frac{\partial e}{\partial t},\Phi-\hat\Phi\right) + (ae,\Phi-\hat\Phi) + B_S(e,\Phi-\hat\Phi;\mathbf{u}), \tag{5.7}$$

where $\hat\Phi\in\mathcal{D}_r(\mathcal{E}_h)$ is an interpolant satisfying (3.4) element-wise. The second and third terms on the right-hand side of (5.7) are bounded, by using the Cauchy–Schwarz inequality and approximation results, as

$$\begin{aligned}\left(\phi\frac{\partial e}{\partial t},\Phi-\hat\Phi\right) &\le K\|e_t\|_0\|\Phi-\hat\Phi\|_0 \le K\frac{h^2}{r^2}\|e_t\|_0\|\Phi\|_2,\\ (ae,\Phi-\hat\Phi) &\le K\|a\|_{L^\infty}\|e\|_0\|\Phi-\hat\Phi\|_0 \le K\frac{h^2}{r^2}\|a\|_{L^\infty}\|e\|_0\|\Phi\|_2.\end{aligned}$$

The last term in (5.7) is composed of eight parts:

$$\begin{aligned}&B_S(e,\Phi-\hat\Phi;\mathbf{u})\\ &\quad= \sum_{E\in\mathcal{E}_h}\int_E\mathbf{D}(\mathbf{u})\nabla e\cdot\nabla(\Phi-\hat\Phi) - \sum_{E\in\mathcal{E}_h}\int_E e\mathbf{u}\cdot\nabla(\Phi-\hat\Phi) - \int_\Omega q^- e(\Phi-\hat\Phi)\\ &\qquad - \sum_{\gamma\in\Gamma_h}\int_\gamma\{\mathbf{D}(\mathbf{u})\nabla e\cdot\mathbf{n}_\gamma\}[\Phi-\hat\Phi] - \sum_{\gamma\in\Gamma_h}\int_\gamma\{\mathbf{D}(\mathbf{u})\nabla(\Phi-\hat\Phi)\cdot\mathbf{n}_\gamma\}[e]\\ &\qquad + \sum_{\gamma\in\Gamma_h}\int_\gamma e^*\mathbf{u}\cdot\mathbf{n}_\gamma[\Phi-\hat\Phi] + \sum_{\gamma\in\Gamma_{h,\mathrm{out}}}\int_\gamma e\mathbf{u}\cdot\mathbf{n}_\gamma(\Phi-\hat\Phi) + J_0^\sigma(e,\Phi-\hat\Phi)\\ &\quad =: \sum_{i=1}^{8}T_i.\end{aligned}$$

Once again, the approximation results and Cauchy–Schwarz inequality yield the estimates for the terms T_1, T_2, and T_3:

$$\begin{aligned}
&T_1 \le K \|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla e\|_0 \|\nabla(\Phi - \hat{\Phi})\|_0 \le K\frac{h}{r}\|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla e\|_0 \|\Phi\|_2,\\
&T_2 \le K\frac{h}{r}\|e\|_0\|\Phi\|_2,\\
&T_3 \le K\frac{h^2}{r^2}\|e\|_0\|\Phi\|_2.
\end{aligned}$$

The term T_7 vanishes because of the assumed no-flow boundary condition. The remaining terms in the bilinear form can be bounded by applying the Cauchy–Schwarz inequality on element faces:

$$\begin{aligned}
&T_4 \le K \sum_{E\in\mathcal{E}_h} \|\nabla e\cdot \mathbf{n}_{\partial E}\|_{0,\partial E}\|\Phi - \hat{\Phi}\|_{0,\partial E} \le K\frac{h^{\frac{3}{2}}}{r^{\frac{3}{2}}}\left(\sum_{E\in\mathcal{E}_h}\|\nabla e\cdot \mathbf{n}_{\partial E}\|^2_{0,\partial E}\right)^{\frac{1}{2}}\|\Phi\|_2,\\
&T_5 \le \sum_{\gamma\in\Gamma_h}\|\{\mathbf{D}(\mathbf{u})\nabla(\Phi - \hat{\Phi})\cdot \mathbf{n}_\gamma\}\|_{0,\gamma}\|[e]\|_{0,\gamma} \le K\frac{h}{r^{\frac{3}{2}}}(J_0^\sigma(e,e))^{\frac{1}{2}}\|\Phi\|_2,\\
&T_6 \le K\frac{h^{\frac{3}{2}}}{r^{\frac{3}{2}}}\left(\sum_{E\in\mathcal{E}_h}\|e\|^2_{0,\partial E}\right)^{\frac{1}{2}}\|\Phi\|_2,\\
&T_8 \le (J_0^\sigma(e,e))^{\frac{1}{2}}(J_0^\sigma(\Phi - \hat{\Phi}, \Phi - \hat{\Phi}))^{\frac{1}{2}} \le K\frac{h}{r^{\frac{1}{2}}}(J_0^\sigma(e,e))^{\frac{1}{2}}\|\Phi\|_2.
\end{aligned}$$

We note that, for conforming meshes with triangles or tetrahedra, terms T_4, T_6, and T_8 vanish if we choose a continuous interpolant $\hat{\Phi}$. Substituting all the estimates back into (5.7), we find that

$$\begin{aligned}
\|e\|^2_{0,\Omega} \le& -\frac{d}{dt}(e,\phi\Phi) + K\frac{h^2}{r^2}\|e_t\|_0\|\Phi\|_2 + K\frac{h^2}{r^2}\|a\|_{L^\infty}\|e\|_0\|\Phi\|_2\\
&+K\frac{h}{r}\|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla e\|_0\|\Phi\|_2 + K\frac{h}{r}\|e\|_0\|\Phi\|_2 + K\frac{h}{r^{\frac{3}{2}-2\delta}}(J_0^\sigma(e,e))^{\frac{1}{2}}\|\Phi\|_2\\
&+K\delta\frac{h^{\frac{3}{2}}}{r^{\frac{3}{2}}}\left(\sum_{E\in\mathcal{E}_h}(\|e\|^2_{0,\partial E} + \|\nabla e\cdot\mathbf{n}\|^2_{0,\partial E})\right)^{\frac{1}{2}}\|\Phi\|_2,
\end{aligned}$$

where $\delta = 0$ for conforming meshes with triangles or tetrahedra, and $\delta = 1/2$ in general.

We complete the proof by integrating (5.8) over the time interval $[0,T]$, applying the Cauchy–Schwarz inequality in $L^2(0,T)$, recalling the regularity bound (5.6), and observing the fact that

$$\begin{aligned}
&(e,\phi\Phi)(0) = (\phi e, \Phi - \hat{\Phi})(0)\\
&\le K\frac{h}{r}\|e\|_0(0)\|\Phi\|_1(0) \le K\frac{h}{r}\|e\|_{L^\infty(0,T;L^2)}\|\Phi\|_{L^\infty(0,T;H^1)}. \qquad \square
\end{aligned}$$

5.2. An $L^2(L^2)$ error estimate for the time derivative of the concentration. To obtain an optimal $L^2(L^2)$ error estimate for the concentration, we need an estimate for its time derivative.

THEOREM 3 ($L^2(L^2)$ error estimate for $\mathbf{c}_t$). *Let the assumptions in Theorem* 1 *hold. Then there exists a constant K, independent of h and r, such that*

$$\left\|\frac{\partial}{\partial t}(C^{DG}-c)\right\|_{L^2(0,T;L^2)} + \|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla(C^{DG}-c)\|_{L^\infty(0,T;L^2)}$$
$$\leq K\frac{h^{\mu-2}}{r^{s-3-\delta}}\|c\|_{L^2(0,T;H^s)} + K\frac{h^{\mu-2}}{r^{s-2}}\|\partial c/\partial t\|_{L^2(0,T;H^{s-1})} + K\frac{h^{\mu-2}}{r^{s-5/2}}\|c_0\|_{s-1},$$

where $\mu = \min(r+1,s)$, $r \geq 1$, $s \geq 2$, $\delta = 0$ for conforming meshes with triangles or tetrahedra, and $\delta = 1/2$ in general.

Proof. Let ξ, ξ^I, and ξ^A be defined by (4.1)–(4.3), respectively. Subtracting (3.8) from (3.9), choosing $w = \partial\xi^A/\partial t$, and integrating the resultant equation over the time interval $[0,t]$, $0 < t \leq T$, we obtain

$$\begin{aligned}\int_0^t \left(\frac{\partial\phi\xi^A}{\partial t}, \frac{\partial\xi^A}{\partial t}\right) + \int_0^t B_S\left(\xi^A, \frac{\partial\xi^A}{\partial t}; \mathbf{u}\right) &= \int_0^t \left(L\left(\frac{\partial\xi^A}{\partial t}; \mathbf{u}, C^{DG}\right) - L\left(\frac{\partial\xi^A}{\partial t}; \mathbf{u}, c\right)\right)\\ &\quad + \int_0^t \left(\frac{\partial\phi\xi^I}{\partial t}, \frac{\partial\xi^A}{\partial t}\right) + \int_0^t B_S\left(\xi^I, \frac{\partial\xi^A}{\partial t}; \mathbf{u}\right).\end{aligned} \tag{5.8}$$

A simple manipulation breaks the bilinear form on the left-hide side of (5.8) into nine components:

$$B_S\left(\xi^A, \frac{\partial\xi^A}{\partial t}; \mathbf{u}\right) = \left(\frac{d}{dt}\sum_{i=1}^7 T_i\right) + T_8 + T_9,$$

where

$$\begin{aligned}\sum_{i=1}^7 T_i &:= \frac{1}{2}\sum_{E\in\mathcal{E}_h}\int_E \mathbf{D}(\mathbf{u})\nabla\xi^A\cdot\nabla\xi^A - \sum_{E\in\mathcal{E}_h}\int_E \xi^A\mathbf{u}\cdot\nabla\xi^A - \frac{1}{2}\int_\Omega q^-\left(\xi^A\right)^2\\ &\quad - \sum_{\gamma\in\Gamma_h}\int_\gamma \left\{\mathbf{D}(\mathbf{u})\nabla\xi^A\cdot\mathbf{n}_\gamma\right\}\left[\xi^A\right] + \sum_{\gamma\in\Gamma_h}\int_\gamma \xi^{A*}\mathbf{u}\cdot\mathbf{n}_\gamma\left[\xi^A\right]\\ &\quad + \frac{1}{2}\sum_{\gamma\in\Gamma_{h,\text{out}}}\int_\gamma \mathbf{u}\cdot\mathbf{n}_\gamma\left(\xi^A\right)^2 + \frac{1}{2}J_0^\sigma\left(\xi^A,\xi^A\right),\\ T_8 &:= \sum_{E\in\mathcal{E}_h}\int_E \frac{\partial\xi^A}{\partial t}\mathbf{u}\cdot\nabla\xi^A,\\ T_9 &:= -\sum_{\gamma\in\Gamma_h}\int_\gamma \frac{\partial\xi^{A*}}{\partial t}\mathbf{u}\cdot\mathbf{n}_\gamma\left[\xi^A\right].\end{aligned}$$

Consequently, the left-hand side of (5.8) may be written as

$$\begin{aligned}&\int_0^t \left(\frac{\partial\phi\xi^A}{\partial t}, \frac{\partial\xi^A}{\partial t}\right) + \int_0^t B_S\left(\xi^A, \frac{\partial\xi^A}{\partial t}; \mathbf{u}\right)\\ &\quad = \int_0^t \left\|\frac{\partial}{\partial t}\sqrt{\phi}\xi^A\right\|_0^2 + \sum_{i=1}^7 T_i(t) - \sum_{i=1}^7 T_i(0) + \int_0^t T_8 + \int_0^t T_9.\end{aligned}$$

It is easy to see that the terms $\int_0^t \|\frac{\partial}{\partial t}\sqrt{\phi}\xi^A\|_0^2$, $T_1(t)$, $T_3(t)$, $T_6(t)$, and $T_7(t)$ are nonnegative. By applying the Cauchy–Schwarz inequality and Theorem 1, the term $T_2(t)$ can be bounded as

$$\begin{aligned}
|T_2(t)| &\le \epsilon |\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|\!|\!|_0^2 + K\|\xi^A\|_0^2 \\
&\le \epsilon |\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|\!|\!|_0^2 + K\|\xi^A\|_{L^\infty(0,T;L^2)}^2 \\
&\le \epsilon |\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|\!|\!|_0^2 + KR_s^2,
\end{aligned}$$

where

$$R_s := \frac{h^{\mu-1}}{r^{s-1-\delta}}|\!|\!|c|\!|\!|_{L^2(0,T;H^s)} + \frac{h^{\mu-1}}{r^{s-1}}(|\!|\!|\partial c/\partial t|\!|\!|_{L^2(0,T;H^{s-1})} + |\!|\!|c_0|\!|\!|_{s-1}).$$

Recalling the definition of the penalty term and applying the Cauchy–Schwarz and inverse inequalities, we may bound the terms T_4 and T_5:

$$\begin{aligned}
|T_4(t)| &\le \frac{\epsilon}{K}\sum_{E\in\mathcal{E}_h}\frac{h}{r^2}\|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A\|_{0,\partial E}^2 + \epsilon J_0^\sigma(\xi^A,\xi^A) \\
&\le \epsilon |\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|\!|\!|_0^2 + \epsilon J_0^\sigma(\xi^A,\xi^A), \\
|T_5(t)| &\le K\sum_{E\in\mathcal{E}_h}\frac{h}{r^2}\|\xi^A\|_{0,\partial E}^2 + \epsilon J_0^\sigma(\xi^A,\xi^A) \le K\|\xi^A\|_0^2 + \epsilon J_0^\sigma(\xi^A,\xi^A) \\
&\le KR_s^2 + \epsilon J_0^\sigma(\xi^A,\xi^A).
\end{aligned}$$

Applications of the approximation results and the continuity of the L_2 projection give

$$\sum_{i=1}^7 |T_i(0)| \le K\frac{h^{2\mu-4}}{r^{2s-5}}|\!|\!|c_0|\!|\!|_{s-1}^2.$$

The Cauchy–Schwarz inequality and Theorem 1 imply

$$\begin{aligned}
\left|\int_0^t T_8\right| &\le \epsilon\left\|\sqrt{\phi}\frac{\partial\xi^A}{\partial t}\right\|_{L^2(0,T;L^2)}^2 + |\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla\xi^A|\!|\!|_{L^2(0,T;L^2)}^2 \\
&\le \epsilon\left\|\sqrt{\phi}\frac{\partial\xi^A}{\partial t}\right\|_{L^2(0,T;L^2)}^2 + KR_s^2.
\end{aligned}$$

An application of the Cauchy–Schwarz and inverse inequalities yields

$$\begin{aligned}
\left|\int_0^t T_9\right| &\le \epsilon\left\|\sqrt{\phi}\frac{\partial\xi^A}{\partial t}\right\|_{L^2(0,T;L^2)}^2 + K\int_0^t J_0^\sigma\left(\xi^A,\xi^A\right) \\
&\le \epsilon\left\|\sqrt{\phi}\frac{\partial\xi^A}{\partial t}\right\|_{L^2(0,T;L^2)}^2 + KR_s^2.
\end{aligned}$$

Collecting the above estimates, we conclude that the left-hide side of (5.8) has the following lower bound:

$$\begin{aligned}
&\int_0^t \left(\frac{\partial \phi \xi^A}{\partial t}, \frac{\partial \xi^A}{\partial t}\right) + \int_0^t B_S\left(\xi^A, \frac{\partial \xi^A}{\partial t}; \mathbf{u}\right) \\
&\quad \geq \frac{1}{2}\int_0^t \left\| \frac{\partial}{\partial t}\sqrt{\phi}\xi^A \right\|_0^2 + \frac{1}{3}|\!|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla \xi^A|\!|\!|_0^2 + \frac{1}{2}\int_\Omega \left|q^-\right| \left(\xi^A\right)^2 \\
&\qquad + \frac{1}{2}\sum_{\gamma\in\Gamma_{h,\mathrm{out}}} \int_\gamma \mathbf{u}\cdot\mathbf{n}_\gamma \left(\xi^A\right)^2 + \frac{1}{3}J_0^\sigma\left(\xi^A, \xi^A\right) \\
&\qquad - KR_s^2 - K\frac{h^{2\mu-4}}{r^{2s-5}}|\!|\!|c_0|\!|\!|_{s-1}^2.
\end{aligned}$$

The first integrand on the right-hand side of (5.8) may be bounded, by using the Cauchy–Schwarz inequality and the Lipschitz continuity of the cut-off operator, as

$$\begin{aligned}
&L\left(\frac{\partial \xi^A}{\partial t}; \mathbf{u}, C^{DG}\right) - L\left(\frac{\partial \xi^A}{\partial t}; \mathbf{u}, c\right) = \int_\Omega \left(r\left(\mathcal{M}(C^{DG})\right) - r\left(\mathcal{M}(c)\right)\right)\frac{\partial \xi^A}{\partial t} \\
&\leq \epsilon \left\| \frac{\partial}{\partial t}\sqrt{\phi}\xi^A \right\|_0^2 + K\left\|\xi\right\|_0^2 \leq \epsilon \left\| \frac{\partial}{\partial t}\sqrt{\phi}\xi^A \right\|_0^2 + KR_s^2.
\end{aligned}$$

An easy application of the Cauchy–Schwarz inequality and approximation results yields the following estimate for the second integrand:

$$\begin{aligned}
&\left(\frac{\partial \phi \xi^I}{\partial t}, \frac{\partial \xi^A}{\partial t}\right) \\
&\leq \epsilon \left\| \frac{\partial}{\partial t}\sqrt{\phi}\xi^A \right\|_0^2 + K\left\| \frac{\partial \xi^I}{\partial t} \right\|_0^2 \leq \epsilon \left\| \frac{\partial}{\partial t}\sqrt{\phi}\xi^A \right\|_0^2 + K\frac{h^{2\mu-2}}{r^{2s-2}}|\!|\!|c_t|\!|\!|_{s-1}^2.
\end{aligned}$$

The third integrand may be decomposed into eight parts:

$$\begin{aligned}
&B_S\left(\xi^I, \frac{\partial \xi^A}{\partial t}; \mathbf{u}\right) \\
&\quad = \sum_{E\in\mathcal{E}_h}\int_E \mathbf{D}(\mathbf{u})\nabla\xi^I\cdot\nabla\frac{\partial \xi^A}{\partial t} - \sum_{E\in\mathcal{E}_h}\int_E \xi^I\mathbf{u}\cdot\nabla\frac{\partial \xi^A}{\partial t} - \int_\Omega q^-\xi^I\frac{\partial \xi^A}{\partial t} \\
&\qquad - \sum_{\gamma\in\Gamma_h}\int_\gamma \left\{\mathbf{D}(\mathbf{u})\nabla\xi^I\cdot\mathbf{n}_\gamma\right\}\left[\frac{\partial \xi^A}{\partial t}\right] - \sum_{\gamma\in\Gamma_h}\int_\gamma \left\{\mathbf{D}(\mathbf{u})\nabla\frac{\partial \xi^A}{\partial t}\cdot\mathbf{n}_\gamma\right\}\left[\xi^I\right] \\
&\qquad + \sum_{\gamma\in\Gamma_h}\int_\gamma \xi^{I*}\mathbf{u}\cdot\mathbf{n}_\gamma\left[\frac{\partial \xi^A}{\partial t}\right] + \sum_{\gamma\in\Gamma_{h,\mathrm{out}}}\int_\gamma \mathbf{u}\cdot\mathbf{n}_\gamma\xi^I\frac{\partial \xi^A}{\partial t} + J_0^\sigma\left(\xi^I, \frac{\partial \xi^A}{\partial t}\right) \\
&\quad =: \sum_{i=1}^{8} S_i.
\end{aligned}$$

The terms S_3 and S_8 are bounded by applying the Cauchy–Schwarz inequality and approximation results:

$$\begin{aligned}
|S_3| &\leq \epsilon \left\| \frac{\partial}{\partial t}\sqrt{\phi}\xi^A \right\|_0^2 + K\frac{h^{2\mu}}{r^{2s}}|\!|\!|c|\!|\!|_s^2, \\
|S_8| &\leq \epsilon J_0^\sigma\left(\xi^A, \xi^A\right) + KJ_0^\sigma\left(\xi^I, \xi^I\right) \\
&\leq \epsilon J_0^\sigma\left(\xi^A, \xi^A\right) + K\frac{h^{2\mu-2}}{r^{2s-3}}|\!|\!|c|\!|\!|_s^2.
\end{aligned}$$

Applications of the Cauchy–Schwarz and inverse inequalities yield the following estimates for the remaining terms:

$$|S_1| + |S_2| + |S_4| + |S_6| + |S_7| \le \epsilon \left\| \frac{\partial}{\partial t} \sqrt{\phi} \xi^A \right\|_0^2 + K \frac{h^{2\mu-4}}{r^{2s-6}} \|\!|c|\!\|_s^2,$$

$$|S_5| \le \epsilon \left\| \frac{\partial}{\partial t} \sqrt{\phi} \xi^A \right\|_0^2 + K \frac{h^{2\mu-4}}{r^{2s-7}} \|\!|c|\!\|_s^2.$$

For conforming meshes with triangles or tetrahedra, we can choose a continuous $\widehat{c}$ to force $S_5 = S_8 = 0$. Combining the bounds for the terms S_i, we obtain

$$\begin{aligned} &\int_0^t B_S\left(\xi^I, \frac{\partial \xi^A}{\partial t}; \mathbf{u}\right) \\ &\quad \le \epsilon \int_0^t \left\| \frac{\partial}{\partial t} \sqrt{\phi} \xi^A \right\|_0^2 + \epsilon R_S^2 + K \frac{h^{2\mu-4}}{r^{2s-6-2\delta}} \|\!|c|\!\|_{L^2(0,T;H^s)}^2. \end{aligned}$$

By back-substituting the estimates into (5.8), we conclude that

$$\begin{aligned} &\int_0^t \left\| \frac{\partial}{\partial t} \sqrt{\phi} \xi^A \right\|_0^2 + \|\!|\mathbf{D}^{\frac{1}{2}}(\mathbf{u}) \nabla \xi^A|\!\|_0^2 + \int_\Omega |q^-| \left(\xi^A\right)^2 \\ &\quad + \sum_{\gamma \in \Gamma_{h,\text{out}}} \int_\gamma \mathbf{u} \cdot \mathbf{n}_\gamma \left(\xi^A\right)^2 + J_0^\sigma\left(\xi^A, \xi^A\right) \\ &\quad \le K \frac{h^{2\mu-4}}{r^{2s-6-2\delta}} \|\!|c|\!\|_{L^2(0,T;H^s)}^2 + K \frac{h^{2\mu-4}}{r^{2s-5}} \|\!|c_0|\!\|_{s-1}^2 + K R_s^2 \\ &\quad \le K \frac{h^{2\mu-4}}{r^{2s-6-2\delta}} \|\!|c|\!\|_{L^2(0,T;H^s)}^2 + K \frac{h^{2\mu-4}}{r^{2s-5}} \|\!|c_0|\!\|_{s-1}^2 + K \frac{h^{2\mu-2}}{r^{2s-2}} \|\!|c_t|\!\|_{L^2(0,T;H^{s-1})}^2. \end{aligned}$$

The theorem follows from the triangle inequality, approximation results, and (4.7). □

5.3. Face error estimates. We also need an error estimate on element faces in order to apply the parabolic lift lemma.

Theorem 4 (face error estimates). *Let the assumptions in Theorem* 1 *hold. Then there exists a constant* K, *independent of* h *and* r, *such that*

$$\begin{aligned} &\left(\sum_{E \in \mathcal{E}_h} \left\| C^{DG} - c \right\|_{L^2(0,T;L^2(\partial E))}^2 \right)^{\frac{1}{2}} + \left(\sum_{E \in \mathcal{E}_h} \left\| \nabla \left(C^{DG} - c\right) \cdot \mathbf{n}_{\partial E} \right\|_{L^2(0,T;L^2(\partial E))}^2 \right)^{\frac{1}{2}} \\ &\quad \le K \frac{h^{\mu - \frac{3}{2}}}{r^{s-2-\delta}} \|\!|c|\!\|_{L^2(0,T;H^s)} + K \frac{h^{\mu-\frac{3}{2}}}{r^{s-2}} \left(\|\!|\partial c/\partial t|\!\|_{L^2(0,T;H^{s-1})} + \|\!|c_0|\!\|_{s-1} \right), \end{aligned}$$

where $\mu = \min(r+1, s)$, $r \ge 1$, $s \ge 2$, $\delta = 0$ *for conforming meshes with triangles or tetrahedra, and* $\delta = 1/2$ *in general.*

Proof. As the first term can be bounded similarly with even sharper estimates, we only present the estimation of the second term, which can be obtained by applying the triangle and inverse inequalities, recalling Theorem 1 and using the approximation results:

$$
\begin{aligned}
&\left(\sum_{E\in\mathcal{E}_h} \left\|\nabla\left(C^{DG}-c\right)\cdot\mathbf{n}_{\partial E}\right\|^2_{L^2(0,T;L^2(\partial E))}\right)^{\frac{1}{2}}\\
&\quad\le \left(\sum_{E\in\mathcal{E}_h} \left\|\nabla\left(C^{DG}-\hat{c}\right)\right\|^2_{L^2(0,T;L^2(\partial E))}\right)^{\frac{1}{2}} + \left(\sum_{E\in\mathcal{E}_h} \left\|\nabla\left(\hat{c}-c\right)\right\|^2_{L^2(0,T;L^2(\partial E))}\right)^{\frac{1}{2}}\\
&\quad\le \frac{r}{h^{\frac{1}{2}}}\left(\sum_{E\in\mathcal{E}_h} \left\|\nabla\left(C^{DG}-\hat{c}\right)\right\|^2_{L^2(0,T;L^2(E))}\right)^{\frac{1}{2}} + K\frac{h^{\mu-\frac{3}{2}}}{r^{s-\frac{3}{2}}}|||c|||_{L^2(0,T;H^s)}\\
&\quad\le \frac{r}{h^{\frac{1}{2}}}\left(\sum_{E\in\mathcal{E}_h} \left\|\nabla\left(C^{DG}-c\right)\right\|^2_{L^2(0,T;L^2(E))}\right)^{\frac{1}{2}} + K\frac{r}{h^{\frac{1}{2}}}\frac{h^{\mu-1}}{r^{s-1}}|||c|||_{L^2(0,T;H^s)}\\
&\quad\le K\frac{h^{\mu-\frac{3}{2}}}{r^{s-2-\delta}}|||c|||_{L^2(0,T;H^s)} + K\frac{h^{\mu-\frac{3}{2}}}{r^{s-2}}\left(|||\partial c/\partial t|||_{L^2(0,T;H^{s-1})} + |||c_0|||_{s-1}\right). \qquad \square
\end{aligned}
$$

5.4. An $L^2(L^2)$ error estimate for the concentration.

THEOREM 5 ($L^2(L^2)$ error estimate for c). *Let the assumptions in Theorem* 1 *hold. Then there exists a constant K, independent of h and r, such that*

$$
\begin{aligned}
(5.9)\qquad &\left\|C^{DG}-c\right\|_{L^2(0,T;L^2)}\\
&\le K\frac{h^{\mu}}{r^{s-1-\delta}}|||c|||_{L^2(0,T;H^s)} + K\frac{h^{\mu}}{r^{s-\delta}}|||\partial c/\partial t|||_{L^2(0,T;H^{s-1})} + K\frac{h^{\mu}}{r^{s-1/2}}|||c_0|||_{s-1},
\end{aligned}
$$

where $\mu=\min(r+1,s)$, $r\ge 1$, $s\ge 2$, $\delta=0$ for conforming meshes with triangles or tetrahedra, and $\delta=1/2$ in general.

Proof. We recall the concentration error ξ in (4.1), and the error equation:

$$
\left(\frac{\partial\phi\xi}{\partial t}, w\right) + B(\xi, w;\mathbf{u}) = L\left(w;\mathbf{u}, C^{DG}\right) - L\left(w;\mathbf{u}, c\right) \qquad \forall w\in\mathcal{D}_r\left(\mathcal{E}_h\right).
$$

We define

$$
a(x,t) = \begin{cases} -\frac{r\left(\mathcal{M}(C^{DG}(x,t))\right)-r(\mathcal{M}(c(x,t)))}{C^{DG}(x,t)-c(x,t)} & \text{if } C^{DG}(x,t)-c(x,t)\neq 0,\\[2ex] 0 & \text{if } C^{DG}(x,t)-c(x,t)=0. \end{cases}
$$

Consequently, we have $L(w;\mathbf{u},C^{DG})-L(w;\mathbf{u},c) = -(a\xi,\,\omega)$. Noting the fact that $a\in L^\infty(0,T;L^\infty)\subset L^2(0,T;L^\infty)$ and recalling Theorems 1, 3, and 4, we obtain (5.9) by applying the parabolic lift argument of Lemma 5. $\square$

6. Optimal estimates in negative norms for the symmetric scheme.

6.1. Error estimates in terms of linear functionals.

We again assume no-flow boundary conditions. Given a function $f\in L^2(0,T;L^2(\Omega))$, we consider a linear functional $F(\cdot)$ of the following form:

$$
F(c) = \int_0^T\int_\Omega c(x,t)f(x,t)dx\,dt.
$$

LEMMA 6 (parabolic lift). *Let $e\in L^2(0,T;H^1(\mathcal{E}_h))$ satisfy* (5.1)–(5.2) *and let the assumptions in Theorem* 1 *hold. We further assume $\phi\in W^{s_1+2,1}_\infty((0,T)\times\Omega)$, $\mathbf{D}_{ij}\in$*

$W_\infty^{s_1+1,0}((0,T)\times\Omega)$, $\mathbf{u}_i \in W_\infty^{s_1}(\Omega)$, $a \in W_\infty^{s_1,0}((0,T)\times\Omega)$, and $q^+ \in W_\infty^{s_1,0}((0,T)\times\Omega)$. Then there exists a constant K, independent of h, r, e, and f, such that

$$
\begin{aligned}
|F(e)| \le K \|f\|_{L^2(0,T;H^{s_1})} \Bigg(& \frac{h^{\mu_1+1}}{r^{s_1+1}} \|e\|_{L^\infty(0,T;L^2)} + \frac{h^{\mu_1+2}}{r^{s_1+2}} \|e_t\|_{L^2(0,T;L^2)} \\
& + \frac{h^{\mu_1+1}}{r^{s_1+1}} \|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla e\|_{L^2(0,T;L^2)} + \frac{h^{\mu_1+1}}{r^{s_1+\frac{3}{2}-2\delta}} \left(\int_0^T J_0^\sigma(e,e) \right)^{\frac{1}{2}} \\
& + \frac{h^{\mu_1+\frac{3}{2}}}{r^{s_1+\frac{3}{2}}} \delta \left(\sum_{E\in\mathcal{E}_h} \left(\|e\|^2_{L^2(0,T;L^2(\partial E))} + \|\nabla e \cdot \mathbf{n}_{\partial E}\|^2_{L^2(0,T;L^2(\partial E))} \right) \right)^{\frac{1}{2}} \Bigg),
\end{aligned}
$$

where $\mu_1 = \min(r-1, s_1)$, $r \ge 1$, $s_1 \ge 0$, $\delta = 0$ for conforming meshes with triangles or tetrahedra, and $\delta = 1/2$ in general.

Proof. We revisit the adjoint parabolic equation (5.3)–(5.5) with e replaced by f. By applying Theorem 2 repeatedly, we obtain a unique solution Φ for (5.3)–(5.5) satisfying

$$
\|\Phi\|_{L^\infty(0,T;H^{s_1+1})} + \|\Phi\|_{L^2(0,T;H^{s_1+2})} \le K \|f\|_{L^2(0,T;H^{s_1})} . \tag{6.1}
$$

We now consider the $L^2(\Omega)$ inner product $(e,\, f)$ at $t \in (0,T]$:

$$
\begin{aligned}
(e,\, f) &= \sum_{E\in\mathcal{E}_h} (e,\, f)_E \\
&= \sum_{E\in\mathcal{E}_h} \left(e, -\frac{\partial \phi\Phi}{\partial t} \right)_E + \sum_{E\in\mathcal{E}_h} (e, \nabla\cdot(-\mathbf{u}\Phi - \mathbf{D}(\mathbf{u})\nabla\Phi))_E + \sum_{E\in\mathcal{E}_h} \left(e, (a+q^+)\Phi \right)_E .
\end{aligned}
$$

Integrating by parts, applying the orthogonality condition (5.1) and observing that $\mathbf{D}(\mathbf{u})\nabla\Phi\cdot\mathbf{n}_{\partial\Omega} = 0$ on $\partial\Omega$, $\nabla\cdot\mathbf{u} = q$, and $[\mathbf{D}(\mathbf{u})\nabla\Phi\cdot\mathbf{n}_\gamma] = [\Phi] = 0$, we conclude that

$$
(e,\, f) = -\frac{d}{dt}(e, \phi\Phi) + \left(\phi\frac{\partial e}{\partial t}, \Phi - \hat{\Phi} \right) + \left(ae, \Phi - \hat{\Phi} \right) + B_S\left(e, \Phi - \hat{\Phi}; \mathbf{u} \right), \tag{6.2}
$$

where we choose an interpolant $\hat{\Phi} \in \mathcal{D}_r(\mathcal{E}_h)$ with element-wise optimal approximation properties (3.4). Applying the Cauchy–Schwarz inequality and approximation results, we obtain estimates for the second and third terms on the right-hand side of (6.2):

$$
\begin{aligned}
\left(\phi\frac{\partial e}{\partial t}, \Phi - \hat{\Phi} \right) &\le K \|e_t\|_0 \|\Phi - \hat{\Phi}\|_0 \le K \frac{h^{\mu_1+2}}{r^{s_1+2}} \|e_t\|_0 \|\Phi\|_{s_1+2} , \\
\left(ae, \Phi - \hat{\Phi} \right) &\le K \|a\|_{L^\infty} \|e\|_0 \|\Phi - \hat{\Phi}\|_0 \le K \frac{h^{\mu_1+2}}{r^{s_1+2}} \|a\|_{L^\infty} \|e\|_0 \|\Phi\|_{s_1+2} .
\end{aligned}
$$

Similar but tedious arguments, together with the inverse inequality and the existence of continuous interpolants for conforming meshes with triangles or tetrahedra, yield a bound for the fourth term:

$$
\begin{aligned}
\left| B_S\left(e, \Phi - \hat{\Phi}; \mathbf{u} \right) \right| \le\; & K \frac{h^{\mu_1+1}}{r^{s_1+1}} \|\mathbf{D}^{\frac{1}{2}}(\mathbf{u})\nabla e\|_0 \|\Phi\|_{s_1+2} + K \frac{h^{\mu_1+1}}{r^{s_1+1}} \|e\|_0 \|\Phi\|_{s_1+2} \\
& + K \frac{h^{\mu_1+1}}{r^{s_1+\frac{3}{2}-2\delta}} \left(J_0^\sigma(e,e) \right)^{\frac{1}{2}} \|\Phi\|_{s_1+2} \\
& + K\delta \frac{h^{\mu_1+\frac{3}{2}}}{r^{s_1+\frac{3}{2}}} \left(\sum_{E\in\mathcal{E}_h} \left(\|e\|^2_{0,\partial E} + \|\nabla e\cdot\mathbf{n}_{\partial E}\|^2_{0,\partial E} \right) \right)^{\frac{1}{2}} \|\Phi\|_{s_1+2} .
\end{aligned}
$$

Observing the fact that

$$\begin{aligned}|F(e) - (e, \phi\Phi)(0)| &= \left| \int_0^T \left((e,\, f) + \frac{d}{dt}(e, \phi\Phi) \right) \right| \\ &\leq \int_0^T \left| (e,\, f) + \frac{d}{dt}(e, \phi\Phi) \right|\end{aligned}$$

and integrating (6.2) over the time interval $[0, T]$, we have

$$\begin{aligned}&|F(e) - (e, \phi\Phi)(0)| \\ &\leq K \left\| \Phi \right\|_{L^2(0,T;H^{s_1+2})} \left(\frac{h^{\mu_1+2}}{r^{s_1+2}} \left\| e_t \right\|_{L^2(0,T;L^2)} \right. \\ &+ \frac{h^{\mu_1+2}}{r^{s_1+2}} \left\| a \right\|_{L^2(0,T;L^\infty)} \left\| e \right\|_{L^\infty(0,T;L^2)} + \frac{h^{\mu_1+1}}{r^{s_1+1}} \left\| \mathbf{D}^{\frac{1}{2}}(\mathbf{u}) \nabla e \right\|_{L^2(0,T;L^2)} \\ &+ \frac{h^{\mu_1+1}}{r^{s_1+1}} \left\| e \right\|_{L^\infty(0,T;L^2)} + \frac{h^{\mu_1+1}}{r^{s_1+\frac{3}{2}-2\delta}} \left(\int_0^T J_0^\sigma(e,e) \right)^{\frac{1}{2}} \\ &\left. + \frac{h^{\mu_1+\frac{3}{2}}}{r^{s_1+\frac{3}{2}}} \delta \left(\sum_{E \in \mathcal{E}_h} \left(\left\| e \right\|^2_{L^2(0,T;L^2(\partial E))} + \left\| \nabla e \cdot \mathbf{n}_{\partial E} \right\|^2_{L^2(0,T;L^2(\partial E))} \right) \right)^{\frac{1}{2}} \right).\end{aligned}$$

The theorem follows from the regularity estimate (6.1) and the fact that

$$\begin{aligned}|(e, \phi\Phi)(0)| &= \left| \left(\phi e, \Phi - \hat{\Phi} \right)(0) \right| \\ &\leq K \frac{h^{\min(r+1, s_1+1)}}{r^{s_1+1}} \left\| e(\cdot, 0) \right\|_0 \left\| \Phi(\cdot, 0) \right\|_{s_1+1} \\ &\leq K \frac{h^{\mu_1+1}}{r^{s_1+1}} \left\| e \right\|_{L^\infty(0,T;L^2)} \left\| \Phi \right\|_{L^\infty(0,T;H^{s_1+1})}. \qquad \square\end{aligned}$$

Theorem 6 (linear functional estimates). *Let the assumptions in Theorem* 1 *hold. In addition, we assume* $\phi \in W_\infty^{s_1+2,1}((0,T) \times \Omega)$, $\mathbf{D}_{ij} \in W_\infty^{s_1+1,0}((0,T) \times \Omega)$, $\mathbf{u}_i \in W_\infty^{s_1}(\Omega)$, $q^+ \in W_\infty^{s_1,0}((0,T) \times \Omega)$, *and that the chemical reaction term has a linear form* $r(c) = k_0 + k_1 c$, *where* $k_0 = k_0(x,t)$ *and* $k_1 = k_1(x,t)$ *are reaction parameters with* $k_1 \in W_\infty^{s_1,0}((0,T) \times \Omega)$. *Then there exists a constant* K, *independent of* h, r, *and* f, *such that*

$$\begin{aligned}\left| F(C^{DG}) - F(c) \right| \leq K &\frac{h^{\mu_1+\mu}}{r^{s_1+s-1-\delta}} \left\| f \right\|_{L^2(0,T;H^{s_1})} \left\| c \right\|_{L^2(0,T;H^s)} \\ &+ K \frac{h^{\mu_1+\mu}}{r^{s_1+s-\delta}} \left\| f \right\|_{L^2(0,T;H^{s_1})} \left\| \partial c / \partial t \right\|_{L^2(0,T;H^{s-1})} \\ &+ K \frac{h^{\mu_1+\mu}}{r^{s_1+s-1/2}} \left\| f \right\|_{L^2(0,T;H^{s_1})} \left\| c_0 \right\|_{s-1},\end{aligned}$$

where $\mu = \min(r+1, s)$, $\mu_1 = \min(r-1, s_1)$, $r \geq 1$, $s \geq 2$, $s_1 \geq 0$, *and* $\delta = 0$ *for conforming meshes with triangles or tetrahedra, and* $\delta = 1/2$ *in general.*

Proof. Recalling the concentration error ξ in (4.1) and defining $a(x,t) = -k_1(x,t)$, we obtain the error equation in the following form, provided that the cut-off constant

M is chosen to be sufficiently large:

$$\left(\frac{\partial \phi \xi}{\partial t}, w\right) + B_S(\xi, w; \mathbf{u}) + (a\xi, w) = 0 \quad \forall w \in \mathcal{D}_r\left(\mathcal{E}_h\right) \quad \forall t \in (0, T].$$

We obtain the desired estimate by applying the parabolic lift of Lemma 6 together with estimates in Theorems 1, 3, and 4. ∎

6.2. Error estimates in negative norms. Assuming m is a positive integer, we define the negative Sobolev norm $\|\cdot\|_{H^{-m}(\Omega)}$ in the usual way:

$$\|c\|_{H^{-m}(\Omega)} = \sup_{v \in C_0^\infty(\Omega) \backslash \{0\}} \frac{|(c, v)|}{\|v\|_{H^m(\Omega)}}.$$

THEOREM 7 (estimates in negative norms). *Let the assumptions in Theorem* 1 *hold. In addition, we assume* $\phi \in W_\infty^{m+2,1}((0,T) \times \Omega)$, $\mathbf{D}_{ij} \in W_\infty^{m+1,0}((0,T) \times \Omega)$, $\mathbf{u}_i \in W_\infty^m(\Omega)$, $q^+ \in W_\infty^{m,0}((0,T) \times \Omega)$, *and that the chemical reaction term has a linear form* $r(c) = k_0 + k_1 c$, *where* $k_0 = k_0(x,t)$ *and* $k_1 = k_1(x,t)$ *are reaction parameters with* $k_1 \in W_\infty^{m,0}((0,T) \times \Omega)$. *Then there exists a constant* K, *independent of* h *and* r, *such that*

$$\begin{aligned} \left\|C^{DG} - c\right\|_{L^2(0,T;H^{-m}(\Omega))} \leq{}& K \frac{h^{\min(r-1,m)+\min(r+1,s)}}{r^{m+s-1-\delta}} \|c\|_{L^2(0,T;H^s)} \\ &+ K \frac{h^{\min(r-1,m)+\min(r+1,s)}}{r^{m+s-\delta}} \|\partial c/\partial t\|_{L^2(0,T;H^{s-1})} \\ &+ K \frac{h^{\min(r-1,m)+\min(r+1,s)}}{r^{m+s-1/2}} \|c_0\|_{s-1}, \end{aligned}$$

where $r \geq 1$, $s \geq 2$, $m \geq 0$, *and* $\delta = 0$ *for conforming meshes with triangles or tetrahedra, and* $\delta = 1/2$ *in general.*

Proof. The theorem follows directly from Theorem 6 and the definition of negative norms. ∎

7. Numerical examples. We consider the problem of (2.1)–(2.4) on a domain $\Omega = (0, 10)^2$ without injection or extraction, i.e., $q = 0$, and with a reaction term $r = r(x,t)$ independent of the concentration c. The domain is divided into two disjoint parts: $\overline{\Omega} = \overline{\Omega}_1 \cup \overline{\Omega}_2$ with $\Omega_1 = \{(x,y) \in \Omega : y < 3 + 0.4x\}$. The porosity ϕ has a constant value of 0.1, and the tensor $\mathbf{D}$ is a constant diagonal tensor with $\mathbf{D}_{ii} = 1.0$. We impose the velocities $\mathbf{u} = (-1, -0.4)$ in Ω_1 and $\mathbf{u} = (0,0)$ in Ω_2. We choose $r(x,t)$, c_B, and c_0 such that the equation has an analytical solution of $c = (1 + \cos(\frac{\pi}{5}x)\cos(\frac{\pi}{5}y))2^{-t/10}$. The penalty parameter is chosen according to the method presented in the next section. The coarsest mesh we take simply consists of the two quadrilateral elements Ω_1 and Ω_2. The simulation time interval is (0, 10], and we use the backward Euler method for time integration with a uniform time step $\Delta t = 0.1$.

7.1. Convergence of h-refinement. We solve the test case using OBB-DG, NIPG, IIPG, and SIPG. We use polynomials of degree $r = 2$ and vary h by uniform refinements starting from the coarsest mesh. The convergence behaviors of h-refinement in the norms of $L^2(L^2)$, $L^\infty(L^2)$, and $L^2(H^1)$ for NIPG are shown in Figure 7.1. It is observed that the errors in all norms are $O(1/n)$, where n is the number of degrees of freedom. As $n \propto 1/h^2$ for two-dimensional spaces, the experimental convergences

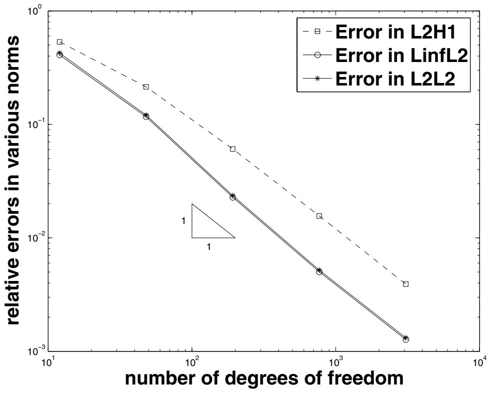


FIG. 7.1. *Convergence of h-refinement for NIPG.*

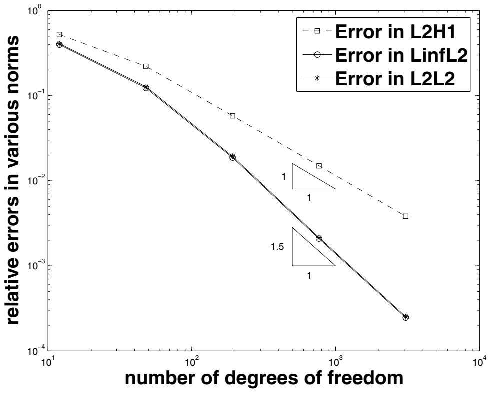


FIG. 7.2. *Convergence of h-refinement for SIPG.*

confirm our theoretical estimates in $L^2(H^1)$. In addition, the numerical results indicate that the errors in NIPG do not converge optimally in $L^\infty(L^2)$ or $L^2(L^2)$. The convergence behaviors of OBB-DG and IIPG (not shown) are nearly identical to those of NIPG. However, unlike NIPG, OBB-DG, and IIPG, the symmetric scheme (SIPG) possesses optimal convergence in all norms of $L^2(L^2)$, $L^\infty(L^2)$, and $L^2(H^1)$, as shown evidently in Figure 7.2, which also validates the predictions from our parabolic lift arguments.

7.2. Convergence of p-refinement. The test case is solved using the four primal DGs on the coarsest mesh with polynomials of degrees r=1, 2, 3, . . . , 10. Figure 7.3 illustrates the convergence behaviors of SIPG in the norms of $L^2(L^2)$, $L^\infty(L^2)$, and $L^2(H^1)$, where the expected exponential convergence rates are achieved. The exponential convergence patterns of OBB-DG, NIPG, and IIPG (not shown) are very similar to those of SIPG. An interesting experimental observation, which is not covered in previous theoretical sections, is that the DG methods with polynomials of odd orders have better performance than those of even orders; this is especially pronounced for OBB-DG.

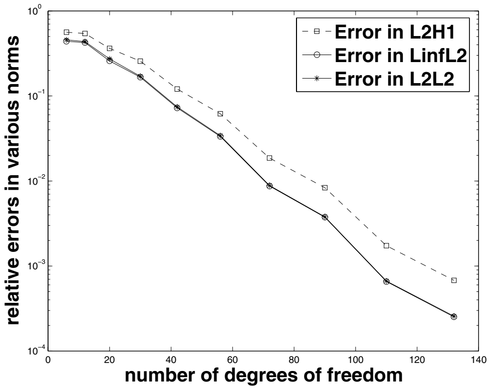


FIG. 7.3. *Convergence of p-refinement for SIPG.*

8. Discussion.

8.1. Penalty parameters for SIPG. Numerical experiments indicate that careful implementations of the penalty terms are crucial to SIPG: not only are the penalty terms necessary for the convergence of SIPG, but also choices of penalty parameters significantly influence the performance of SIPG. Small penalty parameters might result in divergences of the schemes. On the other hand, very large parameters, though ensuring the convergence theoretically, lead to a poor condition number for the resultant linear system, causing numerical difficulties in practice.

Reinvestigating (4.5), we see that it is sufficient to choose $\sigma_\gamma = O(|\mathbf{D}|^{1/2})$, where $|\cdot|$ is a matrix norm. Letting $\sigma_\gamma = \widehat{\sigma}|\mathbf{D}|^{1/2}$ and $\widehat{\sigma} = O(1)$, we have

$$J_0^\sigma(c,w) = \sum_{\gamma\in\Gamma_h} \widehat{\sigma}\sqrt{|\mathbf{D}|}\frac{r^2}{h_\gamma}\int_\gamma [c][w].$$

For most cases, we recommend $\widehat{\sigma} = 1$. It is found that $\widehat{\sigma}$ chosen from (0.1, 10) works well for many test cases. For cases where aspect ratios are very high and/or dispersion-diffusion is highly anisotropic, it is found that the following choice generally gives better results:

$$J_0^\sigma(c,w) = \sum_{\gamma\in\Gamma_h} \widehat{\sigma}\sqrt{|\mathbf{D}\mathbf{n}_\gamma|}\frac{r^2}{h_{m,\gamma}}\int_\gamma [c][w],$$

where $h_{m,\gamma} = \min_{E:\gamma\in E}(\mathrm{meas}(E)/\mathrm{meas}(\gamma))$.

8.2. Reference versus physical polynomial spaces. In the definition (3.1) of the DG space $\mathcal{D}_r(\mathcal{E}_h)$, the local space $\mathbb{P}_r(E)$ is the set of polynomials defined over a physical element E, rather than a reference element $\widehat{E}$. This distinction is unnecessary when E is a triangle or tetrahedron because the transformation from $\widehat{E}$ to E is affine. But for a general quadrilateral or hexahedron, these two spaces are different. We apply DG methods to the test case in section 7 using the uniform p-refinement in the coarsest mesh. Figure 8.1 provides the error ratio $\eta = \|e_r\|_{L^2(0,T;L^2(\Omega))}/\|e_f\|_{L^2(0,T;L^2(\Omega))}$ during the p-refinement, where e_r and e_f denote the DG errors based on the reference and physical spaces, respectively. Clearly, DG solutions based on physical spaces are more accurate than those of reference spaces for high order approximations; this is more significant for OBB-DG than for other primal DGs. This observation suggests

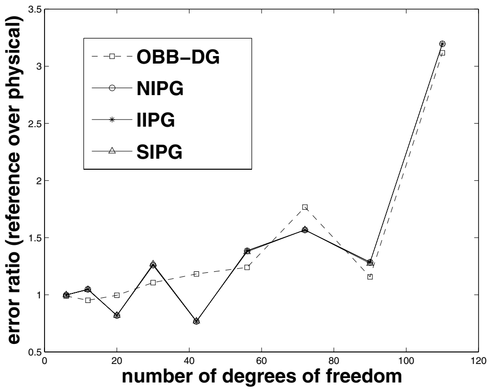


FIG. 8.1. *Comparison of reference versus physical polynomial spaces for DG methods (data of NIPG, SIPG, and IIPG are nearly identical).*

that physical polynomial spaces are preferred in p- and hp-implementations of DGs. It is also noted (not shown) that the improvement of physical over reference spaces is less pronounced on more refined meshes, because the transformation from $\widehat{E}$ to E becomes closer to an affine mapping. Consequently, a choice of physical versus reference spaces does not significantly impact h-versions of DGs.

9. Conclusions. Three primal DG methods with penalty have been analyzed for solving reactive transport problems in porous media. The cut-off operator was introduced in the DG formulations to ensure convergence for general nonlinear kinetic reactions. Error estimates in $L^2(H^1)$ for the concentration were derived for SIPG, NIPG, and IIPG, which are optimal in h and nearly optimal in p. In addition, we established $L^2(H^1)$ concentration error estimates on the element faces as well as $L^2(L^2)$ estimates for time derivatives. A parabolic lift technique for SIPG has been developed, which yields an h-optimal and nearly p-optimal error estimate in $L^2(L^2)$. The same lift technique applied to general linear functionals gives optimal estimates in negative norms. We have also numerically investigated the h- and p-convergence behaviors of OBB-DG, NIPG, IIPG, and SIPG. It was demonstrated that OBB-DG, IIPG, and NIPG possess h-optimal convergence rates in $L^2(H^1)$, but lack the optimality in $L^2(L^2)$ and $L^\infty(L^2)$, whereas SIPG performs h-optimally in the three norms. For smooth problems, exponential convergence rates in p are achieved by the four primal DG methods. In addition, it was observed that DGs with polynomials of odd orders perform better than those of even orders. Implementations of penalty terms are crucial to SIPG and a proper choice of the penalty parameter was proposed. Another important issue in implementations is the selection of physical versus reference spaces, for which we recommended the physical polynomial spaces for p- and hp-versions of DGs. As a future extension, we propose to study error estimates of primal DG methods for transport coupled with kinetic and local-equilibrium reactions and for multiphase flow in porous media.

REFERENCES

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

[2] T. ARBOGAST, S. BRYANT, C. DAWSON, F. SAAF, C. WANG, AND M. WHEELER, *Computational methods for multiphase flow and reactive transport problems arising in subsurface contaminant remediation,* J. Comput. Appl. Math., 74 (1996), pp. 19–32.

[3] D. N. Arnold, *An Interior Penalty Finite Element Method with Discontinuous Elements,* Ph.D. thesis, The University of Chicage, Chicago, IL, 1979.

[4] D. N. Arnold, *An interior penalty finite element method with discontinuous elements,* SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[5] I. Babuska and M. Suri, *The optimal convergence rates of the p version of the finite element methos,* SIAM J. Numer. Anal., 24 (1987), pp. 750–776.

[6] I. Babuska and M. Suri, *The h-p version of the finite element method with quasi-uniform meshes,* RAIRO Model. Math. Anal. Numer., 21 (1987), pp. 199–238.

[7] R. C. Borden and P. B. Bedient, *Transport of dissolved hydrocarbons influenced by oxygen-limited biodegradation 1. theoretical development,* Water Resources Res., 22 (1986), pp. 1973–1982.

[8] S. L. Bryant and K. E. Thompson, *Theory, modeling and experiment in reactive transport in porous media,* Curr. Opin. Colloid Interface Sci., 6 (2001), pp. 217–222.

[9] Z. Chen and H. Chen, *Pointwise error estimates of discontinuous Galerkin methods with penalty for second-order elliptic problems,* SIAM J. Numer. Anal., 42 (2004), pp. 1146–1166.

[10] C. Y. Chiang, C. N. Dawson, and M. F. Wheeler, *Modeling of in-situ biorestoration of organic compounds in groundwater,* Transp. Porous Media, 6 (1991), pp. 667–702.

[11] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, *The development of the discontinuous Galerkin methods,* in First International Symposium on Discontinuous Galerkin Methods, Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, New York, 2000, pp. 3–50.

[12] C. Dawson, S. Sun, and M. F. Wheeler, *Compatible algorithms for coupled flow and transport,* Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2565–2580.

[13] C. N. Dawson, T. F. Russell, and M. F. Wheeler, *Some improved error estimates for the modified method of characteristics,* SIAM J. Numer. Anal., 26 (1989), pp. 1487–1512.

[14] P. Engesgaard and K. L. Kipp, *A geochemical transport model for redox-controlled movement of mineral fronts in groundwater flow systems: A case of nitrate removal by oxidation of pyrite,* Water Resources Res., 28 (1992), pp. 2829–2843.

[15] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems,* SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.

[16] J. S. Kindred and M. A. Celia, *Contaminant transport and biodegradation* 2. *Conceptual model and test simulations,* Water Resources Res., 25 (1989), pp. 1149–1159.

[17] O. A. Ladyzenskaja, V. A. Solonnikov, and N. N. Uralceva, *Linear and Quasilinear Equations of Parabolic Type,* Transl. Math. Monogr. 23, American Mathematical Society, Providence, RI, 1968.

[18] M. G. Larson and A. J. Niklasson, *Analysis of a nonsymmetric discontinuous Galerkin method for elliptic problems: Stability and energy error estimates,* SIAM J. Numer. Anal., 42 (2004), pp. 252–264.

[19] J. T. Oden, I. Babuška, and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems,* J. Comput. Phys., 146 (1998), pp. 491–516.

[20] O. J. Palmer, *Error Estimates for Finite Element Methods Applied to Contaminant Transport Equations,* Ph.D. thesis, Rice University, Houston, 1983.

[21] B. Rivière, *Discontinuous Galerkin Finite Element Methods for Solving the Miscible Displacement Problem in Porous Media,* Ph.D. thesis, The University of Texas at Austin, 2000.

[22] B. Rivière and M. F. Wheeler, *Non conforming methods for transport with nonlinear reaction,* Contemp. Math., 295 (2002), pp. 421–432.

[23] B. Rivière, M. F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems,* SIAM J. Numer. Anal., 39 (2001), pp. 902–931.

[24] J. Rubin, *Transport of reacting solutes in porous media: Relation between mathematical nature of problem formulation and chemical nature of reactions,* Water Resources Res., 19 (1983), pp. 1231–1252.

[25] D. Schötzau and C. Schwab, *Time discretization of parabolic problems by the hp-version of the discontinuous Galerkin finite element method,* SIAM J. Numer. Anal., 38 (2001), pp. 837–875.

[26] D. Schötzau, C. Schwab, and A. Toselli, *Stabilized hp-dgfem for incompressible flow,* Math. Models Methods Appl. Sci., 13 (2003), pp. 1413–1436.

[27] Ch. Schwab, *p- and hp-Finite Element Methods, Theory and Applications in Solid and Fluid Mechanics,* Oxford University Press, Oxford, UK, 1998.

[28] C. I. Steefel and P. Van Cappellen, *Special issue: Reactive transport modeling of natural systems,* J. Hydrol., 209 (1998), pp. 1–388.

[29] S. Sun, *Discontinuous Galerkin Methods for Reactive Transport in Porous Media,* Ph.D. thesis, The University of Texas at Austin, 2003.

[30] S. Sun, B. Rivière, and M. F. Wheeler, *A combined mixed finite element and discontinuous Galerkin method for miscible displacement problems in porous media,* in Recent Progress in Computational and Applied PDEs, Conference Proceedings for the International Conference Held in Zhangjiaje in July 2001, pp. 321–348.

[31] S. Sun and M. F. Wheeler, *Anisotropic and dynamic mesh adaptation for discontinuous Galerkin methods applied to reactive transport,* Comput. Methods Appl. Mech. Engrg., to appear.

[32] S. Sun and M. F. Wheeler, *A posteriori error estimation and dynamic adaptivity for symmetric discontinuous Galerkin approximations of reactive transport problems,* Comput. Methods Appl. Mech. Engrg., to appear.

[33] S. Sun and M. F. Wheeler, *Mesh adaptation strategies for discontinuous Galerkin methods applied to reactive transport problems,* in Proceedings of International Conference on Computing, Communications and Control Technologies (CCCT 2004), Vol. I, H.-W. Chu, M. Savoie, and B. Sanchez, eds., 2004, pp. 223–228.

[34] S. Sun and M. F. Wheeler, *Discontinuous Galerkin methods for coupled flow and reactive transport problems,* Appl. Numer. Math., 52 (2005), pp. 273–298.

[35] S. Sun and M. F. Wheeler, *$L^2(H^1)$ norm a posteriori error estimation for discontinuous Galerkin approximations of reactive transport problems,* J. Sci. Comput., 22 (2005), pp. 501–530.

[36] S. Sun and M. F. Wheeler, *A dynamic, adaptive, locally conservative and nonconforming solution strategy for transport phenomena in chemical engineering,* in Proceedings of American Institute of Chemical Engineers 2004 Annual Meeting, Austin, Texas, 2004.

[37] J. van der Lee and L. De Windt, *Present state and future directions of modeling of geochemistry in hydrogeological systems,* J. Contam. Hydrol., 47/2 (2000), pp. 265–282.

[38] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties,* SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

[39] M. F. Wheeler, S. Sun, O. Eslinger, and B. Rivière, *Discontinuous Galerkin method for modeling flow and reactive transport in porous media,* in Analysis and Simulation of Multifield Problem, W. Wendland, ed., Springer-Verlag, Berlin, 2003, pp. 37–58.

[40] G. T. Yeh and V. S. Tripathi, *A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components,* Water Resources Res., 25 (1989), pp. 93–108.

[41] G. T. Yeh and V. S. Tripathi, *A model for simulating transport of reactive multispecies components: Model development and demonstration,* Water Resources Res., 27 (1991), pp. 3075–3094.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 220–238

STABILITY AND CONVERGENCE OF FINITE-ELEMENT APPROXIMATION SCHEMES FOR HARMONIC MAPS[*]

SÖREN BARTELS[†]

Abstract. This article discusses stability and convergence of approximation schemes for harmonic maps. A finite-element discretization of an iterative algorithm due to F. Alouges is introduced and shown to be stable and convergent in general only on acute-type triangulations. An a posteriori criterion is proposed which allows us to monitor sufficient conditions for weak convergence to a harmonic map on general triangulations and for adaptive mesh refinement. Numerical experiments show that an adaptive strategy automatically refines triangulations in neighborhoods of typical point singularities and thereby underline its efficiency.

Key words. harmonic maps, iterative algorithm, finite element method, weak convergence, liquid crystals, adaptive refinement

AMS subject classifications. 35A40, 65C20, 65N30

DOI. 10.1137/040606594

1. Introduction.

A variational model in the theory of nematic liquid crystals due to Oseen and Frank [29, 22, 13] leads to a minimization of the energy functional

$$I(v) := \frac{1}{2}\int_\Omega k_1|\operatorname{div} v|^2 + k_2|v\cdot\operatorname{curl} v|^2 + k_3|v\times\operatorname{curl} v|^2 + (k_2+k_4)\big(\operatorname{tr}\big[(Dv)^2\big] - (\operatorname{div} v)^2\big)\,\mathrm{d}x$$

over a space of admissible configurations

$$v \in \mathcal{A}(u_{\mathrm{D}}) := \big\{v \in H^1\big(\Omega;\mathbb{R}^3\big) : v|_{\partial\Omega} = u_{\mathrm{D}},\ |v| = 1 \text{ a.e. in } \Omega\big\}.$$

Here, $\Omega \subseteq \mathbb{R}^3$ is a bounded Lipschitz domain and represents the physical domain in which the liquid crystal is embedded, $u_{\mathrm{D}} \in H^{1/2}(\partial\Omega;\mathbb{R}^3)$ with $|u_{\mathrm{D}}| = 1$ almost everywhere on $\partial\Omega$ are given boundary data, and $k_1, k_2, k_3, k_4 \geq 0$ are material- and temperature-dependent constants. A vector field $v \in \mathcal{A}(u_{\mathrm{D}})$ locally represents the mean direction of the molecules that constitute the liquid crystal and a local minimizer of I in $\mathcal{A}(u_{\mathrm{D}})$ defines a stable configuration of the liquid crystal. The pointwise constraint $|v| = 1$ models the physically motivated assumption that in the liquid crystal phase the molecules are rod like with a fixed length.

Existence of (global) minimizers of I in $\mathcal{A}(u_{\mathrm{D}})$ can be established if $\mathcal{A}(u_{\mathrm{D}}) \neq \emptyset$ [14]. Sufficient for $\mathcal{A}(u_{\mathrm{D}}) \neq \emptyset$ is that u_{D} is Lipschitz continuous on $\partial\Omega$ [14]. Owing to the nonconvex constraint $|v| = 1$ uniqueness and higher regularity of solutions cannot be expected [7, 14, 15, 25, 26, 27, 28, 29]. Typically, stable points of I in $\mathcal{A}(u_{\mathrm{D}})$ are not continuous and have point singularities, which correspond to defects in the nematic material. In addition to nonuniqueness and existence of singularities, the nonconvex nature of the problem makes it extremely difficult to numerically approximate stationary points. The crux in the design of numerical schemes lies in a stable realization of the constraint $|v| = 1$. In order to make the main ideas for the approximation of

[*] Received by the editors April 12, 2004; accepted for publication (in revised form) March 16, 2005; published electronically May 27, 2005. This work was supported by Deutsche Forschungsgemeinschaft through the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin, Germany. The work was also supported by a fellowship within the Postdoc-Programme of the German Academic Exchange Service (DAAD).

http://www.siam.org/journals/sinum/43-1/60659.html

[†] Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (sba@math.hu-berlin.de).

the constraint more clear we will only investigate the physically relevant one-constant approximation of I, which assumes $k_1 = k_2 = k_3 = 1$ and $k_4 = 0$ and reduces the minimization problem to the problem of finding harmonic maps

$$\text{(P)} \qquad \begin{cases} \text{Find } u \in \mathcal{A}(u_{\mathrm{D}}) \text{ which is a local minimizer for} \\ E : \mathcal{A}(u_{\mathrm{D}}) \to \mathbb{R},\ v \mapsto \dfrac{1}{2}\displaystyle\int_\Omega |Dv|^2 \, \mathrm{d}x. \end{cases}$$

Solutions of (P) will be called *harmonic maps.* They satisfy the Euler–Lagrange equations

$$-\Delta u = |Du|^2 u, \quad |u| = 1 \qquad \text{in } \Omega.$$

Iterative algorithms for the approximation of harmonic maps have been proposed in [10, 19, 9] and successfully been tested numerically. Convergence of an iterative scheme on a continuous level and stability of a related finite difference discretization have been proved in [1]. The goal of this work is to analyze finite-element discretizations of that algorithm which allows for local mesh refinement and thereby a more efficient resolution of point singularities of solutions. We prove that, in general, finite-element discretizations cannot be expected to be stable and are convergent only on structured triangulations. Sufficient for stability and convergence is that the underlying triangulation is of acute type (cf. Lemma 3.2 for details). We provide an a posteriori criterion that allows us to monitor reliability of the algorithm on general triangulations and gives rise to automatic local mesh refinement. Numerical experiments indicate that adaptive strategies are more efficient when compared to schemes on uniform triangulations. While we restrict the analysis to the one-constant approximation of I we stress that the ideas can be carried over to the full model and refer the reader to [2] for related ideas.

An alternative approach to approximating local minimizers of I consists in regularizing the problem by introducing a penalty term $\varepsilon^{-2}\||v|^2 - 1\|^2_{L^2(\Omega)}$ in I with $0 < \varepsilon \ll 1$ in order to approximate the constraint $|v| = 1$. Difficulties in analyzing such an approach stem from the lack of regularity of minimizers of I and a reliable discretization of the gradient flow of the penalized formulation generally requires very small time step sizes, which limit the practical use. For related approaches and the numerical analysis of more sophisticated models we refer the reader to [3, 4, 20, 21, 23, 12].

The rest of this article is organized as follows. We briefly recall the definition and the main properties of the iterative algorithm of [1] in Section 2. Section 3 discusses finite-element discretizations of that algorithm and gives sufficient a priori conditions for its convergence. A few numerical experiments show the efficiency of the discrete algorithm and are presented in Section 4. Section 5 is devoted to an a posteriori analysis and introduces local refinement indicators. The efficient performance of the resulting adaptive strategy is illustrated by some numerical experiments in Section 6.

2. Alouges' algorithm.

For an initial $u^{(0)} \in \mathcal{A}(u_{\mathrm{D}})$ Alouges' algorithm computes a sequence $\big(u^{(j)} : j \in \mathbb{N}\big) \subseteq H^1(\Omega; \mathbb{R}^3)$ by iterating the following two steps:

(A_1) Let $w^{(j)} \in K_{u^{(j)}}$ satisfy $E\big(u^{(j)} - w^{(j)}\big) \le E\big(u^{(j)} - v\big)$ for all $v \in K_{u^{(j)}}$, where $K_{u^{(j)}} := \big\{v \in H^1\big(\Omega; \mathbb{R}^3\big) : v|_{\partial\Omega} = 0,\ v \cdot u^{(j)} = 0 \text{ a.e. in } \Omega\big\}$.

(A_2) Set $u^{(j+1)} := \dfrac{u^{(j)} - w^{(j)}}{|u^{(j)} - w^{(j)}|}$.

Given any $u^{(j)} \in H^1(\Omega;\mathbb{R}^3)$ step (A$_1$) consists in minimizing an elliptic functional on a subspace of $H^1(\Omega;\mathbb{R}^3)$ and admits a unique solution $w^{(j)} \in K_{u^{(j)}}$. Supposing that $|u^{(j)}| = 1$ almost everywhere in Ω, the definition of $K_{u^{(j)}}$ yields

$$|u^{(j)} - w^{(j)}|^2 = |u^{(j)}|^2 + |w^{(j)}|^2 \geq 1$$

almost everywhere in Ω. Hence, $u^{(j+1)}$ in step (A$_2$) is well defined and satisfies $|u^{(j+1)}| = 1$ almost everywhere in Ω. It is not difficult to verify that for a function $v \in H^1(\Omega;\mathbb{R}^3)$ satisfying $|v| \geq 1$ almost everywhere in Ω there holds

$$E\biggl(\frac{v}{|v|}\biggr) \leq E(v), \tag{2.1}$$

in particular $v/|v| \in H^1(\Omega;\mathbb{R}^3)$ and thus $u^{(j+1)} \in \mathcal{A}(u_{\mathrm{D}})$. Noting that $v \equiv 0 \in K_{u^{(j)}}$ it thus follows that

$$E\bigl(u^{(j+1)}\bigr) = E\biggl(\frac{u^{(j)} - w^{(j)}}{|u^{(j)} - w^{(j)}|}\biggr) \leq E\bigl(u^{(j)} - w^{(j)}\bigr) \leq E\bigl(u^{(j)}\bigr).$$

This is the energy decreasing property of Alouges' algorithm. The main features of the iteration are summarized in the following theorem.

Theorem 2.1 (see [1]). *Let $u^{(0)} \in \mathcal{A}(u_{\mathrm{D}})$. Suppose that the sequences $\bigl(u^{(j)} : j \in \mathbb{N}\bigr)$ and $\bigl(w^{(j)} : j \in \mathbb{N}\bigr)$ are generated by iterating* (A$_1$) *and* (A$_2$)*. Then, for all $j \in \mathbb{N}$ there holds*

$$u^{(j)} \in \mathcal{A}(u_{\mathrm{D}}) \quad \textit{and} \quad E\bigl(u^{(j+1)}\bigr) \leq E\bigl(u^{(j)}\bigr).$$

There holds $w^{(j)} \to 0$ (strongly) in H^1 and there exists a subsequence $(u^{(k)} : k \in \mathbb{N})$ and a harmonic map $u^* \in \mathcal{A}(u_{\mathrm{D}})$ such that $u^{(k)} \rightharpoonup u^*$ (weakly) in H^1.

3. Finite-element discretization and numerical analysis of (A$_1$) and (A$_2$).

In order to make difficulties in a finite-element discretization of (A$_1$) and (A$_2$) more clear we will occasionally consider a two-dimensional situation in this section. Therefore, we assume that $n = 2$ or $n = 3$ and that Ω is a bounded, polygonal or polyhedral, respectively, Lipschitz domain in $\mathbb{R}^n$. Given a regular triangulation $\mathcal{T}$ of Ω into triangles ($n = 2$) or tetrahedra ($n = 3$), let $\mathcal{N}$ denote the set of nodes in $\mathcal{T}$. The lowest order finite-element space related to $\mathcal{T}$ is denoted by $\mathcal{S}^1(\mathcal{T}) \subseteq H^1(\Omega)$. The nodal basis functions $(\varphi_z : z \in \mathcal{N}) \subseteq \mathcal{S}^1(\mathcal{T})$ satisfy $\varphi_z(z) = 1$ and $\varphi_z(z') = 0$ for $z \in \mathcal{N}$ and $z' \in \mathcal{N}\backslash\{z\}$. We define $\mathcal{S}^1_0(\mathcal{T}) := \{v_h \in \mathcal{S}^1(\mathcal{T}) : v_h|_{\partial\Omega} = 0\}$. Throughout this section, m is a positive integer.

The pointwise constraint $|v_h| = 1$ is satisfied solely by functions $v_h \in \mathcal{S}^1(\mathcal{T})^m$, which are constant in Ω. Therefore, assuming $u_{\mathrm{D}} \in C(\partial\Omega;\mathbb{R}^m)$, a reasonable finite-element discretization of (P) replaces $\mathcal{A}(u_{\mathrm{D}})$ by the set

$$\mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}}) := \bigl\{v_h \in \mathcal{S}^1(\mathcal{T})^m : \forall z \in \mathcal{N} \cap \partial\Omega\, v_h(z) = u_{\mathrm{D}}(z), \forall z \in \mathcal{N}\, |v_h(z)| = 1\bigr\}$$

and seeks a local minimizer of E in $\mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$

$$(\mathrm{P}_h) \qquad \begin{cases} \text{Find } u_h \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}}), \text{ which is a local} \\ \text{minimizer for } E \text{ in } \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}}). \end{cases}$$

Existence of solutions for the finite dimensional problem (P$_h$) follows from compactness arguments. The computation of a solution, however, is not obvious.

We propose a discrete version of Alouges' algorithm and state sufficient conditions for its stability and convergence.

Algorithm (A_h). Input: $(\mathcal{T}, u_h^{(0)}, \delta)$, where $\mathcal{T}$ is a regular triangulation of Ω, $u_h^{(0)} \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$ is a starting value, and $\delta > 0$ is a termination parameter.

(a) Set $j := 0$.

(b) Solve the optimization problem

$$\begin{cases} \text{Minimize} & E\big(u_h^{(j)} - v_h\big) \\ \text{subject to} & v_h \in \mathcal{S}_0^1(\mathcal{T})^m \;\text{ and }\; v_h(z) \cdot u_h^{(j)}(z) = 0 \text{ for all } z \in \mathcal{N}. \end{cases}$$

Denote the solution by $w_h^{(j)}$.

(c) If $\|Dw_h^{(j)}\|_{L^2(\Omega)} \le \delta$ set $(u_h, w_h) := (u_h^{(j)}, w_h^{(j)})$ and stop.

(d) Define

$$u_h^{(j+1)} := \sum_{z \in \mathcal{N}} \frac{u_h^{(j)}(z) - w_h^{(j)}(z)}{|u_h^{(j)}(z) - w_h^{(j)}(z)|} \varphi_z.$$

(e) Set $j := j + 1$ and go to (b).

Output: $(u_h, w_h) \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}}) \times \mathcal{S}_0^1(\mathcal{T})^m$.

A discrete version of (2.1) is necessary for stability of step (d) in the algorithm. It will turn out that such an estimate in general only holds on acute-type triangulations.

3.1. Validity and possible failure of an energy decreasing property of $(\mathbf{A}_h)$.

The following definition gives a sufficient criterion for stability of step (d) in Algorithm (A_h).

Definition 3.1. *A regular triangulation $\mathcal{T}$ of Ω is said to satisfy* an energy decreasing condition (ED) *if for all $v_h \in \mathcal{S}^1(\mathcal{T})^m$ satisfying $|v_h(z)| \ge 1$ for all $z \in \mathcal{N}$, $|v_h(z)| = 1$ for all $z \in \mathcal{N} \cap \partial\Omega$, and $w_h \in \mathcal{S}^1(\mathcal{T})^m$ defined by*

$$w_h := \sum_{z \in \mathcal{N}} \frac{v_h(z)}{|v_h(z)|} \varphi_z$$

there holds

$$E(w_h) \le E(v_h).$$

The next lemma implies that acute-type triangulations [17] allow for condition (ED).

Lemma 3.2. *Let $\mathcal{T}$ be a regular triangulation of Ω and suppose that $\int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x \le 0$ for all $z \in \mathcal{N} \backslash \partial\Omega$ and $y \in \mathcal{N} \backslash \{z\}$. Then $\mathcal{T}$ satisfies condition* (ED).

Proof. For $z, y \in \mathcal{N}$ set $k_{zy} := \int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x$. Let $\phi_h \in \mathcal{S}^1(\mathcal{T})^m$ and define $\phi_z := \phi_h(z)$ for all $z \in \mathcal{N}$. Since $\sum_{y \in \mathcal{N}} k_{zy} = 0$ for all $z \in \mathcal{N}$ and since $k_{zy} = k_{yz}$ for all $z, y \in \mathcal{N}$ we have

$$\begin{aligned} \|\nabla\phi_h\|_{L^2(\Omega)}^2 &= \sum_{z,y \in \mathcal{N}} k_{zy} \phi_z \cdot \phi_y = \sum_{z,y \in \mathcal{N}} k_{zy} \phi_z \cdot (\phi_y - \phi_z) \\ &= \frac{1}{2} \sum_{z,y \in \mathcal{N}} k_{zy} \phi_z \cdot (\phi_y - \phi_z) + \frac{1}{2} \sum_{z,y \in \mathcal{N}} k_{zy} \phi_y \cdot (\phi_z - \phi_y) = -\frac{1}{2} \sum_{z,y \in \mathcal{N}} k_{zy} |\phi_z - \phi_y|^2. \end{aligned}$$

Suppose that v_h and w_h are as in Definition 3.1 and let $v_z := v_h(z)$ and $w_z := w_h(z)$ for $z \in \mathcal{N}$. Let $z, y \in \mathcal{N}$ be such that $z \neq y$. If $z \in \mathcal{N}\backslash\partial\Omega$ or $y \in \mathcal{N}\backslash\partial\Omega$ we have $k_{zy} \leq 0$ and hence by Lipschitz continuity of the mapping $\{s \in \mathbb{R}^m : |s| \geq 1\} \to \mathbb{R}^m$, $s \mapsto s/|s|$, with Lipschitz constant 1 that

$$-\frac{1}{2}k_{zy}|w_z - w_y|^2 = -\frac{1}{2}k_{zy}\left|\frac{w_z}{|w_z|} - \frac{w_y}{|w_y|}\right|^2 \leq -\frac{1}{2}k_{zy}|v_z - v_y|^2.$$

If $z, y \in \mathcal{N} \cap \partial\Omega$ we have $w_z = v_z$ and $w_y = v_y$ and hence

$$\|\nabla w_h\|^2_{L^2(\Omega)} = -\frac{1}{2}\sum_{z,y\in\mathcal{N}} k_{zy}|w_z - w_y|^2 \leq -\frac{1}{2}\sum_{z,y\in\mathcal{N}} k_{zy}|v_z - v_y|^2 = \|\nabla v_h\|^2_{L^2(\Omega)},$$

which proves the lemma. □

Remarks 3.3. (i) *Suppose $n = 2$. Given neighboring nodes $z \in \mathcal{N}\backslash\partial\Omega$ and $y \in \mathcal{N}\backslash\{z\}$ let $T_1, T_2 \in \mathcal{T}$ be such that $T_1 \cap T_2$ equals the interior edge connecting z and y. Let $\alpha^{(1)}_{zy}$ and $\alpha^{(2)}_{zy}$ be the angles of T_1 and T_2, respectively, opposite to the edge connecting z and y. There holds* [11]

$$\int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x = -\mathrm{cot}\alpha^{(1)}_{zy} - \mathrm{cot}\alpha^{(2)}_{zy}.$$

Sufficient for $\int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x \leq 0$ is that $\alpha^{(1)}_{zy} + \alpha^{(2)}_{zy} \leq \pi$.

(ii) *Suppose $n = 3$ and let $z \in \mathcal{N}\backslash\partial\Omega$ and $y \in \mathcal{N}\backslash\{z\}$ be such that $z, y \in T$ for some $T \in \mathcal{T}$. Given any $T \in \mathcal{T}$ such that $z, y \in \mathcal{N} \cap T$ let α_{zyT} be the angle between the two faces $F^{(1)}_{zyT}, F^{(2)}_{zyT} \subseteq \partial T$, which do not contain both z and y. There holds* [17]

$$t\int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x = -\sum_{T\in\mathcal{T},\, z,y\in\mathcal{N}\cap T} \frac{|F^{(1)}_{zyT}|\,|F^{(2)}_{zyT}|}{9|T|}\cos\alpha_{zyT},$$

where $|F^{(\ell)}_{zyT}|$ is the surface measure of $F^{(\ell)}_{zyT}$ for $\ell = 1, 2$ and $|T|$ denotes the volume of T. Sufficient for $\int_\Omega \nabla\varphi_z \cdot \nabla\varphi_y \,\mathrm{d}x \leq 0$ is that $\alpha_{zyT} \leq \pi/2$ for all $T \in \mathcal{T}$ such that $z, y \in \mathcal{N} \cap T$.

The conditions of Remark 3.3 are sharp in the sense of the following example.

Example 3.4. *Let $0 < \beta < 1/2$ and $\Omega := (0,1) \times (0,\beta)$. Let*

$$\begin{aligned}
&z_1 := (0,0), && z_2 := (1/2,0), && z_3 := (1,0), && z_4 := (0,\beta),\\
&z_5 := (1/2,\beta), && z_6 := (1,\beta), && z_7 := (1/4,\beta/2), && z_8 := (3/4,\beta/2)
\end{aligned}$$

and $\mathcal{T} := \{T_1, T_2, \ldots, T_8\}$ be defined through

$$\begin{aligned}
&T_1 := \mathrm{conv}\{z_1, z_2, z_7\}, && T_2 := \mathrm{conv}\{z_2, z_8, z_7\}, && T_3 := \mathrm{conv}\{z_2, z_3, z_8\},\\
&T_4 := \mathrm{conv}\{z_3, z_6, z_8\}, && T_5 := \mathrm{conv}\{z_8, z_6, z_5\}, && T_6 := \mathrm{conv}\{z_7, z_8, z_5\},\\
&T_7 := \mathrm{conv}\{z_7, z_5, z_4\}, && T_8 := \mathrm{conv}\{z_1, z_7, z_4\};
\end{aligned}$$

cf. Figure 3.1. Define $s := 1/2 - \beta$ and set $v_j := (1,0)$ for $j = 1, 2, \ldots, 6$, $v_7 := (-1,0)$, and $v_8 := (1,-s)$. Let $v_h, w_h \in \mathcal{S}^1(\mathcal{T})^2$ be such that $v_h(z_j) = v_j$ and $w_h(z_j) = w_j := v_j/|v_j|$ for $j = 1, 2, \ldots, 8$. There holds

$$E(w_h) > E(v_h).$$

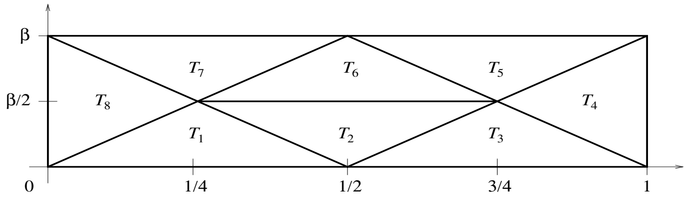


FIG. 3.1. *Triangulation $\mathcal{T}$ in Example* 3.4 *that does not satisfy condition* (ED) *for* $0 < \beta < 1/2$.

Proof. For $j, \ell = 1, 2, \dots, 8$ set $k_{j\ell} := \int_\Omega \nabla\varphi_{z_j} \cdot \nabla\varphi_{z_\ell}\,\mathrm{d}x$. Since $w_j = v_j$ for $j = 1, 2, \dots, 7$ we have (cf. the proof of Lemma 3.2)

$$\delta := \|\nabla v_h\|^2_{L^2(\Omega)} - \|\nabla w_h\|^2_{L^2(\Omega)} = -\frac{1}{2}\sum_{j,\ell=1}^{8} k_{j\ell}\big(|v_j - v_\ell|^2 - |w_j - w_\ell|^2\big)$$
$$= -\sum_{j=1}^{6} k_{j8}\big(|(1,0) - v_8|^2 - |(1,0) - w_8|^2\big) - k_{78}\big(|(-1,0) - v_8|^2 - |(-1,0) - w_8|^2\big).$$

We use $|(1,0) - v_8|^2 = s^2$, $|(-1,0) - v_8|^2 = 4 + s^2$ and abbreviate

$$\kappa_1^2 := |(1,0) - w_8|^2 = 2 - 2/\sqrt{1+s^2}, \quad \kappa_2^2 := |(-1,0) - w_8|^2 = 2 + 2/\sqrt{1+s^2}.$$

Using that $\sum_{j=1}^{8} k_{j8} = 0$ we verify $\sum_{j=1}^{6} k_{j8} = -k_{78} - k_{88}$ and obtain

$$\delta = \big(s^2 - \kappa_1^2\big)(k_{78} + k_{88}) - k_{78}\big(4 + s^2 - \kappa_2^2\big) = k_{88}\big(s^2 - \kappa_1^2\big) - k_{78}\big(4 + \kappa_1^2 - \kappa_2^2\big).$$

Elementary calculations show that

$$k_{88} = \big(12\beta^2 + 5\big)/(4\beta) \quad \text{and} \quad k_{78} = \big(1 - 4\beta^2\big)/(4\beta).$$

With a function ϕ such that $\sqrt{1+s^2} = 1 + \frac{1}{2}s^2 + \phi(s^2)$ we deduce

$$4\beta\sqrt{1+s^2}\delta = (12\beta^2 + 5)\left(\frac{1}{2}s^4 + s^2\phi(s^2) - 2\phi(s^2)\right) - (1 - 4\beta^2)(2s^2 + 4\phi(s^2)).$$

Using that $\beta^2 = \frac{1}{4} - s + s^2$ we verify

$$4\beta\sqrt{1+s^2}\delta = \big(8 - 12s + 12s^2\big)\left(\frac{1}{2}s^4 + s^2\phi(s^2) - 2\phi(s^2)\right) - 16(s - s^2)\left(\frac{1}{2}s^2 + \phi(s^2)\right)$$
$$= -8s^3 + 12s^4 - 6s^5 + 6s^6 + \phi(s^2)(-16 + 8s - 12s^3 + 12s^4)$$
$$= -6s^3(1 - 2s) - 6s^5(1 - s) + 4s\phi(s^2)(2 - 3s^2 + 3s^3) - 2(s^3 + 8\phi(s^2)).$$

Since $0 < s < 1/2$ and $\phi(s^2) < 0$, the first three terms on the right-hand side are negative. A Taylor expansion proves $-s^4/8 \le \phi(s^2)$ and implies that the last term on the right-hand side is nonpositive. This shows $\delta < 0$ and proves the lemma. ∎

We include another sufficient criterion for validity of condition (ED) that allows to construct triangulations of a large class of three-dimensional domains.

LEMMA 3.5. *Suppose $n = 3$ and assume that each $T \in \mathcal{T}$ has three mutually perpendicular edges. Then $\mathcal{T}$ satisfies condition* (ED).

Proof. Let $v_h, w_h \in \mathcal{S}^1(\mathcal{T})^m$ be as in Definition 3.1 and let $T \in \mathcal{T}$. Let $b_1, b_2, b_3 \in \mathbb{R}^3 \backslash \{0\}$ be mutually perpendicular and such that $b_j = z_j - y_j$, $j = 1, 2, 3$, where $z_j, y_j \in \mathcal{N} \cap T$ for $j = 1, 2, 3$. After an appropriate rotation we may assume $b_j = |b_j| e_j$ for $j = 1, 2, 3$, where e_j is the jth canonical basis vector in $\mathbb{R}^3$. Since $s \mapsto s/|s|$ for $s \in \mathbb{R}^m$ with $|s| \geq 1$ is Lipschitz continuous with Lipschitz constant 1 we deduce for $j = 1, 2, 3$ that

$$\left| \frac{\partial w_h|_T}{\partial x_j} \right| = \frac{1}{|z_j - y_j|} \left| \frac{v_h(z_j)}{|v_h(z_j)|} - \frac{v_h(y_j)}{|v_h(y_j)|} \right| \leq \frac{1}{|z_j - y_j|} |v_h(z_j) - v_h(y_j)| = \left| \frac{\partial v_h|_T}{\partial x_j} \right|,$$

which implies the lemma. □

REMARK 3.6. *It can be shown* [16] *that if $n = 3$ and each $T \in \mathcal{T}$ has* 3 *mutually perpendicular edges, which do not pass through the same vertex, then $\mathcal{T}$ is of acute type, i.e., satisfies the conditions of Remark* 3.3 (ii).

The following example defines a triangulation of the unit cube, which satisfies the conditions of Lemma 3.5. It allows to construct triangulations that satisfy condition (ED) of unions of finitely many quadrilaterals. Other constructions and acute-type refinement strategies of tetrahedra can be found in [17, 16].

EXAMPLE 3.7 (see [5]). *Set* $\mathcal{N} := \{z_1, z_2, \ldots, z_8\}$ *for*

$$z_1 := (0,0,0), \quad z_2 := (1,0,0), \quad z_3 := (0,0,1), \quad z_4 := (1,0,1),$$
$$z_5 := (0,1,0), \quad z_6 := (1,1,0), \quad z_7 := (0,1,1), \quad z_8 := (1,1,1).$$

Define $\mathcal{T} := \{T_1, T_2, \ldots, T_6\}$ *with*

$$T_1 := \mathrm{conv}\{z_1, z_2, z_3, z_6\}, \quad T_2 := \mathrm{conv}\{z_2, z_4, z_3, z_6\}, \quad T_3 := \mathrm{conv}\{z_3, z_4, z_8, z_6\},$$
$$T_4 := \mathrm{conv}\{z_3, z_8, z_7, z_6\}, \quad T_5 := \mathrm{conv}\{z_7, z_5, z_3, z_6\}, \quad T_6 := \mathrm{conv}\{z_3, z_5, z_1, z_6\}.$$

Then $\mathcal{T}$ is a regular triangulation of $(0,1)^3$ *and satisfies the assumptions of Lemma* 3.5*; cf. Figure* 3.2.

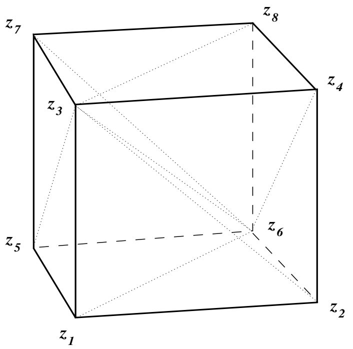


FIG. 3.2. *Triangulation $\mathcal{T}$ of the unit cube defined in Example* 3.7 *such that each element in $\mathcal{T}$ has three mutually perpendicular edges.*

3.2. Well posedness and termination of algorithm (A$_h$)**.** The following lemma shows that all steps in (A$_h$) are well defined and that the algorithm terminates within a finite number of iterations, provided that $\mathcal{T}$ satisfies condition (ED).

LEMMA 3.8. *Suppose $\mathcal{T}$ satisfies condition* (ED). *Given $\delta > 0$ and $u_h^{(0)} \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$, Algorithm* (A$_h$) *with input* $(\mathcal{T}, u_h^{(0)}, \delta)$ *terminates within a finite number*

M of iterations with output $(u_h, w_h) \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}}) \times \mathcal{S}^1_0(\mathcal{T})^m$ such that $\|Dw_h\|_{L^2(\Omega)} \leq \delta$ and

$$E(u_h) \leq E\big(u_h^{(0)}\big).$$

Proof. We proceed by induction to show $u_h^{(j)} \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$ and $E(u_h^{(j+1)}) \leq E(u_h^{(j)})$. Suppose that for some $j \geq 0$ we are given $u_h^{(j)} \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$. The set

$$L^{(j)} := \big\{ v_h \in \mathcal{S}^1_0(\mathcal{T})^m : \forall z \in \mathcal{N}\; v_h(z) \cdot u_h^{(j)}(z) = 0 \big\}$$

is a subspace of $\mathcal{S}^1_0(\mathcal{T})^m$. Hence, there exists $w_h \in L^{(j)}$ such that

$$\int_\Omega Dw_h : Dv_h \,\mathrm{d}x = \int_\Omega Du_h^{(j)} : Dv_h \,\mathrm{d}x \tag{3.1}$$

for all $v_h \in L^{(j)}$. This is equivalent to

$$E\big(u_h^{(j)} - w_h\big) \leq E\big(u_h^{(j)} - v_h\big)$$

for all $v_h \in L^{(j)}$. Thus, $w_h = w_h^{(j)}$ is the unique solution in step (b) of Algorithm (A$_h$). Since $w_h^{(j)}(z) \cdot u_h^{(j)}(z) = 0$ and $|u_h^{(j)}(z)| = 1$ there holds $|u_h^{(j)}(z) - w_h^{(j)}(z)| \geq 1$ for all $z \in \mathcal{N}$. Hence, $u_h^{(j+1)}$ is well defined and $u_h^{(j+1)} \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$. Since $0 \in L^{(j)}$ and $\mathcal{T}$ satisfies condition (ED) there holds $E(u_h^{(j+1)}) \leq E(u_h^{(j)} - w_h^{(j)}) \leq E(u_h^{(j)})$. Equation (3.1) with $v_h = w_h = w_h^{(j)}$ proves $E(u_h^{(j)} - w_h^{(j)}) = E(u_h^{(j)}) - E(w_h^{(j)})$ and a combination with the previous assertion shows

$$0 \leq E\big(w_h^{(j)}\big) \leq E\big(u_h^{(j)}\big) - E\big(u_h^{(j+1)}\big).$$

Since $\big(E(u_h^{(j)}) : j \in \mathbb{N}\big)$ is monotonically decreasing and bounded from below we conclude that it is a Cauchy sequence and hence $\|Dw_h^{(M)}\|_{L^2(\Omega)} \leq \delta$ for M sufficiently large. □

3.3. Convergence for $h \to 0$. The following theorem shows that for a sequence of triangulations with maximal mesh size tending to 0 the sequence of outputs of Algorithm (A$_h$) provides a weakly convergent subsequence whose weak limit is a harmonic map. The important questions whether this weak limit is (globally) energy minimizing in (P) or whether weak convergence can be improved to strong convergence are left for future research.

THEOREM 3.9. *Suppose $u_{\mathrm{D}} \in H^1(\partial\Omega; \mathbb{R}^3)$. Let $(\mathcal{T}_k : k \in \mathbb{N})$ be a sequence of regular triangulations of Ω satisfying condition* (ED) *with maximal mesh sizes $(h_k : k \in \mathbb{N})$ satisfying $h_k \to 0$ for $k \to \infty$ and let $(\delta_k : k \in \mathbb{N})$ be a sequence of positive numbers such that $\delta_k \to 0$ for $k \to \infty$. Suppose that $u_k^{(0)} \in \mathcal{A}_h(\mathcal{T}_k, u_{\mathrm{D}})$ and there exists $C_0 > 0$ such that*

$$\big\|Du_k^{(0)}\big\|_{L^2(\Omega)} \leq C_0$$

for all $k \in \mathbb{N}$. For each $k \in \mathbb{N}$, let (u_k, w_k) be the output of Algorithm (A$_h$) *applied to the input $(\mathcal{T}_k, u_k^{(0)}, \delta_k)$. Then there exists a subsequence $(u_\ell : \ell \in \mathbb{N})$ and a harmonic map $u^* \in \mathcal{A}(u_{\mathrm{D}})$ such that $u_\ell \rightharpoonup u^*$ (weakly) in H^1 and*

$$E(u^*) \leq \liminf_{\ell \to \infty} E(u_\ell).$$

The following lemma is essential in the proof of the theorem. For $c \in \mathbb{R}^3$ and a matrix $A \in \mathbb{R}^{3\times 3}$ with columns $a_1, a_2, a_3 \in \mathbb{R}^3$ we let $c \times A \in \mathbb{R}^{3\times 3}$ be the matrix whose columns equal $c \times a_j$ for $j = 1, 2, 3$.

LEMMA 3.10 (see [8]). *A function $u \in \mathcal{A}(u_{\rm D})$ is a harmonic map if and only if*

$$\int_\Omega \big(u \times Du\big) : D\phi \,\mathrm{d}x = 0 \tag{3.2}$$

for all $\phi \in H^1_0(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$.

Proof. Suppose that u is a harmonic map, i.e., for all $w \in H^1_0(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$ there holds

$$\int_\Omega Du : Dw \,\mathrm{d}x = \int_\Omega |Du|^2 u \cdot w \,\mathrm{d}x.$$

Let $\phi \in H^1_0(\Omega;\mathbb{R}^3) \cap L^\infty(\Omega;\mathbb{R}^3)$ and set $w := u \times \phi$. Using $(a \times b) \cdot c = -(a \times c) \cdot b$ for any $a, b, c \in \mathbb{R}^3$ we verify

$$Du : Dw = Du : \big(Du \times \phi + u \times D\phi\big) = Du : \big(u \times D\phi\big) = -\big(u \times Du\big) : D\phi$$

almost everywhere in Ω and since $u \cdot w = 0$ we find that u satisfies (3.2). Suppose now that u satisfies (3.2) and set $\phi := u \times w$. The identity $(a \times b) \cdot (c \times d) = (a \cdot c)(b \cdot d) - (b \cdot c)(a \cdot d)$ yields

$$\begin{aligned} D(u \times w) : \big(u \times Du\big) &= \big(Du \times w\big) : \big(u \times Du\big) + \big(u \times Dw\big) : \big(u \times Du\big) \\ &= \big(u^T Du\big) \cdot \big(w^T Du\big) - |Du|^2 u \cdot w + |u|^2 Dw : Du - \big(u^T Dw\big) \cdot \big(u^T Du\big) \end{aligned}$$

almost everywhere in Ω. The identity $|u|^2 = 1$ implies $u^T Du = 0$ almost everywhere in Ω. An integration over Ω finishes the proof of the lemma. □

Proof of Theorem 3.9. By Lemma 3.8 and the boundedness of $(u_k^{(0)})$ in H^1 there holds $\|Du_k\|_{L^2(\Omega)} \le \|Du_k^{(0)}\|_{L^2(\Omega)} \le C_0$ for all $k \in \mathbb{N}$. Hence, there exists a subsequence $(u_\ell : \ell \in \mathbb{N})$ and $u^* \in H^1(\Omega;\mathbb{R}^3)$ such that $u_\ell \rightharpoonup u^*$ (weakly) in H^1. Weak lower semicontinuity of E implies $E(u^*) \le \liminf_{\ell\to\infty} E(u_\ell)$. Since $|u_\ell(z)| = 1$ for all $z \in \mathcal{N}_\ell$ we have, by a $\mathcal{T}$ elementwise application of Poincaré's inequality and $|u_\ell| \le 1$ almost everywhere in Ω,

$$\|\, |u_\ell|^2 - 1 \,\|_{L^2(\Omega)} \le C_P h_\ell \big\| 2u_\ell^T Du_\ell \big\|_{L^2(\Omega)} \le 2 C_P C_0 h_\ell.$$

Since $u_\ell \to u^*$ almost everywhere in Ω we deduce $|u^*| = 1$ almost everywhere in Ω. Moreover, we have

$$\big\| u_\ell|_{\partial\Omega} - u_{\rm D} \big\|_{L^2(\partial\Omega)} \le C h_\ell \big\| \partial u_{\rm D}/\partial s \|_{L^2(\partial\Omega)}$$

(here, $\partial u_{\rm D}/\partial s$ denotes the surface gradient of $u_{\rm D}$ along $\partial\Omega$) and compactness of the trace operator as a mapping from $H^1(\Omega;\mathbb{R}^3)$ into $L^2(\partial\Omega;\mathbb{R}^3)$ (cf. [24] for details) implies $u^*|_{\partial\Omega} = u_{\rm D}$. It remains to show that u^* is a harmonic map. For all $\Psi_\ell \in \mathcal{S}^1_0(\mathcal{T}_\ell)^3$ with $\Psi_\ell(z) \cdot u_\ell(z) = 0$ for all $z \in \mathcal{N}_\ell$ there holds by definition of w_ℓ

$$\int_\Omega D(u_\ell - w_\ell) : D\Psi_\ell \,\mathrm{d}x = 0.$$

Given $\phi \in C_0^\infty(\Omega;\mathbb{R}^3)$ let $\Phi_\ell := \phi \times u_\ell$ and choose $\Psi_\ell := \mathcal{I}_\ell(\phi \times u_\ell)$, where $\mathcal{I}_\ell$ denotes the nodal interpolation operator on $\mathcal{T}_\ell$. We then have

$$\int_\Omega Du_\ell : D(\phi \times u_\ell)\,\mathrm{d}x = \int_\Omega D(u_\ell - w_\ell) : D(\Phi_\ell - \Psi_\ell)\,\mathrm{d}x + \int_\Omega Dw_\ell : D\Phi_\ell\,\mathrm{d}x. \tag{3.3}$$

Using

$$Du_\ell : D(\phi \times u_\ell) = Du_\ell : (D\phi \times u_\ell + \phi \times Du_\ell) = Du_\ell : (D\phi \times u_\ell) = D\phi : (u_\ell \times Du_\ell)$$

and $u_\ell \to u^*$ (strongly) in L^2, $Du_\ell \rightharpoonup Du^*$ (weakly) in H^1 we deduce

$$\int_\Omega Du_\ell : D(\phi \times u_\ell)\,\mathrm{d}x = \int_\Omega D\phi : (u_\ell \times Du_\ell)\,\mathrm{d}x \to \int_\Omega D\phi : (u^* \times Du^*)\,\mathrm{d}x. \tag{3.4}$$

Since u_ℓ is $\mathcal{T}$ elementwise affine there holds for each $T \in \mathcal{T}$

$$\begin{aligned}\|D(\Phi_\ell - \Psi_\ell)\|_{L^2(T)} &= \|D\big(\phi \times u_\ell - \mathcal{I}_\ell(\phi \times u_\ell)\big)\|_{L^2(T)} \le Ch_\ell \|D^2(\phi \times u_\ell)\|_{L^2(T)}\\ &\le Ch_\ell\big(\|D^2\phi\|_{L^2(T)} + \|D\phi\|_{L^\infty(\Omega)}\|Du_\ell\|_{L^2(T)}\big),\end{aligned}$$

and hence $\Phi_\ell - \Psi_\ell \to 0$ (strongly) in H^1. Notice that $u_\ell - w_\ell$ is uniformly bounded in H^1 so that

$$\int_\Omega D(u_\ell - w_\ell) : D(\Phi_\ell - \Psi_\ell)\,\mathrm{d}x \to 0. \tag{3.5}$$

Since Φ_ℓ is bounded in H^1 and $w_\ell \to 0$ (strongly) in H^1 we have

$$\int_\Omega Dw_\ell : D\Phi_\ell\,\mathrm{d}x \to 0. \tag{3.6}$$

A combination of (3.3)–(3.6) yields

$$\int_\Omega D\phi : (u^* \times Du^*)\,\mathrm{d}x = 0,$$

which, according to Lemma 3.10, shows that u^* is a harmonic map. □

4. Numerical experiments I.

In this section we report on some numerical experiments. We first discuss the implementation of Algorithm (A_h).

4.1. Uzawa iteration for the efficient solution of step (b).

Step (b) of Algorithm (A_h) requires the solution of a quadratic optimization problem with linear constraints. This can be solved directly, but may be inefficient. We thus propose the use of an Uzawa iteration. The optimization problem may be rewritten as a saddle point problem and the related optimality conditions read

$$(\mathrm{SP}_h) \qquad \begin{cases}\text{Find } w_h \in \mathcal{S}_0^1(\mathcal{T})^3 \text{ and } \lambda \in \mathbb{R}^{\mathcal{K}} \text{ such that, for all } v_h \in \mathcal{S}_0^1(\mathcal{T})^3,\\ \displaystyle\int_\Omega Dw_h : Dv_h\,\mathrm{d}x + \sum_{z\in\mathcal{K}} \lambda_z u_h^{(j)}(z)\cdot v_h(z) = \int_\Omega Du_h^{(j)} : Dv_h\,\mathrm{d}x,\\ w_h(z)\cdot u_h^{(j)}(z) = 0 \quad \text{for all } z \in \mathcal{K}.\end{cases}$$

Here, $\mathcal{K} := \mathcal{N} \cap \Omega$ denotes the set of free nodes in $\mathcal{N}$. The problem can be recast as

$$(\mathrm{SP}'_h) \qquad \begin{cases}\text{Find } x \in \mathbb{R}^{3N'} \text{ and } \lambda \in \mathbb{R}^{N'} \text{ such that}\\ A'x + B^T\lambda = b,\\ Bx = 0.\end{cases}$$

In this formulation, $x \in \mathbb{R}^{3N'}$ contains the values of w_h in the free nodes and we set $N' := \operatorname{card}(\mathcal{K})$. The constraint $u_h(z) \cdot w_h(z) = 0$, $z \in \mathcal{K}$, is realized by the matrix $B \in \mathbb{R}^{3N' \times N'}$. The positive definite matrix $A' \in \mathbb{R}^{3N' \times 3N'}$ is the restriction of A to $\mathcal{S}^1_0(\mathcal{T})^3$, where A is the stiffness matrix defined through the nodal basis in $\mathcal{S}^1(\mathcal{T})^3$. Finally, b is given by the restriction of $A\underline{u}$ to the free nodes, assuming that $\underline{u}$ contains the nodal values of $u_h^{(j)}$ in $\mathcal{N}$. The efficient iterative solution of (SP'_h) is realized by an Uzawa algorithm with conjugate directions and an LU decomposition of A' (cf., e.g., [6]).

4.2. Numerical examples. For the first numerical experiments we specify (P) in the following example.

EXAMPLE 4.1. *Set* $\Omega := (-1/2, 1/2)^3$ *and* $u_{\mathrm{D}}(x) := x/|x|$, $x \in \partial\Omega$. *Then,* $u(x) = x/|x|$, $x \in \Omega$, *is the unique solution of* (P) [18].

In order to satisfy the conditions that guarantee convergence in Theorem 3.9 we construct triangulations of Ω that satisfy condition (ED) by scaling, translating, and assembling copies of the triangulation $\mathcal{T}$ from Example 3.7.

EXAMPLE 4.2. *Given an integer* $k \geq 1$ *set* $h_k := 1/k$,

$$C_k := \big\{h_k(\ell, m, n) : 0 \leq \ell, m, n \leq k-1\big\} - (1,1,1)/2,$$

and define, with $\tilde{\mathcal{T}}$ *from Example* 3.7,

$$\mathcal{T}_k := \big\{c + h_k T : c \in C_k,\, T \in \mathcal{T}\big\}.$$

Then, $\mathcal{T}_k$ *is a regular triangulation of* $\Omega = (-1/2, 1/2)^3$ *with maximal mesh size* $\sqrt{3}/k$ *and satisfies condition* (ED).

We used four triangulations $\mathcal{T}_k$, specified through $k = 4, 8, 16, 32$ in Example 4.1, with $3N'_k = 375, 2187, 14739, 107811$ degrees of freedom (i.e., N'_k free nodes in $\mathcal{T}_k$). We set $\delta_k := 10^{-4}/\log_2(k)$ and define initial functions $u_k^{(0)} \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$ by

$$u_k^{(0)}(z) := \begin{cases} z/|z| & \text{for } z \in \mathcal{N}_k \cap \partial\Omega, \\ (0,1,0) & \text{for } z \in \mathcal{N}_k \cap \Omega. \end{cases}$$

In all experiments the Uzawa iteration was stopped when the ℓ^2 norm of the residual Bx in (SP'_h) was less than 10^{-6}. In most of the experiments this stopping criterion was satisfied after at most 20 iterations.

Figure 4.1 displays the decay of the energy $E(u_k^{(j)})$, $j = 1, 2, \ldots$, in the iteration of Algorithm (A_h) with input $(\mathcal{T}_k, u_k^{(0)}, \delta_k)$ for $k = 4, 8, 16, 32$. The plot shows that the decrease in the energy is largest for the first few iterations. This yields the conjecture that the choice of the termination criteria $\delta_k = 10^{-4}/\log_2 k$ is inefficient in this example if one is only interested in an asymptotic behavior for $h \to 0$.

Figure 4.2 shows the projection of the vector fields $u_{32}^{(j)}(0, \cdot, \cdot)$ obtained from Algorithm (A_h) onto $\{(x, y, z) \in \mathbb{R}^3 : x = 0\}$ in $(-1/2, 1/2)^2$ for $j = 0, 10, 50, 315$. We observe that only a few iterations are needed to rotate vectors in such a way that only one degree-1 singularity is present. The subsequent iterations move this singularity to the origin. After 317 iterations Algorithm (A_h) with input $(\mathcal{T}_{32}, u_{32}^{(0)}, \delta_{32})$ terminates and the nodal values of the output u_{32} appear to be very close to the exact solution away from 0. The value of the numerical solution at 0, where the exact solution has a singularity, has no particular meaning and seems to depend on the triangulation and the initial value.

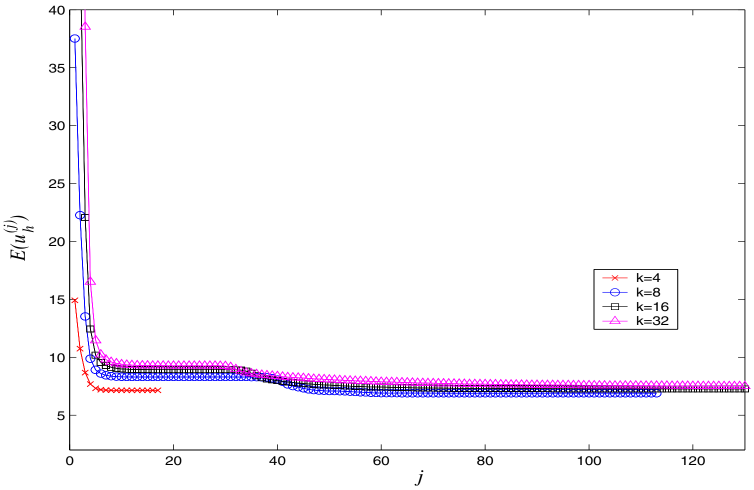


FIG. 4.1. *Decay of the energy in the iteration of Algorithm* (A_h) *on* $\mathcal{T}_k$ *with* $k = 4, 8, 16, 32$ *in Example* 4.1 *and initial* $u_k^{(0)} \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$.

We assume that our definition of $u_k^{(0)}$ is suboptimal as it admits large gradients in a neighborhood of $\partial\Omega$. In particular, this choice does not satisfy $\|Du_k^{(0)}\|_{L^2(\Omega)} \le C_0$ for $h_k \to 0$. However, even if for all $z \in \mathcal{K}$, $\xi(z)$ is a random unit vector in $\mathbb{R}^3$ and the starting value $\tilde{u}_k^{(0)} \in \mathcal{A}_h(\mathcal{T}, u_{\mathrm{D}})$ is defined by

$$\tilde{u}_k^{(0)}(z) := \begin{cases} z/|z| & \text{for } z \in \mathcal{N}_k \cap \partial\Omega, \\ \xi(z) & \text{for } z \in \mathcal{N}_k \cap \Omega, \end{cases}$$

then we observe in Figure 4.3 that the energy still decreases rapidly in the first iterations and becomes stationary almost as fast as for the previous choice. We assume that the number of iterations depends on the initial energy and can be reduced with an optimal choice of $u_k^{(0)}$. Indeed, the proof of Lemma 3.8 shows that the sequence of corrections $w_k^{(j)}$ satisfies for all $\ell \ge 0$

$$\sum_{j=0}^{\ell} \|Dw_k^{(j)}\|_{L^2(\Omega)}^2 \le \|Du_k^{(0)}\|_{L^2(\Omega)}^2,$$

and assuming that $\|Du_k^{(0)}\|_{L^2(\Omega)} \le C_0$ (for a k-independent constant $C_0 > 0$) then the number of iterations can be expected to grow less fast than in the presented experiments.

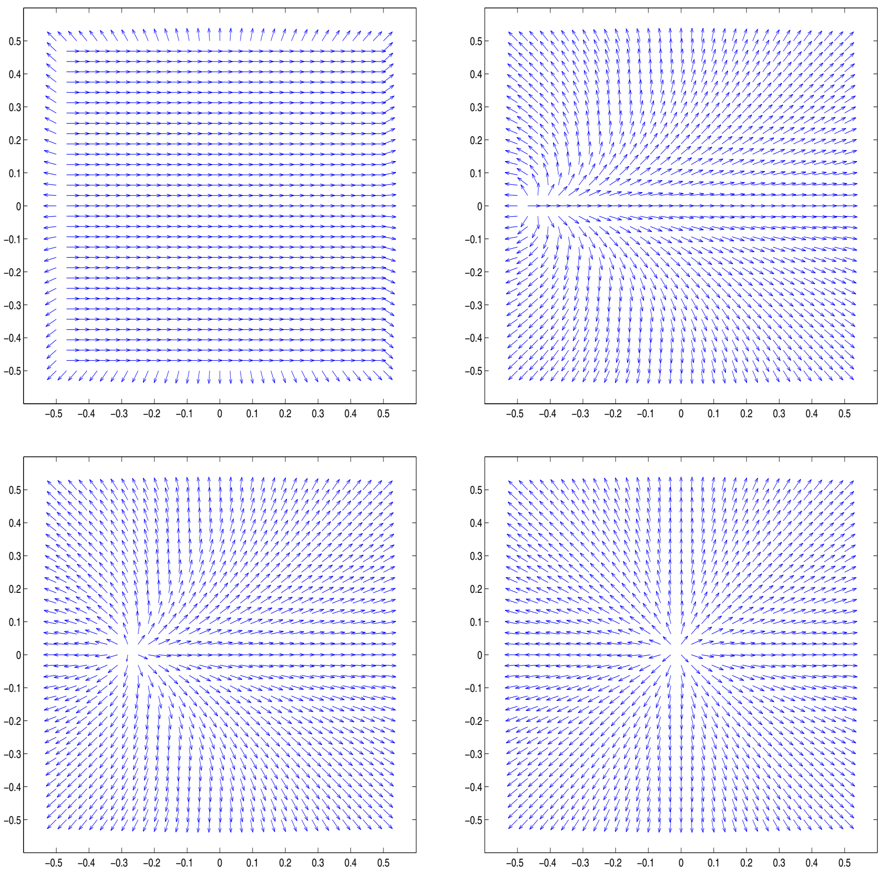


FIG. 4.2. *Projection of the vector fields* $u_{32}^{(j)}(0,\cdot,\cdot)$ *onto* $\{(x,y,z)\in\mathbb{R}^3 : x=0\}$ *in* $(-1/2,1/2)^2$ *for* $j=0,10,50,315$ *in Example* 4.1 *and initial* $u_k^{(0)}\in\mathcal{A}_h(\mathcal{T},u_{\mathrm{D}})$.

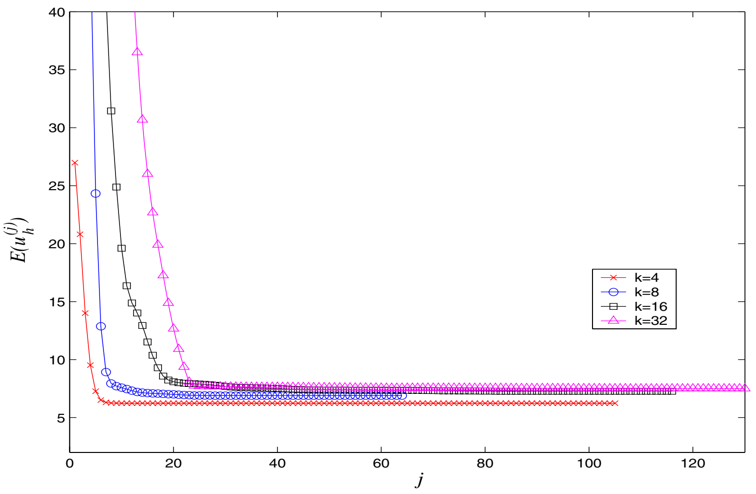


FIG. 4.3. *Decay of the energy in the iteration of Algorithm* (A_h) *on* $\mathcal{T}_k$ *with* $k=4,8,16,32$ *in Example* 4.1 *with random initial data* $\tilde{u}_k^{(0)}\in\mathcal{A}_h(\mathcal{T},u_{\mathrm{D}})$.

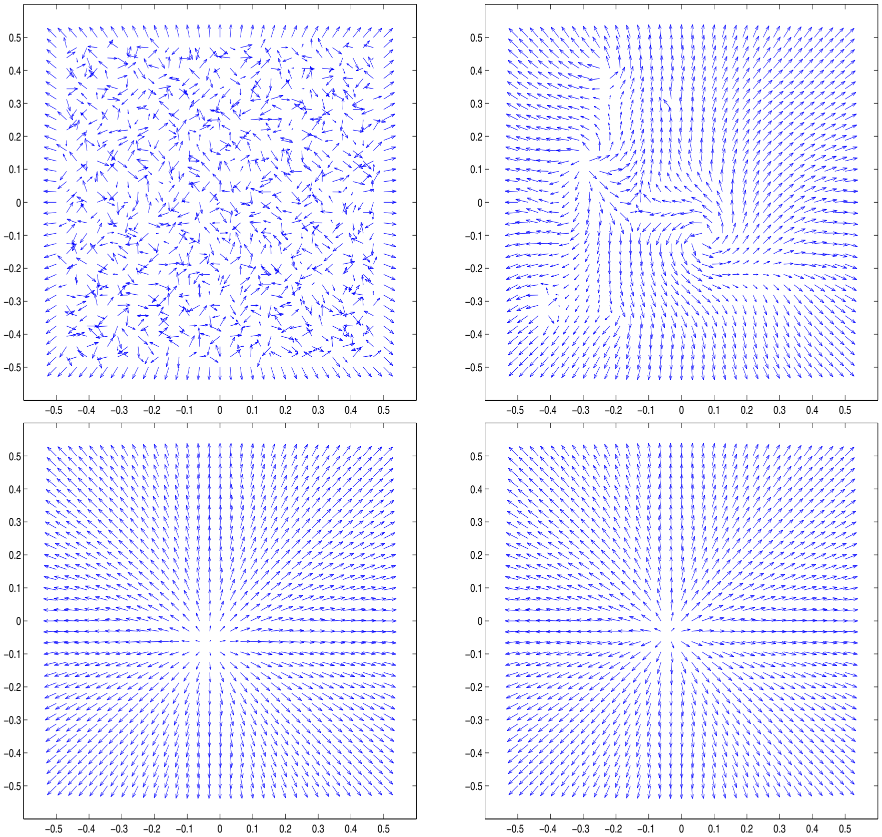


FIG. 4.4. *Projection of the vector fields* $u_{32}^{(j)}(0,\cdot,\cdot)$ *onto* $\{(x,y,z)\in\mathbb{R}^3 : x=0\}$ *in* $(-1/2,1/2)^2$ *for* $j=0,10,50,165$ *in Example* 4.1 *with initial data* $\tilde{u}_{32}^{(0)}\in\mathcal{A}_h(\mathcal{T},u_{\mathrm{D}})$.

Figure 4.4 shows the projection of the vector fields $u_{32}^{(j)}(0,\cdot,\cdot)$ onto $\{(x,y,z)\in\mathbb{R}^3 : x=0\}$ in $(-1/2,1/2)^2$ for $j=0,10,50,165$ produced by Algorithm (A_h) with initial data $\tilde{u}_{32}^{(0)}$. We observe that the algorithm immediately changes the highly unordered initial configuration into a more stable one; after 10 iterations only one degree-1 singularity with high symmetry can be seen. The subsequent iterations move the singularity to the origin.

5. Local refinement criteria. The main assertion of this section is a modification of the assumptions of Theorem 3.9. It replaces the assumption that the maximal mesh size tends to 0 and that the employed triangulations are of acute type by the weaker assumption that certain computable quantities tend to 0. Moreover, the assertion is independent of a particular scheme since the computable quantities are entirely determined by an approximation u_h.

Given a regular triangulation $\mathcal{T}$ of Ω let $h_{\mathcal{T}}\in L^\infty(\Omega)$ be the $\mathcal{T}$-elementwise constant function satisfying $h_{\mathcal{T}}|_T=\operatorname{diam}(T)$ for all $T\in\mathcal{T}$. $\mathcal{F}$ denotes the set of all faces in $\mathcal{T}$ and $h_{\mathcal{F}}$ is defined on $\cup\mathcal{F}$ through $h_{\mathcal{F}}|_F:=\operatorname{diam}(F)$ for all $F\in\mathcal{F}$. For a $\mathcal{T}$-elementwise smooth (e.g., $\mathcal{T}$-elementwise constant) function $\Sigma\in L^\infty(\Omega;\mathbb{R}^{3\times 3})$ we set

$$\big[\Sigma\cdot n_{\mathcal{F}}\big]|_F := \big((\Sigma|_{T_2})|_F-(\Sigma|_{T_1})|_F\big)\cdot n_F,$$

where $F \in \mathcal{F} \cap \Omega$, $T_1, T_2 \in \mathcal{T}$ such that $T_1 \cap T_2 = F$, and $n_F \in \mathbb{R}^3$ is the unit normal vector to F pointing from T_1 into T_2.

Definition 5.1. *Given any $u_h \in \mathcal{S}^1(\mathcal{T})^3$ let $w_h \in \mathcal{S}^1_0(\mathcal{T})^3$ satisfy $w_h(z) \cdot u_h(z) = 0$ for all $z \in \mathcal{N}$ and*

$$\int_\Omega Dw_h : Dv_h \, \mathrm{d}x = \int_\Omega Du_h : Dv_h \, \mathrm{d}x$$

for all $v_h \in \mathcal{S}^1_0(\mathcal{T})^3$ with $v_h(z) \cdot u_h(z) = 0$ for all $z \in \mathcal{N}$. For each $T \in \mathcal{T}$ set

$$\begin{aligned} \eta_1(T, u_h)^2 &:= \big\| |u_h|^2 - 1 \big\|^2_{L^2(T)} + \big\| u_h|_{\partial\Omega} - u_{\mathrm{D}} \big\|^2_{L^2(\partial T \cap \partial\Omega)}, \\ \eta_2(T, u_h)^2 &:= \big\| h_{\mathcal{F}}^{1/2} [D(u_h - w_h) \cdot n_{\mathcal{F}}] \big\|^2_{L^2(\partial T \cap \Omega)} + \| Dw_h \|^2_{L^2(T)}. \end{aligned}$$

Note that the following assertion does not assume that u_h is obtained by Algorithm (A_h), that the maximal mesh sizes tend to 0, or that the triangulations satisfy condition (ED).

Proposition 5.2. *Suppose that $(\mathcal{T}_k : k \in \mathbb{N})$ is a sequence of regular triangulations of Ω, and let $(u_k : k \in \mathbb{N}) \subseteq \mathcal{S}^1(\mathcal{T}_k)^3$ be such that $\|Du_k\|_{L^2(\Omega)} \le C_1$ for some $C_1 > 0$ and all $k \ge 0$. Suppose that*

$$\sum_{T \in \mathcal{T}_k} \eta_1(T, u_k)^2 + \eta_2(T, u_k)^2 \to 0 \quad \textit{for } k \to \infty.$$

Then there exists a subsequence $(u_\ell : \ell \in \mathbb{N})$ and a harmonic map $u^* \in \mathcal{A}(u_{\mathrm{D}})$ such that $u_\ell \rightharpoonup u^*$ (weakly) in H^1 and

$$E(u^*) \le \liminf_{\ell \to \infty} E(u_\ell). \tag{5.1}$$

Proof. The boundedness of $\|Du_k\|^2_{L^2(\Omega)} + \|u_k\|^2_{L^2(\Omega)}$ implies the existence of a weakly convergent subsequence $(u_\ell : \ell \in \mathcal{N})$ and a weak limit $u^* \in H^1(\Omega; \mathbb{R}^3)$. Since $\sum_{T \in \mathcal{T}} \eta_1(T, u_\ell)^2 \to 0$ one verifies that $u^* \in \mathcal{A}(u_{\mathrm{D}})$. The weak lower semicontinuity of E proves (5.1). It remains to show that u^* is a harmonic map. Given any $\phi \in C_0^\infty(\Omega; \mathbb{R}^3)$ we set $\Phi_\ell := \phi \times u_\ell$ and let $\Psi_\ell := \mathcal{I}_\ell \Phi_\ell$ be the nodal interpolant of Φ_ℓ on $\mathcal{T}_\ell$. As in the proof of Theorem 3.9 we have to show that

$$\int_\Omega Du_\ell : D(\phi \times u_\ell) \, \mathrm{d}x = \int_\Omega D(u_\ell - w_\ell) : D(\Phi_\ell - \Psi_\ell) \, \mathrm{d}x + \int_\Omega Dw_\ell : D\Phi_\ell \, \mathrm{d}x \to 0.$$

A $\mathcal{T}_\ell$-elementwise integration by parts and standard interpolation estimates yield

$$\begin{aligned} \int_\Omega D(u_\ell - w_\ell) : D(\Phi_\ell - \Psi_\ell) \, \mathrm{d}x &= \sum_{F \in \mathcal{F}_\ell,\, F \subseteq \Omega} \int_F [D(u_h - w_h) \cdot n_{\mathcal{F}_\ell}] \cdot (\Phi_\ell - \Psi_\ell) \, \mathrm{d}s \\ &\le C \Big(\sum_{T \in \mathcal{T}} \eta_2(T, u_\ell)^2 \Big)^{1/2} \|D\Phi_\ell\|_{L^2(\Omega)}. \end{aligned}$$

Hölder's inequality implies

$$\int_\Omega Dw_\ell : D\Phi_\ell \, \mathrm{d}x \le \Big(\sum_{T \in \mathcal{T}_k} \eta_2(T, u_k)^2 \Big)^{1/2} \|D\Phi_\ell\|_{L^2(\Omega)}.$$

The proof of Theorem 3.9 shows

$$\int_\Omega Du_\ell : D(\phi \times u_\ell)\,\mathrm{d}x \to \int_\Omega D\phi : (u^* \times Du^*)\,\mathrm{d}x.$$

A combination of the assertions with Lemma 3.10 and $\|D\Phi_\ell\|_{L^2(\Omega)} \le C$ shows that u^* is a harmonic map. □

6. Numerical experiments II.

6.1. Adaptive algorithm. Proposition 5.2 motivates the following adaptive mesh refinement algorithm. It realizes uniform mesh refinement for $\Theta = 0$ and adaptive mesh refinement for $\Theta = 1/2$. The idea is to iterate steps (b) and (d) of Algorithm (A_h) as long as the energy is significantly decreasing. A termination criterion that may be based on smallness of the local refinement indicators $\eta_j(T, u_h)$ can easily be included.

Algorithm (A_h^Θ). Input: $(\mathcal{T}, u_h^{(0)}, \kappa)$, where $\mathcal{T}$ is a regular triangulation of Ω, $u_h^{(0)} \in \mathcal{A}_h(\mathcal{T}, u_\mathrm{D})$, and $\kappa > 0$.

(a) Set $j := 0$.

  (a1) Solve the optimization problem

$$\begin{cases} \text{Minimize} & E\Big(u_h^{(j)} - v_h\Big) \\ \text{subject to} & v_h \in \mathcal{S}_0^1(\mathcal{T})^3 \ \text{ and } \ v_h(z) \cdot u_h^{(j)}(z) = 0 \text{ for all } z \in \mathcal{N}. \end{cases}$$

  Denote the solution by $w_h^{(j)}$.

  (a2) Define

$$u_h^{(j+1)} := \sum_{z \in \mathcal{N}} \frac{u_h^{(j)}(z) - w_h^{(j)}(z)}{|u_h^{(j)}(z) - w_h^{(j)}(z)|}\varphi_z.$$

  (a3) If $E(u_h^{(j+1)}) \le E(u_h^{(j)}) - \kappa$ set $j := j+1$ and go to (a1).

(b) Set $u_h := u_h^{(j)}$.

(c) For each $T \in \mathcal{T}$ compute $\eta(T)^2 := \sum_{j=1}^2 \eta_j(T, u_h)^2$.

(d) Mark all $T \in \mathcal{T}$ which satisfy $\eta(T) \ge \Theta \max_{S \in \mathcal{T}} \eta(S)$ for refinement and generate a new regular triangulation $\mathcal{T}'$ such that all marked elements are refined.

(e) Set $\mathcal{T} := \mathcal{T}'$, construct $u_h^{(0)} \in \mathcal{A}_h(\mathcal{T}, u_\mathrm{D})$ by interpolating nodal values of u_h, and go to (a).

6.2. Numerical example. We ran Algorithm (A_h^Θ) with $\Theta = 0$ and $\Theta = 1/2$ in Example 4.1 and an initial triangulation of Ω into five tetrahedra. We chose the termination criterion $\kappa := 10^{-4}$ for the iteration in step (a) of Algorithm (A_h^Θ). The mesh refinement was realized by a bisection strategy for $\Theta = 1/2$ and by uniform (red) refinement for $\Theta = 0$.

The left plot in Figure 6.1 displays the L^2 error $\|u - u_h\|_{L^2(\Omega)}$ for uniform and adaptive mesh refinement with the iterates u_h of Algorithm (A_h^Θ). We used a logarithmic scaling on both axes to identify a relation between the number of degrees of freedom and the L^2 error.

We observe that the L^2 error is significantly smaller at comparable numbers of degrees of freedom when the refinement indicators of Proposition 5.2 are used to refine

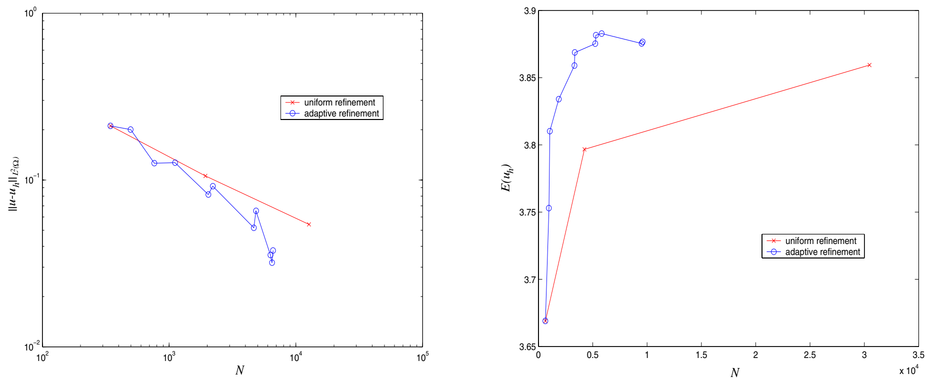


FIG. 6.1. *L^2 error for uniform and adaptive mesh refinement in Example* 4.1 *(left). Discrete energies $E(u_h)$ for uniform and adaptive mesh refinement (right).*

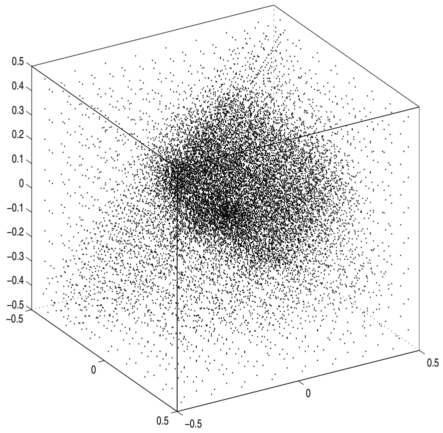


FIG. 6.2. *Midpoints of tetrahedra (indicated by dots) in the adaptively generated triangulation $\mathcal{T}$ after four iterations of Algorithm* $(\mathrm{A}_h^{1/2})$ *in Example* 4.1.

the mesh locally. Moreover, the experimental convergence rate for uniform meshes is only $\mathcal{O}(h)$ (owing to $h = N^{-1/3}$ for uniform meshes) instead of the optimal convergence rate $\mathcal{O}(h^2)$. The adaptive refinement strategy leads to an improved experimental convergence rate. The right plot in Figure 6.1 displays the discrete energies $E(u_h)$ for uniform and adaptive mesh refinement and we observe that the adaptive strategy reaches a stable value for a smaller number of degrees of freedom than the uniform refinement strategy.

Figure 6.2 displays the adapted triangulation generated by four iterations of Algorithm $(\mathrm{A}_h^{1/2})$. The dots in the plot indicate the location of a midpoint of a tetrahedron and we observe a refinement toward the origin, where the exact solution has a point singularity.

6.3. Instability of a degree-2 singularity. The final numerical example discusses a situation that leads to more than one degree-1 singularity.

EXAMPLE 6.1 (see [1, 14]). *Let $\pi_s : S^2 \to \mathbb{C}$ denote the stereographic projection of the unit sphere $S^2 \subseteq \mathbb{R}^3$ into the complex numbers $\mathbb{C}$ and let $\omega(z) := z^2$. Set*

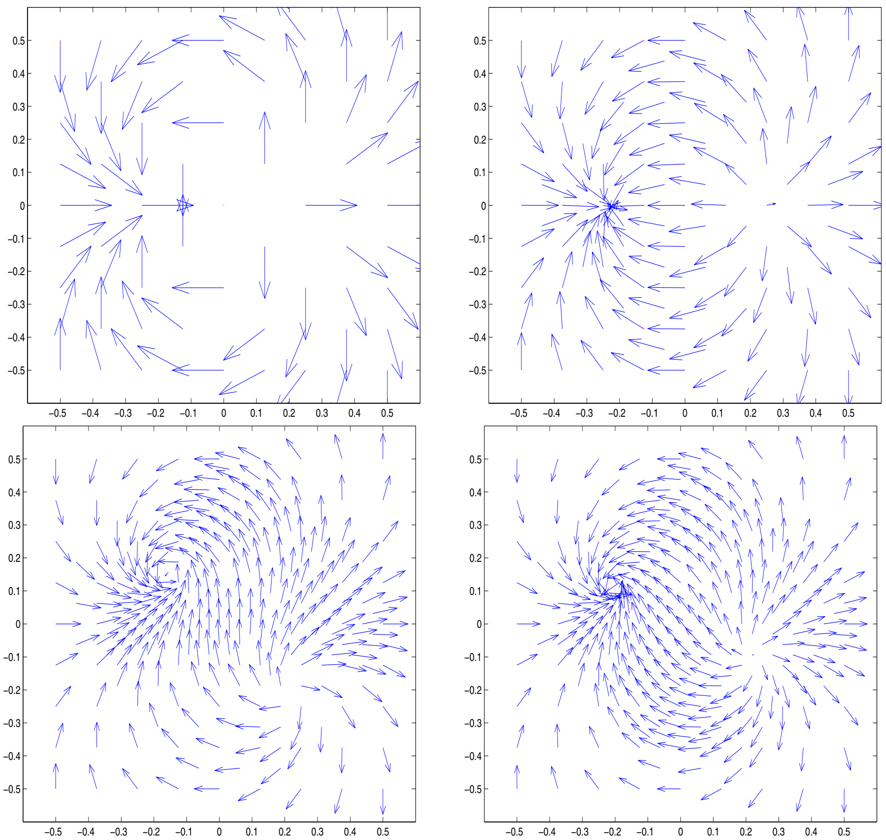


Fig. 6.3. *Intermediate solutions* $u_h(\cdot,\cdot,0)$ *in* $(-1/2,1/2)^2$ *after* $0,4,8,$ *and* 12 *iterations of Algorithm* $(\mathrm{A}_h^{1/2})$ *in Example* 6.1.

$\Omega := (-1/2,1/2)^3$ *and* $u_{\mathrm{D}}(x) := \pi_s^{-1} \circ \omega \circ \pi_s(x/|x|)$ *for* $x \in \partial\Omega$*.*

We employed Algorithm (A_h^{Θ}) with $\Theta = 1/2$ in Example 6.1 with an initial triangulation of Ω into five tetrahedra. We defined an initial function $u_h^{(0)}$ by nodal interpolation of the initial data. Figure 6.3 displays projections of intermediate solutions restricted to $\{(x,y,0) : -1/2 \le x,y \le 1/2\}$ on the adapted meshes after 0, 4, 8, and 12 iterations of the algorithm. We observe that the initial degree-2 singularity splits into two degree-1 singularities and the mesh is refined mostly between the two singularities in which the discrete vector field has a large gradient.

Acknowledgments. The author is grateful to one anonymous referee for pointing out the proof of Lemma 3.2. The author is also thankful to G.K. Dolzmann and R.H. Nochetto for stimulating discussions. The author wishes to thank J. Bolte for sharing his knowledge on three-dimensional mesh refinement.

REFERENCES

[1] F. Alouges, *A new algorithm for computing liquid crystal stable configurations: the harmonic mapping case*, SIAM J. Numer. Anal., 34 (1997), pp. 1708–1726.

[2] F. Alouges and J. M. Ghidaglia, *Minimizing Oseen-Frank energy for nematic liquid crystals: algorithms and numerical results*, Ann. Inst. H. Poincaré Phys. Théor., 66 (1997), pp. 411–447.

[3] S. Bartels, *Robust A Priori Error Analysis for the Approximation of Degree-One Ginzburg-Landau Vortices*, Preprint, 2004, www.math.umd.edu/˜sba.

[4] S. Bartels, *A posteriori error analysis for time-dependent Ginzburg-Landau type equations*, Numer. Math., 99 (2005), pp. 557–583.

[5] J. Bey, *Tetrahedral grid refinement*, Computing, 55 (1995), pp. 355–378.

[6] D. Braess, *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics*, 2nd ed., Cambridge University Press, Cambridge, 2001.

[7] H. Brezis, J.-M. Coron, and E. Lieb, *Harmonic maps with defects*, Comm. Math. Phys., 107 (1986), pp. 649–705.

[8] Y. Chen, *The weak solutions to the evolution problem of harmonic maps*, Math. Z. 201 (1989), pp. 69–74.

[9] R. Cohen, R. Hardt, D. Kinderlehrer, S.-Y. Lin, and M. Luskin, *Minimum energy configurations for liquid crystals: Computational results*, in Theory and Applications of Liquid Crystals, IMA Vol. Math. Appl. 5, Springer-Verlag, New York, 1987, pp. 99–122.

[10] R. Cohen, S. -Y. Lin, and M. Luskin, *Relaxation and gradient methods for molecular orientation in liquid crystals*, Comput. Phys. Comm., 53 (1989), pp. 455–465.

[11] Q. Du, R. A. Nicolaides, and X. Wu, *Analysis and convergence of a covolume approximation of the Ginzburg-Landau model of superconductivity*, SIAM J. Numer. Anal., 35 (1998), pp. 1049–1072.

[12] Q. Du, M. D. Gunzburger, and J. S. Peterson, *Analysis and approximation of the Ginzburg-Landau model of superconductivity*, SIAM Rev., 34 (1992), pp. 54–81.

[13] F.C. Frank, *On the theory of liquid crystals*, Discuss. Faraday Soc., 25 (1958), pp. 19–28.

[14] R. Hardt, D. Kinderlehrer, and F.-H. Lin, *Existence and partial regularity of static liquid crystal configurations*, Comm. Math. Phys., 105 (1986), pp. 547–570.

[15] S. Hildebrandt, H. Kaul, and K.-O. Widman, *Dirichlet's boundary value problem for harmonic mappings of Riemannian manifolds*, Math. Z., 147 (1976), pp. 225–236.

[16] S. Korotov and M. Křížek, *Acute type refinements of tetrahedral partitions of polyhedral domains*, SIAM J. Numer. Anal., 39 (2001), pp. 724–733.

[17] M. Křížek and L. Qun, *On diagonal dominance of stiffness matrices in* $3D$, East-West J. Numer. Math., 3 (1995), pp. 59–69.

[18] F.-H. Lin, *A remark about the map* $x/|x|$, C. R. Acad. Sci. Paris, 305 (1987), pp. 529–531.

[19] S.-Y. Lin and M. Luskin, *Relaxation methods for liquid crystal problems*, SIAM J. Numer. Anal., 26 (1989), pp. 1310–1326.

[20] C. Liu and N. J. Walkington, *Approximation of liquid crystal flows*, SIAM J. Numer. Anal., (2002), pp. 725–741.

[21] C. Liu and N. J. Walkington, *Mixed methods for the approximation of liquid crystal flows*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 205–222.

[22] C. W. Oseen, *The theory of liquid crystals*, Trans. Faraday Soc., 29 (1933), pp. 883–899.

[23] A. Prohl, *Computational Micromagnetism*, Adv. Numer. Math., B. G. Teubner, Stuttgart, 2001.

[24] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Texts Appl. Math., Springer-Verlag, New York, 1993.

[25] T. Rivìere, *Existence of Infinitely Many Weakly Harmonic Maps into Spheres for Nonconstant Boundary Data*, Preprint 9419, Centre de Mathématiques et de Leurs Applications, Cachan, France, 1994.

[26] T. Rivìere, *Everywhere discontinuous harmonic maps into spheres*, Acta Math., 175 (1995), pp. 197–226.

[27] R. Schoen and K. Uhlenbeck, *A regularity theory for harmonic maps*, J. Differential Geom., 17 (1982), pp. 307–335.

[28] M. Struwe, *On the evolution of harmonic mappings of Riemannian surfaces*, Comment. Math. Helv., 60 (1985), pp. 558–581.

[29] E. G. Virga, *Variational Theories for Liquid Crystals*, Applied Mathematics and Mathematical Computation 8, Chapman & Hall, London, 1994.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 239–258

ERROR ANALYSIS OF PRESSURE-CORRECTION SCHEMES FOR THE TIME-DEPENDENT STOKES EQUATIONS WITH OPEN BOUNDARY CONDITIONS[*]

J. L. GUERMOND[†], P. MINEV[‡], AND J. SHEN[§]

Abstract. The incompressible Stokes equations with prescribed normal stress (open) boundary conditions on part of the boundary are considered. It is shown that the standard pressure-correction method is not suitable for approximating the Stokes equations with open boundary conditions, whereas the rotational pressure-correction method yields reasonably good error estimates. These results appear to be the first ever published for splitting schemes with open boundary conditions. Numerical results in agreement with the error estimates are presented.

Key words. Navier–Stokes and Stokes equations, incompressibility, pressure-correction methods, open boundary conditions, finite elements, spectral approximations

AMS subject classifications. 65M12, 35Q30, 76D05

DOI. 10.1137/040604418

1. Introduction.

In this paper we consider the time-dependent Navier–Stokes equations with normal stress boundary conditions prescribed on parts of the boundary. These conditions are usually imposed to model outflow boundaries or free surfaces. For Newtonian flows, the boundary conditions in question take the form

$$\left[\mathsf{p}n - \nu(\nabla \mathsf{u} + (\nabla \mathsf{u})^T)n\right]|_\Gamma = b,$$

where u is the velocity vector field, p is the pressure, Γ is the boundary of the domain Ω, n is the unit outward normal, and b is the prescribed data.

There are numerous ways to discretize the time-dependent incompressible Navier–Stokes equations in time. Undoubtedly, the most popular one consists of using projection methods. Most of these techniques are based on the original ideas of Chorin [2] and Temam [22]. They are usually fractional step methods composed of two substeps such that either the Laplacian of the velocity or the pressure gradient is made explicit in one substep and (implicitly) corrected in the other substep. In both cases, one substep always consists of the projection of some vector field onto a divergence-free space. Following the terminology introduced in [11], a scheme is classified as a pressure-correction (resp., velocity-correction) method if the pressure gradient (resp., Laplacian of the velocity) is treated explicitly in one substep and (implicitly) corrected in the other substep. In the present paper we restrict ourselves to pressure-correction methods. Each of the above two classes of methods has a standard form and a rotational form (see [9, 10]), and each of them can be implemented either in algebraic

[*] Received by the editors February 24, 2004; accepted for publication (in revised form) August 31, 2004; published electronically May 27, 2005.
http://www.siam.org/journals/sinum/43-1/60441.html

[†] LIMSI (CNRS-UPR 3152), BP 133, 91403, Orsay, France (guermond@limsi.fr). The work of this author was supported by CNRS and Texas Institute for Computational and Applied Mathematics, Austin, TX, under a TICAM Visiting Faculty Fellowship.

[‡] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB T6G 2G1, Canada (minev@ualberta.ca). The work of this author was supported by an NSERC research grant.

[§] Department of Mathematics, Purdue University, West Lafayette, IN 47907 (shen@math.purdue.edu). The work of this author was partially supported by NFS grant DMS-0311915.

form (cf. [4, 5, 15]) or in differential form. However, to the best of our knowledge, no rigorous error analysis of any of these schemes with open boundary conditions is available in the literature. Moreover, there is some confusion in the literature over the performance of these methods with this type of boundary condition. The aim of this paper is to discuss some of these issues and to derive error estimates.

We show that the standard pressure-correction schemes, implemented either in algebraic form or in differential form (in fact, they can be shown to be equivalent), are not suitable for approximating the Navier–Stokes equations supplemented with open boundary conditions. However, we show that the rotational pressure-correction schemes yield reasonable error estimates. More precisely, assuming full regularity of the Stokes problem, the second-order rotational pressure-correction method yields $\mathcal{O}(\Delta t^{3/2})$ convergence rate for the velocity in the L^2-norm and $\mathcal{O}(\Delta t)$ convergence rate for both the velocity in the H^1-norm and the pressure in the L^2-norm. These estimates deteriorate if the Stokes problem does not possesses full regularity, as is probably the case in three dimensions.

2. Preliminaries.

We shall consider the time-dependent Navier–Stokes equations on a finite time interval $[0,T]$ and in an open, connected, bounded domain $\Omega \subset \mathbb{R}^d$ ($d = 2$, or 3) with a boundary Γ sufficiently smooth. We assume that the following nontrivial partition holds: $\Gamma = \Gamma_1 \cup \Gamma_2$, $\Gamma_1 \cap \Gamma_2 = \emptyset$, $\text{meas}(\Gamma_1) \neq \emptyset$, $\text{meas}(\Gamma_2) \neq \emptyset$.

2.1. Notation.

We denote by $H^m(\Omega)$ and $\|\cdot\|_m$ ($m = 0, \pm 1, \ldots$) the standard Sobolev spaces and norms. In particular, the norm and inner product of $L^2(\Omega) = H^0(\Omega)$ are denoted by $\|\cdot\|_0$ and $(\cdot,\cdot)$, respectively. We shall also make use of fractional Sobolev spaces $H^s(\Omega)$ which are defined by interpolation. To account for homogeneous Dirichlet boundary conditions on Γ_1, we define

$$X = \{v \in H^1(\Omega) : v|_{\Gamma_1} = 0\}. \tag{2.1}$$

Owing to the Poincaré inequality, $\|\nabla v\|_0$ is a norm equivalent to $\|v\|_1$ for all $v \in X$. Henceforth, we redefine the norm $\|\cdot\|_1$ in X such that $\|v\|_1 := \|\nabla v\|_0$.

We introduce two spaces of incompressible vector fields,

$$H = \{v \in L^2(\Omega)^d;\ \nabla\cdot v = 0;\ v \cdot n|_{\Gamma_1} = 0\}, \tag{2.2}$$

$$V = \{v \in H^1(\Omega)^d;\ \nabla\cdot v = 0;\ v|_{\Gamma_1} = 0\}, \tag{2.3}$$

and we define P_H to be the L^2-orthogonal projection onto H, i.e.,

$$(u - P_H u, v) = 0 \quad \forall u \in L^2(\Omega)^d,\ \forall v \in H. \tag{2.4}$$

We also denote

$$N = \{q \in H^1(\Omega);\ q|_{\Gamma_2} = 0\}. \tag{2.5}$$

The following well-known lemma plays a key role in the analysis of projection methods.

LEMMA 2.1. *The following orthogonal decomposition of $L^2(\Omega)^d$ holds:*

$$L^2(\Omega)^d = H \oplus \nabla N. \tag{2.6}$$

Since the nonlinear term in the Navier–Stokes equations has a marginal influence on the splitting error, we shall hereafter consider only the time-dependent Stokes equations written in terms of velocity, u, and pressure, p:

$$\begin{cases} \partial_t \mathsf{u} + A\mathsf{u} + \nabla \mathsf{p} = f \quad \text{in } \Omega \times [0,T], \\ \nabla \cdot \mathsf{u} = 0 \quad \text{in } \Omega \times [0,T], \\ \mathsf{u}|_{\Gamma_1} = 0, \quad \text{and} \quad (\mathsf{p}n - \nu(D\mathsf{u})n)|_{\Gamma_2} = 0 \quad \text{in } [0,T], \\ \mathsf{u}|_{t=0} = u_0 \quad \text{in } \Omega. \end{cases} \tag{2.7}$$

Henceforth, the operators A and D may assume one of the two following forms:

$$Av = -2\nu \nabla \cdot Dv, \tag{2.8}$$

$$Dv = \begin{cases} \frac{1}{2}\nabla v, & \text{case 1}, \\ \frac{1}{2}(\nabla v + (\nabla v)^T), & \text{case 2}. \end{cases} \tag{2.9}$$

We recall that the symmetric positive definite bilinear form

$$a(u,v) = \nu(Du, Dv) \tag{2.10}$$

induces a norm on X that is equivalent to the H^1-norm. We denote by α the coercivity constant of a:

$$a(v,v) \geq \alpha \|\nabla v\|_0^2 \quad \forall v \in X. \tag{2.11}$$

In case 1, $\alpha = \nu$, whereas in case 2, $\alpha = c\nu$, where c is a constant that can be derived by using a Korn inequality; see, e.g., [1].

To simplify our presentation, we assume that the unique solution (u, p) to the above system is as smooth as needed.

To perform the temporal discretization of the problem, we define $\Delta t > 0$ to be a time step and we set $t^k = k\Delta t$ for $0 \leq k \leq K = [T/\Delta t]$. Let $\phi^0, \phi^1, \ldots, \phi^K$ be a sequence of functions in some Hilbert space E. We denote by $\phi_{\Delta t}$ this sequence, and we use the following discrete norms:

$$\|\phi_{\Delta t}\|_{\ell^2(E)} := \left(\Delta t \sum_{k=0}^{K} \|\phi^k\|_E^2 \right)^{1/2}, \qquad \|\phi_{\Delta t}\|_{\ell^\infty(E)} := \max_{0 \leq k \leq K} \left(\|\phi^k\|_E \right). \tag{2.12}$$

We denote by c a generic constant that is independent of small parameters like ϵ, Δt, and h but possibly depends on the data and the solution. We shall use the expression $A \lesssim B$ to say that there exists a generic constant c such that $A \leq cB$.

Let μ be a positive real number. We shall repeatedly make use of the following interpolation result, whose proof is fairly standard and so we omit it due to the space limitation.

Lemma 2.2. *For all $0 \leq s \leq 1$, there exists an operator $\mathcal{I}_{\mu,s} : H^s(\Omega) \longrightarrow H_0^1(\Omega)$ such that for all r in $H^s(\Omega)$ we have*

$$\|r - \mathcal{I}_{\mu,s} r\|_0 \lesssim \mu^{\frac{s}{2}} \|r\|_{H^s(\Omega)}, \tag{2.13}$$

$$\|\mathcal{I}_{\mu,s} r\|_1 \lesssim \mu^{-1+\frac{s}{2}} \|r\|_{H^s(\Omega)}. \tag{2.14}$$

2.1.1. The inverse of the Stokes operator and its regularity index. In this section we recall properties of the inverse of the Stokes operator. Let X' be the dual space of X. We denote by $\langle \cdot, \cdot \rangle$ the duality pairing between X' and X. The inverse of the Stokes operator, which we shall denote by $S : X' \longrightarrow X$, is defined as follows. For all v in X', $S(v) \in X$ is the solution to the dual problem

$$\begin{cases} a(w, S(v)) - (r, \nabla \cdot w) = \langle v, w \rangle & \forall w \in X, \\ (q, \nabla \cdot S(v)) = 0 \qquad \forall q \in L^2(\Omega). \end{cases} \tag{2.15}$$

Obviously, we have

$$\forall v \in X', \qquad \|S(v)\|_1 + \|r\|_0 \leq c\|v\|_{X'}. \tag{2.16}$$

It is well known that when Dirichlet boundary conditions on the velocity are enforced on the entire boundary and Ω is smooth or convex, we have $\|r\|_1 \lesssim \|v\|_0$ (see, for instance, [23]). In the present case, where boundary conditions are mixed, it is a nontrivial task to determine the regularity of r. It is generally expected that the H^1-regularity does not hold in the three-dimensional case. However, it is possible that regularity in some fractional Sobolev space holds. To account for this, we make the following definition.

Definition 2.1 (regularity index of the Stokes operator). *The regularity index of the Stokes operator is the largest number, s, such that for all $v \in L^2(\Omega)^d$, the solution $r \in L^2(\Omega)$ to the dual Stokes problem* (2.15) *satisfies* $\|r\|_{H^s(\Omega)} \lesssim \|v\|_0$.

We observe from (2.16) that $s \geq 0$, and it is clear that $s \leq 1$. Hence, the case $s = 0$ is referred to as *no regularity* while the case $s = 1$ is referred to as *full regularity.* We refer to [14] for techniques to evaluate this index in two dimensions.

The operator S has interesting properties, as listed below.

Lemma 2.3. *For all v in X, all $0 < \gamma < 1$, and all $0 < \mu < 1$, we have*

$$a(v, S(v)) \geq (1 - \gamma)\|v\|_0^2 - c(\gamma)\left(\mu^{2\alpha_1}\|\nabla \cdot v\|_0^2 + \mu^{-2\alpha_2}\|v - P_H v\|_0^2\right),$$

with $\alpha_1 = \frac{s}{2}$ and $\alpha_2 = 1 - \frac{s}{2}$ and s being the regularity index of the Stokes operator. In particular, for all $v \in V$, $(\nabla S(v), \nabla v) = \|v\|_0^2$.

Proof. Owing to the definition of $S(v)$ and to the fact $\mathcal{I}_{\varepsilon,s} r$ is zero on Γ_2, we have

$$\begin{aligned} a(v, S(v)) &= \|v\|_0^2 + (r, \nabla \cdot v) \\ &= \|v\|_0^2 + (r - \mathcal{I}_{\mu,s} r, \nabla \cdot v) + (\nabla \mathcal{I}_{\mu,s} r, v) \\ &= \|v\|_0^2 + (r - \mathcal{I}_{\mu,s} r, \nabla \cdot v) + (\nabla \mathcal{I}_{\mu,s} r, v - P_H v) \\ &\geq \|v\|_0^2 - \left(\mu^{\alpha_1}\|\nabla \cdot v\|_0 + \mu^{-\alpha_2}\|v - P_H v\|_0\right)\|r\|_{H^s(\Omega)}. \end{aligned}$$

Then using the fact that s is the regularity index of the Stokes operator (see Definition 2.1), we derive the desired bound. ◻

Lemma 2.4. *The bilinear form $X' \times X' \ni (v, w) \longmapsto \langle S(v), w \rangle := a(S(v), S(w)) \in \mathbb{R}$ induces a seminorm on X' that we denote by $|\cdot|_\star$, and*

$$\forall v \in X', \qquad |v|_\star = a(S(v), S(v))^{1/2} \lesssim \|v\|_{X'}.$$

Proof. It is clear that the bilinear form is symmetric, $\langle S(v), w \rangle = a(S(v), S(w)) = \langle S(w), v \rangle$, and positive, $\langle S(v), v \rangle = a(S(v), S(v))$; hence, $\langle S(v), w \rangle$ induces a seminorm on X'. Furthermore, $|v|_\star^2 = \langle S(v), v \rangle = a(S(v), S(v)) \lesssim \|v\|_{X'}^2$. The proof is complete. ◻

3. Standard pressure-correction methods. For purely Dirichlet boundary conditions, the second-order pressure-correction scheme is known to be one-order more accurate than the original projection scheme of Chorin–Temam (cf. [25, 3, 21, 7]). Using the second-order backward difference formula (BDF2) to discretize the time derivative, the second-order pressure-correction scheme takes the following form:

Set $u^0 = u_0$, $p^0 = \mathsf{p}|_{t=0}$, which can be computed from the data, and compute $(\tilde{u}^1, u^1, p^1)$ by using the scheme below with BDF2 replaced by the backward Euler formula. Then, for $k \geq 1$, compute $(\tilde{u}^{k+1}, u^{k+1}, p^{k+1})$ such that

$$\text{(3.1)} \qquad \begin{cases} \dfrac{3\tilde{u}^{k+1} - 4u^k + u^{k-1}}{2\Delta t} + A\tilde{u}^{k+1} + \nabla p^k = f(t^{k+1}), \\ \tilde{u}^{k+1}|_{\Gamma_1} = 0 \quad \text{and} \quad (p^k n - \nu(D\tilde{u}^{k+1})n)|_{\Gamma_2} = 0 \end{cases}$$

and

$$\text{(3.2)} \qquad \begin{cases} \dfrac{3u^{k+1} - 3\tilde{u}^{k+1}}{2\Delta t} + \nabla(p^{k+1} - p^k) = 0, \\ \nabla \cdot u^{k+1} = 0, \\ u^{k+1} \cdot n|_{\Gamma_1} = 0 \quad \text{and} \quad (p^{k+1} - p^k)|_{\Gamma_2} = 0. \end{cases}$$

The first substep accounts for viscous effects, whereas the second one accounts for incompressibility. The second substep is usually referred to as the projection step, for it is a realization of the identity $u^{k+1} = P_H \tilde{u}^{k+1}$. We emphasize that it is essential, for stability considerations, that $(p^{k+1} - p^k)|_{\Gamma_2} = 0$ is enforced. Otherwise, (3.2) can not be interpreted as a projection step. Note that the boundary conditions in (3.2) lead to the series of equalities

$$\text{(3.3)} \qquad \begin{aligned} &\frac{\partial}{\partial n} p^{k+1}|_{\Gamma_1} = \frac{\partial}{\partial n} p^k|_{\Gamma_1} = \cdots = \frac{\partial}{\partial n} p^1|_{\Gamma_1}, \\ &p^{k+1}|_{\Gamma_2} = p^k|_{\Gamma_2} = \cdots = p^1|_{\Gamma_2}, \end{aligned}$$

which are certainly inaccurate since they are almost never satisfied by the exact solution. In the purely Dirichlet case, i.e., $\Gamma_2 = \emptyset$, it is possible to deduce a reasonably good approximation result for the pressure in the L^2-norm. But when $\Gamma_2 \neq \emptyset$ the pressure approximation is severely degraded.

Not being aware of any published convergence result for the scheme (3.1)–(3.2), we shall prove the following result.

THEOREM 3.1. *If* (u, p)*, the solution to* (2.7)*, is smooth enough in space and time, the solution to* (3.1)–(3.2) *satisfies the following error estimates:*

$$\begin{aligned} \|\mathsf{p}_{\Delta t} - p_{\Delta t}\|_{\ell^\infty(L^2(\Omega))} + \|\mathsf{u}_{\Delta t} - \tilde{u}_{\Delta t}\|_{\ell^\infty(H^1(\Omega)^d)} &\lesssim \Delta t^{\frac{1}{2}}, \\ \|\mathsf{u}_{\Delta t} - u_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)} + \|\mathsf{u}_{\Delta t} - \tilde{u}_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)} &\lesssim \Delta t^{\frac{s+1}{2}}, \end{aligned}$$

where s is the regularity index of the Stokes operator.

Proof. As will become clear in the course of the proof, using BDF2 instead of the backward Euler formula does not improve the accuracy in the presence of open boundary conditions. So to simplify the presentation, we consider the backward Euler formula for the time derivative:

$$\text{(3.4)} \qquad \begin{cases} \dfrac{\tilde{u}^{k+1} - u^k}{\Delta t} + A\tilde{u}^{k+1} + \nabla p^k = f(t^{k+1}), \\ \tilde{u}^{k+1}|_{\Gamma_1} = 0 \quad \text{and} \quad (p^k n - \nu(D\tilde{u}^{k+1})n)|_{\Gamma_2} = 0 \end{cases}$$

and

$$\begin{cases} \dfrac{u^{k+1} - \tilde{u}^{k+1}}{\Delta t} + \nabla(p^{k+1} - p^k) = 0, \\ \nabla \cdot u^{k+1} = 0, \\ u^{k+1} \cdot n|_{\Gamma_1} = 0 \quad \text{and} \quad (p^{k+1} - p^k)|_{\Gamma_2} = 0. \end{cases} \tag{3.5}$$

Technically, the proof is very similar to those in Shen [21] and Guermond [6]; hence we show only those steps where the consistency error is degraded.

Let us introduce the interpolation operator $\mathcal{I}_{\Delta t,1} : H^1(\Omega) \longmapsto H^1_0(\Omega)$ defined in Lemma 2.2. This operator is such that for all r in $H^1(\Omega)$,

$$\|\mathcal{I}_{\Delta t,1} r - r\|_0 \lesssim \Delta t^{\frac{1}{2}} \|r\|_1, \tag{3.6}$$

$$\|\nabla \mathcal{I}_{\Delta t,1} r\|_0 \lesssim \Delta t^{-\frac{1}{2}} \|r\|_1. \tag{3.7}$$

Without introducing any essential extra error, we can take $p^0 = \mathcal{I}_{\Delta t,1} \mathsf{p}|_{t=0}$, which implies $p^k|_{\Gamma_2} = 0$ for all k.

Now we introduce the following notation:

$$\begin{cases} e^k = \mathsf{u}(t^k) - u^k, & \tilde{e}^k = \mathsf{u}(t^k) - \tilde{u}^k, \\ \psi^k = \mathcal{I}_{\Delta t,1} \mathsf{p}(t^{k+1}) - p^k, & q^k = \mathcal{I}_{\Delta t,1} \mathsf{p}(t^k) - p^k. \end{cases}$$

The weak form of the error equation that corresponds to the viscous step (3.4) is given by

$$\begin{aligned} \frac{1}{\Delta t}(\tilde{e}^{k+1} - e^k, v) + a(\tilde{e}^{k+1}, v) - (\psi^k, \nabla \cdot v) = (R(t^{k+1}), v) \\ + (\mathsf{p}(t^{k+1}) - \mathcal{I}_{\Delta t,1} \mathsf{p}(t^{k+1}), \nabla \cdot v) \quad \forall v \in X, \end{aligned}$$

where $R(t^{k+1}) = \frac{1}{\Delta t}(u(t^{k+1}) - u(t^k)) - u_t(t^{k+1}) = \mathcal{O}(\Delta t)$. Note that the surface integrals resulting from the integration by parts cancel on both Γ_1 and Γ_2 due to the boundary conditions in (3.4).

Taking $v = 2\Delta t \tilde{e}^{k+1}$ in the above equation and using (3.6), we can derive

$$2\Delta t(\mathsf{p}(t^{k+1}) - \mathcal{I}_{\Delta t,1} \mathsf{p}(t^{k+1}), \nabla \cdot \tilde{e}^{k+1}) \lesssim \Delta t^2 + \alpha \Delta t \|\tilde{e}^{k+1}\|_1^2,$$

$$\|\tilde{e}^{k+1}\|_0^2 + \|\tilde{e}^{k+1} - e^k\|_0^2 + \alpha \Delta t \|\tilde{e}^{k+1}\|_1^2 - 2\Delta t(\psi^k, \nabla \cdot \tilde{e}^{k+1}) \le \|e^k\|_0^2 + c\Delta t^2. \tag{3.8}$$

Note that the consistency error is degraded at this step; more precisely, a Δt factor is already missing in the above estimate.

The error equation corresponding to the projection step (3.5) can be written as

$$\begin{cases} \dfrac{1}{\Delta t} e^{k+1} + \nabla q^{k+1} = \dfrac{1}{\Delta t} \tilde{e}^{k+1} + \nabla \psi^k, \\ \nabla \cdot e^{k+1} = 0, \\ e^{k+1} \cdot n|_{\Gamma_1} = 0 \quad \text{and} \quad q^{k+1}|_{\Gamma_2} = 0. \end{cases}$$

Taking the square of the first relation above and multiplying the result by Δt^2, we infer

$$\|e^{k+1}\|_0^2 + \Delta t^2 \|\nabla q^{k+1}\|_0^2 = \|\tilde{e}^{k+1}\|_0^2 + \Delta t^2 \|\nabla \psi^k\|_0^2 - 2\Delta t(\psi^k, \nabla \cdot \tilde{e}^{k+1}). \tag{3.9}$$

Note that integration by parts can be performed on both sides owing to the fact that $q^{k+1}|_{\Gamma_2} = 0 = \psi^k|_{\Gamma_2}$. Now we have

$$\begin{aligned}\Delta t^2 \|\nabla \psi^k\|_0^2 &= \Delta t^2 \|\nabla q^k + \nabla(\mathcal{I}_{\Delta t}(\mathsf{p}(t^{k+1}) - \mathsf{p}(t^k)))\|_0^2,\\ &\le \Delta t^2 (\|\nabla q^k\|_0^2 + c\Delta t^{1-\frac{1}{2}} \|\nabla q^k\|_0 + c'\Delta t^{2(1-\frac{1}{2})})\\ &\le \Delta t^2 (1+\Delta t)\|\nabla q^k\|_0^2 + c\Delta t^2,\end{aligned}$$

where the consistency error is also degraded by a factor of $\mathcal{O}(\Delta t)$. Combining this result and the previous one, we have

$$\begin{aligned}\|e^{k+1}\|_0^2 + \Delta t^2 \|\nabla q^{k+1}\|_0^2 \le \|\tilde{e}^{k+1}\|_0^2 + \Delta t^2(1+\Delta t)\|\nabla q^k\|_0^2\\ - 2\Delta t(\psi^{k+1}, \nabla\cdot \tilde{e}^{k+1}) + c\Delta t^2.\end{aligned} \tag{3.10}$$

The first error estimate of the theorem is obtained by combining (3.8) and (3.10), using the discrete Gronwall lemma, and repeating the whole argument for time increments. The second estimate can be derived by a duality argument similar to that used in the proof of Lemma 4.4. □

REMARK 3.1. *Note that the error on the pressure in the L^2-norm is $\mathcal{O}(\Delta t^{\frac{1}{2}})$, whereas it is $\mathcal{O}(\Delta t)$ when Dirichlet boundary conditions are enforced on the whole boundary. It is clear that the artificial Dirichlet boundary condition* (3.3) *is responsible for this poor convergence property. Since using an inexact factorization (cf.* [4, 5, 15, 16, 17, 13]*) of the discrete Stokes operator does not enforce the Dirichlet boundary condition on* Γ_2 explicitly*, some authors have argued that the inexact factorization scheme does not suffer from the error due to the artificial Dirichlet boundary condition. However, it can be shown (see* [12] *for details) that the inexact factorization scheme actually enforces the artificial Dirichlet boundary condition* weakly *and hence suffers from the same accuracy loss as its PDE counterpart. In other words, mere algebraic manipulations cannot overcome essential difficulties encountered in functional analysis.*

REMARK 3.2. *Note that the need to integrate by parts the term* $2\Delta t(\nabla\psi^{k+1}, \tilde{e}^{k+1})$ *in* (3.9) *is critical, and it is made possible by enforcing the homogeneous Dirichlet boundary condition on the pressure at* Γ_2 *in the projection step* (3.2)*.*

We finish this section by recalling that to simulate outflow boundary conditions, an alternative set of conditions is $p|_{\Gamma_2} = 0$, $u \times n|_{\Gamma_2} = 0$. This set of conditions is not equivalent to the zero normal stress conditions studied above. Nevertheless, an interesting property of these boundary conditions is that they are compatible with the pressure-correction algorithm (3.1)–(3.2); i.e., they yield near optimal convergence rates. We refer to [8] for other technical details on this matter.

4. Rotational pressure-correction methods.

In this section, we show that the rotational pressure-correction scheme introduced in [24] improves, by a factor of $\Delta t^{1/2}$, the error estimates of the standard pressure-correction scheme. It is proved in [11, 10] that when Dirichlet boundary conditions are enforced on the entire boundary, the same improvement holds. The main result is stated in Theorem 4.1.

4.1. Rotational form.

When applied to problems with open boundary conditions on Γ_2, the rotational pressure-correction scheme takes the following form:

Set $u^0 = u_0$, $p^0 = \mathsf{p}|_{t=0}$, which can be computed from the data, and compute $(\tilde{u}^1, u^1, p^1)$ by using the scheme shown below with BDF2 replaced by the backward

Euler formula. Then, for $k \geq 1$, compute $(\tilde{u}^{k+1}, u^{k+1}, p^{k+1})$ such that

$$\left\{\begin{array}{l}\dfrac{3\tilde{u}^{k+1} - 4u^k + u^{k-1}}{2\Delta t} + A\tilde{u}^{k+1} + \nabla p^k = f(t^{k+1}), \\ \tilde{u}^{k+1}|_{\Gamma_1} = 0, \quad (p^k n - \nu(D\tilde{u}^{k+1})n)|_{\Gamma_2} = 0,\end{array}\right. \tag{4.1}$$

$$\left\{\begin{array}{l}\dfrac{3u^{k+1} - 3\tilde{u}^{k+1}}{2\Delta t} + \nabla\phi^{k+1} = 0, \\ \nabla\cdot u^{k+1} = 0, \\ u^{k+1}\cdot n|_{\Gamma_1} = 0, \quad \phi^{k+1}|_{\Gamma_2} = 0.\end{array}\right. \tag{4.2}$$

$$\phi^{k+1} = p^{k+1} - p^k + \chi\nabla\cdot\tilde{u}^{k+1}, \tag{4.3}$$

where χ is a tunable positive coefficient.

Remark 4.1. *As originally introduced in* [24]*, the coefficient χ was taken to be equal to α, defined in* (2.11)*, which is simply ν in the Newtonian case. The analysis performed in* [11, 10] *shows that this choice is sufficient to guarantee stability and convergence when Dirichlet boundary conditions are enforced. However, when natural boundary conditions are enforced on parts of the boundary, the analysis (see below) shows that χ should be chosen such that*

$$0 < \chi < 2\alpha \inf_{v\in X} \frac{\|\nabla v\|^2}{\|\nabla\cdot v\|^2}. \tag{4.4}$$

Owing to the inequality $\|\nabla\cdot v\|^2 \leq d\|\nabla v\|^2$, where d is the space dimension, it is sufficient to choose

$$0 < \chi < \frac{2}{d}\alpha. \tag{4.5}$$

4.2. A corresponding singularly perturbed system. To better understand the behavior of the scheme (4.1)–(4.3), we examine first a singularly perturbed system corresponding to the limiting case as $\Delta t \to 0$ (with $\varepsilon \sim \Delta t$). This system of PDEs is obtained by eliminating u^k from (4.1)–(4.2) and dropping some higher-order terms in ε:

$$\left\{\begin{array}{l}\partial_t u^\varepsilon + Au^\varepsilon + \nabla p^\varepsilon = f, \quad u^\varepsilon|_{\Gamma_1} = 0, \quad (p^\varepsilon n - \nu(Du^\varepsilon)n)|_{\Gamma_2} = 0, \\ \nabla\cdot u^\varepsilon - \varepsilon\nabla^2\phi^\varepsilon = 0, \quad \dfrac{\partial\phi^\varepsilon}{\partial n}|_{\Gamma_1} = 0, \quad \phi^\varepsilon|_{\Gamma_2} = 0, \\ \varepsilon\partial_t p^\varepsilon = \phi^\varepsilon - \chi\nabla\cdot u^\varepsilon,\end{array}\right. \tag{4.6}$$

with $u^\varepsilon|_{t=0} = u(0)$ and $p^\varepsilon(0) = p(0)$.

4.2.1. An estimate on $\nabla\cdot u^\varepsilon$. The following lemma is the key to obtaining improved error estimates.

Lemma 4.1. *Provided* u *and* p *are smooth enough in time and space, we have*

$$\|\nabla\cdot u^\varepsilon\|_{L^\infty(L^2(\Omega)^d)} + \sqrt{\varepsilon}\|\nabla\phi^\varepsilon\|_{L^\infty(L^2(\Omega))} \lesssim \varepsilon^{\frac{5}{4}}.$$

Proof. We set $e = u^\varepsilon - \mathsf{u}$ and $q = p^\varepsilon - \mathsf{p}$. Subtracting (4.6) from (2.7), we find

$$e_t + Ae + \nabla q = 0; \quad e|_{\Gamma_1} = 0, \quad (qn - \nu(De)n)|_{\Gamma_2} = 0, \tag{4.7}$$

$$\nabla\cdot e - \varepsilon\nabla^2\phi^\varepsilon = 0, \quad \frac{\partial\phi^\varepsilon}{\partial n}|_{\Gamma_1} = 0, \quad \phi^\varepsilon|_{\Gamma_2} = 0, \tag{4.8}$$

$$\varepsilon q_t = \phi^\varepsilon - \chi\nabla\cdot e - \varepsilon\mathsf{p}_t, \tag{4.9}$$

with $e(0) = 0$ and $q(0) = 0$.

Taking the inner product of the time derivative of (4.7) with e_t, we find

$$\frac{1}{2}\partial_t\|e_t\|_0^2 + \alpha\|\nabla e_t\|_0^2 - (q_t, \nabla\cdot e_t) \leq 0. \tag{4.10}$$

The inner product of (4.9) with $\nabla\cdot e_t$ yields

$$(q_t, \nabla\cdot e_t) = \frac{1}{\varepsilon}(\phi^\varepsilon, \nabla\cdot e_t) - (\mathsf{p}_t, \nabla\cdot e_t) - \frac{\chi}{2\varepsilon}\partial_t\|\nabla\cdot e\|^2, \tag{4.11}$$

and the inner product of the time derivative of (4.8) with ϕ^ε yields

$$\frac{1}{\varepsilon}(\phi^\varepsilon, \nabla\cdot e_t) = -(\nabla\phi^\varepsilon_t, \nabla\phi^\varepsilon). \tag{4.12}$$

The above two relations lead to

$$(q_t, \nabla\cdot e_t) = -\frac{1}{2}\partial_t\|\nabla\phi^\varepsilon\|_0^2 - (\mathsf{p}_t, \nabla\cdot e_t) - \frac{\chi}{2\varepsilon}\partial_t\|\nabla\cdot e\|^2. \tag{4.13}$$

Substituting this expression into (4.10) we obtain

$$\frac{1}{2}\partial_t\|e_t\|_0^2 + \alpha\|\nabla e_t\|_0^2 + \frac{1}{2}\partial_t\|\nabla\phi^\varepsilon\|_0^2 + \frac{\chi}{2\varepsilon}\partial_t\|\nabla\cdot e\|_0^2 \leq -(\mathsf{p}_t, \nabla\cdot e_t). \tag{4.14}$$

At this point, one would like to replace $\nabla\cdot e_t$ by $\varepsilon\nabla^2\phi^\varepsilon_t$ in $(\mathsf{p}_t, \nabla\cdot e_t)$ and integrate by parts. The integration by parts is not possible since neither p_t nor $\partial_n\phi^\varepsilon_t$ is zero at the boundary Γ_2. To account for this fact, we introduce the interpolation operator $\mathcal{J}_\varepsilon : H^1(\Omega) \longmapsto H^1_0(\Omega) \subset N$ such that $\mathcal{J}_\varepsilon = \mathcal{I}_{\sqrt{\varepsilon},1}$, where $\mathcal{I}_{\mu,s}$ has been defined in Lemma 2.2. Recall that for all r in $H^1(\Omega)$, Lemma 2.2 (with $\mu = \sqrt{\varepsilon}$, $s = 1$) yields

$$\|\mathcal{J}_\varepsilon r - r\|_0 \lesssim \varepsilon^{\frac{1}{4}}\|r\|_1, \qquad \|\nabla\mathcal{J}_\varepsilon r\|_0 \lesssim \varepsilon^{-\frac{1}{4}}\|r\|_1. \tag{4.15}$$

We rewrite (4.14) as

$$\frac{1}{2}\partial_t\left(\|e_t\|_0^2 + \|\nabla\phi^\varepsilon\|^2 + \frac{\chi}{\varepsilon}\|\nabla\cdot e\|_0^2\right) + \alpha\|\nabla e_t\|_0^2 = -(\mathsf{p}_t - \mathcal{J}_\varepsilon\mathsf{p}_t, \nabla\cdot e_t) + \varepsilon(\nabla\mathcal{J}_\varepsilon\mathsf{p}_t, \nabla\phi^\varepsilon_t).$$

Note that we used the fact that $\mathcal{J}_\varepsilon\mathsf{p}_t$ is zero at Γ_2 to integrate by parts. This is the key argument in this proof. Since $e(0) = 0$ and $q(0) = 0$, we infer $e_t(0) = 0$. Since $\nabla\cdot u^\varepsilon(0) = \nabla\cdot u(0) = 0$, we derive from (4.8) that $\phi^\varepsilon(0) = 0$. By integrating in time between 0 and t, we obtain

$$\begin{aligned}
\frac{1}{2}&\left(\|e_t\|_0^2 + \|\nabla\phi^\varepsilon\|_0^2 + \frac{\chi}{\varepsilon}\|\nabla\cdot e\|_0^2\right) + \alpha\int_0^t \|\nabla e_t\|_0^2 d\tau \\
&\leq -(\mathsf{p}_t - \mathcal{J}_\varepsilon\mathsf{p}_t, \nabla\cdot e) + \int_0^t (\mathsf{p}_{\tau\tau} - \mathcal{J}_\varepsilon\mathsf{p}_{\tau\tau}, \nabla\cdot e)d\tau \\
&\quad + \varepsilon(\nabla\mathcal{J}_\varepsilon\mathsf{p}_t, \nabla\phi^\varepsilon) - \int_0^t \varepsilon(\nabla\mathcal{J}_\varepsilon\mathsf{p}_{\tau\tau}, \nabla\phi^\varepsilon)d\tau \\
&\leq \frac{1}{4}\left(\frac{\chi}{\varepsilon}\|\nabla\cdot e\|_0^2 + \|\nabla\phi^\varepsilon\|_0^2\right) + \int_0^t \left(\frac{\chi}{\varepsilon}\|\nabla\cdot e\|_0^2 + \|\nabla\phi^\varepsilon\|_0^2\right)d\tau \\
&\quad + c\,\varepsilon\|\mathsf{p}_t - \mathcal{J}_\varepsilon\mathsf{p}_t\|^2_{L^\infty(0,t;L^2(\Omega))} + c'\,\varepsilon^2\|\mathcal{J}_\varepsilon\mathsf{p}_t\|^2_{L^\infty(0,t;H^1(\Omega))} \\
&\quad + c\,\varepsilon\|\mathsf{p}_{tt} - \mathcal{J}_\varepsilon\mathsf{p}_{tt}\|^2_{L^2(0,t;L^2(\Omega))} + c'\,\varepsilon^2\|\mathcal{J}_\varepsilon\mathsf{p}_{tt}\|^2_{L^2(0,t;H^1(\Omega))}.
\end{aligned}$$

Using the estimates (4.15), we infer

$$\frac{1}{4}\left(\|e_t\|_0^2+\|\nabla\phi^\varepsilon\|_0^2+\frac{\chi}{\varepsilon}\|\nabla\cdot e\|_0^2\right)+\alpha\int_0^t\|\nabla e_t\|_0^2d\tau\le\int_0^t\left(\frac{\chi}{\varepsilon}\|\nabla\cdot e\|_0^2+\|\nabla\phi^\varepsilon\|_0^2\right)d\tau+c\,\varepsilon^{\frac{3}{2}}.$$

An application of the Gronwall lemma leads to

$$\|e_t\|_0^2+\|\nabla\phi^\varepsilon\|_0^2+\frac{\chi}{\varepsilon}\|\nabla\cdot e\|_0^2+\int_0^t\|\nabla e_\tau\|_0^2d\tau\lesssim\varepsilon^{\frac{3}{2}}. \tag{4.16}$$

The proof is complete. □

4.2.2. L^2-estimate on the velocity. An estimation of the error on the velocity in the L^2-norm is given by the following lemma.

LEMMA 4.2. *Provided* u *and* p *are smooth enough in time and space, then*

$$\|\mathsf{u}-u^\varepsilon\|_{L^2(L^2(\Omega)^d)}\lesssim\varepsilon^{\frac{5+s}{4}}, \tag{4.17}$$

where s is the regularity index of the Stokes operator.

Proof. We multiply (4.7) by $S(e)$. Owing to Lemma 2.4 we infer

$$\frac{1}{2}\partial_t|e|_\star^2+a(e,S(e))=0.$$

Using Lemma 2.3 with $\mu=\sqrt{\varepsilon}$, we obtain

$$\frac{1}{2}\partial_t|e|_\star^2+\frac{1}{2}\|e\|_0^2\lesssim\varepsilon^{\alpha_1}\|\nabla\cdot e\|_0^2+\varepsilon^{-\alpha_2}\|e-P_He\|_0^2.$$

From the definition of ϕ^ε, it is clear that $\varepsilon\nabla\phi^\varepsilon=e-P_He$; we then derive from the estimates in Lemma 4.1 that

$$\frac{1}{2}\partial_t|e|_\star^2+\frac{1}{2}\|e\|_0^2\lesssim\varepsilon^{\alpha_1}\|\nabla\cdot e\|_0^2+\varepsilon^{1-\alpha_2}\varepsilon\|\nabla\phi^\varepsilon\|_0^2\lesssim\varepsilon^{\frac{5}{2}}(\varepsilon^{\alpha_1}+\varepsilon^{1-\alpha_2}).$$

Since $\alpha_1=1-\alpha_2$, we find

$$\frac{1}{2}\partial_t|e|_\star^2+\frac{1}{2}\|e\|_0^2\lesssim\varepsilon^{\frac{5}{2}+\alpha_1}=\varepsilon^{\frac{5+s}{2}}.$$

The proof is completed using an integration in time. □

4.3. Error estimates for the time discrete case. The main result in this paper is the following.

THEOREM 4.1. *Let $0<\chi<\frac{2\alpha}{d}$. Assuming that the solution to (2.7) is smooth enough in time and space, the solution $(u^k,\tilde{u}^k,p^k)$ to (4.1)–(4.3) satisfies the estimates*

$$\|\mathsf{u}_{\Delta t}-u_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)}+\|\mathsf{u}_{\Delta t}-\tilde{u}_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)}\lesssim\Delta t^{\frac{5+s}{4}},$$
$$\|\mathsf{u}_{\Delta t}-\tilde{u}_{\Delta t}\|_{\ell^2(H^1(\Omega)^d)}+\|\mathsf{p}_{\Delta t}-p_{\Delta t}\|_{\ell^2(L^2(\Omega))}\lesssim\Delta t^{\frac{3+s}{4}},$$

where s is the regularity index of the Stokes operator.

REMARK 4.2. *With full Stokes regularity, i.e., $s=1$, the L^2-norm of the error on the velocity is $\mathcal{O}(\Delta t^{\frac{3}{2}})$, and the H^1-norm of the error on the velocity and the L^2-norm of the error on the pressure are $\mathcal{O}(\Delta t)$. In view of Lemma 4.1 and of the first estimate in Lemma 4.3, we believe that the H^1-estimates can be improved up to $\mathcal{O}(\Delta t^{\frac{5}{4}})$ by a* *sophisticated argument using weighted seminorms in time as in* [18, 20]*. However, the details of this proof are beyond the scope of this paper. Numerical results reported in section* 5 *seem to confirm this conjecture, at least in two dimensions.*

The proof of Theorem 4.1 is carried out in a way similar to that of Theorem 4.1 in [10], but since there are several important differences in the proofs of the underlying lemmas, we give all the details. In particular the error analysis reveals why a homogeneous Dirichlet boundary condition must be enforced on ϕ^{k+1} on Γ_2; it explains also the origin of the factor χ in (4.3).

Let us first introduce some notation. For any sequence $\varphi^0, \varphi^1, \dots$, we set

$$\delta_t\varphi^k = \varphi^k - \varphi^{k-1}, \quad \delta_{tt}\varphi^k = \delta_t(\delta_t\varphi^k), \quad \delta_{ttt}\varphi^k = \delta_t(\delta_{tt}\varphi^k),$$

and

$$\begin{cases} e^k = \mathsf{u}(t^k) - u^k, & \tilde{e}^k = \mathsf{u}(t^k) - \tilde{u}^k, \\ \psi^k = \mathsf{p}(t^{k+1}) - p^k, & q^k = \mathsf{p}(t^k) - p^k. \end{cases} \tag{4.18}$$

It is straightforward to show that $(\tilde{u}^1, u^1, p^1)$ obtained by using the scheme (4.1)–(4.3), with BDF2 replaced by backward Euler, satisfies the following estimates:

$$\begin{aligned} \|e^1\|_0 + \|\tilde{e}^1\|_0 + \Delta t^{\frac{1}{2}}(\|\nabla e^1\|_0 + \|\nabla \tilde{e}^1\|_0) &\lesssim \Delta t^2, \\ \|q^1\|_0 &\lesssim \Delta t. \end{aligned} \tag{4.19}$$

Note that for any bilinear form $(\cdot,\cdot)$ and any sequences $a^0, a^1, \dots$, and $b^0, b^1, \dots$, the following holds:

$$\delta_t(a^{k+1}, b^{k+1}) = (\delta_t a^{k+1}, b^{k+1}) + (a^k, \delta_t b^{k+1}). \tag{4.20}$$

The error estimates of Theorem 4.1 are proved through a succession of lemmas. The following result is the discrete counterpart of Lemma 4.1.

Lemma 4.3. *Under the hypotheses of Theorem* 4.1*, we have*

$$\begin{aligned} \|\nabla\cdot\tilde{u}_{\Delta t}\|_{\ell^\infty(L^2(\Omega))} + \sqrt{\Delta t}\|\nabla\phi_{\Delta t}\|_{\ell^\infty(L^2(\Omega))} &\lesssim \Delta t^{\frac{5}{4}}, \\ \|\delta_t\tilde{e}_{\Delta t}\|_{\ell^2(H^1(\Omega)^d)} &\lesssim \Delta t^{\frac{7}{4}}, \\ \|\delta_t\tilde{e}_{\Delta t} - \delta_t e_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)} &\lesssim \Delta t^{\frac{9}{4}}. \end{aligned}$$

Proof. Upon defining

$$R^k = \partial_t\mathsf{u}(t^k) - \frac{3\mathsf{u}(t^k) - 4\mathsf{u}(t^{k-1}) + \mathsf{u}(t^{k-2})}{2\Delta t}, \tag{4.21}$$

then, for $k \geq 2$, the equations that control the time increments of the errors are

$$\begin{cases} \dfrac{3\delta_t\tilde{e}^{k+1} - 4\delta_t e^k + \delta_t e^{k-1}}{2\Delta t} + A\delta_t\tilde{e}^{k+1} + \nabla\delta_t\psi^k = \delta_t R^{k+1}, \\ \delta_t\tilde{e}^{k+1}|_{\Gamma_1} = 0, \quad (\delta_t\psi^k n - \nu(D\delta_t\tilde{e}^{k+1})n)|_{\Gamma_2} = 0 \end{cases} \tag{4.22}$$

and

$$\begin{cases} \dfrac{3}{2\Delta t}\delta_t e^{k+1} - \nabla\phi^{k+1} = \dfrac{3}{2\Delta t}\delta_t\tilde{e}^{k+1} - \nabla\phi^k, \\ \nabla\cdot\delta_t e^{k+1} = 0, \\ \delta_t e^{k+1}\cdot n|_{\Gamma_1} = 0, \quad \phi^{k+1}|_{\Gamma_2} = \phi^k|_{\Gamma_2} = 0. \end{cases} \tag{4.23}$$

We take the inner product of (4.22) with $4\Delta t\,\delta_t\tilde{e}^{k+1}$ and obtain

$$\begin{aligned}2(\delta_t\tilde{e}^{k+1}, 3\delta_t\tilde{e}^{k+1} - 4\delta_t e^k + \delta_t e^{k-1}) &+ 4\alpha\Delta t\|\nabla\delta_t\tilde{e}^{k+1}\|_0^2 \\ &- 4\Delta t(\nabla\cdot\delta_t\tilde{e}^{k+1}, \delta_t\psi^k) = 4\Delta t(\delta_t\tilde{e}^{k+1}, \delta_t R^{k+1}) \\ &\leq \gamma\alpha\Delta t\|\nabla\delta_t\tilde{e}^{k+1}\|_0^2 + c\Delta t^7,\end{aligned}\tag{4.24}$$

where γ will be chosen later, and we have used the coercivity of the bilinear form a together with the fact that $\|\delta_t R^{k+1}\|_0 \lesssim \Delta t^3$. Note also that we have used the inequality $2ab \leq \gamma a^2 + b^2/\gamma$, which holds for all $\gamma > 0$. We shall repeatedly use this standard trick hereafter without mentioning it anymore.

Let us denote $I = 2(\delta_t\tilde{e}^{k+1}, 3\delta_t\tilde{e}^{k+1} - 4\delta_t e^k + \delta_t e^{k-1})$; then we have

$$\begin{aligned}I = 6(\delta_t\tilde{e}^{k+1}, \delta_t\tilde{e}^{k+1} - \delta_t e^{k+1}) &+ 2(\delta_t\tilde{e}^{k+1} - \delta_t e^{k+1}, 3\delta_t e^{k+1} - 4\delta_t e^k + \delta_t e^{k-1}) \\ + 2(\delta_t e^{k+1}, 3\delta_t e^{k+1} &- 4\delta_t e^k + \delta_t e^{k-1}).\end{aligned}$$

Let I_1, I_2, and I_3 be the three terms in the right-hand side. Using the algebraic identities

$$2(a^{k+1}, a^{k+1} - a^k) = |a^{k+1}|^2 + |a^{k+1} - a^k|^2 - |a^k|^2,\tag{4.25}$$

$$\begin{aligned}2(a^{k+1}, 3a^{k+1} - 4a^k + a^{k-1}) = |a^{k+1}|^2 &+ |2a^{k+1} - a^k|^2 + |\delta_{tt}a^{k+1}|^2 \\ &- |a^k|^2 - |2a^k - a^{k-1}|^2,\end{aligned}\tag{4.26}$$

we derive

$$\begin{aligned}I_1 &= 3\|\delta_t\tilde{e}^{k+1}\|_0^2 + 3\|\delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}\|_0^2 - 3\|\delta_t e^{k+1}\|_0^2, \\ I_3 &= \|\delta_t e^{k+1}\|_0^2 + \|2\delta_t e^{k+1} - \delta_t e^k\|_0^2 + \|\delta_{ttt}e^{k+1}\|_0^2 - \|\delta_t e^k\|_0^2 - \|2\delta_t e^k - \delta_t e^{k-1}\|_0^2.\end{aligned}$$

Owing to (4.23) and using the fact that $e^k \in H$, we derive the following equality:

$$\frac{3}{2\Delta t}I_2 = -2(\nabla\delta_t\phi^{k+1}, 3\delta_t e^{k+1} - 4\delta_t e^k + \delta_t e^{k-1}) = 0.$$

Collecting all the above results, we obtain

$$\begin{aligned}3\|\delta_t\tilde{e}^{k+1}\|_0^2 - 3\|\delta_t e^{k+1}\|_0^2 &+ \|\delta_t e^{k+1}\|_0^2 + \|2\delta_t e^{k+1} - \delta_t e^k\|_0^2 \\ &+ 3\|\delta_t e^{k+1} - \delta_t\tilde{e}^{k+1}\|_0^2 + \|\delta_{ttt}e^{k+1}\|_0^2 \\ &+ (4-\gamma)\alpha\Delta t\|\nabla\delta_t\tilde{e}^{k+1}\|_0^2 - 4\Delta t(\nabla\cdot\delta_t\tilde{e}^{k+1}, \delta_t\psi^k) \\ &\leq c\,\Delta t^7 + \|\delta_t e^k\|_0^2 + \|2\delta_t e^k - \delta_t e^{k-1}\|_0^2.\end{aligned}\tag{4.27}$$

Taking the square of (4.23) and integrating over the domain, we obtain

$$\begin{aligned}3\|\delta_t e^{k+1}\|_0^2 + \frac{4}{3}\Delta t^2\|\nabla\phi^{k+1}\|_0^2 = 3\|\delta_t\tilde{e}^{k+1}\|_0^2 &+ \frac{4}{3}\Delta t^2\|\nabla\phi^k\|_0^2 \\ &+ 4\Delta t(\nabla\cdot\delta_t\tilde{e}^{k+1}, \phi^k).\end{aligned}\tag{4.28}$$

Note that integration by parts on $(\delta_t e^{k+1}, \nabla\phi^{k+1})$ and $(\delta_t\tilde{e}^{k+1}, \nabla\phi^k)$ is legitimate because both $\phi^{k+1}|_{\Gamma_2}$ and $\phi^k|_{\Gamma_2}$ are zero. Since $\phi^k = p^k - p^{k-1} - \chi\nabla\cdot\tilde{e}^k$, we can bound the inner product in the right-hand side of (4.28) as follows:

$$\begin{aligned}4\Delta t(\nabla\cdot\delta_t\tilde{e}^{k+1}, \phi^k) &= 4\Delta t(\nabla\cdot\delta_t\tilde{e}^{k+1}, p^k - p^{k-1} - \chi\nabla\cdot\tilde{e}^k) \\ &= 2\chi\Delta t(-\|\nabla\cdot\tilde{e}^{k+1}\|_0^2 + \|\nabla\cdot\tilde{e}^k\|_0^2 + \|\nabla\cdot\delta_t\tilde{e}^{k+1}\|_0^2) \\ &\quad - 4\Delta t(\nabla\cdot\delta_t\tilde{e}^{k+1}, \delta_t\psi^k) + 4\Delta t(\nabla\cdot\delta_t\tilde{e}^{k+1}, \delta_t\mathsf{p}(t^{k+1})).\end{aligned}\tag{4.29}$$

To control the troublesome term $\Delta t\|\nabla\cdot\delta_t\tilde{e}^{k+1}\|_0^2$ we use

$$\chi\|\nabla\cdot v\|_0^2 \le 2\gamma'\alpha\|\nabla v\|_0^2 \quad \forall v \in X. \tag{4.30}$$

Due to the condition χ, (4.4), we know that the constant γ' is such that $0 < \gamma' < 1$. Summing (4.27), (4.28), and (4.29), and using (4.30), we finally obtain

$$\begin{aligned}&\|\delta_t e^{k+1}\|_0^2 + \|2\delta_t e^{k+1} - \delta_t e^k\|_0^2 + \frac{4}{3}\Delta t^2\|\nabla\phi^{k+1}\|_0^2 + 2\chi\Delta t\|\nabla\cdot\tilde{e}^{k+1}\|_0^2\\ &\qquad + (4-4\gamma'-\gamma)\alpha\Delta t\|\nabla\delta_t\tilde{e}^{k+1}\|_0^2 + 3\|\delta_t(e^{k+1}-\tilde{e}^{k+1})\|_0^2 + \|\delta_{ttt}e^{k+1}\|_0^2\\ &\quad \le \|\delta_t e^k\|_0^2 + \|2\delta_t e^k - \delta_t e^{k-1}\|_0^2 + \frac{4}{3}\Delta t^2\|\nabla\phi^k\|_0^2 + 2\chi\Delta t\|\nabla\cdot\tilde{e}^k\|_0^2\\ &\qquad + 4\Delta t(\nabla\cdot\delta_t\tilde{e}^{k+1}, \delta_t\mathsf{p}(t^{k+1})) + c\Delta t^7.\end{aligned} \tag{4.31}$$

At this point, we are formally at the same stage as (4.14). To integrate by parts in time the term $(\nabla\cdot\delta_t\tilde{e}^{k+1}, \delta_t\mathsf{p}(t^{k+1}))$, we use (4.20) as follows:

$$(\nabla\cdot\delta_t\tilde{e}^{k+1}, \delta_t\mathsf{p}(t^{k+1})) = \delta_t(\nabla\cdot\tilde{e}^{k+1}, \delta_t\mathsf{p}(t^{k+1})) - (\nabla\cdot\tilde{e}^k, \delta_{tt}\mathsf{p}(t^{k+1})).$$

Next, we use the interpolation operator defined in (4.15). Let us denote $\mathcal{R}^{k+1} = \mathsf{p}(t^{k+1}) - \mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1}))$ (where $\mathcal{J}_{\Delta t} = \mathcal{I}_{\sqrt{\Delta t},1}$). Then we have

$$\frac{1}{\Delta t}\|\delta_{tt}\mathcal{R}^{k+1}\|_0^2 + \|\nabla\delta_{tt}\mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1}))\|_0^2 \lesssim \Delta t^{\frac{7}{2}}.$$

Since $\mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1}))$ is zero on Γ_2, we have

$$\begin{aligned}(\nabla\cdot\delta_t\tilde{e}^{k+1}, \delta_t\mathsf{p}(t^{k+1})) &= \delta_t(\nabla\cdot\tilde{e}^{k+1}, \delta_t\mathcal{R}^{k+1}) + \delta_t(\nabla\cdot\tilde{e}^{k+1}, \delta_t\mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1})))\\ &\quad - (\nabla\cdot\tilde{e}^k, \delta_{tt}\mathcal{R}^{k+1}) - (\nabla\cdot\tilde{e}^k, \delta_{tt}\mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1})))\\ &= \delta_t(\nabla\cdot\tilde{e}^{k+1}, \delta_t\mathcal{R}^{k+1}) + \frac{2\Delta t}{3}\delta_t(\nabla\phi^{k+1}, \nabla\delta_t\mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1})))\\ &\quad - (\nabla\cdot\tilde{e}^k, \delta_{tt}\mathcal{R}^{k+1}) - \frac{2\Delta t}{3}(\nabla\phi^k, \nabla\delta_{tt}\mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1})))\\ &\le \delta_t(\nabla\cdot\tilde{e}^{k+1}, \delta_t\mathcal{R}^{k+1}) + \frac{2\Delta t}{3}\delta_t(\nabla\phi^{k+1}, \nabla\delta_t\mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1})))\\ &\quad + \frac{\chi\Delta t}{2}\|\nabla\cdot\tilde{e}^k\|_0^2 + \frac{\Delta t^2}{3}\|\nabla\phi^k\|_0^2 + c\Delta t^{\frac{7}{2}}.\end{aligned}$$

By inserting this bound into (4.31), we obtain

$$\begin{aligned}&\|\delta_t e^{k+1}\|_0^2 + \|2\delta_t e^{k+1} - \delta_t e^k\|_0^2 + \frac{4}{3}\Delta t^2\|\nabla\phi^{k+1}\|_0^2 + 2\chi\Delta t\|\nabla\cdot\tilde{e}^{k+1}\|_0^2\\ &\qquad + (4-4\gamma'-\gamma)\alpha\Delta t\|\nabla\delta_t\tilde{e}^{k+1}\|_0^2 + 3\|\delta_t(e^{k+1}-\tilde{e}^{k+1})\|_0^2 + \|\delta_{ttt}e^{k+1}\|_0^2\\ &\quad \le \|\delta_t e^k\|_0^2 + \|2\delta_t e^k - \delta_t e^{k-1}\|_0^2\\ &\qquad + \frac{4}{3}\Delta t^2(1+\Delta t)\|\nabla\phi^k\|_0^2 + 2\chi\Delta t(1+\Delta t)\|\nabla\cdot\tilde{e}^k\|_0^2\\ &\qquad + 4\Delta t\delta_t(\nabla\cdot\tilde{e}^{k+1}, \delta_t\mathcal{R}^{k+1}) + \frac{8\Delta t^2}{3}\delta_t(\nabla\phi^{k+1}, \nabla\delta_t\mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1}))) + c\,\Delta t^{\frac{9}{2}}.\end{aligned}$$

Summing up the relation above for $l = 2, \dots, k$ and taking into account (4.19), we obtain

$$\begin{aligned}
&\|\delta_t e^{k+1}\|_0^2 + \|2\delta_t e^{k+1} - \delta_t e^k\|_0^2 + \frac{4}{3}\Delta t^2 \|\nabla \phi^{k+1}\|_0^2 + 2\chi \Delta t \|\nabla\cdot \tilde{e}^{k+1}\|_0^2 \\
&\qquad + (4 - 4\gamma' - \gamma)\alpha \Delta t \sum_{l=2}^{k} \|\nabla \delta_t \tilde{e}^{l+1}\|_0^2 + 3\sum_{l=2}^{k} \|\delta_t e^{l+1} - \delta_t \tilde{e}^{l+1}\|_0^2 \\
&\quad \le c\left(\|\delta_t e^2\|_0^2 + \|2\delta_t e^2 - \delta_t e^1\|_0^2 + \Delta t^2 \|\nabla\phi^2\|_0^2 + \Delta t \|\nabla\cdot \tilde{e}^2\|_0^2 + \Delta t^{\frac{7}{2}}\right) \\
&\qquad + \Delta t \sum_{l=2}^{k} \left(\frac{4}{3}\Delta t^2 \|\nabla\phi^l\|_0^2 + 2\chi\Delta t \|\nabla\cdot \tilde{e}^l\|_0^2\right) \\
&\qquad - 4\Delta t(\nabla\cdot \tilde{e}^{k+1}, \delta_t \mathcal{R}^{k+1}) - \frac{8\Delta t^2}{3}(\nabla\phi^{k+1}, \nabla \delta_t \mathcal{J}_{\Delta t}(\mathsf{p}(t^{k+1}))) \\
&\qquad + 4\Delta t(\nabla\cdot \tilde{e}^2, \delta_t \mathcal{R}^2) + \frac{8\Delta t^2}{3}(\nabla\phi^2, \nabla\delta_t \mathcal{J}_{\Delta t}(\mathsf{p}(t^2))) \\
&\quad \le c\Delta t^{\frac{7}{2}} + \frac{2}{3}\Delta t^2 \|\nabla\phi^{k+1}\|_0^2 + \chi\Delta t \|\nabla\cdot \tilde{e}^{k+1}\|_0^2 \\
&\qquad + \Delta t \sum_{l=2}^{k} \left(\frac{4}{3}\Delta t^2 \|\nabla\phi^l\|_0^2 + 2\chi\Delta t\|\nabla\cdot\tilde{e}^l\|_0^2\right).
\end{aligned}$$

Since $0 < \gamma' < 1$, we can choose γ such that $4 - 4\gamma' - \gamma \ge 0$. Then an application of the discrete Gronwall lemma yields the desired result. □

REMARK 4.3. *Note that to balance the term* $-(\nabla\cdot\delta_t \tilde{e}^{k+1}, \psi^k)$ *in* (4.27) *it is necessary to integrate by parts the term* $(\delta_t \tilde{e}^{k+1}, \nabla\phi^k)$ *in* (4.28). *This is possible only because the Dirichlet boundary condition* $\phi^k|_{\Gamma_2} = 0$ *is enforced. This fact is the main reason why we enforce a homogeneous Dirichlet boundary condition on* ϕ^{k+1} *in* (4.2). *This argument shows the importance of the error analysis (or stability analysis) performed in the proof of Lemma* 4.3. *The necessity of the Dirichlet boundary condition also becomes clear when one understands that* (4.2) *is a realization of* $u^{k+1} = P_H \tilde{u}^{k+1}$, *since the orthogonal complement of* H *is* ∇N *according to Lemma* 2.1.

REMARK 4.4. *The introduction of the parameter* χ *together with the bound* (4.4) *is justified by step* (4.30). *Whether the bound* (4.4) *is sharp is not yet clear.*

LEMMA 4.4. *Under the hypotheses of Theorem* 4.1, *we have*

$$\|\mathsf{u}_{\Delta t} - \tilde{u}_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)} + \|\mathsf{u}_{\Delta t} - u_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)} \lesssim \Delta t^{\frac{5+s}{4}}.$$

Proof. By using the relation $e^l = \tilde{e}^l + \frac{2\Delta t}{3}\nabla\phi^l$, for all $l \ge 2$, one obtains

$$(4.32)\qquad \begin{cases} \dfrac{3\tilde{e}^{k+1} - 4\tilde{e}^k + \tilde{e}^{k-1}}{2\Delta t} + A\tilde{e}^{k+1} + \nabla\gamma^k = R^{k+1}, \\ \tilde{e}^{k+1}|_{\Gamma_1} = 0, \quad (\gamma^k n - \nu(D\tilde{e}^{k+1})n)|_{\Gamma_2} = 0, \end{cases}$$

where $\nabla\gamma^k$ stands for the collection of all the gradient terms.

As in the time continuous case, we make use of the inverse Stokes operator. By taking the inner product of (4.32) with $4\Delta t S(\tilde{e}^{k+1})$ and using the identity (4.26), we obtain

$$\begin{aligned}
&|\tilde{e}^{k+1}|_\star^2 + |2\tilde{e}^{k+1} - \tilde{e}^k|_\star^2 + |\delta_{tt}\tilde{e}^{k+1}|_\star^2 + 4\Delta t\, a(\tilde{e}^{k+1}, S(\tilde{e}^{k+1})) \\
&\qquad\qquad = 4\Delta t\,(R^{k+1}, S(\tilde{e}^{k+1})) + |\tilde{e}^k|_\star^2 + |2\tilde{e}^k - \tilde{e}^{k-1}|_\star^2.
\end{aligned}$$

Using Lemma 2.3 with $\mu = \sqrt{\Delta t}$ and Lemma 4.3, we infer

$$\begin{aligned}4a(\tilde{e}^{k+1}, S(\tilde{e}^{k+1})) &\geq 2\|\tilde{e}^{k+1}\|_0^2 - c(\Delta t^{\alpha_1}\,\|\nabla\cdot\tilde{e}^{k+1}\|_0^2 + \Delta t^{-\alpha_2}\,\|\tilde{e}^{k+1} - e^{k+1}\|^2)\\ &\geq 2\|\tilde{e}^{k+1}\|_0^2 - c(\Delta t^{\alpha_1}\,\|\nabla\cdot\tilde{e}^{k+1}\|_0^2 + \Delta t^{1-\alpha_2}\,\Delta t\|\nabla\phi^{k+1}\|^2\\ &\geq 2\|\tilde{e}^{k+1}\|_0^2 - c\Delta t^{\alpha_1+\frac{5}{2}} \geq 2\|\tilde{e}^{k+1}\|_0^2 - c\Delta t^{\frac{5+s}{2}}.\end{aligned}$$

We also derive from the Cauchy–Schwarz inequality and (2.16) that

$$4\Delta t(R^{k+1}, S(\tilde{e}^{k+1})) \leq c\Delta t\|R^{k+1}\|_{X'}^2 + \Delta t\|\tilde{e}^{k+1}\|_0^2 \leq c'\Delta t^5 + \Delta t\|\tilde{e}^{k+1}\|_0^2.$$

Combining these two estimates, we obtain

$$|\tilde{e}^{k+1}|_\star^2 + |2\tilde{e}^{k+1} - \tilde{e}^k|_\star^2 + \Delta t\|\tilde{e}^{k+1}\|_0^2 \leq |\tilde{e}^k|_\star^2 + |2\tilde{e}^k - \tilde{e}^{k-1}|_\star^2 + c\Delta t^{1+\frac{5+s}{2}}.$$

The desired result is now an easy consequence of the discrete Gronwall lemma. The estimate on $\|\mathsf{u}_{\Delta t} - u_{\Delta t}\|_0$ is obtained by using the triangular inequality $\|\mathsf{u}_{\Delta t} - u_{\Delta t}\|_0 \leq \|\mathsf{u}_{\Delta t} - \tilde{u}_{\Delta t}\|_0 + \frac{2\Delta t}{3}\|\nabla\phi_{\Delta t}\|_0$ (derived from (4.2)) and Lemma 4.3. □

The key for obtaining improved estimates on $\|\tilde{e}_{\Delta t}\|_{\ell^2(H^1(\Omega)^d)}$ and $\|q_{\Delta t}\|_{\ell^2(L^2(\Omega))}$ is to derive an improved estimate on $\frac{1}{2\Delta t}(3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1})$. To this end, for any sequence of functions $\phi^0, \phi^1, \ldots$, we define

$$D_t\phi^{k+1} := \frac{1}{2}(3\phi^{k+1} - 4\phi^k + \phi^{k-1}).$$

Lemma 4.5. *Under the hypotheses of Theorem* 4.1, *we have*

$$\Delta t^{-1}\|(D_t\tilde{e})_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)} \lesssim \Delta t^{\frac{3+s}{4}}.$$

Proof. We use the same argument as in the proof of the L^2-estimate, but we use it on the time increment $\delta_t\tilde{e}^{k+1}$. For $k \geq 2$ we have

$$\frac{3\delta_t\tilde{e}^{k+1} - 4\delta_t\tilde{e}^k + \delta_t\tilde{e}^{k-1}}{2\Delta t} + A\delta_t\tilde{e}^{k+1} + \nabla\delta_t\gamma^{k+1} = \delta_t R^{k+1}.$$

Taking the inner product of the above relation with $4\Delta t S(\delta_t\tilde{e}^{k+1})$, using Lemma 2.3 with $\mu = \sqrt{\Delta t}$, and repeating the same arguments as in the previous lemma, we obtain

$$\begin{aligned}&|\delta_t\tilde{e}^{k+1}|_\star^2 + |2\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^k|_\star^2 + |\delta_{ttt}\tilde{e}^{k+1}|_\star^2 + \Delta t\|\delta_t\tilde{e}^{k+1}\|_0^2\\ &\quad\leq c\Delta t\|\delta_t R^{k+1}\|_0^2 + c\Delta t(\Delta t^{\alpha_1}\|\nabla\cdot\delta_t\tilde{e}^{k+1}\|_0^2 + \Delta t^{-\alpha_2}\|\delta_t\tilde{e}^{k+1} - \delta_t e^{k+1}\|_0^2)\\ &\qquad + |\delta_t\tilde{e}^k|_\star^2 + |2\delta_t\tilde{e}^k - \delta_t\tilde{e}^{k-1}|_\star^2.\end{aligned}$$

Applying the discrete Gronwall lemma, and using the initial estimates and Lemma 4.3, we obtain

$$\|\delta_t\tilde{e}_{\Delta t}\|_{l^2(L^2(\Omega)^d)}^2 \lesssim \Delta t^{\frac{7+s}{2}}.$$

We conclude by using the fact that $2D_t\tilde{e}^{k+1} = 3\delta_t\tilde{e}^{k+1} - \delta_t\tilde{e}^k$. □

We are now in position to prove the remaining claims in Theorem 4.1.

Lemma 4.6. *Under the hypotheses of Theorem* 4.1, *we have*

$$\|\mathsf{u}_{\Delta t} - \tilde{u}_{\Delta t}\|_{\ell^2(H^1(\Omega)^d)} + \|\mathsf{p}_{\Delta t} - p_{\Delta t}\|_{\ell^2(L^2(\Omega))} \lesssim \Delta t^{\frac{3+s}{4}}.$$

Proof. By adding the viscous step and the projection step, it is clear that we have

$$\begin{cases} A\tilde{e}^{k+1} + \nabla(q^{k+1} + \chi\nabla\cdot\tilde{e}^{k+1}) = h^{k+1}, \\ \nabla\cdot\tilde{e}^{k+1} = g^{k+1}, \quad \tilde{e}^{k+1}|_{\Gamma_1} = 0, \quad ((q^{k+1} + \chi\nabla\cdot\tilde{e}^{k+1})n - (D\tilde{e}^{k+1})n)|_{\Gamma_2} = 0, \end{cases} \tag{4.33}$$

where

$$h^{k+1} = R^{k+1} - \frac{D_t e^{k+1}}{\Delta t}, \qquad g^{k+1} = -\frac{2\Delta t}{3}\nabla^2\phi^{k+1}. \tag{4.34}$$

Owing to Lemma 4.3, we have

$$\|g^{k+1}\|_0 = \|\nabla\cdot\tilde{e}^{k+1}\|_0 \lesssim \Delta t^{\frac{5}{4}} \qquad \forall k. \tag{4.35}$$

Since $e^k = P_H\tilde{e}^k$, owing to Lemma 4.5, we infer

$$\Delta t^{-1}\|\delta_t e_{\Delta t}\|_{l^2(L^2(\Omega)^d)} \leq \Delta t^{-1}\|\delta_t\tilde{e}_{\Delta t}\|_{l^2(L^2(\Omega)^d)} \lesssim \Delta t^{\frac{3+s}{4}}.$$

Hence, we have

$$\|h_{\Delta t}\|_{\ell^2(X')} \lesssim \|R_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)} + \Delta t^{-1}\|D_t\tilde{e}_{\Delta t}\|_{\ell^2(L^2(\Omega)^d)} \lesssim \Delta t^{\frac{3+s}{4}}. \tag{4.36}$$

Now, we apply the following standard stability result for nonhomogeneous Stokes systems to (4.33) (cf. [23]):

$$\|\tilde{e}^{k+1}\|_1 + \|(q^{k+1} + \chi\nabla\cdot\tilde{e}^{k+1})\|_0 \lesssim \|h^{k+1}\|_{X'} + \|g^{k+1}\|_0. \tag{4.37}$$

Owing to (4.35) and (4.36), we derive

$$\|\tilde{e}_{\Delta t}\|_{\ell^2(H^1(\Omega)^d)} + \|(q + \chi\nabla\cdot\tilde{e})\|_{\ell^2(L^2(\Omega))} \lesssim \Delta t^{\frac{3+s}{4}}.$$

Then, from

$$\|q^{k+1}\|_0 \leq \|q^{k+1} + \chi\nabla\cdot\tilde{e}^{k+1}\|_0 + \chi\|\nabla\cdot\tilde{e}^{k+1}\|_0,$$

we derive $\|q_{\Delta t}\|_{l^2(L^2(\Omega))} \lesssim \Delta t^{\frac{3+s}{4}}$. □

Thus, all the results in Theorem 4.1 have been proved.

5. Numerical results and discussions.

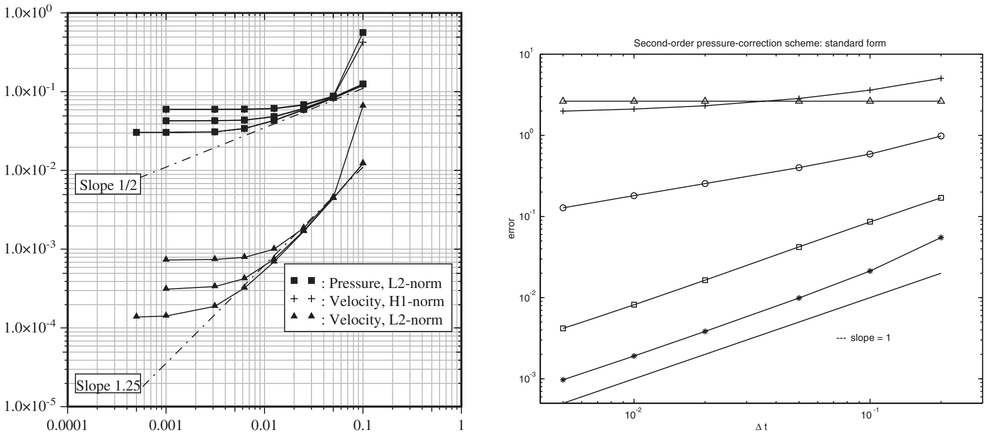
5.1. Standard pressure-correction scheme. We take the exact solution $(\mathsf{u}_1, \mathsf{u}_2, \mathsf{p})$ of the linearized Navier–Stokes equations to be

$$\mathsf{u}_1(x,y,t) = \sin x\sin(y+t), \ \mathsf{u}_2(x,y,t) = \cos x\cos(y+t), \ \mathsf{p}(x,y,t) = \cos x\sin(y+t).$$

We set $\Omega =]0,1[^2$, $\Gamma_2 = \{(x,y) \in \Gamma, \, x = 0\}$. This solution satisfies the following open boundary conditions:

$$-\partial_x\mathsf{u}_2|_{\Gamma_2} = 0, \qquad \mathsf{p} - \partial_x\mathsf{u}_1|_{\Gamma_2} = 0.$$

To confirm the results in Theorem 3.1, we have carried out convergence tests in time using $\mathbb{P}_2/\mathbb{P}_1$ finite elements as well as the $\mathbb{P}_N^2 \times \mathbb{P}_{N-2}$ Legendre–Galerkin method [19] (where $\mathbb{P}_k$ denotes the space of polynomials of degree less than or equal



Finite elements with $h = 1/20$, $1/40$, and $1/80$. Legendre–Galerkin with $N = 40$.

FIG. 5.1. *Errors vs.* Δt, *standard pressure-correction scheme: Note that the curves corresponding to the error on the velocity in* H^1*-norm and the pressure in* L^2*-norm almost coincide.*

to k). We use the standard BDF2 pressure-correction scheme, which enforces a homogeneous Dirichlet boundary condition on the pressure increment at the open boundary in the projection step.

For the finite elements, the errors at $t = 1$ for three meshes ($h = 1/20$, $1/40$, $1/80$) and $5.10^{-4} \leq \Delta t \leq 10^{-1}$ are reported in the left panel of Figure 5.1. Note that the error for small time steps is dominated by the spatial discretization error. The reference slope represents the asymptotic convergence rate as $h \to 0$.

For the Legendre–Galerkin method, the results with $N = 40$ are reported in the right panel of Figure 5.1. For the range of time steps explored, the spatial discretization error is negligible compared to the time discretization error.

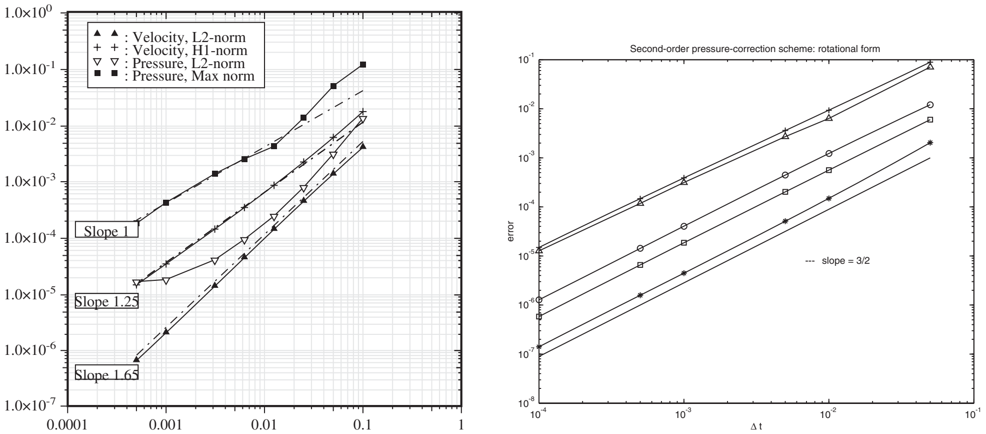
These tests clearly indicate that the L^2-error of the velocity (resp., the pressure) is of order Δt (resp., $\Delta t^{\frac{1}{2}}$), which are consistent with Theorem 3.1.

5.2. Rotational pressure-correction scheme. We again use the analytical solution described above to test the time accuracy of the rotational pressure-correction scheme (4.1)–(4.3).

We first report the results with $\mathbb{P}_2/\mathbb{P}_1$ finite elements. We use $h = 1/80$ to guarantee that the error in space is significantly smaller than the splitting error. The results are reported in the left panel of Figure 5.2. The convergence rate of the error on the velocity in the L^2-norm is close to $\mathcal{O}(\Delta t^{3/2})$, and that of the H^1-norm behaves like $\mathcal{O}(\Delta t^{5/4})$, which is higher than the $\mathcal{O}(\Delta t)$ rate predicted by Theorem 4.1 (see Remark 4.2 and Lemma 4.3). The convergence rate of the error on the pressure in the L^∞-norm is $\mathcal{O}(\Delta t)$, and that of the L^2-norm is between $\mathcal{O}(\Delta t)$ and $\mathcal{O}(\Delta t^{\frac{3}{2}})$. These rates are mostly consistent with the error estimates in Theorem 4.1. The accuracy saturation observed for small time steps comes from the spatial discretization error.

The results using the Legendre–Galerkin method are reported in the right panel of Figure 5.2. We note that the convergence rate for the error on the velocity in the L^2-norm is of order $O(\Delta t^{\frac{3}{2}})$, as predicted by Theorem 4.1. The convergence rates on all the other quantities are also close to $O(\Delta t^{\frac{3}{2}})$, which is higher than what Theorem 4.1 predicts (see Remark 4.2).

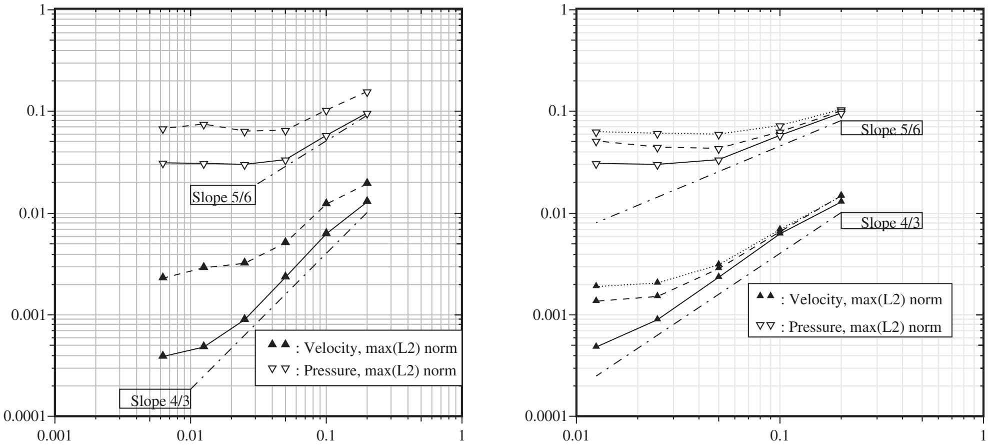
To complete this series of tests, we have performed convergence tests in three



Velocity: (▲) L^2-norm; (+) H^1-norm.
Pressure: (▽) L^2-norm; (■) L^∞-norm.

Velocity: (*) L^2-norm; (□) H^1-norm.
Pressure: (○) L^2; (+) H^1; (△) L^∞.

FIG. 5.2. *Rotational pressure-correction scheme: Left, finite elements; errors at $t = 1$ vs. Δt (using $h = 1/80$). Right, spectral method; error vs. Δt with $N = 40$ fixed.*



$h = 1/40$. Standard form (dashes) vs. rotational form (solid line)

$h = 1/10$, $1/20$, $1/40$. Rotational form, $\chi = 2/3$.

FIG. 5.3. *Pressure-correction scheme with $\mathbb{P}_2/\mathbb{P}_1$ finite elements in three dimensions. Errors vs. Δt. Velocity: (▲) L^2-norm. Pressure: (▽) L^2-norm.*

dimensions using $\mathbb{P}_2/\mathbb{P}_1$ finite elements. The boundary conditions and the source term in the Stokes equations are set so that the solution is given by

$$\mathsf{u}_1(x,y,z,t) = \sin x \sin(y+z+t),\ \mathsf{u}_2(x,y,z,t) = \cos x \cos(y+z+t),$$
$$\mathsf{u}_3(x,y,z,t) = \cos(x)\sin(y+t),\ \mathsf{p}(x,y,t) = \cos x \sin(y+z+t).$$

Both the standard and the rotational forms of the BDF2 pressure-correction scheme were tested. We show in Figure 5.3 the maximum in time of the L^2-norm of the errors on the velocity and the pressure for both schemes. On the left panel we compare the standard and rotational forms of the scheme using $h = 1/40$. Unfortunately, using a higher uniform resolution in space was not possible due to the high cost of the computations. The grid with a stepsize $h = 1/40$ already contains close to

500,000 $\mathbb{P}_2$ nodes. On the right panel we show the errors for the rotational form of the scheme using three different meshes: $h = 1/10$, $1/20$, $1/40$. The convergence rates of the standard version of the scheme are clearly lower than those of the rotational form. The slopes for both the velocity and the pressure errors obtained with the rotational form of the scheme are slightly lower than the best possible estimate following from the claim of Theorem 4.1. The rates $\mathcal{O}(\Delta t^{\frac{4}{3}})$ and $\mathcal{O}(\Delta t^{\frac{5}{6}})$ seem to correspond to a regularity index $s < 1$.

6. Concluding remarks. In this paper, we have analyzed pressure-correction schemes for approximating the incompressible Navier–Stokes equations with prescribed normal stress boundary conditions enforced on parts of the boundary. Our conclusions are twofold.

First, we have shown that the convergence rates of standard pressure-correction methods are too poor to be recommendable for approximating the Navier–Stokes equations in these circumstances. The main reason for the poor accuracy is that an *artificial* homogeneous Dirichlet boundary condition on the pressure has to be imposed to ensure stability.

Second, we have shown that the rotational pressure-correction method leads to reasonably good error estimates. More precisely, assuming full regularity of the Stokes problem, we have shown that the second-order rotational pressure-correction method yields $\mathcal{O}(\Delta t^{3/2})$ accuracy for the velocity in the L^2-norm and $\mathcal{O}(\Delta t)$ accuracy for the velocity in the H^1-norm and the pressure in the L^2-norm. To the best of our knowledge, the results presented in this paper are the first published convergence estimates for a splitting method solving the time-dependent Stokes equations with open boundary conditions.

Finally, it is clear that even though the second-order rotational pressure-correction method yields the best error estimates to date, these are still suboptimal and more research is needed to find a splitting scheme with better properties.

REFERENCES

[1] S. C. Brenner and R. L. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer, New York, 1994.

[2] A. J. Chorin, *On the convergence of discrete approximations to the Navier-Stokes equations*, Math. Comp., 23 (1969), pp. 341–353.

[3] W. E and J.-G. Liu, *Projection method.* I. *Convergence and numerical boundary layers*, SIAM J. Numer. Anal., 32 (1995), pp. 1017–1057.

[4] P. M. Gresho and S. T. Chan, *On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via finite element method that also introduces a nearly consistent mass matrix. Part* I, Internat. J. Numer. Methods Fluids, 11 (1990), pp. 587–620.

[5] P. M. Gresho and S. T. Chan, *On the theory of semi-implicit projection methods for viscous incompressible flow and its implementation via finite element method that also introduces a nearly consistent mass matrix. Part* II, Internat. J. Numer. Methods Fluids, 11 (1990), pp. 621–659.

[6] J.-L. Guermond, *Some practical implementations of projection methods for Navier-Stokes equations*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 637–667. Also in C. R. Acad. Sci. Paris Sér. I Math., 319 (1994), pp. 887–892.

[7] J.-L. Guermond, *Un résultat de convergence d'ordre deux en temps pour l'approximation des équations de Navier-Stokes par une technique de projection incrémentale*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 169–189. Also in C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 1329–1332.

[8] J.-L. Guermond and L. Quartapelle, *On the approximation of the unsteady Navier–Stokes equations by finite element projection methods*, Numer. Math., 80 (1998), pp. 207–238.

[9] J. L. Guermond and J. Shen, *Velocity-correction projection methods for incompressible flows*, SIAM J. Numer. Anal., 41 (2003), pp. 112–134.

[10] J.-L. Guermond and J. Shen, *On the error estimates for the rotational pressure-correction projection methods*, Math. Comp., 73 (2004), pp. 1719–1737.

[11] J. L. Guermond and J. Shen, *Quelques résultats nouveaux sur les méthodes de projection*, C. R. Acad. Sci. Paris Sér. I Math., 333 (2001), pp. 1111–1116.

[12] J. L. Guermond, P. Minev, and J. Shen, *An overview of projection methods for incompressible flows*, Comput. Methods Appl. Mech. Engrg., to appear.

[13] M. J. Lee, B. D. Oh, and Y. B. Kim, *Canonical fractional-step methods and consistent boundary conditions for the incompressible Navier–Stokes equations*, J. Comput. Phys., 168 (2001), pp. 73–100.

[14] M. Orlt and A.-M. Sändig, *Regularity of viscous Navier-Stokes flows in nonsmooth domains*, in Boundary Value Problems and Integral Equations in Nonsmooth Domains (Luminy, 1993), S. Nicaise, M. Costabel, and M. Dauge, eds., Lecture Notes in Pure and Appl. Math. 167, Marcel Dekker, New York, 1995, pp. 185–201.

[15] J. B. Perot, *An analysis of the fractional step method*, J. Comput. Phys., 108 (1993), pp. 51–58.

[16] A. Quarteroni, F. Saleri, and A. Veneziani, *Analysis of the Yosida method for the incompressible Navier-Stokes equations*, J. Math. Pures Appl. (9), 78 (1999), pp. 473–503.

[17] A. Quarteroni, F. Saleri, and A. Veneziani, *Factorization methods for the numerical approximation of Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 505–526.

[18] R. Rannacher, *On Chorin's projection method for the incompressible Navier-Stokes equations*, in The Navier-Stokes Equations II—Theory and Numerical Methods (Oberwolfach, 1991), Lecture Notes in Math. 1530, Springer, Berlin, 1992, pp. 167–183.

[19] J. Shen, *Efficient spectral-Galerkin method.* I. *Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[20] J. Shen, *A new pseudo-compressibility method for the Navier-Stokes equations*, Appl. Numer. Math., 21 (1996), pp. 71–90.

[21] J. Shen, *On error estimates of projection methods for the Navier-Stokes equations: Second-order schemes*, Math. Comp., 65 (1996), pp. 1039–1065.

[22] R. Temam, *Sur l'approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires* II, Arch. Ration. Mech. Anal., 33 (1969), pp. 377–385.

[23] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland, Amsterdam, 1984.

[24] L. J. P. Timmermans, P. D. Minev, and F. N. Van De Vosse, *An approximate projection scheme for incompressible flow using spectral elements*, Internat. J. Numer. Methods Fluids, 22 (1996), pp. 673–688.

[25] J. van Kan, *A second-order accurate pressure-correction scheme for viscous incompressible flow*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 870–891.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 259–279

POLYNOMIAL FITTING FOR EDGE DETECTION IN IRREGULARLY SAMPLED SIGNALS AND IMAGES*

RICK ARCHIBALD[†], ANNE GELB[‡], AND JUNGHO YOON[§]

Abstract. We propose a new edge detection method that is effective on multivariate irregular data in any domain. The method is based on a local polynomial annihilation technique and can be characterized by its convergence to zero for any value away from discontinuities. The method is numerically cost efficient and entirely independent of any specific shape or complexity of boundaries. Application of the *minmod* function to the edge detection method of various orders ensures a high rate of convergence away from the discontinuities while reducing the inherent oscillations near the discontinuities. It further enables distinction of jump discontinuities from steep gradients, even in instances where only sparse nonuniform data is available. These results are successfully demonstrated in both one and two dimensions.

Key words. *minmod* function, multivariate edge detection, Newton divided differencing, nonuniform grids

AMS subject classifications. 41A25, 41A45, 41A63

DOI. 10.1137/S0036142903435259

1. Introduction.

Edge detection is of fundamental importance in image analysis. In particular, a reliable and efficient edge detection method can both provide the possibility of processing an image with high accuracy as well as serve to simplify the analysis of images by drastically reducing the amount of data to be processed. Among the many common criteria relevant to edge detector performance, there are two very important issues. The first and most obvious issue is the possibility of failing to find real edge points and/or falsely identifying nonedge points. Regardless of specific types of data (regular or irregular) and domains, it is imperative that the edges occurring in the image should not be missed, and that there be no spurious responses. This is critical since the edges of the image constitute piecewise smooth regions. Hence errors in edge identification could also have drastic consequences on image reconstruction. The second issue is the necessity for simple implementation and cost efficiency.

To address these issues, this study constructs an edge detection method based on local Taylor expansions. Indeed, several well-known methods exist in the univariate case (see, e.g., [2], [3], [4], [5], [11], and the references therein). However, for the bivariate case, particularly with irregular points, no successful method has thus far been developed. Recent developments (see [1] and [15]) in essentially nonoscillatory (ENO) and weighted essentially nonoscillatory (WENO) methods for hyperbolic conservation laws and Hamilton–Jacobi equations on multidimensional unstructured meshes also utilize Taylor expansions to determine regions of analyticity. For uniform sampling,

*Received by the editors August 28, 2003; accepted for publication (in revised form) September 3, 2004; published electronically May 27, 2005. This research was supported in part by NSF grants CNS 0324957, DMS 0107428, EAR 0222327, and KRF-2002-070-C00014.

http://www.siam.org/journals/sinum/43-1/43525.html

[†] Department of Neuroscience, Brown Medical School, Box 1953, Providence, RI 02912 (Richard_Archibald@brown.edu). The research of this author was supported in part by the SSERC.

[‡] Department of Mathematics and Statistics, Arizona State University, Tempe, Arizona 85287-1804 (ag@math.la.asu.edu). The research of this author was supported in part by NIH grant EB 02533-01.

[§] Department of Mathematics, Ewha W. University, Seoul, 120-750, South Korea (yoon@math.ewha.ac.kr).

we note that while ENO and WENO methods identify stencils yielding the "most" smooth polynomial interpolations, they do not distinguish between, say, steep gradients and edges. In our method, Taylor expansions are used for the exclusive purpose of determining the true edges of an image by incorporating various orders and stencil sizes.

In this paper we present an edge detection method for multivariate irregular data that has the following desirable properties: (I) It can be applied to any irregular data in any domain. (II) It is independent of any specific shape of discontinuities in both the univariate and bivariate cases. (III) The method depends only on locally sampled signals, making it easy to implement numerically, since for each point our scheme needs only to solve a simple matrix and no global system of equations needs to be solved. (IV) It has a fast rate of convergence to zero away from the discontinuities. The benefit of the last property is that the edge detection method will be able to distinguish jumps from steep gradients more readily than methods of slower convergence. There will be additional considerations, as it will become apparent that high order edge detection methods produce more oscillations in the neighborhoods of the jump discontinuities. To distinguish true jump locations from neighborhood oscillations, we find from [13] that the *minmod* function (see Definition 3.1), typically used for reducing oscillations in the presence of shocks in numerical solutions for conservation laws (see, e.g., [7]), may help to reduce oscillations in the presence of jump discontinuities. We extend this idea to our edge detection method for the case of multivariate irregular data and moreover provide a proof for its convergence rate to zero away from the discontinuities, which has previously not been accomplished.

This study is primarily concerned with the detection of jump discontinuities (or fault detection). While it is important to consider the effects of a noisy environment on an edge detection method, it is beyond the scope of this introductory paper. Hence we leave the study of noise for future investigations.

This paper is organized as follows: In section 2 we present the formulation of the edge detection method. In section 3 we use this formulation to construct an edge detection method for the one-dimensional case and employ the *minmod* function to the edge detection method. Section 4 is devoted to analyzing the behavior of the edge detection method in two dimensions. Finally, some numerical algorithms are provided in Appendices A and B.

2. General formulation for edge detection.

Let us first introduce the following notation, which will be used throughout this paper:

For $x = (x_1, \dots, x_d)$ in $\mathbb{R}^d$, $|x| := (x_1^2 + x_2^2 + \cdots + x_d^2)^{1/2}$ stands for its Euclidean norm. For any finite set of points $\mathcal{S}$ in $\mathbb{R}^d$ we use the notation $K_{\mathcal{S}}$ for the convex hull of the set $\mathcal{S}$. We denote by $\mathbb{N} := \{1, 2, \dots, \}$ the set of natural numbers and by $\mathbb{Z}_+ := \{0, 1, 2, \dots, \}$ the set of nonnegative integers. For any $\alpha \in \{(\alpha_1, \dots, \alpha_d) : \alpha_1, \dots, \alpha_d \in \mathbb{Z}_+\} := \mathbb{Z}_+^d$, we set $|\alpha|_1 := \sum_{k=1}^d \alpha_k$, and $\alpha! := \alpha_1! \cdots \alpha_d!$. Throughout α is a multivariate nonnegative integer that will change dimension based upon the dimension under discussion. We denote a uniform grid of density h as $h\mathbb{Z}^d := \{h\alpha | \alpha \in \mathbb{Z}^d\}$. We use the usual notation $\lceil s \rceil$ to indicate the smallest integer greater than or equal to s. For any $m \in \mathbb{Z}_+$, Π_m denotes the space of all polynomials of degree $\le m$ in $d \in \mathbb{N}$ variables where the dimension of Π_m is denoted by

$$m_d := \binom{m+d}{d}. \tag{2.1}$$

We recall the Dirac delta function

$$\delta_{i,j} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases} \tag{2.2}$$

Finally, throughout this paper we use A to represent an arbitrary constant that may change value.

We introduce an edge detection method on the set of irregularly distributed points in a bounded domain Ω in $\mathbb{R}^d$. Let $\mathcal{S}$ be a set of discrete points in Ω and let f be a piecewise smooth function known only on $\mathcal{S}$. In order to identify the jump discontinuities of f, we construct a function $L_m f$, $m \in \mathbb{N}$, which can be characterized by the asymptotical convergence property,

$$L_m f(x) \longrightarrow 0,$$

for any x away from discontinuities, with the convergence rate depending in part on the given positive integer m. The choice of m is user dependent, but a higher number m provides a faster rate of convergence in smooth regions of f.

The edge detection method presented here is based on a local polynomial annihilation property. The general form of $L_m f$ is given by the following two step method. In the first step, for any $x \in \Omega$, we choose a set

$$\mathcal{S}_x := \mathcal{S}_{m_d,x} := \{x_1, \ldots, x_{m_d}\}, \tag{2.3}$$

which is a local set of m_d (2.1) points around x. In practice, though the dimension d can be arbitrary, we consider only the case $d \leq 2$ and note that for $d > 2$ the method is the same although the numerical algorithms are more complicated.

In order to annihilate polynomials up to degree $m-1$, we solve a linear system for the coefficients $c_j(x)$, $j = 1, \ldots, m_d$, given by

$$\sum_{x_j \in \mathcal{S}_x} c_j(x) p_i(x_j) = \sum_{|\alpha|_1 = m} p_i^{(\alpha)}(x), \quad \alpha \in \mathbb{Z}_+^d, \tag{2.4}$$

where p_i, $i = 1, \ldots, m_d$, is a basis of Π_m. Note that the solution (2.4) exists and is unique. Our edge detector $L_m f$ is defined, using the solution of (2.4), as

$$L_m f(x) = \frac{1}{q_{m,d}(x)} \sum_{x_j \in \mathcal{S}_x} c_j(x) f(x_j). \tag{2.5}$$

Here $q_{m,d}(x)$ is a suitable normalization factor depending on m, the dimension d, and the local set $\mathcal{S}_x$ (2.3). In the following sections, we will determine $q_{m,d}(x)$ specifically for $d = 1, 2$. Indeed, for the univariate case the normalization factor $q_{m,d}(x)$ is important in detecting the jump amount at a discontinuity. However, in the multivariate case, it is no longer meaningful since the jump amount varies depending on the directions at a discontinuity. In this case $q_{m,d}(x)$ can be used to estimate the magnitude of the jump in its normal direction, as will be explained later.

It is evident from (2.5) that $L_m f$ is *local* in the sense that it employs data only in a small neighborhood of x. It is also apparent that (2.5) detects edges regardless of the geometrical aspects of the discontinuities. Furthermore, we will show that $L_m f(x)$ converges to zero away from the discontinuities with a certain rate depending on m and the local smoothness of the function f.

3. Edge detection in one dimension.

3.1. Formulation.

3.1. Formulation. Throughout section 3, let f be a piecewise smooth function on an interval $[a, b]$, known only at the finite discrete points

$$\mathcal{S} \subset [a, b], \quad \#\mathcal{S} =: N < \infty,$$

which we will call "nodes." Suppose that f has jump discontinuities with well-defined one-sided limits, and let

$$J = \{\xi : a \le \xi \le b\}$$

denote the set of jump discontinuities of f in $[a, b]$. We define the local jump function corresponding to f as

$$[f](x) := f(x+) - f(x-),$$

where $f(x+)$ and $f(x-)$ are the right- and left-hand limits of the function f at x. Clearly, if f is continuous at x, then $[f](x) = 0$ and for any $\xi \in J$, $[f](\xi) = f(\xi+) - f(\xi-) \neq 0$.

The ability to find the locations and corresponding amplitudes of the jump discontinuities depends on the accuracy of the approximation to the jump function $[f](x)$. Hence we construct $L_m f(x)$ in (2.5) to be an approximation to $[f](x)$ such that

$$L_m f(x) \longrightarrow \begin{cases} [f](\xi) & \text{if } x_{j-1} \le \xi, x \le x_j \text{ for } \xi \in J, \\ 0 & \text{if } I_x \cap J = \emptyset, \end{cases} \tag{3.1}$$

where I_x is the smallest closed interval such that $\mathcal{S}_x \subset I_x$, with $\mathcal{S}_x$ defined in (2.3). In this way, a jump discontinuity $\xi \in J$ is identified by its enclosed cell, $x_{j-1} \le \xi \le x_j$, and the convergence rate of the approximation $L_m f(x)$ to the jump function $[f](x)$ is given in terms of

$$h(x) := \max\{|x_i - x_{i-1}| \ : \ x_{i-1}, x_i \in \mathcal{S}_x\}. \tag{3.2}$$

Clearly $h(x)$ is dependent upon the density of $\mathcal{S}_x$.

The function $L_m f$ for the univariate case is defined as follows: For the given positive integer m, we choose a local set $\mathcal{S}_x$ of m_1 points around x. Here m_1 is the dimension of Π_m in $\mathbb{R}$ as given by (2.1), i.e., $m_1 = m + 1$. The coefficients utilized in the edge detection method are determined by the solution of the linear system

$$\sum_{x_j \in \mathcal{S}_x} c_j(x) p_i(x_j) = p_i^{(m)}(x), \quad i = 1, \ldots, m_1, \tag{3.3}$$

where p_i, $i = 1, \ldots, m_1$, is a basis of Π_m. Clearly, the coefficients $c_j(x)$ are uniquely determined by the local set $\mathcal{S}_x$, and are of order $\mathcal{O}(h(x)^{-m})$ as $h(x) \to 0$. Fortunately, an explicit formula exists for $c_j(x)$ that will be described later in Theorem 3.2.

Next, by defining

$$\mathcal{S}_x^+ := \{x_j \in \mathcal{S}_x | x_j \ge x\} \quad \text{and} \quad \mathcal{S}_x^- := \mathcal{S}_x \setminus \mathcal{S}_x^+, \tag{3.4}$$

we set the normalization factor in (2.5) as

$$q_m(x) := q_{m,1}(x) := \sum_{x_j \in \mathcal{S}_x^+} c_j(x), \tag{3.5}$$

such that $q_m(x) \neq 0$. Note from (3.3) that it is clear that $q_m(x)$ is of order $\mathcal{O}(h(x)^{-m})$ as well.

Finally, the edge detection method (2.5) in the one-dimensional case is

$$L_m f(x) = \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x} c_j(x) f(x_j). \tag{3.6}$$

There is no restriction in choosing the sets $\mathcal{S}_x^+$ and $\mathcal{S}_x^-$ in (3.4), but from a practical point of view, a good choice is to put almost the same numbers of nodes on each side of x. For instance, if m is odd, one may choose $\mathcal{S}_x^+$ and $\mathcal{S}_x^-$ such that $\#\mathcal{S}_x^+ = \#\mathcal{S}_x^-$. These sets will have to be adjusted near the boundary of the domain, and naturally will become more one-sided.

In order for $L_m f$ in (3.6) to be successful, it should approximate the jump function $[f](x)$ with high accuracy. Theorem 3.1 shows that $L_m f(x)$ converges to zero away from the jump discontinuities of f with a certain rate depending on m and the local smoothness of f.

THEOREM 3.1. *Let $m \in \mathbb{N}$ and $L_m f(x)$ be defined as in* (3.6) *using a local set $\mathcal{S}_x$ with $\#\mathcal{S}_x = m_1 = m + 1$. Then we have*

$$L_m f(x) = \begin{cases} [f](\xi) + \mathcal{O}(h(x)) & \text{if } x_{j-1} \le \xi, x \le x_j, \\ \mathcal{O}(h^{\min(m,k)}(x)) & \text{if } f \in C^k(I_x) \text{ for } k > 0, \end{cases}$$

where $h(x)$ is given in (3.2) *and I_x is the smallest closed interval such that $\mathcal{S}_x \subset I_x$.*

Proof. Assume first that $f \in C^k(I_x)$ for some $k > 0$. Denote $k_m := \min(k, m) > 0$ and let $T_{k_m-1}f$ be the Taylor expansion of f of degree $k_m - 1$ around x, namely,

$$T_{k_m-1}f(\cdot) = \sum_{\alpha=0}^{k_m-1} (\cdot - x)^\alpha f^{(\alpha)}(x)/\alpha!.$$

Since $T_{k_m-1}f$ is a polynomial of degree less than m, the definition of $c_j(x)$ in (3.3) implies that

$$\sum_{x_j \in \mathcal{S}_x} c_j(x) T_{k_m-1} f(x_j) = 0. \tag{3.7}$$

By rewriting $f = T_{k_m-1}f + R_{k_m-1}f$, where $R_{k_m-1}f$ is the remainder of Taylor expansion, it follows from (3.7) that

$$\begin{aligned} |L_m f(x)| &= \left| \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x} c_j(x) R_{k_m-1} f(x_j) \right| \\ &= \left| \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x} c_j(x) (x_j - x)^{k_m} f^{(k_m)}(\zeta_j)/k_m! \right| \\ &\le A h^{k_m}(x) \frac{1}{|q_m(x)|} \sum_{x_j \in \mathcal{S}_x} |c_j(x)| \end{aligned}$$

for some ζ_j between x and x_j, where the last inequality is implied since $|x - x_j| \le mh(x)$ for any $x_j \in \mathcal{S}_x$ and $|f^{(k_m)}(x)/k_m!|$ is bounded for $x \in I_x$. Since both $c_j(x)$ and $q_m(x)$ are $\mathcal{O}(h^{-m}(x))$, it is clear that $|L_m f(x)| \le A h^{k_m}(x)$.

Next, consider the case that $x_{j-1} \le \xi, x \le x_j$ for $\xi \in J$ and $x_{j-1}, x_j \in \mathcal{S}_x$. Without loss of generality, assume that ξ is the only discontinuity of f in a neighborhood I_ξ and $\mathcal{S}_x \subset I_\xi$. Invoking the notation $\mathcal{S}_x^+$ and $\mathcal{S}_x^-$ in (3.4), we have

$$\begin{aligned} L_m f(x) &= \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x^+} c_j(x) f(x_j) + \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x^-} c_j(x) f(x_j) \\ &= \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x^+} c_j(x) \left[f(\xi^+) + (x_j - \xi) f'(\zeta_j^+) \right] \\ &\quad + \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x^-} c_j(x) \left[f(\xi^-) + (x_j - \xi) f'(\zeta_j^-) \right] \end{aligned}$$

for some ζ_j^+ and ζ_j^-. Since (2.4) implies that $\sum_{x_j \in \mathcal{S}_x} c_j(x) = 0$, it follows that

$$\sum_{x_j \in \mathcal{S}_x^+} c_j(x) = - \sum_{x_j \in \mathcal{S}_x^-} c_j(x).$$

Utilization of (3.5) yields

$$L_m f(x) = (f(\xi^+) - f(\xi^-)) + \mathcal{O}(h(x))$$

to complete the proof. □

Next, Theorem 3.2 establishes the relationship between the edge detection method (3.6) and Newton divide differences, which are frequently employed to determine smooth regions in finite difference schemes (see, for example, [6], [9], and [12]). This beneficial relationship provides an explicit formula for the coefficients $c_j(x)$ without solving the linear system (3.3). Denoting $\mathcal{S}_x =: \{x_1, \ldots, x_{m_1}\}$ with $m_1 = m+1$, recall the definition of the m_1th degree Newton divided difference for a smooth function $f(x)$ on $\mathcal{S}_x$:

$$\begin{aligned} f[\mathcal{S}_x] := f[x_1, x_2, \ldots, x_{m_1}] &= \frac{f[x_1, x_2, \ldots, x_{m_1 - 1}] - f[x_2, x_3, \ldots, x_{m_1}]}{x_1 - x_{m_1}} \\ &= \sum_{j=1}^{m_1} \frac{f(x_j)}{\omega_j(\mathcal{S}_x)} = \frac{f^{(m)}(\xi)}{m!}, \end{aligned} \tag{3.8}$$

where $\xi \in (x_1, x_{m_1})$ and

$$\omega_j(\mathcal{S}_x) := \prod_{\substack{i=1 \\ i \ne j}}^{m_1} (x_j - x_i). \tag{3.9}$$

THEOREM 3.2. *Under the same conditions and notation as Theorem* 3.1, *the coefficients $c_j(x)$ can be directly solved as*

$$c_j(x) = \frac{m!}{\omega_j(\mathcal{S}_x)}, \quad j = 1, \ldots, m_1, \tag{3.10}$$

with $\omega_j(\mathcal{S}_x)$ in (3.9). *Furthermore, the $L_m f(x)$ in* (3.6) *can be expressed as*

$$L_m f(x) = \frac{m!}{q_m(x)} f[\mathcal{S}_x].$$

Proof. Since the coefficients $c_j(x)$ that solve (3.3) are independent of the basis of Π_m, it is enough to consider the basis $p_i(x) = x^{i-1}$ for $i = 1, \dots, m_1$. The mth derivative of these basis functions satisfies

$$p_i^{(m)}(x) = m!\delta_{i,m_1} \quad \text{for all } x \in \mathbb{R}. \tag{3.11}$$

By (3.8), it is possible to conclude that $p_i[\mathcal{S}_x] = \delta_{i,m_1}$, yielding $m!p_i[\mathcal{S}_x] = p_i^{(m)}(x)$, $i = 1, \dots, m_1$. Hence by (3.8) we have

$$m! \sum_{j=1}^{m_1} \frac{p_i(x_j)}{\omega_j(\mathcal{S}_x)} = p_i^{(m)}(x), \quad i = 1, \dots, m_1,$$

indicating that the coefficients $c_j(x)$ can also be formulated by (3.10).

Given the direct representation (3.10) of the coefficients $c_j(x)$ as the solution of the linear system (3.3), the edge detection method (3.6) can be expressed as

$$\begin{aligned} L_m f(x) &= \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x} c_j(x) f(x_j), \\ &= \frac{m!}{q_m(x)} \sum_{x_j \in \mathcal{S}_x} \frac{f(x_j)}{\omega_j(\mathcal{S}_x)} = \frac{m!}{q_m(x)} f[\mathcal{S}_x], \end{aligned}$$

finishing the proof. □

Remark 3.1. Let $x \in (a, b)$ be fixed and let I_x^+ be the smallest closed interval such that $\mathcal{S}_x^+ \subset I_x^+$. Choosing $f = \chi_{I_x^+}$ with $\chi_{I_x^+}$ the characteristic function on I_x^+, Theorem 3.2 implies that the normalization factor $q_m(x)$ in (3.5) can be written as

$$\begin{aligned} q_m(x) &= \sum_{x_j \in \mathcal{S}_x^+} c_j(x) \\ &= \sum_{x_j \in \mathcal{S}_x} c_j(x) \chi_{I_x^+}(x_j) = m! \chi_{I_x^+}[\mathcal{S}_x]. \end{aligned}$$

Thus, the assumption $q_m(x) \neq 0$ is reasonable.

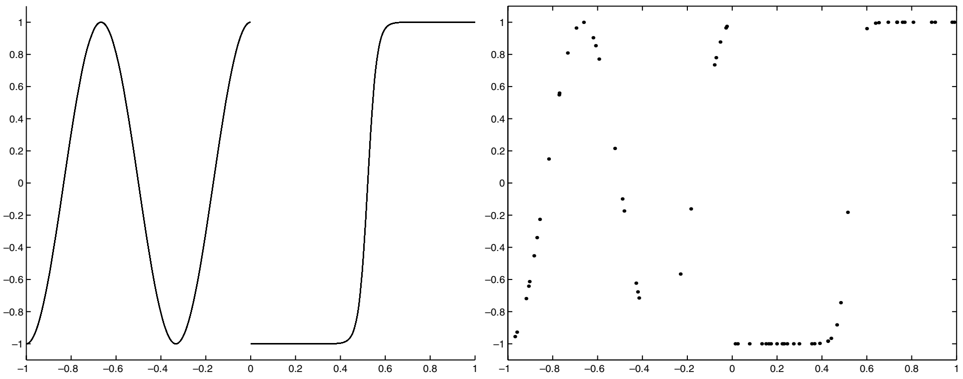
Remark 3.2. As discussed above, for any given data $(\mathcal{S}, f|_{\mathcal{S}})$, the evaluation of $L_m f(x)$ involves only finite local data $(\mathcal{S}_x, f|_{\mathcal{S}_x})$ around x. Accordingly, if $\mathcal{S}$ is a set of irregular points, the coefficients $c_j(x)$ may vary depending on the location x. However, if the given set $\mathcal{S}$ is uniform, say,

$$\mathcal{S} := \{a + nh \mid n = 0, \dots, N\}, \quad h = \frac{b-a}{N} > 0, \tag{3.12}$$

then there exists only one set of coefficients $c_j(x) = c_j$, $j = 1, \dots, m_1$, which are independent of the position x inside $[a, b]$ (but away from the boundary). Specifically, (3.10) can be directly applied to obtain the coefficients

$$c_j = \frac{m!}{\omega_j(\mathcal{S}_x)} = \frac{m!}{h \prod_{i=1, i \neq j}^{m_1} (j - i)}, \quad j = 1, \dots, m_1,$$

which in turn yields the normalization factor $q_m = q_m(x)$ in (3.5). Hence $\frac{c_j}{q_m}$ in (3.6) is bounded and independent of both h and x, and consequently the numerical computation of (3.6) is further simplified, while keeping the same convergence properties in Theorem 3.1.



(a) **(b)**

FIG. 3.1. (a) *Graph of* $f_1(x)$. (b) *Random sampling of* $f_1(x)$ *on* $N = 64$ *points.*

To demonstrate the efficacy of the edge detection method $L_m f(x)$, let us consider the following example.

Example 3.1.

$$f(x) := \begin{cases} \cos(3\pi x), & -1 \le x < 0, \\ \frac{2}{1+3e^{-50x+25}} - 1, & 0 < x \le 1. \end{cases} \tag{3.13}$$

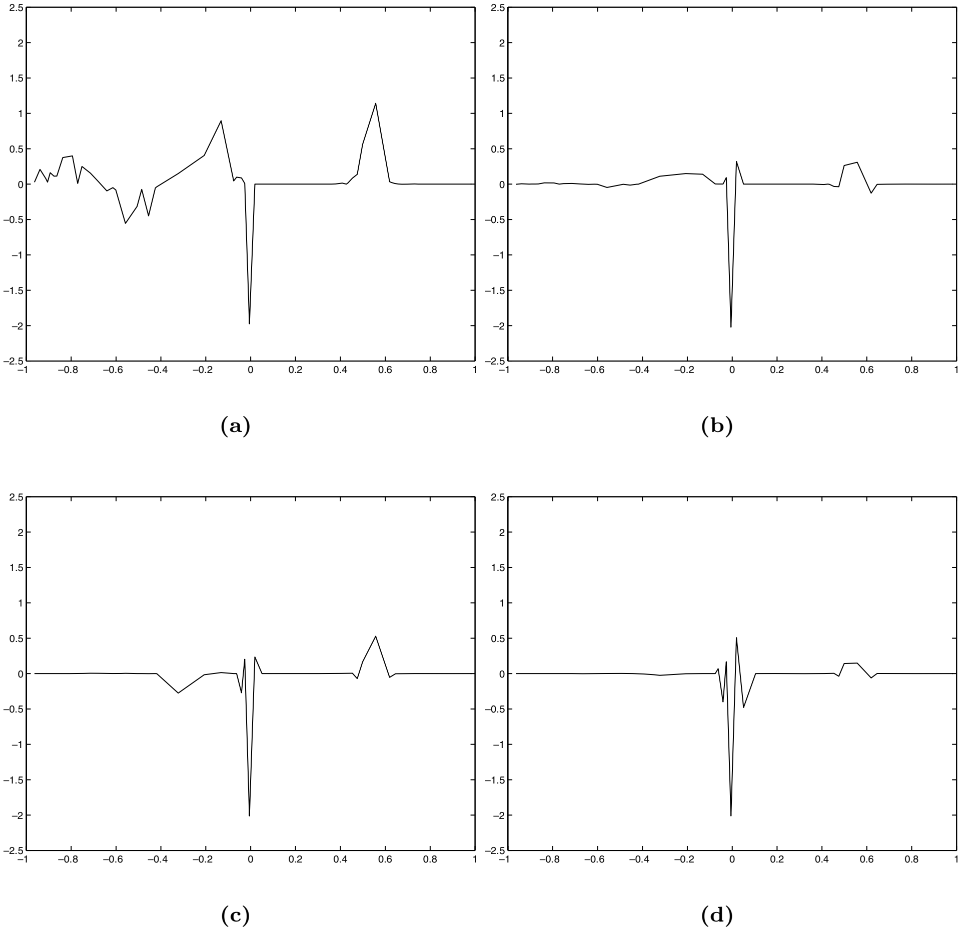
The function $f(x)$ has an edge at $x = 0$ and the corresponding jump function

$$[f](x) = \begin{cases} -2 & \text{if } x = 0, \\ 0 & \text{else.} \end{cases} \tag{3.14}$$

We wish to approximate $[f](x)$, based on the scattered grid point values generated randomly by MATLAB and depicted in Figure 3.1(b). Figure 3.2 demonstrates the application of $L_m f(x)$ for $m = 1, 3, 4$, and 6.

Observe in Figure 3.2 that the application of (3.6) encounters some problems in the approximation of jump functions. Specifically, as m increases, oscillations that occur in the neighborhood of a jump discontinuity can be misidentified as true edges. On the other hand, for smaller m, there is a risk of identifying a steep gradient as an edge, especially in regions where the scattered grid points are far apart. We wish to avoid the possibility of misidentification due either to the low resolution problems associated with the low order edge detection or to the oscillations inherent in the high order case. Presented in the following section is the *minmod* edge detection method that helps to prevent the edge detection method (3.6) from misidentifying edges.

3.2. ***Minmod*** **edge detection in one dimension.** It was observed in [13] that the *minmod* function, typically used in numerical conservation laws to reduce oscillations (see, e.g., [7]), could also be applied to distinguish true jump discontinuities from neighborhood oscillations. In what follows we describe the oscillating behavior near the jump discontinuities that results from using our local edge detector (3.6). Thus motivated, we extend the use of the *minmod* function and incorporate various orders of m for nonuniform grids. Moreover, we provide the proof of its convergence rate to zero away from the discontinuities. We will refer to this technique as the *minmod* edge detection method.



(a) **(b)**

(c) **(d)**

FIG. 3.2. *The edge detection method* $L_m f(x)$ *given by* (3.6) *for* (a) $m = 1$, (b) $m = 3$, (c) $m = 4$, *and* (d) $m = 6$.

The next theorem describes the behavior of the edge detection method (3.6) in the neighborhoods of discontinuities and motivates the need for further refinement by the *minmod* function.

THEOREM 3.3. *Let* $m \in \mathbb{N}$ *and* $L_m f(x)$ *be defined as in* (3.6) *using* $\mathcal{S}_x$ *with* $\#\mathcal{S}_x = m_1$, *and let*

$$Q_m(\xi, x) := \sum_{x_j \in \mathcal{S}_\xi^+} c_j(x) \tag{3.15}$$

with $\mathcal{S}_\xi^+ := \{x_j \in \mathcal{S}_x | x_j \geq \xi\}$ *as given in* (3.4). *Then*

$$L_m f(x) = \begin{cases} \frac{Q_m(\xi,x)}{q_m(x)}[f](\xi) + \mathcal{O}(h(x)) & \text{if } I_x \cap \xi \neq \emptyset \text{ for } \xi \in J, \\ \mathcal{O}(h^{\min(m,k)}(x)) & \text{if } f \in C^k(I_x) \text{ for } k > 0. \end{cases}$$

Here, I_x *is the smallest closed interval such that the local set* $\mathcal{S}_x \subset I_x$.

Proof. Assume first that $I_x \cap J = \emptyset$ and therefore $f \in C^k(I_x)$ for some $k > 0$. It is therefore possible to conclude by Theorem 3.1 that $L_m f(x) = \mathcal{O}(h^{\min(m,k)}(x))$.

Next, let us consider the case that $I_x \cap J \neq \emptyset$. Without loss of generality, assume that ξ is the only discontinuity of f in a neighborhood I_x. Invoking the notation $\mathcal{S}_\xi^+$ and $\mathcal{S}_\xi^-$ in (3.4), we see that

$$\begin{aligned} L_m f(x) &= \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_\xi^+} c_j(x) f(x_j) + \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_\xi^-} c_j(x) f(x_j) \\ &= \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_\xi^+} c_j(x) \left[f(\xi^+) + (x_j - \xi) f'(\zeta_j) \right] \\ &\quad + \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_\xi^-} c_j(x) \left[f(\xi^-) + (x_j - \xi) f'(\zeta_j) \right] \end{aligned}$$

with ζ_j between x_j and ξ. Here, from the condition $\sum_{x_j \in \mathcal{S}_x} c_j(x) = 0$ in (3.3), it is clear that

$$Q_m(\xi, x) := \sum_{j \in \mathcal{S}_\xi^+} c_j(x) = - \sum_{j \in \mathcal{S}_\xi^-} c_j(x).$$

Hence

$$L_m f(x) = \frac{Q_m(\xi, x)}{q_m(x)} (f(\xi^+) - f(\xi^-)) + \mathcal{O}(h(x)),$$

which completes the proof. □

The behavior characterized in Theorem 3.3 is visible in Figure 3.2 as m increases in (3.6). Specifically, the edge detection method approximates the jump function with high order outside the neighborhoods of the discontinuities. Unfortunately, inside the neighborhoods of the discontinuities the edge detection method oscillates according to the fraction $\frac{Q_m(\xi,x)}{q_m(x)}[f](\xi)$.

The *minmod* edge detection method, as defined below, uses the *minmod* function to exploit the characteristics of the edge detection method of various orders both inside and outside the neighborhoods of the discontinuities to ensure the highest order of convergence possible away from the discontinuity, as well as to reduce the oscillations inside the neighborhoods of discontinuities.

DEFINITION 3.1. *For a given finite set $\mathcal{M} \subset \mathbb{N}$ of positive integers, consider the set $L_\mathcal{M} f = \{L_m f : \mathbb{R} \to \mathbb{R} \,|\, m \in \mathcal{M}\}$. The minmod function is defined by*

$$(3.16)\quad MM\Big(L_\mathcal{M} f(x)\Big) = \begin{cases} \min\limits_{m \in \mathcal{M}} L_m f(x) & \text{if } L_m f(x) > 0 \text{ for all } m \in \mathcal{M}, \\ \max\limits_{m \in \mathcal{M}} L_m f(x) & \text{if } L_m f(x) < 0 \text{ for all } m \in \mathcal{M}, \\ 0 & \text{otherwise}. \end{cases}$$

Theorem 3.4 characterizes the convergence of the *minmod* function applied to the set of edge detectors $L_m f$ of various order m and demonstrates its ability to distinguish jump discontinuities from neighborhood oscillations.

THEOREM 3.4. *If $\mathcal{M} = \{1, 2, \ldots, \mu\}$, we have*

$$MM\Big(L_\mathcal{M} f(x)\Big) = \begin{cases} [f](\xi) + \mathcal{O}(h(x)) & \text{if } x_{j-1} \leq \xi, x \leq x_j, \\ \mathcal{O}(h^{\min(\mathcal{M}_x, k)}(x)) & \text{if } f \in C^k(I_x), \end{cases}$$

where I_x is the smallest closed interval such that $\mathcal{S}_x \subset I_x$ with $\#\mathcal{S}_x \leq \binom{\mathcal{M}_x+1}{1}$, and $\mathcal{M}_x$ is defined by

$$\mathcal{M}_x := \max\big\{m \in \mathcal{M} \,|\, \#\mathcal{S}_x = m_1,\ I_x \cap J = \emptyset\big\}. \tag{3.17}$$

Proof. For $x \in [a,b]$, assume without loss of generality that $x_{j-1} \leq x \leq x_j$ for some $x_{j-1}, x_j \in \mathcal{S}$. If there exists $\xi \in J$ such that $x_{j-1} \leq \xi \leq x_j$, then by Theorem 3.1 we have

$$L_m f(x) = [f](\xi) + \mathcal{O}(h(x))$$

for any $m \in \mathcal{M}$. Therefore, it is possible to conclude that

$$MM\big(L_{\mathcal{M}} f(x)\big) = [f](\xi) + \mathcal{O}(h(x)).$$

If $J \cap [x_{j-1}, x_j] = \emptyset$, then by definition we have $\mathcal{M}_x \geq 1$. Also from the definition of $\mathcal{M}_x$, for any $m \in \mathcal{M}$ such that $m \leq \mathcal{M}_x$ and $\#\mathcal{S}_x = m_1$ we have $I_x \cap J = \emptyset$. Therefore, Theorem 3.1 implies that $L_m f(x) = \mathcal{O}(h^{\min(m,k)}(x))$, yielding

$$MM\big(L_{\mathcal{M}} f(x)\big) = \mathcal{O}(h^{\min(\mathcal{M}_x,k)}(x))$$

to complete the proof. ∎

By including $1 \in \mathcal{M}$ in Theorem 3.4, first order convergence is ensured at edges, even in the case where edges are in neighboring centers. Large values are also included in the set $\mathcal{M}$ so that there will be a high order of convergence away from the discontinuity.

The *minmod* edge detection method relaxes the assumption for edge resolution in Theorem 3.1, specifically that edge detection is possible only if a maximum of one edge is contained in each local set, or equivalently

$$\#\big([x_j, x_{j+m_1}] \cap J\big) \leq 1 \quad \text{for } j = 1, \ldots, N - m_1, \tag{3.18}$$

where J is the set of discontinuities of f on $[a,b]$. In this case, only a certain density of edges can be resolved, i.e., only one discontinuity can be resolved for each m_1 points. Furthermore, the order of the method is restricted to the "closeness" of the edges in terms of their grid point location. Theorem 3.4 relaxes this assumption so that edge resolution is possible if J, the set of discontinuities of f on $[a,b]$, satisfies

$$\#\big([x_j, x_{j+1}] \cap J\big) \leq 1 \quad \text{for } j = 1, \ldots, N - 1, \tag{3.19}$$

i.e., the edges can occur at neighboring grid point values. If this requirement is not satisfied, the problem is clearly underresolved.

The superior convergence properties of the *minmod* edge detection method for Example 3.1 with $\mathcal{M} = \{1, 2, \ldots, 6\}$ are evident in Figure 3.3. Of particular interest is the ability of the *minmod* edge detection method to resolve the local jump function even when the first order approximation, as displayed in Figure 3.2(a), detects edges in smooth regions that are artifacts of the variability of the function and sparse sampling. Residual small oscillations that are still evident can be removed by a thresholding process.

The algorithm in Appendix A details the one-dimensional edge detection computation of Examples 3.1, where the particular choice of local sets, reconstruction grid points, and basis functions are specified.

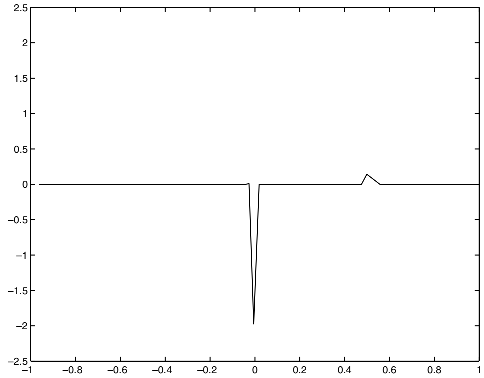


FIG. 3.3. *The minmod edge detection method, $MM(L_{\mathcal{M}}f(x))$, for Example* 3.1. *Here* $\mathcal{M} = \{1, 2, \ldots, 6\}$.

4. Edge detection in two dimensions.

4.1. Formulation. Throughout section 4, let f be a piecewise smooth function on a domain $\Omega \subset \mathbb{R}^2$ known only on the set of discrete nodes

$$\mathcal{S} \subset \Omega, \quad \#\mathcal{S} =: N < \infty.$$

Though d indicates an arbitrary dimension, here we consider only the bivariate case. The higher-dimensional case can be similarly constructed with more complicated numerical algorithms.

In two dimensions a jump discontinuity at $x = \xi$ is identified by its enclosed points (i.e., triangular points) and is characterized by the convergence property away from the discontinuities. Specifically, the enclosed points can be defined by the Delaunay triangulation for $\mathcal{S}$ that consists of the set of lines connecting each point to its natural neighbors [10]. These sets of lines form elementary triangles whose vertices consist of points in $\mathcal{S}$. A triangle is considered elementary if every combination of vertex pairs are natural neighbors. Let the number of elementary triangles of the set $\mathcal{S}$ be defined as N_T. We denote the set of vertices of all elementary triangles in the Delaunay triangulation of $\mathcal{S}$ as

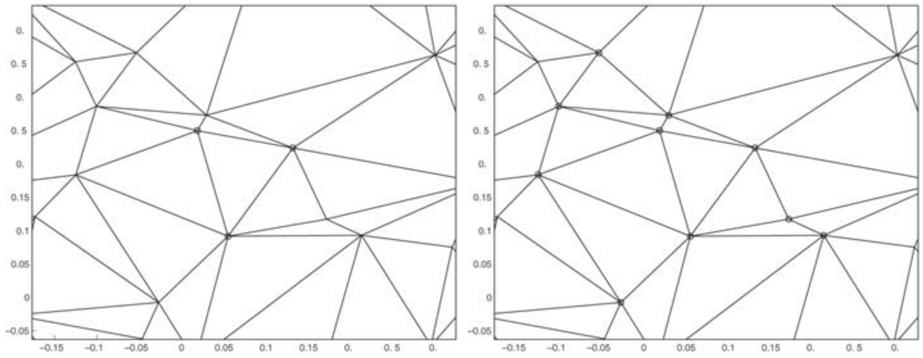
$$\mathcal{T}_{\mathcal{S}} = \left\{ T_j \middle| T_j := \{x_1^j, x_2^j, x_3^j\} \subset \mathcal{S} \text{ for } j = 1, \ldots, N_T \right\}, \tag{4.1}$$

where T_j is the set of vertices for an elementary triangle.

Since discontinuities are identified at specific points by their enclosed cells, the local sets $\mathcal{S}_x$ are chosen to include points that characterize these cells. For arbitrary $x \in \Omega$, we can assume without loss of generality that $x \in K_{T_j} \subset \Omega$. Recall that K_{T_j} is the convex hull of the set of vertices $T_j \in \mathcal{T}_{\mathcal{S}}$. Therefore the local set $\mathcal{S}_x$ for arbitrary $x \in \Omega$ can now be defined specifically to include the set that characterizes its enclosed points as

$$\mathcal{S}_x := T_j \cup \mathcal{S}_{T_j}, \tag{4.2}$$

where $T_j \in \mathcal{T}_{\mathcal{S}}$, $x \in K_{T_j}$, and $\mathcal{S}_{T_j}$ is the set of the $m_2 - 3$ closest points to x in the set $\mathcal{S} \setminus T_j$. To illustrate how $\mathcal{S}_x$ is chosen, Figure 4.1 depicts a region of the Delaunay



(a) **(b)**

FIG. 4.1. *A region of the Delaunay triangulation of* 256 *randomly sampled points on* $[-1,1]\times[-1,1]$. *(The region is enlarged for visibility purposes.)* (a) *The elementary triangle* T_j *that satisfies* $x \in K_{T_j}$*, represented by circles.* (b) *Pictorial representation of the local set* $\mathcal{S}_x := T_j \cup \mathcal{S}_{T_j}$*, where* $\#\mathcal{S}_x = 10$*.*

triangulation of 256 randomly sampled points on $[-1,1]\times[-1,1]$. Figure 4.1(a) displays a point x and the elementary triangle T_j that satisfies $x \in K_{T_j}$. Figure 4.1(b) exhibits a local set $\mathcal{S}_x$ as defined in (4.2), where $\#\mathcal{S}_x = 10$.

In order to quantify the convergence rate of the edge detection method, we define

$$h(x) := \max_{x \in K_{\mathcal{S}_x}} \min_{x_j \in \mathcal{S}_x} |x - x_j|, \tag{4.3}$$

which is dependent upon the density of the local set $\mathcal{S}_x$.

Recall that for a given positive integer m, the dimension of Π_m in $\mathbb{R}^2$ is denoted by m_2 (2.1). If $\mathcal{S}_x$ is a local set of m_2 points around x, the function $L_m f$ is given by

$$L_m f(x) = \frac{1}{q_{m,2}(x)} \sum_{x_j \in \mathcal{S}_x} c_j(x) f(x_j), \tag{4.4}$$

where the coefficients $c_j(x)$, $j = 1, \ldots, m_2$, are dependent upon the local set $\mathcal{S}_x$ and satisfy the linear system

$$\sum_{x_j \in S_x} c_j(x) p_i(x_j) = \sum_{|\alpha|_1 = m} p_i^{(\alpha)}(x), \quad i = 1, \ldots, m_2, \quad \alpha \in \mathbb{Z}_+^2. \tag{4.5}$$

Here p_i, $i = 1, \ldots, m_2$, form a basis of Π_m. Further illustration of the application of (4.5) for a particular basis of Π_m is detailed in Appendix B. It is easy to check that $c_j(x) = \mathcal{O}(h(x)^{-m})$, implying that $q_{m,2}(x) = \mathcal{O}(h(x)^{-m})$ as well. Recall that in the one-dimensional case, the constant $q_{m,1}$ is used to determine the jump amplitude at a discontinuity. In the bivariate case, the jump amplitude may vary depending on the paths through a given discontinuity, so quantifying the jump amount at such discontinuity points is not meaningful. However, in the case where jump discontinuities arise locally along a simple curve, we can estimate the jump magnitude in the normal direction with a suitable $q_{m,2}(x)$ and then apply the *minmod* edge detection method from (3.16) to pinpoint the edges. This will be discussed further in section 4.2. For now we limit our discussion to detecting edges without consideration of their jump amounts.

Since $c_j(x) = \mathcal{O}(h(x)^{-m})$, it is possible to bound $L_m f$ uniformly by defining

$$q_m(x) = q_{m,2}(x) := \sum_{x_j \in \mathcal{P}_x} c_j(x), \tag{4.6}$$

where $\mathcal{P}_x$ can be a suitable subset of $\mathcal{S}_x$ such that $q_m(x) \neq 0$. This will be discussed later following Definition 4.1, where we will see that the versatility of $\mathcal{P}_x$ can be utilized to provide a good approximation to the jump magnitudes in the normal directions of the edges in the multivariate case.

Theorem 4.1 establishes the convergence rate of $L_m f(x)$, defined in (4.4), away from the discontinuities of f.

THEOREM 4.1. *Suppose f is a piecewise smooth function on a domain Ω in $\mathbb{R}^2$ known only on discrete nodes $\mathcal{S}$. Let J denote the set of jump discontinuities of f in Ω, and let $L_m f$ be defined as in* (4.4) *with $m \in \mathbb{N}$. Then if $f \in C^k(K_{\mathcal{S}_x})$ for some $k > 0$, we have*

$$L_m f(x) = \mathcal{O}(h^{\min(k,m)}(x)).$$

Proof. The technique of proving Theorem 3.1 is adapted in a straightforward fashion to prove this theorem. Assuming that $f \in C^k(K_{\mathcal{S}_x})$ for some $k > 0$, we define $k_m := \min(k, m)$ and then separate f into two parts:

$$f = T_{k_m-1} f + R_{k_m-1} f,$$

where $T_{k_m-1} f$ is the Taylor polynomial of f of degree $(k_m - 1)$ around x, namely,

$$T_{k_m-1} f(y) = \sum_{|\alpha|_1 \leq k_m - 1} (y - x)^\alpha D^{(\alpha)} f(x)/\alpha!, \tag{4.7}$$

and $R_{k_m-1} f$ is its remainder. Then from the definition of $c_j(x)$ in (4.5) we see that

$$\sum_{x_j \in \mathcal{S}_x} c_j(x) T_{k_m-1}(x_j) = 0,$$

leading to the relation

$$\begin{aligned} L_m f(x) &= \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x} c_j(x) R_{k_m-1} f(x_j) \\ &= \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{S}_x} c_j(x) \sum_{|\alpha|_1 = k_m} (x_j - x)^\alpha D^{(\alpha)} f(\zeta_j)/\alpha! \end{aligned}$$

for some ζ_j between x_j and x. Since $c_j(x)$ and $q_m(x)$ are both $\mathcal{O}(h(x)^{-m})$, we obtain the relation $L_m f(x) = \mathcal{O}(h^{k_m}(x))$, which completes the proof. □

Remark 4.1. As in the univariate case, if the data are given on a uniform grid $\mathcal{S}$, we can find a unique set of coefficients $c_j(x) = c_j$, $j = 1, \ldots, m_2$, with m_2 given in (2.1), and apply it to construct $L_m f(x)$, regardless of the position x (away from the boundary of Ω) and the density $h(x)$ of points. Let $\mathbf{U}$ be a set of integers around the origin with $\#\mathbf{U} = m_2$, and assume that for any x, the shape of the stencil of S_x is the same as $\mathbf{U}$; i.e., there exists $\nu(x) \in h\mathbb{Z}^2 \cap \Omega$ such that

$$S_x = \nu(x) + h\mathbf{U}, \quad h > 0. \tag{4.8}$$

Solving the linear system

$$\sum_{j\in\mathbf{U}} c_j \frac{j^\alpha}{\alpha!} = \delta_{m_2,|\alpha|_1}, \quad \alpha \in \mathbb{Z}_+^2, \quad |\alpha|_1 \le m_2, \tag{4.9}$$

we define $L_m f$ as follows:

$$L_m f(x) = \frac{1}{q_m} \sum_{j\in\mathbf{U}} c_j f(\nu(x) + jh), \tag{4.10}$$

where q_m is also independent of x. Note that c_j and q_m are bounded by a constant, while if $\mathcal{S}_x$ is a scattered data set, they are $\mathcal{O}(h^{-m}(x))$. A straightforward application of the proof in Theorem 4.1 shows that for the uniform case with $h = h(x)$, we have

$$L_m f(x) = \mathcal{O}(h^{\min(k,m)}).$$

4.2. *Minmod* edge detection in two dimensions. As in the one-dimensional case, the utilization of the *minmod* edge detection method increases the area of convergence away from the discontinuities of f. Theorem 4.1 establishes that for a certain order m, the edge detection method $L_m f(x)$ defined in (4.4) converges to zero away from the discontinuities if $K_{\mathcal{S}_x} \cap J = \emptyset$. Here J denotes the set of jump discontinuities of f in Ω. Theorem 4.2 demonstrates that the *minmod* edge detection method converges to zero away from the discontinuities if $K_{T_j} \cap J = \emptyset$, where $T_j \in \mathcal{T}_{\mathcal{S}}$ is defined in (4.1). Clearly this is an improvement since $K_{T_j} \subset K_{\mathcal{S}_x}$.

THEOREM 4.2. *If $x \in K_{T_j}$ and $K_{T_j} \cap J = \emptyset$ for some $T_j \in \mathcal{T}_{\mathcal{S}}$* (4.1), *then the minmod edge detection method* (3.16) *for the set $\mathcal{M} = \{1, 2, \dots, \mu\}$ has the property*

$$MM\big(L_{\mathcal{M}} f(x)\big) = \mathcal{O}(h^{\min(\mathcal{M}_x,k)}(x)),$$

where $\mathcal{M}_x$ is defined as

$$\mathcal{M}_x := \max\big\{m \in \mathcal{M} \,:\, K_{\mathcal{S}_x} \cap J = \emptyset,\ \#\mathcal{S}_x = m_2\big\}, \tag{4.11}$$

and $f \in C^k(K_{\mathcal{S}_x})$ for some $k > 0$ with $\#\mathcal{S}_x \le \binom{\mathcal{M}_x+2}{2}$.

Proof. Assume that $x \in K_{T_j}$ and $K_{T_j} \cap J = \emptyset$ for some $T_j \in \mathcal{T}_{\mathcal{S}}$. Since $K_{T_j} \cap J = \emptyset$ we have $\mathcal{M}_x \ge 1$. Then for any $m \in \mathcal{M}$ such that $m \le \mathcal{M}_x$, the corresponding local set $\mathcal{S}_x$ such that $\#\mathcal{S}_x = m_2$ will satisfy $\mathcal{S}_x \cap J = \emptyset$. Theorem 4.1 then gives $L_m f(x) = \mathcal{O}(h^{\min(m,k)}(x))$. Therefore

$$MM\big(L_{\mathcal{M}} f(x)\big) = \mathcal{O}(h^{\min(\mathcal{M}_x,k)}(x)),$$

which finishes the proof. □

As in the case of one dimension, the choice of $\mathcal{M}$ in Theorem 4.2, an arbitrary set of positive integers, is purposeful. By including $1 \in \mathcal{M}$, first order convergence is ensured at the neighboring cells of discontinuities. Large values are also included in the set $\mathcal{M}$ so that there will be a high order of convergence away from the discontinuities.

Recall that for any particular point $x \in \Omega$, the normalization factor $q_m(x)$ in (4.6) is defined for a subset $\mathcal{P}_x \subset \mathcal{S}_x$ such that $q_m(x) \ne 0$. Theorem 4.3 demonstrates that for a particular $\mathcal{P}_x$ the *minmod* edge detection method will provide a good approximation to the jump magnitudes in the normal directions of the edges. To accomplish this approximation, we provide the following definition.

Definition 4.1. *For an arbitrary point $x \in \Omega$ of a piecewise smooth function f, define the subset $\mathcal{P}_x$ of the local set $\mathcal{S}_x \subset \mathcal{S}$ as*

$$\mathcal{P}_x = \arg\max_{\mathcal{P}}\{\#\mathcal{P}|\mathcal{P} \subset \mathcal{S}_x, \textit{ and } f \in C^k(K_{\mathcal{P}}) \textit{ for some } k > 0\}. \tag{4.12}$$

Therefore $\mathcal{P}_x$ is the largest subset of the local set $\mathcal{S}_x$ such that $f \in C^k(K_{\mathcal{P}_x})$ for some $k > 0$. (A technique for approximating this particular $\mathcal{P}_x$ is provided in Appendix B.*)*

As in the one-dimensional case, $q_m(x)$ can be considered as a generalized version of divided difference for the characteristic function $\chi_{\mathcal{P}_x}$ on $\mathcal{S}_x$ (see the "Remark" following Theorem 3.2). Hence, the condition $q_m(x) \neq 0$ is reasonable. Further, it is assumed without loss of generality in the following analysis that for a small enough local set, if $\mathcal{P}_x \neq \mathcal{S}_x$, then $f \in C^k(K_{\mathcal{P}_x^c \cap \mathcal{S}_x})$ for some $k > 0$. Here $\mathcal{P}_x^c$ indicates the complement of the set $\mathcal{P}_x$. This assumption is similar to the one-dimensional case, where it is assumed that each local set contains at most one discontinuity. If this assumption is not true, the problem is clearly underresolved.

Theorem 4.3 characterizes the *minmod* edge detection method for the two-dimensional edge detection function $|L_m f|$ in (4.4). In this case, we use the absolute value of (4.4) since the jump amplitude may vary depending on paths through a given discontinuity.

Theorem 4.3. *For each $m \in \mathbb{N}$, define $L_m f$ as in* (4.4) *with $q_m(x)$ given in* (4.6)*. If $x \in K_{T_j}$ for some $T_j \in \mathcal{T}_{\mathcal{S}}$, then the minmod edge detection method* (3.16) *for the set $\mathcal{M} = \{1, 2, \ldots, \mu\}$ has the property*

$$MM\big(|L_{\mathcal{M}} f(x)|\big) = \begin{cases} [F](x) + \mathcal{O}(h(x)) & \textit{if } K_{T_j} \cap J \neq \emptyset, \\ \mathcal{O}(h^{\min(\mathcal{M}_x, k)}(x)) & \textit{if } K_{T_j} \cap J = \emptyset, \end{cases}$$

where $\mathcal{M}_x$ is defined in (4.11)*, $f \in C^k(K_{\mathcal{S}_x})$ for some $k > 0$ with $\#\mathcal{S}_x \leq \binom{\mathcal{M}_x+2}{2}$, and*

$$[F](x) := \max\big\{|f(u) - f(v)| \,:\, u \in K_{\mathcal{P}_x} \cap K_{T_j},\ v \in K_{\mathcal{P}_x^c \cap \mathcal{S}_x} \cap K_{T_j}\big\}. \tag{4.13}$$

Proof. For any given integer $m \in \mathcal{M}$, choose the local set $\mathcal{S}_x$ such that $\#\mathcal{S}_x = m_2$. Assume first $K_{T_j} \cap J \neq \emptyset$. Then clearly $\mathcal{P}_x, \mathcal{P}_x^c \cap \mathcal{S}_x \neq \emptyset$. Now, for some $\beta_j, \gamma_j \in (0, 1)$, we have

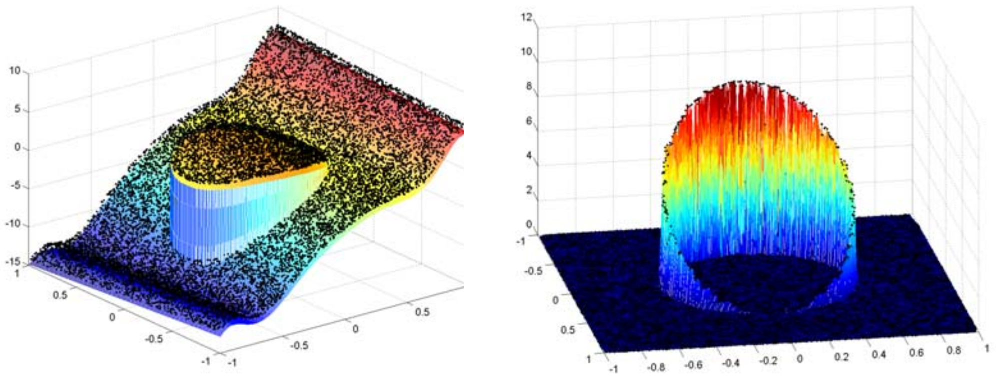
$$\begin{aligned} |L_m f(x)| &= \left| \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{P}_x} c_j(x) f(x_j) + \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{P}_x^c \cap \mathcal{S}_x} c_j(x) f(x_j) \right| \\ &= \Bigg| \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{P}_x} c_j(x) \Bigg[f(u) + \sum_{|\alpha|_1 = 1} (x_j - u)^\alpha D^\alpha f(u + \beta_j(x_j - u)) \Bigg] \\ &\quad + \frac{1}{q_m(x)} \sum_{x_j \in \mathcal{P}_x^c \cap \mathcal{S}_x} c_j(x) \Bigg[f(v) + \sum_{|\alpha|_1 = 1} (x_j - v)^\alpha D^\alpha f(v + \beta_j(x_j - u)) \Bigg] \Bigg| \end{aligned}$$

for any $u \in K_{\mathcal{P}_x} \cap K_{T_j}$ and $v \in K_{\mathcal{P}_x^c \cap \mathcal{S}_x} \cap K_{T_j}$. From the condition $\sum_{j \in \mathcal{S}_x} c_j(x) = 0$, we see from (4.6) that

$$q_m(x) = \sum_{x_j \in \mathcal{P}_x} c_j(x) = - \sum_{x_j \in \mathcal{P}_x^c \cap \mathcal{S}_x} c_j(x),$$

and therefore

$$|L_m f(x)| = |f(u) - f(v)| + \mathcal{O}(h(x)).$$



(a) **(b)**

FIG. 4.2. (a) $f_1(x)$ *from Example* 4.1 *sampled on random points with* $\#\mathcal{S} = 128^2$. (b) *The minmod edge detection method of* $MM(L_{\mathcal{M}} f_1(x))$ *for* $\mathcal{M} = \{1, 2, 3, 4\}$.

Since u and v are arbitrary points in the sets $K_{\mathcal{P}_x} \cap K_{T_j}$ and $K_{\mathcal{P}_x^c \cap \mathcal{S}_x} \cap K_{T_j}$, respectively, we obtain that $|L_m f(x)| = [F](x) + \mathcal{O}(h(x))$, where $[F](x)$ is defined in (4.13), yielding

$$MM\big(|L_{\mathcal{M}} f(x)|\big) = [F](x) + \mathcal{O}(h(x)).$$

Next, assume that $K_{T_j} \cap J = \emptyset$. By definition, $\mathcal{M}_x \geq 1$. Then for any $m \in \mathcal{M}$ such that $m \leq \mathcal{M}_x$, the corresponding local set $\mathcal{S}_x$ such that $\#\mathcal{S}_x = m_2$ will satisfy $\mathcal{S}_x \cap J = \emptyset$. Theorem 4.1 then gives $L_m f(x) = \mathcal{O}(h^{\min(m,k)}(x))$. Clearly it can be concluded that

$$MM\big(|L_{\mathcal{M}} f(x)|\big) = \mathcal{O}(h^{\min(\mathcal{M}_x,k)}(x)),$$

which finishes the proof. □

To demonstrate the efficacy of the *minmod* edge detection method in two dimensions we consider the following example.

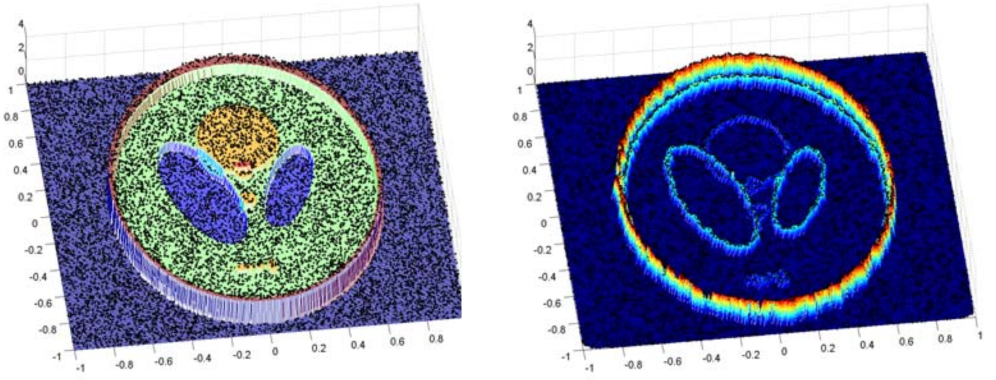
Example 4.1.

$$f_1(x) := f_1(u,v) := \begin{cases} uv + \cos(2\pi u^2) - \sin(2\pi u^2) & \text{if } u^2 + v^2 \leq \frac{1}{4}, \\ 10u - 5 + uv + \cos(2\pi u^2) - \sin(2\pi u^2) & \text{if } u^2 + v^2 > \frac{1}{4} \end{cases}$$

for $-1 \leq u, v \leq 1$.

Note that the edges comprise the circle $u^2 + v^2 = \frac{1}{4}$ with the exception of $u = \frac{1}{2}$, where the function is smooth. Figure 4.2(a) shows $f_1(x)$ sampled on a MATLAB randomly generated data set $\mathcal{S}$ with $\#\mathcal{S} = 128^2$. Figure 4.2(b) displays the results of applying the *minmod* edge detection method to $L_m f_1$ with $m \in \mathcal{M} = \{1, 2, 3, 4\}$. Of particular interest is the ability of the *minmod* edge detection method to resolve the positions and magnitudes in the normal direction of the edges even in areas of sparse sampling and steep gradients.

Let us now turn our attention to a practical example often used as a benchmark test for edge detection in magnetic resonance imaging (MRI).



(a) **(b)**

FIG. 4.3. (a) $f_2(x)$ *from Example* 4.2 *sampled on random points with* $\#\mathcal{S} = 128^2$. (b) *The minmod edge detection method of* $MM(L_{\mathcal{M}} f_2(x))$ *for* $\mathcal{M} = \{1, 2, 3, 4\}$.

Example 4.2. The so-called Shepp–Logan phantom, $f_2(x)$, defined in Appendix C.

Note that the edges of the Shepp–Logan phantom comprise various ellipses of different sizes and orientations, some of which overlap. Figure 4.3(a) shows the Shepp–Logan phantom (denoted as $f_2(x)$), sampled on a MATLAB randomly generated data set $\mathcal{S}$ with $\#\mathcal{S} = 128^2$. Figure 4.3(b) displays the results of applying the *minmod* edge detection method on $L_m f_2(x)$ with $m \in \mathcal{M} = \{1, 2, 3, 4\}$. Of particular interest is the ability of the *minmod* edge detection method to resolve edges that reside in neighboring centers.

The algorithm in Appendix B details the two-dimensional edge detection computation for Examples 4.1 and 4.2, where the particular choice of local sets, reconstruction points, and basis functions are specified. Although no formal computational cost studies were conducted, our experiments indicate that the two-dimensional algorithm experiences minimal increase in computational effort.

5. Concluding remarks. In this paper we have introduced an edge detection method (2.5) based on a local polynomial annihilation property on a set of irregularly distributed points in a bounded domain $\Omega \subset \mathbb{R}^d$. The method successfully captures discontinuities that are identified by their enclosed cells by characterizing the convergence away from the discontinuities. Although the convergence of the edge detection method can be of high order away from discontinuities, there are problematic oscillations in the neighborhoods of discontinuities. The *minmod* function (3.6) for one-dimensional global edge detection methods enables the distinction of jump discontinuities from neighborhood oscillations by the effective use of the information intrinsic to the edge detection approximation. The resulting *minmod* edge detection method ensures the highest rate of convergence *up to* the enclosed points of discontinuities.

The edge detection method described in our study is local, numerically cost efficient, and entirely independent of any specific shape or complexity of boundaries. Furthermore, it demonstrates the ability to detect edges of piecewise smooth functions with steep gradients as well as in low resolution environments with sparse, nonuniform sampling. For uniformly distributed points, the cost of computation is significantly reduced since the coefficients in the edge detection method are constant for every type of local stencil.

This study is concerned with the detection of jump discontinuities. Our future work will focus on integrating this method to real signals and images in various scientific disciplines, where noise, poor resolution, and numerical efficiency all become critical issues. We also are currently generalizing our method to determine jump discontinuities in the derivatives, critical for resolving texture in images.

Appendix A. One-dimensional edge detection algorithm. For any $x \in [a,b]$, let $\mathcal{S}_x$ be the closest $m_1 = m+1$ points to x in $\mathcal{S}$. As a basis of Π_m, choose $p_i(x) = x^{i-1}$ for $i = 1, \ldots, m_1$. The *minmod* function for $\mathcal{M} = \{1, 2, \ldots, \mu\}$ will be reconstructed on the points $x_{j+\frac{1}{2}} = \frac{1}{2}(x_{j+1} + x_j)$ with $j = 1, \ldots, N-1$.

```
for $m = 1$ to $\mu$ and $j = 1$ to $N-1$
    step 1. For each $x_{j+\frac{1}{2}}$, define $\mathcal{S}^+_{x_{j+\frac{1}{2}}} = \{x_n | x_n > x_{j+\frac{1}{2}}\}$ and set $r = \#\mathcal{S}^+_{x_{j+\frac{1}{2}}}$.
    step 2. Calculate the coefficients
            $c_i(x_{j+\frac{1}{2}}) = \frac{m!}{\omega_i(\mathcal{S}_{x_{j+\frac{1}{2}}})}, \quad i = 1, \ldots, m_1,$
        where $\omega_i(\mathcal{S}_{x_{j+\frac{1}{2}}})$ is defined as in (3.9).
    step 3. Calculate the normalization factor
            $q_m(x_{j+\frac{1}{2}}) = \sum_{i=m_1-r+1}^{m_1} c_i(x_{j+\frac{1}{2}}).$
    step 4. Compute the jump function
            $L_m f(x_{j+\frac{1}{2}}) = \frac{1}{q_m(x_{j+\frac{1}{2}})} \sum_{i=1}^{m_1} c_i(x_{j+\frac{1}{2}}) f(x_{i+j+r-m_1}).$
end ($m$, $j$)
step 5. Apply minmod edge detection method $MM\big(L_{\mathcal{M}} f(x_{j+\frac{1}{2}})\big)$.
```

Appendix B. Two-dimensional edge detection algorithm. Let $\mathcal{S} := \{x_j := (u_j, v_j) \,|\, j = 1, \ldots, N\} \subset \Omega$ and choose $p_\alpha(x) = u^{\alpha_1} v^{\alpha_2}$ for $x = (u,v)$ and $\alpha = (\alpha_1, \alpha_2) \in \mathbb{Z}^2_+$ such that $|\alpha|_1 \leq m$ as a basis of Π_m. The *minmod* function for $\mathcal{M} = \{1, 2, \ldots, \mu\}$ will be reconstructed on the set

$$\mathcal{D}_{\mathcal{T}_\mathcal{S}} = \left\{ \bar{x}_j \,\middle|\, \bar{x}_j = \frac{\sum_{i=1}^{3} x_i^j}{3}, \text{ where } T_j = \{x_1^j, x_2^j, x_3^j\} \in \mathcal{T}_\mathcal{S} \text{ for } j = 1, \ldots, N_T \right\}. \tag{B.1}$$

```
for $m = 1$ to $\mu$ and $j = 1$ to $N_T$
    step 1. For $\bar{x}_j$ in (B.1), determine $\mathcal{S}_{\bar{x}_j}$ as in (4.2) with $\#\mathcal{S}_{\bar{x}_j} = m_2 = \binom{m+2}{2}$.
            Set $\mathcal{S}_{\bar{x}_j} = \{x_1, \ldots, x_{m_2}\}$ such that $f(x_1) \leq f(x_2) \leq \cdots \leq f(x_{m_2})$.
    step 2. Solve the linear system
            $\sum_{x_i \in S_{\bar{x}_j}} c_i(\bar{x}_j) p_\alpha(x_i) = \begin{cases} 0 & \text{if } \alpha_1 + \alpha_2 < m, \\ \alpha_1! \alpha_2! & \text{if } \alpha_1 + \alpha_2 = m \end{cases}$
        for $\alpha_1, \alpha_2 = 0, \ldots, m$, such that $\alpha_1 + \alpha_2 \leq m$.
    step 3. Calculate $q_m(\bar{x}_j)$ as in (4.6). Here the subset $\mathcal{P}_x$ (4.12) of $\mathcal{S}_{\bar{x}_j}$ is computed as
                $\mathcal{P}_x = \{x_1, \ldots, x_r\}$,
            where
                $|f(x_{r+1}) - f(x_r)| = \max_{i=1,\ldots,m_2-1} |f(x_{i+1}) - f(x_i)|$.
            If such $r$ is not unique, choose the smallest one.
    step 4. Calculate $L_m f(\bar{x}_j) = \frac{1}{q_m(\bar{x}_j)} \sum_{x_i \in \mathcal{S}_{\bar{x}_j}} c_i(\bar{x}_j) f(x_i)$.
end $(m, j)$
step 5. Apply minmod edge detection method $MM\big(L_{\mathcal{M}} f(\bar{x}_j)\big)$.
```

Appendix C. Shepp–Logan phantom algorithm. The Shepp–Logan phantom is a piecewise smooth function on the domain $\Omega = [-1,1] \times [-1,1]$ in $\mathbb{R}^2$. For any arbitrary point $(u,v) \in \Omega$ the value of the Shepp–Logan phantom $z = f(u,v)$ is calculated as follows:

```
for each point $(u,v)$
    let $z = 0$, $\xi_1 = (u - .22)\cos(.4\pi) + v\sin(.4\pi)$, $\eta_1 = -(u - .22)\sin(.4\pi) + v\cos(.4\pi)$, $\xi_2 = (u+.22)\cos(.6\pi) + v\sin(.6\pi)$, and $\eta_2 = -(u+.22)\sin(.6\pi) + v\cos(.6\pi)$.
    if $(\frac{u}{.69})^2 + (\frac{v}{.92})^2 \leq 1$,
        then $z = 2$.
    if $(\frac{u}{.06624})^2 + (\frac{v+.0184}{.874})^2 \leq 1$,
        then $z = z - .98$.
    if $(\frac{\xi_1}{.31})^2 + (\frac{\eta_1}{.11})^2 \leq 1$ or $(\frac{\xi_2}{.41})^2 + (\frac{\eta_2}{.16})^2 \leq 1$,
        then $z = z - .02$.
    if $(\frac{u-.35}{.3})^2 + (\frac{v}{.6})^2 \leq 1$, or $(\frac{u}{.21})^2 + (\frac{v-.35}{.25})^2 \leq 1$, or $(\frac{u}{.046})^2 + (\frac{v-.1}{.046})^2 \leq 1$, or $(\frac{u}{.046})^2 + (\frac{v+.1}{.046})^2 \leq 1$, or $(\frac{u+.08}{.046})^2 + (\frac{v+.605}{.023})^2 \leq 1$, or $(\frac{u}{.023})^2 + (\frac{v+.605}{.023})^2 \leq 1$, or $(\frac{u-.06}{.023})^2 + (\frac{v+.605}{.023})^2 \leq 1$,
        then $z = z + .01$.
end.
```

REFERENCES

[1] R. Abgrall, *On essentially non-oscillatory schemes on unstructured meshes: Analysis and implementation*, J. Comput. Phys., 114 (1994), pp. 45–54.

[2] J. Canny, *A computational approach to edge detection*, IEEE Trans. Pattern Anal. Machine Intell., 8 (1986), pp. 679–698.

[3] A. Gelb and E. Tadmor, *Detection of edges in spectral data*, Appl. Comput. Harmon. Anal., 7 (1999), pp. 101–135.

[4] A. Gelb and E. Tadmor, *Detection of edges in spectral data* II. *Nonlinear enhancement*, SIAM J. Numer. Anal., 38 (2000), pp. 1389–1408.

[5] M. Gökmen and A. Jain, *$\lambda\tau$-space representation of images and generalized edge detector*, IEEE Trans. Pattern Anal. Machine Intell., 19 (1997), pp. 545–563.

[6] A. Harten, B. Engquist, S. Osher, and S. Chakarvarthy, *Uniformly high order essentially non-oscillatory schemes*, III, J. Comput. Phys., 71 (1987), pp. 231–303.

[7] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Texts, Cambridge, UK, 2002.

[8] Z.-P. Liang and P. C. Lauterbur, *Principles of Magnetic Resonance Imaging: A Signal Processing Perspective*, IEEE Press, New York, 2000, pp. 241–243.

[9] X.-D. Liu, S. Osher, and T. Chan, *Weighted essentially non-oscillatory schemes*, J. Comput. Phys., 115 (1994), pp. 200–212.

[10] F. P. Preparata and M. I. Shamos, *Computational Geometry: An Introduction*, Springer-Verlag, New York, 1985.

[11] S. Mallat and W. Hwang, *Singularity detection and processing with wavelets*, IEEE Trans. Inform. Theory, 38 (1992), pp. 617–643.

[12] C.-W. Shu, *Essentially Non-oscillatory and Weighted Essentially Non-oscillatory Schemes for Hyperbolic Conservation Laws*, NASA CR-97-206253 ICASE Report 97-65, 1997.

[13] E. Tadmor, *Private communication*, 2003.

[14] E. T. Whittaker and G. Robinson, *The Calculus of Observation: A Treatise on Numerical Mathematics*, 4th ed., Dover, New York, 1967.

[15] Y.-T. Zhang and C.-W. Shu, *High-order WENO schemes for Hamilton–Jacobi equations on triangular meshes*, SIAM J. Sci. Comput., 24 (2003), pp. 1005–1030.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 280–302

A DISCONTINUOUS PETROV–GALERKIN METHOD WITH LAGRANGIAN MULTIPLIERS FOR SECOND ORDER ELLIPTIC PROBLEMS*

PAOLA CAUSIN[†] AND RICCARDO SACCO[‡]

Abstract. We present a discontinuous Petrov–Galerkin (DPG) method for the finite element discretization of second order elliptic boundary value problems. The novel approach emanates from a one-element weak formulation of the differential problem. This procedure, which is typical of discontinuous Galerkin (DG) methods, is based on introducing variables defined in the interior and on the boundary of the element. The interface variables are suitable Lagrangian multipliers that enforce interelement continuity of the solution and of its normal derivative, thus providing the proper connection between neighboring elements. The internal variables can be eliminated in favor of the interface variables using static condensation to end up with a system of reduced size in the sole Lagrangian multipliers. A stability and convergence analysis of the novel formulation is carried out and its connection with mixed-hybrid and DG methods is explored. Numerical tests on several benchmark problems are included to validate the convergence performance and the flux-conservation properties of the DPG method.

Key words. Petrov–Galerkin formulations, mixed and hybrid finite element methods, discontinuous Galerkin methods, elliptic problems

AMS subject classifications. 65N12, 65N30, 65N15

DOI. 10.1137/S0036142903427871

1. Introduction and motivation.

Recent years have seen an ever increasing use, development, and analysis of *discontinuous methods* in the approximation of boundary value problems. Within this active research area, discontinuous Galerkin (DG) formulations certainly occupy a prominent position (we refer to [19] for a survey on the state-of-the-art of the literature on DG methods) and their success in the approximation of hyperbolic problems has extended their use to cover the case of parabolic and elliptic equations.

A considerable impulse in the direction of extending the use of DG methods to parabolic and elliptic equations is due to the contributions given in [6, 7], where discontinuous finite elements of high order are used in the numerical solution of the compressible Navier–Stokes equations. Two methodological aspects in [6, 7] are of particular importance for their influence on later research activity.

The first aspect is the technique used to accommodate the viscous terms arising in the momentum and energy balance equations within the structure of the DG formulations traditionally devoted to hyperbolic problems. The technique consists of introducing a new unknown, related to the gradient of the conservative variables, and then providing a consistent approximation for the new unknown. This strategy is closely related to classical *mixed methods* and is one of the starting motivations of the work conducted, although in different directions, in [2, 3, 16] and in the present article.

[*] Received by the editors May 13, 2003; accepted for publication (in revised form) July 12, 2004; published electronically May 27, 2005. This work was partially supported by the MURST Cofin 2001 Project "Metodi Numerici Avanzati per Equazioni alle Derivate Parziali di Interesse Applicativo."
http://www.siam.org/journals/sinum/43-1/42787.html

[†] INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt B.P. 105, 78153 Le Chesnay Cedex, France (paola.causin@inria.fr).

[‡] Dipartimento di Matematica "F. Brioschi," Politecnico di Milano, via Bonardi 9, 20133 Milano, Italy (riccardo.sacco@mate.polimi.it).

The second aspect is the extension of the concept and use of *numerical fluxes* in the treatment of boundary terms arising from integration by parts of the equations at the element level. Numerical fluxes are a key ingredient of any performing DG formulation and must be properly designed to impart stability and accuracy to the approximation. This is usually done by borrowing their expression from finite volume techniques, as discussed in [3] in the case of DG methods applied to the numerical solution of elliptic boundary value problems. The choice of numerical fluxes in DG methods is not trivial since it must be tailored to the problem at hand, leading in some cases to an involved implementation of the resulting scheme, a drawback that is quite common to many high-order finite volume formulations.

The motivation of the discontinuous Petrov–Galerkin (DPG) method proposed in the present article strongly arises from this latter observation. It is indeed a fact that the *values of the variables on the element boundaries* (or an appropriate representation of them) are the ingredients to be used to provide the necessary coupling between neighboring elements. Having this clearly in mind, an alternative approach to numerical flux definition may be pursued by introducing *independent interface variables* that are single-valued functions solely defined on element boundaries (*hybrid interface variables*). The hybrid interface variables are suitable Lagrangian multipliers that enforce the continuity of the displacement (the scalar variable of the problem) and of the normal stress (the vector variable of the problem) across the interfaces of the finite element triangulation. By doing so, proper interelement connection can be enforced without needing to exhibit any specific upfront recipe for the numerical flux. Therefore, the DPG method establishes a connection between DG and hybrid methods, a connection that is presently the object of analogous research activity by many authors in different areas (see, for example, [21, 22, 20]).

The DPG method was proposed in [10], where a stability and convergence analysis of the formulation was carried out in one spatial dimension. Then, the method was applied to the numerical solution of scalar advective-diffusive models [12, 11] and of fluid-mechanical problems in both compressible and incompressible regimes [17].

In the present article we carry out the theoretical analysis of the stability and convergence properties of the novel formulation applied to the solution of an elliptic boundary value model problem in two spatial dimensions, an aspect that was still lacking. We also discuss the efficient computer implementation of the scheme, thus strengthening the connection between the DPG methodology and classical DG and mixed-hybrid approaches. Numerical results are then shown to demonstrate the convergence and conservation properties of the novel formulation.

The paper is organized as follows: in section 2 we introduce the one-element weak formulation that is the starting point of the DPG approach. In section 3 we set up the formulation at the continuous level and we carry out its stability analysis. In section 4 we introduce the corresponding approximation and in section 5 we discuss the construction of appropriate finite element spaces, addressing in particular the case of the element of lowest degree (DPG_0) for which we carry out a stability and error analysis in section 6. We address the issue of an efficient implementation of the DPG_0 formulation in section 7. In section 8 we present some numerical results to validate the convergence performance, while in section 9 we assess the conservation properties of the DPG method. Finally, in section 10 we end with some concluding remarks.

2. One-element formulation of the elliptic model problem.

We consider the following elliptic model problem:

$$-\mathrm{div}\,\nabla\, u = f \quad \text{in } \Omega, \qquad u = g_D \quad \text{on } \Gamma_D, \qquad \nabla\, u \cdot \boldsymbol{n} = g_N \quad \text{on } \Gamma_N, \tag{2.1}$$

where Ω is an open bounded set of $\mathbb{R}^2$ with Lipschitz continuous boundary $\Gamma = \partial\Omega$ such that $\Gamma = \Gamma_D \cup \Gamma_N$, $\Gamma_D \neq \emptyset$, and where f, g_D, and g_N are given functions. Problem (2.1) will be referred to as the *primal formulation* and u as the *primal unknown*. Upon introducing the auxiliary unknown $\boldsymbol{\sigma} = \nabla u$, problem (2.1) may be rewritten as the first order system

$$\left\{ \begin{array}{llll} -\mathrm{div}\,\boldsymbol{\sigma} = f & \text{in}\,\Omega, & \boldsymbol{\sigma} = \nabla u & \text{in}\,\Omega, \\ u = g_D & \text{on}\,\Gamma_D, & \boldsymbol{\sigma}\cdot\boldsymbol{n} = g_N & \text{on}\,\Gamma_N. \end{array} \right. \tag{2.2}$$

Problem (2.2) will be referred to as the *mixed formulation* of (2.1). In this latter context we shall refer in a generalized sense to the mixed unknowns u and $\boldsymbol{\sigma}$ as *displacements* and *stresses*, respectively.

Given a triangulation $\mathcal{T}_h$ of Ω made of triangles, we consider the following *one-element* weak form of problem (2.2) (see section 3.1 for the notation):

$\forall\, K \in \mathcal{T}_h$, find $(\boldsymbol{\sigma}^K, u^K)$ such that

(2.3)
$$\left\{ \begin{array}{ll} \displaystyle\int_K \boldsymbol{\sigma}^K \cdot \boldsymbol{q}^K\,dx + \int_K u^K\,\mathrm{div}\,\boldsymbol{q}^K\,dx - \int_{\partial K} u_{\partial K}\,\boldsymbol{q}_{\partial K}\cdot\boldsymbol{n}_K\,ds = 0 & \forall\,\boldsymbol{q}^K, \\ \displaystyle\int_K \boldsymbol{\sigma}^K \cdot \nabla v^K\,dx - \int_{\partial K} \boldsymbol{\sigma}_{\partial K}\cdot\boldsymbol{n}_K\,v_{\partial K}\,ds = \int_K f^K\,v^K\,dx & \forall v^K, \end{array} \right.$$

where $\boldsymbol{\sigma}^K, u^K, \boldsymbol{q}^K$, and v^K belong to spaces of smooth vector and scalar functions defined on K and where the symbols $\boldsymbol{\sigma}_{\partial K}$ and $u_{\partial K}$ represent the traces on ∂K of $\boldsymbol{\sigma}^K$ and u^K, respectively, properly accounting for the boundary conditions. Notice that a formal integration by parts has been performed on both equations in (2.2).

System (2.3) is a general setting from which both discontinuous Galerkin and hybrid formulations can be derived, the latter after a suitable use of integration by parts in $(2.3)_1$ or $(2.3)_2$. The common factor shared by DG and hybrid formulations relies on the role played by the variables traced on the element interfaces that are the connectors demanded to preserve the proper coupling between K and its neighbors.

In DG methods the interelement constraints are enforced by defining on ∂K specific expressions for $\boldsymbol{\sigma}_{\partial K}$ and $u_{\partial K}$ as functions of the internal variables, the so-called *numerical fluxes* (see [2, 3, 16]).

In hybrid formulations the variables traced on the element interfaces are instead suitable Lagrange multipliers and are *additional* unknowns of the problem. In particular, primal mixed-hybrid methods [35, 37] are obtained by integrating by parts $(2.3)_1$, while dual mixed-hybrid methods [14] are obtained by integrating by parts $(2.3)_2$. System (2.3) is thus in dual-primal mixed-hybrid form. In both cases a symmetric Galerkin formulation is obtained from the nonsymmetric formulation (2.3) and only *one* Lagrangian multiplier is introduced, with the conclusion that in hybrid formulations a different numerical treatment is applied to the displacement and stress fields.

The choice of introducing independent interface unknowns as interelement connectors, thus avoiding the need of ad hoc definition of the numerical fluxes, while preserving at the same time a completely parithetic (and discontinuous) approximation of u and $\boldsymbol{\sigma}$ on $\mathcal{T}_h$, as in DG formulations, is the main idea underlying the DPG method discussed in the forthcoming sections.

3. The DPG formulation.

In what follows we introduce the DPG formulation and carry out a stability and convergence analysis of the method.

3.1. Notation and functional setting. We let $\overline{\Omega} = \cup \overline{K}$ be a regular partition $\mathcal{T}_h$ of the domain Ω into triangular elements K (see [18]); i.e., we suppose that there exists a constant $\sigma \geq 1$ such that $(h_K/\rho_K) \leq \sigma$ for all $K \in \mathcal{T}_h$, h_K being the diameter of K and $\rho_K = \sup\{\mathrm{diam}(S) \,|\, S \,\text{is a ball contained in}\, K\}$. We let $\mathcal{E}_h$ be the set of the edges of $\mathcal{T}_h$, and the edge shared by the elements K and K' will be referred to as $e_{K-K'}$. For each element $K \in \mathcal{T}_h$, we denote by ∂K the Lipschitz continuous boundary of K and by $\boldsymbol{n}_K$ the unit outward normal vector along the boundary ∂K. We also let $\partial K_D = \partial K \cap \Gamma_D$, $\partial K_N = \partial K \cap \Gamma_N$. Moreover, if v is any function defined in Ω, we denote by v^K its restriction to the element K and by $v_{\partial K}$ its restriction to the element boundary ∂K. Similarly, if η is any function defined on $\mathcal{E}_h$, we denote by $\eta_{\partial K}$ its restriction on $\mathcal{E}_h \cap \partial K$.

Given an integer $m \geq 0$ and the real numbers $p, q \in [1, \infty)$, we define the following local space:

$$W^{m,p}(K) = \{v \in L^p(K) \,|\, D^\alpha v \in L^p(K) \,\forall \alpha, |\alpha| \leq m\} \qquad \forall K \in \mathcal{T}_h,$$

endowed with the usual norm and seminorm $||v||_{m,p,K}$ and $|v|_{m,p,K}$. When $p = 2$, $W^{m,2}(K)$ is the usual $H^m(K)$ Sobolev space (see [27, 28]), and the simplified notation $||.||_{m,K}$ and $|.|_{m,K}$ will be used. We also introduce the local space

$$W_q(\mathrm{div}; K) = \left\{\boldsymbol{\tau} \in (L^q(K))^2 \,|\, \mathrm{div}\, \boldsymbol{\tau} \in L^q(K)\right\},$$

endowed with the usual graph norm $||\boldsymbol{\tau}||_{q,\mathrm{div},K}$. When $q = 2$, the space $W_2(\mathrm{div}; K)$ is the Sobolev space $H(\mathrm{div}; K)$ (see [14]).

From now on, p and q will be chosen to be conjugate numbers, i.e., $1/p + 1/q = 1$. It will be useful in what follows to consider the space of the traces on ∂K of functions $v \in W^{1,p}(K)$ and $\boldsymbol{\tau} \in W_q(\mathrm{div}; K)$. Notice that the trace $v_{\partial K}$ belongs to the space $W^{1/q,p}(\partial K)$, while the normal trace $\boldsymbol{\tau} \cdot \boldsymbol{n}|_{\partial K}$ belongs to the space $W^{-1/q,q}(\partial K)$; the spaces $W^{1/q,p}(\partial K)$ and $W^{-1/q,q}(\partial K)$ are endowed with the following norms:

$$(3.1) \qquad ||\boldsymbol{\tau} \cdot \boldsymbol{n}||_{-1/q,q,\partial K} = \sup_{v \in W^{1,p}(K)} \frac{\langle \boldsymbol{\tau} \cdot \boldsymbol{n}, v\rangle_{\partial K}}{||v||_{1,p,K}} \quad \forall \boldsymbol{\tau} \in W_q(\mathrm{div}; K)$$

and

$$(3.2) \qquad ||v||_{1/q,p,\partial K} = \sup_{\boldsymbol{\tau} \in W_q(\mathrm{div};K)} \frac{\langle \boldsymbol{\tau} \cdot \boldsymbol{n}, v\rangle_{\partial K}}{||\boldsymbol{\tau}||_{W_q(\mathrm{div};K)}} \qquad \forall v \in W^{1,p}(K).$$

Note that for $p = q = 2$ the quantity in (3.1) is the standard norm $||\boldsymbol{\tau} \cdot \boldsymbol{n}||_{-1/2,\partial K}$.

3.2. DPG weak formulation. Proceeding along the same lines as in [23, 24], we assume henceforth that $\frac{4}{3} < p < 2$, and thus $2 < q < 4$.

We introduce the trial function spaces

$$\Sigma = (L^q(\Omega))^2, \qquad U = L^q(\Omega),$$

$$\Lambda = \Bigg\{\lambda \in \prod_{K \in \mathcal{T}_h} W^{1/q,p}(\partial K), \lambda^K = \lambda^{K'} \text{ on } e_{K-K'} \;\forall K, K' \in \mathcal{T}_h,$$
$$\lambda^K = g_D \text{ on } \partial K_D \;\forall K \in \mathcal{T}_h\Bigg\},$$

$$M = \Bigg\{\mu \in \prod_{K \in \mathcal{T}_h} H^{-1/2}(\partial K),\; \mu^K + \mu^{K'} = 0 \text{ on } e_{K-K'} \;\forall K, K' \in \mathcal{T}_h,$$
$$\mu^K = g_N \text{ on } \partial K_N \;\forall K \in \mathcal{T}_h\Bigg\}$$

and the test function spaces $W = \prod_{K\in\mathcal{T}_h} W_q(\mathrm{div}; K)$ and $V = \prod_{K\in\mathcal{T}_h} H^1(K)$. We set $X = (U \times \Lambda)$, $Y = (\Sigma \times M)$ and we introduce the compact notation $\widetilde{u} = (u; \lambda)$ and $\widetilde{\boldsymbol{\sigma}} = (\boldsymbol{\sigma}; \mu)$. The DPG weak formulation of problem (2.1) is obtained from (2.3) by introducing the hybrid variables λ and μ to represent the values $u_{\partial K}$ and $\sigma_{\partial K}$, respectively, and by summing up on the triangles and reads:

find $(\widetilde{u}, \widetilde{\boldsymbol{\sigma}}) \in \big(X \times Y\big)$ such that

$$\left\{\begin{array}{lclcll} a(\widetilde{\boldsymbol{\sigma}}, \boldsymbol{q}) & + & b_1(\widetilde{u}, \boldsymbol{q}) & = & 0 & \forall \boldsymbol{q} \in W, \\ b_2(\widetilde{\boldsymbol{\sigma}}, v) & & & = & (f, v) & \forall v \in V, \end{array}\right. \tag{3.3}$$

where $(\cdot,\cdot)$ is the usual L^2 product and where we have set

$$a(\widetilde{\boldsymbol{\sigma}}, \boldsymbol{q}) = \sum_{K\in\mathcal{T}_h} \int_K \boldsymbol{\sigma} \cdot \boldsymbol{q}\, dx, \qquad b_1(\widetilde{u}, \boldsymbol{q}) = \sum_{K\in\mathcal{T}_h} \left(\int_K u \,\mathrm{div}\boldsymbol{q}\, dx - \int_{\partial K} \lambda \boldsymbol{q} \cdot \boldsymbol{n}\, ds \right),$$

$$b_2(\widetilde{\boldsymbol{\sigma}}, v) = \sum_{K\in\mathcal{T}_h} \left(\int_K \boldsymbol{\sigma} \cdot \nabla v\, dx - \int_{\partial K} \mu\, v\, ds \right).$$

Due to the *simultaneous* presence of the two Lagrangian multipliers λ and μ, the resulting scheme lacks the formal symmetry of a standard Galerkin mixed-hybrid formulation and becomes a DPG method for the numerical approximation of second order boundary value problems. It is characterized by a completely equal treatment of the mixed variables u and $\boldsymbol{\sigma}$. Indeed, since the integration by parts has relaxed all the regularity requirements on u and $\boldsymbol{\sigma}$ at the expense of more regular test functions $\boldsymbol{q}$ and v, an equal-order interpolation for these internal fields is allowed in the finite element approximation of (3.3).

Remark 3.1. The choice of the unknowns and of the corresponding functional setting in (3.3) allows one to interpret the DPG formulation as a (suitable) hybridization of the dual-mixed problem (2.2), the hybridization procedure being carried out here "on the continuous level" rather than "on the discrete level" as done in the classical reference work [1]. The advantage of dual-mixed formulations with hybridization on standard dual-hybrid formulations is that the former yield an approximation of the normal stresses that is *continuous* on each edge of $\mathcal{T}_h$, while the latter ensure flux conservation only in an average sense over the patch of elements surrounding each node of the triangulation. In [1], this goal is achieved by introducing a hybrid variable that appears only as a discrete function (the so-called λ-trick), while with the present functional setting, functions in $W^{1/q,p}(\partial K)$ need not be continuous at the vertices of ∂K, which allows for a *fully discontinuous* approximation of the hybrid variable over $\mathcal{E}_h$. Notice that the extra amount of regularity ($q > 2$) required to achieve the desired conservation properties has no practical limiting consequences on the choice of the finite element spaces for the approximation of functions in $W_q(\mathrm{div}; K)$.

Remark 3.2. Given the above functional spaces, it will be possible to prove the (weak) coerciveness of $a(\cdot,\cdot)$, $b_1(\cdot,\cdot)$, and $b_2(\cdot,\cdot)$ in norms weaker than the natural ones (cf. Propositions 3.3, 3.4, and 3.5, Corollary 3.7, and the corresponding discrete ones in section 6.1). This, however, will be enough [23, 24] to establish uniqueness of the weak solution of (3.3) and of its corresponding finite element discretization (see section 6.1), as well as superconvergence error estimates for the approximation λ_h of the hybrid variable λ in a natural boundary norm (see section 6.3).

Remark 3.3. Whenever continuous test functions $\boldsymbol{q}$ and v are used in (3.3), we recover the dual-primal method proposed and analyzed in [30]. Therefore, the DPG method can be fully regarded as a hybridization of the above-mentioned scheme. Moreover, the mixed system obtained by taking continuous test functions on Ω for $\boldsymbol{\sigma}$ and u in $(3.4)_1$ and $(3.6)_1$ yields, upon summing over $\mathcal{T}_h$, the primal-dual formulation proposed and analyzed in [39].

Remark 3.4. Equation $(3.3)_1$ may be thought of as derived from the following integral form:

$$\int_K (\boldsymbol{\sigma}^K - \nabla u^K) \cdot \boldsymbol{q}^K \, dx + \int_{\partial K} (u_{\partial K} - \lambda_{\partial K}) \eta_{\partial K} \, ds = 0 \quad \forall \, \boldsymbol{q}, \eta, \tag{3.4}$$

where $\boldsymbol{q}$ and η are smooth enough test functions. Choosing $\eta_{\partial K} = \boldsymbol{q}^K \cdot \boldsymbol{n}_K|_{\partial K}$ and integrating by parts yields

$$\int_K (\boldsymbol{\sigma}^K \cdot \boldsymbol{q}^K + u^K \operatorname{div} \boldsymbol{q}^K) \, dx - \int_{\partial K} \lambda_{\partial K} \, \boldsymbol{q}_{\partial K} \cdot \boldsymbol{n}_K \, ds = 0 \qquad \forall \boldsymbol{q}. \tag{3.5}$$

Similarly, $(3.3)_2$ may be thought of as derived from the following integral form:

$$\int_K (\operatorname{div} \boldsymbol{\sigma}^K + f^K) v^K \, dx + \int_{\partial K} (\boldsymbol{\sigma}_{\partial K} \cdot \boldsymbol{n}_K - \mu_{\partial K}) \, \xi_{\partial K} \, ds = 0 \quad \forall \, v, \xi, \tag{3.6}$$

where v and ξ are smooth enough test functions. Choosing $\xi_{\partial K} = v_{\partial K}$ and integrating by parts yields

$$\int_K \boldsymbol{\sigma}^K \cdot \nabla \, v^K \, dx - \int_{\partial K} \mu_{\partial K} \, v_{\partial K} \, ds = 0 \qquad \forall v. \tag{3.7}$$

The DPG weak formulation can then be formally interpreted as a mixed-hybrid virtual work principle where *nonvanishing* virtual variations $\delta u = v$ and $\delta(\boldsymbol{\sigma} \cdot \boldsymbol{n}) = \boldsymbol{q} \cdot \boldsymbol{n}$ are allowed on ∂K_D and ∂K_N, respectively (see, e.g., [4] for an extensive discussion of this topic).

3.3. Existence and uniqueness of the DPG solution. In this section we prove the existence and uniqueness of the solution of problem (3.3). To do so, we make use of the generalized saddle point problem theory introduced in [33] and further developed in [8]. For ease of presentation, we assume henceforth that $\Gamma \equiv \Gamma_D$ with $g_D = 0$.

Lemma 3.1 (existence). *Assume that the solution* $\overline{u}$ *of problem* (2.1) *is such that* $\overline{\boldsymbol{\sigma}} \in (L^q(\Omega))^2 \cap H(\mathrm{div};\Omega)$, *where* $\overline{\boldsymbol{\sigma}} = \nabla \overline{u}$. *Then, we have that* $\widetilde{u} = (\overline{u}, \overline{u}|_{\mathcal{E}_h})$ *and* $\widetilde{\boldsymbol{\sigma}} = (\overline{\boldsymbol{\sigma}}, (\overline{\boldsymbol{\sigma}} \cdot \boldsymbol{n})|_{\mathcal{E}_h})$ *is a solution of problem* (3.3).

Proof. Taking $v \in H_0^1(\Omega)$ in $(3.3)_2$ yields (see [14, Chap. 3, Prop. 1.1])

$$\sum_{K \in \mathcal{T}_h} \int_{\partial K} \mu v \, ds = 0 \qquad \forall v \in H_0^1(\Omega).$$

Then $(3.3)_2$ becomes

$$\int_\Omega \boldsymbol{\sigma} \cdot \nabla v \, dx = \int_\Omega f \, v \, dx \qquad \forall v \in H_0^1(\Omega),$$

from which it follows that $\overline{\boldsymbol{\sigma}} = \nabla \overline{u}$ is a solution. Similarly, taking $\boldsymbol{q} \in W_q(\mathrm{div};\Omega)$ in $(3.3)_1$ yields

$$\sum_{K \in \mathcal{T}_h} \int_{\partial K} \lambda \, \boldsymbol{q} \cdot \boldsymbol{n} \, ds = 0 \qquad \forall \boldsymbol{q} \in W_q(\mathrm{div};\Omega).$$

Then $(3.3)_1$ becomes

$$\int_\Omega \nabla\overline{u}\cdot\boldsymbol{q}\,dx + \int_\Omega u\,\mathrm{div}\,\boldsymbol{q}\;dx\; = 0 \qquad \forall\boldsymbol{q}\in W_q(\mathrm{div};\Omega),$$

for which $\overline{u}$ is a solution.

Let us now go back to the hybrid fields; integrating by parts the first term in the bilinear form $b_2(\cdot,\cdot)$, we obtain

$$\sum_{K\in\mathcal{T}_h}\left(\int_{\partial K} v\,\overline{\boldsymbol{\sigma}}\cdot\boldsymbol{n}\,ds - \int_{\partial K}\mu\,v\,ds\right) = 0 \qquad \forall v\in V,$$

which shows that $\mu|_{\partial K} = \overline{\boldsymbol{\sigma}}\cdot\boldsymbol{n}|_{\partial K}$ is a solution (recall indeed that $\overline{\boldsymbol{\sigma}}\in H(\mathrm{div};\Omega)$ implies $\overline{\boldsymbol{\sigma}}^K\cdot\boldsymbol{n}_K + \overline{\boldsymbol{\sigma}}^{K'}\cdot\boldsymbol{n}_{K'} = 0 \quad \forall e_{K-K'}$). Proceeding similarly with the bilinear form $b_1(\cdot,\cdot)$, we obtain

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}(\overline{u}-\lambda)\boldsymbol{q}\cdot\boldsymbol{n}\,ds = 0 \qquad \forall\boldsymbol{q}\in W,$$

which shows that $\lambda|_{\partial K} = \overline{u}|_{\partial K}$ is a solution. The consistency of the continuous DPG formulation with the original problem is thus proved. □

Before dealing with the issue of the uniqueness of the solution of problem (3.3), we state the following useful property that is an extension of the Helmholtz decomposition principle to the present functional setting (cf. [14, Chap. VII, Prop. 3.4 and Remark 3.3] and [25]).

PROPOSITION 3.2. *Every function $\boldsymbol{w}\in(L^q(\Omega))^2$ admits the orthogonal decomposition*

$$\boldsymbol{w} = \nabla\xi\oplus\mathrm{curl}\,\phi,$$

where $\xi\in W_0^{1,q}(\Omega), \phi\in W^{1,q}(\Omega)\setminus\mathbb{R}$, and $\mathrm{curl}\,\phi = (\frac{\partial\phi}{\partial x_1}, -\frac{\partial\phi}{\partial x_2})^T$.

Using the above proposition, we can characterize the null spaces $\mathcal{K}_1$ and $\mathcal{K}_2$ associated with the bilinear forms $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$ as follows:

$$\mathcal{K}_1 = \{\boldsymbol{q}\in W\,|\,b_1(\widetilde{u},\boldsymbol{q}) = 0\,\forall\widetilde{u}\in X\} = \{\boldsymbol{q}\in W_q(\mathrm{div};\Omega)\,|\,\mathrm{div}\,\boldsymbol{q} = 0 \text{ in }\Omega\},$$

$$\mathcal{K}_2 = \{\widetilde{\boldsymbol{\sigma}}\in Y\,|\,b_2(\widetilde{\boldsymbol{\sigma}},v) = 0\,\forall v\in V\} = \{\boldsymbol{\sigma}\in\mathcal{K}_1;\,\mu = \boldsymbol{\sigma}\cdot\boldsymbol{n}\text{ on }\partial K\;\forall K\in\mathcal{T}_h\}.$$

The continuous bilinear forms $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$ induce the following orthogonal decompositions in terms of the closed subspaces $\mathcal{K}_1$ and $\mathcal{K}_2$, respectively:

$$W = \mathcal{K}_1\oplus\mathcal{W}_1 \quad\text{and}\quad Y = \mathcal{K}_2\oplus\mathcal{W}_2,$$

where $\mathcal{W}_1 = \mathcal{K}_1^\perp$ and $\mathcal{W}_2 = \mathcal{K}_2^\perp$.

In order to prove the uniqueness of the solution of the DPG weak formulation, let us check the weak coerciveness of $a(\cdot,\cdot)$ and the inf-sup condition for $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$.

PROPOSITION 3.3. *There exists a constant $\delta > 0$ such that*

$$\sup_{\boldsymbol{q}\in\mathcal{K}_1} a(\widetilde{\boldsymbol{\sigma}},\boldsymbol{q}) \geq \delta||\boldsymbol{\sigma}||_{0,\Omega}\,||\boldsymbol{q}||_{0,\Omega} \quad \forall\widetilde{\boldsymbol{\sigma}}\in\mathcal{K}_2, \tag{3.8}$$

$$\sup_{\widetilde{\boldsymbol{\sigma}}\in\mathcal{K}_2} a(\widetilde{\boldsymbol{\sigma}},\boldsymbol{q}) > 0 \quad \forall \boldsymbol{q}\in\mathcal{K}_1, \boldsymbol{q}\neq 0. \tag{3.9}$$

Proof. Let $\widetilde{\boldsymbol{\sigma}} \in \mathcal{K}_2$ and take $\boldsymbol{q}^* \in \mathcal{K}_1$. Condition (3.8) is immediately verified with $\delta = 1$ by taking $\boldsymbol{q}^\star = \widetilde{\boldsymbol{\sigma}}$ since $a(\widetilde{\boldsymbol{\sigma}},\boldsymbol{q}^*) = ||\boldsymbol{\sigma}||_{0,\Omega}\,||\boldsymbol{q}^*||_{0,\Omega}\ \forall\,\widetilde{\boldsymbol{\sigma}} \in \mathcal{K}_2$. Now let $\boldsymbol{q} \in \mathcal{K}_1$, $\boldsymbol{q} \neq 0$, and $\widetilde{\boldsymbol{\sigma}}^* \in \mathcal{K}_2$. Taking $\widetilde{\boldsymbol{\sigma}}^* = \boldsymbol{q}$, condition (3.9) is immediately verified. □

Proposition 3.4. *There exists a constant* $\gamma_1 > 0$ *such that*

$$\sup_{\boldsymbol{q}\in W} b_1(\widetilde{u},\boldsymbol{q}) \geq \gamma_1||u||_{0,\Omega}||\boldsymbol{q}||_{0,\Omega} \quad \forall \widetilde{u}\in X. \tag{3.10}$$

Proof. Let $\widetilde{u} \in X$ and $\boldsymbol{q}^* = \nabla w$, where w is the solution of the Dirichlet problem

$$\triangle w = u \quad \text{in } \Omega, \qquad w = 0 \quad \text{on } \Gamma.$$

Since $u \in L^q(\Omega)$, we have $\boldsymbol{q}^* \in W_q(\mathrm{div};\Omega)$, and there exists a constant C such that $||\boldsymbol{q}^*||_{0,\Omega} \leq C||u||_{0,\Omega}$. Moreover, it is easy to verify that $b_1(\widetilde{u},\boldsymbol{q}^*) = ||u||^2_{0,\Omega}$ from which (3.10) immediately follows. □

Proposition 3.5. *There exists a constant* $\gamma_2 > 0$ *such that*

$$\sup_{\widetilde{\boldsymbol{\sigma}}\in X} b_2(\widetilde{\boldsymbol{\sigma}},v) \geq \gamma_2||\boldsymbol{\sigma}||_{0,\Omega}||v||_{0,\Omega} \quad \forall v\in V. \tag{3.11}$$

Proof. Let $v \in V$ and $\boldsymbol{\sigma}^* \in W_q(\mathrm{div};\Omega)$, $\mathrm{div}\boldsymbol{\sigma}^* \neq 0$ and $\mu^* = \boldsymbol{\sigma}^*\cdot\boldsymbol{n}$ on ∂K, $\forall K \in \mathcal{T}_h$. Then, after integrating by parts over each element $K \in \mathcal{T}_h$, we have

$$b_2((\boldsymbol{\sigma}^*;\mu^*),v) = -\sum_{K\in\mathcal{T}_h}\int_K v\,\mathrm{div}\boldsymbol{\sigma}^*\,dx,$$

from which we prove (3.11) following the same proof as in Proposition 3.4. □

The following theorem is an immediate consequence of the previous results.

Theorem 3.6. *Under the regularity assumptions stated in Lemma* 3.1*, problem* (3.3) *admits a unique solution* $(\widetilde{u},\widetilde{\boldsymbol{\sigma}}) \in (X \times Y)$.

Having proved Propositions 3.3, 3.4, and 3.5, we can state the following stability result.

Corollary 3.7. *The solution* $(\widetilde{\boldsymbol{\sigma}},\widetilde{u})$ *of the DPG problem* (3.3) *satisfies the estimate*

$$|||\widetilde{u}|||_X \leq K_1||f||_{0,\Omega}, \qquad |||\widetilde{\boldsymbol{\sigma}}|||_Y \leq K_2||f||_{0,\Omega},$$

where $|||\widetilde{u}|||^2_X = ||u||^2_{0,\Omega} + ||\lambda||^2_\Lambda$, $|||\widetilde{\boldsymbol{\sigma}}|||^2_Y = ||\sigma||^2_{0,\Omega} + ||\mu||^2_M$, *and where* $K_1 = (c_1^2 + c_2^2)^{1/2}$, $K_2 = (c_3^2 + c_4^2)^{1/2}$, *with*

$$c_1 = \frac{\delta(2+\gamma_2)+2}{\gamma_2\delta}, \quad c_2 = \frac{2}{\delta}, \quad c_3 = \frac{\delta(2+\gamma_2)+2}{\gamma_2\delta}, \quad c_4 = \frac{2}{\delta}.$$

4. The DPG finite element approximation. Given the finite-dimensional spaces

$$X_h \subset X, \quad Y_h \subset Y, \qquad \text{and} \qquad W_h \subset W, \quad V_h \subset V,$$

the DPG finite element approximation of problem (2.1) reads:

find $(\widetilde{u}_h, \widetilde{\boldsymbol{\sigma}}_h) \in \big(X_h \times Y_h\big)$ such that

$$\left\{ \begin{array}{llll} a(\widetilde{\boldsymbol{\sigma}}_h, \boldsymbol{q}_h) \quad + \quad b_1(\widetilde{u}_h, \boldsymbol{q}_h) & = & 0 & \forall \boldsymbol{q}_h \in W_h, \\ b_2(\widetilde{\boldsymbol{\sigma}}_h, v_h) & = & (f, v_h) & \forall v_h \in V_h. \end{array} \right. \tag{4.1}$$

We have to define the spaces X_h, V_h, W_h, V_h and specify their degrees of freedom.

The choice of the finite element spaces is absolutely nontrivial in mixed Petrov–Galerkin formulations. The idea is first to lay down the properties we want the trial finite element spaces to satisfy and then to select accordingly the discrete test finite element spaces in order to end up with a stable and convergent approximate scheme.

4.1. Trial finite element spaces. The objectives we want the discrete approximation to achieve are the highest possible level of discontinuity and an equal-order interpolation for u_h and $\boldsymbol{\sigma}_h$ and for λ_h and μ_h, respectively. The motivation for adopting equal-order interpolation for both mixed and hybrid variables is that by doing so the numerical performance of a scheme may be significantly enhanced. As a matter of fact, mixed formulations can be interpreted as a *phase-space* approach. Established approaches in dynamics problems applications suggest that an equal-order treatment of the two fields is the right key to achieving correct energy conservation (see [9]).

A natural choice for both internal and interface unknown fields is to consider on each triangle K polynomial finite elements of equal order, respectively, in K and on each edge of ∂K. Henceforth we let k be a nonnegative integer and we denote by $\mathbb{P}_k(K)$ the space of all polynomials of degree $\leq k$ on K and by $R_k(\partial K)$ the space of all functions defined over the boundary ∂K of K whose restrictions to any side $e \in \partial K$ are polynomials of degree $\leq k$. Notice that functions in $R_k(\partial K)$ need not be continuous at the vertices of K.

We take on each triangle $K \in \mathcal{T}_h$

$$X_h^k(K) = \mathbb{P}_k(K) \times R_k(\partial K), \qquad Y_h^k(K) = (\mathbb{P}_k(K))^2 \times R_k(\partial K), \tag{4.2}$$

and we set

$$X_h^k = \prod_{K \in \mathcal{T}_h} X_h^k(K), \qquad Y_h^k = \prod_{K \in \mathcal{T}_h} Y_h^k(K), \tag{4.3}$$

where functions belonging to $R_k(\partial K)$ are single-valued on each internal edge and satisfy the appropriate boundary conditions on Γ_D and Γ_N, respectively. For brevity of notation we also set $\mathcal{X}_h^k(K) = X_h^k(K) \times Y_h^k(K)$ and $\mathcal{X}_h^k = X_h^k \times Y_h^k$.

4.2. Test finite element spaces. Let us now address the issue of properly choosing the finite element test spaces for the DPG approximation. We will start by setting up necessary conditions for the dimension of the test finite element spaces in order for the linear system arising from (4.1) to be a square one. The stability of the approximation will provide a sufficient criterion for explicitly selecting the discrete test functions.

We start with performing a count of the total degrees of freedom corresponding to the choice (4.2) as a function of the polynomial degree k. Subtracting the total number of constraints enforced by the definition of the hybrid field finite element spaces from the previously obtained amount provides the total number of equations that must be written to end up with a *square* algebraic linear system for each value of k. Denoting by NE, Ned, Ni, and Nb the number of triangles, edges, internal edges, and boundary edges, respectively, we have

$$\dim(\mathcal{X}_h^k) = \frac{3}{2}(k+1)(k+6)\texttt{NE}, \tag{4.4}$$

while the total number of constraints is $\texttt{Nc} = (\text{k+1})(2\ \texttt{Ni} + \texttt{Nb})$.

Applying Euler's theorem ($\texttt{Ned} = (3\ \texttt{NE} + \texttt{Nb})/2$), we can express the total number of constraints as a function of NE as $\texttt{Nc} = 3(\text{k+1})\texttt{NE}$, from which it follows that the dimension of the global finite element test space

$$\mathcal{V}_h^k(K) = W_h^k(K) \times V_h^k(K), \qquad \mathcal{V}_h^k = W_h^k \times V_h^k$$

that is needed to end up with a square linear system for each value of k is

$$\dim(\mathcal{V}_h^k) = \dim(\mathcal{X}_h^k) - \texttt{Nc} = \frac{3}{2}(k+1)(k+4)\texttt{NE}. \tag{4.5}$$

Looking at (4.4) and (4.5) it clearly appears that for each k, the degrees of freedom for both trial and test spaces as well as the total number of constraints can all be expressed as a function of the sole number of mesh triangles NE. Therefore, the proper design of the finite element test function spaces can be carried out at the single element level. Precisely, denoting by $\texttt{Nc}(K)$ the number of constraints on triangle K, relation (4.5) can be written at the element level as

$$\dim(\mathcal{V}_h^k(K)) = \dim(\mathcal{X}_h^k(K)) - \texttt{Nc}(K) = \frac{3}{2}(k+1)(k+4) \quad \forall K \in \mathcal{T}_h. \tag{4.6}$$

This equation expresses the balance between degrees of freedom, constraints, and number of equations that must be fulfilled independently on each single element K. Based on these constraints, we start in the next section with the construction of the finite element test space $\mathcal{X}_h^k(K)$ in the lowest degree case $k = 0$. The resulting local finite element space will be denoted as $\text{DPG}_0(K) = \mathcal{X}_h^0(K) \times \mathcal{V}_h^0(K)\ \forall K \in \mathcal{T}_h$.

5. Choice of the finite element spaces.

In this section we discuss in detail the lowest order finite element approximation DPG_0 and then we use this procedure as a guideline for the generation of higher order elements.

5.1. DPG_0 finite element approximation.

Setting $k = 0$, relation (4.6) gives

$$\dim(\mathcal{V}_h^k(K)) = \dim(W_h^0(K)) + \dim(V_h^0(K)) = 6 \qquad \forall K \in \mathcal{T}_h.$$

The minimal choice for the scalar finite element test space is $V_h^0(K) = \mathbb{P}_1(K)$. By doing so, 3 degrees of freedom are left for the vector finite element test space, which can be conveniently saturated by setting $W_h^0(K) = \mathbb{RT}_0(K)$, where $\mathbb{RT}_k(K) = (\mathbb{P}_k(K))^2 \oplus \boldsymbol{x}\mathbb{P}_k(K)\ \forall K \in \mathcal{T}_h$ is the Raviart–Thomas finite element space of degree k [36]. The DPG_0 local finite element space is then defined on each element as

$$\text{DPG}_0(K) = \underbrace{\left(\mathbb{P}_0(K) \times R_0(\partial K) \times (\mathbb{P}_0(K))^2 \times R_0(\partial K)\right)}_{\mathcal{X}_h^0(K)} \times \underbrace{\left(\mathbb{RT}_0(K) \times \mathbb{P}_1(K)\right)}_{\mathcal{V}_h^0(K)}. \tag{5.1}$$

5.2. Higher order DPG methods. Higher order finite elements will be consistently denoted as $\mathrm{DPG}_k(K) = \mathcal{X}_h^k(K) \times \mathcal{V}_h^k(K)\ \forall K \in \mathcal{T}_h$. Under the assumption that the local finite element trial space is defined as in (4.2)–(4.3), the question is how to construct a suitable test finite element space such that the following conditions are satisfied:

1. the dimension of the test finite element space is

$$\dim(\mathcal{V}_h^k(K)) = \dim(W_h^k(K)) + \dim(V_h^k(K)) = \frac{3}{2}(k+1)(k+4) \quad \forall K \in \mathcal{T}_h; \tag{5.2}$$

2. the following inf-sup condition is verified:

$$\mu_h \in R_k(\partial K), \quad \int_{\partial K} \mu_h\, v_h\, ds = 0 \qquad \forall v \in V_h^k(K) \quad \text{implies} \quad \mu_h = 0. \tag{5.3}$$

The first condition ensures that we end up with a square system. The second condition forces a restriction on the minimum order of the polynomial space for v_h. There is no need of such a condition for the term $\int_{\partial K} \lambda_h\, \boldsymbol{q}_h \cdot \boldsymbol{n}\, ds$ since both λ_h and the normal traces of functions in $W_h^k(K)$ are *discontinuous* on ∂K.

In the construction procedure, we must distinguish between the case when k is an even or an odd integer. Indeed, using the results stated by Lemmas 4 and 6 in [35] (where the same compatibility problem occurs), we have that condition (5.3) holds if

$$\begin{cases} v_h \in \mathbb{P}_{k+1}(K) & \text{for } k \text{ even, } k \geq 0, \\ v_h \in \widehat{\mathbb{P}}(K),\ \mathbb{P}_{k+1}(K) \subset \widehat{\mathbb{P}}(K) \subset \mathbb{P}_{k+2}(K) & \text{for } k \text{ odd, } k \geq 1. \end{cases} \tag{5.4}$$

In the first case (k even), the family of finite element spaces is immediately built by setting $V_h^k(K) = \mathbb{P}_{k+1}(K)$ and then suitably saturating the degrees of freedom implied by (5.2)

$$\dim(W_h^k(K)) = \dim(\mathcal{V}_h^k(K)) - \dim(V_h^k(K)) = k^2 + 5k + 3 \qquad \forall K \in \mathcal{T}_h \tag{5.5}$$

by choosing $W_h^k(K) = \mathbb{BDFM}_{k+1}(K)$, $k \geq 0$ (for the definition of this space and its properties, see [13, 14]).

The situation is more complicated when k is odd. In this case the choice $v_h \in \mathbb{P}_{k+1}(K)$ is not allowed by (5.4), while the choice $v_h \in \mathbb{P}_{k+2}(K)$ is acceptable but unnecessarily expensive. In [35] it has been shown that in order to satisfy condition (5.4) it is sufficient to enrich the space $\mathbb{P}_{k+1}(K)$ with a single additional degree of freedom suitably excerpted from the space $\mathbb{P}_{k+2}(K)$. Setting thus $V_h^k(K) = \widehat{\mathbb{P}}(K)$, relation (5.2) becomes for each $K \in \mathcal{T}_h$

$$\dim(W_h^k(K)) = \dim(\mathcal{V}_h^k(K)) - \dim(V_h^k(K)) = k^2 + 5k + 2. \tag{5.6}$$

A possible choice is then

$$W_h^k(K) = \mathbb{RT}_k(K) \oplus B_{k-1}(K) \oplus B_k(K) \oplus \cdots \oplus B_{2k-3}(K), \qquad k \geq 1,$$

where $(k+1)(k+3)$ degrees of freedom are saturated by the $\mathbb{RT}_k$ space and the remaining $(k-1)$ degrees of freedom are saturated by adding $(k-1)$ bubble functions B_l defined as (see [38]) $B_l(K) = \{\boldsymbol{q} \mid \boldsymbol{q} = \mathrm{curl}\,(b_K w)\}$, with $w \in \mathbb{P}_l(K)$, $b_K = \prod_{i=1}^{3} z_i(\boldsymbol{x})$, z_i, $i = 1, 2, 3$, being the barycentric coordinates in K.

To summarize, the family of DPG$_k$ finite element spaces is defined as

$$X_h^k(K) = \mathbb{P}_k(K) \times R_k(\partial K), \qquad Y_h^k(K) = (\mathbb{P}_k(K))^2 \times R_k(\partial K), \quad k = 0, 1, 2, \dots,$$

and letting $m = 0, 1, 2, \dots,$ we have

$$\begin{cases} W_h^{2m}(K) = \mathbb{BDFM}_{2m+1}(K), \quad V_h^{2m}(K) = \mathbb{P}_{2m+1}(K), \qquad k = 2m, \\ W_h^{2m+1}(K) = \mathbb{RT}_{2m+1}(K) \oplus B_{2m}(K) \cdots \oplus B_{4m-1}(K),\ V_h^{2m+1}(K) = \widehat{\mathbb{P}}(K), \\ \qquad k = 2m+1. \end{cases}$$

Notice that with the above choices the matrix arising from the term $\int_{\partial K} \lambda_h \boldsymbol{q}_h \cdot \boldsymbol{n}\, ds$ is always square and nonsingular since both $\lambda_h|_{\partial K}$ and $\boldsymbol{q}_h \cdot \boldsymbol{n}|_{\partial K}$ belong to the same finite element space due to the properties of the $\mathbb{RT}_k$ and $\mathbb{BDFM}_{k+1}$ spaces (see [14, Chap. 3]).

6. Stability and convergence analysis of the approximate DPG$_0$ solution.

In this section and in the remainder of the article, we focus our attention on the member of the DPG$_k$ family of lowest degree, the DPG$_0$ finite element. Numerical results on the convergence performance of higher order elements of the family can be found in [10].

6.1. Existence and uniqueness of the DPG$_0$ solution.

We start proving the uniqueness (and thus the existence) of the solution of problem (4.1). To do so, we characterize the discrete null spaces $\mathcal{K}_1^h$ and $\mathcal{K}_2^h$ as

$$\begin{aligned} \mathcal{K}_1^h &= \{\boldsymbol{q}_h \in W_h \,|\, b_1(\widetilde{u}_h, \boldsymbol{q}_h) = 0 \,\forall \widetilde{u}_h \in X_h\} \\ &= \{\boldsymbol{q}_h \in W_h \,|\, \boldsymbol{q}_K \cdot \boldsymbol{n}_K + \boldsymbol{q}_{K'} \cdot \boldsymbol{n}_{K'} = 0 \ \forall e_{K-K'},\ \operatorname{div} \boldsymbol{q}_h = 0 \text{ in } \Omega\}, \\ \mathcal{K}_2^h &= \{\widetilde{\boldsymbol{\sigma}}_h \in Y_h \,|\, b_2(\widetilde{\boldsymbol{\sigma}}_h, v_h) = 0 \,\forall v \in V_h\} \\ &= \{\boldsymbol{\sigma}_h \in \Sigma \,|\, \boldsymbol{\sigma}_K \cdot \boldsymbol{n}_K + \boldsymbol{\sigma}_{K'} \cdot \boldsymbol{n}_{K'} = 0 \ \forall e_{K-K'};\ \mu_h = \boldsymbol{\sigma}_h \cdot \boldsymbol{n} \text{ on } \partial K \ \forall K \in \mathcal{T}_h\}. \end{aligned}$$

The continuous bilinear forms $b_1(\cdot, \cdot)$ and $b_2(\cdot, \cdot)$ induce the following orthogonal decompositions in terms of the closed subspaces $\mathcal{K}_1^h$ and $\mathcal{K}_2^h$, respectively:

$$W_h = \mathcal{K}_1^h \oplus \mathcal{W}_1^h \quad \text{and} \quad Y_h = \mathcal{K}_2^h \oplus \mathcal{W}_2^h, \quad \text{where } \mathcal{W}_1^h = \mathcal{K}_1^{h,\perp},\ \mathcal{W}_2^h = \mathcal{K}_2^{h,\perp}.$$

Moreover, the following properties hold.

PROPOSITION 6.1. *There exists a constant $\delta' > 0$ independent of h such that*

$$\begin{aligned} &\sup\nolimits_{\boldsymbol{q}_h \in \mathcal{K}_1^h} a(\widetilde{\boldsymbol{\sigma}}_h, \boldsymbol{q}_h) \geq \delta' ||\boldsymbol{\sigma}_h||_{0,\Omega}\, ||\boldsymbol{q}_h||_{0,\Omega} \quad \forall \widetilde{\boldsymbol{\sigma}}_h \in \mathcal{K}_2^h, \\ &\sup\nolimits_{\boldsymbol{\sigma}_h \in \mathcal{K}_2^h} a(\widetilde{\boldsymbol{\sigma}}_h, \boldsymbol{q}_h) > 0 \quad \forall \boldsymbol{q}_h \in \mathcal{K}_1^h, \boldsymbol{q}_h \neq 0. \end{aligned} \tag{6.1}$$

Proof. Let $\widetilde{\boldsymbol{\sigma}}_h \in \mathcal{K}_2^h$ and take $\boldsymbol{q}_h^* \in \mathcal{K}_1^h$. Condition $(6.1)_1$ is immediately verified with $\delta' = 1$ by taking $\boldsymbol{q}_h^* = \widetilde{\boldsymbol{\sigma}}_h$ since $a(\widetilde{\boldsymbol{\sigma}}_h, \boldsymbol{q}_h^*) = ||\boldsymbol{\sigma}_h||_{0,\Omega}\, ||\boldsymbol{q}_h^*||_{0,\Omega}\ \forall\, \widetilde{\boldsymbol{\sigma}}_h \in \mathcal{K}_2^h$. Now let $\boldsymbol{q}_h \in \mathcal{K}_1^h$, $\boldsymbol{q}_h \neq 0$ and $\widetilde{\boldsymbol{\sigma}}_h^* \in \mathcal{K}_2^h$. Taking $\widetilde{\boldsymbol{\sigma}}_h^* = \boldsymbol{q}_h$, condition $(6.1)_2$ is immediately verified. □

PROPOSITION 6.2. *There exists a constant $\gamma_1' > 0$ independent of h such that*

$$\sup_{\boldsymbol{q}_h \in W_h} b_1(\widetilde{u}_h, \boldsymbol{q}_h) \geq \gamma_1' ||u_h||_{0,\Omega} ||\boldsymbol{q}_h||_{0,\Omega} \quad \forall \widetilde{u}_h \in X_h. \tag{6.2}$$

Proof. Let $\widetilde{u}_h \in X_h$ and take $\boldsymbol{q}^* \in W_h$ such that $\boldsymbol{q}^*_K \cdot \boldsymbol{n}_K + \boldsymbol{q}^*_{K'} \cdot \boldsymbol{n}_{K'} = 0$ on $e_{K-K'}$ $\forall K, K' \in \mathcal{T}_h$. Then using Lemma 7.2.1 of [34] and noting that $||\boldsymbol{q}_h||_{H(\text{div};\Omega)} \geq ||\boldsymbol{q}_h||_{0,\Omega}$, condition (6.2) immediately follows. □

To prove the discrete inf-sup condition for the bilinear form $b_2(\cdot,\cdot)$, we need to introduce the space $\widetilde{V} = \prod_{K\in\mathcal{T}_h}(H^1(K) \setminus \mathbb{R})$, equipped with the norm $|||v|||_{\widetilde{V}} = (\sum_{K\in\mathcal{T}_h} |v|^2_{1,K})^{1/2}$. Notice that this norm is indeed a norm on the space $\widetilde{V}$ and is equivalent to the norm $||\cdot||_V$ for functions $v \in \widetilde{V}$. We also consider a finite-dimensional approximation of $\widetilde{V}$, i.e., the space $\widetilde{V}_h \subset \widetilde{V}$ defined as $\widetilde{V}_h = \prod_{K\in\mathcal{T}_h} \mathbb{P}_1(K) \setminus \mathbb{R}$.

PROPOSITION 6.3. *There exists a constant $\gamma'_2 > 0$ independent of h such that*

$$\sup_{\widetilde{\boldsymbol{\sigma}}_h \in Y_h} b_2(\widetilde{\boldsymbol{\sigma}}_h, v_h) \geq \gamma'_2 ||\boldsymbol{\sigma}_h||_{0,\Omega} |||v_h|||_{\widetilde{V}} \quad \forall v_h \in \widetilde{V}_h. \tag{6.3}$$

Proof. Let $v_h \in \widetilde{V}_h$. Take $\boldsymbol{\sigma}^*_h \in \Sigma_h$ such that $\boldsymbol{\sigma}^*_h = \nabla v_h$ on K $\forall v_h \in \widetilde{V}_h$, $\forall K \in \mathcal{T}_h$ and set $\mu^*_h \equiv 0$. Then we have $b_2(\widetilde{\boldsymbol{\sigma}}^*_h, v_h) = ||\boldsymbol{\sigma}^*_h||_{0,\Omega} |||v_h|||_{\widetilde{V}}$, and (6.3) immediately follows with $\gamma'_2 = 1$. □

Remark 6.1. The above proof reveals that the choice $v_h =$ constant on each $K \in \mathcal{T}_h$ by itself does not allow us to state condition (6.3). However, taking $v_h = 1$ on K and equal to zero elsewhere is a possible and significant choice since it provides the local conservation property of the DPG formulation $-\int_{\partial K} \mu_h \, ds = \int_K f \, dx$. Moreover, a global conservation property can be shown to hold as well by taking $v_h \equiv 1$ on Ω, yielding the relation $-\int_\Gamma \mu_h \, ds = \int_\Omega f \, dx$. A detailed discussion of this subject will be carried out in section 9.

Remark 6.2. Choosing $\boldsymbol{\sigma}^*_h = 0$, the bilinear form $b_2(\cdot,\cdot)$ yields the familiar relation of primal hybrid formulations $\sum_{K\in\mathcal{T}_h} \int_K \mu_h v_h = 0$ $\forall v_h \in V_h$, which admits the unique solution $\mu_h \equiv 0$.

The following theorem is an immediate consequence of the previous results.

THEOREM 6.4. *The DPG$_0$ approximation of problem* (2.2) *admits a unique solution* $(\widetilde{u}_h, \widetilde{\boldsymbol{\sigma}}_h) \in (X_h \times Y_h)$.

6.2. Error estimates.

In the following sections we establish optimal error estimates for the mixed variables u_h and $\boldsymbol{\sigma}_h$ and for the hybrid variables λ_h and μ_h.

6.2.1. Projection operators.

In view of the error analysis of the DPG formulation, it is useful to introduce some approximation operators. We denote by P_K the projection operator from $L^2(K)$ onto $\mathbb{P}_0(K)$ satisfying the approximation property

$$||v - P_K v||_{0,K} \leq Ch|v|_{1,K} \qquad \forall v \in H^1(K). \tag{6.4}$$

From the operator P_K, for all $v \in L^2(\Omega)$, we construct the global operator P_h as

$$P_h v|_K = P_K v \qquad \forall K \in \mathcal{T}_h. \tag{6.5}$$

We also need to introduce the projection operator ρ^0_h from $\prod_{K\in\mathcal{T}_h} L^2(\partial K)$ onto $R_0(\partial K)$ such that, for all $\lambda \in \prod_{K\in\mathcal{T}_h} L^2(\partial K)$, we have

$$\int_{\partial K} (\rho^0_h \lambda - \lambda) r_0 \, ds = 0 \qquad \forall r_0 \in R_0(\partial K), \qquad \forall K \in \mathcal{T}_h. \tag{6.6}$$

Remark 6.3. The operator ρ^0_h is well defined since, by Sobolev's embedding theorem [27, 28], we have that $W^{1/q,p}(\partial K) \hookrightarrow L^2(\partial K)$.

6.2.2. Error estimates for the mixed variables. The following optimal error estimates hold.

THEOREM 6.5. *Let $(u, \boldsymbol{\sigma})$ be the solution of* (3.3) *and let $(u_h, \boldsymbol{\sigma}_h)$ be the solution of* (4.1). *If $\boldsymbol{\sigma} \in (H^1(\Omega))^2$, then there exists a positive constant C independent of h such that*

$$\begin{aligned} ||u - u_h||_{0,\Omega} &\leq Ch(|u|_{1,\Omega} + |\boldsymbol{\sigma}|_{1,\Omega}), \\ ||\boldsymbol{\sigma} - \boldsymbol{\sigma}_h||_{0,\Omega} &\leq Ch|\boldsymbol{\sigma}|_{1,\Omega}. \end{aligned} \tag{6.7}$$

Proof. From (3.3) and (4.1) we have

$$\left\{ \begin{array}{lclcll} a(\widetilde{\boldsymbol{\sigma}} - \widetilde{\boldsymbol{\sigma}}_h, \boldsymbol{q}_h) & + & b_1(\widetilde{u} - \widetilde{u}_h, \boldsymbol{q}_h) & = & 0 & \forall \boldsymbol{q}_h \in W_h, \\ b_2(\widetilde{\boldsymbol{\sigma}} - \widetilde{\boldsymbol{\sigma}}_h, v_h) & & & = & 0 & \forall v_h \in V_h. \end{array} \right. \tag{6.8}$$

Taking $\boldsymbol{q}_h \in \mathcal{K}_1^h$, the first relation in (6.8) becomes

$$a(\widetilde{\boldsymbol{\sigma}} - \widetilde{\boldsymbol{\sigma}}_h, \boldsymbol{q}_h) = 0 \qquad \forall \boldsymbol{q}_h \in \mathcal{K}_1^h. \tag{6.9}$$

Let us introduce the decomposition $\widetilde{\boldsymbol{\sigma}}_h = (\boldsymbol{\sigma}_h, \mu_h) = \widetilde{\boldsymbol{\sigma}}_h^0 + \widetilde{\boldsymbol{\sigma}}_h^\perp$, where $\widetilde{\boldsymbol{\sigma}}_h^0 = (\boldsymbol{\sigma}_h^0, \mu_h^0) \in \mathcal{K}_2^h$ and $\widetilde{\boldsymbol{\sigma}}_h^\perp = (\boldsymbol{\sigma}_h^\perp, \mu_h^\perp) \in \mathcal{W}_2^h$. Introducing the projection operator $\widetilde{\Pi}_h = ((P_h)^2, \rho_h^0)$, where P_h and ρ_h^0 have been defined in (6.5) and (6.6), respectively, and using the decomposition $\sigma_h = (\boldsymbol{\sigma}_h^0 + \boldsymbol{\sigma}_h^\perp)$, equation (6.9) reads

$$a((\widetilde{\Pi}_h \boldsymbol{\sigma})^0 - \boldsymbol{\sigma}_h^0, \boldsymbol{q}_h) = a(\widetilde{\Pi}_h \boldsymbol{\sigma} - \boldsymbol{\sigma}, \boldsymbol{q}_h) + a(\boldsymbol{\sigma}_h^\perp - (\widetilde{\Pi}_h \boldsymbol{\sigma})^\perp, \boldsymbol{q}_h) \qquad \forall \boldsymbol{q}_h \in \mathcal{K}_1^h,$$

which, using the coercivity and continuity of the bilinear form $a(\cdot,\cdot)$, yields

$$||(\widetilde{\Pi}_h \boldsymbol{\sigma})^0 - \boldsymbol{\sigma}_h^0||_{0,\Omega} \leq (||\widetilde{\Pi}_h \boldsymbol{\sigma} - \boldsymbol{\sigma}||_{0,\Omega} + ||(\widetilde{\Pi}_h \boldsymbol{\sigma})^\perp - \boldsymbol{\sigma}_h^\perp||_{0,\Omega}). \tag{6.10}$$

Now we need to bound the quantity $||(\widetilde{\Pi}_h \boldsymbol{\sigma})^\perp - \boldsymbol{\sigma}_h^\perp||_{0,\Omega}$. Using (6.5) into $(6.8)_2$, we get

(6.11)
$$b_2(\widetilde{\Pi}_h \widetilde{\boldsymbol{\sigma}} - \widetilde{\boldsymbol{\sigma}}_h, v_h) = b_2(\widetilde{\Pi}_h \widetilde{\boldsymbol{\sigma}} - \widetilde{\boldsymbol{\sigma}}, v_h) = -\sum_{K \in \mathcal{T}_h} \int_{\partial K} (\rho_h^0 \mu - \mu) v_h \, ds \qquad \forall v_h \in V_h.$$

Recalling Lemma 9 in [35], we have that

$$\int_{\partial K} (\boldsymbol{\sigma} \cdot \boldsymbol{n} - \rho_h^0 \mu) v \, ds \leq C \frac{h_K}{\rho_K} |\boldsymbol{\sigma}|_{1,K} |v|_{1,K} \qquad \forall v \in H^1(K).$$

Using this latter relation in (6.11) and the discrete inf-sup condition for $b_2(\cdot,\cdot)$, we get the estimate

$$||(\widetilde{\Pi}_h \boldsymbol{\sigma})^\perp - \boldsymbol{\sigma}_h^\perp||_{0,\Omega} \leq Ch|\boldsymbol{\sigma}|_{1,\Omega}. \tag{6.12}$$

Now, gathering (6.10) and (6.12) and using the triangle inequality, we get $(6.7)_2$.

Let us now prove $(6.7)_1$. Taking $\boldsymbol{q}_h \in \mathcal{W}_1^h$ in the first equation of (6.8) we get

$$b_1(\widetilde{u} - \widetilde{u}_h, \boldsymbol{q}_h) = a(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}, \boldsymbol{q}_h) \qquad \forall \boldsymbol{q}_h \in \mathcal{W}_1^h.$$

Introducing the projection operator $\widetilde{P}_h = (P_h, \rho_h^0)$, we write the latter relation as

$$b_1(\widetilde{P}_h \widetilde{u} - \widetilde{u}_h, \boldsymbol{q}_h) = a(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}, \boldsymbol{q}_h) \qquad \forall \boldsymbol{q}_h \in \mathcal{W}_1^h.$$

Then using the discrete inf-sup condition for $b_1(\cdot,\cdot)$, we get

$$||P_h u - u_h||_{0,\Omega} \leq C||\boldsymbol{\sigma} - \boldsymbol{\sigma}_h||_{0,\Omega} \leq Ch|\boldsymbol{\sigma}|_{1,\Omega}. \tag{6.13}$$

Eventually, using $(6.7)_2$, (6.4), and the triangle inequality, we get $(6.7)_1$. □

6.3. Error estimates for the hybrid variables. In this section, we derive a priori error estimates for the discretization errors associated with the hybrid variables λ_h and μ_h. In doing this, we shall prove some equivalence results between the DPG_0 method and hybrid formulations, both of primal and dual type.

To start with, we let $W_{h,0}^{NC}$ denote the set of nonconforming functions in V_h that are affine on each $K \in \mathcal{T}_h$ and are continuous at the midpoint of each edge and vanish at the midpoint of each edge of Γ. Then, we define $u_h^* \in W_{h,0}^{NC}$ as the piecewise linear nonconforming function such that

$$\int_{\partial K} u_h^* \, \eta_h \, ds = \int_{\partial K} \lambda_h \, \eta_h \, ds \qquad \forall \eta_h \in R_0(\partial K), \qquad \forall K \in \mathcal{T}_h, \tag{6.14}$$

which implies in particular that $u_h^*(x_{MP,i}) = \lambda_{h,e_i}$ for each edge $e_i \in \partial K$, $x_{MP,i}$ being the coordinate vector of the midpoint of e_i. Using the fact that $\boldsymbol{q}_h \cdot \boldsymbol{n}_K \in R_0(\partial K)$ and (6.14), we get

$$\int_{\partial K} \lambda_h \boldsymbol{q}_h \cdot \boldsymbol{n}_K \, ds = \int_{\partial K} u_h^* \boldsymbol{q}_h \cdot \boldsymbol{n}_K \, ds = \int_K u_h^* \mathrm{div} \boldsymbol{q}_h \, dx + \int_K \boldsymbol{q}_h \cdot \nabla u_h^* \, dx \ \ \forall \boldsymbol{q}_h \in W_h(K).$$

Substituting this latter expression in $(4.1)_1$, we obtain

$$\int_K (\boldsymbol{\sigma}_h - \nabla u_h^*) \cdot \boldsymbol{q}_h \, dx + \int_K (u_h - u_h^*) \mathrm{div} \boldsymbol{q}_h \, dx = 0 \qquad \forall \boldsymbol{q}_h \in W_h(K). \tag{6.15}$$

Taking $q_h \in (\mathbb{P}_0(K))^2$ in (6.15) yields

$$\boldsymbol{\sigma}_h^K = \nabla u_h^* \qquad \forall K \in \mathcal{T}_h, \tag{6.16}$$

while taking $q_h = (x, y)^T$ in (6.15) and using (6.14) yields

$$u_h^K = \frac{\int_K u_h^* \, dx}{|K|} = P_K u_h^* = \frac{1}{3} \sum_{i=1}^{3} \lambda_{h,e_i} \qquad \forall K \in \mathcal{T}_h. \tag{6.17}$$

Relation (6.17) shows that u_h^K is the *average value* of u_h^* on K and thus the average value of the hybrid variables λ_h on the edges of the element. Let us now consider $(4.1)_2$ and take $v_h \in W_{h,0}^{NC}$. Equation $(4.1)_2$ becomes

$$\sum_{K \in \mathcal{T}_h} \int_K \nabla u_h^* \cdot \nabla v_h \, dx = \sum_{K \in \mathcal{T}_h} \int_K f v_h \, dx \qquad \forall v_h \in W_{h,0}^{NC}. \tag{6.18}$$

Relation (6.18) shows that u_h^* actually *coincides* with the solution $u_h^{NC} \in W_{h,0}^{NC}$ of problem (2.1), which is obtained with the nonconforming finite element approximation (see [35]).

Then, the following error estimate can be proved.

THEOREM 6.6. *Let u be the solution of problem* (2.1) *such that $u \in H^2(\Omega) \cap H_0^1(\Omega)$, and let u_h^* be the solution of problem* (6.18) *such that* (6.14) *holds. Then, under the assumption that the polygonal domain Ω is convex, we have (see* [35]*)*

$$||u - u_h^*||_{0,\Omega} \leq C h^2 |u|_{2,\Omega}. \tag{6.19}$$

The above theorem is a superconvergence result for the piecewise linear nonconforming extension over Ω of the hybrid variable λ_h computed by the DPG_0 formulation. Moreover, the estimate (6.19) can be regarded as the counterpart for the DPG formulation of Theorem 2.2 in [1] valid for the dual-mixed method with hybridization.

Considering again $(4.1)_2$ and taking this time $v_h \in V_h$, we obtain

$$\int_{\partial K} \mu_h \, v_h \, ds = \int_K \nabla u_h^* \cdot \nabla v_h \, dx - \int_K f v_h \, dx \qquad \forall v_h \in V_h(K), \quad \forall K \in \mathcal{T}_h, \tag{6.20}$$

which coincides with the postprocessing procedure discussed in [37, section 19] for the primal-hybrid formulation. This result actually demonstrates that the hybrid field μ_h computed by the DPG$_0$ approximation coincides with the field $\mathbf{p}_h \cdot \boldsymbol{n}$ computed by the primal-hybrid formulation. We then have the following result.

Theorem 6.7. *Under the assumptions of Theorem* 6.6 *and the condition stated in Remark* 6.2*, we have*

$$||\boldsymbol{\sigma} \cdot \boldsymbol{n} - \mu_h||_{-1/2,h} \leq Ch|\boldsymbol{\sigma}|_{1,\Omega},$$

where $\forall \xi \in R_0(\partial K)$ *we define the norm* $||\xi||_{-1/2,h} = (\sum_{e\in\mathcal{E}_h} |e|||\xi||_{0,e}^2)^{1/2}$ *(see* [1]*).*

Proceeding along the above guideline, it is possible to further explore the connection existing between the DPG$_0$ formulation and the dual-mixed method. In view of establishing this connection, we assume henceforth f to be piecewise constant over $\mathcal{T}_h$. Under this hypothesis, we can use the following result proved in [29]:

$$\begin{aligned} u_h^{DM} - P_K u_h^* &= u_h^{DM} - u_h^K = -\frac{f^K}{4}\left(|x_{CG,K}^2| - \frac{1}{|K|}\int_K |x|^2 \, dx\right) \\ &= \frac{1}{144} f^K \sum_{i=1}^{3} |e_i|^2 = \mathcal{O}(h_K^2) \qquad \forall K \in \mathcal{T}_h, \end{aligned} \tag{6.21}$$

where $x_{CG,K}$ is the coordinate vector of the center of gravity of K and $u_h^{DM} \in \mathbb{P}_0(K)$ is the solution computed by the dual-mixed method. Using the result (6.21) and recalling the standard estimates for the dual-mixed approximation (see [23, 14]) gives by the triangle inequality the following result.

Theorem 6.8. *Let* $(u, \boldsymbol{\sigma})$ *be the solution of* (2.1) *and* $(u_h, \boldsymbol{\sigma}_h)$ *be the solution of* (4.1)*. If the triangulation* $\mathcal{T}_h$ *is uniformly regular and* $\boldsymbol{\sigma} \in (H^1(\Omega))^2$, $\mathrm{div}\boldsymbol{\sigma} \in H^1(\Omega)$, *then*

$$||P_h u - u_h||_{0,\Omega} \leq Ch^2(|\boldsymbol{\sigma}|_{1,\Omega}| + |\mathrm{div}\boldsymbol{\sigma}|_{1,\Omega}). \tag{6.22}$$

Relation (6.22) can be interpreted as a superconvergence result for u_h at the center of gravity of each triangle K. This latter result also allows us to derive an optimal estimate for the quantity $||\rho_h^0 \lambda - \lambda_h||_{1/q,p,\partial K}$. To proceed, we first need to recall the following result [23].

Lemma 6.9. *For all* $T \in W^{-1/q,q}(\partial K)$ *there exists a unique* $\boldsymbol{q}_h \in \mathbb{RT}_0(K)$ *such that* $\forall K \in \mathcal{T}_h$ *we have*

$$\int_{\partial K} (\boldsymbol{q}_h \cdot \boldsymbol{n} - T) \, r_0 \, ds = 0 \qquad \forall r_0 \in R_0(\partial K). \tag{6.23}$$

Furthermore, if $\mathcal{T}_h$ *is uniformly regular, then there is a constant* C *independent of* K *such that*

$$||\boldsymbol{q}_h||_{0,K} \leq Ch^{2/p-1}||T||_{-1/q,q,\partial K}, \quad ||\mathrm{div}\boldsymbol{q}_h||_{0,K} \leq Ch^{2/p-2}||T||_{-1/q,q,\partial K}, \tag{6.24}$$

where the norm $||.||_{-1/q,q,\partial K}$ *has been defined in* (3.1)*.*

For the definition of a uniformly regular triangulation, see [18].

We are now in a position to state the following result.

THEOREM 6.10. *Under the assumptions of Theorem* 6.8*, we have*

$$||\rho_h^0\lambda - \lambda_h||_{1/q,p,\partial K} \le Ch^{2/p}\,(|\boldsymbol{\sigma}|_{1,\Omega} + |\mathrm{div}\boldsymbol{\sigma}|_{1,\Omega}) \qquad \forall K \in \mathcal{T}_h. \tag{6.25}$$

Proof. Let $\boldsymbol{q}$ be any element of $W_q(\mathrm{div};K)$ and let $\overline{\boldsymbol{q}}_h \in \mathbb{RT}_0(K)$ be defined by (6.23) with $T = \boldsymbol{q}\cdot\boldsymbol{n}|_{\partial K}$ and such that $\boldsymbol{q}_h|_K = \overline{\boldsymbol{q}}_h$, $\boldsymbol{q}_h|_{K'} = 0\ \forall K' \neq K$, $\forall K \in \mathcal{T}_h$. Subtracting the first equation of (4.1) from the first equation of (3.3), we get

$$\int_{\partial K} \overline{\boldsymbol{q}}_h \cdot \boldsymbol{n}(\lambda - \lambda_h)\,ds = \int_K (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot\overline{\boldsymbol{q}}_h\,dx + \int_K (u - u_h)\,\mathrm{div}\overline{\boldsymbol{q}}_h\,dx,$$

which can be written as

$$\int_{\partial K} \overline{\boldsymbol{q}}_h \cdot \boldsymbol{n}(\rho_h^0\lambda - \lambda_h)\,ds = \int_K (\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\cdot\overline{\boldsymbol{q}}_h\,dx + \int_K (P_K u - u_h)\,\mathrm{div}\overline{\boldsymbol{q}}_h\,dx.$$

Owing to the definition of $\overline{\boldsymbol{q}}_h$, using (6.24), (6.22), and the definition (3.2), we eventually get the estimate (6.25). □

Since $p < 2$, estimate (6.25) can be regarded as a superconvergence property for λ_h.

To conclude our equivalence analysis, we show that

$$\mu_h^K = \boldsymbol{\sigma}_h^{DM}\cdot\boldsymbol{n}_K \qquad \forall K \in \mathcal{T}_h, \tag{6.26}$$

where $\boldsymbol{\sigma}_h^{DM} \in \mathbb{RT}_0(K)$ is the solution computed by the dual-mixed method. We recall the following result proved in [29]:

$$\boldsymbol{\sigma}_h \equiv \nabla u_h^* = \boldsymbol{\sigma}_h^{DM} + (x - x_{CG})\frac{f^K}{2} \qquad \forall K \in \mathcal{T}_h.$$

Substituting the above relation into $(4.1)_2$, integrating by parts, and observing that $\mathrm{div}\boldsymbol{\sigma}_h^{DM} + f^K = 0$, we obtain

$$\int_{\partial K} (\mu_h^K - \boldsymbol{\sigma}_h^{DM}\cdot\boldsymbol{n}_K)v_h\,dx = 0 \qquad \forall v_h \in \mathbb{P}_1(K),$$

which clearly implies $\mu_h = \boldsymbol{\sigma}_h^{DM}\cdot\boldsymbol{n}_K$ on ∂K. This result shows that the values μ_h are actually the degrees of freedom of the variable $\boldsymbol{\sigma}_h^{DM}$ and as such provide a simple procedure to recover a self-equilibrated stress field within each element satisfying interelement traction reciprocity.

6.4. Elliptic problem with variable coefficients.

We come now to briefly addressing the extension of the DPG$_0$ method to the case of an elliptic model problem with variable coefficients. With this aim, we consider the Poisson problem

$$-\mathrm{div}(a(x)\nabla u) = f \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad \Gamma, \tag{6.27}$$

where $a = a(x)$ is a symmetric positive definite matrix-valued function. The mixed form of (6.27) reads

$$-\mathrm{div}\,\boldsymbol{\sigma} = f \quad \text{in} \quad \Omega, \qquad \boldsymbol{\sigma} = a(x)\nabla u \quad \text{in} \quad \Omega, \qquad u = 0 \quad \text{on} \quad \Gamma. \tag{6.28}$$

In this case, the discrete formulation (4.1) becomes find $(u_h, \lambda_h; \boldsymbol{\sigma}_h, \mu_h) \in (X_h \times Y_h)$ such that

(6.29)
$$\begin{cases} \displaystyle\sum_{K\in\mathcal{T}_h} \left(\int_K a^{-1}\boldsymbol{\sigma}_h \cdot \boldsymbol{q}_h \, dx + \int_K u_h \mathrm{div}\boldsymbol{q}_h \, dx - \int_{\partial K} \lambda_h \, \boldsymbol{q}_h \cdot \boldsymbol{n} \, ds \right) = 0 & \forall \boldsymbol{q}_h \in W_h, \\ \displaystyle\sum_{K\in\mathcal{T}_h} \left(\int_K \boldsymbol{\sigma}_h \cdot \nabla v_h \, dx - \int_{\partial K} \mu_h \, v_h \, ds \right) = \int_\Omega f v_h \, dx & \forall v_h \in V_h. \end{cases}$$

We again let u_h^* be a nonconforming approximation of the solution u of problem (6.27) satisfying (6.14). Taking $\boldsymbol{q}_h \in (\mathbb{P}_0(K))^2$ in $(6.29)_1$ and integrating by parts as done in the case $a(x) \equiv 1$, we obtain

$$\int_K \left(a^{-1}\boldsymbol{\sigma}_h \cdot \boldsymbol{q}_h - \nabla u_h^* \cdot \boldsymbol{q}_h \right) dx = 0 \qquad \forall K \in \mathcal{T}_h.$$

Upon introducing the *harmonic average* of $a(x)$ defined as $\widetilde{\alpha}_K := (\frac{1}{|K|}\int_K (a(x))^{-1}\, dx)^{-1}$, we immediately get the equivalence

$$\boldsymbol{\sigma}_h = \widetilde{\alpha}^{-1}\nabla u_h^* \qquad \forall K \in \mathcal{T}_h.$$

Taking $v_h \in W_{h,0}^{NC}$ in $(6.29)_2$ and using the previous relation yields

$$\sum_{K\in\mathcal{T}_h} \int_K \widetilde{\alpha}_K^{-1} \nabla u_h^* \cdot \nabla v_h \, dx = \sum_{K\in\mathcal{T}_h} \int_K f v_h \, dx \qquad \forall v_h \in W_{h,0}^{NC}, \tag{6.30}$$

demonstrating that u_h^* turns out to be the nonconforming approximation of problem (6.28) *with harmonic averaging of the coefficient* $a(x)$. It is relevant to observe that u_h^* actually differs from the solution u_h^{NC} of the standard nonconforming approximation of problem (6.28), which would simply read

$$\sum_{K\in\mathcal{T}_h} \int_K \overline{a}_K \, \nabla u_h^{NC} \cdot \nabla v_h \, dx = \sum_{K\in\mathcal{T}_h} \int_K f v_h \, dx \qquad \forall v_h \in W_{h,0}^{NC},$$

where $\overline{a}_K := \frac{1}{|K|}\int_K a(x)\, dx$ is the usual average of $a(x)$ on K. In the presence of strong variations of the coefficient a, the harmonic average is well known to provide superior accuracy and stability compared to the standard average (see [5, 1, 15, 31]).

7. Computer implementation of the DPG method. The object of the present section is to discuss an efficient computer implementation of the DPG$_0$ method. The main issue is to reduce the dimension of the algebraic linear system arising from (4.1). To start with, we consider the following system of 6 equations in 9 unknowns that arises from the contribution of each triangle in (4.1):

$$\begin{bmatrix} A & B & C & \emptyset \\ D & \emptyset & \emptyset & E \end{bmatrix} \begin{pmatrix} \boldsymbol{\sigma} \\ \mathbf{u} \\ \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{f} \end{pmatrix}, \tag{7.1}$$

where the bold symbols represent the vectors of unknowns and given data, and $\mathbf{f}$ is the right-hand side integral in $(4.1)_2$.

On the one hand, one can exploit the nature of hybrid formulations of the DPG method performing a static condensation of the internal variables in favor of the hybrid variables. Defining the new variable $\widehat{\boldsymbol{\sigma}} = [\boldsymbol{\sigma}, \mathbf{u}]^T$, system (7.1) can be rewritten as

$$\begin{bmatrix} \mathcal{A} & C & \emptyset \\ \mathcal{D} & \emptyset & E \end{bmatrix} \begin{pmatrix} \widehat{\boldsymbol{\sigma}} \\ \boldsymbol{\lambda} \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{f} \end{pmatrix}, \qquad \text{with} \quad \mathcal{A} = [A,\, B]\,, \qquad \mathcal{D} = [D,\, \emptyset]\,. \tag{7.2}$$

The (3×3) matrix $\mathcal{A}$ is nonsingular, so that $\widehat{\boldsymbol{\sigma}}$ can be eliminated in favor of the sole edge variable $\boldsymbol{\lambda}$, obtaining the following reduced system of 3 equations in 6 interface unknowns:

$$-\mathcal{D}\mathcal{A}^{-1}C\boldsymbol{\lambda} + E\boldsymbol{\mu} = \mathbf{f} \qquad \forall K \in \mathcal{T}_h, \tag{7.3}$$

which is the algebraic form of (6.20). The matrix E is square and nonsingular because it emanates from the bilinear form $\int_{\partial K} \mu_h \, v_h \, ds$, $\mu_h \in R_0(\partial K)$, $v_h \in \mathbb{P}_1(K)$, which satisfies the discrete inf-sup condition. Therefore, we can eliminate $\boldsymbol{\mu}$ in favor of the sole unknown $\boldsymbol{\lambda}$:

$$\boldsymbol{\mu} = E^{-1}\mathcal{D}\mathcal{A}^{-1}C\boldsymbol{\lambda} + E^{-1}\mathbf{f} \qquad \forall K \in \mathcal{T}_h. \tag{7.4}$$

Enforcing the condition that the hybrid variable μ_h is single-valued on each internal edge yields a square symmetric and positive definite linear system for the sole unknown $\boldsymbol{\lambda}$ of dimension `Ni`.

On the other hand, one can instead exploit the DG nature of the DPG method, eliminating the hybrid variables (counterpart of the numerical fluxes) in favor of the internal variables. Since both E and C are square nonsingular matrices, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ can be eliminated in favor of $\widehat{\boldsymbol{\sigma}}$, obtaining on each element $K \in \mathcal{T}_h$

$$\boldsymbol{\lambda} = \mathcal{A}\widehat{\boldsymbol{\sigma}}, \qquad \boldsymbol{\mu} = E^{-1}(\mathbf{f} - \mathcal{D}\widehat{\boldsymbol{\sigma}}). \tag{7.5}$$

After some algebra, one sees that the first relation in (7.5) expresses λ_h on each element as a discrete Taylor expansion of u about the center of gravity of the element, while the second relation in (7.5) represents a *conservative finite volume-like* discretization of the equilibrium equation. Enforcing now the hybrid variables to be single-valued on each internal edge of the triangulation, we end up with a square nonsingular linear algebraic system of dimension 3`NE` in the sole internal (and fully discontinuous over $\mathcal{T}_h$) unknown vector $\widehat{\boldsymbol{\sigma}}$.

It is interesting to compare the computational effort associated with the two implementation approaches discussed above. Using Euler's theorem (2 `Ni` + `Nb` = 3`NE`) shows that the cost of the solution of the system in the discontinuous unknown $\widehat{\boldsymbol{\sigma}}$ *is (asymptotically) twice* the cost associated with the solution of the system in the sole hybrid unknown $\boldsymbol{\lambda}$. The superior efficiency of this latter formulation makes it preferable in computations, and it is the reason why it has been used in all of the numerical examples shown in the next sections.

8. Numerical results.

In this section we present the results obtained by applying the DPG$_0$ formulation to the numerical solution of two elliptic model problems.

8.1. Elliptic model problem 1.

We consider problem (2.1) on the unit square with $\Gamma \equiv \Gamma_D$, such that the exact solution is the "bubble function" $u = x(1-x)y(1-y)$, with the right-hand side f computed accordingly. In Figure 8.1 (left) we show the computed convergence rates using four different unstructured meshes for the quantities $||u - u_h||_{0,\Omega}$, ($\circ$); $||P_h u - u_h||_{0,\Omega}$, ($\nabla$); $||u - u_h^*||_{0,\Omega}$, ($\square$); and $||\rho_h^0 \lambda - \lambda_h||_{-1/2,h}$, ($*$), while in Figure 8.1 (right) we show the computed convergence rates for the quantities $||\boldsymbol{\sigma} - \boldsymbol{\sigma}_h||_{0,\Omega}$, ($\square$), and $||\rho_h^0 \mu - \mu_h||_{-1/2,h}$, ($\circ$). The computed errors are in agreement with the theoretical estimates of section 6.2.

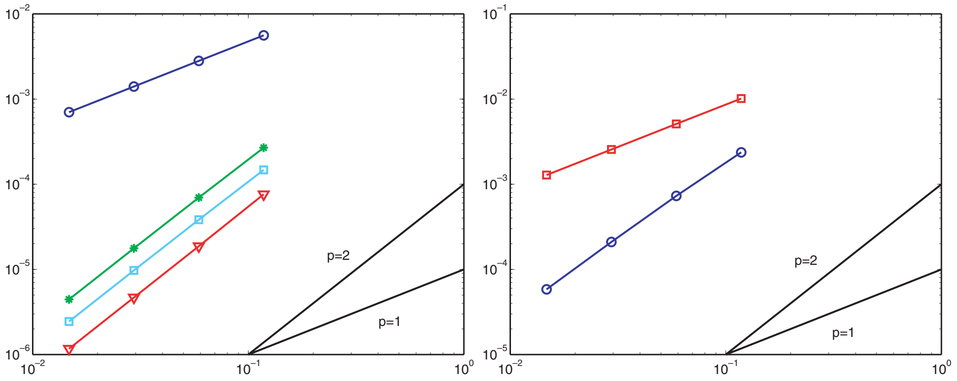


FIG. 8.1. *Error norms for the elliptic model problem* 1.

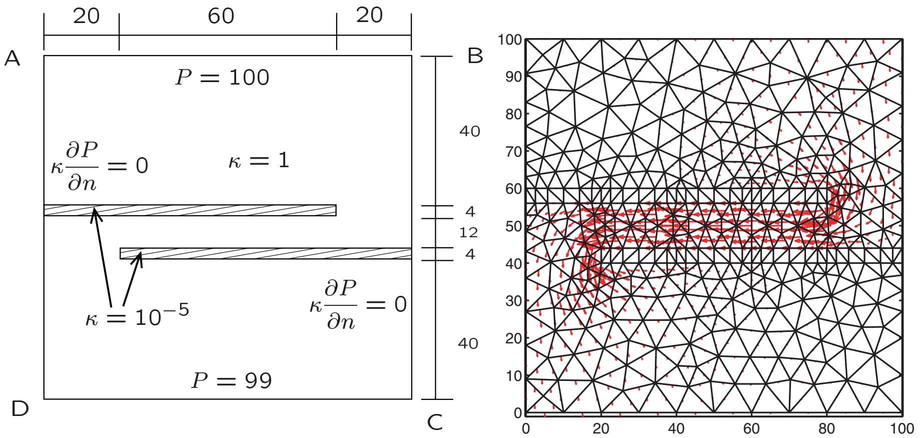


FIG. 8.2. *Problem setting for the flow in a porous medium (left) and associated velocity field (right).*

8.2. Elliptic model problem 2. We study the problem of a two-dimensional steady flow system in a porous medium modeled by Darcy's law [32]: find the hydraulic potential P and the associated velocity field $q = \kappa \nabla P$, where κ is the hydraulic conductivity tensor such that

$$\left\{ \begin{array}{llll} -\operatorname{div} q = 0 & \text{in} \quad \Omega, & q = \kappa \nabla P & \text{in} \quad \Omega, \\ P = P_D & \text{on} \quad \Gamma_D, & q \cdot n = q_N & \text{on} \quad \Gamma_N. \end{array} \right. \tag{8.1}$$

In Figure 8.2 (left) we show the computational domain, the boundary conditions, and the piecewise constant values of κ which are seen to attain strong variations on Ω. In Figure 8.2 (right), we show the velocity field represented as an $\mathbb{RT}_0$ finite element function reconstructed over $\overline{\Omega}$ from the computed values μ_h as in (6.20). The continuity of the normal component of the velocity field across interelement edges is a crucial property when computing the flow streamlines (see [32] for a discussion of this issue).

9. Conservation properties of the DPG method. The present section is aimed at enlightening through a numerical example the conservation properties of the DPG method. We observe that

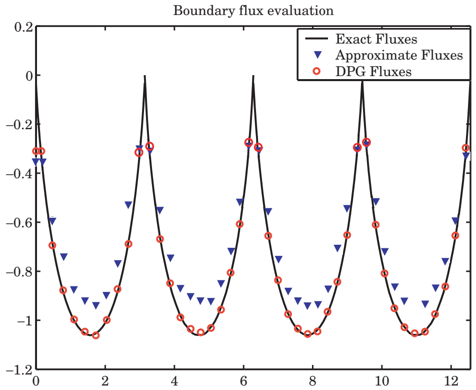


FIG. 9.1. *Exact fluxes on* Γ *compared with the fluxes computed by the* CG *method with no postprocessing and with the interface field* μ_h *computed by the* DPG$_0$ *method.*

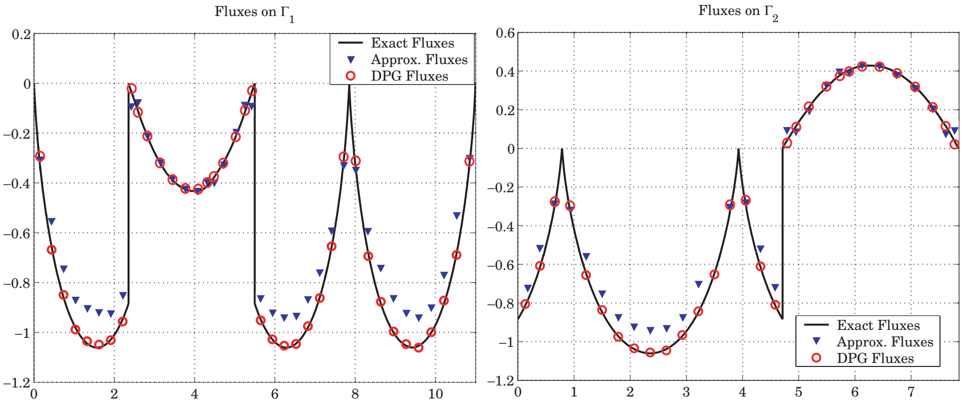


FIG. 9.2. *Exact fluxes on the boundaries* Γ_1 *(left) and* Γ_2 *(right) compared with the fluxes computed by the* CG *method with no postprocessing and interface fields computed by the* DPG$_0$ *method.*

1. integral global conservation is achieved by taking in (4.1) $\boldsymbol{q}_h \equiv [1,1]^T$ and $v_h \equiv 1$ in Ω, respectively;
2. integral local conservation is achieved by taking in (4.1) $\boldsymbol{q}_h = [1,1]^T$ and $v_h = 1$, in any subdomain $\mathcal{S} \subseteq \Omega$ and zero elsewhere, respectively.

In the standard continuous Galerkin (CG) method neither the first nor the second choice for the test function is admissible [26]. Recovering fluxes that enjoy the desired conservation properties requires a postprocessing procedure, thus adding additional computational cost to the basic CG discretization. Moreover, if, for example, a nodal flux approach is used as in [26], overshoots and undershoots appear when a node coincides with an endpoint of the interface, since there the flux is artificially enforced to be continuous.

To numerically assess these concepts, we solve the Poisson equation on the domain $\Omega = [0,\pi] \times [0,\pi]$ with $u = 0$ on $\Gamma = \partial\Omega$ and $f = 1$. To test local conservation properties, we split Ω into the subdomains $\Omega_1 = [0, \frac{3}{4}\pi] \times [0,\pi]$, $\Omega_2 = [\frac{3}{4}\pi, \pi] \times [0,\pi]$ such that $\Omega = \Omega_1 \cup \Omega_2$ and with boundaries Γ_1 and Γ_2, respectively. From the exact solution of the problem (see [40]), we compute the fluxes $\boldsymbol{\sigma} \cdot \boldsymbol{n} = \nabla u \cdot \boldsymbol{n}$ on Γ (Figure 9.1), Γ_1 (Figure 9.2, left), and Γ_2 (Figure 9.2, right) and we compare them with the numerical fluxes obtained from the displacement field u_{CG} of a piecewise linear CG approximation and with the field μ_h obtained from the DPG_0 (using the same grid). The DPG fluxes are accurate and do not exhibit spurious oscillations at the endpoints of the boundaries. Moreover, the global equilibrium $\int_\Omega f\,dx + \int_\Gamma \boldsymbol{\sigma}\cdot\boldsymbol{n}\,ds = 0$ is verified to machine precision by the DPG approximation.

10. Conclusions. In this article we have presented the DPG method for the finite element discretization scheme of second order elliptic boundary value problems. A stability and convergence analysis of the novel formulation has been carried out, and numerical results have been shown to validate the computational performance of the novel formulation. Introducing the DPG formulation has established a clear connection between mixed-hybrid and DG methods. This result is the motivation and starting point for future investigations and applications of the novel scheme to deal with more general problems.

Acknowledgments. The authors wish to thank Professor F. Brezzi and Professor L. D. Marini for their time in fruitful discussions and their suggestions. Moreover, the observations, remarks, and suggestions of the referees are also gratefully acknowledged.

REFERENCES

[1] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing, and error estimates*, M2AN Math. Model. Numer. Anal., 19 (1985), pp. 7–32.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Discontinuous Galerkin methods for elliptic problems*, Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin, 2000, pp. 89–101.

[3] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[4] S. N. Atluri, P. Tong, and H. Murakawa, *Recent studies in hybrid and mixed finite element methods in mechanics*, in Hybrid and Mixed Finite Element Methods, S. N. Atluri, R. Gallagher, and O. C. Zienkiewicz, eds., John Wiley and Sons, New York, 1983, pp. 51–71.

[5] I. Babuska and J. E. Osborn, *Generalized finite element methods: Their performance and their relation to mixed methods*, SIAM J. Numer. Anal., 20 (1983), pp. 510–536.

[6] F. Bassi, G. Mariotti, S. Pedinotti, S. Rebay, and M. Savini, *A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows*, in Proceedings of the 2nd European Conference on Turbomachinery Fluid Dynamics and Thermodynamics, R. Decuypere and G. Dibelius, eds., 1997, pp. 99–108.

[7] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[8] C. Bernardi, C. Canuto, and Y. Maday, *Generalized inf-sup condition for Chebyshev spectral approximation of the Stokes problem*, SIAM J. Numer. Anal., 25 (1988), pp. 1237–1271.

[9] M. Borri, C. L. Bottasso, and P. Mantegazza, *Basic features of the time finite element approach for dynamics*, Meccanica, 27 (1992), pp. 119–130.

[10] C. L. Bottasso, S. Micheletti, and R. Sacco, *The discontinuous Petrov-Galerkin method for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 3391–3409.

[11] C. L. Bottasso, P. Causin, S. Micheletti, and R. Sacco, *The discontinuous Petrov-Galerkin finite element method*, in Proceedings of the Sixth U.S. National Congress on Computational Mechanics, Dearborn, MI, 2001, p. 401.

[12] C. L. Bottasso, S. Micheletti, and R. Sacco, *A multiscale formulation of the discontinuous Petrov–Galerkin method for advective-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 2819–2838.

[13] F. Brezzi, J. Douglas, M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 581–604.

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

[15] F. Brezzi, L. D. Marini, and P. Pietra, *Two-dimensional exponential fitting and applications to drift-diffusion models*, SIAM J. Numer. Anal., 26 (1989), pp. 1342–1355.

[16] P. Castillo, B. Cockburn, I. Perugia, and D. Schötzau, *Local discontinuous Galerkin method for elliptic problems*, Comm. Numer. Methods Engrg., 18 (2002), pp. 69–75.

[17] P. Causin and R. Sacco, *Mixed-hybrid Galerkin and Petrov-Galerkin finite element formulations in continuum mechanics*, in Proceedings of the 5th World Congress on Computational Mechanics (WCCM V), F. H. A. Mang and J. Eberhardsteiner, eds., Vienna University of Technology, Vienna, 2002, http://wccm.tuwien.ac.at.

[18] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[19] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods: Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. E. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 3–50.

[20] C. Dawson and V. Aizinger, *Upwind-mixed methods for transport equations*, Comput. Geosci., 11 (1999), pp. 93–110.

[21] C. Farhat, I. Harari, and L. Franca, *The discontinuous enrichment method*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6455–6479.

[22] C. Farhat, I. Harari, and U. Hetmaniuk, *A discontinuous Galerkin method with Lagrange multipliers for the solution of Helmholtz problems in the mid-frequency regime*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 1389–1419.

[23] M. Farhloul and M. Fortin, *A new mixed finite element for the Stokes and elasticity problems*, SIAM J. Numer. Anal., 30 (1993), pp. 971–990.

[24] M. Farhloul and M. Fortin, *Dual hybrid methods for the elasticity and the Stokes problems: A unified approach*, Numer. Math., 76 (1997), pp. 419–440.

[25] V. Girault and P. A. Raviart, *Finite Element Approximation of Navier–Stokes Equations*, Lecture Notes in Math. 749, Springer-Verlag, Berlin, 1979.

[26] T. J. R. Hughes, G. Engel, L. Mazzei, and M. G. Larson, *The continuous Galerkin method is locally conservative*, J. Comput. Phys., 163 (2000), pp. 467–488.

[27] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. 1, Travaux et Recherches Mathématiques 17, Dunod, Paris, 1968.

[28] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Vol. 2, Travaux et Recherches Mathématiques 18, Dunod, Paris, 1968.

[29] L. D. Marini, *An inexpensive method for the evaluation of the solution of the lowest order Raviart–Thomas mixed method*, SIAM J. Numer. Anal., 22 (1985), pp. 493–496.

[30] S. Micheletti and R. Sacco, *Dual-primal mixed finite elements for elliptic problems*, Numer. Methods Partial Differential Equations, 17 (2001), pp. 137–151.

[31] S. Micheletti, R. Sacco, and F. Saleri, *On some mixed finite element methods with numerical integration*, SIAM J. Sci. Comput., 23 (2001), pp. 245–270.

[32] R. Mosé, P. Ackerer, P. Siegel, and G. Chavent, *Application of the mixed hybrid finite element approximation in a groundwater flow model: Luxury or necessity?*, Water Resources Research, 30 (1994), pp. 3001–3012.

[33] R. A. Nicolaides, *Existence, uniqueness and approximation for generalized saddle point problems*, SIAM J. Numer. Anal., 19 (1982), pp. 349–357.

[34] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.

[35] P. A. Raviart and J. M. Thomas, *Primal hybrid finite element methods for* 2*nd order elliptic equations*, Math. Comp., 31 (1977), pp. 391–413.

[36] P. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, Berlin, 1977, pp. 292–315.

[37] J. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, Handb. Numer. Anal. II, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 523–639.

[38] R. Stenberg, *A family of mixed finite elements for the elasticity problem*, Numer. Math., 53 (1988), pp. 513–538.

[39] J. M. Thomas and D. Trujillo, *Finite volume methods for elliptic problems: Convergence on unstructured meshes*, in Numerical Methods in Mechanics, Pitman Res. Notes Math. Ser. 371, Longman, Harlow, UK, 1997, pp. 163–174.

[40] H. F. Weinberger, *A First Course in Partial Differential Equations with Complex Variables and Transform Methods*, Blaisdell and Ginn, New York, Toronto, London, 1965.

SIAM J. Numer. Anal.
Vol. 43, No. 1, pp. 303–317

CONVERGENCE ANALYSIS OF FULLY DISCRETE FINITE VOLUME METHODS FOR MAXWELL'S EQUATIONS IN NONHOMOGENEOUS MEDIA[*]

ERIC T. CHUNG[†] AND BJORN ENGQUIST[‡]

Abstract. We will consider both explicit and implicit fully discrete finite volume schemes for solving three-dimensional Maxwell's equations with discontinuous physical coefficients on general polyhedral domains. Stability and convergence for both schemes are analyzed. We prove that the schemes are second order accurate in time. Both schemes are proved to be first order accurate in space for the Voronoi–Delaunay grids and second order accurate for nonuniform rectangular grids. We also derive explicit expressions for the dependence on the physical parameters in all estimates.

Key words. convergence, stability, finite volume, fully discrete, Maxwell equations

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

DOI. 10.1137/S0036142903435442

1. Introduction.

The finite volume method (FVM) has been developed as a practical compromise between the finite difference method (FDM) and the finite element method (FEM) in the numerical solutions of electromagnetic problems [1, 2, 3, 4, 6, 9, 10, 11, 15, 16]. It allows for unstructured grids, as the FEM does, but it is typically explicit and as computationally efficient as the FDM.

Even though FEMs provide excellent tools for solving electromagnetic problems on geometrically complex domains, marching techniques applied to FEMs produce implicit schemes. In order to obtain an explicit scheme to solve Maxwell's equations on domains which are geometrically complicated, FVMs are considered. There has been serious stability problems with some FVM approximations of the Maxwell's equations. In this paper we will present a class of FVMs for which we prove stability and convergence. We will in particular consider inhomogeneous media with discontinuous coefficients. The methods presented here are based on the semidiscrete techniques introduced by Chung, Du, and Zou [4, 6]. Rigorous analysis of fully discrete FVMs for electromagnetics is rarely seen in the literature. It is the purpose of this paper to provide stability and convergence analysis of two fully discrete FVMs. The first one is the leapfrog scheme. We will analyze this scheme by using the approach suggested in Nicolaides and Wang [13], where an FVM for Maxwell's equations with constant coefficients is considered. The second one is the Crank–Nicolson scheme. This scheme is unconditionally stable, so that larger time steps can be used. In addition to stability and convergence, we also prove existence and uniqueness of solution of the discrete problem resulting from the implicit Crank–Nicolson discretization.

Let us first introduce notation and formulate the differential equations to be approximated. Let Ω be a polyhedral domain in $\mathbb{R}^3$ and $T > 0$. Then the electric

[*] Received by the editors September 26, 2003; accepted for publication (in revised form) September 3, 2004; published electronically May 27, 2005.

http://www.siam.org/journals/sinum/43-1/43544.html

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

[‡] PACM, Department of Mathematics, Princeton University, Princeton, NJ 08544 (engquist@math.princeton.edu).

field $\mathbf{E}(\mathbf{x},t)$ and the magnetic field $\mathbf{H}(\mathbf{x},t)$ satisfy the Maxwell's equations

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

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

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

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

where $\mathbf{J}(\mathbf{x},t)$ and $\rho(\mathbf{x},t)$ are the applied current density and the charge density, respectively. We also have the constitutive relations $\mathbf{D}=\varepsilon\mathbf{E}$ and $\mathbf{B}=\mu\mathbf{H}$, where $\mathbf{D}$ and $\mathbf{B}$ are the electric flux density and the magnetic flux density, respectively. The paper is concerned with the case where the domain Ω is composed of two distinct dielectric materials. Let $\Omega_1\subset\Omega$ be a polyhedral subdomain and $\Omega_2=\Omega\backslash\bar{\Omega}_1$. We assume the electric permittivity ε and the magnetic permeability μ are piecewise constant functions such that $\varepsilon=\varepsilon_i$ and $\mu=\mu_i$ in Ω_i, where ε_i and μ_i are positive constants for $i=1,2$. We supplement the system (1.1)–(1.4) with the perfect conductor boundary condition $\mathbf{E}\times\mathbf{n}=0$ and initial condition. On the interface $\Gamma := \partial\Omega_1$, $\mathbf{E}$ and $\mathbf{H}$ satisfy

$$[\mathbf{E}\times\mathbf{m}]=0, \quad [\varepsilon\mathbf{E}\cdot\mathbf{m}]=\rho_\Gamma, \tag{1.5}$$

$$[\mathbf{H}\times\mathbf{m}]=0, \quad [\mu\mathbf{H}\cdot\mathbf{m}]=0, \tag{1.6}$$

where $\rho_\Gamma(\mathbf{x})$ is the surface charge density, $\mathbf{m}$ is the unit normal of $\partial\Omega_1$, and $[f] := f_2|_\Gamma - f_1|_\Gamma$ is the jump of the function f across the interface with $f_i=f|_{\Omega_i}$ for $i=1,2$.

The rest of the paper is organized as follows. In section 2, we will describe the triangulation of the domain as well as some discrete vector spaces and operators defined on it. In section 3, we will briefly summarize the finite volume spatial discretization of Maxwell's equations. In section 4, we will derive both explicit and implicit fully discrete finite volume schemes. They are based on leapfrog and Crank–Nicolson time discretization, respectively. Convergence and stability will be analyzed in case of general tetrahedral grids. In section 5, we extend the schemes to nonuniform rectangular grids. It can be shown that the spatial convergence is one order higher. We will also provide an example in one space dimension to show the optimality of our estimates in this section.

2. Discrete vector spaces. In this section, we will briefly summarize the triangulation of the domain as well as discrete vector spaces and operators defined on it. The domain Ω is triangulated by the standard Voronoi–Delaunay triangulation (cf. [8]). We will refer to the Delaunay triangulation, which contains tetrahedral cells, as the primal grid and the Voronoi triangulation, which contains polyhedral cells and are formed by connecting the circumcenters of adjacent primal cells, as the dual grid. The primal grid is chosen so that the faces of the primal cells align with the interface Γ. For a more detailed description and assumption of the triangulation, see Chung, Du, and Zou [6]. Throughout this paper, we use K to represent a generic constant which is independent of the mesh size and the physical coefficients ε and μ.

Now, we will discuss the discrete vector spaces and operators. Let F_1 and M_1 be the number of interior primal faces and interior primal edges, respectively. Then F_1 is also the number of dual edges and M_1 is also the number of dual faces. The individual primal face and edge will be denoted by κ_j and σ_k with $s_j=|\kappa_j|$ and $h_k=|\sigma_k|$. The corresponding quantities related to the dual grid are denoted by primed form such as κ_j', σ_k', s_j', and h_k'. Let $(\cdot,\cdot)$ be the standard Euclidean inner product. For any u and

v in $\mathbb{R}^{F_1}$, we define the inner product $(u, v)_W := (Su, D'v)$, where $S := \mathrm{diag}(s_j)$ and $D' := \mathrm{diag}(\bar{h}'_j)$ with $\bar{h}'_j := (\mu_1^{-1} a_j + \mu_2^{-1}(1 - a_j))h'_j$. Here a_j, $0 \leq a_j \leq 1$, is the ratio of the length of the portion of σ'_j that belongs to Ω_1 over the length of σ'_j. We also define an inner product in $\mathbb{R}^{M_1}$ by $(u, v)_{W'} = (S'u, Dv)$, where $S' := \mathrm{diag}(\bar{s}'_j)$ and $D := \mathrm{diag}(h_j)$ with $\bar{s}'_j := (\varepsilon_1 b_j + \varepsilon_2(1 - b_j))s'_j$ and b_j, $0 \leq b_j \leq 1$, is the ratio of the area of the portion of κ'_j that belongs to Ω_1 over the area of κ'_j.

Each primal and dual edge is assigned a direction. Direction is also assigned to each primal and dual face such that it is the same as the corresponding dual and primal edge, respectively. We say $\sigma'_j \in \partial\kappa'_i$ is oriented positively along $\partial\kappa'_i$ if the direction of σ'_j agrees with the one of $\partial\kappa'_i$ formed by the right-hand rule with the thumb pointing in the direction of σ_i. Otherwise, we say σ'_j is oriented negatively along $\partial\kappa'_i$. We define an $F_1 \times M_1$ matrix G as

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

Then, for $w \in \mathbb{R}^{M_1}$ and $v \in \mathbb{R}^{F_1}$, we define $Cw := GDw$ and $C'v := G^T D'v$, which also satisfy the following discrete Green's formula:

$$(Cw, D'v) = (C'v, Dw). \tag{2.1}$$

See [5, 6, 12, 13, 14] for more details about these operators and related topics.

3. FVM. The FVM proposed in Chung and Zou [4] can be stated as follows:

Find $E \in \mathbb{R}^{M_1}$ and $B \in \mathbb{R}^{F_1}$ such that

$$S' \frac{dE}{dt} - C'B = \tilde{J}, \tag{3.1}$$

$$S \frac{dB}{dt} + CE = 0, \tag{3.2}$$

where $\tilde{J} \in \mathbb{R}^{M_1}$ is defined by $(\tilde{J})_j := \int_{\kappa'_j} \mathbf{J} \cdot \mathbf{n} \, d\sigma$. Convergence and stability analysis of the semidiscrete scheme (3.1)–(3.2) was given in Chung, Du, and Zou [6]. In the following sections, we will give our main results in this paper, which are the full discretizations of the semidiscrete scheme (3.1)–(3.2).

For subsequent analysis, derived in [4], we have

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

$$\bar{s}'_j \frac{d}{dt}(E'_f)_j - (C'B'_e)_{\kappa'_j} = \tilde{J}_j, \tag{3.4}$$

where B_f, B'_e, E'_f, and E_e are average quantities defined in Chung, Du, and Zou [4, 6].

4. Fully discrete schemes. In this section, we will give two different time discretizations for (3.1)–(3.2). First, we will consider an explicit scheme, which is a standard leapfrog scheme. With a stability condition, this scheme can be shown to be first order convergent in space and second order convergent in time. Second, we will consider an implicit scheme, which is a Crank–Nicolson time discretization of

(3.1)–(3.2). It can be shown that this method is unconditionally stable with the same rate of convergence as the explicit counterpart.

Consider a uniform partition of $[0,T]$ with N_T subintervals. Let $\Delta t := T/N_T$. For $n = 0, 1, \dots, N_T - 1$, we define $t_n := n\Delta t$ and $t_{n+\frac{1}{2}} := (n + \frac{1}{2})\Delta t$. For subsequent analysis, we define

$$c_m := (\min(\varepsilon_1, \varepsilon_2)\min(\mu_1, \mu_2))^{-\frac{1}{2}}.$$

We also denoted by M_2 the maximum of the ratios of maximum to minimum side lengths over adjacent tetrahedra and by M_3 the maximum number of sides over all dual faces. Furthermore, by $\min(h)$ we mean the minimum side length over all primal and dual edges.

4.1. Explicit scheme. We will apply a leapfrog scheme to (3.1)–(3.2). We approximate $\mathbf{E}(t)$ at t_n and $\mathbf{B}(t)$ at $t_{n+\frac{1}{2}}$ with approximations denoted by E^n and $B^{n+\frac{1}{2}}$, respectively. Given $(E^n, B^{n+\frac{1}{2}})$, the next approximation $(E^{n+1}, B^{n+\frac{3}{2}})$ will be obtained by solving

$$S'(E^{n+1} - E^n) - \Delta t C' B^{n+\frac{1}{2}} = \tilde{J}^{n+\frac{1}{2}}, \tag{4.1}$$

$$S(B^{n+\frac{3}{2}} - B^{n+\frac{1}{2}}) + \Delta t C E^{n+1} = 0, \tag{4.2}$$

where $\tilde{J}^{n+\frac{1}{2}} := \int_{n\Delta t}^{(n+1)\Delta t} \tilde{J}(s)\, ds$. The initial conditions are given by $E^0 := E_e(t_0)$ and $B^{\frac{1}{2}} := B_e'(t_{\frac{1}{2}})$. The value of $B_e'(t_{\frac{1}{2}})$ can be calculated by using (1.2) and the Taylor series method. Now, we have the following stability estimate.

THEOREM 4.1. *Under the stability condition*

$$\delta := \Delta t c_m \frac{M_2 M_3^{\frac{1}{2}}}{\min(h)} < 1, \tag{4.3}$$

the fully discrete scheme (4.1)–(4.2) *is stable. Moreover, the following stability estimate holds for* $1 \le k \le N_T - 1$*:*

$$\begin{aligned} &\|E^k\|_{W'}^2 + \|B^{k+\frac{1}{2}}\|_W^2 \\ &\quad\le \frac{2(1+\delta)}{1-\delta}(\|E^0\|_{W'}^2 + \|B^{\frac{1}{2}}\|_W^2) + \frac{4T}{(1-\delta)^2}\int_0^T \|S'^{-1}\tilde{J}(t)\|_{W'}^2\, dt. \end{aligned} \tag{4.4}$$

Proof. Multiplying (4.1) by $D(E^{n+1} + E^n)$ and (4.2) by $D'(B^{n+\frac{3}{2}} + B^{n+\frac{1}{2}})$ yields

$$(E^{n+1} - E^n, E^{n+1} + E^n)_{W'} - \Delta t(C'B^{n+\frac{1}{2}}, D(E^{n+1} + E^n)) = R^{n+\frac{1}{2}},$$

$$(B^{n+\frac{3}{2}} - B^{n+\frac{1}{2}}, B^{n+\frac{3}{2}} + B^{n+\frac{1}{2}})_W + \Delta t(CE^{n+1}, D'(B^{n+\frac{3}{2}} + B^{n+\frac{1}{2}})) = 0,$$

where $R^{n+\frac{1}{2}} := (\tilde{J}^{n+\frac{1}{2}}, D(E^{n+1} + E^n))$. Let k be an integer satisfying $1 \le k \le N_T - 1$. Adding all equations from $n = 0$ to $n = k - 1$ and using (2.1), we have

$$\begin{aligned} &\|E^k\|_{W'}^2 + \|B^{k+\frac{1}{2}}\|_W^2 \\ &= \|E^0\|_{W'}^2 + \|B^{\frac{1}{2}}\|_W^2 + \Delta t(C'B^{\frac{1}{2}}, DE^0) - \Delta t(DE^k, C'B^{k+\frac{1}{2}}) + \sum_{n=0}^{k-1} R^{n+\frac{1}{2}}. \end{aligned} \tag{4.5}$$

Notice that

$$\Delta t(C'B^{\frac{1}{2}}, DE^0) \leq \Delta t\|(DS^{'-1})^{\frac{1}{2}}C'(SD')^{-\frac{1}{2}}\|_2\|B^{\frac{1}{2}}\|_W\|E^0\|_{W'}.$$

By the definition of the matrix 2-norm, $\|(DS^{'-1})^{\frac{1}{2}}C'(SD')^{-\frac{1}{2}}\|_2$ is bounded above by the square root of the largest eigenvalue of $(SD')^{-\frac{1}{2}}D'GDS^{'-1}G^TD'(SD')^{-\frac{1}{2}}$. By Gershgorin's theorem, the largest eigenvalue of GG^T is bounded above by $3M_3$, as every primal face has 3 sides and every dual face has at most M_3 sides. So,

$$\|(DS^{'-1})^{\frac{1}{2}}C'(SD')^{-\frac{1}{2}}\|_2 \leq c_m \max_{1\leq j\leq F_1}\left(\frac{\max_j(h)}{\min_j(h)^2}\right)(3M_3)^{\frac{1}{2}},$$

where $\max_j(h)$ and $\min_j(h)$ are the local maximum and local minimum of side lengths around a primal face κ_j, respectively. So, we have

$$\Delta t(C'B^{\frac{1}{2}}, DE^0) \leq \Delta t c_m \frac{M_2M_3^{\frac{1}{2}}}{\min(h)}(\|B^{\frac{1}{2}}\|_W^2 + \|E^0\|_{W'}^2).$$

A similar inequality holds for $\Delta t(C'B^{k+\frac{1}{2}}, DE^k)$. Hence, (4.5) can be written as

$$\|E^k\|_{W'}^2 + \|B^{k+\frac{1}{2}}\|_W^2 \leq \frac{1+\delta}{1-\delta}(\|E^0\|_{W'}^2 + \|B^{\frac{1}{2}}\|_W^2) + \frac{1}{1-\delta}\sum_{n=0}^{k-1}R^{n+\frac{1}{2}}.$$

Now, (4.4) follows from the Cauchy–Schwarz inequality. □

We are now in a position to study the convergence theory of the method (4.1)–(4.2). To do so, for brevity, we define

$$\mathbf{e}(E)^n := E^n - E_e(t_n), \quad \mathbf{e}(B)^n := B^n - B'_e(t_n), \tag{4.6}$$

$$\mathbf{f}(E)^n := E^n - E'_f(t_n), \quad \mathbf{f}(B)^n := B^n - B_f(t_n). \tag{4.7}$$

Then we subtract (4.1) by (3.4) and (4.2) by (3.3) to obtain

$$S'(\mathbf{f}(E)^{n+1} - \mathbf{f}(E)^n) - \Delta t C'\mathbf{e}(B)^{n+\frac{1}{2}} = P_1^{n+\frac{1}{2}}, \tag{4.8}$$

$$S(\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}) + \Delta t C\mathbf{e}(E)^{n+1} = P_2^{n+1}, \tag{4.9}$$

where

$$P_1^{n+\frac{1}{2}} := \tilde{J}^{n+\frac{1}{2}} - S'(E'_f(t_{n+1}) - E'_f(t_n)) + \Delta t C'B'_e(t_{n+\frac{1}{2}}), \tag{4.10}$$

$$P_2^{n+1} := -S(B_f(t_{n+\frac{3}{2}}) - B_f(t_{n+\frac{1}{2}})) - \Delta t C E_e(t_{n+1}). \tag{4.11}$$

Multiplying (4.8) by $D\mathbf{e}(E)^n + D\mathbf{e}(E)^{n+1}$ and (4.9) by $D'\mathbf{e}(B)^{n+\frac{3}{2}} + D'\mathbf{e}(B)^{n+\frac{1}{2}}$, adding the two resulting equations, and using (2.1), (4.6), and (4.7), we have

$$\begin{aligned}
&(\mathbf{e}(E)^{n+1} - \mathbf{e}(E)^n, \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'} \\
&\quad + (\mathbf{e}(B)^{n+\frac{3}{2}} - \mathbf{e}(B)^{n+\frac{1}{2}}, \mathbf{e}(B)^{n+\frac{3}{2}} + \mathbf{e}(B)^{n+\frac{1}{2}})_W \\
&= ((E'_f(t_{n+1}) - E_e(t_{n+1})) - (E'_f(t_n) - E_e(t_n)), \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'} \\
&\quad + ((B_f(t_{n+\frac{3}{2}}) - B'_e(t_{n+\frac{3}{2}})) - (B_f(t_{n+\frac{1}{2}}) - B'_e(t_{n+\frac{1}{2}})), \mathbf{e}(B)^{n+\frac{3}{2}} + \mathbf{e}(B)^{n+\frac{1}{2}})_W \\
&\quad + (P_1^{n+\frac{1}{2}}, D(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})) + (P_2^{n+1}, D'(\mathbf{e}(B)^{n+\frac{3}{2}} + \mathbf{e}(B)^{n+\frac{1}{2}})) \\
&\quad + \Delta t(C'\mathbf{e}(B)^{n+\frac{1}{2}}, D\mathbf{e}(E)^n) - \Delta t(D\mathbf{e}(E)^{n+1}, C'\mathbf{e}(B)^{n+\frac{3}{2}}).
\end{aligned}$$

Adding from $n = 0$ to $n = k - 1$, we have

$$\begin{aligned}
&\|\mathbf{e}(E)^k\|_{W'}^2 + \|\mathbf{e}(B)^{k+\frac{1}{2}}\|_W^2 \\
&= -\Delta t(D\mathbf{e}(E)^k, C'\mathbf{e}(B)^{k+\frac{1}{2}}) \\
&\quad + \sum_{n=0}^{k-1}\Big\{((E_f'(t_{n+1}) - E_e(t_{n+1})) - (E_f'(t_n) - E_e(t_n)), \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'} \\
&\quad + ((B_f(t_{n+\frac{3}{2}}) - B_e'(t_{n+\frac{3}{2}})) - (B_f(t_{n+\frac{1}{2}}) - B_e'(t_{n+\frac{1}{2}})), \mathbf{e}(B)^{n+\frac{1}{2}} + \mathbf{e}(B)^{n+\frac{3}{2}})_W \\
&\quad + (P_1^{n+\frac{1}{2}}, D(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})) + (P_2^{n+1}, D'(\mathbf{e}(B)^{n+\frac{1}{2}} + \mathbf{e}(B)^{n+\frac{3}{2}}))\Big\}.
\end{aligned} \tag{4.12}$$

Next, we will prove the following truncation error for the discretization of the semidiscrete scheme (3.1)–(3.2) by leapfrog time stepping. We will use a dot to represent the time derivative.

LEMMA 4.2. *Assume* $(\mathbf{B}, \mathbf{E}) \in H^2(0, T; W^{1,p}(\Omega_r)^3)^2$ *for* $2 < p \leq 3$ *and* $r = 1, 2$. *Then for general unstructured grids,*

$$\sum_{n=0}^{k-1} \|S'^{-1}P_1^{n+\frac{1}{2}}\|_{W'} \leq c_m K(\Delta t)^2 \sum_{r=1}^{2} \|\mu_r^{-\frac{1}{2}}\mathbf{B}\|_{H^2(0,T;W^{1,p}(\Omega_r)^3)}, \tag{4.13}$$

$$\sum_{n=0}^{k-1} \|S^{-1}P_2^{n+1}\|_W \leq c_m K(\Delta t)^2 \sum_{r=1}^{2} \|\varepsilon_r^{\frac{1}{2}}\mathbf{E}\|_{H^2(0,T;W^{1,p}(\Omega_r)^3)}. \tag{4.14}$$

Proof. We will give a proof for (4.13). The proof for (4.14) can be done in a similar way. For the jth dual face, using (3.4) and (4.10), we have

$$(P_1^{n+\frac{1}{2}})_j = -\int_{n\Delta t}^{(n+1)\Delta t} (C'B_e')_j(s)\, ds + \Delta t (C'B_e')_j(t_{n+\frac{1}{2}}).$$

By the Sobolev embedding theorem, $(P_1^{n+\frac{1}{2}})_j$ defines a bounded linear functional on $W^{2,1}(n\Delta t, (n+1)\Delta t)$ and vanishes for any linear functions in time. So by the Bramble–Hilbert lemma [7] and the standard scale change technique, we have

$$|(P_1^{n+\frac{1}{2}})_j| \leq K(\Delta t)^2 \int_{n\Delta t}^{(n+1)\Delta t} |(C'\ddot{B}_e')_j(s)|\, ds.$$

Using the tangential continuity of $\mathbf{H}$ across the interface Γ, we conclude that $(C'\ddot{B}_e)_j$ vanishes for constant functions in space. By the Bramble–Hilbert lemma and scale change argument,

$$|(C'\ddot{B}_e')_j| \leq Kh^{2-\frac{3}{p}} |\ddot{\mathbf{H}}|_{W^{1,p}(\tau_k' \cup \tau_l')^3},$$

where τ_k' and τ_l' are the two dual cells sharing the same dual face κ_j'. So, we have

$$|(P_1^{n+\frac{1}{2}})_j| \leq Kh^{2-\frac{3}{p}} (\Delta t)^2 \int_{n\Delta t}^{(n+1)\Delta t} |\ddot{\mathbf{H}}|_{W^{1,p}(\tau_k' \cup \tau_l')^3}\, ds.$$

Hence, the result (4.13) follows from the definitions of c_m and the W'-norm. □

The following theorem gives the main result of this section. It states that the explicit scheme (4.1)–(4.2) is first order convergent in space and second order convergent in time under an assumption on the regularity of the true solution and a CFL stability condition.

THEOREM 4.3. *Assume that* $(\mathbf{E},\mathbf{B}) \in H^2(0,T;W^{1,p}(\Omega_r)^3)^2$, $2 < p \le 3$, $r = 1,2$, *is the solution to* (1.1)–(1.2) *and* $(E^n, B^{n+\frac{1}{2}})$ *is the solution to the explicit fully discrete scheme* (4.1)–(4.2). *Then, under the stability condition* (4.3),

$$\begin{aligned}&\max_{0\le n\le N_T-1}(\|E^n - E_e(t_n)\|_{W'} + \|B^{n+\frac{1}{2}} - B'_e(t_{n+\frac{1}{2}})\|_W)\\ &\le \frac{K}{1-\delta}(h + c_m(\Delta t)^2)\sum_{r=1}^{2}\|(\varepsilon_r^{\frac{1}{2}}\mathbf{E}, \mu_r^{-\frac{1}{2}}\mathbf{B})\|_{H^2(0,T;W^{1,p}(\Omega_r)^3)^2}.\end{aligned} \tag{4.15}$$

Proof. The proof is based on (4.12). First, by resembling the techniques used in the proof of Theorem 4.1, we have

$$\Delta t(D\mathbf{e}(E)^k, C'\mathbf{e}(B)^{k+\frac{1}{2}}) \le \Delta t c_m \frac{M_2 M_3^{\frac{1}{2}}}{\min(h)}(\|\mathbf{e}(B)^{k+\frac{1}{2}}\|_W^2 + \|\mathbf{e}(E)^k\|_{W'}^2).$$

Second, by integrating in time and the definition of the W'-norm, we have

$$\|(E'_f(t_{n+1}) - E_e(t_{n+1})) - (E'_f(t_n) - E_e(t_n))\|_{W'}^2 \le \Delta t \int_{n\Delta t}^{(n+1)\Delta t} \|\dot{E}'_f - \dot{E}_e\|_{W'}^2\, ds.$$

Using Theorem 5.1 of [6],

$$\begin{aligned}&((E'_f(t_{n+1}) - E_e(t_{n+1})) - (E'_f(t_n) - E_e(t_n)), \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'}\\ &\le Kh\sqrt{\Delta t}\max_{0\le n\le k}\|\mathbf{e}(E)^n\|_{W'}\sum_{r=1}^{2}\|\varepsilon_r^{\frac{1}{2}}\dot{\mathbf{E}}\|_{L^2(n\Delta t,(n+1)\Delta t;W^{1,p}(\Omega_r)^3)}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&((B_f(t_{n+\frac{3}{2}}) - B'_e(t_{n+\frac{3}{2}}) - (B_f(t_{n+\frac{1}{2}}) - B'_e(t_{n+\frac{1}{2}})), \mathbf{e}(B)^{n+\frac{1}{2}} + \mathbf{e}(B)^{n+\frac{3}{2}})_W\\ &\le Kh\sqrt{\Delta t}\max_{0\le n\le k}\|\mathbf{e}(B)^{n+\frac{1}{2}}\|_W\sum_{r=1}^{2}\|\mu_r^{-\frac{1}{2}}\dot{\mathbf{B}}\|_{L^2((n+\frac{1}{2})\Delta t,(n+\frac{3}{2})\Delta t;W^{1,p}(\Omega_r)^3)}.\end{aligned}$$

Third, by the Cauchy–Schwarz inequality,

$$\begin{aligned}(P_1^{n+\frac{1}{2}}, D(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})) &\le \|S^{'-1}P_1^{n+\frac{1}{2}}\|_{W'}\|\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1}\|_{W'},\\ (P_2^{n+1}, D'(\mathbf{e}(B)^{n+\frac{1}{2}} + \mathbf{e}(B)^{n+\frac{3}{2}})) &\le \|S^{-1}P_2^{n+1}\|_W\|\mathbf{e}(B)^{n+\frac{1}{2}} + \mathbf{e}(B)^{n+\frac{3}{2}}\|_W.\end{aligned}$$

Now, by using (4.13)–(4.14) and collecting all of the above results, we have shown the desired convergence estimate. □

4.2. Implicit scheme. In this section, we will consider an implicit scheme for (3.1)–(3.2). We will apply a standard Crank–Nicolson time discretization. The approximate solutions of $\mathbf{E}(t)$ and $\mathbf{B}(t)$ will be obtained at t_n and denoted by E^n and

B^n, respectively. Given (E^n, B^n), the next approximation (E^{n+1}, B^{n+1}) will be obtained by solving

$$S'(E^{n+1} - E^n) - \frac{\Delta t}{2}(C'B^n + C'B^{n+1}) = \tilde{J}^{n+\frac{1}{2}}, \tag{4.16}$$

$$S(B^{n+1} - B^n) + \frac{\Delta t}{2}(CE^n + CE^{n+1}) = 0. \tag{4.17}$$

That is, in each time step, we need to solve a linear system $Ax = b$ for x with

$$A := \begin{pmatrix} S' & -\frac{\Delta t}{2}C' \\ \frac{\Delta t}{2}C & S \end{pmatrix} \quad \text{and} \quad b := \begin{pmatrix} \tilde{J}^{n+\frac{1}{2}} + S'E^n + \frac{\Delta t}{2}C'B^n \\ SB^n - \frac{\Delta t}{2}CE^n \end{pmatrix}. \tag{4.18}$$

The question of well-posedness of problem (4.16)–(4.17) will be given by the following theorem. We emphasize here that the scheme is unconditionally stable with respect to the time step Δt.

THEOREM 4.4. *The implicit fully discrete scheme* (4.16)–(4.17) *is well-posed. Moreover, the following stability inequality holds for* $1 \leq n \leq N_T$*:*

$$\|E^n\|^2_{W'} + \|B^n\|^2_W \leq 2\|E^0\|^2_{W'} + 2\|B^0\|^2_W + 4T\int_0^T \|S^{'-1}\tilde{J}(t)\|^2_{W'}\, dt. \tag{4.19}$$

Proof. We first show that the system (4.16)–(4.17) has a unique solution. To do this, we prove that all eigenvalues of the matrix A in (4.18) are nonzero. We rewrite A as

$$A = \begin{pmatrix} S' & 0 \\ 0 & S \end{pmatrix} + \frac{\Delta t}{2}\begin{pmatrix} 0 & -C' \\ C & 0 \end{pmatrix}.$$

By the definitions of C and C', it suffices to consider the matrix

$$A_1 := \begin{pmatrix} S'D^{-1} & -\frac{\Delta t}{2}G^T \\ \frac{\Delta t}{2}G & SD^{'-1} \end{pmatrix}.$$

Let $x = (x_1 \; x_2)$, where $x_1 \in \mathbb{C}^{M_1}$ and $x_2 \in \mathbb{C}^{F_1}$, be an eigenvector of A_1 of unit length. Then

$$x^* A_1 x = (x_1)^* S'D^{-1}x_1 + (x_2)^* SD^{'-1}x_2 + \Delta t\, \mathrm{Im}((x_2)^* G x_1) i,$$

where $*$ denotes the conjugate transpose and $i = \sqrt{-1}$. Hence, all eigenvalues of A_1 have positive real part, which shows that the matrix A_1 is invertible.

To prove the stability estimate, we multiply (4.16) by $D(E^{n+1} + E^n)$ and (4.17) by $D'(B^{n+1} + B^n)$ and add up the two resulting equations using (2.1) to get

$$\|E^{n+1}\|^2_{W'} + \|B^{n+1}\|^2_W = \|E^n\|^2_{W'} + \|B^n\|^2_W + (\tilde{J}^{n+\frac{1}{2}}, D(E^{n+1} + E^n)). \tag{4.20}$$

By induction, we have, for any $0 \leq n \leq N_T$,

$$\|E^n\|^2_{W'} + \|B^n\|^2_W = \|E^0\|^2_{W'} + \|B^0\|^2_W + \sum_{k=1}^{n}(\tilde{J}^{k-\frac{1}{2}}, D(E^k + E^{k-1})).$$

Now, (4.19) follows from a standard application of Cauchy–Schwarz inequality. □

If we compare (4.5) and (4.20), we see that the appearance of terms of the form $\Delta t(DE^k, C'B^{k+\frac{1}{2}})$ makes the explicit scheme (4.1)–(4.2) conditionally stable, while the absence of these terms makes the implicit scheme (4.16)–(4.17) unconditionally stable.

We are now in a position to study the convergence theory of the method (4.16)–(4.17). Applying the same technique used to prove (4.12), we have

$$\begin{aligned}
&\|\mathbf{e}(E)^k\|_{W'}^2 + \|\mathbf{e}(B)^k\|_W^2 \\
&= \sum_{n=0}^{k-1} \Big\{((E'_f(t_{n+1}) - E_e(t_{n+1})) - (E'_f(t_n) - E_e(t_n)), \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'} \\
&\quad + ((B'_f(t_{n+1}) - B_e(t_{n+1})) - (B'_f(t_n) - B_e(t_n)), \mathbf{e}(B)^n + \mathbf{e}(B)^{n+1})_W \\
&\quad + (Q_1^{n+\frac{1}{2}}, D(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})) + (Q_2^{n+\frac{1}{2}}, D'(\mathbf{e}(B)^n + \mathbf{e}(B)^{n+1}))\Big\},
\end{aligned} \tag{4.21}$$

where

$$Q_1^{n+\frac{1}{2}} := \tilde{J}^{n+\frac{1}{2}} - S'(E'_f(t_{n+1}) - E'_f(t_n)) + \frac{\Delta t}{2}(C'B'_e(t_n) + C'B'_e(t_{n+1})), \tag{4.22}$$

$$Q_2^{n+\frac{1}{2}} := -S(B_f(t_{n+1}) - B_f(t_n)) - \frac{\Delta t}{2}(CE_e(t_n) + CE_e(t_{n+1})). \tag{4.23}$$

Similar to Lemma 4.2, we have the following consistency error estimate of the Crank–Nicolson time discretization of (4.16)–(4.17). The proof is similar to that of Lemma 4.2.

LEMMA 4.5. *Assume that* $(\mathbf{B}, \mathbf{E}) \in H^2(0,T; W^{1,p}(\Omega_r)^3)^2$ *for* $2 < p \leq 3$ *and* $r = 1, 2$. *Then on general unstructured grids,*

$$\sum_{n=0}^{k-1} \|S^{'-1} Q_1^{n+\frac{1}{2}}\|_{W'} \leq c_m K (\Delta t)^2 \sum_{r=1}^{2} \|\varepsilon_r^{\frac{1}{2}} \mathbf{E}\|_{H^2(0,T;W^{1,p}(\Omega_r)^3)}, \tag{4.24}$$

$$\sum_{n=0}^{k-1} \|S^{-1} Q_2^{n+\frac{1}{2}}\|_{W} \leq c_m K (\Delta t)^2 \sum_{r=1}^{2} \|\mu_r^{-\frac{1}{2}} \mathbf{B}\|_{H^2(0,T;W^{1,p}(\Omega_r)^3)}. \tag{4.25}$$

The following theorem gives the main result in this section. It can be proved by using a technique similar to that used to prove Theorem 4.3.

THEOREM 4.6. *Assume that* $(\mathbf{E}, \mathbf{B}) \in H^2(0,T; W^{1,p}(\Omega_r)^3)^2$, $2 < p \leq 3$, $r = 1, 2$, *is the solution to* (1.1)–(1.2) *and* (E^n, B^n) *is the solution to the implicit fully discrete scheme* (4.16)–(4.17). *Then*

$$\begin{aligned}
&\max_{0 \leq n \leq N_T} (\|E^n - E_e(t_n)\|_{W'} + \|B^n - B'_e(t_n)\|_W) \\
&\leq K(h + c_m(\Delta t)^2) \sum_{r=1}^{2} \|(\varepsilon_r^{\frac{1}{2}} \mathbf{E}, \mu_r^{-\frac{1}{2}} \mathbf{B})\|_{H^2(0,T;W^{1,p}(\Omega_r)^3)^2}.
\end{aligned} \tag{4.26}$$

5. Analysis on rectangular grids. In this section, we will consider the explicit finite volume scheme (4.1)–(4.2) and the implicit finite volume scheme (4.16)–(4.17) on rectangular grids. It is clear that all primal and dual elements are cuboids and all faces are rectangles. All definitions that we made in previous sections can be made in exactly the same way on nonuniform rectangular grids. For instance, $M_3 = 4$ for rectangular grids. It can be shown that the two schemes (4.1)–(4.2) and (4.16)–(4.17) are second order convergent in space. The second order convergence comes from the fact that the circumcenter of a cuboid is also its barycenter. In section 5.2, we give a one-dimensional counterexample to show that without taking the barycenter as dual node the scheme may reduce to first order.

5.1. Convergence analysis. We will first consider convergence analysis on the explicit fully discrete scheme (4.1)–(4.2) for nonuniform rectangular grids. It is easy to show that

$$\begin{aligned}
&\|\mathbf{e}(E)^k\|_{W'}^2 + \|\mathbf{f}(B)^{k+\frac{1}{2}}\|_W^2 \\
= &-\Delta t(D\mathbf{e}(E)^k, C'\mathbf{e}(B)^{k+\frac{1}{2}}) \\
&+ \sum_{n=0}^{k-1}\Big\{((E'_f(t_{n+1}) - E_e(t_{n+1})) - (E'_f(t_n) - E_e(t_n)), \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'} \\
&- (\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}, (B_f(t_{n+\frac{3}{2}}) - B'_e(t_{n+\frac{3}{2}})) + (B_f(t_{n+\frac{1}{2}}) - B'_e(t_{n+\frac{1}{2}})))_W \\
&+ (P_1^{n+\frac{1}{2}}, D(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})) + (P_2^{n+1}, D'(\mathbf{e}(B)^{n+\frac{1}{2}} + \mathbf{e}(B)^{n+\frac{3}{2}}))\Big\}.
\end{aligned} \tag{5.1}$$

The following theorem gives the convergence result for the explicit finite volume scheme (4.1)–(4.2) on nonuniform rectangular grids.

Theorem 5.1. *Assume that*

$$\begin{aligned}
&\mathbf{B}(\mathbf{x}, t) \in H^1(0, T; H^3(\Omega_r)^3), \\
&\mathbf{E}(\mathbf{x}, t) \in H^1(0, T; H^3(\Omega_r)^3) \cap H^2(0, T; W^{2,p}(\Omega_r)^3),
\end{aligned}$$

for $p > 3$ *and* $r = 1, 2$, *is the solution to* (1.1)–(1.2) *and* $(E^n, B^{n+\frac{1}{2}})$ *is the solution to the explicit fully discrete scheme* (4.1)–(4.2) *on nonuniform rectangular grids. Then, under the stability condition* $\delta < \frac{1}{2}$,

$$\begin{aligned}
&\max_{0\le n\le N_T-1}(\|E^n - E_e(t_n)\|_{W'} + \|B^{n+\frac{1}{2}} - B_f(t_{n+\frac{1}{2}})\|_W) \\
&\le \frac{K}{1-\delta}(h^2 + c_m(\Delta t)^2) \\
&\times \sum_{r=1}^{2}\Big\{\|(\varepsilon_r^{\frac{1}{2}}\mathbf{E}, \mu_r^{-\frac{1}{2}}\mathbf{B})\|_{H^1(0,T;H^3(\Omega_r)^3)} + \|\varepsilon_r^{\frac{1}{2}}\mathbf{E}\|_{H^2(0,T;W^{2,p}(\Omega_r)^3)}\Big\}.
\end{aligned} \tag{5.2}$$

Proof. The proof is based on (5.1).

(i) To begin, notice that the first term on the right-hand side of (5.1) can be estimated as follows:

$$\begin{aligned}
&\Delta t(D\mathbf{e}(E)^k, C'\mathbf{e}(B)^{k+\frac{1}{2}}) \\
&= \Delta t(D\mathbf{e}(E)^k, C'\mathbf{f}(B)^{k+\frac{1}{2}}) + \Delta t(D\mathbf{e}(E)^k, C'(B_f(t_{k+\frac{1}{2}}) - B'_e(t_{k+\frac{1}{2}}))).
\end{aligned}$$

Applying the technique used in proving Theorem 4.1, we have

$$\Delta t(D\mathbf{e}(E)^k, C'\mathbf{f}(B)^{k+\frac{1}{2}}) \le \Delta t c_m \frac{M_2 M_3^{\frac{1}{2}}}{\min(h)}(\|\mathbf{e}(E)^k\|_{W'}^2 + \|\mathbf{f}(B)^{k+\frac{1}{2}}\|_W^2).$$

Applying (2.1) and [6, Lemma 5.2], we have

$$\begin{aligned}&\Delta t(D\mathbf{e}(E)^k, C'(B_f(t_{k+\frac{1}{2}}) - B'_e(t_{k+\frac{1}{2}})))\\ &= \Delta t(C\mathbf{e}(E)^k, D'(B_f(t_{k+\frac{1}{2}}) - B'_e(t_{k+\frac{1}{2}})))\\ &= \Delta t(C\mathbf{e}(E)^k, D'u^{k+\frac{1}{2}}) + \Delta t(C\mathbf{e}(E)^k, \xi^{k+\frac{1}{2}})\\ &= \Delta t(D\mathbf{e}(E)^k, C'u^{k+\frac{1}{2}}) + \Delta t(D\mathbf{e}(E)^k, C'D^{'-1}\xi^{k+\frac{1}{2}}),\end{aligned}$$

where $\|u^{k+\frac{1}{2}}\|_W$ and $\|D^{'-1}\xi^{k+\frac{1}{2}}\|_W$ are $O(h^2)$. So, we obtain

$$\begin{aligned}&\Delta t(D\mathbf{e}(E)^k, C'(B_f(t_{k+\frac{1}{2}}) - B'_e(t_{k+\frac{1}{2}})))\\ &\le \Delta t c_m \frac{M_2 M_3^{\frac{1}{2}}}{\min(h)}(2\|\mathbf{e}(E)^k\|^2_{W'} + \|u^{k+\frac{1}{2}}\|^2_W + \|D^{'-1}\xi^{k+\frac{1}{2}}\|^2_W).\end{aligned}$$

(ii) By Lemma 5.3 in [6] with $\phi = \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1}$, we have

$$\begin{aligned}&(\dot{E}'_f(t_n) - \dot{E}_e(t_n), \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'}\\ &= (\dot{v}, \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'} + (D'\dot{w}, C(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1}))\\ &\quad + (S^{'-1}\dot{\lambda}, \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'}.\end{aligned}$$

So, we have

$$\begin{aligned}&((E'_f(t_{n+1}) - E_e(t_{n+1})) - (E'_f(t_n) - E_e(t_n)), \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'}\\ &= \int_{n\Delta t}^{(n+1)\Delta t} \Big\{(\dot{v}, \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'} + (D'\dot{w}, C(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1}))\\ &\quad + (S^{'-1}\dot{\lambda}, \mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})_{W'}\Big\}\, ds\\ &\le 2(\max_{0\le n\le k}\|\mathbf{e}(E)^n\|_{W'}) \times \int_{n\Delta t}^{(n+1)\Delta t} \|\dot{v} + S^{'-1}\dot{\lambda}\, ds\|_{W'}\\ &\quad + \int_{n\Delta t}^{(n+1)\Delta t} (D'\dot{w}, C(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1}))\, ds.\end{aligned}$$

By the definition of the W'-norm and the estimate of $\|\dot{v}\|_{W'}$ from [6],

$$\begin{aligned}\int_{n\Delta t}^{(n+1)\Delta t} \|\dot{v}\, ds\|^2_{W'} &\le \Delta t \int_{n\Delta t}^{(n+1)\Delta t} \|\dot{v}\|^2_{W'}\, ds\\ &\le K h^4 \Delta t \sum_{r=1}^{2} \|\varepsilon_r^{\frac{1}{2}}\dot{\mathbf{E}}\|^2_{L^2(n\Delta t,(n+1)\Delta t;H^3(\Omega_r)^3)}.\end{aligned}$$

Similarly, by the definition of the W'-norm and the estimate of $\|S^{'-1}\dot{\lambda}\|_{W'}$ from [6], we have

$$\int_{n\Delta t}^{(n+1)\Delta t} \|S^{'-1}\dot{\lambda}\, ds\|^2_{W'} \le K h^4 \Delta t \sum_{r=1}^{2} \|\varepsilon_r^{\frac{1}{2}}\dot{\mathbf{E}}\|^2_{L^2(n\Delta t,(n+1)\Delta t;H^3(\Omega_r)^3)}.$$

Using (4.9) and the mean value theorem for integral, we have

$$\begin{aligned}
&\int_{n\Delta t}^{(n+1)\Delta t} (D'\dot{w}, C(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1}))\, ds \\
&= -\frac{1}{\Delta t}\int_{n\Delta t}^{(n+1)\Delta t} \Big\{(D'\dot{w}, S(\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n-\frac{1}{2}})) - (D'\dot{w}, P_2^n + P_2^{n+1})\Big\}\, ds \\
&= -(D'\dot{w}(\eta^{n+\frac{1}{2}}), S(\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n-\frac{1}{2}})) + \frac{1}{\Delta t}\int_{n\Delta t}^{(n+1)\Delta t} (D'\dot{w}, P_2^n + P_2^{n+1})\, ds,
\end{aligned}$$

where $t_n < \eta^{n+\frac{1}{2}} < t_{n+1}$. Applying summation by parts,

$$\begin{aligned}
&\sum_{n=0}^{k-1}(D'\dot{w}(\eta^{n+\frac{1}{2}}), S(\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n-\frac{1}{2}})) \\
&= (D'\dot{w}(\eta^{k-\frac{3}{2}}), S\mathbf{f}(B)^{k-\frac{1}{2}}) + (D'\dot{w}(\eta^{k-\frac{1}{2}}), S\mathbf{f}(B)^{k+\frac{1}{2}}) \\
&\quad - \sum_{n=1}^{k-2}(D'(\dot{w}(\eta^{n+\frac{3}{2}}) - \dot{w}(\eta^{n-\frac{1}{2}})), S\mathbf{f}(B)^{n+\frac{1}{2}}).
\end{aligned}$$

The first two terms on the right-hand side of the previous equation can be bounded by using the estimate of $\|\dot{w}\|_W$ from [6]. For the third term, we notice that

$$|\dot{w}(\eta^{n+\frac{3}{2}}) - \dot{w}(\eta^{n-\frac{1}{2}})|^2 \leq 3\Delta t \int_{(n-1)\Delta t}^{(n+2)\Delta t} |\ddot{w}|^2\, ds.$$

By the definition of the W-norm and the estimate of $\|\dot{w}\|_W$ from [6],

$$\|\dot{w}(\eta^{n+\frac{3}{2}}) - \dot{w}(\eta^{n-\frac{1}{2}})\|_W^2 \leq Kh^4\Delta t \sum_{r=1}^{2} \|\varepsilon_r^{\frac{1}{2}}\ddot{\mathbf{E}}\|^2_{L^2((n-1)\Delta t,(n+2)\Delta t;W^{2,p}(\Omega_r)^3)}.$$

By the Sobolev embedding theorem and the estimate of $\|\dot{w}\|_W$ from [6], we get

$$\begin{aligned}
&\frac{1}{\Delta t}\int_{n\Delta t}^{(n+1)\Delta t} (D'\dot{w}, P_2^n + P_2^{n+1})\, ds \\
&\leq \max_{0\leq s\leq T} \|\dot{w}(s)\|_W \|S^{-1}(P_2^n + P_2^{n+1})\|_W \\
&\leq Kh^2 \|S^{-1}(P_2^n + P_2^{n+1})\|_W \sum_{r=1}^{2} \|\varepsilon_r^{\frac{1}{2}}\mathbf{E}\|_{H^2(0,T;W^{2,p}(\Omega_r)^3)},
\end{aligned}$$

where the last term can be estimated by (4.14).

(iii) Using (4.9) and (4.11), we have

$$\begin{aligned}
&S(\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}) \\
&= -\Delta t C\mathbf{e}(E)^{n+1} + P_2^{n+1} \\
&= -\Delta t C\mathbf{e}(E)^{n+1} - \int_{(n+\frac{1}{2})\Delta t}^{(n+\frac{3}{2})\Delta t} S\dot{B}_f\, ds - \Delta t CE_e(t_{n+1}).
\end{aligned}$$

By (3.3), we have $S(\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}) = C\phi$ with

$$\phi = -\Delta t\mathbf{e}(E)^{n+1} + \int_{(n+\frac{1}{2})\Delta t}^{(n+\frac{3}{2})\Delta t} E_e(s)\, ds - \Delta t E_e(t_{n+1}).$$

Hence, by Lemma 5.2 from [6], we have

$$\begin{aligned}&(\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}, B_f(t_{n+\frac{3}{2}}) - B'_e(t_{n+\frac{3}{2}}))_W \\ &= (\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}, u^{n+\frac{3}{2}})_W + (\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}, D^{'-1}\xi^{n+\frac{3}{2}})_W.\end{aligned}$$

Applying summation by parts, we obtain

$$\begin{aligned}&\sum_{n=0}^{k-1} (\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}, u^{n+\frac{3}{2}})_W \\ &\quad = (\mathbf{f}(B)^{k+\frac{1}{2}}, u^{k+\frac{1}{2}})_W - \sum_{n=1}^{k-1} (\mathbf{f}(B)^{n+\frac{1}{2}}, u^{n+\frac{3}{2}} - u^{n+\frac{1}{2}})_W.\end{aligned}$$

We remark that a similar result holds for $(\mathbf{f}(B)^{n+\frac{3}{2}} - \mathbf{f}(B)^{n+\frac{1}{2}}, D^{'-1}\xi^{n+\frac{3}{2}})_W$.

(iv) We have

$$(P_1^{n+\frac{1}{2}}, D(\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1})) \leq \|S^{'-1}P_1^{n+\frac{1}{2}}\|_{W'}\|\mathbf{e}(E)^n + \mathbf{e}(E)^{n+1}\|_{W'},$$

which can be estimated by using (4.13).

(v) Notice that

$$\begin{aligned}&(P_2^{n+1}, D'(\mathbf{e}(B)^{n+\frac{1}{2}} + \mathbf{e}(B)^{n+\frac{3}{2}})) \\ &= (P_2^{n+1}, D'(\mathbf{f}(B)^{n+\frac{1}{2}} + \mathbf{f}(B)^{n+\frac{3}{2}})) \\ &\quad + (P_2^{n+1}, D'((B_f(t_{n+\frac{3}{2}}) - B'_e(t_{n+\frac{3}{2}})) + (B_f(t_{n+\frac{1}{2}}) - B'_e(t_{n+\frac{1}{2}})))).\end{aligned}$$

The first term on the right-hand side can be estimated by using (4.14). Notice that P_2^{n+1} can be written as $C\tilde{\phi}$ for some $\tilde{\phi}$. So, by Lemma 5.2 from [6], we have

$$(P_2^{n+1}, D'(B_f(t_{n+\frac{3}{2}}) - B'_e(t_{n+\frac{3}{2}}))) = (P_2^{n+1}, D'u^{n+\frac{3}{2}}) + (P_2^{n+1}, \xi^{n+\frac{3}{2}}),$$

which can be estimated by using (4.14) and the estimates of $\|u\|_W$ and $\|D^{'-1}\xi\|_W$ from [6]. □

Finally, we will state, without proof, the convergence estimate for the implicit fully discrete scheme (4.16)–(4.17) for nonuniform rectangular grids.

THEOREM 5.2. *Assume that*

$$\begin{aligned}&\mathbf{B}(\mathbf{x}, t) \in H^1(0, T; H^3(\Omega_r)^3), \\ &\mathbf{E}(\mathbf{x}, t) \in H^1(0, T; H^3(\Omega_r)^3) \cap H^2(0, T; W^{2,p}(\Omega_r)^3),\end{aligned}$$

for $p > 3$ *and* $r = 1, 2$, *is the solution to* (1.1)–(1.2) *and* (E^n, B^n) *is the solution to the implicit fully discrete scheme* (4.16)–(4.17) *on nonuniform rectangular grids. Then*

$$\begin{aligned}&\max_{0\leq n\leq N_T} (\|E^n - E_e(t_n)\|_{W'} + \|B^n - B_f(t_n)\|_W) \\ &\leq K(h^2 + c_m(\Delta t)^2) \\ &\quad \times \sum_{r=1}^{2} \Big\{ \|(\varepsilon_r^{\frac{1}{2}}\mathbf{E}, \mu_r^{-\frac{1}{2}}\mathbf{B})\|_{H^1(0,T;H^3(\Omega_r)^3)} + \|\varepsilon_r^{\frac{1}{2}}\mathbf{E}\|_{H^2(0,T;W^{2,p}(\Omega_r)^3)} \Big\}.\end{aligned} \tag{5.3}$$

5.2. A counterexample. We see from the previous sections that the spatial convergence of the FVM on rectangular grids is one order higher than that on general unstructured mesh. One main difference between these two grids is that the circumcenter of a rectangle coincides with its barycenter. Without choosing the barycenter as dual node, the scheme remains first order accurate. We will show this by considering the following one-dimensional example. Consider the system

$$\begin{aligned} u_t &= v_x, \\ v_t &= u_x, \end{aligned}$$

with initial conditions

$$u(x,0) = 0, \quad v(x,0) = \frac{1}{2}x^2.$$

The exact solution to this problem is given by

$$u(x,t) = xt, \quad v(x,t) = \frac{1}{2}(x^2 + t^2).$$

Now, we consider a uniform grid with mesh size h. A dual node within each primal interval is chosen such that the distance between the left end point and the dual node is αh for fixed $0 < \alpha < 1$. Therefore, the distance between the right end point and the dual node is $(1-\alpha)h$. Also, the distance between two consecutive dual nodes is h. We consider the semidiscrete scheme

$$\begin{aligned} \frac{d}{dt}u_j &= \frac{v'_{j+1} - v'_j}{h}, \\ \frac{d}{dt}v'_j &= \frac{u_j - u_{j-1}}{h}, \end{aligned}$$

where u_j is defined corresponding to the primal node x_j, while v'_j is defined corresponding to the dual node x'_j. By a direct computation, it can be shown that the following is a solution to the semidiscrete scheme:

$$\begin{aligned} u_j &= x_j t + (\alpha^2 - (1-\alpha)^2)ht, \\ v'_j &= \frac{1}{2}((x'_j)^2 + t^2), \end{aligned}$$

resulting in an $O(h)$ error.

REFERENCES

[1] A. Chatterjee, L. C. Kempel, and J. L. Volakis, *Finite Element Method for Electromagnetics: Antennas, Microwave Circuits, and Scattering Applications*, IEEE Press, New York, 1998.

[2] J. S. Chen and K. S. Yee, *The finite-difference time-domain and the finite-volume time-domain methods in solving Maxwell's equations*, IEEE Trans. Antennas Propagat., 45 (1997), pp. 354–363.

[3] Z. Chen, Q. Du, and J. Zou, *Finite element methods with matching and nonmatching meshes for Maxwell equations with discontinuous coefficients*, SIAM J. Numer. Anal., 37 (2000), pp. 1542–1570.

[4] T. S. Chung and J. Zou, *A finite volume method for Maxwell's equations with discontinuous physical coefficients*, Int. J. Appl. Math., 7 (2001), pp. 201–224.

[5] E. T. Chung and J. Zou, *The eigenvalues and eigenspaces of some discrete div- and curl-related operators*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 1149–1160.

[6] E. T. Chung, Q. Du, and J. Zou, *Convergence analysis on a finite volume method for Maxwell's equations in nonhomogeneous media*, SIAM J. Numer. Anal., 41 (2003), pp. 37–63.

[7] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978; reprinted as Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[8] S. Fortune, *Voronoi diagrams and Delaunay triangulations*, in Computing in Euclidean Geometry, World Scientific, Singapore, 1992, pp. 193–233.

[9] J. Jin, *The Finite Element Method in Electromagnetics*, John Wiley, New York, 2002.

[10] N. K. Madsen, *Divergence preserving discrete surface integral methods for Maxwell's curl equations using non-orthogonal unstructured grids*, J. Comput. Phys., 119 (1995), pp. 34–45.

[11] P. Monk and E. Süli, *A convergence analysis of Yee's scheme on nonuniform grids*, SIAM J. Numer. Anal., 31 (1994), pp. 393–412.

[12] R. A. Nicolaides, *Direct discretization of planar div-curl problems*, SIAM J. Numer. Anal., 29 (1992), pp. 32–56.

[13] R. A. Nicolaides and D. Q. Wang, *Convergence analysis of a covolume scheme for Maxwell's equations in three dimensions*, Math. Comp., 67 (1998), pp. 947–963.

[14] R. A. Nicolaides and X. Wu, *Covolume solutions of three-dimensional div-curl equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2195–2203.

[15] A. Taflove, *Computational Electrodynamics*, Artech House, Boston, MA, 1995.

[16] K. S. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. Antennas Propagat., 14 (1966), pp. 302–307.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 318–339

A LEAST-SQUARES MIXED FINITE ELEMENT METHOD FOR BIOT'S CONSOLIDATION PROBLEM IN POROUS MEDIA[*]

JOHANNES KORSAWE[†] AND GERHARD STARKE[†]

Abstract. A least-squares mixed finite element method for the coupled problem of flow and deformation is presented and analyzed in this paper. For the analysis, we restrict ourselves to fully saturated conditions for the flow process and to a linearly elastic material law for the deformation process. This is known in the literature as Biot's consolidation problem. For simplicity, the analysis is presented for the problem in two space dimensions. Our least-squares approach is motivated by the fact that all process variables, i.e., fluid pressure and flux as well as displacement field and stress tensor, are approximated directly by suitable finite element spaces. Ellipticity of the corresponding variational formulation is proven for the stationary case as well as for the subproblems arising at each step of an implicit time discretization in the general time-dependent case. Standard H^1-conforming piecewise linear and quadratic finite elements are used for the fluid pressure and for (each component of) the displacement, respectively. For the flux and stress components, the $H(\text{div})$-conforming Raviart–Thomas spaces (of lowest order) are used. Computational results are presented for some two-dimensional test problems.

Key words. Biot's consolidation, least-squares finite element method, porous media

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/S0036142903432929

1. Introduction.

In recent years, a lot of effort has been dedicated to the theoretical and numerical treatment of models for fluid flow and deformation in porous media. Our model will be based on the classical phenomenological approach by Biot [2] using the concept of effective stresses. Modern formulations based on multiphase mixture theories were developed in the last 30 years; see the monograph by de Boer [12] for an overview. Much of current research concerned with the numerical treatment of such coupled problems is devoted to the extension to two-phase flow and elastoplastic deformation models. The classical Biot consolidation problem assumes a fully saturated porous medium and a linearly elastic material law leading to a linear parabolic system. We restrict our attention to this linear problem since our aim is to analyze a least-squares mixed finite element method in the simplest possible situation. The novelty of our least-squares approach is the introduction of approximation spaces for the fluid flux and the stress tensor in addition to the primary variables fluid pressure and displacement field.

The numerical treatment of Biot's consolidation problem by the Taylor–Hood finite element spaces was studied by Murad and Loula in [16] and [17]. This work was continued with a detailed analytical investigation in the contribution by Murad, Thomée, and Loula [18]. A general reference for the use of the finite element method for the numerical simulation of fluid flow and deformation processes in porous media is the monograph by Lewis and Schrefler [15]. A nonlinearly elastic material law for fluid-saturated porous solids is considered in Ehlers and Eipper [13]. Wang and Kolditz [21] investigate the numerical treatment of elastoplastic material behavior

[*] Received by the editors August 7, 2003; accepted for publication (in revised form) August 17, 2004; published electronically May 27, 2005.
http://www.siam.org/journals/sinum/43-1/43292.html

[†] Institut für Angewandte Mathematik, Universität Hannover, Welfengarten 1, 30167 Hannover, Germany (jkorsawe@ifam.uni-hannover.de, starke@ifam.uni-hannover.de). The work of the first author was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant STA 402/8.

using a Drucker–Prager model for some two-dimensional plane strain problems under fully saturated conditions. Wieners et al. [22] consider saturated porous medium flow in three spatial dimensions where the material behavior of the porous skeleton is assumed to be elastoviscoplastic. Taylor–Hood elements are used in [22] for the approximation of the displacement field and fluid pressure.

Our purpose in this paper is the derivation and study of a least-squares finite element approach to the coupled mechanical and flow problem. The least-squares approach introduces finite element spaces for the approximation of all the process variables involved in our model. In the case of fluid flow in deformable porous media these consist of fluid pressure and flux as well as displacement field and stress tensor. At first sight this appears to increase the computational work by introducing more variables compared to standard approaches involving only pressure and displacements. However, the introduction of these variables does, in general, lead to significant simplifications at various stages of the solution algorithm. First, the variables coupling the flow and deformation processes are used directly to describe the problem, which makes it straightforward to derive the underlying variational formulation. In this way, all the variables of interest can be approximated directly and, for fluid flux and for the stress tensor, more accurately than by postprocessing from the results of the standard formulation. Moreover, the finite element spaces used for approximating the different process variables can be chosen independently since no inf-sup compatibility conditions are required in the least-squares finite element approach. Finally, the local evaluation of the least-squares functional provides an a posteriori error estimator at no additional cost.

The equivalence of the least-squares functional to the displacement components and fluid pressure measured in the H^1 norm and to the stress components and fluid flux measured in the $H(\mathrm{div})$ norm is the main analytical contribution of this paper. This ellipticity result ensures that the boundary value problem under consideration is well-posed and leads to finite element approximation estimates in a straightforward way. In the stationary case, the proof rests upon the results in [7] for the flow part and [9, 6] for the deformation part. For the time-dependent case a more careful analysis is carried out to establish ellipticity of the corresponding least-squares formulation. Standard H^1-conforming linear and quadratic finite elements are then used for the fluid pressure and the displacement components, respectively. The fluid flux and the stress components are approximated by the $H(\mathrm{div})$-conforming Raviart–Thomas spaces of lowest order.

For the mechanical part of the model we restrict ourselves to linear elastic material behavior. For linear elasticity a least-squares mixed finite element approach using displacement and stress as process variables has been proposed recently in [9, 10, 6]. This approach was extended to incompressible Newtonian flow in [8]. The first-order system studied in [8] has some similarities to the deformation part of the consolidation problem under investigation here. For the fluid flow part of the model we assume fully saturated conditions for our analysis. However, variably saturated porous media can also be treated using a least-squares mixed finite element approach for variably saturated subsurface flow, which was studied in [20] and [19]. The combination of these two methods for the treatment of the coupled problem of variably saturated fluid flow and mechanical deformation has been tested numerically in [14].

The remainder of this paper is organized as follows. In section 2 we present the least-squares mixed formulation of Biot's consolidation problem. The ellipticity of the bilinear form associated with the least-squares functional for the stationary case is proved in section 3. Ellipticity for the least-squares formulation resulting from an implicit Euler time discretization of the time-dependent problem is established in section 4. Section 5 provides the appropriate finite element spaces with the corresponding approximation properties and discusses the issue of error estimation. Finally, the results of computational experiments are presented in section 6.

2. Least-squares formulation of Biot's consolidation problem.

We start this section with a review of the least-squares formulation of linear elasticity due to [9] and [6]. We consider the equations of linear elasticity in the form

$$\begin{aligned} \operatorname{div} \boldsymbol{\sigma} &= \mathbf{0} \text{ in } \Omega, \\ \boldsymbol{\sigma} - \mathcal{C}\boldsymbol{\varepsilon}(\mathbf{u}) &= \mathbf{0} \text{ in } \Omega, \\ \mathbf{u} &= \mathbf{0} \text{ on } \Gamma_D^{\text{elas}}, \\ \boldsymbol{\sigma} \cdot \mathbf{n} &= \mathbf{g} \text{ on } \Gamma_N^{\text{elas}}, \end{aligned} \tag{2.1}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded domain with the boundary divided into two disjoint parts Γ_D^{elas} and Γ_N^{elas} such that $\overline{\Gamma}_D^{\text{elas}} \cup \overline{\Gamma}_N^{\text{elas}} = \partial\Omega$. For simplicity we assume that Γ_D^{elas} and Γ_N^{elas} are both of positive length. The operator $\mathcal{C}$ is the linear mapping from strains to stresses given by

$$\mathcal{C}\boldsymbol{\varepsilon} = 2\mu\, \boldsymbol{\varepsilon} + \lambda\, (\operatorname{tr} \boldsymbol{\varepsilon})\, \mathbf{I}\,,\ \boldsymbol{\varepsilon} \in \mathbb{R}^{2\times 2}, \tag{2.2}$$

with the Lamé constants λ, μ. We allow $\lambda > 0$ to be arbitrarily large; i.e., the treatment of nearly incompressible materials is possible, but we assume that μ is on the order of one. Note that this can be achieved by a suitable rescaling of the displacements $\mathbf{u}$. The strain tensor in linear elasticity is given by

$$\boldsymbol{\varepsilon}(\mathbf{u}) = \begin{bmatrix} \partial_1 u_1 & \frac{1}{2}(\partial_2 u_1 + \partial_1 u_2) \\ \frac{1}{2}(\partial_2 u_1 + \partial_1 u_2) & \partial_2 u_2 \end{bmatrix}, \tag{2.3}$$

the symmetric gradient of $\mathbf{u}$. The traction boundary conditions in (2.1) are treated by extending the boundary values $\mathbf{g}$ on Γ_N^{elas} to a function $\boldsymbol{\sigma}^N \in H(\text{div}, \Omega)^2$. Such an extension exists if we assume $\mathbf{g} \in H^{-1/2}(\Gamma_N)$ and can be constructed as follows: Denote by $\mathbf{G} \in H^{1/2}(\Gamma_N)$ the antiderivative (componentwise) of $\mathbf{g}$ along Γ_N parametrized with respect to arc length (which implies that the tangential derivative $\mathbf{n} \times \nabla \mathbf{G} = \mathbf{g}$). Using well-known lifting theorems (see, e.g., [11, section IV.4]) an extension $\boldsymbol{\Psi} \in H^1(\Omega)$ with $\boldsymbol{\Psi} = \mathbf{G}$ on Γ_N is obtained. It can easily be verified that $\boldsymbol{\sigma}^N = \nabla^\perp \mathbf{G}$ has the desired properties (note that even $\operatorname{div} \boldsymbol{\sigma}^N = 0$ holds). This leads to the problem of finding

$$\begin{aligned} \hat{\boldsymbol{\sigma}} \in H_{\Gamma_N^{\text{elas}}}(\text{div}, \Omega)^2 &= \{\boldsymbol{\tau} \in H(\text{div}, \Omega)^2 : \boldsymbol{\tau} \cdot \mathbf{n} = \mathbf{0} \text{ on } \Gamma_N^{\text{elas}}\}, \\ \mathbf{u} \in H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2 &= \{\mathbf{v} \in H^1(\Omega)^2 : \mathbf{v} = \mathbf{0} \text{ on } \Gamma_D^{\text{elas}}\} \end{aligned}$$

with $\boldsymbol{\sigma} = \boldsymbol{\sigma}^N + \hat{\boldsymbol{\sigma}}$ and $\mathbf{u}$ satisfying the system (2.1).

From now on, the standard norm on $L^2(\Omega)$ (or $L^2(\Omega)^2$, $L^2(\Omega)^{2\times 2}$, respectively) is abbreviated by $\|\cdot\|$. Introducing the least-squares functional

$$\mathcal{F}_{\text{elas}}(\boldsymbol{\sigma}, \mathbf{u}) = \|\operatorname{div} \boldsymbol{\sigma}\|^2 + \|\mathcal{C}^{-1/2}\boldsymbol{\sigma} - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u})\|^2\,, \tag{2.4}$$

our aim is then to minimize $\mathcal{F}_{\text{elas}}(\boldsymbol{\sigma}, \mathbf{u})$ among all $\boldsymbol{\sigma} = \boldsymbol{\sigma}^N + \hat{\boldsymbol{\sigma}}$ with $\hat{\boldsymbol{\sigma}} \in H_{\Gamma_N^{\text{elas}}}(\text{div}, \Omega)^2$ and $\mathbf{u} \in H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2$. The scaling by $\mathcal{C}^{1/2}$ is necessary for the proper balance between both terms of the functional in the nearly incompressible case. The theoretical foundation of this choice of scaling is given in [9] and [6].

The fully saturated flow problem in the absence of deformation reads

$$\begin{aligned} \operatorname{div}\mathbf{w} &= 0 \text{ in } \Omega, \\ \mathbf{w} + \kappa\,\nabla p &= \mathbf{0} \text{ in } \Omega, \\ p &= p^D \text{ on } \Gamma_D^{\text{flow}}, \\ \mathbf{w}\cdot\mathbf{n} &= 0 \text{ on } \Gamma_N^{\text{flow}}. \end{aligned} \tag{2.5}$$

The splitting of the boundary $\partial\Omega = \overline{\Gamma}_D^{\text{flow}} \cup \overline{\Gamma}_N^{\text{flow}}$ into a Dirichlet and Neumann part will, in general, not be the same as for the elasticity problem. We assume, however, that Γ_D^{flow} has positive length. In (2.5), $\mathbf{w}$ denotes the volumetric flux of the fluid and p the pressure potential in Ω. Note that, in this definition, the true pressure in the porous medium may be recovered from p by adding the effect of the fluid weight due to gravity. Moreover, the diffusion parameter κ denotes the quotient of permeability and fluid viscosity. Assuming $p^D \in H^{1/2}(\Gamma_D)$, we may extend these boundary values to a function in $H^1(\Omega)$ also denoted by p^D (see again [11, section IV.4]). Our aim is to find $\mathbf{w} \in H_{\Gamma_N^{\text{flow}}}(\text{div}, \Omega)$ and $p = p^D + \hat{p}$ with $\hat{p} \in H^1_{\Gamma_D^{\text{flow}}}(\Omega)$ such that the least-squares functional

$$\mathcal{F}_{\text{flow}}(\mathbf{w}, p) = \|\text{div}\,\mathbf{w}\|^2 + \|\mathbf{w} + \kappa\nabla p\|^2 \tag{2.6}$$

is minimized. Here we assume that the percolation parameter κ is on the order of one, which can be achieved by a suitable rescaling of the pressure p. κ may vary in the domain Ω as long as $\underline{\kappa} \le \kappa \le \overline{\kappa}$ holds with positive constants $\underline{\kappa}, \overline{\kappa}$.

The consolidation process is described by a model which couples flow and deformation. The system (2.1), modeling the elastic deformation, needs to be modified in such away that it includes the stress field caused by the fluid pressure. To this end, the stress part $\boldsymbol{\sigma}$ associated with the material deformation, called effective stress, is introduced as a process variable, and the momentum balance equation is modified such that it incorporates stresses and forces connected to the fluid pressure. This leads to

$$\begin{aligned} \text{div}\,(\boldsymbol{\sigma} - p\mathbf{I}) &= \mathbf{0} \text{ in } \Omega\,, \\ \boldsymbol{\sigma} - \mathcal{C}\varepsilon(\mathbf{u}) &= \mathbf{0} \text{ in } \Omega\,, \\ \mathbf{u} &= \mathbf{0} \text{ on } \Gamma_D^{\text{elas}}\,, \\ \boldsymbol{\sigma}\cdot\mathbf{n} &= \mathbf{g} \text{ on } \Gamma_N^{\text{elas}}\,. \end{aligned} \tag{2.7}$$

Note that the displacement field $\mathbf{u}$ of this model is defined with respect to the equilibrium under gravity. The true effective stress in the porous medium may be computed from $\boldsymbol{\sigma}$ by adding the effect of its own weight due to gravity. The mass balance equation in the fluid flow model (2.5) needs to take care of the change in pore space due to the displacement, which leads to

$$\begin{aligned} \text{div}\,\mathbf{w} + \partial_t \text{div}\,\mathbf{u} &= 0 \text{ in } \Omega, \\ \mathbf{w} + \kappa\,\nabla p &= \mathbf{0} \text{ in } \Omega, \\ p &= p^D \text{ on } \Gamma_D^{\text{flow}}, \\ \mathbf{w}\cdot\mathbf{n} &= 0 \text{ on } \Gamma_N^{\text{flow}}\,. \end{aligned} \tag{2.8}$$

Note again that in this model p has the character of a hydraulic potential and does not represent the true pressure in the porous medium. The pressure includes an additional term due to the effect of the weight of the fluid under gravity. Using an implicit time discretization, one obtains from this system

$$\begin{aligned} \operatorname{div} \mathbf{w} + \frac{1}{k}(\operatorname{div} \mathbf{u} - \operatorname{div} \mathbf{u}^{\text{old}}) &= 0 \text{ in } \Omega, \\ \mathbf{w} + \kappa \nabla p &= \mathbf{0} \text{ in } \Omega, \\ p &= p^D \text{ on } \Gamma_D^{\text{flow}}, \\ \mathbf{w} \cdot \mathbf{n} &= 0 \text{ on } \Gamma_N^{\text{flow}}, \end{aligned} \tag{2.9}$$

where k is the time-step length and $\mathbf{u}^{\text{old}}$ denotes the displacement field at the previous time-step. The least-squares functional associated with the coupled problem (2.7), (2.9) is given by

$$\begin{aligned} \mathcal{F}(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{w}, p; \mathbf{u}^{\text{old}}) = {} & \|\operatorname{div} (\boldsymbol{\sigma} - p\mathbf{I})\|^2 + \|\mathcal{C}^{-1/2}\boldsymbol{\sigma} - \mathcal{C}^{1/2}\varepsilon(\mathbf{u})\|^2 \\ & + \left\|\operatorname{div} \mathbf{w} + \frac{1}{k}(\operatorname{div} \mathbf{u} - \operatorname{div} \mathbf{u}^{\text{old}})\right\|^2 + \|\mathbf{w} + \kappa \nabla p\|^2 . \end{aligned} \tag{2.10}$$

The least-squares variational formulation consists in finding $\boldsymbol{\sigma} = \boldsymbol{\sigma}^N + \hat{\boldsymbol{\sigma}}$ with $\hat{\boldsymbol{\sigma}} \in H_{\Gamma_N^{\text{elas}}}(\operatorname{div}, \Omega)^2$, $\mathbf{u} \in H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2$, $\mathbf{w} \in H_{\Gamma_N^{\text{flow}}}(\operatorname{div}, \Omega)$, and $p = p^D + \hat{p}$ with $\hat{p} \in H^1_{\Gamma_D^{\text{flow}}}(\Omega)$ such that $\mathcal{F}(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{w}, p; \mathbf{u}^{\text{old}})$ is minimized.

We abbreviate the standard inner product on $L^2(\Omega)$ (or $L^2(\Omega)^2$, $L^2(\Omega)^{2\times 2}$, respectively) by $(\,\cdot\,,\,\cdot\,)$. Associated with the quadratic least-squares functional (2.10) is then the bilinear form

$$\begin{aligned} \mathcal{B}(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{w}, p; \boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q) = {} & (\operatorname{div} (\boldsymbol{\sigma} - p\mathbf{I}), \operatorname{div} (\boldsymbol{\tau} - q\mathbf{I})) \\ & + (\mathcal{C}^{-1/2}\boldsymbol{\sigma} - \mathcal{C}^{1/2}\varepsilon(\mathbf{u}), \mathcal{C}^{-1/2}\boldsymbol{\tau} - \mathcal{C}^{1/2}\varepsilon(\mathbf{v})) \\ & + \left(\operatorname{div} \mathbf{w} + \frac{1}{k}\operatorname{div} \mathbf{u}, \operatorname{div} \mathbf{z} + \frac{1}{k}\operatorname{div} \mathbf{v}\right) \\ & + (\mathbf{w} + \kappa \nabla p, \mathbf{z} + \kappa \nabla q) . \end{aligned} \tag{2.11}$$

The quadratic minimization problem above is then equivalent to the following variational problem: Find $\hat{\boldsymbol{\sigma}} \in H_{\Gamma_N^{\text{elas}}}(\operatorname{div}, \Omega)^2$, $\mathbf{u} \in H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2$, $\mathbf{w} \in H_{\Gamma_N^{\text{flow}}}(\operatorname{div}, \Omega)$, and $\hat{p} \in H^1_{\Gamma_D^{\text{flow}}}(\Omega)$ such that

$$\begin{aligned} & \mathcal{B}(\hat{\boldsymbol{\sigma}}, \mathbf{u}, \mathbf{w}, \hat{p}; \boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q) \\ & = - (\operatorname{div}\boldsymbol{\sigma}^N, \operatorname{div} (\boldsymbol{\tau} - q\mathbf{I})) - (\mathcal{C}^{-1/2}\boldsymbol{\sigma}^N, \mathcal{C}^{-1/2}\boldsymbol{\tau} - \mathcal{C}^{1/2}\varepsilon(\mathbf{v})) \\ & \quad + \left(\frac{1}{k}\operatorname{div} \mathbf{u}^{\text{old}}, \operatorname{div} \mathbf{z} + \frac{1}{k}\operatorname{div} \mathbf{v}\right) - (\kappa \nabla p^D, \mathbf{z} + \kappa \nabla q) \end{aligned} \tag{2.12}$$

for all $\boldsymbol{\tau} \in H_{\Gamma_N^{\text{elas}}}(\operatorname{div}, \Omega)^2$, $\mathbf{v} \in H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2$, $\mathbf{z} \in H_{\Gamma_N^{\text{flow}}}(\operatorname{div}, \Omega)$, and $q \in H^1_{\Gamma_D^{\text{flow}}}(\Omega)$.

3. Ellipticity of the least-squares formulation: Stationary case. In this section, we prove that the bilinear form (2.11) is coercive and continuous on the product space $H_{\Gamma_N^{\text{elas}}}(\operatorname{div}, \Omega)^2 \times H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2 \times H_{\Gamma_N^{\text{flow}}}(\operatorname{div}, \Omega) \times H^1_{\Gamma_D^{\text{flow}}}(\Omega)$. More precisely, in order to get approximation results which are uniform with respect to the Lamé parameters μ and λ, our aim is to show coercivity and continuity with constants which are independent of these parameters. To this end, we introduce the norm

$$
\begin{aligned}
|||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q)||| = \Big(&\|\mathrm{div}\,\boldsymbol{\tau}\|^2 + \|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 + \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&+ \|\mathrm{div}\,\mathbf{z}\|^2 + \|\mathbf{z}\|^2 + \|\nabla q\|^2\Big)^{1/2} .
\end{aligned}
\tag{3.1}
$$

Under our assumption on the boundary conditions, Korn's inequality, i.e.,

$$
\|\mathbf{v}\|^2 + \|\nabla \mathbf{v}\|^2 \le C_K \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 , \tag{3.2}
$$

for all $\mathbf{v} \in H^1_{\Gamma_D^{\mathrm{elas}}}(\Omega)^2$ holds (cf. [3, section VI.3]). Moreover, the Poincaré–Friedrichs inequality

$$
\|q\|^2 \le C_F \|\nabla q\|^2 \tag{3.3}
$$

is satisfied for all $q \in H^1_{\Gamma_D^{\mathrm{flow}}}(\Omega)$ (cf. [3, section II.1]). Therefore, $|||(\cdot,\cdot,\cdot,\cdot)|||$ defined in (3.1) is a norm on the product space $H_{\Gamma_N^{\mathrm{elas}}}(\mathrm{div},\Omega)^2 \times H^1_{\Gamma_D^{\mathrm{elas}}}(\Omega)^2 \times H_{\Gamma_N^{\mathrm{flow}}}(\mathrm{div},\Omega) \times H^1_{\Gamma_D^{\mathrm{flow}}}(\Omega)$.

Since the bilinear form $\mathcal{B}(\cdot,\cdot,\cdot,\cdot;\cdot,\cdot,\cdot,\cdot)$ is symmetric, the Cauchy–Schwarz inequality gives us

$$
\mathcal{B}(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{w}, p; \boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q) \le \mathcal{F}(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{w}, p; \mathbf{0})^{1/2} \mathcal{F}(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0})^{1/2} . \tag{3.4}
$$

It is therefore sufficient to show that

$$
\alpha\, |||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q)|||^2 \le \mathcal{F}(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0}) \le \beta\, |||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q)|||^2 \tag{3.5}
$$

holds for all $(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q) \in H_{\Gamma_N^{\mathrm{elas}}}(\mathrm{div},\Omega)^2 \times H^1_{\Gamma_D^{\mathrm{elas}}}(\Omega)^2 \times H_{\Gamma_N^{\mathrm{flow}}}(\mathrm{div},\Omega) \times H^1_{\Gamma_D^{\mathrm{flow}}}(\Omega)$ with positive constants α, β which are independent of the Lamé parameter λ.

The following lemma states these equivalence results for the least-squares functionals (2.4) and (2.6) associated with the deformation and flow subproblems, respectively. Our analysis of the coupled problem will be based on these results.

LEMMA 3.1. *There are positive constants* α_{elas} *and* β_{elas} *such that*

$$
\alpha_{\mathrm{elas}} |||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{0}, 0)|||^2 \le \mathcal{F}_{\mathrm{elas}}(\boldsymbol{\tau}, \mathbf{v}) \le \beta_{\mathrm{elas}} |||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{0}, 0)|||^2 \tag{3.6}
$$

holds for all $(\boldsymbol{\tau}, \mathbf{v}) \in H_{\Gamma_N^{\mathrm{elas}}}(\mathrm{div},\Omega)^2 \times H^1_{\Gamma_D^{\mathrm{elas}}}(\Omega)^2$ *uniformly as* $\lambda \to \infty$.

Furthermore, there are positive constants α_{flow} *and* β_{flow} *such that*

$$
\alpha_{\mathrm{flow}} |||(\mathbf{0}, \mathbf{0}, \mathbf{z}, q)|||^2 \le \mathcal{F}_{\mathrm{flow}}(\mathbf{z}, q) \le \beta_{\mathrm{flow}} |||(\mathbf{0}, \mathbf{0}, \mathbf{z}, q)|||^2 \tag{3.7}
$$

holds for all $(\mathbf{z}, q) \in H_{\Gamma_N^{\mathrm{flow}}}(\mathrm{div},\Omega) \times H^1_{\Gamma_D^{\mathrm{flow}}}(\Omega)$.

The equivalence (3.6) for the linear elasticity problem has been established in [9] and [6]. The stationary flow problem has been studied in [7], where (3.7) has been proven.

The least-squares functional associated with the stationary problem is given by

$$
\begin{aligned}
\mathcal{F}_{\mathrm{stat}}(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{w}, p) = &\|\mathrm{div}\,(\boldsymbol{\sigma} - p\mathbf{I})\|^2 + \|\mathcal{C}^{-1/2}\boldsymbol{\sigma} - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u})\|^2 \\
&+ \|\mathrm{div}\,\mathbf{w}\|^2 + \|\mathbf{w} + \kappa\nabla p\|^2 .
\end{aligned}
\tag{3.8}
$$

Our aim is to prove that this functional is equivalent to $|||(\cdot,\cdot,\cdot,\cdot)|||^2$ uniformly in the Lamé parameters, which is the statement of the following theorem.

Theorem 3.2. *There are positive constants α_{stat} and β_{stat} such that*

$$\alpha_{\text{stat}}|||(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q)|||^2 \le \mathcal{F}_{\text{stat}}(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q) \le \beta_{\text{stat}}|||(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q)|||^2 \tag{3.9}$$

holds for all $(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q) \in H_{\Gamma_N^{\text{elas}}}(\text{div},\Omega)^2 \times H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2 \times H_{\Gamma_N^{\text{flow}}}(\text{div},\Omega) \times H^1_{\Gamma_D^{\text{flow}}}(\Omega)$ *uniformly as* $\lambda \to \infty$.

Proof. Since div $(q\mathbf{I}) = \nabla q$, we have

$$\mathcal{F}_{\text{stat}}(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q) = \mathcal{F}_{\text{elas}}(\boldsymbol{\tau},\mathbf{v}) + \mathcal{F}_{\text{flow}}(\mathbf{z},q) + \|\nabla q\|^2 - 2(\text{div}\,\boldsymbol{\tau},\nabla q)\,. \tag{3.10}$$

The proof of the lower bound in (3.9) is based on

$$\begin{aligned}\mathcal{F}_{\text{stat}}(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q) &\ge \mathcal{F}_{\text{elas}}(\boldsymbol{\tau},\mathbf{v}) + \mathcal{F}_{\text{flow}}(\mathbf{z},q) + \left(1-\frac{1}{\eta}\right)\|\nabla q\|^2 - \eta\|\text{div}\,\boldsymbol{\tau}\|^2\\ &\ge (1-\eta)\|\text{div}\,\boldsymbol{\tau}\|^2 + \|\mathcal{C}^{-1/2}\boldsymbol{\tau} - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2\\ &\quad + \alpha_{\text{flow}}\left(\|\text{div}\,\mathbf{z}\|^2 + \|\mathbf{z}\|^2\right) + \left(\alpha_{\text{flow}} + 1 - \frac{1}{\eta}\right)\|\nabla q\|^2\,,\end{aligned}$$

where we are still free to choose $\eta \in (0,1)$. Choosing η in such a way that

$$\frac{1}{\eta} - 1 = \frac{1}{2}\alpha_{\text{flow}}\ ,\ \text{i.e.,}\ \eta = \frac{2}{2+\alpha_{\text{flow}}},$$

leads to

$$\begin{aligned}\mathcal{F}_{\text{stat}}(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q) &\ge (1-\eta)\mathcal{F}_{\text{elas}}(\boldsymbol{\tau},\mathbf{v}) + \frac{1}{2}\alpha_{\text{flow}}|||(\mathbf{0},\mathbf{0},\mathbf{z},q)|||^2\\ &\ge \frac{\alpha_{\text{flow}}}{2+\alpha_{\text{flow}}}\alpha_{\text{elas}}|||(\boldsymbol{\tau},\mathbf{v},\mathbf{0},0)|||^2 + \frac{1}{2}\alpha_{\text{flow}}|||(\mathbf{0},\mathbf{0},\mathbf{z},q)|||^2\\ &\ge \min\left\{\frac{\alpha_{\text{flow}}}{2+\alpha_{\text{flow}}}\alpha_{\text{elas}}, \frac{1}{2}\alpha_{\text{flow}}\right\}|||(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q)|||^2\,.\end{aligned}$$

For the upper bound in (3.9), we use (3.6), (3.7), and Cauchy–Schwarz inequality to obtain from (3.10)

$$\begin{aligned}\mathcal{F}_{\text{stat}}(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q) &\le \beta_{\text{elas}}|||(\boldsymbol{\tau},\mathbf{v},\mathbf{0},0)|||^2 + \beta_{\text{flow}}|||(\mathbf{0},\mathbf{0},\mathbf{z},q)|||^2 + 2\|\nabla q\|^2 + \|\text{div}\,\boldsymbol{\tau}\|^2\\ &\le (\beta_{\text{elas}}+1)|||(\boldsymbol{\tau},\mathbf{v},\mathbf{0},0)|||^2 + (\beta_{\text{flow}}+2)|||(\mathbf{0},\mathbf{0},\mathbf{z},q)|||^2\\ &\le \max\{\beta_{\text{elas}}+1,\beta_{\text{flow}}+2\}|||(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q)|||^2\,. \qquad \square\end{aligned}$$

4. Ellipticity of the least-squares formulation: Time-dependent case. For the analysis of the time-dependent least-squares functional (2.10) we make the assumption that

$$\Gamma_N^{\text{elas}} = \Gamma_D^{\text{flow}} =: \Gamma_1 \quad \text{and} \quad \Gamma_D^{\text{elas}} = \Gamma_N^{\text{flow}} =: \Gamma_2\,.$$

This has the physical meaning that the normal stress is prescribed on the same part of the boundary as the fluid pressure. Similarly, the displacements and the normal flux are set to zero on the same boundary segments. This is not unrealistic from a physical point of view even though more general situations may occur in practice (see the examples in section 6). We introduce as new variables

$$\tilde{\boldsymbol{\sigma}} = \boldsymbol{\sigma} - p\mathbf{I} \quad \text{and} \quad \tilde{\mathbf{w}} = \mathbf{w} + \frac{1}{k}\mathbf{u} \tag{4.1}$$

and observe that

$$\begin{aligned} \boldsymbol{\sigma} \in H_{\Gamma_1}(\mathrm{div},\Omega)^2 \, , \; p \in H^1_{\Gamma_1}(\Omega) \quad &\Longrightarrow \quad \tilde{\boldsymbol{\sigma}} \in H_{\Gamma_1}(\mathrm{div},\Omega)^2 \, , \\ \mathbf{w} \in H_{\Gamma_2}(\mathrm{div},\Omega) \, , \; \mathbf{u} \in H^1_{\Gamma_2}(\Omega)^2 \quad &\Longrightarrow \quad \tilde{\mathbf{w}} \in H_{\Gamma_2}(\mathrm{div},\Omega) \, . \end{aligned}$$

For the time-dependent case, the behavior of the least-squares formulation for different time-step sizes is of interest. In order to eliminate the dependence of the equivalence constants on the time-step length as much as possible, we scale the least-squares functional (2.10) with k as

$$\begin{aligned} \mathcal{F}_k(\boldsymbol{\sigma}, \mathbf{u}, \mathbf{w}, p; \mathbf{u}^{\mathrm{old}}) = \|\mathrm{div}\,(\boldsymbol{\sigma} - p\mathbf{I})\|^2 + \|\mathcal{C}^{-1/2}\boldsymbol{\sigma} - \mathcal{C}^{1/2}\varepsilon(\mathbf{u})\|^2 \\ + k\left\|\mathrm{div}\,\mathbf{w} + \frac{1}{k}(\mathrm{div}\,\mathbf{u} - \mathrm{div}\,\mathbf{u}^{\mathrm{old}})\right\|^2 + k\|\mathbf{w} + \kappa\nabla p\|^2 \, . \end{aligned} \tag{4.2}$$

Note that it is sufficient to consider the case $\kappa \equiv 1$ in the analysis since

$$\begin{aligned} \|\mathbf{w} + \kappa\nabla p\|^2 &\geq \underline{\kappa}\|\kappa^{-1/2}\mathbf{w} + \kappa^{1/2}\nabla p\|^2 = \underline{\kappa}\left(\kappa^{-1}\|\mathbf{w}\|^2 + 2(\mathbf{w}, \nabla p) + \kappa\|\nabla p\|^2\right) \\ &\geq \underline{\kappa}\min\{\overline{\kappa}^{-1}, \underline{\kappa}\}\left(\|\mathbf{w}\|^2 + 2(\mathbf{w}, \nabla p) + \|\nabla p\|^2\right) \\ &= \min\{\underline{\kappa}\overline{\kappa}^{-1}, \underline{\kappa}^2\}\|\mathbf{w} + \nabla p\|^2 \, , \end{aligned}$$

$$\begin{aligned} \|\mathbf{w} + \kappa\nabla p\|^2 &\leq \overline{\kappa}\|\kappa^{-1/2}\mathbf{w} + \kappa^{1/2}\nabla p\|^2 = \overline{\kappa}\left(\kappa^{-1}\|\mathbf{w}\|^2 + 2(\mathbf{w}, \nabla p) + \kappa\|\nabla p\|^2\right) \\ &\leq \overline{\kappa}\max\{\underline{\kappa}^{-1}, \overline{\kappa}\}\left(\|\mathbf{w}\|^2 + 2(\mathbf{w}, \nabla p) + \|\nabla p\|^2\right) \\ &= \min\{\overline{\kappa}\underline{\kappa}^{-1}, \overline{\kappa}^2\}\|\mathbf{w} + \nabla p\|^2 \, . \end{aligned}$$

Our analysis is then carried out with respect to the scaled norm

$$\begin{aligned} |||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q)|||_k = \Big(\|\mathrm{div}\,\boldsymbol{\tau}\|^2 + \|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2 + \|\mathcal{C}^{1/2}\varepsilon(\mathbf{v})\|^2 \\ + k\left(\|\mathrm{div}\,\mathbf{z}\|^2 + \|\mathbf{z}\|^2 + \|\nabla q\|^2\right)\Big)^{1/2} \, . \end{aligned} \tag{4.3}$$

The following lemma states that it is sufficient to prove ellipticity with respect to the transformed variables.

LEMMA 4.1. *For the norm* $|||(\,\cdot\,,\,\cdot\,,\,\cdot\,,\,\cdot\,)|||_k$ *defined in* (4.3) *and under the transformation* (4.1)*, the equivalence*

$$\frac{1}{\gamma}|||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q)|||_k^2 \leq |||(\tilde{\boldsymbol{\tau}}, \mathbf{v}, \tilde{\mathbf{z}}, q)|||_k^2 \leq \gamma|||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q)|||_k^2 \tag{4.4}$$

for all $(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q), (\tilde{\boldsymbol{\tau}}, \mathbf{v}, \tilde{\mathbf{z}}, q) \in H_{\Gamma_1}(\mathrm{div},\Omega)^2 \times H^1_{\Gamma_2}(\Omega)^2 \times H_{\Gamma_2}(\mathrm{div},\Omega) \times H^1_{\Gamma_1}(\Omega)$ *holds* *with*

$$\gamma = \max\left\{2\,,\ 1+\frac{4C_K}{k}, 1+\frac{2}{k}+\frac{2C_F}{k(\lambda+\mu)}\right\}\ .$$

Here, C_K and C_F denote the constants from Korn's inequality (3.2) *and the Poincaré–Friedrichs inequality* (3.3), *respectively.*

Proof. The left inequality in (4.4) follows from

$$\begin{aligned}
|||(\boldsymbol{\tau},\mathbf{v},\mathbf{z},q)|||_k^2 &= \left|\left|\left|\left(\tilde{\boldsymbol{\tau}}+q\mathbf{I},\mathbf{v},\tilde{\mathbf{z}}-\frac{1}{k}\mathbf{v},q\right)\right|\right|\right|_k^2 \\
&= \|\mathrm{div}\,(\tilde{\boldsymbol{\tau}}+q\mathbf{I})\|^2+\|\mathcal{C}^{-1/2}(\tilde{\boldsymbol{\tau}}+q\mathbf{I})\|^2+\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&+ k\left(\|\mathrm{div}\,(\tilde{\mathbf{z}}-\frac{1}{k}\mathbf{v})\|^2+\|\tilde{\mathbf{z}}-\frac{1}{k}\mathbf{v}\|^2+\|\nabla q\|^2\right) \\
&\le 2\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2+2\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2+\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2+\frac{2}{k}\|\mathrm{div}\,\mathbf{v}\|^2+\frac{2}{k}\|\mathbf{v}\|^2 \\
&+ 2k\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2+2k\|\tilde{\mathbf{z}}\|^2+(k+2)\|\nabla q\|^2+2\|\mathcal{C}^{-1/2}(q\mathbf{I})\|^2 \\
&\le 2\left(\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2+\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2\right)+\left(1+\frac{4C_K}{k}\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&+ 2k\left(\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2+\|\tilde{\mathbf{z}}\|^2\right)+\left(1+\frac{2}{k}+\frac{2C_F}{k(\lambda+\mu)}\right)k\|\nabla q\|^2\ .
\end{aligned}$$

For the last inequality, we have used

$$\begin{aligned}
\|\mathrm{div}\,\mathbf{v}\|^2+\|\mathbf{v}\|^2 &\le 2\|\nabla\mathbf{v}\|^2+\|\mathbf{v}\|^2\le 2C_K\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2\ , \\
\|\mathcal{C}^{-1/2}(q\,\mathbf{I})\|^2 &= (q\,\mathbf{I},\mathcal{C}^{-1}(q\,\mathbf{I}))=\frac{1}{\lambda+\mu}\|q\|^2\le\frac{C_F}{\lambda+\mu}\|\nabla q\|^2
\end{aligned}$$

(recall from (2.2) that $\mathcal{C}\,\mathbf{I}=2(\lambda+\mu)\,\mathbf{I}$). The right inequality in (4.4) follows along the same lines from

$$|||(\tilde{\boldsymbol{\tau}},\mathbf{v},\tilde{\mathbf{z}},q)|||_k^2 = \left|\left|\left|\left(\boldsymbol{\tau}-q\mathbf{I},\mathbf{v},\mathbf{z}+\frac{1}{k}\mathbf{v},q\right)\right|\right|\right|_k^2\ . \qquad \square$$

We prove the equivalence of the functional with respect to the transformed variables under (4.1). The following lemma will be useful in the analysis.

Lemma 4.2. *The antisymmetric part of the stress tensor,*

$$\mathrm{as}\,\tilde{\boldsymbol{\tau}}=\frac{1}{2}(\tilde{\boldsymbol{\tau}}-\tilde{\boldsymbol{\tau}}^T),$$

satisfies

$$\|\mathrm{as}\,\tilde{\boldsymbol{\tau}}\|^2\le 2\mu\|\mathcal{C}^{-1/2}(\tilde{\boldsymbol{\tau}}+q\,\mathbf{I})-\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2\ . \tag{4.5}$$

Proof. Using the special structure of $\mathcal{C}$ in (2.2), we are led to

$$\begin{aligned}
&\|\mathcal{C}^{-1/2}(\tilde{\boldsymbol{\tau}} + q\,\mathbf{I}) - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&= (\mathcal{C}^{-1}(\tilde{\boldsymbol{\tau}} + q\,\mathbf{I}), \tilde{\boldsymbol{\tau}} + q\,\mathbf{I}) - 2(\tilde{\boldsymbol{\tau}} + q\,\mathbf{I}, \boldsymbol{\varepsilon}(\mathbf{v})) + (\mathcal{C}\boldsymbol{\varepsilon}(\mathbf{v}), \boldsymbol{\varepsilon}(\mathbf{v})) \\
&= \left(\mathcal{C}^{-1}\begin{pmatrix} \tilde{\tau}_{11} + q & \tilde{\tau}_{12} \\ \tilde{\tau}_{21} & \tilde{\tau}_{22} + q \end{pmatrix}, \begin{pmatrix} \tilde{\tau}_{11} + q & \tilde{\tau}_{12} \\ \tilde{\tau}_{21} & \tilde{\tau}_{22} + q \end{pmatrix}\right) \\
&- 2\left(\begin{pmatrix} \tilde{\tau}_{11} + q & \tilde{\tau}_{12} \\ \tilde{\tau}_{21} & \tilde{\tau}_{22} + q \end{pmatrix}, \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{pmatrix}\right) + \left(\mathcal{C}\begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{pmatrix}, \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{pmatrix}\right) \\
&= \frac{1}{2\mu}\left(\begin{pmatrix} * & \tilde{\tau}_{12} \\ \tilde{\tau}_{21} & * \end{pmatrix}, \begin{pmatrix} * & \tilde{\tau}_{12} \\ \tilde{\tau}_{21} & * \end{pmatrix}\right) \\
&- 2\left(\begin{pmatrix} * & \tilde{\tau}_{12} \\ \tilde{\tau}_{21} & * \end{pmatrix}, \begin{pmatrix} * & \varepsilon_{12} \\ \varepsilon_{21} & * \end{pmatrix}\right) + 2\mu\left(\begin{pmatrix} * & \varepsilon_{12} \\ \varepsilon_{21} & * \end{pmatrix}, \begin{pmatrix} * & \varepsilon_{12} \\ \varepsilon_{21} & * \end{pmatrix}\right) .
\end{aligned}$$

This leads to

$$\begin{aligned}
\|\mathcal{C}^{-1/2}\boldsymbol{\tau} - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 &= \left\|\begin{pmatrix} * & \frac{1}{\sqrt{2\mu}}\tau_{12} - \sqrt{2\mu}\varepsilon_{12} \\ \frac{1}{\sqrt{2\mu}}\tau_{21} - \sqrt{2\mu}\varepsilon_{21} & * \end{pmatrix}\right\|^2 \\
&\geq \left\|\frac{1}{\sqrt{2\mu}}\tau_{12} - \sqrt{2\mu}\varepsilon_{12}\right\|^2 + \left\|\frac{1}{\sqrt{2\mu}}\tau_{21} - \sqrt{2\mu}\varepsilon_{21}\right\|^2 \\
&\geq \frac{1}{2}\left\|\frac{1}{\sqrt{2\mu}}(\tau_{12} - \tau_{21}) - \sqrt{2\mu}(\varepsilon_{12} - \varepsilon_{21})\right\|^2 \\
&= \frac{1}{2}\left\|\frac{1}{\sqrt{2\mu}}(\tau_{12} - \tau_{21})\right\|^2 = \frac{1}{4\mu}\|\tau_{12} - \tau_{21}\|^2 = \frac{1}{2\mu}\|\text{as}\,\boldsymbol{\tau}\|^2 . \qquad \square
\end{aligned}$$

THEOREM 4.3. *There are positive constants $\tilde{\alpha}$ and $\tilde{\beta}$ such that*

$$\tilde{\alpha}|||(\tilde{\boldsymbol{\tau}}, \mathbf{v}, \tilde{\mathbf{z}}, q)|||_k^2 \leq \mathcal{F}_k(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0}) \leq \tilde{\beta}|||(\tilde{\boldsymbol{\tau}}, \mathbf{v}, \tilde{\mathbf{z}}, q)|||_k^2 \tag{4.6}$$

holds for all $(\tilde{\boldsymbol{\tau}}, \mathbf{v}, \tilde{\mathbf{z}}, q) \in H_{\Gamma_1}(\text{div}, \Omega)^2 \times H^1_{\Gamma_2}(\Omega)^2 \times H_{\Gamma_2}(\text{div}, \Omega) \times H^1_{\Gamma_1}(\Omega)$ uniformly as $\lambda \to \infty$.

Proof. Writing the functional with respect to the transformed variables gives

$$\begin{aligned}
&\mathcal{F}_k(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0}) \\
&= \|\text{div}\,\tilde{\boldsymbol{\tau}}\|^2 + \|\mathcal{C}^{-1/2}(\tilde{\boldsymbol{\tau}} + q\mathbf{I}) - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 + k\|\text{div}\,\tilde{\mathbf{z}}\|^2 + k\left\|\tilde{\mathbf{z}} + \nabla q - \frac{1}{k}\mathbf{v}\right\|^2 .
\end{aligned}$$

The right inequality in (4.6) follows from straightforward application of the Cauchy–Schwarz inequality, and we may therefore restrict our attention to the lower bound.

Integration by parts, the elementary estimates

$$2(\text{tr}\,\tilde{\boldsymbol{\tau}}, q) \leq \frac{\xi}{2}\|\text{tr}\,\tilde{\boldsymbol{\tau}}\|^2 + \frac{2}{\xi}\|q\|^2 \quad \text{and} \quad 2(\tilde{\mathbf{z}}, \mathbf{v}) \leq k\eta\|\tilde{\mathbf{z}}\|^2 + \frac{1}{k\eta}\|\mathbf{v}\|^2$$

with constants $\xi, \eta \in (0, 1)$ to be selected appropriately, and the inequality

$$\|\mathrm{as}\,\tilde{\boldsymbol{\tau}}\|^2 \le 2\mu\|\mathcal{C}^{-1/2}(\tilde{\boldsymbol{\tau}} + q\,\mathbf{I}) - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2$$

shown in Lemma 4.2 will be the basis of the following chain of inequalities. With suitable positive constants A, B, and D to be chosen later, we have

$$\begin{aligned}
&\max\{A, 1+B, D\}\, \mathcal{F}_k(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0}) \\
&\ge A\,\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + (1+B)\|\mathcal{C}^{-1/2}(\tilde{\boldsymbol{\tau}} + q\,\mathbf{I}) - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&+ Dk\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2 + k\left\|\tilde{\mathbf{z}} + \nabla q - \frac{1}{k}\mathbf{v}\right\|^2 \\
&= A\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + B\|\mathcal{C}^{-1/2}(\tilde{\boldsymbol{\tau}} + q\,\mathbf{I}) - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 + \|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}} - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&+ \|\mathcal{C}^{-1/2}(q\,\mathbf{I})\|^2 + Dk\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2 + k\|\tilde{\mathbf{z}} + \nabla q\|^2 + \frac{1}{k}\|\mathbf{v}\|^2 \\
&+ 2(\mathrm{tr}\,(\mathcal{C}^{-1}\tilde{\boldsymbol{\tau}}) - \mathrm{div}\,\mathbf{v}, q) - 2(\tilde{\mathbf{z}} + \nabla q, \mathbf{v}) \\
&\ge A\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + \frac{B}{2\mu}\|\mathrm{as}\,\tilde{\boldsymbol{\tau}}\|^2 + \|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}} - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 + \frac{1}{k}\|\mathbf{v}\|^2 \\
&+ Dk\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2 + k\|\tilde{\mathbf{z}} + \nabla q\|^2 + \frac{1}{\lambda+\mu}\|q\|^2 + 2\left(\frac{1}{2(\lambda+\mu)}\mathrm{tr}\,\tilde{\boldsymbol{\tau}}, q\right) - 2(\tilde{\mathbf{z}}, \mathbf{v}) \\
&\ge A\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + \frac{B}{2\mu}\|\mathrm{as}\,\tilde{\boldsymbol{\tau}}\|^2 + \|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}} - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 - \frac{\xi}{4(\lambda+\mu)}\|\mathrm{tr}\,\tilde{\boldsymbol{\tau}}\|^2 \\
&- \frac{1}{k}\left(\frac{1}{\eta} - 1\right)\|\mathbf{v}\|^2 + Dk\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2 + k\left(\|\tilde{\mathbf{z}} + \nabla q\|^2 - \eta\|\tilde{\mathbf{z}}\|^2\right) \\
&- \left(\frac{1}{\xi} - 1\right)\frac{1}{\lambda+\mu}\|q\|^2 =: \mathcal{G}_1(\tilde{\boldsymbol{\tau}}, \mathbf{v}) + \mathcal{G}_2(\tilde{\mathbf{z}}, q)\,.
\end{aligned}$$

This implies

$$\mathcal{F}_k(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0}) \ge \min\left\{\frac{1}{A}, \frac{1}{1+B}, \frac{1}{D}\right\}\left(\mathcal{G}_1(\tilde{\boldsymbol{\tau}}, \mathbf{v}) + \mathcal{G}_2(\tilde{\mathbf{z}}, q)\right), \tag{4.7}$$

and we are left with estimating $\mathcal{G}_1$ and $\mathcal{G}_2$ separately.

For the first term, we use

$$\begin{aligned}
\|\mathrm{tr}\,\tilde{\boldsymbol{\tau}}\|^2 &= 2(\lambda+\mu)(\mathrm{tr}\,(\mathcal{C}^{-1}\tilde{\boldsymbol{\tau}}), \mathrm{tr}\,\tilde{\boldsymbol{\tau}}) \\
&= 4(\lambda+\mu)\left(\frac{1}{2}(\mathrm{tr}\,(\mathcal{C}^{-1}\tilde{\boldsymbol{\tau}}))\,\mathbf{I}, \frac{1}{2}(\mathrm{tr}\,\tilde{\boldsymbol{\tau}})\,\mathbf{I}\right) \\
&\le 4(\lambda+\mu)(\mathcal{C}^{-1}\tilde{\boldsymbol{\tau}}, \tilde{\boldsymbol{\tau}}) = 4(\lambda+\mu)\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2\,,
\end{aligned}$$

$$\begin{aligned}
(\tilde{\boldsymbol{\tau}}, \boldsymbol{\varepsilon}(\mathbf{v})) &= (\tilde{\boldsymbol{\tau}} - \mathrm{as}\,\tilde{\boldsymbol{\tau}}, \boldsymbol{\varepsilon}(\mathbf{v})) = (\tilde{\boldsymbol{\tau}} - \mathrm{as}\,\tilde{\boldsymbol{\tau}}, \nabla\mathbf{v}) \\
&= -(\mathrm{div}\,\tilde{\boldsymbol{\tau}}, \mathbf{v}) - (\mathrm{as}\,\tilde{\boldsymbol{\tau}}, \nabla\mathbf{v}),
\end{aligned}$$

and Korn's inequality (3.2) to obtain

$$\begin{aligned}
\mathcal{G}_1(\tilde{\boldsymbol{\tau}},\mathbf{v}) &\geq A\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + \frac{B}{2\mu}\|\mathrm{as}\,\tilde{\boldsymbol{\tau}}\|^2 + \|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}} - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 - \xi\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2 \\
&\quad - \frac{C_K}{k}\left(\frac{1}{\eta}-1\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&= A\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + \frac{B}{2\mu}\|\mathrm{as}\,\tilde{\boldsymbol{\tau}}\|^2 + (1-\xi)\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2 - 2(\tilde{\boldsymbol{\tau}},\boldsymbol{\varepsilon}(\mathbf{v})) \\
&\quad + \left(1 - \frac{C_K}{k}\left(\frac{1}{\eta}-1\right)\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&= A\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + \frac{B}{2\mu}\|\mathrm{as}\,\tilde{\boldsymbol{\tau}}\|^2 + (1-\xi)\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2 + 2(\mathrm{div}\,\tilde{\boldsymbol{\tau}},\mathbf{v}) \\
&\quad + 2(\mathrm{as}\,\tilde{\boldsymbol{\tau}},\nabla\mathbf{v}) + \left(1 - \frac{C_K}{k}\left(\frac{1}{\eta}-1\right)\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&= \left(A - \frac{1}{\rho}\right)\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + (1-\xi)\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2 + \left(1 - \frac{C_K}{k}\left(\frac{1}{\eta}-1\right)\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\
&\quad + \left\|\frac{1}{\rho^{1/2}}\mathrm{div}\,\tilde{\boldsymbol{\tau}} + \rho^{1/2}\mathbf{v}\right\|^2 - \rho\|\mathbf{v}\|^2 \\
&\quad + \left\|\left(\frac{B}{2\mu}\right)^{1/2}\mathrm{as}\,\tilde{\boldsymbol{\tau}} + \left(\frac{2\mu}{B}\right)^{1/2}\nabla\mathbf{v}\right\|^2 - \frac{2\mu}{B}\|\nabla\mathbf{v}\|^2 \\
&\geq \left(A - \frac{1}{\rho}\right)\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + (1-\xi)\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2 \\
&\quad + \left(1 - \frac{C_K}{k}\left(\frac{1}{\eta}-1\right) - C_K\max\left\{\rho,\frac{2\mu}{B}\right\}\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \,.
\end{aligned}$$

Choosing

$$\eta = \frac{1}{1+\frac{k}{4C_K}}\ ,\ \rho = \frac{1}{4C_K}\ ,\ A = 8C_K\ ,\ B = \frac{2\mu}{\rho} = 8\mu C_K$$

leads to

$$\mathcal{G}_1(\tilde{\boldsymbol{\tau}},\mathbf{v}) \geq 4C_K\|\mathrm{div}\,\tilde{\boldsymbol{\tau}}\|^2 + (1-\xi)\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2 + \frac{1}{2}\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \,. \tag{4.8}$$

Note that $\xi \in (0,1)$ is still free to be chosen appropriately for the estimation of $\mathcal{G}_2$ below.

For the second term, the Poincaré–Friedrichs inequality (3.3) leads to

$$\begin{aligned}
\mathcal{G}_2(\tilde{\mathbf{z}},q) &\geq Dk\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2 + k\left(\|\tilde{\mathbf{z}}+\nabla q\|^2 - \eta\|\tilde{\mathbf{z}}\|^2\right) - \left(\frac{1}{\xi}-1\right)\frac{C_F}{\lambda+\mu}\|\nabla q\|^2 \\
&= Dk\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2 + k(1-\eta)\|\tilde{\mathbf{z}}\|^2 + 2k(\tilde{\mathbf{z}},\nabla q) + \left(k - \left(\frac{1}{\xi}-1\right)\frac{C_F}{\lambda+\mu}\right)\|\nabla q\|^2 \\
&= Dk\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2 + k(1-\eta)\|\tilde{\mathbf{z}}\|^2 - 2k(\mathrm{div}\,\tilde{\mathbf{z}},q) + \left(k - \left(\frac{1}{\xi}-1\right)\frac{C_F}{\lambda+\mu}\right)\|\nabla q\|^2 \\
&\geq \frac{Dk}{2}\|\mathrm{div}\,\tilde{\mathbf{z}}\|^2 + k(1-\eta)\|\tilde{\mathbf{z}}\|^2 + \left(k - \left(\frac{1}{\xi}-1\right)\frac{C_F}{\lambda+\mu} - k\frac{2C_F}{D}\right)\|\nabla q\|^2 \,.
\end{aligned}$$

Here, we may choose

$$\xi = \frac{1}{1 + \frac{k(\lambda+\mu)}{3C_F}} \ , \ D = 6C_F,$$

which implies

$$\mathcal{G}_2(\tilde{\mathbf{z}}, q) \geq k \left(\frac{D}{2} \|\operatorname{div} \tilde{\mathbf{z}}\|^2 + (1-\eta)\|\tilde{\mathbf{z}}\|^2 + \frac{1}{3}\|\nabla q\|^2 \right) . \tag{4.9}$$

Therefore,

$$\mathcal{G}_1(\tilde{\boldsymbol{\tau}}, \mathbf{v}) + \mathcal{G}_2(\tilde{\mathbf{z}}, q) \geq \min \left\{ \frac{A}{2}, 1-\xi, \frac{1}{4}, \frac{D}{2}, 1-\eta, \frac{1}{3} \right\} \ |||(\tilde{\boldsymbol{\tau}}, \mathbf{v}, \tilde{\mathbf{z}}, q)|||_k^2,$$

which, together with (4.7), completes our proof. □

Finally, our main result follows easily from Theorems 4.3 and 4.1.

THEOREM 4.4. *There are positive constants* α *and* β *such that*

$$\alpha |||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q)|||_k^2 \leq \mathcal{F}_k(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0}) \leq \beta |||(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q)|||_k^2 \tag{4.10}$$

holds for all $(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q) \in H_{\Gamma_1}(\operatorname{div}, \Omega)^2 \times H^1_{\Gamma_2}(\Omega)^2 \times H_{\Gamma_2}(\operatorname{div}, \Omega) \times H^1_{\Gamma_1}(\Omega)$ *uniformly as* $\lambda \to \infty$.

Remark. The equivalence constants in (4.10) and (4.6) still depend on the time-step length k. The behavior of the finite element approximation obtained with the least-squares method for small k is of particular interest in connection with the reduced regularity of Biot's consolidation problem for $t \to 0$ analyzed in detail in [18]. A closer inspection of the proof of Theorem 4.3 shows that (4.6) can be sharpened to the following inequalities with constants α^*, β^* now independent of k:

$$\begin{aligned} \mathcal{F}_k(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0}) \geq \alpha^* (&\|\operatorname{div} \tilde{\boldsymbol{\tau}}\|^2 + k\|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2 + \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \\ &+ k\|\operatorname{div} \tilde{\mathbf{z}}\|^2 + k^2\|\tilde{\mathbf{z}}\|^2 + k\|\nabla q\|^2) , \end{aligned} \tag{4.11}$$

$$\begin{aligned} \mathcal{F}_k(\boldsymbol{\tau}, \mathbf{v}, \mathbf{z}, q; \mathbf{0}) \leq \beta^* \Big(&\|\operatorname{div} \tilde{\boldsymbol{\tau}}\|^2 + \|\mathcal{C}^{-1/2}\tilde{\boldsymbol{\tau}}\|^2 + \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 + \frac{1}{k}\|\mathbf{v}\|^2 \\ &+ k\|\operatorname{div} \tilde{\mathbf{z}}\|^2 + k\|\tilde{\mathbf{z}}\|^2 + k\|\nabla q\|^2 + \|q\|^2 \Big) . \end{aligned} \tag{4.12}$$

5. Finite element approximation. The finite element approximation of the least-squares formulation consists in minimizing the least-squares functional in (2.10) with respect to suitable finite-dimensional subspaces of $H_{\Gamma_N^{\text{elas}}}(\operatorname{div}, \Omega)^2 \times H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2 \times H_{\Gamma_N^{\text{flow}}}(\operatorname{div}, \Omega) \times H^1_{\Gamma_D^{\text{flow}}}(\Omega)$. In analogy to the derivation in section 2, the minimization with respect to these finite-dimensional spaces leads to a linear variational problem of the form (2.12) with respect to these finite-dimensional spaces. This leaves us with the question of choosing appropriate finite element spaces for

$$\hat{\boldsymbol{\sigma}} \in H_{\Gamma_N^{\text{elas}}}(\operatorname{div}, \Omega)^2 \ , \ \mathbf{u} \in H^1_{\Gamma_D^{\text{elas}}}(\Omega)^2 \ , \ \mathbf{w} \in H_{\Gamma_N^{\text{flow}}}(\operatorname{div}, \Omega) \ , \ \hat{p} \in H^1_{\Gamma_D^{\text{flow}}}(\Omega) \ .$$

In contrast to the situation for mixed finite element approaches based on a saddle point formulation, the finite element spaces do not have to satisfy the compatibility condition by Ladyshenskaja–Babuška–Brezzi. Instead the spaces can be chosen completely independently of each other in the least-squares approach. The convergence of the least-squares mixed finite element method is only governed by the ellipticity of the least-squares functional, i.e., its equivalence to a meaningful norm of the error, and by the approximation properties of the finite element spaces.

Our finite element spaces are based on a sequence of triangulations $\{\mathcal{T}_h\}$ of the domain Ω. Suitable finite element spaces for $H(\mathrm{div},\Omega)$ are the Raviart–Thomas spaces given by piecewise polynomials of the form

$$\mathbf{s}_h|_T = \begin{pmatrix} p_{m-1}^{(I)} \\ p_{m-1}^{(II)} \end{pmatrix} + \mathbf{x} p_{m-1}^{(III)}$$

on each triangle $T \in \mathcal{T}_h$, where $p_{m-1}^{(I)}$, $p_{m-1}^{(II)}$, and $p_{m-1}^{(III)}$ denote polynomials of degree $m-1$. For $m=1$, the case we use in our numerical computations presented in section 6, this implies

$$\mathbf{s}_h|_T = \begin{pmatrix} \alpha_T \\ \beta_T \end{pmatrix} + \gamma_T \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} .$$

The Raviart–Thomas space of degree m is also characterized as those polynomials of degree m (componentwise) with the property that the trace of $\mathbf{s}_h \cdot \mathbf{n}$ on each edge is a polynomial of degree $m-1$. Raviart–Thomas spaces are used in our computations for the approximation of each component of the stress tensor and for the fluid flux. Note that the boundary values for $\boldsymbol{\sigma} \cdot \mathbf{n}$ and $\mathbf{w} \cdot \mathbf{n}$ can easily be prescribed for these spaces since the degrees of freedom in the above basis representation are associated with these quantities. For the approximation of each component of the displacement and for the hydraulic potential, standard conforming piecewise polynomial spaces are used.

We investigate the theoretical approximation properties of our least-squares approach on a sequence of triangulations under sufficient regularity assumptions if we use the finite element spaces mentioned above. For the Raviart–Thomas elements, we have

$$\begin{aligned} \|\boldsymbol{\sigma} - \mathbf{P}_h\boldsymbol{\sigma}\| \le Ch^m|\boldsymbol{\sigma}|_{m,\Omega} \, , \quad & \|\mathrm{div}\,(\boldsymbol{\sigma} - \mathbf{P}_h\boldsymbol{\sigma})\| \le Ch^m|\mathrm{div}\,\boldsymbol{\sigma}|_{m,\Omega} \, , \\ \|\mathbf{w} - \mathbf{P}_h\mathbf{w}\| \le Ch^m|\mathbf{w}|_{m,\Omega} \, , \quad & \|\mathrm{div}\,(\mathbf{w} - \mathbf{P}_h\mathbf{w})\| \le Ch^m|\mathrm{div}\,\mathbf{w}|_{m,\Omega} \, , \end{aligned}$$

with a suitable interpolation operator $\mathbf{P}_h$ (cf. [5, Proposition III.3.9]). The standard finite element interpolation estimates (see, for example, [3, section II.6]) imply

$$\begin{aligned} \|\mathbf{u} - \mathbf{Q}_h\mathbf{u}\| \le Ch^m|\mathbf{u}|_{m,\Omega} \, , \quad & \|\nabla\,(\mathbf{u} - \mathbf{Q}_h\mathbf{u})\| \le Ch^m|\mathbf{u}|_{m+1,\Omega} \, , \\ \|p - Q_hp\| \le Ch^m|p|_{m,\Omega} \, , \quad & \|\nabla\,(p - Q_hp)\| \le Ch^m|p|_{m+1,\Omega} \, , \end{aligned}$$

where Q_h denotes the standard interpolation operator.

With (4.12) the above interpolation estimates lead to

$$\begin{aligned} \mathcal{F}_k(\boldsymbol{\sigma}_h, \mathbf{u}_h, \mathbf{w}_h, p_h; \mathbf{u}^{\mathrm{old}}) &= \mathcal{F}_k(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h, \mathbf{u} - \mathbf{u}_h, \mathbf{w} - \mathbf{w}_h, p - p_h; \mathbf{0}) \\ &\le \beta^* C^2 h^{2m} \Bigg(|\mathrm{div}\,\tilde{\boldsymbol{\sigma}}|_{m,\Omega}^2 + k|\tilde{\boldsymbol{\sigma}}|_{m,\Omega}^2 + |\mathbf{u}|_{m+1,\Omega}^2 + \frac{1}{k}|\mathbf{u}|_{m,\Omega}^2 \\ &\qquad + k|\mathrm{div}\,\tilde{\mathbf{w}}|_{m,\Omega}^2 + k|\tilde{\mathbf{w}}|_{m,\Omega}^2 + k|p|_{m+1,\Omega}^2 + |p|_{m,\Omega}^2 \Bigg) \\ &\le \beta^* C^2 h^{2m} \left(|\mathbf{u}|_{m+1,\Omega}^2 + \frac{1}{k}|\mathbf{u}|_{m,\Omega}^2 + k|\mathrm{div}\,\mathbf{w}|_{m,\Omega}^2 + k|p|_{m+1,\Omega}^2 + |p|_{m,\Omega}^2 \right) . \end{aligned} \tag{5.1}$$

If we combine this with (4.11), we obtain

$$\begin{aligned}&\|\mathrm{div}\ (\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}_h)\| + k^{1/2}\|\mathcal{C}^{-1/2}(\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}_h)\| + \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u} - \mathbf{u}_h)\| \\ &\quad + k^{1/2}\|\mathrm{div}\ (\tilde{\mathbf{w}} - \tilde{\mathbf{w}}_h)\| + k\|\tilde{\mathbf{w}} - \tilde{\mathbf{w}}_h\| + k^{1/2}\|\nabla(p - p_h)\| \\ &\le \sqrt{\frac{\beta^*}{\alpha^*}} C h^m \left(|\mathbf{u}|_{m+1,\Omega} + \frac{|\mathbf{u}|_{m,\Omega}}{k^{1/2}} + k^{1/2}|\mathrm{div}\ \mathbf{w}|_{m,\Omega} + k^{1/2}|p|_{m+1,\Omega} + |p|_{m,\Omega} \right) .\end{aligned}$$

In particular, for the approximation of the primal variables,

$$\begin{aligned}&\|\nabla(\mathbf{u} - \mathbf{u}_h)\| + k^{1/2}\|\nabla(p - p_h)\| \\ &\le \sqrt{\frac{\beta^*}{\alpha^*}} C h^m \left(|\mathbf{u}|_{m+1,\Omega} + \frac{|\mathbf{u}|_{m,\Omega}}{k^{1/2}} + k^{1/2}|\mathrm{div}\ \mathbf{w}|_{m,\Omega} + k^{1/2}|p|_{m+1,\Omega} + |p|_{m,\Omega} \right)\end{aligned} \tag{5.2}$$

holds.

In comparison, the standard mixed approach is based on the variational formulation of finding $\mathbf{u} \in H^1_{\Gamma_2}(\Omega)^2$ and $p \in p^D + H^1_{\Gamma_1}(\Omega)$ such that

$$\begin{aligned}(\mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}), \boldsymbol{\varepsilon}(\mathbf{v})) - (p, \mathrm{div}\ \mathbf{v}) &= (\mathbf{g}, \mathbf{v})_{\Gamma_1} - (p^D, \mathbf{n} \cdot \mathbf{v})_{\Gamma_1} \\ -(\mathrm{div}\ \mathbf{u}, q) - k(\kappa\, \nabla p, \nabla q) &= -(\mathrm{div}\ \mathbf{u}^{\mathrm{old}}, q)\end{aligned} \tag{5.3}$$

holds for all $\mathbf{v} \in H^1_{\Gamma_2}(\Omega)^2$ and $q \in H^1_{\Gamma_1}(\Omega)$. By $(\,\cdot\,,\,\cdot\,)_{\Gamma_1}$, the inner product in $L^2(\Gamma_1)$ is meant. If a compatible pair of finite element spaces satisfying the inf-sup condition is used, then approximation bounds are obtained which are uniform as $k \to 0$. In particular, for the Taylor–Hood elements combining piecewise polynomials of degree m for $\mathbf{u}$ with piecewise polynomials of degree $m-1$ for p, the estimates

$$\begin{aligned}\|\nabla(\mathbf{u} - \mathbf{u}_h)\| &\le C h^m |\mathbf{u}|_{m+1,\Omega}\ , \\ \|p - p_h\| + k^{1/2}\|\nabla(p - p_h)\| &\le C \left(h^m + k^{1/2} h^{m-1} \right) |p|_{m,\Omega}\end{aligned} \tag{5.4}$$

for $m \ge 2$ can be deduced from the results (for the case $k = 0$) in [4, section 12.6]. The regularity results in [18, Lemma 2.2] imply for the solution of the initial-boundary value problem (2.7), (2.8)

$$|\mathbf{u}(t)|_{m+1,\Omega} + |p(t)|_{m,\Omega} \le C_R t^{-m/2}\ , \tag{5.5}$$

where the constant C_R depends on the boundary data $\mathbf{g}$ and p^D. Of course, (5.5) only holds under suitable assumptions on the boundary conditions and on the smoothness of $\partial\Omega$. Based on these regularity estimates, the a priori estimates (5.4) for the mixed method lead to

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\| + \|p - p_h\| + k^{1/2}\|\nabla(p - p_h)\| \le C \left(\left(\frac{h}{k^{1/2}} \right)^m + \left(\frac{h}{k^{1/2}} \right)^{m-1} \right) . \tag{5.6}$$

In comparison, for our least-squares approach we obtain

$$\|\nabla(\mathbf{u} - \mathbf{u}_h)\| + k^{1/2}\|\nabla(p - p_h)\| \le C \left(\frac{h}{k^{1/2}} \right)^m \tag{5.7}$$

using (5.2). Note, however, that our motivation for the least-squares approach in this paper is the simultaneous approximation for the stress and flux variables as well as the availability of the least-squares functional as an a posteriori estimator for adaptive refinement of the finite element triangulation.

Table 6.1
Material parameters for numerical experiments.

Parameter	Value
Young's modulus E	$30 \ \frac{kN}{m^2}$
Poisson ratio ν	0.2
percolation coefficient κ	$10^{-4} \ \frac{m^4}{kNs}$
boundary condition σ_0	$1 \ \frac{kN}{m^2}$

6. Computational experiments. In this section, we study the behavior of the approximate solution of Biot's consolidation problem by the least-squares formulation (2.12). To this end, we consider two test examples with different properties in a two-dimensional domain. Both examples are taken from [17].

Example 1. The first example is a two-dimensional simulation of a one-dimensional consolidation problem in a column of porous soil of depth 1 with rigid and impermeable walls. This column shall be loaded on the top by a pressure σ_0 and is free to drain. This one-dimensional problem has an analytic solution which is given in terms of an infinite sum (see [17]). It is therefore perfectly suited for an illustration of the effects at the beginning of the simulation for different time-step sizes k.

For the two-dimensional simulation of the one-dimensional consolidation problem, a box of 0.1 m width and 1 m depth with the origin located on the upper left corner is used as simulation domain Ω with the following boundary conditions:

$$\begin{array}{llll} p = 0, & \sigma_{xy} = 0, & \sigma_{yy} = -\sigma_0 & \text{on } \{(x,y) \in \partial\Omega : y = 0\}, \\ w = 0, & u_x = 0, & \sigma_{yx} = 0 & \text{on } \{(x,y) \in \partial\Omega : x = 0\}, \\ w = 0, & u_x = 0, & \sigma_{yx} = 0 & \text{on } \{(x,y) \in \partial\Omega : x = 0.1\}, \\ w = 0, & \sigma_{xy} = 0, & u_y = 0 & \text{on } \{(x,y) \in \partial\Omega : y = -1\}. \end{array}$$

As initial conditions, $u = 0$ is prescribed in Ω for $t = 0$. The material parameters are given in Table 6.1.

The computations are based on the lowest-order case of finite element spaces ($m = 1$ in the notation of section 5) except for the displacement, which is approximated by piecewise quadratic functions. Our observations indicate that the overall finite element approximation is not satisfactory if only piecewise linear functions are used for the displacement. Note, however, that the choice of quadratic finite elements does not lead to uniform convergence for nearly incompressible materials. Uniform approximation in the incompressible limit could be achieved by the use of nonconforming finite elements. A detailed study of this topic for the elasticity subproblem using the least-squares functional (2.4), including computational results, is given in [6].

For $t \to 0$, the pressure front involves large gradients, as shown in Figure 6.1. This is caused by the Dirichlet boundary condition $p = 0$ at $y = 0$, which contradicts the asymptotic behavior $p \to \sigma_0$ for $t \to 0$ for all $x \in (0, 1)$. Due to this behavior, the length of the time-step influences the choice of an initial grid, which is fine enough to resolve this front. For this purpose, a nonuniform initial coarse triangulation is used in our computations which is refined near the top of the domain. Six steps of adaptive refinement are then performed starting from this initial triangulation of 1725 degrees of freedom for the time-step lengths $k = 100, 1$, and 0.001. The local evaluation of the least-squares functional is used as an a posteriori error estimator for the adaptive refinement process (cf. [1]). Tables 6.2, 6.3 and 6.4 demonstrate the effect of adaptive refinement by means of reduction of the least-squares functional. We give the total number of degrees of freedom (DOF) and the value of the least-squares

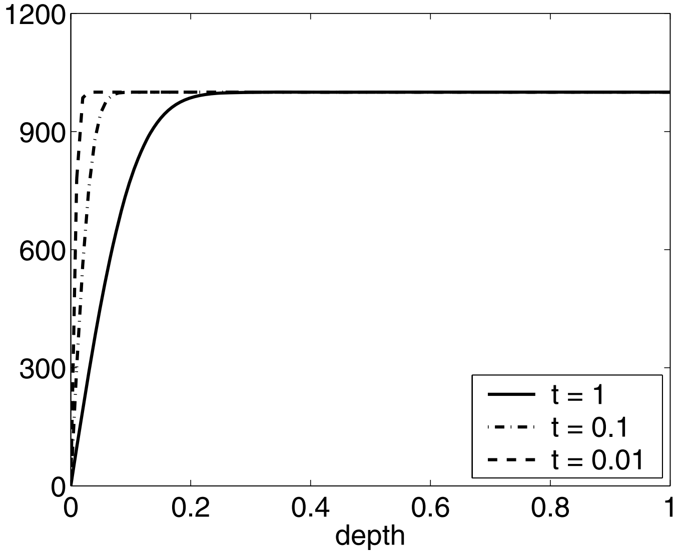


FIG. 6.1. *Pressure front for times* $t = 1, 0.1,$ *and* 0.01.

functional (LSF). Moreover, the ratio

$$\rho = \frac{\mathrm{LSF}|_{\text{current level}}}{\mathrm{LSF}|_{\text{previous level}}} \cdot \frac{\mathrm{DOF}|_{\text{current level}}}{\mathrm{DOF}|_{\text{previous level}}}$$

is listed in these tables. Ideally, for piecewise linear finite elements in two dimensions, the reduction rate of the least-squares functional

$$\frac{\mathrm{LSF}|_{\text{current level}}}{\mathrm{LSF}|_{\text{previous level}}} ,$$

which represents the square of the approximation error, should be inversely proportional to the increase in the number of unknowns

$$\frac{\mathrm{DOF}|_{\text{current level}}}{\mathrm{DOF}|_{\text{previous level}}} .$$

In other words, ρ should be roughly 1 for an optimal order of convergence in the adaptive process. In each of the tables, these numbers are given for times $t = k$ and $t = 10\,k$ to account for the time dependence of the problem.

The numbers show a nearly optimal convergence of the computed approximation according to a reduction of the least-squares functional proportional to the inverse of the increase in number of degrees of freedom (i.e., $\rho \approx 1$ in Tables 6.2–6.4). The estimates (5.1) and (5.5) from the previous section imply

$$\mathcal{F}_k(\boldsymbol{\sigma}_h, \mathbf{u}_h, \mathbf{w}_h, p_h; \mathbf{u}^{\text{old}}) \le \frac{C}{k} h^2 .$$

Fortunately, Tables 6.2, 6.3, and 6.4 show that the dependence of the functional minimum is by far not as dramatic as suggested by this formula in the actual computations.

Example 2. The second example is a true two-dimensional footing problem as given also in [17]. The simulation domain is a 8×5 m block of porous soil as given in

Table 6.2
Example 1: *Approximation for* $k = 100$.

	$t = k$			$t = 10\,k$		
Level	DOF	LSF	ρ	DOF	LSF	ρ
0	1725	1.7834e-04	-	1725	2.4235e-09	-
1	2373	8.2168e-05	0.634	2418	1.0607e-09	0.614
2	3471	4.6874e-05	0.834	3390	6.3210e-10	0.835
3	5073	2.8543e-05	0.890	4767	4.1312e-10	0.919
4	7431	1.7257e-05	0.886	6900	2.4882e-10	0.872
5	11022	1.1527e-05	0.991	9780	1.6488e-10	0.939
6	15504	8.3054e-06	1.014	13830	1.2034e-10	1.032
7	22461	5.4254e-06	0.946	19464	8.2040e-11	0.959

Table 6.3
Example 1: *Approximation for* $k = 1$.

	$t = k$			$t = 10\,k$		
Level	DOF	LSF	ρ	DOF	LSF	ρ
0	1725	5.4981e-03	-	1725	1.9241e-03	-
1	2328	3.1445e-03	0.772	2355	6.1131e-04	0.434
2	3336	2.1637e-03	0.986	3507	3.9092e-04	0.952
3	4668	1.5917e-03	1.029	4758	1.9642e-04	0.682
4	6837	1.1903e-03	1.095	6765	1.3249e-04	0.959
5	10163	8.1069e-04	1.012	10203	9.3223e-05	1.061
6	14692	5.7497e-04	1.025	14910	5.7600e-05	0.903
7	21550	4.2002e-04	1.071	21651	3.8023e-05	0.959

Table 6.4
Example 1: *Approximation for* $k = 0.001$.

	$t = k$			$t = 10\,k$		
Level	DOF	LSF	ρ	DOF	LSF	ρ
0	1725	3.8117e-02	-	1725	3.6349e-02	-
1	2490	3.1330e-02	1.186	2553	2.7627e-02	1.125
2	3669	2.3306e-02	1.096	3939	1.8255e-02	1.019
3	5433	1.7161e-02	1.090	6162	1.1972e-02	1.026
4	7773	1.2844e-02	1.071	9114	8.2663e-03	1.021
5	10660	9.8728e-03	1.054	14487	6.2077e-03	1.194
6	15619	7.5567e-03	1.121	21562	4.5356e-03	1.087
7	22054	5.6481e-03	1.055	30802	3.6467e-03	1.149

Figure 6.2. At the base of this domain the soil is assumed to be fixed and impervious, while at the upper left part of the domain a uniform load of intensity σ_0 is applied in a strip of length 1 m. The simulated domain then constitutes the right half of a model which is cut in the middle along its symmetry axis.

For the material parameters and the applied load the same values as in the previous example are used. As the initial condition we use $u = 0$ in Ω for $t = 0$, and the boundary data is given as follows:

$$\begin{array}{llll}
p = 0, & \sigma_{xy} = 0, & \sigma_{yy} = -\sigma_0 & \text{on } \{(x,y) \in \partial\Omega : y = 0, x \le 1\}, \\
p = 0, & \sigma_{xy} = 0, & \sigma_{yy} = 0 & \text{on } \{(x,y) \in \partial\Omega : y = 0, x > 1\}, \\
w = 0, & u_x = 0, & \sigma_{yx} = 0 & \text{on } \{(x,y) \in \partial\Omega : x = 0\}, \\
w = 0, & u_x = 0, & \sigma_{yx} = 0 & \text{on } \{(x,y) \in \partial\Omega : x = 8\}, \\
w = 0, & \sigma_{xy} = 0, & u_y = 0 & \text{on } \{(x,y) \in \partial\Omega : y = -5\}.
\end{array}$$

The finite element approximation spaces are also identical to those for Example 1.

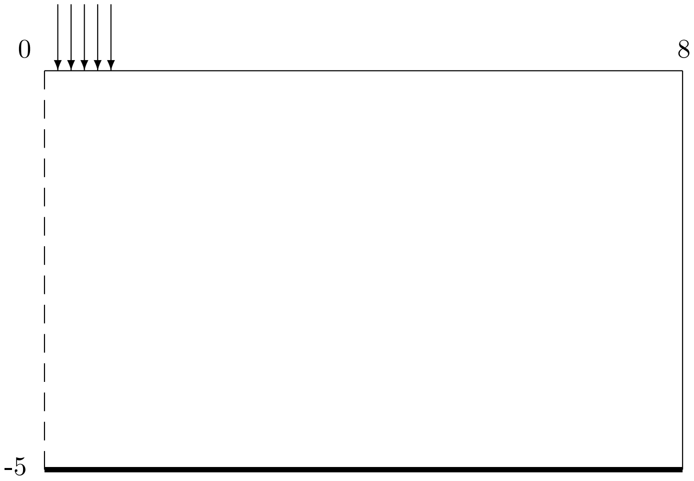


FIG. 6.2. *Example* 2: *Simulated domain with applied loads.*

TABLE 6.5
Example 2: *Approximation of stationary problem.*

Level	DOF	LSF	ρ
0	600	6.6029e-02	-
1	872	2.8771e-02	0.633
2	1379	1.5896e-02	0.874
3	2158	9.8770e-03	0.972
4	3349	6.4977e-03	1.021
5	5372	4.1945e-03	1.035
6	8135	2.8438e-03	1.027
7	12421	1.9623e-03	1.054
8	18775	1.3011e-03	1.002

For the consolidation of the material, the following three different stages have been considered in [17]. For $t = 0$, an incompressible Stokes system needs to be solved, and then the consolidation process starts. The stationary state of the system can finally be obtained by the solution of decoupled deformation and flow problems.

Table 6.5 shows the results for the approximation of the stationary problem using eight steps of adaptive refinement. In this example, the stationary solution is actually $p \equiv 0$ for the fluid flow problem and reduces to an elastic deformation problem. The nearly optimal behavior of the adaptive algorithm can be seen from the numbers in Table 6.5.

The solution of the time-dependent problem is shown for various t in Figure 6.3. The graphs show the increasingly steeper gradients with a singular behavior for $t \to 0$.

That the time-dependent problems become more difficult for decreasing k due to this singular nature of the problem can be observed in Tables 6.6, 6.7, and 6.8 where the numerical results are listed for time-step lengths $k = 10000$, 100, and 1 and for two different times $t = k$ and $t = 10\,k$. Six steps of adaptive refinement with an initial grid of 1348 degrees of freedom are used. The notation is the same as in Example 1.

The triangulation after four steps of adaptive refinement for $k = 1$ and $t = 3$ is shown in Figure 6.4.

If we compare our results with those of Murad and Loula [17], we see a qualita-

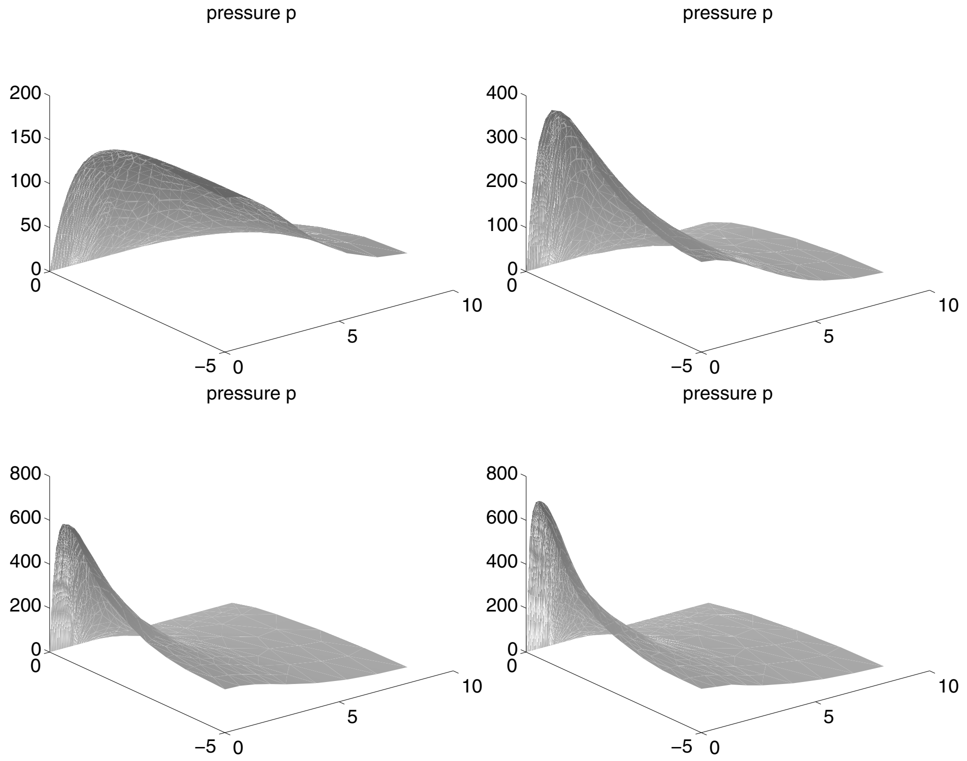


FIG. 6.3. *Pressure front for times* $t = 1000$, 100, 10, *and* 1.

TABLE 6.6

Example 2: *Approximation for* $k = 10000$.

	$t = k$			$t = 10\,k$		
Level	DOF	LSF	ρ	DOF	LSF	ρ
0	1348	2.4550e-02	-	1348	2.4646e-02	-
1	2043	1.2671e-02	0.782	2043	1.2741e-02	0.784
2	3225	7.7514e-03	0.966	3225	7.7915e-03	0.965
3	5022	4.9485e-03	0.994	5040	4.9251e-03	0.988
4	7884	3.3426e-03	1.060	7892	3.3626e-03	1.069
5	12189	2.1854e-03	1.011	12235	2.1872e-03	1.008
6	19003	1.4009e-03	0.999	18795	1.4257e-03	1.001

TABLE 6.7

Example 2: *Approximation for* $k = 100$.

	$t = k$			$t = 10\,k$		
Level	DOF	LSF	ρ	DOF	LSF	ρ
0	1348	4.3685e-02	-	1348	2.9035e-02	-
1	1971	2.5940e-02	0.868	2053	1.4850e-02	0.779
2	3045	1.8177e-02	1.083	3164	9.6709e-03	1.004
3	4868	1.1921e-02	1.049	4915	6.0689e-03	0.975
4	7343	8.0240e-03	1.015	7602	4.1349e-03	1.054
5	11086	5.4508e-03	1.026	11695	2.7232e-03	1.013
6	16675	3.7071e-03	1.023	17930	1.8196e-03	1.024

TABLE 6.8
Example 2: *Approximation for* $k = 1$.

	$t = k$			$t = 10\,k$		
Level	DOF	LSF	ρ	DOF	LSF	ρ
0	1348	2.2133e-01	-	1348	1.7277e-01	-
1	1981	1.7234e-01	1.144	2178	1.0841e-01	1.014
2	3223	1.3144e-01	1.241	3631	7.6777e-02	1.181
3	5128	1.0526e-01	1.274	5466	5.6833e-02	1.114
4	7746	8.5335e-02	1.225	8730	4.0558e-02	1.140
5	11831	6.7470e-02	1.208	13301	2.7934e-02	1.049
6	17558	5.4181e-02	1.192	20830	1.8262e-02	1.024

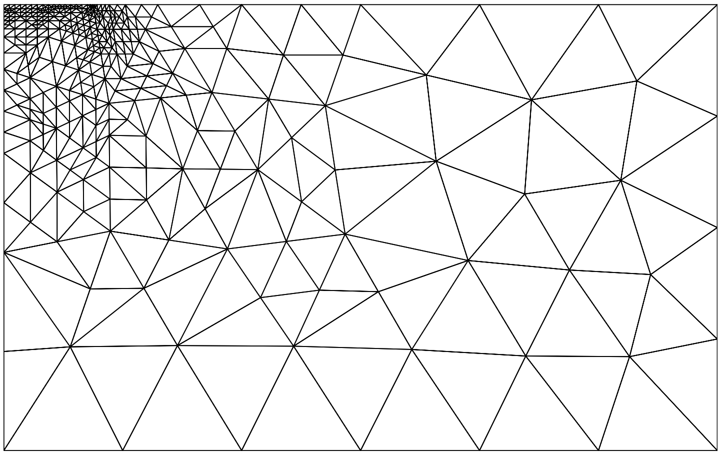


FIG. 6.4. *Adaptively refined grid for* $k = 1$ *and* $t = 3$.

tively similar behavior. The oscillations for small t which are present in the numerical results in [17], however, are avoided in our least-squares formulation. Adaptive refinement near the load boundary is crucial for the good performance of our least-squares approach as is the proper scaling of the individual terms in the least-squares functional.

Acknowledgments. The numerous discussions with Olaf Kolditz and Wenqing Wang at the Center for Applied Geosciences, University of Tübingen, have been of great importance for the work reported in this paper.

REFERENCES

[1] M. BERNDT, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Local error estimates and adaptive refinement for first-order system least squares*, Electron. Trans. Numer. Anal., 6 (1997), pp. 35–43.

[2] M. A. BIOT, *General theory of three-dimensional consolidation*, J. Appl. Phys., 12 (1941), pp. 155–164.

[3] D. BRAESS, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.

[4] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer, New York, 2002.

[5] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.

[6] Z. Cai, J. Korsawe, and G. Starke, *An adaptive least squares mixed finite element method for the stress-displacement formulation of linear elasticity*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 132–148.

[7] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[8] Z. Cai, B. Lee, and P. Wang, *Least-squares methods for incompressible Newtonian fluid flow: Linear stationary problems*, SIAM J. Numer. Anal., 42 (2004), pp. 843–859.

[9] Z. Cai and G. Starke, *First-order system least squares for the stress-displacement formulation: Linear elasticity*, SIAM J. Numer. Anal., 41 (2003), pp. 715–730.

[10] Z. Cai and G. Starke, *Least-squares methods for linear elasticity*, SIAM J. Numer. Anal., 42 (2004), pp. 826–842.

[11] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology, Volume* 2: *Functional and Variational Methods*, Springer, Berlin, 1988.

[12] R. de Boer, *Theory of Porous Media*, Springer, Berlin, 2000.

[13] W. Ehlers and G. Eipper, *Finite elastic deformations in liquid-saturated and empty porous solids*, Transp. Porous Media, 34 (1999), pp. 179–191.

[14] J. Korsawe and G. Starke, *A least squares mixed finite element method for variably saturated subsurface flow in deformable porous media*, in Proceedings of the 4th Workshop on Porous Media, O. Kolditz, S. Bauer, and M. Xie, eds., Center for Applied Geosciences, University of Tübingen, Germany, 2003.

[15] R. W. Lewis and B. A. Schrefler, *The Finite Element Method in the Static and Dynamic Deformation and Consolidation of Porous Media*, Wiley, Chichester, UK, 1998.

[16] M. Murad and A. Loula, *Improved accuracy in finite element analysis of Biot's consolidation problem*, Comput. Methods Appl. Mech. Engrg., 95 (1992), pp. 359–382.

[17] M. Murad and A. Loula, *On stability and convergence of finite element approximations of Biot's consolidation problem*, Internat. J. Numer. Methods Engrg., 37 (1994), pp. 645–667.

[18] M. A. Murad, V. Thomée, and A. F. D. Loula, *Asymptotic behavior of semidiscrete finite-element approximations of Biot's consolidation problem*, SIAM J. Numer. Anal., 33 (1996), pp. 1065–1083.

[19] G. Starke, *Gauss–Newton multilevel methods for least-squares finite element computations of variably saturated subsurface flow*, Computing, 64 (2000), pp. 323–338.

[20] G. Starke, *Least-squares mixed finite element solution of variably saturated subsurface flow problems*, SIAM J. Sci. Comput., 21 (2000), pp. 1869–1885.

[21] W. Wang and O. Kolditz, *Numerical Analysis of Elasto-plastic Consolidation in Porous Media*, Tech. report, Center for Applied Geosciences, University of Tübingen, Germany, 2003.

[22] C. Wieners, M. Ammann, S. Diebels, and W. Ehlers, *Parallel* 3*-d simulations for porous media models in soil mechanics*, Comput. Mech., 29 (2002), pp. 75–87.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 340–362

ON LEAST-SQUARES FINITE ELEMENT METHODS FOR THE POISSON EQUATION AND THEIR CONNECTION TO THE DIRICHLET AND KELVIN PRINCIPLES[*]

PAVEL BOCHEV[†] AND MAX GUNZBURGER[‡]

Abstract. Least-squares finite element methods for first-order formulations of the Poisson equation are not subject to the inf-sup condition and lead to stable solutions even when all variables are approximated by equal-order continuous finite element spaces. For such elements, one can also prove optimal convergence in the "energy" norm (equivalent to a norm on $H^1(\Omega) \times H(\Omega, \text{div})$) for all variables and optimal L^2 convergence for the scalar variable. However, showing optimal L^2 convergence for the flux has proven to be impossible without adding the redundant curl equation to the first-order system of partial differential equations. In fact, numerical evidence strongly suggests that nodal continuous flux approximations do not posses optimal L^2 accuracy. In this paper, we show that optimal L^2 error rates for the flux can be achieved without the curl constraint, provided that one uses the div-conforming family of Brezzi–Douglas–Marini or Brezzi–Douglas–Duran–Fortin elements. Then, we proceed to reveal an interesting connection between a least-squares finite element method involving $H(\Omega, \text{div})$-conforming flux approximations and mixed finite element methods based on the classical Dirichlet and Kelvin principles. We show that such least-squares finite element methods can be obtained by approximating, through an L^2 projection, the Hodge operator that connects the Kelvin and Dirichlet principles. Our principal conclusion is that when implemented in this way, a least-squares finite element method combines the best computational properties of finite element methods based on each of the classical principles.

Key words. least-squares finite element methods, Kelvin principle, Dirichlet principle, BDM spaces, BDDF spaces, RT spaces, mixed methods, Poisson equation

AMS subject classifications. 76D05, 76D07, 65F10, 65F30

DOI. 10.1137/S003614290443353X

1. Introduction.

Stable mixed finite element methods for the Poisson equation[1] (written in a first-order form in terms of a scalar variable and a flux) require the use of finite element spaces that satisfy an appropriate inf-sup condition. For methods based on the Dirichlet principle, the inf-sup condition can be easily satisfied but for the dual Kelvin principle, it imposes complicated restrictions on the choice of spaces; see [11]. In either case, it is well known that pairs of standard nodal-based, continuous finite element spaces fail the inf-sup condition and lead to unstable mixed methods. It is also well known that the inf-sup condition is circumvented if one uses such simple element pairs in finite element methods based on L^2 least-squares variational principles. Ever since such least-squares finite element methods for first-order formulations of the Poisson equation were first considered in [24], this fact

[*] Received by the editors August 25, 2003; accepted for publication September 3, 2004; published electronically June 14, 2005. This work was funded by the Applied Mathematical Sciences program, U.S. Department of Energy, Office of Energy Research, and performed at Sandia National Labs, a multiprogram laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC-94AL85000.

http://www.siam.org/journals/sinum/43-1/43353.html

[†] Computational Mathematics and Algorithms Department, Sandia National Laboratories, Albuquerque, NM 87185-1110 (pbboche@sandia.gov).

[‡] School of Computational Science, Florida State University, Tallahassee, FL 32306-4120 (gunzburg@csit.fsu.edu). The research of this author was supported in part by CSRI, Sandia National Laboratories, under contract 18407.

[1] Although we consider only the Poisson problem, much of what we discuss can be easily extended to more general second-order elliptic partial differential equations.

has been deemed as an important advantage of those methods over mixed Galerkin methods.

Already in [24], optimal L^2 error estimates for least-square finite element methods were established for the scalar variable; however, there, no optimal L^2 convergence results were obtained for nodal approximations of the flux. This situation persisted in all subsequent analyses: optimal L^2 error estimates for the flux could not be obtained[2] without the addition of a "redundant" curl equation; see, e.g., [13, 14, 15, 25, 27]. Moreover, computational studies in [16] strongly suggested that optimal L^2 convergence may in fact be nearly impossible if one uses pairs of standard nodal-based continuous finite element spaces. A notable exception was a case studied in [16] for which optimal L^2 error estimates for both the scalar variable and the flux were obtained when these variables were approximated by continuous nodal spaces built on a criss-cross grid. The key to their proof was the validity of the grid decomposition property (GDP) which was established for the criss-cross grid in [17]. So far, the criss-cross grid remains the only known case of a continuous nodal-based finite element space for which the GDP can be verified. More importantly, it was shown in [17] (see also [7]) that the GDP is necessary and sufficient for the stability of the mixed finite element method based on the Kelvin principle.

The correlation between the stability of mixed finite element methods and the optimal accuracy of least-squares finite element methods, established in [16], opens up the intriguing possibility that optimal L^2 accuracy for the flux may be obtainable for a least-squares finite element method, provided that this variable is approximated by $H(\Omega, \mathrm{div})$-conforming elements that are stable for mixed finite element methods based on the Kelvin principle. Today, the stability of mixed finite element methods based on the Kelvin principle is well understood, and many examples of stable finite element pairs are known. The first goal of our study is to show that the use of some of these spaces in a least-squares finite element method will indeed help to improve the L^2 accuracy of the flux approximation. Our second goal is to offer a new perspective on least-squares principles as resulting from a choice for the approximation of the Hodge $*$-operator that connects two mutually dual "energy" principles. Among other things, such an interpretation shows, in our context, why the use of $H(\Omega, \mathrm{div})$-conforming elements is in fact more natural than the use of equal-order C^0 spaces.

While our conclusions may disappoint the adherents of equal-order implementations, our results do not void least-squares finite element methods as a viable computational alternative. To the contrary, they demonstrate that when implemented correctly, a *least-squares finite element method combines the best computational properties of finite element methods based on both the Dirichlet and Kelvin principles,* and at the same time manages to avoid the indefinite linear systems that make the latter more difficult to solve. Although we reach this conclusion in the specific context of mixed and least-squares finite element methods for the Poisson problem, the idea of defining the latter type of method so that it inherits the best characteristics of a pair of mixed methods that are related through duality may have considerably wider application.

In the rest of this section, we briefly review the notation used throughout the paper. Then, in section 2.1, we recall the Dirichlet and Kelvin principles and the

[2] A somewhat different situation exists for negative-norm-based least-squares finite element methods, for which it is known that the L^2 accuracy of the flux is optimal with respect to the spaces used; however, for such methods, no error bound for the divergence of the flux could be established; see [10].

associated first-order div-grad formulation of the boundary value problem for the Poisson equation. There, in the context of the Kelvin principle, we also review basic definitions and properties of stable $H(\Omega, \mathrm{div})$-conforming mixed finite elements spaces for the flux and show that they satisfy the GDP. For the sake of brevity, we restrict attention to the well-known Raviart–Thomas (RT), Brezzi–Douglas–Marini (BDM), and Brezzi–Douglas–Duran–Fortin (BDDF) classes of affine families of finite elements. Section 3 deals with least-squares finite element methods for first-order formulations of the Poisson problem. After a brief review of known error estimates in $H^1(\Omega) \times H(\Omega, \mathrm{div})$, we turn our attention to the L^2 accuracy and the rarely discussed case of least-squares finite element methods using RT, BDM, or BDDF approximations of the flux. We show that BDM and BDDF spaces lead to optimal convergence of the flux in L^2. In section 4, we offer an interpretation of such least-squares finite element methods which is derived with the help of some notions from exterior calculus and differential forms.

1.1. Notation. Throughout, Ω denotes a bounded region in $\mathbb{R}^n$, $n = 2, 3$, with a Lipschitz continuous boundary $\Gamma = \partial\Omega$. We assume that Γ consists of two disjoint parts denoted by Γ_D and Γ_N. For $p > 0$, $H^p(\Omega)$ denotes the Sobolev space of order p with norm and inner product denoted by $\|\cdot\|_p$ and $(\cdot,\cdot)_p$, respectively. When $p = 0$, we use the standard notation $L^2(\Omega)$. The symbol $|\cdot|_k$, $0 \le k \le p$, denotes the kth seminorm on $H^p(\Omega)$. Vector-valued functions and vector analogues of the Sobolev spaces are denoted by lower- and upper case bold-face font, respectively, e.g., $\mathbf{u}$, $\mathbf{H}^1(\Omega)$, $\mathbf{L}^2(\Omega)$, etc. We recall the space

$$H(\Omega, \mathrm{div}) = \{\mathbf{u} \in \mathbf{L}^2(\Omega) \mid \nabla\cdot\mathbf{u} \in L^2(\Omega)\},$$

which is a Hilbert space when equipped with the norm

$$\|\mathbf{u}\|_{H(\Omega,\mathrm{div})} = (\|\mathbf{u}\|_0^2 + \|\nabla\cdot\mathbf{u}\|_0^2)^{1/2}.$$

To deal with the boundary conditions, we introduce the spaces

$$H^1_D(\Omega) = \{\phi \in H^1(\Omega) \mid \phi = 0 \quad \text{on } \Gamma_D\}$$

and

$$H_N(\Omega, \mathrm{div}\,) = \{\mathbf{v} \in H(\Omega, \mathrm{div}) \mid \mathbf{v}\cdot\mathbf{n} = 0 \quad \text{on } \Gamma_N\}.$$

Details about all the notation just introduced may be found, e.g., in [11, 18].

Throughout, we will refer to the problem

$$-\Delta\phi + \gamma\phi = f \quad \text{in } \Omega, \qquad \phi = 0 \quad \text{on } \Gamma_D, \quad \text{and} \quad \partial\phi/\partial n = 0 \quad \text{on } \Gamma_N \tag{1.1}$$

as the *Poisson problem*, even though that terminology is usually reserved for the case $\gamma = 0$.

2. Mixed finite element methods for the Poisson problem.

So as to provide a background for some of the discussions of sections 3 and 4 concerning least-squares finite element methods, we consider, in this section, primal and dual mixed finite element methods for the Poisson problem.

2.1. The generalized Dirichlet and Kelvin principles. The Dirichlet and Kelvin principles arise in a variety of applications. Mathematically, they provide two variational formulations for the Poisson problem and also form the basis for defining mixed finite element methods for approximations of the solution of that problem.

2.1.1. The generalized Dirichlet principle. Consider the functional

$$D(\phi, \mathbf{w}; f) = \frac{1}{2}\int_\Omega \left(|\mathbf{w}|^2 + \gamma|\phi|^2\right) d\Omega - \int_\Omega f\phi \, d\Omega$$

and the minimization problem

$$\min_{(\phi,\mathbf{w})\in H^1_D(\Omega)\times\nabla H^1_D(\Omega)} D(\phi, \mathbf{w}; f) \qquad \text{subject to} \qquad \mathbf{w} + \nabla\phi = 0, \tag{2.1}$$

where $\gamma \geq 0$ is a given function that is assumed to satisfy $\|\gamma\|_{L^\infty(\Omega)} \leq C$ for some constant $C \geq 0$. The minimization principle (2.1) is known as the (generalized) *Dirichlet principle*.[3] Although the constraint $\mathbf{w} + \nabla\phi = 0$ can be directly substituted into the functional to eliminate the flux $\mathbf{w}$,[4] it will be more profitable for our discussions to continue to consider the form (2.1).

With the help of a Lagrange multiplier $\mathbf{u}$ to enforce the constraint $\mathbf{w} + \nabla\phi = 0$ and the Lagrangian functional

$$L_D(\phi, \mathbf{w}, \mathbf{u}; f) = \frac{1}{2}\int_\Omega \left(|\mathbf{w}|^2 + \gamma|\phi|^2\right) d\Omega - \int_\Omega f\phi \, d\Omega - \int_\Omega \mathbf{u}\cdot(\mathbf{w} + \nabla\phi)\, d\Omega,$$

the constrained minimization problem (2.1) can be transformed into the unconstrained optimization problem of determining saddle-points $(\phi, \mathbf{w}, \mathbf{u}) \in H^1_D(\Omega) \times \nabla H^1_D(\Omega) \times \nabla H^1_D(\Omega)$ of $L_D(\phi, \mathbf{w}, \mathbf{u}; f)$. It is not difficult to see that the optimality system obtained by setting the first variations of $L_D(\phi, \mathbf{w}, \mathbf{u}; f)$ to zero is given by the following: seek $(\phi, \mathbf{w}, \mathbf{u}) \in H^1_D(\Omega) \times \nabla H^1_D(\Omega) \times \nabla H^1_D(\Omega)$ such that

$$\begin{cases} \displaystyle\int_\Omega \mathbf{w}\cdot\mathbf{v}\, d\Omega + \int_\Omega \nabla\phi\cdot\mathbf{v}\, d\Omega = 0 & \forall\, \mathbf{v} \in \nabla H^1_D(\Omega), \\ \displaystyle\int_\Omega (\mathbf{w} - \mathbf{u})\cdot\mathbf{q}\, d\Omega = 0 & \forall\, \mathbf{q} \in \nabla H^1_D(\Omega), \\ \displaystyle -\int_\Omega \mathbf{u}\cdot\nabla\psi\, d\Omega + \int_\Omega \gamma\phi\psi\, d\Omega = \int_\Omega f\psi\, d\Omega & \forall\, \psi \in H^1_D(\Omega). \end{cases} \tag{2.3}$$

The first and second equations may be easily combined to yield the simplified system

$$\begin{cases} \displaystyle\int_\Omega \mathbf{u}\cdot\mathbf{v}\, d\Omega + \int_\Omega \nabla\phi\cdot\mathbf{v}\, d\Omega = 0 & \forall\, \mathbf{v} \in \nabla H^1_D(\Omega), \\ \displaystyle\int_\Omega \nabla\psi\cdot\mathbf{u}\, d\Omega - \int_\Omega \gamma\psi\phi\, d\Omega = -\int_\Omega f\psi\, d\Omega & \forall\, \psi \in H^1_D(\Omega), \end{cases} \tag{2.4}$$

involving only $\phi \in H^1_D(\Omega)$ and $\mathbf{u} \in \nabla H^1_D(\Omega)$.

If solutions to the constrained minimization problem (2.1) or, equivalently, of (2.4), are sufficiently smooth, then without much difficulty one obtains that

[3] For $f = 0$, $\gamma = 0$, and appropriate boundary conditions, the Dirichlet principle in the inviscid fluid mechanics setting states that among all irrotational velocity fields, the one that minimizes the kinetic energy is the solenoidal one. In the solid mechanics setting, $\mathbf{w}$ is a tensor, and a simplified version of (2.1) is the *energy minimization principle.*

[4] This results in the certainly more familiar form for the (generalized) Dirichlet principle:

$$\min_{\phi\in H^1_D(\Omega)} \widetilde{D}(\phi; f), \quad \text{where} \quad \widetilde{D}(\phi; f) = \frac{1}{2}\int_\Omega (|\nabla\phi|^2 + \gamma|\phi|^2)\, d\Omega - \int_\Omega f\psi\, d\Omega. \tag{2.2}$$

$$\begin{cases} \nabla\cdot\mathbf{u}+\gamma\phi = f \quad \text{and} \quad \mathbf{u}+\nabla\phi=\mathbf{0} \quad \text{in } \Omega, \\ \phi = 0 \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{u}\cdot\mathbf{n}=0 \quad \text{on } \Gamma_N. \end{cases} \tag{2.5}$$

Note that from the second equation of (2.3) we also have that $\mathbf{w}=\mathbf{u}$. Eliminating the flux $\mathbf{u}$ from (2.5) (again assuming that sufficient smoothness is available), one obtains the Poisson problem[5] (1.1) for the scalar variable ϕ.

2.1.2. The generalized Kelvin principle. Now consider the functional

$$K(\lambda,\mathbf{u}) = \frac{1}{2}\int_\Omega \left(|\mathbf{u}|^2+\gamma|\lambda|^2\right) d\Omega$$

and the minimization problem

$$\min_{\lambda\in L^2(\Omega),\ \mathbf{u}\in H_N(\Omega,\mathrm{div})} K(\lambda,\mathbf{v}) \quad \text{subject to } \nabla\cdot\mathbf{u}+\gamma\lambda=f. \tag{2.7}$$

The minimization principle (2.7) is known as the (generalized) *Kelvin principle*;[6] it is dual to the (generalized) Dirichlet principle.[7]

With the help of a Lagrange multiplier ϕ to enforce the constraint and the Lagrangian functional

$$L_K(\lambda,\mathbf{u},\phi;f) = \frac{1}{2}\int_\Omega \left(|\mathbf{u}|^2+\gamma|\lambda|^2\right) d\Omega - \int_\Omega \phi(\nabla\cdot\mathbf{u}+\gamma\lambda-f)\, d\Omega,$$

the constrained minimization problem (2.7) can be transformed into the unconstrained problem of determining saddle-points $(\lambda,\mathbf{u},\phi)\in L^2(\Omega)\times H_N(\Omega,\mathrm{div})\times L^2(\Omega)$ of $L_K(\lambda,\mathbf{u},\phi;f)$. It is not difficult to see that the optimality system obtained by setting

[5] Note that since $\nabla\psi\in\mathbf{L}^2(\Omega)$, one can easily combine the two equations in (2.4) to yield the more familiar weak formulation

$$\int_\Omega \nabla\phi\cdot\nabla\psi\, d\Omega + \int_\Omega \gamma\psi\phi\, d\Omega = \int_\Omega f\psi\, d\Omega \qquad \forall\,\psi\in H^1_D(\Omega) \tag{2.6}$$

for the Poisson problem (1.1). Again, it will be more profitable for our discussion to continue to use (2.4) instead of the more familiar form (2.6).

[6] Setting $f=0$ and $\gamma=0$ and allowing for an inhomogeneous boundary condition for $\mathbf{u}\cdot\mathbf{n}$, the Kelvin principle for inviscid flows states that, among all incompressible velocity fields, the one that minimizes the kinetic energy is irrotational. In structural mechanics (where $\mathbf{u}$ is a tensor), a simplified version of (2.7) is known as the *complimentary energy principle*.

[7] Unlike the case of the Dirichlet principle, if $\gamma=0$, one cannot directly use the constraint $\nabla\cdot\mathbf{u}+\gamma\lambda=f$ to eliminate one of the variables. If $\gamma>0$, then it is possible to use the constraint to eliminate the scalar variable λ. In fact, in the latter case we are led to the problem

$$\min_{\mathbf{u}\in H_N(\Omega,\mathrm{div})} \widetilde{K}(\mathbf{u};f), \quad \text{where} \quad \widetilde{K}(\mathbf{u};f)=\frac{1}{2}\int_\Omega \left(|\mathbf{u}|^2+\frac{1}{\gamma}|\nabla\cdot\mathbf{u}-f|^2\right) d\Omega. \tag{2.8}$$

Comparing (2.2) and (2.8), we already see a big difference between the Kelvin and Dirichlet principles, in addition to the obvious difficulty seen in (2.8) for the case $\gamma=0$. The functional $\widetilde{D}(\cdot;f)$ in (2.2) involves all first derivatives of the scalar variable ϕ, which is why we can minimize it over the space $H^1_D(\Omega)$. On the other hand, the functional $\widetilde{K}(\cdot;f)$ in (2.8) only involves the combination $\nabla\cdot\mathbf{u}$ of first derivatives of the flux $\mathbf{u}$, which is why we can minimize it only with respect to a subspace of $H(\Omega,\mathrm{div})$, and not with respect to $\mathbf{H}^1(\Omega)$.

the first variations of $L(\lambda, \mathbf{u}, \phi; f)$ to zero is given by the following: seek $(\lambda, \mathbf{u}, \phi) \in L^2(\Omega) \times H_N(\Omega, \mathrm{div}\,) \times L^2(\Omega)$ such that

$$\text{(2.9)} \quad \begin{cases} \displaystyle\int_\Omega \mathbf{u} \cdot \mathbf{v}\, d\Omega - \int_\Omega \phi \nabla \cdot \mathbf{v}\, d\Omega = 0 & \forall\, \mathbf{v} \in H_N(\Omega, \mathrm{div}\,), \\ \displaystyle\int_\Omega \gamma(\lambda - \phi)\mu\, d\Omega = 0 & \forall\, \mu \in L^2(\Omega), \\ \displaystyle -\int_\Omega \psi \nabla \cdot \mathbf{u}\, d\Omega - \int_\Omega \gamma\psi\lambda\, d\Omega = -\int_\Omega f\psi\, d\Omega & \forall\, \psi \in L^2(\Omega). \end{cases}$$

For $\gamma \neq 0$, the second and third equations may be easily combined to yield the simplified system

$$\text{(2.10)} \quad \begin{cases} \displaystyle\int_\Omega \mathbf{u} \cdot \mathbf{v}\, d\Omega - \int_\Omega \phi \nabla \cdot \mathbf{v}\, d\Omega = 0 & \forall\, \mathbf{v} \in H_N(\Omega, \mathrm{div}\,), \\ \displaystyle -\int_\Omega \psi \nabla \cdot \mathbf{u}\, d\Omega - \int_\Omega \gamma\psi\phi\, d\Omega = -\int_\Omega f\psi\, d\Omega & \forall\, \psi \in L^2(\Omega), \end{cases}$$

involving only $\phi \in L^2(\Omega)$ and $\mathbf{u} \in H_N(\Omega, \mathrm{div}\,)$. If $\gamma = 0$, then (2.9) directly reduces to (2.10) so that the latter holds for any $\gamma \geq 0$.

The duality of the Dirichlet and Kelvin principles extends to the optimality systems (2.4) and (2.10). For example, they are respectively described using the dual operators ∇ and $-\nabla\cdot$. The domain of ∇ is all of $H^1_D(\Omega)$, while its range is a constrained subspace of $\mathbf{L}^2(\Omega)$ consisting of irrotational functions. In contrast, the domain of $\nabla\cdot$ is a constrained subspace of $\mathbf{L}^2(\Omega)$ and its range is all of $L^2(\Omega)$. We note again the difference between the domain spaces of the two operators: $H^1_D(\Omega)$ involves all first derivatives of the scalar variable, while $H_N(\Omega, \mathrm{div})$ only involves a combination of first derivatives of the flux.

2.2. Stable mixed finite element spaces. Finite element approximations of the mixed problems (2.4) and (2.10) are not stable unless the spaces chosen to approximate ϕ and $\mathbf{u}$ satisfy the inf-sup condition. To keep our presentation reasonably short and devoid of unnecessary technical details, we focus on affine families of stable spaces defined on simplicial triangulations $\mathcal{T}_h$ of the domain Ω into elements $\mathcal{K}$. In two dimensions, $\mathcal{K}$ are triangles, and in three dimensions, they are tetrahedra. The symbol $P_k(\mathcal{K})$ denotes the space of all polynomials of degree less than or equal to k defined on $\mathcal{K}$.

Nodal C^0 finite element spaces built from mth degree polynomials, $m \geq 1$, are denoted by[8] $\mathcal{W}^0_m(\Omega)$. We recall that there exists an interpolation operator $\mathcal{I}_0$ into $\mathcal{W}^0_m(\Omega)$ such that for any $\phi \in H^{m+1}(\Omega)$,

$$\text{(2.11)} \qquad \|\phi - \mathcal{I}_0\phi\|_0 + h\|\nabla(\phi - \mathcal{I}_0\phi)\|_0 \leq Ch^{m+1}|\phi|_{m+1}.$$

[8] The reasoning leading to the choice of notation $\mathcal{W}^i_k$ for the finite element spaces we employ will become clear later.

We denote by $\mathcal{W}_m^1(\Omega)$ the space $\nabla(\mathcal{W}_m^0(\Omega))$.[9] We will use the pair of finite element spaces $\mathcal{W}_m^0(\Omega)$ and $\mathcal{W}_m^1(\Omega)$ to discretize the Dirichlet principle.

For the Kelvin principle, we will use the[10] BDM_k and RT_k spaces on Ω that are built from the individual element spaces

$$\mathrm{BDM}_k(\mathcal{K}) = (P_k(\mathcal{K}))^n \quad \text{and} \quad \mathrm{RT}_k(\mathcal{K}) = (P_k(\mathcal{K}))^n + \mathbf{x}P_k(\mathcal{K})$$

in a manner that ensures the continuity of the normal component across element boundaries; see [11, pp. 113–116] for details and definitions of the corresponding element degrees of freedom. Since BDM_k and RT_k both contain complete polynomials of degree k, their approximation properties in L^2 are the same. In particular, one can show that for either the BDM_k or RT_k spaces there exists an interpolation operator $\mathcal{I}_2$ such that

$$\|\mathbf{u} - \mathcal{I}_2\mathbf{u}\|_0 \leq Ch^r|\mathbf{u}|_r \quad \forall\, \mathbf{u} \in \mathbf{H}^r(\Omega) \text{ and } 1 \leq r \leq k+1. \tag{2.12}$$

Since RT_k also contains the higher-degree polynomial component $\mathbf{x}P_k(\mathcal{K})$, it has better accuracy in $H(\Omega, \mathrm{div})$ than does BDM_k. Note, however, that this additional component does not help to improve the L^2 accuracy of RT_k spaces because it does not increase the order of complete polynomials contained in RT_k to $k+1$. In summary, we have the following estimates for the error in the divergence of the interpolant (see [11, p. 132]):

$$\|\nabla \cdot (\mathbf{u} - \mathcal{I}_2\mathbf{u})\|_0 \leq Ch^k\|\nabla \cdot \mathbf{u}\|_k \qquad \text{for } \mathrm{BDM}_k \text{ spaces} \tag{2.13}$$

and

$$\|\nabla \cdot (\mathbf{u} - \mathcal{I}_2\mathbf{u})\|_0 \leq Ch^{k+1}\|\nabla \cdot \mathbf{u}\|_{k+1} \qquad \text{for } \mathrm{RT}_k \text{ spaces.} \tag{2.14}$$

In what follows, we will denote by $\mathcal{W}_k^2(\Omega)$ the RT and BDM spaces having *equal approximation orders with respect to the divergence operator*, i.e.,

$$\mathcal{W}_k^2(\Omega) = \{\mathbf{v} \in H(\Omega, \mathrm{div}) | \mathbf{v}|_{\mathcal{K}} \in \mathcal{W}_k^2(\mathcal{K})\},$$

where $\mathcal{W}_k^2(\mathcal{K})$ is one of the finite element spaces $\mathrm{RT}_{k-1}(\mathcal{K})$ or $\mathrm{BDM}_k(\mathcal{K})$. We can now combine (2.13) and (2.14) into a single statement: there exists an interpolation operator $\mathcal{I}_2$ into $\mathcal{W}_k^2(\Omega)$ such that

$$\|\nabla \cdot (\mathbf{u} - \mathcal{I}_2\mathbf{u})\|_0 \leq Ch^k\|\nabla \cdot \mathbf{u}\|_k. \tag{2.15}$$

Note, however, that from (2.12) we have that the interpolation operator $\mathcal{I}_2$ into $\mathcal{W}_k^2(\Omega)$ satisfies

$$\|\mathbf{u} - \mathcal{I}_2\mathbf{u}\|_0 \leq Ch^r|\mathbf{u}|_r \qquad \begin{cases} \text{for } 1 < r \leq k & \text{if } \mathcal{W}_k^2(\mathcal{K}) = \mathrm{RT}_{k-1}, \\ \text{for } 1 < r \leq k+1 & \text{if } \mathcal{W}_k^2(\mathcal{K}) = \mathrm{BDM}_k. \end{cases} \tag{2.16}$$

[9] In our setting, $\mathcal{W}_m^1(\Omega)$ is a space of vector-valued functions that are discontinuous with respect to the simplicial triangulation $\mathcal{T}_h$ and whose components belong to a subspace $P_{m-1}(\mathcal{K})$ in each $\mathcal{K}$. Functions belonging to $\mathcal{W}_m^1(\Omega)$ must be curl-free within each element $\mathcal{K}$ (since they are gradients of function belonging to $\mathcal{W}_m^0(\Omega)$), so that, except for $m = 1$, they are not complete $(m-1)$st degree polynomials. However, the precise, explicit characterization of $\mathcal{W}_m^1(\Omega)$, e.g., the construction of a basis, is not difficult (using their irrotational property), and moreover, as we shall see, it turns out not to be necessary in practice. For future reference, we note that functions belonging to the approximating space $\mathcal{W}_m^0(\Omega)$ for the scalar variable are continuous across element boundaries, so that the tangential components of functions belonging to the approximating space $\mathcal{W}_m^1(\Omega) = \nabla(\mathcal{W}_m^0(\Omega))$ for the flux are *automatically* also continuous across element boundaries.

[10] To simplify notation, from now on we will denote both the BDM and BDDF spaces simply by BDM.

We denote by $\mathcal{W}_k^3(\Omega)$ the space $\nabla\cdot(\mathcal{W}_k^2(\Omega))$. For mixed finite element methods based on the Kelvin principle, we will use the finite element spaces RT_{k-1} or BDM_k to approximate the flux. For characterizations of these spaces and the associated spaces $\mathcal{W}_k^3(\Omega) = \nabla\cdot(\mathcal{W}_k^2(\Omega))$ for the scalar variable, see [11].

2.2.1. Stable mixed finite element spaces for the Dirichlet principle. A mixed finite element method based on the Dirichlet principle may be defined by discretizing (2.4), i.e.,

$$\text{(2.17)}\qquad \begin{cases} \displaystyle\int_\Omega \mathbf{u}_h\cdot\mathbf{v}_h\,d\Omega + \int_\Omega \nabla\phi_h\cdot\mathbf{v}_h\,d\Omega = 0 & \forall\,\mathbf{v}_h\in\mathcal{W}_m^1(\Omega),\\ \displaystyle\int_\Omega \nabla\psi_h\cdot\mathbf{u}_h\,d\Omega - \int_\Omega \gamma\psi_h\phi_h\,d\Omega = -\int_\Omega f\psi_h\,d\Omega & \forall\,\psi\in\mathcal{W}_m^0(\Omega). \end{cases}$$

Since $\mathcal{W}_m^1(\Omega)\equiv\nabla(\mathcal{W}_m^0(\Omega))$, note that, even at the discrete level, we may again eliminate the flux approximation to obtain the equivalent discrete problem

$$\text{(2.18)}\qquad \int_\Omega \nabla\phi_h\cdot\nabla\psi_h\,d\Omega + \int_\Omega \gamma\psi_h\phi_h\,d\Omega = \int_\Omega f\psi_h\,d\Omega \qquad \forall\,\psi\in\mathcal{W}_m^0(\Omega),$$

which we recognize as the standard Galerkin discretization of (2.2) or (2.6). In fact, using the pair of spaces $\mathcal{W}_m^0(\Omega)$ and $\mathcal{W}_m^1(\Omega)$ for approximating the scalar variable and the flux, respectively, in the discretization (2.17) of (2.4) is equivalent[11] to using the scalar space $\mathcal{W}_m^0(\Omega)$ in the standard Galerkin discretization (2.18) of (2.2) and then letting the approximation of the flux be the gradient of the resulting approximation of the scalar variable.

In this way we see that for discretizations of (2.4), i.e., the Dirichlet principle, the required inf-sup condition is completely benign in the sense that it can be avoided by eliminating the flux approximation $\mathbf{u}_h$ from (2.17), then solving (2.18) for the approximation ϕ_h of the scalar variable using a standard continuous nodal finite element space $\mathcal{W}_m^0(\Omega)$, and, at the end, determining the approximation to the flux from the exact relation $\mathbf{u}_h = -\nabla\phi_h$. The required inf-sup condition is implicitly satisfied by the pair of spaces $\mathcal{W}_m^0(\Omega)$ and $\mathcal{W}_m^1(\Omega) = \nabla(\mathcal{W}_m^0(\Omega))$. If one insists on solving (2.4), then one needs to explicitly produce a basis for $\mathcal{W}_m^1(\Omega)$; this is easily accomplished.

From either (2.17) or (2.18) one obtains, for the Dirichlet principle, that if $\phi\in H^{m+1}(\Omega)\cap H_D^1(\Omega)$, then

$$\text{(2.19)}\qquad \|\phi-\phi_h\|_0 \le h^{m+1}\|\phi\|_{m+1},$$

while the flux approximation is less accurate:

$$\text{(2.20)}\qquad \|\mathbf{u}-\mathbf{u}_h\|_0 = \|\nabla(\phi-\phi_h)\|_0 \le h^m\|\phi\|_{m+1}.$$

2.2.2. Stable mixed finite element spaces for the Kelvin principle. For discretizations of (2.10), i.e., the Kelvin principle, the inf-sup condition is much more onerous in the sense that defining a pair of stable finite element spaces for the scalar variable and the flux is not so straightforward a matter.

[11] Here, by *equivalent* we mean that they yield exactly the same approximate solutions.

The mixed finite element method associated with (2.10), i.e., the Kelvin principle, is given by the following: seek $(\phi_h, \mathbf{u}_h) \in \mathcal{W}_k^3(\Omega) \times \mathcal{W}_k^2(\Omega)$ such that

$$\text{(2.21)} \qquad \begin{cases} \displaystyle\int_\Omega \mathbf{u}_h \cdot \mathbf{v}_h \, d\Omega - \int_\Omega \phi_h \nabla \cdot \mathbf{v}_h \, d\Omega = 0 & \forall\, \mathbf{v}_h \in \mathcal{W}_k^2(\Omega), \\ \displaystyle\int_\Omega \psi_h \nabla \cdot \mathbf{u}_h \, d\Omega + \int_\Omega \gamma \psi_h \psi_h \, d\Omega = \int_\Omega f \psi_h \, d\Omega & \forall\, \psi_h \in \mathcal{W}_k^3(\Omega). \end{cases}$$

For the $\gamma = 0$ case, we refer to [11] for a proof that $(\mathcal{W}_k^3(\Omega), \mathcal{W}_k^2(\Omega))$ is a stable pair for the mixed finite element problem (2.21). Moreover, one can show [11, Proposition 1.2, p. 139] that for any sufficiently regular exact solution of (2.10) one has the error estimate

$$\text{(2.22)} \qquad \|\mathbf{u} - \mathbf{u}_h\|_0 \le Ch^r \|\mathbf{u}\|_r \qquad \begin{cases} \text{for } 1 < r \le k & \text{if } \mathcal{W}_k^2(\mathcal{K}) = \mathrm{RT}_{k-1}, \\ \text{for } 1 < r \le k+1 & \text{if } \mathcal{W}_k^2(\mathcal{K}) = \mathrm{BDM}_k, \end{cases}$$

while the error in the divergence is of the same order in both cases,

$$\text{(2.23)} \qquad \|\nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|_0 \le Ch^r \|\nabla \cdot \mathbf{u}\|_r \qquad \text{for } 1 < r \le k,$$

as is the error in the scalar variable:

$$\text{(2.24)} \qquad \|\phi - \phi_h\|_0 \le Ch^r (\|\phi\|_r + \|\mathbf{u}\|_r) \qquad \text{for } 1 < r \le k.$$

These results also hold for the $\gamma > 0$ case, since the mixed finite element problem (2.21) is identical to what one obtains for penalty methods for the $\gamma = 0$ case; see, e.g., [11, 18], for details.

We have thus seen that the duality between the Dirichlet and Kelvin principles propagates to their numerical approximations by mixed finite element methods that themselves have, in a sense, complementary computational properties. For example, for the Dirichlet principle, one directly approximates the scalar variable in the H^1-conforming finite element space $W_m^0(\Omega)$, and the flux is approximated in the finite element space $W_m^1(\Omega) = \nabla(W_m^0(\Omega))$. With respect to $L^2(\Omega)$ norms, the mixed approximation ϕ_h to ϕ satisfies the optimal bound (2.19), while the approximation $\mathbf{u}_h$ of the flux $\mathbf{u}$ is less accurate; see (2.20). For the Kelvin principle, the situation is reversed in the sense that now one directly approximates the flux in the $H_N(\Omega, \mathrm{div}\,)$-conforming finite element space $\mathcal{W}_k^2(\Omega)$ and the scalar variable in $\mathcal{W}_k^3(\Omega) = \nabla \cdot (\mathcal{W}_k^2(\Omega))$. The approximation $\mathbf{u}_h$ to $\mathbf{u}$ now satisfies the optimal bound (2.22), while the scalar approximation is less accurate when $\mathcal{W}_k^2(\Omega) = \mathrm{BDM}_k$.

We have also seen the differences in how easily one can satisfy the inf-sup condition for mixed methods based on the two principles. From (2.18), one sees that for the Dirichlet principle one can essentially avoid the inf-sup condition, or, if one insists on using the mixed formulation (2.17), one can easily construct a stable pair of spaces. This is closely related to the fact that the null space of the gradient consists of the constant function and is trivial to approximate. On the other hand, for the Kelvin principle, one has to construct a pair of finite element spaces such that the space for approximating the scalar variable is the divergence of the space for approximating the flux and the latter is a subspace of $H(\Omega, \mathrm{div})$. This is a much more difficult construction since the divergence operator has a decidedly nontrivial null space that is much harder to approximate than the (trivial) null space of the gradient. Compared to the finite element subspaces that can be used for approximations of the Dirichlet principle, for the Kelvin principle this leads to the need to define more complicated finite element subspaces for the flux such as the RT and BDM spaces or continuous piecewise linear subspaces based on the criss-cross grid.

It is important to note that if one uses C^0 finite element spaces for both the scalar variable and the flux, then (2.17) and (2.21) are identical discrete systems. It is well known that this leads to unstable approximations, so that one cannot use such pairs of finite element spaces in mixed methods derived from either the Dirichlet or Kelvin principles.

2.3. The grid decomposition property. We continue our study of mixed methods based on the Kelvin principle by showing that the spaces $\mathcal{W}_k^2(\Omega)$ satisfy the GDP.[12]

THEOREM 2.1. *For every $\mathbf{u}_h \in \mathcal{W}_k^2(\Omega)$, there exist $\mathbf{w}_h, \mathbf{z}_h$ in $\mathcal{W}_k^2(\Omega)$ such that*[13]

$$\mathbf{u}_h = \mathbf{w}_h + \mathbf{z}_h, \tag{2.25}$$

$$\nabla \cdot \mathbf{z}_h = 0, \tag{2.26}$$

$$\int_\Omega \mathbf{w}_h \cdot \mathbf{z}_h \, d\Omega = 0, \tag{2.27}$$

$$\|\mathbf{w}_h\|_0 \le C(\|\nabla \cdot \mathbf{u}_h\|_{-1} + h\|\nabla \cdot \mathbf{u}_h\|_0). \tag{2.28}$$

Proof. Given a $\mathbf{u}_h \in \mathcal{W}_k^2(\Omega)$, define $\mathbf{w}_h$ to be a solution of the following mixed problem: seek $(\phi_h, \mathbf{w}_h) \in \mathcal{W}_k^3(\Omega) \times \mathcal{W}_k^2(\Omega)$ such that

$$\begin{aligned} \int_\Omega \mathbf{w}_h \cdot \mathbf{v}_h \, d\Omega - \int_\Omega \phi_h \nabla \cdot \mathbf{v}_h \, d\Omega = 0 \quad \forall\, \mathbf{v}_h \in \mathcal{W}_k^2(\Omega), \\ \int_\Omega \psi_h \nabla \cdot \mathbf{w}_h \, d\Omega = \int_\Omega \psi_h \nabla \cdot \mathbf{u}_h \, d\Omega \quad \forall\, \psi_h \in \mathcal{W}_k^3(\Omega). \end{aligned} \tag{2.29}$$

The second component is then defined as the algebraic complement

$$\mathbf{z}_h = \mathbf{u}_h - \mathbf{w}_h \tag{2.30}$$

of $\mathbf{u}_h$. Therefore, the first GDP property (2.25) is trivially satisfied.

To prove (2.26), we use the second equation in (2.29) to conclude that

$$\int_\Omega \psi_h \nabla \cdot \mathbf{z}_h \, d\Omega = \int_\Omega \psi_h (\nabla \cdot \mathbf{u}_h - \nabla \cdot \mathbf{w}_h) \, d\Omega = 0 \quad \forall\, \psi_h \in \mathcal{W}_k^3(\Omega).$$

Assume now that $\nabla \cdot \mathbf{z}_h \neq 0$. From the definition of $\mathcal{W}_k^3(\Omega)$, it follows that the divergence operator is a surjective mapping $\mathcal{W}_k^2(\Omega) \mapsto \mathcal{W}_k^3(\Omega)$. Therefore, there exists

[12] An analogous "GDP" can be defined in the context of the Dirichlet principle; it requires that for every $\phi_h \in \mathcal{W}_k^0(\Omega)$ there exist $\lambda_h, \chi_h \in \mathcal{W}_k^0(\Omega)$ such that $\phi_h = \lambda_h + \chi_h$, $\nabla \chi_h = 0$, $\int_\Omega \lambda_h \chi_h d\Omega = 0$, and $\|\lambda_h\|_0 \le C(\|\nabla \phi_h\|_{-1} + h\|\nabla \phi_h\|_0)$. Of course, these conditions are trivially satisfied since $\nabla \chi_h = 0$ and $\chi_h \in \mathcal{W}_k^0(\Omega)$ imply that $\chi_h = 0$ and therefore $\lambda_h = \phi_h$. Again, the fact that the null space of the gradient operator with respect to $H_D^1(\Omega)$ is trivial plays a crucial role in the triviality of the GDP for the Dirichlet principle. On the other hand, for the Kelvin principle, the fact that the null space of the divergence operator with respect to $H_N(\Omega, \mathrm{div}\,)$ is decidedly not trivial also plays a crucial role in the GDP for that principle. All this, of course, is related to the observations made above about the inf-sup conditions for the two principles.

[13] In its original form (see [17]), the GDP was formulated without the term $h\|\nabla \cdot \mathbf{u}_h\|_0$ in (2.28). However, thanks to the multiplicative h factor, this term will not affect the L^2 error rates.

a nonzero element $\widehat{\psi}_h \in \mathcal{W}_k^3(\Omega)$ such that $\widehat{\psi}_h = \nabla \cdot \mathbf{z}_h$. Then,

$$0 = \int_\Omega \widehat{\psi}_h \nabla \cdot \mathbf{z}_h \, d\Omega = \int_\Omega \widehat{\psi}_h \widehat{\psi}_h \, d\Omega \neq 0,$$

a contradiction.

To show that $\mathbf{w}_h$ and $\mathbf{z}_h$ are orthogonal, we use the first equation in (2.29) with $\mathbf{v}_h = \mathbf{z}_h$:

$$\int_\Omega \mathbf{w}_h \cdot \mathbf{z}_h \, d\Omega = \int_\Omega \phi_h \nabla \cdot \mathbf{z}_h \, d\Omega = 0.$$

To prove the last GDP property (2.28), we will need the solution $(\phi, \mathbf{w})$ of the first-order problem

$$\begin{cases} \nabla \cdot \mathbf{w} = \nabla \cdot \mathbf{u}_h \quad \text{and} \quad \mathbf{w} + \nabla \phi = \mathbf{0} \quad \text{in } \Omega, \\ \phi = 0 \quad \text{on } \Gamma_D \quad \text{and} \quad \mathbf{w} \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N. \end{cases}$$

It will be also necessary to assume that this problem has full elliptic regularity, i.e., $\mathbf{w} \in \mathbf{H}^1(\Omega)$ and $\phi \in H^2(\Omega)$. Lastly, we recall the a priori bounds

$$\|\mathbf{w}\|_0 \leq \|\nabla \cdot \mathbf{u}_h\|_{-1} \quad \text{and} \quad \|\mathbf{w}\|_1 \leq \|\nabla \cdot \mathbf{u}_h\|_0.$$

Then, from (2.22)

$$\|\mathbf{w} - \mathbf{w}_h\|_0 \leq Ch\|\mathbf{w}\|_1.$$

Using this error estimate, the a priori bounds, and the triangle inequality yields that

$$\begin{aligned} \|\mathbf{w}_h\|_0 &\leq \|\mathbf{w}_h - \mathbf{w}\|_0 + \|\mathbf{w}\|_0 \\ &\leq Ch\|\mathbf{w}\|_1 + \|\nabla \cdot \mathbf{u}_h\|_{-1} \leq Ch\|\nabla \cdot \mathbf{u}_h\|_0 + \|\nabla \cdot \mathbf{u}_h\|_{-1}. \qquad \square \end{aligned}$$

It was shown in [17] that the GDP, i.e., (2.25)–(2.28), along with the relation $\mathcal{W}_k^3(\Omega) = \nabla \cdot (\mathcal{W}_k^2(\Omega))$, are necessary and sufficient for the stability of a mixed finite element method based on the Kelvin principle.

3. Least-squares finite element methods. A least-squares finite element method for the Poisson equation replaces the search for saddle-points of the Lagrangian functional, either $L_D(\phi, \mathbf{w}, \mathbf{u}, f)$ or $L_K(\lambda, \mathbf{u}, \phi, f)$, by a search for the unconstrained global minimizer of the quadratic functional

$$J(\phi, \mathbf{u}; f) = \frac{1}{2}\big(\|\nabla \cdot \mathbf{u} + \gamma\phi - f\|_0^2 + \|\nabla\phi + \mathbf{u}\|_0^2\big). \tag{3.1}$$

The least-squares variational principle

$$\min_{(\phi,\mathbf{u}) \in H_D^1(\Omega) \times H_N(\Omega,\mathrm{div})} J(\phi, \mathbf{u}; f) \tag{3.2}$$

then has a solution that minimizes the L^2 residuals of the first-order system (2.5). It is clear that this solution coincides with the solution of (2.5) or, equivalently, (1.1), and that it can be determined from the following first-order optimality system for (3.2): seek $(\phi, \mathbf{u}) \in H_D^1(\Omega) \times H_N(\Omega, \mathrm{div}\,)$ such that

$$Q((\phi, \mathbf{u}); (\psi, \mathbf{v})) = \mathcal{F}(\psi, \mathbf{v}) \quad \forall\, (\psi, \mathbf{v}) \in H_D^1(\Omega) \times H_N(\Omega, \mathrm{div}), \tag{3.3}$$

where

$$Q((\phi,\mathbf{u});(\psi,\mathbf{v})) = \int_\Omega (\nabla\cdot\mathbf{u}+\gamma\phi)(\nabla\cdot\mathbf{v}+\gamma\psi)\,d\Omega + \int_\Omega (\nabla\phi+\mathbf{u})\cdot(\nabla\psi+\mathbf{v})\;d\Omega \tag{3.4}$$

and

$$\mathcal{F}(\psi,\mathbf{v}) = \int_\Omega f(\nabla\cdot\mathbf{v}+\gamma\psi)\,d\Omega. \tag{3.5}$$

To define a least-squares finite element method, we restrict (3.2) to the conforming subspace $\mathcal{W}_m^0(\Omega)\times\mathcal{W}_k^2(\Omega)\subset H_D^1(\Omega)\times H_N(\Omega,\mathrm{div}\,)$. The least-squares finite element approximation is then obtained from the following discrete optimality system: seek $(\phi_h,\mathbf{u}_h)\in\mathcal{W}_m^0(\Omega)\times\mathcal{W}_k^2(\Omega)$ such that

$$Q((\phi_h,\mathbf{u}_h);(\psi_h,\mathbf{v}_h)) = \mathcal{F}(\psi_h,\mathbf{v}_h) \quad \forall\,(\psi_h,\mathbf{v}_h)\in\mathcal{W}_m^0(\Omega)\times\mathcal{W}_k^2(\Omega). \tag{3.6}$$

The next theorem states that

$$|||\psi,\mathbf{v}||| = (Q((\psi,\mathbf{v});(\psi,\mathbf{v})))^{1/2}$$

is an equivalent norm on $H_D^1(\Omega)\times H_N(\Omega,\mathrm{div}\,)$. We call it the *energy norm* corresponding to the least-squares principle.

THEOREM 3.1. *There exist positive constants C_1 and C_2 such that for any $(\psi,\mathbf{v})\in H_D^1(\Omega)\times H_N(\Omega,\mathrm{div})$,*

$$C_1\big(\|\psi\|_1^2+\|\mathbf{v}\|_{H(\Omega,\mathrm{div})}^2\big) \le |||\psi,\mathbf{v}|||^2 \le C_2\big(\|\psi\|_1^2+\|\mathbf{v}\|_{H(\Omega,\mathrm{div})}^2\big). \tag{3.7}$$

For a proof, see any of [12, 13, 14, 27]. Theorem 3.1 implies that both the continuous variational problem (3.3) and its finite element restriction (3.6) are uniquely solvable and that their solutions are bounded by the norm of the data.

Note for later use that (3.3) and (3.6) imply the standard finite element orthogonality relation

$$Q((\phi-\phi_h,\mathbf{u}-\mathbf{u}_h);(\psi_h,\mathbf{v}_h)) = 0 \quad \forall\,(\psi_h,\mathbf{v}_h)\in\mathcal{W}_m^0(\Omega)\times\mathcal{W}_k^2(\Omega). \tag{3.8}$$

3.1. Error estimates in $H^1(\Omega)\times H(\Omega,\mathrm{div})$.

In this section, we review the convergence properties of least-squares finite element methods for the Poisson equation with respect to the $H^1(\Omega)\times H(\Omega,\mathrm{div})$ norm. Most of the details are omitted, as the proofs follow by standard elliptic finite element arguments.

THEOREM 3.2. *Assume that the solution $(\phi,\mathbf{u})$ of (3.3) satisfies $(\phi,\mathbf{u})\in H_D^1(\Omega)\cap H^{m+1}(\Omega)\times H_N(\Omega,\mathrm{div}\,)\cap\mathbf{H}^{k+1}(\Omega)$ for some integers $k,m\ge 1$. Let $(\phi_h,\mathbf{u}_h)\in\mathcal{W}_m^0(\Omega)\times\mathcal{W}_k^2(\Omega)$ be the solution of the least-squares finite element problem (3.6). Then, there exists a constant $C>0$ such that*

$$\|\phi-\phi_h\|_1+\|\mathbf{u}-\mathbf{u}_h\|_{H(\Omega,\mathrm{div})} \le C\left(h^k\|\mathbf{u}\|_{k+1}+h^m\|\phi\|_{m+1}\right). \tag{3.9}$$

The error estimate (3.9) remains valid when $\mathbf{u}_h$ is approximated by the C^0 space $(P_k(\Omega))^n$.

Proof. Since $(\phi_h, \mathbf{u}_h)$ is a projection with respect to the energy norm $|||\cdot|||$,

$$|||\phi - \phi_h; \mathbf{u} - \mathbf{u}_h||| \leq |||\phi - \psi_h; \mathbf{u} - \mathbf{v}_h||| \quad \forall\, \psi_h \in \mathcal{W}_m^0(\Omega), \;\; \mathbf{v}_h \in \mathcal{W}_k^2(\Omega).$$

Then, (3.9) easily follows from the norm equivalence relation (3.7) and the approximation theoretic estimates (2.11)–(2.15). □

Theorem 3.2 shows that the errors in $\mathbf{u}_h$ and ϕ_h will be equilibrated whenever $k = m$. For example, if any of the pairs (RT_0, P_1), (BDM_1, P_1), or $((P_1)^n, P_1)$ are used in the least-squares finite element method, the a priori bound (3.9) specializes to

$$\|\phi - \phi_h\|_1 + \|\mathbf{u} - \mathbf{u}_h\|_{H(\Omega,\mathrm{div})} \leq Ch\left(\|\mathbf{u}\|_2 + \|\phi\|_2\right).$$

Therefore, the asymptotic accuracy of all three pairs in the norm of $H^1(\Omega)\times H(\Omega,\mathrm{div})$ is the same. For this reason, in the implementation of the least-squares finite element method, one usually chooses the C^0 pair $((P_1)^n, P_1)$ because it is the easiest to implement. Indeed, the ability to use equal-order interpolation has been often cited as a primary reason for choosing to use least-squares finite element methods. Nevertheless, the C^0 pair is not flawless because optimal L^2 norm errors for the flux approximation have proven impossible to obtain without augmenting (2.5) with an additional redundant curl constraint equation. Also, as we have already mentioned, numerical studies in [16] indicate that the L^2 convergence of the flux is indeed suboptimal with C^0 finite element spaces.

The curl constraint, first introduced in the least-squares finite element setting in [15] and subsequently utilized by many others (see, e.g., [12, 13, 14, 25]), makes the least-squares functional norm-equivalent on $H^1(\Omega) \times \mathbf{H}^1(\Omega)$. However, in some situations the curl equation may unduly restrict the range of the differential operator and should be avoided. In the next section, we will see that if the nodal approximation of the flux is replaced by an approximation in $\mathcal{W}_k^2(\Omega)$, it may be possible to recover optimal L^2 convergence rates without adding the curl constraint. As in [16], the key to this is the GDP.

3.2. Error estimates in L^2.

Throughout this section, we let $(\phi, \mathbf{u})$ and $(\phi_h, \mathbf{u}_h) \in \mathcal{W}_m^0(\Omega) \times \mathcal{W}_k^2(\Omega)$ denote the solutions of (3.3) and (3.6), respectively. We assume that the solution of the problem

$$-\Delta\psi + \gamma\psi = \eta \quad \text{in } \Omega, \qquad \psi = 0 \quad \text{on } \Gamma_D, \qquad \frac{\partial\psi}{\partial n} = 0 \quad \text{on } \Gamma_D \tag{3.10}$$

satisfies the regularity estimate

$$\|\psi\|_{s+2} \leq C\|\eta\|_s \quad \text{for } s = 0, 1 \text{ and } \forall\, \eta \in H^s(\Omega). \tag{3.11}$$

This additional regularity is necessary since our L^2 error estimates are based on a duality argument.

3.2.1. L^2 error estimates for the scalar variable.

Our first lemma bounds the negative norm of the error in the first equation in (2.5) in terms of the energy norm of the total error. Note that (3.11) of course implies that $\|\nabla\psi\|_{s+1} \leq C\|\eta\|_s$ for $s = 0, 1$.

LEMMA 3.3. *Let $(\phi_h, \mathbf{u}_h)$ be a least-squares finite element approximation of $(\phi, \mathbf{u})$. Then,*

$$\|\nabla \cdot (\mathbf{u} - \mathbf{u}_h) + \gamma(\phi - \phi_h)\|_{-1} \leq Ch|||\phi - \phi_h, \mathbf{u} - \mathbf{u}_h|||. \tag{3.12}$$

Proof. Let $\eta \in H_0^1(\Omega)$ be an arbitrary function, let ψ solve the boundary value problem (3.10), and let $\mathbf{v} = -\nabla\psi$. One then obtains

$$\begin{aligned}
&\int_\Omega (\nabla\cdot(\mathbf{u}-\mathbf{u}_h) + \gamma(\phi-\phi_h))\,\eta\, d\Omega \\
&\qquad = \int_\Omega (\nabla\cdot(\mathbf{u}-\mathbf{u}_h) + \gamma(\phi-\phi_h))(\nabla\cdot\mathbf{v} + \gamma\psi)\, d\Omega \\
&\qquad = Q(\phi-\phi_h, \mathbf{u}-\mathbf{u}_h; \psi, \mathbf{v}) = Q(\phi-\phi_h, \mathbf{u}-\mathbf{u}_h; \psi - \mathcal{I}_0\psi, \mathbf{v}-\mathcal{I}_2\mathbf{v}) \\
&\qquad = \int_\Omega (\nabla\cdot(\mathbf{u}-\mathbf{u}_h) + \gamma(\phi-\phi_h))(\nabla\cdot(\mathbf{v}-\mathcal{I}_2\mathbf{v}) + \gamma(\psi-\mathcal{I}_0\psi)\, d\Omega \\
&\qquad\quad + \int_\Omega (\nabla(\phi-\phi_h) + (\mathbf{u}-\mathbf{u}_h))\cdot(\nabla(\psi-\mathcal{I}_0\psi) + (\mathbf{v}-\mathcal{I}_2\mathbf{v}))\, d\Omega \\
&\qquad \le C((\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\|_0 + \|\phi-\phi_h\|_0)(\|\nabla\cdot(\mathbf{v}-\mathcal{I}_2\mathbf{v})\|_0 + \|\psi-\mathcal{I}_0\psi\|_0) \\
&\qquad\qquad + (\|\nabla(\phi-\phi_h)\|_0 + \|\mathbf{u}-\mathbf{u}_h\|_0)(\|\nabla(\psi-\mathcal{I}_0\psi)\|_0 + \|\mathbf{v}-\mathcal{I}_2\mathbf{v}\|_0)),
\end{aligned}$$

where we have successively used $\nabla\psi + \mathbf{v} = 0$, the error orthogonality (3.8), the definition of $Q(\cdot,\cdot;\cdot,\cdot)$, and the Cauchy–Schwarz inequality. Using the interpolation error estimates (2.11)–(2.15) and the regularity assumption (3.11), we have that

$$\left.\begin{aligned}
\|\nabla\cdot(\mathbf{v}-\mathcal{I}_2\mathbf{v})\|_0 \le Ch\|\mathbf{v}\|_2 \le Ch\|\eta\|_1 \\
\|\mathbf{v}-\mathcal{I}_2\mathbf{v}\|_0 \le Ch\|\mathbf{v}\|_2 \le Ch\|\eta\|_1 \\
\|\nabla(\psi-\mathcal{I}_0\psi)\|_0 \le Ch\|\psi\|_2 \le Ch\|\eta\|_1 \\
\|\psi-\mathcal{I}_0\psi\|_0 \le Ch^2\|\psi\|_2 \le Ch^2\|\eta\|_1
\end{aligned}\right\}\qquad \forall\,\eta \in H_0^1(\Omega).$$

Combining the last two sets of results, we easily obtain that, for all $\eta \in H_0^1(\Omega)$,

$$\int_\Omega (\nabla\cdot(\mathbf{u}-\mathbf{u}_h) + \gamma(\phi-\phi_h))\,\eta\, d\Omega \le Ch(\|\phi-\phi_h\|_1 + \|\mathbf{u}-\mathbf{u}_h\|_{H(\Omega,\mathrm{div})})\|\eta\|_1,$$

while the left inequality in (3.7) gives

$$\int_\Omega (\nabla\cdot(\mathbf{u}-\mathbf{u}_h) + \gamma(\phi-\phi_h))\,\eta\, d\Omega \le Ch|||\phi-\phi_h, \mathbf{u}-\mathbf{u}_h|||\ \|\eta\|_1 \quad \forall\,\eta \in H_0^1(\Omega).$$

The lemma follows by taking a supremum over $\eta \in H_0^1(\Omega)$. □

Next, we bound the L^2 error in ϕ_h by the energy norm.

LEMMA 3.4. *Let $(\phi_h, \mathbf{u}_h)$ be a least-squares finite element approximation of $(\phi, \mathbf{u})$. Then,*

$$\|\phi-\phi_h\|_0 \le Ch|||\phi-\phi_h, \mathbf{u}-\mathbf{u}_h|||. \tag{3.13}$$

Proof. Let ψ solve the boundary value problem

$$\begin{cases} -\triangle\psi + \gamma\psi = \phi - \phi_h & \text{in } \Omega, \\ \psi = 0 \text{ on } \Gamma_D \quad \text{and} \quad \partial\psi/\partial n = 0 \text{ on } \Gamma_N. \end{cases} \tag{3.14}$$

The regularity assumption (3.11) implies that

$$\|\psi\|_2 \le C\|\phi-\phi_h\|_0. \tag{3.15}$$

Using the definition of ψ, integration by parts, and the definition of the least-squares form (3.4) yields

$$\begin{aligned}\|\phi-\phi_h\|_0^2 &= \int_\Omega (\phi-\phi_h)(-\triangle\psi+\gamma\psi)\,d\Omega = \int_\Omega (\nabla(\phi-\phi_h)\cdot\nabla\psi+\gamma(\phi-\phi_h)\psi)\,d\Omega\\ &= \int_\Omega (\nabla(\phi-\phi_h)+(\mathbf{u}-\mathbf{u}_h))\cdot\nabla\psi\,d\Omega\\ &\quad -\int_\Omega (\mathbf{u}-\mathbf{u}_h)\cdot\nabla\psi\,d\Omega+\int_\Omega \gamma(\phi-\phi_h)\psi\,d\Omega\\ &= Q(\phi-\phi_h,\mathbf{u}-\mathbf{u}_h;\psi,0)+\int_\Omega (1-\gamma)(\nabla\cdot(\mathbf{u}-\mathbf{u}_h)+\gamma(\phi-\phi_h))\,\psi\,d\Omega.\end{aligned}$$

Using the error orthogonality (3.8), the definition (3.14) for ψ, and the regularity assumption (3.15) yields

$$\begin{aligned}&Q(\phi-\phi_h,\mathbf{u}-\mathbf{u}_h;\psi,0) = Q(\phi-\phi_h,\mathbf{u}-\mathbf{u}_h;\psi-\mathcal{I}_0\psi,0)\\ &\quad\le |||\phi-\phi_h,\mathbf{u}-\mathbf{u}_h|||\,\|\psi-\mathcal{I}_0\psi\|_1 \le Ch|||\phi-\phi_h,\mathbf{u}-\mathbf{u}_h|||\,\|\psi\|_2\\ &\quad\le Ch|||\phi-\phi_h,\mathbf{u}-\mathbf{u}_h|||\,\|\phi-\phi_h\|_0.\end{aligned}$$

In addition, the definition (3.14) for ψ, (3.15), and (3.12) imply that

$$\begin{aligned}&\int_\Omega (1-\gamma)(\nabla\cdot(\mathbf{u}-\mathbf{u}_h)+\gamma(\phi-\phi_h))\,\psi\,d\Omega \le \|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)+\gamma(\phi-\phi_h)\|_{-1}\|\psi\|_1\\ &\quad\le C\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)+\gamma(\phi-\phi_h)\|_{-1}\|\phi-\phi_h\|_0\\ &\quad\le Ch|||\phi-\phi_h,\mathbf{u}-\mathbf{u}_h|||\,\|\phi-\phi_h\|_0.\end{aligned}$$

The lemma easily follows by combining the last three results. □

COROLLARY 3.5. *Assume that the regularity assumption* (3.11) *is satisfied, and assume that the solution* $(\phi,\mathbf{u})$ *of* (3.3) *satisfies* $(\phi,\mathbf{u})\in H^1_D(\Omega)\cap H^{m+1}(\Omega)\times H_N(\Omega,\mathrm{div}\,)\cap\mathbf{H}^{k+1}(\Omega)$ *for some integers* $k,m\ge 1$. *Let* $(\phi_h,\mathbf{u}_h)\in\mathcal{W}^0_m(\Omega)\times\mathcal{W}^2_k(\Omega)$ *be the solution of the least-squares finite element problem* (3.6). *Then, there exists a constant* $C>0$ *such that*

$$\|\phi-\phi_h\|_0\le C(h^{k+1}\|\mathbf{u}\|_{k+1}+h^{m+1}\|\phi\|_{m+1}). \tag{3.16}$$

Proof. The corollary follows simply by a direct application of (3.7) and (3.9) to (3.13). □

The optimal L^2 error bound (3.16) for the scalar variable does not depend on whether or not the finite element space for the flux satisfies (2.25)–(2.28), i.e., the GDP. Thus, it remains valid even when equal-order C^0 finite element functions are used for the flux approximations, a result first shown in [24]. On the other hand, we will see that the GDP is needed if one wants to improve the L^2 accuracy of the flux.

3.2.2. L^2 error estimate for the flux.

Ultimately, the final L^2 error estimates for approximations to the flux depend on whether $\mathcal{W}^2_k(\Omega)$ represents the RT_{k-1} or the BDM_k family. To this end, we need the following result.

LEMMA 3.6. *Let* $(\phi_h,\mathbf{u}_h)\in\mathcal{W}^0_m(\Omega)\times\mathcal{W}^2_k(\Omega)$ *be the least-squares finite element approximation defined by* (3.6). *Then,*[14]

$$\|\mathbf{u}-\mathbf{u}_h\|_0\le C\big(h|||\phi-\phi_h,\mathbf{u}-\mathbf{u}_h|||+h\|\nabla\cdot(\mathbf{u}-\mathbf{v}_h)\|_0+\|\mathbf{u}-\mathbf{v}_h\|_0\big) \tag{3.17}$$

[14] It is clear from the proof of this lemma that it holds for any flux approximation that satisfies the GDP.

for any $\mathbf{v}_h \in \mathcal{W}_k^2(\Omega)$.

Proof. Let $\mathbf{v}_h$ be an arbitrary element of $\mathcal{W}_k^2(\Omega)$. From Theorem 2.1, we know that there exist $\mathbf{z}_h$ and $\mathbf{w}_h$, also in $\mathcal{W}_k^2(\Omega)$, such that

$$\mathbf{u}_h - \mathbf{v}_h = \mathbf{w}_h + \mathbf{z}_h$$

and the properties (2.26)–(2.28) hold. We recall for later use that

$$\|\nabla \cdot \mathbf{v}\|_{-1} \leq \|\mathbf{v}\|_0 \quad \forall\, \mathbf{v} \in H_N(\Omega, \mathrm{div}). \tag{3.18}$$

We now bound the two GDP components of $\mathbf{u}_h - \mathbf{v}_h$ in L^2. To estimate $\|\mathbf{w}_h\|_0$, we successively use (2.28), (3.18), (3.12), (3.7), and (3.13) to obtain

$$\begin{aligned}
\|\mathbf{w}_h\|_0 &\leq C(\|\nabla \cdot (\mathbf{u}_h - \mathbf{v}_h)\|_{-1} + h\|\nabla \cdot (\mathbf{u}_h - \mathbf{v}_h)\|_0) \\
&\leq C(\|\nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|_{-1} + h\|\nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|_0 \\
&\qquad + \|\nabla \cdot (\mathbf{u} - \mathbf{v}_h)\|_{-1} + h\|\nabla \cdot (\mathbf{u} - \mathbf{v}_h)\|_0) \\
&\leq C(\|\nabla \cdot (\mathbf{u} - \mathbf{u}_h) + \gamma(\phi - \phi_h)\|_{-1} + h\|\nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|_0 \\
&\qquad + \|\gamma(\phi - \phi_h)\|_{-1} + \|\mathbf{u} - \mathbf{v}_h\|_0 + h\|\nabla \cdot (\mathbf{u} - \mathbf{v}_h)\|_0) \\
&\leq C(h|||\phi - \phi_h, \mathbf{u} - \mathbf{u}_h||| + \|\phi - \phi_h\|_0 + ||\mathbf{u} - \mathbf{v}_h\|_0 + h\|\nabla \cdot (\mathbf{u} - \mathbf{v}_h)\|_0) \\
&\leq C(h|||\phi - \phi_h, \mathbf{u} - \mathbf{u}_h||| + ||\mathbf{u} - \mathbf{v}_h\|_0 + h\|\nabla \cdot (\mathbf{u} - \mathbf{v}_h)\|_0).
\end{aligned}$$

To estimate $\|\mathbf{z}_h\|_0$, we use the error orthogonality (3.8) with $\psi_h = 0$ and $\mathbf{v}_h = \mathbf{z}_h$. Since from (2.26) we have that $\nabla \cdot \mathbf{z}_h = 0$, this identity reduces to

$$\int_\Omega (\nabla(\phi - \phi_h) + (\mathbf{u} - \mathbf{u}_h)) \cdot \mathbf{z}_h \, d\Omega = 0,$$

from which integrating by parts and again using $\nabla \cdot \mathbf{z}_h = 0$ yields

$$\int_\Omega (\mathbf{u} - \mathbf{u}_h) \cdot \mathbf{z}_h \, d\Omega = 0.$$

Using this result and the orthogonality of $\mathbf{z}_h$ and $\mathbf{w}_h$ (see (2.27)), we obtain

$$\begin{aligned}
\|\mathbf{z}_h\|_0^2 = \int_\Omega \mathbf{z}_h \cdot \mathbf{z}_h \, d\Omega &= \int_\Omega (\mathbf{z}_h + \mathbf{w}_h) \cdot \mathbf{z}_h \, d\Omega \\
= \int_\Omega (\mathbf{u}_h - \mathbf{v}_h) \cdot \mathbf{z}_h \, d\Omega &= \int_\Omega (\mathbf{u} - \mathbf{v}_h) \cdot \mathbf{z}_h \, d\Omega,
\end{aligned}$$

so that

$$\|\mathbf{z}_h\|_0 \leq \|\mathbf{u} - \mathbf{v}_h\|_0.$$

To complete the proof, we note that

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq \|\mathbf{u} - \mathbf{v}_h\|_0 + \|\mathbf{u}_h - \mathbf{v}_h\|_0 \leq \|\mathbf{u} - \mathbf{v}_h\|_0 + \|\mathbf{w}_h\|_0 + \|\mathbf{z}_h\|_0$$

and then use the bounds on $\mathbf{z}_h$ and $\mathbf{w}_h$. ▯

Let us now inspect (3.17). The first term on the right-hand side is exactly the same one as in the L^2 bound (3.13) for the scalar variable. Let us further assume that the approximation orders of the spaces used for the scalar variable and the flux are equilibrated, i.e., $\phi_h \in \mathcal{W}_r^0(\Omega)$ and $\mathbf{u}_h \in \mathcal{W}_r^2(\Omega)$ for some $r \geq 1$. Then,

$$|||\phi - \phi_h, \mathbf{u} - \mathbf{u}_h||| \leq Ch^r(\|\phi\|_{r+1} + \|\mathbf{u}\|_{r+1}).$$

The additional factor h multiplying this term in (3.17) will increase the order of that term to $r+1$, just as in (3.16). However, (3.17) contains the two additional terms $h\|\nabla \cdot (\mathbf{u} - \mathbf{v}_h)\|_0$ and $\|\mathbf{u} - \mathbf{v}_h\|_0$. Recall that $\mathcal{W}_r^2(\Omega)$ represents RT and BDM spaces that are equilibrated with respect to the divergence error. Therefore, after setting $\mathbf{v}_h = \mathcal{I}_2\mathbf{u}$, from (2.15) it follows that

$$\|\nabla \cdot (\mathbf{u} - \mathbf{v}_h)\|_0 \leq Ch^r\|\mathbf{u}\|_{r+1}.$$

After multiplication by h, the order of this term also increases to $r+1$. However, the order of the last term will depend on whether $\mathcal{W}_r^2(\Omega)$ represents a BDM or RT space. Indeed, from (2.16),

$$\|\mathbf{u} - \mathbf{v}_h\|_0 \leq C\begin{cases} h^r\|\mathbf{u}\|_r & \text{if } \mathcal{W}_r^2(\Omega) = \mathrm{RT}_{r-1}, \\ h^{r+1}\|\mathbf{u}\|_{r+1} & \text{if } \mathcal{W}_r^2(\Omega) = \mathrm{BDM}_r. \end{cases}$$

The next corollary summarizes these observations.

Corollary 3.7. *Assume that the regularity assumption* (3.11) *is satisfied, and assume that the solution* $(\phi, \mathbf{u})$ *of* (3.3) *satisfies* $(\phi, \mathbf{u}) \in H_D^1(\Omega) \cap H^{r+1}(\Omega) \times H_N(\Omega, \mathrm{div}\,) \cap \mathbf{H}^{r+1}(\Omega)$ *for some integer* $r \geq 1$. *Let* $(\phi_h, \mathbf{u}_h) \in \mathcal{W}_r^0(\Omega) \times \mathcal{W}_r^2(\Omega)$ *be the solution of the least-squares finite element problem* (3.6). *Then, there exists a constant* $C > 0$ *such that*

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq C\begin{cases} h^r(\|\mathbf{u}\|_{r+1} + \|\phi\|_{r+1}) & \textit{if } \mathcal{W}_r^2(\Omega) = \mathrm{RT}_{r-1}, \\ h^{r+1}(\|\mathbf{u}\|_{r+1} + \|\phi\|_{r+1}) & \textit{if } \mathcal{W}_r^2(\Omega) = \mathrm{BDM}_r. \end{cases} \tag{3.19}$$

Consider, for example, the lowest-order case for which $r = 1$, $\mathcal{W}_1^0(\Omega) = P_1$, and $\mathcal{W}_1^2(\Omega)$ is either RT_0 or BDM_1. If the least-squares finite element method is implemented with RT_0 elements, (3.19) specializes to

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq h(\|\mathbf{u}\|_2 + \|\phi\|_2).$$

If instead we use BDM_1 elements, we then obtain the improved error bound

$$\|\mathbf{u} - \mathbf{u}_h\|_0 \leq h^2(\|\mathbf{u}\|_2 + \|\phi\|_2).$$

It is worth repeating that the reason for this difference in the L^2 errors is the structure of the RT spaces. Since $\mathrm{RT}_0 = (P_0)^n + \mathbf{x}P_0$, the approximation properties of RT_0 in L^2 are the same as those of P_0. However, it is easy to see that, thanks to the extra term $\mathbf{x}P_0$, $\nabla \cdot (\mathrm{RT}_0) = P_0$; i.e., the divergence of $\mathbf{u}$ is approximated to the same order as the field itself. For numerical examples with least-squares methods that illustrate this feature of RT spaces, we refer to [2].

4. Least-squares finite element methods and duality. We already saw that a least-squares finite element method implemented using equal-order C^0 finite element spaces approximates the scalar variable with the same accuracy as a Galerkin (or, equivalently, a mixed) method for the Dirichlet principle. However, the approximation properties of the Kelvin principle are only partially inherited in the sense that the accuracy in the approximation to the divergence of the flux is recovered, but the accuracy in the flux approximation itself may be of one order less. This should not be too much of a surprise because C^0 elements provide stable discretization only for the Dirichlet principle (with the exception of the criss-cross grid; see [16]). While least-squares minimization is stable enough to allow for the approximation of scalar variables and the flux by equal-order C^0 finite element spaces, it cannot completely recover from the fact that such spaces are unstable for the Kelvin principle.

The key observation from section 3.2 is that a least-squares finite element method can inherit the computational properties of *both the Dirichlet and the Kelvin* principles, provided the scalar variable and the flux are approximated by finite element spaces that are stable with respect to these two principles. Then, as our analysis showed, least-squares finite element solutions recover the accuracy of the Dirichlet principle for the scalar variable and the accuracy of the Kelvin principle for the flux.

In a way, we see that, implemented in this particular manner, the least-squares finite element method represents a balanced mixture of the two principles. Below, we provide an explanation of this observation using the apparatus of differential form calculus, albeit in a simplified form and without an explicit reference to differential forms on manifolds. For consistency, in what follows, $H(\Omega, \mathrm{grad})$, $H(\Omega, \mathbf{curl})$, and $H(\Omega, \mathrm{div})$ denote spaces of square integrable functions whose gradients, curls, and divergences, respectively, are also square integrable.[15]

The De Rham differential complex

$$\mathbb{R} \hookrightarrow H(\Omega, \mathrm{grad}) \overset{\nabla}{\longmapsto} H(\Omega, \mathbf{curl}) \overset{\nabla\times}{\longmapsto} H(\Omega, \mathrm{div}) \overset{\nabla\cdot}{\longmapsto} L^2(\Omega) \longmapsto 0 \tag{4.1}$$

is an exact sequence of spaces in the sense that each operator maps the space on its left to the kernel of the next operator in the sequence, and the last mapping is a surjection. We will now start to use the identifications

$$\mathcal{W}^0(\Omega) = H(\Omega, \mathrm{grad}),\ \mathcal{W}^1(\Omega) = H(\Omega, \mathbf{curl}),\ \mathcal{W}^2(\Omega) = H(\Omega, \mathrm{div}),\ \mathcal{W}^3(\Omega) = L^2(\Omega)$$

to indicate that these function spaces are comprised of *proxies* for differential forms of orders 0, 1, 2, and 3, respectively.[16] Exact sequences of finite element spaces provide piecewise polynomial approximations of the proxies. Commonly used terminologies for the finite element subspaces of $\mathcal{W}^0$, $\mathcal{W}^1$, $\mathcal{W}^2$, and $\mathcal{W}^3$ are nodal, edge, face, and volume (or discontinuous) elements, respectively.

Differential forms have always played a fundamental role in classical mechanics and numerical methods for Hamiltonian systems; see, e.g., [3, 6]. Their place as an abstraction tool for discretization of elliptic boundary value problems was perhaps first recognized in [21], while [8, 9] further affirmed their importance in computational electromagnetism.

Subsequently, the idea that a stable partial differential equation discretization can be developed using a discrete equivalent of the De Rham complex has been exploited by many researchers in finite element, finite volume, and finite difference methods

[15] Here we treat the case of $n = 3$; similar developments can be carried out for the two-dimensional case.

[16] This should explain the seemingly peculiar choice of notation introduced earlier in the paper.

[1, 5, 19, 20, 22, 23, 26]; see [4] for a more extensive bibliography. In particular, for second-order elliptic problems, a key tool for encoding their structure is provided by the *factorization* or *Tonti diagrams*; see [19, 20]. Essentially, these diagrams break the problem into topological relations between different spaces in a De Rham complex connected by metric relations expressed by the Hodge $*$-operator, a linear map $\mathcal{W}^k(\Omega) \mapsto \mathcal{W}^{n-k}(\Omega)$. The factorization diagram for the Poisson problem (see, e.g., [20, 26]) is

$$\begin{tikzcd}
\mathcal{W}^0(\Omega) & \phi \arrow[r, "\nabla"] \arrow[d, leftrightarrow, "\xi = *\phi"'] & -\mathbf{u} \arrow[d, leftrightarrow, "\mathbf{q} = *\mathbf{u}"] & \mathcal{W}^1(\Omega) \\
\mathcal{W}^3(\Omega) & \boxed{\mathrm{f}} - \gamma\xi & \mathbf{q} \arrow[l, "\nabla\cdot"'] & \mathcal{W}^2(\Omega)
\end{tikzcd} \tag{4.2}$$

We will refer to the relation and the variables on the top of the diagram as the *primal* variables and equilibrium equation. The dual variables and their "equilibrium" equation are represented by the bottom part of the diagram. The dual and primal variables serve as proxies for 0, 1 and 2, 3 differential forms, respectively.

The horizontal links in (4.2) correspond to the differential equations

$$\nabla\phi = -\mathbf{u} \quad \text{and} \quad \nabla\cdot\mathbf{q} = -\gamma\xi + f,$$

while the vertical links provide the "constitutive" relations

$$\xi = *\phi \quad \text{and} \quad \mathbf{q} = *\mathbf{u}.$$

The importance of structures such as (4.2) stems from the fact that they encode fundamental relationships between spaces and operators that are required for the stability of discretizations; see, e.g., [1, 4, 9, 20, 22].

Let us now show that the Dirichlet and Kelvin principles are obtained from (4.2) by the approximation of the Hodge operator by an identity operator and subsequent elimination of the dual or the primal variables, respectively.

If the dual variables are substituted by the primal ones according to

$$\xi = \phi \quad \text{and} \quad \mathbf{q} = \mathbf{u},$$

then the dual equation in (4.2) must be modified to account for the fact that $\mathbf{u}$ is a proxy of a 1-form, rather than of a 2-form. As such, $\mathbf{u}$ is in the domain of the curl operator but not in the domain of the divergence operator. Thus, in the dual equilibrium equation, we replace the divergence operator by a weak divergence operator defined through the following variational statement:

$$\widetilde{\nabla}\cdot\, : \mathcal{W}^1(\Omega) \mapsto \mathcal{W}^0(\Omega), \quad \widetilde{\nabla}\cdot\mathbf{u} = \phi,$$

if and only if

$$\int_\Omega \phi\psi\, d\Omega = -\int_\Omega \mathbf{u}\cdot\nabla\psi\, d\Omega \quad \forall\, \psi \in \mathcal{W}^0(\Omega).$$

This changes the original factorization diagram to one in terms of only the primal variables:

$$\begin{tikzcd}
\mathcal{W}^0(\Omega) & \phi \arrow[r, "\nabla"] \arrow[d, "\eta = \phi"'] & -\mathbf{u} \arrow[d, "\mathbf{v} = \mathbf{u}"] & \mathcal{W}^1(\Omega) \\
\mathcal{W}^0(\Omega) & \boxed{\mathrm{f}} - \gamma\phi & \mathbf{u} \arrow[l, "\widetilde{\nabla}\cdot"'] & \mathcal{W}^1(\Omega)
\end{tikzcd} \tag{4.3}$$

The partial differential equation system represented by this diagram is[17]

$$
\begin{aligned}
\nabla\phi + \mathbf{u} = 0 \qquad & \text{in } \mathcal{W}^1(\Omega),\\
-\int_\Omega \mathbf{u}\cdot\nabla\psi - \gamma\phi\psi\, d\Omega = \int_\Omega f\psi\, d\Omega \quad & \forall\,\psi \in \mathcal{W}^0(\Omega).
\end{aligned}
$$

One recognizes (4.3) as the optimality system (2.4) for the Dirichlet principle. The diagram (4.3) can be viewed as a representation of this principle.

If instead the primal variables are eliminated according to

$$\phi = \xi \quad \text{and} \quad \mathbf{u} = \mathbf{q},$$

then the primal equilibrium equation in (4.2) must be modified to account for the fact that ξ is a proxy of a 3-form, rather than for a 0-form. As such, ξ is not in the domain of the gradient operator, which therefore must be replaced by a weak one:

$$\widetilde{\nabla} : \mathcal{W}^3(\Omega), \mapsto \mathcal{W}^2(\Omega), \quad \widetilde{\nabla}\xi = \mathbf{q},$$

if and only if

$$\int_\Omega \mathbf{q}\cdot\mathbf{v}\, d\Omega = -\int_\Omega \xi\nabla\cdot\mathbf{v}\, d\Omega \quad \forall\,\mathbf{v} \in \mathcal{W}^2(\Omega).$$

The factorization diagram in terms of the dual variables is then given by

$$
\begin{tikzcd}
\mathcal{W}^3(\Omega) & & \xi \arrow[r, "\widetilde{\nabla}"] & -\mathbf{q} & & \mathcal{W}^2(\Omega) \\
(4.4) & \phi = \xi & \arrow[u] & \arrow[u] & \mathbf{u} = \mathbf{q} & \\
\mathcal{W}^3(\Omega) & & \boxed{\mathrm{f}} - \gamma\xi & \mathbf{q} \arrow[l, "\nabla\cdot"'] & & \mathcal{W}^2(\Omega)
\end{tikzcd}
$$

The problem represented by this diagram is

$$
\begin{aligned}
\int_\Omega \mathbf{q}\cdot\mathbf{v}\, d\Omega - \int_\Omega \xi\nabla\cdot\mathbf{v}\, d\Omega = 0 \quad & \forall\,\mathbf{v} \in \mathcal{W}^2(\Omega),\\
\nabla\cdot\mathbf{v} + \gamma\xi = f \qquad & \text{in } \mathcal{W}^3(\Omega).
\end{aligned}
$$

Now the second equation is an exact relation, and we see that, by elimination of the primal variables, we recover the optimality system (2.10) for the Kelvin principle.

It is now clear that each of the classical variational principles for the system (1.1) can be derived from (4.2) by elimination of one of the sets of variables (primal or dual) and relaxation of the complementary equilibrium equation. Elimination of variables, on the other hand, can be interpreted as approximation of the Hodge $*$-operator by an identity. This, of course, immediately leads to the following question: what kinds of variational principles can be obtained by using other ways of approximating the Hodge operator? Here, we will focus on one particular method wherein this operator is replaced by an L^2 projection. Not surprisingly, we will see that this approximation leads eventually to a least-squares principle for the first-order formulation of (1.1), but one that is necessarily implemented with spaces for the scalar inherited from the

[17] The first equation can also be stated in variational form; see (2.4). However, we write it in algebraic form to stress the fact that it represents an exact relationship.

Dirichlet principle, and, for the flux, from the Kelvin principle. Thus, in a sense, the least-squares method, when implemented in this manner, is indeed a mixture of the two classical principles that combines their best properties.

The idea is to keep both the primal and dual sets of variables, but to replace the Hodge operator by an optimization problem that penalizes the discrepancy between these sets. Then, the primal and dual equations become linear constraints that must be satisfied by the minimizers of this functional. Therefore, we are led to the following constrained optimization problem: seek $(\phi, \mathbf{u}, \xi, \mathbf{q})$ in $W^0(\Omega) \times \mathcal{W}^1(\Omega) \times \mathcal{W}^2(\Omega) \times \mathcal{W}^3(\Omega)$ such that

$$\mathcal{J}(\phi, \mathbf{u}, \xi, \mathbf{q}) = \frac{1}{2}(\|\xi - \phi\|_0^2 + \|\mathbf{q} - \mathbf{u}\|_0^2) \quad \mapsto \quad \min \tag{4.5}$$

subject to

$$\nabla \phi + \mathbf{u} = 0 \quad \text{and} \quad \nabla \cdot \mathbf{q} + \gamma \xi = f. \tag{4.6}$$

In this problem, the Hodge operator is approximated by the L^2 projections

$$(*_0) : \mathcal{W}^0(\Omega) \mapsto \mathcal{W}^3(\Omega) \quad \text{and} \quad (*_1) : \mathcal{W}^1(\Omega) \mapsto \mathcal{W}^2(\Omega)$$

defined implicitly via the optimization process.

It is possible to solve (4.5)–(4.6) by using Lagrange multipliers to enforce the constraints. However, a better strategy (that also reduces the number of variables) is to note that the constraint equations can be satisfied exactly if the spaces chosen for ϕ_h, $\mathbf{u}_h$, ξ_h, and $\mathbf{q}_h$ are from a discrete exact sequence. It is also important to note that primal and dual variables can be approximated by discrete exact sequences that are not necessarily defined on the same mesh. Thus, assume that for the primal side we have chosen $\mathcal{W}^0_m(\Omega)$ and $\mathcal{W}^1_m(\Omega)$ to approximate ϕ and $\mathbf{u}$, respectively, while for the dual side we work with the spaces $\mathcal{W}^2_k(\Omega)$ and $\mathcal{W}^3_k(\Omega)$ to approximate $\mathbf{q}$ and ξ, respectively. In this manner, each set of variables is represented in the discrete problem by an internal approximation, and we can use the equilibrium equations (rather than the constitutive relations) to eliminate ξ_h and $\mathbf{u}_h$. This leads to the following discrete minimization problem in terms of ϕ_h and $\mathbf{q}_h$ only:

$$\min_{\mathcal{W}^0_m(\Omega) \times \mathcal{W}^2_k(\Omega)} \frac{1}{2}(\|\nabla \cdot \mathbf{q}_h + \gamma \phi_h - f\|_0^2 + \|\mathbf{q}_h + \nabla \phi_h\|_0^2). \tag{4.7}$$

While this problem appears identical to a least-squares formulation derived directly from (2.5), the manner in which it was obtained retains the information about the origins of the different variables. In particular, we see that in (4.7), the scalar variable is inherited from the primal Dirichlet principle, while the flux is inherited from the dual Kelvin principle. As was shown in section 3.2, when this is taken into account in the choice of approximating finite element spaces, the computational properties of both principles are recovered by (4.7). This is perhaps the most important point of our discussion. Another important distinction between (4.7) and a nodal-based implementation of a least-squares principle is that (4.7) leads to a conservative approximation in the following sense. Once ϕ_h and $\mathbf{q}_h$ are found, we can recover the eliminated dual and primal variables so as to obtain four fields ϕ_h, $\mathbf{u}_h$, ξ_h, and $\mathbf{q}_h$ that exactly satisfy the relations

$$\nabla \phi_h + \mathbf{u}_h = 0 \quad \text{and} \quad \nabla \cdot \mathbf{q}_h + \gamma \xi_h = \Pi_3 f.$$

The operator Π_3 that appears above is the L^2 projection into the subspace $\mathcal{W}^3_k$ of $L^2(\Omega)$, while the discrete Hodge operators can be identified with L^2 projections from nodal to discontinuous elements and from edge to face elements, respectively.

5. Conclusions. We have demonstrated that least-squares finite element methods for the first-order Poisson equation can combine the best properties of the classical Dirichlet and Kelvin principles if their implementation uses spaces consistent with the origins of the scalar variable and the flux. In particular, we have shown that a least-squares formulation can be viewed as resulting from a particular choice in the approximation of the Hodge operator. From this point of view, the scalar variable is inherited from the Dirichlet principle and requires approximation by nodal elements. The flux is inherited from the Kelvin principle and must be approximated by $H(\Omega, \text{div})$ conforming families to enable recovery of optimal L^2 rates *without* the addition of curl constraints.

When implemented in this manner, the least-squares finite element method can be deemed superior to both the classical Galerkin and mixed methods because, on the one hand, it provides optimal approximation of all fields with the possibility of recovering an approximation that is conservative in the sense explained earlier, while, on the other hand, it leads to symmetric and positive definite algebraic systems of equations.

Acknowledgment. The authors wish to acknowledge their debt to George Fix, from whom they learned much about mixed and least-squares finite element methods and about science and mathematics in general. Without his guidance and contributions, much of this paper might not have been possible.

REFERENCES

[1] D. Arnold, *Differential complexes and numerical stability,* in Proceedings of the International Congress of Mathematicians, Beijing 2002, Li Tatsien, ed., World Scientific, River Edge, NJ, 2002, Vol. 1.

[2] D. N. Arnold, D. Boffi, and R. S. Falk, *Quadrilateral H*(div) *finite elements*, SIAM J. Numer. Anal., 42 (2005), pp. 2429–2451.

[3] V. Arnold, *Mathematical Methods of Classical Mechanics*, Springer, Berlin, 1989.

[4] P. Bochev, *A discourse on variational and geometric aspects of stability of discretizations,* in Proceedings of the 33th Computational Fluid Dynamics Lecture Series, H. Deconinck, ed., Von Karman Institute of Fluid Dynamics, Lecture Series 2003–05, ISSN0377-8312.

[5] P. B. Bochev and A. C. Robinson, *Matching algorithms with physics: Exact sequences of finite element spaces*, in Collected Lectures on the Preservation of Stability Under Discretization, D. Estep and S. Tavener, eds., SIAM, Philadelphia, 2002, pp. 145–165.

[6] P. Bochev and C. Scovel, *On quadratic invariants and simplectic structure,* BIT, 34 (1994), pp. 337–345.

[7] D. Boffi, F. Brezzi, and L. Gastaldi, *On the problem of spurious eigenvalues in the approximation of linear elliptic problems in mixed form*, Math. Comp., 69 (2000), pp. 121–140.

[8] A. Bossavit, *A rationale for "edge-elements" in 3D fields computations*, IEEE Trans. Magnetics, 24 (1988), pp. 74–79.

[9] A. Bossavit, *Whitney forms: A class of finite elements for three dimensional computations in electromagnetism*, IEEE Proceedings, 135 (1988), pp. 493–500.

[10] J. Bramble, R. Lazarov, and J. Pasciak, *A least squares approach based on a discrete minus one inner product for first order systems,* Math. Comp., 66 (1997), pp. 935–955.

[11] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, Berlin, 1991.

[12] Z. Cai, R. Lazarov, T. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[13] G. Carey and A. Pehlivanov, *Error estimates for least-squares mixed finite elements,* Math. Model Numer. Anal., 28 (1994), pp. 499–516

[14] C. L. Chang, *Finite element approximation for grad-div type systems in the plane,* SIAM J. Numer. Anal., 29 (1992), pp. 452–461.

[15] C. Chang and M. Gunzburger, *A finite element method for first order elliptic systems in three dimensions,* Appl. Math. Comp., 23 (1987), pp. 171–184.

[16] G. Fix, M. Gunzburger, and R. Nicolaides, *On finite element methods of the least-squares type,* Comput. Math. Appl., 5 (1979), pp. 87–98.

[17] G. Fix, M. Gunzburger, and R. Nicolaides, *On mixed finite element methods for first-order elliptic systems,* Numer. Math., 37 (1981), pp. 29–48.

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

[19] R. Hiptmair, *Canonical construction of finite element spaces,* Math. Comp., 68 (1999), pp. 1325–1346.

[20] R. Hiptmair, *Discrete Hodge operators,* Numer. Math., 90 (2001), pp. 265–289.

[21] M. Hyman and C. Scovel, *Deriving Mimetic Difference Approximations to Differential Operators Using Algebraic Topology,* unpublished internal report, Los Alamos National Laboratory, Los Alamos, NM, 1988.

[22] J. Hyman and M. Shashkov, *Natural discretizations for the divergence, gradient, and curl on logically rectangular grids,* Int. J. Comput Math. Appl., 33 (1997), pp. 88–104.

[23] J. Hyman and M. Shashkov, *Adjoint operators for the natural discretizations of the divergence, gradient, and curl on logically rectangular grids,* Appl. Numer. Math., 25 (1997), pp. 413–442.

[24] D. Jesperson, *A least-squares decomposition method for solving elliptic equations,* Math. Comp., 31 (1977), pp. 873–880.

[25] B. Jiang and L. Povinelli, *Optimal least-squares finite element methods for elliptic problems,* Comp. Methods Appl. Mech. Engrg., 102 (1993), pp. 199–212.

[26] C. Mattiussi, *An analysis of finite volume, finite element and finite difference methods using some concepts from algebraic topology,* J. Comput. Phys., 133 (1997), pp. 289–309.

[27] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems,* SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

SIAM J. Numer. Anal.
Vol. 43, No. 1, pp. 363–385

RELIABLE AND EFFICIENT APPROXIMATION OF POLYCONVEX ENVELOPES*

SÖREN BARTELS†

Abstract. An iterative algorithm that approximates the polyconvex envelope f^{pc} of a given function $f : \mathbb{R}^{n\times m} \to \mathbb{R}$, i.e., the largest function below f which is convex in all minors, is established. Also presented are a rigorous error analysis with a focus on reliability and optimal orders of convergence, an efficient strategy that reduces the large number of unknowns, as well as numerical experiments.

Key words. nonconvex variational problem, calculus of variations, quasi convexity, relaxed variational problems, microstructure, adaptive algorithm

AMS subject classifications. 65N12, 65N15, 65N30

DOI. 10.1137/S0036142903428840

1. Introduction.

A nonconvex variational problem due to [BJ] modeling phase transitions in crystalline solids and allowing for microstructure reads

$$\text{(M)} \qquad \text{Minimize} \quad I(u) := \int_\Omega f(x, u, \nabla u)\, dx \quad \text{among } u \in \mathcal{A}$$

for a bounded Lipschitz domain $\Omega \subseteq \mathbb{R}^n$, $p \geq 1$, a (nonconvex) continuous energy density $f : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^{n\times m} \to \mathbb{R}$ satisfying p-growth conditions, and a space of admissible deformations $\mathcal{A} \subseteq W^{1,p}(\Omega; \mathbb{R}^m)$ containing boundary conditions. Since I may not be weakly lower semicontinuous, minimizing sequences develop oscillations in the gradient variable, and their weak limits do not in general minimize I (see, e.g., [Da2, M, R2]). Together with a Young measure generated by a minimizing sequence in the sense of [B2], weak limits contain the most relevant information about microscopic and macroscopic effects. Moreover, each weak limit of a minimizing sequence is a solution of a relaxed problem in which f is replaced by its quasi-convex envelope f^{qc} (see, e.g., [Da2, M, R2]). In general, it is not possible to compute f^{qc} explicitly or even approximately in order to define the relaxed problem. Therefore, it is desirable to know upper and lower bounds for f^{qc}, and it is the aim of this paper to establish a reliable and efficient algorithm that computes a lower bound. Numerical schemes for the approximation of upper bounds can be found in [Do, DW, Ba2].

Error estimates for the approximation of (M) are available for the case either that $\mathcal{A}$ contains affine boundary conditions on $\partial\Omega$ defined through certain $F \in \mathbb{R}^{n\times m}$ (see, e.g., [L, CM, BP]) or that f^{qc} is convex (see, e.g., [NW2, CP1, CP2, CR, Ba1]). In the first case theoretical convergence rates for the approximation of (M) and thereby of $f^{qc}(\cdot, \cdot, F)$ are stated, but, owing to mesh-dependent oscillations, those approaches cannot be expected to lead to efficient numerical algorithms. In the second case efficient algorithms are available, but the proposed numerical schemes are restricted

*Received by the editors May 27, 2003; accepted for publication (in revised form) August 27, 2004; published electronically June 14, 2005.

http://www.siam.org/journals/sinum/43-1/42884.html

†Department of Mathematics, University of Maryland, College Park, MD 20742 (sba@math.umd.edu). The author gratefully acknowledges support by the DFG through the priority program 1095 “Analysis, Modeling and Simulation of Multiscale Problems.”

to scalar problems. An algorithm that checks for different notions of convexity for a class of functions can be found in [DH].

By computing a (polyconvex) Young measure solution for (M) with affine boundary conditions, our iterative algorithm approximates the polyconvex envelope [B1, Da1] f^{pc} of f as a lower bound for f^{qc}. A straightforward discretization linearizes nonlinear constraints and results in a large but linear optimization problem. We show that for a large class of functions f the approximation is very accurate. The efficient iterative strategy for the solution of the linear optimization problem is based on results in [R1] that state sharp estimates on the support of a (polyconvex) Young measure solution for (M). Moreover, the strategy employs and generalizes a multilevel scheme of [CR] for the approximation of scalar nonconvex variational problems.

The proposed algorithm can be employed for the simultaneous (polyconvex) relaxation and approximation of nonconvex variational problems. This approach results in discrete problems with two numerical scales that reflect microscopic and macroscopic effects. We refer to [NW1, HH, ML, Kr, Ba1, Do, DW] for related numerical experiments. Moreover, in combination with the algorithms of [Do, DW, Ba2] for the approximation of an upper bound, the results of this paper allow are to numerically check for equality of polyconvex and rank-1 convex envelopes.

The rest of the paper is organized as follows. We present the approximation scheme with an error estimate in section 2. Some preliminaries in section 3 lead to the proof of the main result, which is given in section 4. Section 5 is devoted to a reliable and efficient algorithm that realizes the approximation scheme. Numerical experiments that illustrate the high efficiency and accuracy of the proposed algorithm are reported in section 6. Section 7 discusses the effective numerical solution of (M) based on the approximation of polyconvex envelopes.

2. Approximation scheme and main results. Throughout this article we suppose that f in (M) is independent of x and u; i.e., $f : \mathbb{R}^{n\times m} \to \mathbb{R}$, is continuous, and satisfies, for certain $c_f > 0$, $c_f' \geq 0$, $p > 0$, and all $F \in \mathbb{R}^{n\times m}$,

$$f(F) \geq c_f |F|^p - c_f'. \tag{2.1}$$

The polyconvex envelope f^{pc} of f is for $F \in \mathbb{R}^{n\times m}$ given by (see [B1, Da1])

$$f^{pc}(F) = \inf\left\{ \sum_{\ell=1}^{\tau+1} \varrho_\ell f(A_\ell) : A_\ell \in \mathbb{R}^{n\times m},\ \varrho_\ell \geq 0,\ \sum_{\ell=1}^{\tau+1} \varrho_\ell = 1,\ \sum_{\ell=1}^{\tau+1} \varrho_\ell T(A_\ell) = T(F) \right\}.$$

Here, $T(A) \in \mathbb{R}^\tau$ is a vector containing all minors of the matrix $A \in \mathbb{R}^{n\times m}$ in a fixed order and τ denotes its length; there holds $|T(A)| \leq c_T |A|_\infty^{\min\{n,m\}}$ if $|\cdot|_\infty$ denotes the maximum norm and $|A|_\infty \geq 1$. Choosing a set of points $\mathcal{N}_{d,r} := d\mathbb{Z}^{n\times m} \cap \overline{B_r(0)}$ for $r \geq d > 0$ and $B_r(0) := \{A \in \mathbb{R}^{n\times m} : |A|_\infty < r\}$ such that $F \in \operatorname{conv}\mathcal{N}_{d,r}$, an approximation of $f^{pc}(F)$ reads

$$\begin{aligned} f^{pc}_{d,r}(F) := \inf\Bigg\{ &\sum_{A\in\mathcal{N}_{d,r}} \theta_A f(A) : \forall A \in \mathcal{N}_{d,r},\ \theta_A \geq 0, \\ &\sum_{A\in\mathcal{N}_{d,r}} \theta_A = 1,\ \sum_{A\in\mathcal{N}_{d,r}} \theta_A T(A) = T(F) \Bigg\}. \end{aligned}$$

The latter infimum defines a finite-dimensional *linear* optimization problem and admits a solution and a Lagrange multiplier $\lambda^F_{d,r} \in \mathbb{R}^\tau$ associated with the constraint

$\sum_{A\in\mathcal{N}_{d,r}}\theta_A T(A)=T(F)$. Our main results concerning the approximation of polyconvex envelopes are summarized in Theorem A. We refer to section 4 for more general assertions and to [BKK] for conditions that ensure $f^{pc}\in C^{1,\alpha}_{loc}(\mathbb{R}^{n\times m})$ together with explicit bounds on $|f^{pc}|_{C^{1,\alpha}(B_d(F))}$.

THEOREM A. *Suppose that $F\in\operatorname{conv}\mathcal{N}_{d,r}$, $p\ge\min\{n,m\}=:n\wedge m$, $r\ge 1$, the computable a posteriori condition*

$$c_T(n\wedge m)|\lambda^F_{d,r}|\le pc_f r^{p-n\wedge m}\quad\text{and}\quad c_T|\lambda^F_{d,r}|r^{n\wedge m}-c_f r^p+c_f'\le\lambda^F_{d,r}\cdot T(F)-f^{pc}_{d,r}(F)$$

is satisfied, and $f\in C^{1,\alpha}_{loc}(\mathbb{R}^{n\times m})$ for some $\alpha\in[0,1]$. Then $f^{pc}_{d,r}(F)=\tilde f^{pc}_d(F)$ for a polyconvex function $\tilde f^{pc}_d:\mathbb{R}^{n\times m}\to\mathbb{R}$, $f^{pc}_{d,r}(F)=f^{pc}_{d,s}(F)$ for all $s\ge r$, and there exists $r'\ge r$ such that

$$|f^{pc}_{d,r}(F)-f^{pc}(F)|\le c_1\,d^{1+\alpha}|f|_{C^{1,\alpha}(B_{r'}(0))}.$$

If, additionally, $\alpha>0$ and $f^{pc}\in C^{1,\alpha}_{loc}(\mathbb{R}^{n\times m})$, then

$$|\lambda^F_{d,r}\cdot DT(F)-Df^{pc}(F)|\le c_2\,d^{\alpha}\big(|f|_{C^{1,\alpha}(B_{r'}(0))}+|f^{pc}|_{C^{1,\alpha}(B_d(F))}\big).$$

The constants $c_1,c_2>0$ depend only on n and m.

It can be shown that $\lambda^F_{d,r}$ and $f^{pc}_{d,r}(F)$ remain bounded for $r\to\infty$, so that the a posteriori condition of the theorem is satisfied if $p>n\wedge m$ and if r is large enough. The direct computation of $f^{pc}_{d,r}(F)$ requires the solution of a linear optimization problem with $(r/d)^{nm}$ unknowns and would therefore be very expensive. The combination of an active set strategy (due to [CR] for $\min\{n,m\}=1$) with local grid refinement and coarsening to avoid checking a maximum principle in all nodes of $\mathcal{N}_{d,r}$ leads to a very efficient but still reliable iterative algorithm that computes $f^{pc}_{d,r}(F)$.

3. Preliminaries. Throughout this article, $|\cdot|$ denotes the Frobenius norm of a vector or a matrix in $\mathbb{R}^n$, $\mathbb{R}^m$, $\mathbb{R}^\tau$, or $\mathbb{R}^{n\times m}$; e.g., for $A\in\mathbb{R}^{n\times m}$ with entries $(A)_{j,k}\in\mathbb{R}$ for $j=1,\dots,n$ and $k=1,\dots,m$

$$|A|^2=\sum_{j=1}^n\sum_{k=1}^m(A)^2_{j,k}.$$

The maximum norm of a vector or a matrix is denoted by $|\cdot|_\infty$, e.g., $|A|_\infty=\max_{j,k}|(A)_{j,k}|$; there holds $|v|_\infty\le|v|\le\sqrt{\ell}|v|_\infty$ for all $v\in\mathbb{R}^\ell$.

Given $r>0$ and $G\in\mathbb{R}^\ell$, we set $B_r(G):=\{A\in\mathbb{R}^\ell:|A-G|_\infty<r\}$ and, for a positive parameter $d>0$ with $d\le r$, define (cf. the left plot in Figure 1)

$$\mathcal{N}_{d,r}:=d\mathbb{Z}^{n\times m}\cap\overline{B_r(0)}\subseteq\mathbb{R}^{n\times m};$$

$\mathbb{Z}$ denotes the set of all integers. We let $\omega_{d,r}$ be the interior of the union of all closed (nm)-dimensional cubes $Q\subseteq\overline{B_r(0)}$ with vertices in $\mathcal{N}_{d,r}$, and define a uniform triangulation $\mathcal{T}_{d,r}$ of $\omega_{d,r}$ by setting (cf. the left plot in Figure 1)

$$\mathcal{T}_{d,r}:=\big\{Q\subseteq\overline{\omega}_{d,r}:Q\text{ is a closed cube with vertices in }\mathcal{N}_{d,r}\text{ and edges of length }d\big\}.$$

Note that each $Q\in\mathcal{T}_{d,r}$ is the convex hull of 2^{nm} nodes $M_1,\dots,M_{2^{nm}}\in\mathcal{N}_{d,r}$; i.e., $Q=\operatorname{conv}\{M_1,\dots,M_{2^{nm}}\}$. To $\mathcal{T}_{d,r}$ we associate the set of continuous $\mathcal{T}_{d,r}$-elementwise (nm)-linear functions

$$\mathcal{S}^1(\mathcal{T}_{d,r}):=\big\{v_h\in C(\overline{\omega}_{d,r}):\forall Q\in\mathcal{T}_{d,r},\ v_h|_Q\text{ is a polynomial of partial degree}\le 1\big\}.$$

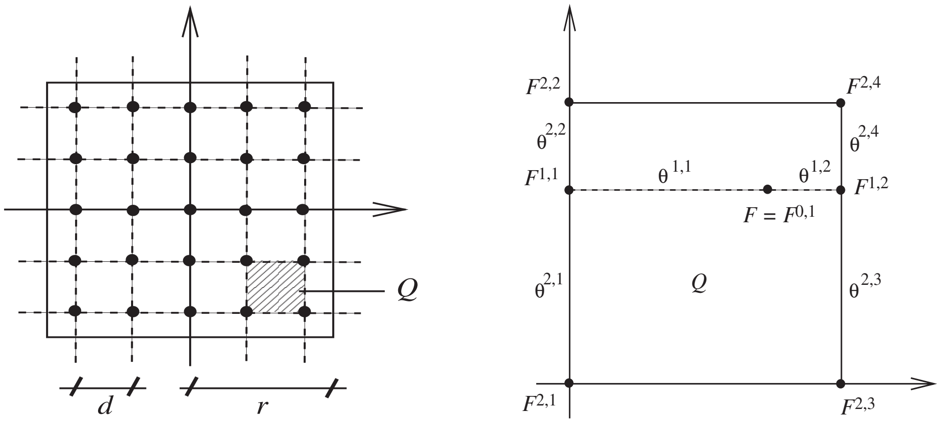


FIG. 1. *Left: set of nodes $\mathcal{N}_{d,r}$ (filled circles) and $Q \in \mathcal{T}_{d,r}$ (shaded small box). Right: decomposition of the matrix F in the proof of Lemma* 3.1 *for $n = 2$ and $m = 1$.*

The nodal interpolation operator $\mathcal{I}_{d,r}$ on $\mathcal{T}_{d,r}$ is for $v \in C(\overline{\omega}_{d,r})$ defined by

$$\mathcal{I}_{d,r} v := \sum_{A \in \mathcal{N}_{d,r}} v(A) \varphi_A.$$

Here, for each $A \in \mathcal{N}_{d,r}$ the function $\varphi_A \in \mathcal{S}^1(\mathcal{T}_{d,r})$ satisfies $\varphi_A(A) = 1$ and $\varphi_A(B) = 0$ for all $B \in \mathcal{N}_{d,r} \setminus \{A\}$. There exists $c_{\mathcal{I}} > 0$ such that

$$\|\mathcal{I}_{d,r} g - g\|_{L^\infty(\omega_{d,r})} \le c_{\mathcal{I}} d^{1+\alpha} |g|_{C^{1,\alpha}(B_r(0))} \tag{3.1}$$

for $\alpha \in (0, 1]$ and $g \in C^{1,\alpha}_{loc}(\mathbb{R}^{n\times m})$ or for $\alpha = 0$ and a locally Lipschitz continuous function $g : \mathbb{R}^{n\times m} \to \mathbb{R}$; $|g|_{C^{1,\alpha}(B_r(0))}$ denotes the α-Hölder constant of Dg on $B_r(0)$ if $\alpha > 0$, i.e.,

$$|g|_{C^{1,\alpha}(B_r(0))} := \sup_{G,H \in B_r(0)} \frac{|Dg(G) - Dg(H)|}{|G - H|^\alpha},$$

and the Lipschitz constant $|g|_{C^{1,0}(B_r(0))} := |g|_{Lip,r} = |g|_{Lip(B_r(0))}$ of g on $B_r(0)$ if $\alpha = 0$.

The operator $T : \mathbb{R}^{n\times m} \to \mathbb{R}^\tau$ for $A \in \mathbb{R}^{n\times m}$ is defined by

$$\begin{aligned} T(A) = \big((A)_{1,1}, \dots, (A)_{1,m}, (A)_{2,1}, \dots, (A)_{2,m}, \dots,& \\ (A)_{n,1}, \dots, (A)_{n,m}, \operatorname{adj}_2 A, \dots, \operatorname{adj}_{n\wedge m} A\big),& \end{aligned}$$

where for $2 \le \ell \le n \wedge m = \min\{n, m\}$, $\operatorname{adj}_\ell A$ is a vector containing all $\ell \times \ell$ minors of A and

$$\tau := \tau(n, m) = \sum_{\ell=1}^{n\wedge m} \sigma(\ell) \quad \text{for} \quad \sigma(\ell) = \frac{m!\, n!}{(\ell!)^2 (m-\ell)!(n-\ell)!}.$$

There exists $c_T > 0$ (which depends on n and m) such that $|T(A)| \le c_T |A|_\infty^{n\wedge m}$ for all $A \in \mathbb{R}^{n\times m}$ with $|A|_\infty \ge 1$; for $n = m = 2$ we have $\operatorname{adj}_2 A = \det A$ and we can choose $c_T = 2\sqrt{2}$.

The following observation is of central importance in our analysis. It shows that the values of the nodal basis functions in $\mathcal{S}^1(\mathcal{T}_{d,r})$ define a rank-1 decomposition of a matrix $F \in \overline{\omega}_{d,r}$.

LEMMA 3.1. *Let $Q \in \mathcal{T}_{d,r}$ and $M_1, \dots, M_{2^{nm}} \in \mathcal{N}_{d,r}$ be such that $Q = \operatorname{conv}\{M_1, \dots, M_{2^{nm}}\}$. Let $F \in Q$ and define $\theta_\iota := \varphi_{M_\iota}(F) \geq 0$, $\iota = 1, \dots, 2^{nm}$. There holds*

$$\sum_{\iota=1}^{2^{nm}} \theta_\iota = 1 \quad \textit{and} \quad \sum_{\iota=1}^{2^{nm}} \theta_\iota T(M_\iota) = T(F). \tag{3.2}$$

Proof. We construct convex coefficients that satisfy (3.2) and then show that they equal $\varphi_{M_\iota}(F)$. Suppose first that $d = 1$ and $Q = [0,1]^{n\times m}$ is the unit cube in $\mathbb{R}^{n\times m}$ so that $\{M_1, \dots, M_{2^{nm}}\} = \{0,1\}^{n\times m}$ are the vertices of Q. Set $F^{0,1} := F$ and $\varrho^{0,1} := 1$. Then, for $j = 1, \dots, n$ and $k = 1, \dots, m$ set $\ell := (j-1)m + k$ and define $F^{\ell,2\iota-1}, F^{\ell,2\iota} \in \mathbb{R}^{n\times m}$, $\iota = 1, \dots, 2^{\ell-1}$, by setting, for $j' = 1, \dots, n$ and $k' = 1, \dots, m$,

$$\begin{aligned} \big(F^{\ell,2\iota-1}\big)_{j',k'} &:= \begin{cases} \big(F^{\ell-1,\iota}\big)_{j',k'} & \text{for } (j',k') \neq (j,k), \\ 0 & \text{for } (j',k') = (j,k), \end{cases} \\ \big(F^{\ell,2\iota}\big)_{j',k'} &:= \begin{cases} \big(F^{\ell-1,\iota}\big)_{j',k'} & \text{for } (j',k') \neq (j,k), \\ 1 & \text{for } (j',k') = (j,k). \end{cases} \end{aligned}$$

Moreover, set $\theta^{\ell,2\iota-1} := 1 - (F^{\ell-1,\iota})_{j,k}$ and $\theta^{\ell,2\iota} := (F^{\ell-1,\iota})_{j,k}$ and

$$\varrho^{\ell,2\iota-1} := \varrho^{\ell-1,\iota}\big(1 - \big(F^{\ell-1,\iota}\big)_{j,k}\big), \qquad \varrho^{\ell,2\iota} := \varrho^{\ell-1,\iota}\big(F^{\ell-1,\iota}\big)_{j,k}. \tag{3.3}$$

(The right plot in Figure 1 schematically displays the decomposition for $n = 2$ and $m = 1$.) The decomposition of F has the following properties:

(i) $\big\{F^{nm,\iota} : \iota = 1, \dots, 2^{nm}\big\} = \{0,1\}^{n\times m}$;

(ii) $\theta^{\ell,2\iota-1}, \theta^{\ell,2\iota} \geq 0$, $\theta^{\ell,2\iota-1} + \theta^{\ell,2\iota} = 1$, and $F^{\ell-1,\iota} = \theta^{\ell,2\iota-1}F^{\ell,2\iota-1} + \theta^{\ell,2\iota}F^{\ell,2\iota}$ for $\ell = 1, \dots, nm$ and $\iota = 1, \dots, 2^{\ell-1}$;

(iii) $\operatorname{rank}(F^{\ell,2\iota-1} - F^{\ell,2\iota}) = 1$ for $\ell = 1, \dots, nm$ and $\iota = 1, \dots, 2^{\ell-1}$;

(iv) $\varphi_{F^{nm,\iota}}(F) = \varrho^{nm,\iota}$ for $\iota = 1, \dots, 2^{nm}$.

The proofs of (i)–(iii) follow directly from the decomposition. To verify (iv) we note that, according to (3.3), each $\varrho^{nm,\iota}$, $\iota = 1, \dots, 2^{nm}$, defines a polynomial in F of partial degree ≤ 1. Moreover, if $F \in \{0,1\}^{n\times m}$, then $F = F^{nm,\iota}$ for some $\iota \in \{1, \dots, 2^{nm}\}$, and by construction we then have $\varrho^{nm,\iota} = 1$ and $\varrho^{nm,\iota'} = 0$ for $\iota' \in \{1, \dots, 2^{nm}\} \setminus \{\iota\}$. This proves (iv).

Set $\theta_\iota := \varrho^{nm,\iota}$ for $\iota = 1, \dots, 2^{nm}$. The assertion of the lemma (for $Q = \operatorname{conv}\{0,1\}^{n\times m}$) follows from an induction over $\ell = 1, \dots, 2^{nm}$ with (i)–(iv) and the fact that T is affine along rank-1 connections. The case $Q \neq [0,1]^{n\times m}$ follows with a dilation and a translation from the special case. ∎

4. Proof of Theorem A. This section is devoted to the proof of Theorem A, which follows from several propositions that state more general results. The first proposition is a partial version of Theorem A but does not state sufficient conditions for an efficient choice of r. Throughout this section we consider a fixed $F \in \mathbb{R}^{n\times m}$ and assume that either $\alpha = 0$ and f is locally Lipschitz continuous or $\alpha \in (0,1]$ and $f \in C^{1,\alpha}_{loc}(\mathbb{R}^{n\times m})$.

PROPOSITION 4.1. *There exists $r' = r'(F) > 0$ such that*

$$|f^{pc}(F) - f^{pc}_{d,r'}(F)| \leq 2c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))}.$$

Proof. Let $t>0$ be such that $|f|_{C^{1,\alpha}(B_t(0))}>0$. By definition of $f^{pc}(F)$ there exist $\varrho_\ell \geq 0$ and $A_\ell \in \mathbb{R}^{n\times m}$, $\ell=1,\dots,\tau+1$, such that $\sum_{\ell=1}^{\tau+1}\varrho_\ell=1$, $\sum_{\ell=1}^{\tau+1}\varrho_\ell T(A_\ell)=T(F)$, and

$$f^{pc}(F)\leq \sum_{\ell=1}^{\tau+1}\varrho_\ell f(A_\ell)\leq f^{pc}(F)+c_{\mathcal{I}}d^{1+\alpha}|f|_{C^{1,\alpha}(B_t(0))}. \tag{4.1}$$

Choose $r'\geq t$ such that $A_1,\dots,A_{\tau+1}\in\omega_{d,r'}$. For each $\ell=1,\dots,\tau+1$, Lemma 3.1 guarantees the existence of $M_\iota^{(\ell)}\in\mathcal{N}_{d,r'}$ and $\theta_{M_\iota^{(\ell)}}=\varphi_{M_\iota^{(\ell)}}(A_\ell)\geq 0$, $\iota=1,\dots,2^{nm}$, such that $\sum_{\iota=1}^{2^{nm}}\theta_{M_\iota^{(\ell)}}=1$, $\sum_{\iota=1}^{2^{nm}}\theta_{M_\iota^{(\ell)}}T(M_\iota^{(\ell)})=T(A_\ell)$, and $A_\ell\in Q_\ell=\operatorname{conv}\{M_1^{(\ell)},\dots,M_{2^{nm}}^{(\ell)}\}\in\mathcal{T}_{d,r'}$. For $A\in\mathcal{N}_{d,r'}$ and $B\in\mathbb{R}^{n\times m}$ let $\chi_A(B)=1$ if $A=B$ and $\chi_A(B)=0$ otherwise. Setting for each $A\in\mathcal{N}_{d,r'}$

$$\tilde\theta_A=\sum_{\ell=1}^{\tau+1}\sum_{\iota=1}^{2^{nm}}\varrho_\ell\,\theta_{M_\iota^{(\ell)}}\chi_A(M_\iota^{(\ell)}) \tag{4.2}$$

defines a feasible $(\tilde\theta_A : A\in\mathcal{N}_{d,r'})$ to compute $f^{pc}_{d,r'}(F)$ so that

$$f^{pc}(F)\leq f^{pc}_{d,r'}(F)=\sum_{A\in\mathcal{N}_{d,r'}}\theta_A f(A)\leq\sum_{A\in\mathcal{N}_{d,r'}}\tilde\theta_A f(A), \tag{4.3}$$

where $(\theta_A : A\in\mathcal{N}_{d,r'})$ is optimal in $f^{pc}_{d,r'}(F)$. Since $|f|_{C^{1,\alpha}(B_t(0))}\leq|f|_{C^{1,\alpha}(B_{r'}(0))}$ and $\theta_{M_\iota^{(\ell)}}=\varphi_{M_\iota^{(\ell)}}(A_\ell)$, estimates (3.1) and (4.1)–(4.3) imply

$$\begin{aligned}0&\leq f^{pc}_{d,r'}(F)-f^{pc}(F)\\&\leq\sum_{A\in\mathcal{N}_{d,r'}}\tilde\theta_A f(A)-\sum_{\ell=1}^{\tau+1}\varrho_\ell f(A_\ell)+c_{\mathcal{I}}d^{1+\alpha}|f|_{C^{1,\alpha}(B_{r'}(0))}\\&=\sum_{\ell=1}^{\tau+1}\varrho_\ell\left(\sum_{\iota=1}^{2^{nm}}\theta_{M_\iota^{(\ell)}}f(M_\iota^{(\ell)})-f(A_\ell)\right)+c_{\mathcal{I}}d^{1+\alpha}|f|_{C^{1,\alpha}(B_{r'}(0))}\\&=\sum_{\ell=1}^{\tau+1}\varrho_\ell\big(\mathcal{I}_{d,r'}f(A_\ell)-f(A_\ell)\big)+c_{\mathcal{I}}d^{1+\alpha}|f|_{C^{1,\alpha}(B_{r'}(0))}\\&\leq 2c_{\mathcal{I}}d^{1+\alpha}|f|_{C^{1,\alpha}(B_{r'}(0))},\end{aligned} \tag{4.4}$$

which proves the proposition. □

Remark 4.1. The assumption $f\in C^{1,\alpha}_{loc}(\mathbb{R}^{n\times m})$ can be replaced by $f\in C^{1,\alpha}_{loc}(U_d)$, where $U_d=\{G\in\mathbb{R}^{n\times m}:\inf_{H\in U}|G-H|_\infty<d\}$ for $U=\{H\in\mathbb{R}^{n\times m}:f^{pc}(H)=f(H)\}$.

The subsequent lemma states the Kuhn–Tucker optimality conditions for the linear optimization problem that defines $f^{pc}_{d,r}(F)$ and which we will refer to through $f^{pc}_{d,r}(F)$. In particular, the lemma characterizes the Lagrange multiplier $\lambda^F_{d,r}$ mentioned in section 2. The equations have first been employed in the context of relaxation in the calculus of variations in [R1, R2] and have further been exploited in the numerical approximation of scalar nonconvex variational problems in [CR, Ba1].

LEMMA 4.1. *There exists* $\lambda^F_{d,r}\in\mathbb{R}^\tau$ *such that*

$$\max_{A\in\mathcal{N}_{d,r}}\big(\lambda^F_{d,r}\cdot T(A)-f(A)\big)\leq\lambda^F_{d,r}\cdot T(F)-f^{pc}_{d,r}(F). \tag{4.5}$$

Conversely, any $(\theta_A : A \in \mathcal{N}_{d,r})$ *that is feasible in* $f^{pc}_{d,r}(F)$ *is optimal if there exists* $\lambda^F_{d,r} \in \mathbb{R}^\tau$ *such that* (4.5) *holds with* $f^{pc}_{d,r}(F)$ *replaced by* $\sum_{A\in\mathcal{N}_{d,r}} \theta_A f(A)$. □

Employing the optimality conditions of Lemma 4.1, we can state sufficient conditions that ensure that $r \le r'$ is large enough so that $f^{pc}_{d,r}(F) = f^{pc}_{d,r'}(F)$, where r' is as in Proposition 4.1. If $p > n \wedge m$, condition (4.6) below can be employed as an a posteriori criterion to iteratively enlarge r.

PROPOSITION 4.2. *Let* $\lambda^F_{d,r} \in \mathbb{R}^\tau$ *be as in Lemma* 4.1, *suppose* $p \ge n \wedge m$, $r \ge 1$, *and*

$$\begin{aligned} c_T(n\wedge m)|\lambda^F_{d,r}| &\le p c_f r^{p-n\wedge m} \quad \textit{and} \\ c_T|\lambda^F_{d,r}| r^{n\wedge m} - c_f r^p + c'_f &\le \lambda^F_{d,r}\cdot T(F) - f^{pc}_{d,r}(F). \end{aligned} \tag{4.6}$$

Then, there holds

$$\max_{A\in d\mathbb{Z}^{n\times m}} \big(\lambda^F_{d,r}\cdot T(A) - f(A)\big) \le \lambda^F_{d,r}\cdot T(F) - f^{pc}_{d,r}(F) \tag{4.7}$$

and $f^{pc}_{d,r'}(F) = f^{pc}_{d,r}(F)$ *for all* $r' \ge r$.

Proof. Since $c_T(n\wedge m)|\lambda^F_{d,r}| \le p c_f r^{p-n\wedge m}$ and $p \ge n\wedge m$, the mapping

$$\{t \in \mathbb{R} : t \ge r\} \to \mathbb{R}, \quad t \mapsto c_f t^p - c'_f - c_T|\lambda^F_{d,r}| t^{n\wedge m}$$

is monotonically increasing. Since $|T(G)| \le c_T |G|^{n\wedge m}_\infty$ for all $G \in \mathbb{R}^{n\times m}$ with $|G|_\infty \ge 1$, we have, for all $A \in d\mathbb{Z}^{n\times m} \setminus \mathcal{N}_{d,r}$, i.e., for all $A \in d\mathbb{Z}^{n\times m}$ with $|A|_\infty > r \ge 1$,

$$\lambda^F_{d,r}\cdot T(A) - f(A) \le c_T|\lambda^F_{d,r}||A|^{n\wedge m}_\infty - c_f|A|^p_\infty + c'_f \le c_T|\lambda^F_{d,r}| r^{n\wedge m} - c_f r^p + c'_f.$$

Then, the second inequality in (4.6) implies, for all $A \in d\mathbb{Z}^{n\times m} \setminus \mathcal{N}_{d,r}$,

$$f(A) \ge \lambda^F_{d,r}\cdot T(A) - \lambda^F_{d,r}\cdot T(F) + f^{pc}_{d,r}(F),$$

while the optimality conditions (4.5) guarantee, for all $A \in \mathcal{N}_{d,r}$,

$$f(A) \ge \lambda^F_{d,r}\cdot T(A) - \lambda^F_{d,r}\cdot T(F) + f^{pc}_{d,r}(F).$$

The last two estimates prove (4.7). Let $r' \ge r$, and let $(\tilde\theta_A : A \in \mathcal{N}_{d,r'})$ be a solution to $f^{pc}_{d,r'}(F)$. Employing (4.7) and $\sum_{A\in\mathcal{N}_{d,r'}} \tilde\theta_A T(A) = T(F)$, we infer

$$f^{pc}_{d,r'}(F) = \sum_{A\in\mathcal{N}_{d,r'}} \tilde\theta_A f(A) \ge \sum_{A\in\mathcal{N}_{d,r'}} \tilde\theta_A\big(\lambda^F_{d,r}\cdot T(A) - \lambda^F_{d,r}T(F) + f^{pc}_{d,r}(F)\big) = f^{pc}_{d,r}(F).$$

The obvious estimate $f^{pc}_{d,r'}(F) \le f^{pc}_{d,r}(F)$ concludes the proof. □

We now turn to estimates for $\lambda^F_{d,r}$, for which we need to construct a polyconvex extension of $f^{pc}_{d,r}(F)$ to $\mathbb{R}^{n\times m}$. Note that the subsequent definition of this extension does not depend on F; i.e., it depends only on d.

DEFINITION 4.1. *Let* $\tilde f_d$ *be the nodal interpolant of* f *on* $\mathbb{R}^{n\times m}$ *with respect to nodes in* $d\mathbb{Z}^{n\times m}$; *i.e.,* $\tilde f_d(A) = f(A)$ *for all* $A \in d\mathbb{Z}^{n\times m}$, $\tilde f_d$ *is continuous, and* $\tilde f_d|_{\tilde Q}$ *is* (nm)*-linear for each cube* $\tilde Q \subseteq \mathbb{R}^{n\times m}$ *with vertices in* $d\mathbb{Z}^{n\times m}$ *and edges of length* d. *Then let* $\tilde f^{pc}_d$ *be the polyconvex envelope of* $\tilde f_d$; *i.e., for* $A \in \mathbb{R}^{n\times m}$,

$$\tilde f^{pc}_d(A) = \inf\left\{\sum_{\ell=1}^{\tau+1} \varrho_\ell \tilde f_d(A_\ell) : A_\ell \in \mathbb{R}^{n\times m},\ \varrho_\ell \ge 0,\ \sum_{\ell=1}^{\tau+1}\varrho_\ell = 1,\ \sum_{\ell=1}^{\tau+1}\varrho_\ell T(A_\ell) = T(A)\right\}.$$

The following lemma shows that $\tilde{f}_d^{pc}$ can be approximated arbitrarily well by convex combinations of nodal values of f.

LEMMA 4.2. *For all $\varepsilon > 0$ and all $A \in \mathbb{R}^{n\times m}$ there exist $B_\kappa \in d\mathbb{Z}^{n\times m}$, $\gamma_\kappa \geq 0$, $\kappa = 1, \dots, 2^{nm}(\tau+1)$, such that $\sum_{\kappa=1}^{2^{nm}(\tau+1)} \gamma_\kappa = 1$, $\sum_{\kappa=1}^{2^{nm}(\tau+1)} \gamma_\kappa T(B_\kappa) = T(A)$, and*

$$\tilde{f}_d^{pc}(A) \leq \sum_{\kappa=1}^{2^{nm}(\tau+1)} \gamma_\kappa f(B_\kappa) \leq \tilde{f}_d^{pc}(A) + \varepsilon. \tag{4.8}$$

Remark 4.2. Employing optimality conditions for the minimization problem $\tilde{f}_d^{pc}(A)$ in Definition 4.1, one can show that the infimum is attained if $p > n \wedge m$. In this case one may choose $\varepsilon = 0$ in Lemma 4.2.

Proof of Lemma 4.2. By definition of $\tilde{f}_d^{pc}(A)$ there exist $A_\ell \in \mathbb{R}^{n\times m}$ and $\varrho_\ell \geq 0$, $\ell = 1, \dots, \tau+1$, such that $\sum_{\ell=1}^{\tau+1} \varrho_\ell = 1$, $\sum_{\ell=1}^{\tau+1} \varrho_\ell T(A_\ell) = T(A)$, and

$$\sum_{\ell=1}^{\tau+1} \varrho_\ell \tilde{f}_d(A_\ell) \leq \tilde{f}_d^{pc}(A) + \varepsilon.$$

For each $\ell = 1, \dots, \tau+1$ let $\tilde{Q}_\ell = \operatorname{conv}\{M_1^{(\ell)}, \dots, M_{2^{nm}}^{(\ell)}\}$ be a cube in $\mathbb{R}^{n\times m}$ with vertices $M_1^{(\ell)}, \dots, M_{2^{nm}}^{(\ell)} \in d\mathbb{Z}^{n\times m}$, edges of length d, and such that $A_\ell \in \tilde{Q}_\ell$. By Lemma 3.1 (with $r = \tilde{r}$ for some $\tilde{r}$ large enough so that $A_\ell \in \tilde{Q}_\ell \subseteq \omega_{d,\tilde{r}}$), there exist $\theta_{M_\iota^{(\ell)}} \geq 0$, $\iota = 1, \dots, 2^{nm}$, such that $\sum_{\iota=1}^{2^{nm}} \theta_{M_\iota^{(\ell)}} = 1$, $\sum_{\iota=1}^{2^{nm}} \theta_{M_\iota^{(\ell)}} T(M_\iota^{(\ell)}) = T(A_\ell)$, and

$$\tilde{f}_d(A_\ell) = \sum_{\iota=1}^{2^{nm}} \varphi_{M_\iota^{(\ell)}}(A_\ell) f(M_\iota^{(\ell)}) = \sum_{\iota=1}^{2^{nm}} \theta_{M_\iota^{(\ell)}} f(M_\iota^{(\ell)}).$$

This implies

$$\sum_{\ell=1}^{\tau+1} \sum_{\iota=1}^{2^{nm}} \varrho_\ell \theta_{M_\iota^{(\ell)}} f(M_\iota^{(\ell)}) \leq \tilde{f}_d^{pc}(A) + \varepsilon,$$

which, after appropriate relabeling, is (4.8). □

The following assertion is due to Ball [B1].

LEMMA 4.3. *There exist convex functions $\hat{f}, \hat{f}_d : \mathbb{R}^\tau \to \mathbb{R}$ such that*

$$f^{pc} = \hat{f} \circ T \qquad \text{and} \qquad \tilde{f}_d^{pc} = \hat{f}_d \circ T.$$

For $g = f$ or $g = \tilde{f}_d$ the function $\hat{g} = \hat{f}$ or $\hat{g} = \hat{f}_d$, respectively, can be defined by

$$\hat{g}(X) = \inf\left\{ \sum_{\ell=1}^{\tau+1} \varrho_\ell g(A_\ell) : A_\ell \in \mathbb{R}^{n\times m},\ \varrho_\ell \geq 0,\ \sum_{\ell=1}^{\tau+1} \varrho_\ell = 1,\ \sum_{\ell=1}^{\tau+1} \varrho_\ell T(A_\ell) = X \right\}. \qquad \square$$

Remark 4.3. The function $\hat{g}$ is not unique, and the presented formula can be found in [Da2].

An estimate for the difference between $\hat{f}$ and $\hat{f}_d$ follows immediately.

LEMMA 4.4. (i) *Let $B^\tau = \{E_1, \dots, E_\tau\}$ be the canonical basis in $\mathbb{R}^\tau$. There exists $r' > 0$ such that, for all $E \in \pm B^\tau$, there holds*

$$|\hat{f}_d(T(F) + dE) - \hat{f}(T(F) + dE)| \leq 2c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))}.$$

(ii) *Let $B^{n\times m} = \{E_1, \dots, E_{nm}\}$ be the canonical basis in $\mathbb{R}^{n\times m}$. There exists $r' > 0$ such that, for all $E \in \pm B^{n\times m}$, there holds*

$$|\hat{f}_d(T(F+dE)) - f^{pc}(F+dE)| \le 2c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))}.$$

Proof. (i) Let $E \in \pm B^\tau$, and let $t > 0$ be such that $|f|_{C^{1,\alpha}(B_t(0))} > 0$. By definition of $\hat{f}$ and $\hat{f}_d$ there exist A_ℓ, $A_\ell^{(d)} \in \mathbb{R}^{n\times m}$ and ϱ_ℓ, $\varrho_\ell^{(d)} \ge 0$, $\ell = 1, \dots, \tau+1$, such that $\sum_{\ell=1}^{\tau+1} \varrho_\ell = \sum_{\ell=1}^{\tau+1} \varrho_\ell^{(d)} = 1$, $\sum_{\ell=1}^{\tau+1} \varrho_\ell A_\ell = \sum_{\ell=1}^{\tau+1} \varrho_\ell^{(d)} A_\ell^{(d)} = T(F) + dE$, and

$$\hat{f}(T(F)+dE) \le \sum_{\ell=1}^{\tau+1} \varrho_\ell f(A_\ell) \le \hat{f}(T(F)+dE) + c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_t(0))}$$

as well as

$$\hat{f}_d(T(F)+dE) \le \sum_{\ell=1}^{\tau+1} \varrho_\ell^{(d)} f(A_\ell^{(d)}) \le \hat{f}_d(T(F)+dE) + c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_t(0))}.$$

Let $r' \ge t$ be such that $A_\ell, A_\ell^{(d)} \in \omega_{d,r'}$ for all $\ell = 1, \dots, \tau+1$. If $\hat{f}(T(F)+dE) \ge \hat{f}_d(T(F)+dE)$, then

$$\begin{aligned} 0 &\le \hat{f}(T(F)+dE) - \hat{f}_d(T(F)+dE) \\ &\le \sum_{\ell=1}^{\tau+1} \varrho_\ell^{(d)} \big(f(A_\ell^{(d)}) - f_d(A_\ell^{(d)}) \big) + c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))} \\ &\le 2c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))}. \end{aligned}$$

Otherwise, if $\hat{f}(T(F)+dE) \le \hat{f}_d(T(F)+dE)$, then

$$\begin{aligned} 0 &\le \hat{f}_d(T(F)+dE) - \hat{f}(T(F)+dE) \\ &\le \sum_{\ell=1}^{\tau+1} \varrho_\ell \big(f_d(A_\ell) - f(A_\ell) \big) + c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))} \\ &\le 2c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))}. \end{aligned}$$

Choosing r' maximal so that for each $E \in \pm B^\tau$ one of the last two estimates holds proves the first part of the lemma.

(ii) The proof of the second assertion is similar to the proof of (i) and uses the fact that $\hat{f} \circ T = f^{pc}$. □

The next lemma is the key observation for the estimates for $\lambda_{d,r}^F$ for which we employ the concept of subgradients. Some elementary facts about the subgradient are cited in the following remark.

Remark 4.4 (see [C]). Let $h : \mathbb{R}^\ell \to \mathbb{R}$ be a continuous convex function. For $v_0 \in \mathbb{R}^\ell$ define

$$\partial h(v_0) := \{ s \in \mathbb{R}^\ell : \forall v \in \mathbb{R}^\ell, s \cdot (v - v_0) \le h(v) - h(v_0) \}.$$

The following hold: (i) If $g : \mathbb{R} \to \mathbb{R}^\ell$ is affine and $g(0) = v_0$, then $\partial(h \circ g)(0) = \partial h(v_0) \cdot Dg(0)$. (ii) If $\ell = 1$, then $\partial h(v_0) \subseteq \big[(h(v_0) - h(v_0 - s))/s, (h(v_0+s) - h(v_0))/s \big]$ for all $s > 0$. (iii) If $h(v_0) \le h(v)$ for all $v \in \mathbb{R}^\ell$, then $0 \in \partial h(v_0)$.

Lemma 4.5. *Suppose that*

$$\max_{A \in d\mathbb{Z}^{n\times m}} \big(\lambda_{d,r}^F \cdot T(A) - f(A) \big) \le \lambda_{d,r}^F \cdot T(F) - f_{d,r}^{pc}(F).$$

Then it holds that $\tilde{f}_d^{pc}(F) = f_{d,r}^{pc}(F)$ *and* $\lambda_{d,r}^F \in \partial \hat{f}_d(T(F))$.

Proof. We show that $\tilde{f}_d^{pc}(F) = f_{d,r}^{pc}(F)$ and

$$\max_{A\in\mathbb{R}^{n\times m}} \big(\lambda_{d,r}^F \cdot T(A) - \tilde{f}_d^{pc}(A)\big) \le \lambda_{d,r}^F \cdot T(F) - \tilde{f}_d^{pc}(F). \tag{4.9}$$

Then the asserted inclusion is deduced from these observations as follows. Let $\varepsilon > 0$. For $X \in \mathbb{R}^\tau$ there exist $A_\ell \in \mathbb{R}^{n\times m}$, $\varrho_\ell \ge 0$, $\ell = 1,\dots,\tau+1$, such that $\sum_{\ell=1}^{\tau+1} \varrho_\ell = 1$, $\sum_{\ell=1}^{\tau+1} \varrho_\ell T(A_\ell) = X$, and

$$\hat{f}_d(X) \ge \sum_{\ell=1}^{\tau+1} \varrho_\ell \tilde{f}_d(A_\ell) - \varepsilon.$$

Using $\tilde{f}_d^{pc} \le \tilde{f}_d$, $\tilde{f}_d^{pc}(F) = \hat{f}_d(T(F))$, and (4.9), we deduce

$$\begin{aligned}
\hat{f}_d(X) - \lambda_{d,r}^F \cdot X &\ge \sum_{\ell=1}^{\tau+1} \varrho_\ell \big(\tilde{f}_d(A_\ell) - \lambda_{d,r}^F \cdot T(A_\ell)\big) - \varepsilon \\
&\ge \sum_{\ell=1}^{\tau+1} \varrho_\ell \big(\tilde{f}_d^{pc}(A_\ell) - \lambda_{d,r}^F \cdot T(A_\ell)\big) - \varepsilon \\
&\ge \tilde{f}_d^{pc}(F) - \lambda_{d,r}^F \cdot T(F) - \varepsilon \\
&= \hat{f}_d(T(F)) - \lambda_{d,r}^F \cdot T(F) - \varepsilon.
\end{aligned}$$

By arbitrariness of $\varepsilon > 0$, the convex function $X \mapsto \hat{f}_d(X) - \lambda_{d,r}^F \cdot X$, $X \in \mathbb{R}^\tau$, has a minimum in $T(F)$ so that (iii) in Remark 4.4 implies the asserted inclusion.

To verify $\tilde{f}_d^{pc}(F) = f_{d,r}^{pc}(F)$ we note that $\tilde{f}_d^{pc}(F) \le f_{d,r}^{pc}(F)$ and let $\varepsilon > 0$. By Lemma 4.2 there exist $B_\kappa \in d\mathbb{Z}^{n\times m}$, $\gamma_\kappa \ge 0$, $\kappa = 1,\dots,2^{nm}(\tau+1)$, such that $\sum_{\kappa=1}^{2^{nm}(\tau+1)} \gamma_\kappa = 1$, $\sum_{\kappa=1}^{2^{nm}(\tau+1)} \gamma_\kappa T(B_\kappa) = T(F)$, and

$$\sum_{\kappa=1}^{2^{nm}(\tau+1)} \gamma_\kappa f(B_\kappa) \le \tilde{f}_d^{pc}(F) + \varepsilon.$$

The hypothesis of the lemma implies, for $\kappa = 1,\dots,2^{nm}(\tau+1)$,

$$f(B_\kappa) \ge \lambda_{d,r}^F \cdot T(B_\kappa) - \lambda_{d,r}^F \cdot T(F) + f_{d,r}^{pc}(F),$$

and this estimate yields

$$\tilde{f}_d^{pc}(F) \ge \sum_{\kappa=1}^{2^{nm}(\tau+1)} \gamma_\kappa f(B_\kappa) - \varepsilon \ge f_{d,r}^{pc}(F) - \varepsilon,$$

which, by arbitrariness of $\varepsilon > 0$, shows $\tilde{f}_d^{pc}(F) \ge f_{d,r}^{pc}(F)$ and hence yields $\tilde{f}_d^{pc}(F) = f_{d,r}^{pc}(F)$.

To prove (4.9), let $A^* \in \mathbb{R}^{n\times m}$ be maximal in the left-hand side of (4.9). For $\varepsilon > 0$ Lemma 4.2 guarantees the existence of $C_\kappa \in d\mathbb{Z}^{n\times m}$ and $\delta_\kappa \ge 0$, $\kappa = 1,\dots,2^{nm}(\tau+1)$, such that $\sum_{\kappa=1}^{2^{nm}(\tau+1)} \delta_\kappa = 1$, $\sum_{\kappa=1}^{2^{nm}(\tau+1)} \delta_\kappa T(C_\kappa) = T(A^*)$, and

$$\sum_{\kappa=1}^{2^{nm}(\tau+1)} \delta_\kappa f(C_\kappa) \le \tilde{f}_d^{pc}(A^*) + \varepsilon.$$

Then the hypothesis of the lemma and $\tilde{f}_d^{pc}(F) = f_{d,r}^{pc}(F)$ imply

$$\begin{aligned}\max_{A\in\mathbb{R}^{n\times m}}\big(\lambda_{d,r}^F\cdot T(A)-\tilde{f}_d^{pc}(A)\big) &\le \lambda_{d,r}^F\cdot T(A^*)-\tilde{f}_d^{pc}(A^*)\\ &\le \sum_{\kappa=1}^{2^{nm}(\tau+1)}\delta_\kappa\big(\lambda_{d,r}^F\cdot T(C_\kappa)-f(C_\kappa)\big)+\varepsilon\\ &\le \sum_{\kappa=1}^{2^{nm}(\tau+1)}\delta_\kappa\max_{A\in d\mathbb{Z}^{n\times m}}\big(\lambda_{d,r}^F\cdot T(A)-f(A)\big)+\varepsilon\\ &\le \lambda_{d,r}^F\cdot T(F)-f_{d,r}^{pc}(F)+\varepsilon\\ &= \lambda_{d,r}^F\cdot T(F)-\tilde{f}_d^{pc}(F)+\varepsilon,\end{aligned}$$

which, by arbitrariness of $\varepsilon > 0$, is (4.9) and therefore concludes the proof. ◻

Provided that $\hat{f}$ is of class $C_{loc}^{1,\alpha}$, we have an estimate for $|\lambda_{d,r}^F - D\hat{f}(T(F))|$.

PROPOSITION 4.3. *Assume that the hypothesis of Lemma* 4.5 *is satisfied, suppose that* $\alpha > 0$, *and let* $\hat{f} \in C_{loc}^{1,\alpha}(\mathbb{R}^\tau)$. *There exists* $r' > 0$ *such that*

$$|\lambda_{d,r}^F - D\hat{f}(T(F))| \le 4\sqrt{\tau}c_{\mathcal{I}}d^\alpha|f|_{C^{1,\alpha}(B_{r'}(0))} + \sqrt{\tau}d^\alpha|\hat{f}|_{C^{1,\alpha}(B_d(T(F)))}.$$

Proof. Let $B^\tau = \{E_1,\dots,E_\tau\}$ be the canonical basis in $\mathbb{R}^\tau$. Lemma 4.5 proves

$$\begin{aligned}|\lambda_{d,r}^F - D\hat{f}(T(F))| &\le \sup_{S\in\partial\hat{f}_d(T(F))}|S-D\hat{f}(T(F))|\\ &\le \sqrt{\tau}\sup_{S\in\partial\hat{f}_d(T(F))}|S-D\hat{f}(T(F))|_\infty\\ &= \sqrt{\tau}\sup_{S\in\partial\hat{f}_d(T(F))}\max_{E\in\pm B^\tau}|(S-D\hat{f}(T(F)))\cdot E|.\end{aligned}$$

Let $S \in \partial\hat{f}_d(T(F))$ and $E \in \pm B^\tau$. Since $\hat{f}_d$ is convex, (i) and (ii) in Remark 4.4 show $S\cdot E \in \partial\hat{f}_d(T(F))\cdot E \subseteq [S_-(E), S_+(E)]$ for

$$S_\pm(E) = \pm\frac{\hat{f}_d(T(F)\pm dE)-\hat{f}_d(T(F))}{d}.$$

Assume without loss of generality $S_-(E) \le S_+(E)$. Lemma 4.4 and Proposition 4.1 (note that $\hat{f}(T(F)) = f^{pc}(F)$ and $\hat{f}_d(T(F)) = f_{d,r}^{pc}(F)$) prove

$$S_+(E) \le \frac{\hat{f}(T(F)+dE)-\hat{f}(T(F))+4c_{\mathcal{I}}d^{1+\alpha}|f|_{C^{1,\alpha}(B_{r'}(0))}}{d}$$

and

$$S_-(E) \ge \frac{\hat{f}(T(F))-\hat{f}(T(F)-dE)-4c_{\mathcal{I}}d^{1+\alpha}|f|_{C^{1,\alpha}(B_{r'}(0))}}{d}.$$

From these estimates, the mean value theorem, and Hölder continuity of $D\hat{f}$ we infer

$$|(S-D\hat{f}(T(F)))\cdot E| \le 4c_{\mathcal{I}}d^\alpha|f|_{C^{1,\alpha}(B_{r'}(0))} + d^\alpha|\hat{f}|_{C^{1,\alpha}(B_d(T(F)))},$$

which concludes the proof. ◻

It is not known under which conditions there holds $\hat{f} \in C^{1,\alpha}_{loc}(\mathbb{R}^\tau)$ or under which conditions there exists a convex function $\tilde{f} \in C^{1,\alpha}_{loc}(\mathbb{R}^\tau)$ such that $\tilde{f} = f^{pc} \circ T$. A result in [BKK] shows that if

$$\liminf_{G\to\infty} \frac{f(G)}{|G|^p} > 0 \quad \text{and} \quad \limsup_{G\to\infty} \frac{f(G)}{|G|^{p+1}} = 0$$

or

$$\frac{f(G)}{|G|^p} \to \infty \text{ as } G \to \infty \quad \text{and} \quad \limsup_{G\to\infty} \frac{f(G)}{|G|^{p+1}} < \infty$$

and if there exists $c > 0$ such that, for all $G \in \mathbb{R}^{n\times m}$, there exists $S \in \partial f(G)$ such that

$$f(G+H) - f(H) - S \cdot H \le c \max\{f(G), 1\} |H|^{1+\alpha}$$

for all $H \in B_1(0)$, then $f^{pc} \in C^{1,\alpha}_{loc}(\mathbb{R}^{n\times m})$. Observing that $D\hat{f}(T(F)) \cdot DT(F) = Df^{pc}(F)$, a small modification (that does not use *any* regularity of $\hat{f}$) of the proof of Proposition 4.3 yields the following result.

Proposition 4.4. *Assume that the hypothesis of Lemma* 4.5 *is satisfied, suppose that* $\alpha > 0$, *and let* $f^{pc} \in C^{1,\alpha}_{loc}(\mathbb{R}^{n\times m})$. *Then, there exists* $r' > 0$ *such that*

$$|\lambda^F_{d,r} \cdot DT(F) - Df^{pc}(F)| \le 4\sqrt{nm} c_{\mathcal{I}} d^\alpha |f|_{C^{1,\alpha}(B_{r'}(0))} + \sqrt{nm} d^\alpha |f^{pc}|_{C^{1,\alpha}(B_d(F))}.$$

Proof. Let $B^{n\times m} = \{E_1, \dots, E_{n\times m}\}$ be the canonical basis in $\mathbb{R}^{n\times m}$. Lemma 4.5 proves

$$|\lambda^F_{d,r} \cdot DT(F) - Df^{pc}(F)| \le \sqrt{nm} \sup_{S\in\partial\hat{f}_d(T(F))} \max_{E\in\pm B^{n\times m}} |(S \cdot DT(F) - Df^{pc}(F)) \cdot E|.$$

Let $S \in \partial\hat{f}_d(T(F))$ and $E \in \pm B^{n\times m}$. Convexity of $t \mapsto \hat{f}_d(T(F+tE))$ shows $(S \cdot DT(F)) \cdot E \in [S_-(E), S_+(E)]$ (cf. Remark 4.4) for

$$S_\pm(E) = \pm \frac{\hat{f}_d(T(F \pm dE)) - \hat{f}_d(T(F))}{d}.$$

Assuming $S_-(E) \le S_+(E)$, Lemma 4.4, $\hat{f}(T(F)) = f^{pc}_{d,r}(F)$, and Proposition 4.1 show

$$S_+(E) \le \frac{f^{pc}(F+dE) - f^{pc}(F) + 4c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))}}{d}$$

and

$$S_-(E) \ge \frac{f^{pc}(F) - f^{pc}(F-dE) - 4c_{\mathcal{I}} d^{1+\alpha} |f|_{C^{1,\alpha}(B_{r'}(0))}}{d}.$$

These estimates, the mean value theorem, and the Hölder continuity of Df^{pc} imply

$$|(S \cdot DT(F) - Df^{pc}(F)) \cdot E| \le 4c_{\mathcal{I}} d^\alpha |f|_{C^{1,\alpha}(B_{r'}(0))} + d^\alpha |f^{pc}|_{C^{1,\alpha}(B_d(F))}$$

and thereby prove the proposition. ∎

Proof of Theorem A. This is a combination of Propositions 4.1, 4.2, and 4.4 and Lemma 4.5. ∎

5. Efficient computation of $f_{d,r}^{pc}(F)$. As mentioned in the introduction, the direct computation of $f_{d,r}^{pc}(F)$ is very expensive. To reduce the number of unknowns we use a multilevel scheme with local mesh refinement and coarsening.

5.1. Grid coarsening and local refinement. The following propositions define criteria that allow us to add nodes to and remove nodes from $\mathcal{N}_{d,r}$ to obtain a set of nodes that leads to a good approximation of $f_{d/2,r}^{pc}(F)$, provided that we computed (an approximation of) $f_{d,r}^{pc}(F)$. The first assertion allows us to remove nodes $A \in \mathcal{N}_{d/2,r}$ that cannot lead to volume fractions θ_A larger than a given threshold $\delta = c/M$ for a known constant c and a (large) parameter M.

PROPOSITION 5.1. *Let $(\theta_A : A \in \mathcal{N}_{d,r})$ be feasible and optimal for $f_{d,r}^{pc}(F)$ with corresponding multiplier $\lambda_{d,r}^F \in \mathbb{R}^\tau$. For each $A \in \mathcal{N}_{d,r}$ let $M(A) \geq M > 0$ and set*

$$Z = \{A \in \mathcal{N}_{d,r} : \lambda_{d,r}^F \cdot T(A) - f(A) \leq \lambda_{d,r}^F \cdot T(F) - f_{d,r}^{pc}(F) - d\,M(A)\}.$$

Let

$$Z' = \{A' \in \mathcal{N}_{d/2,r} : \exists A \in Z,\, |A - A'| \leq d\}.$$

Then for any $(\theta'_{A'} : A' \in \mathcal{N}_{d/2,r})$ that is feasible and optimal for $f_{d/2,r}^{pc}(F)$ there holds

$$\sum_{A' \in Z'} \theta'_{A'} \leq \frac{(|f|_{Lip,r} + |\lambda_{d,r}^F||T|_{Lip,r})}{M}.$$

Proof. For $A' \in Z'$ and $A \in Z$ such that $|A - A'| \leq d$, there holds

$$\begin{aligned} f(A') &\geq f(A) - d|f|_{Lip,r} \\ &\geq f_{d,r}^{pc}(F) + \lambda_{d,r}^F \cdot T(A) - \lambda_{d,r}^F \cdot T(F) + dM(A) - d|f|_{Lip,r} \\ &\geq f_{d,r}^{pc}(F) + \lambda_{d,r}^F \cdot T(A') - \lambda_{d,r}^F \cdot T(F) + dM(A) - d|f|_{Lip,r} - d|\lambda_{d,r}^F||T|_{Lip,r}. \end{aligned}$$

Let $B' \in \mathcal{N}_{d/2,r} \setminus Z'$ and $B \in \mathcal{N}_{d,r}$ be such that $|B' - B| \leq d$. Employing (4.5) and arguing similarly as in the previous estimate, we infer

$$f(B') \geq f_{d,r}^{pc}(F) + \lambda_{d,r}^F \cdot T(B') - \lambda_{d,r}^F \cdot T(F) - d|f|_{Lip,r} - d|\lambda_{d,r}^F||T|_{Lip,r}.$$

Let $(\theta'_{A'} : A' \in \mathcal{N}_{d/2,r})$ be feasible and optimal for $f_{d/2,r}^{pc}(F)$. The previous two estimates imply

$$\begin{aligned} f_{d/2,r}^{pc}(F) = \sum_{A' \in \mathcal{N}_{d/2,r}} \theta_{A'} f(A') &= \sum_{A' \in Z'} \theta_{A'} f(A') + \sum_{B' \in \mathcal{N}_{d/2,r} \setminus Z'} \theta_{B'} f(B') \\ &\geq f_{d,r}^{pc}(F) - (|f|_{Lip,r} + |\lambda_{d,r}^F||T|_{Lip,r})d + \sum_{A' \in Z'} \theta_{A'} Md. \end{aligned}$$

Using $f_{d/2,r}^{pc}(F) \leq f_{d,r}^{pc}(F)$, we are then able to deduce $\sum_{A' \in Z'} \theta_{A'} \leq (|f|_{Lip,r} + |\lambda_{d,r}^F||T|_{Lip,r})/M$. □

A more efficient and even more reliable assertion can be formulated if we have explicit estimates for $|f_{d,r}^{pc}(F) - f_{d/2,r}^{pc}(F)|$ and $|\lambda_{d,r}^F - \lambda_{d/2,r}^F|$. Sufficient for this is the knowledge of r' in Theorem A and $\hat{f} \in C_{loc}^{1,1}(\mathbb{R}^\tau)$ with $\hat{f}$ as in Lemma 4.3.

PROPOSITION 5.2. *For each $A \in \mathcal{N}_{d,r}$ let $M(A) > 0$, suppose that $|\lambda^F_{d,r} - \lambda^F_{d/2,r}| \leq c_{LM} d$, and let Z and Z' be as in Proposition* 5.1. *Assume that for all $A \in \mathcal{N}_{d,r}$ there holds*

$$c_{LM}|T|_{Lip,r'} + |\lambda^F_{d,r}|\,|T|_{Lip,r'} + 4c_{\mathcal{I}}|f|_{Lip,r'} + c_{LM}|T(A)| + c_{LM}|T(F)| < M(A). \tag{5.1}$$

Then for any $(\theta_A : A \in \mathcal{N}_{d/2,r})$ that is feasible and optimal for $f^{pc}_{d/2,r}(F)$ there holds $\theta_{A'} = 0$ for all $A' \in Z'$.

Proof. Let $A' \in Z'$ and $A \in Z$ such that $|A - A'| \leq d$. By the hypotheses and by Lipschitz continuity of f and T there holds

$$\begin{aligned}
&\lambda^F_{d/2,r} \cdot T(A') - f(A') \\
&= \lambda^F_{d,r} \cdot T(A) - f(A) + (\lambda^F_{d/2,r} - \lambda^F_{d,r}) \cdot T(A') + \lambda^F_{d,r} \cdot (T(A') - T(A)) + (f(A) - f(A')) \\
&\leq \lambda^F_{d,r} \cdot T(A) - f(A) + c_{LM} d|T(A')| + d|\lambda^F_{d,r}||T|_{Lip,r'} + d|f|_{Lip,r'} \\
&\leq \lambda^F_{d,r} \cdot T(F) - f^{pc}_{d,r}(F) - dM(A) + c_{LM} d|T(A')| + d|\lambda^F_{d,r}||T|_{Lip,r'} + d|f|_{Lip,r'}.
\end{aligned}$$

The definitions of Z and Z', Proposition 4.1, and again the assumed estimate for $|\lambda^F_{d,r} - \lambda^F_{d/2,r}|$ show

$$\begin{aligned}
&\lambda^F_{d,r} \cdot T(F) - f^{pc}_{d,r}(F) - dM(A) + c_{LM} d|T(A')| + d|\lambda^F_{d,r}||T|_{Lip,r'} + d|f|_{Lip,r'} \\
&= \lambda^F_{d/2,r} \cdot T(F) - f^{pc}_{d/2,r}(F) + \big(\lambda^F_{d,r} - \lambda^F_{d/2,r}\big) \cdot T(F) + \big(f^{pc}_{d/2,r}(F) - f^{pc}_{d,r}(F)\big) \\
&\quad - dM(A) + c_{LM} d|T(A')| + d|\lambda^F_{d,r}||T|_{Lip,r'} + 2c_{\mathcal{I}} d|f|_{Lip,r'} \\
&\leq \lambda^F_{d/2,r} \cdot T(F) - f^{pc}_{d/2,r}(F) + c_{LM} d|T(F)| + d|\lambda^F_{d,r}||T|_{Lip,r'} - dM(A) \\
&\quad + 2c_{\mathcal{I}} d|f|_{Lip,r'} + c_{LM} d|T(F)| + 2c_{\mathcal{I}} d|f|_{Lip,r'}.
\end{aligned}$$

Employing the Lipschitz continuity of T once more proves

$$\begin{aligned}
&\lambda^F_{d/2,r} \cdot T(F) - f^{pc}_{d/2,r}(F) \\
&\qquad + d\big(c_{LM}|T(A')| + |\lambda^F_{d,r}||T|_{Lip,r'} + 2c_{\mathcal{I}} d|f|_{Lip,r'} + c_{LM}|T(F)| - M(A)\big) \\
&\quad \leq \lambda^F_{d/2,r} \cdot T(F) - f^{pc}_{d/2,r}(F) + d\big(c_{LM}|T|_{Lip,r'} + c_{LM}|T(A)| + |\lambda^F_{d,r}||T|_{Lip,r'} \\
&\qquad\qquad + 4c_{\mathcal{I}}|f|_{Lip,r'} + c_{LM}|T(F)| - M(A)\big).
\end{aligned}$$

In view of (5.1), the last three estimates imply, for all $A' \in Z'$,

$$\lambda^F_{d/2,r} \cdot T(A') - f(A') < \lambda^F_{d/2,r} \cdot T(F) - f^{pc}_{d/2,r}(F). \tag{5.2}$$

The optimality conditions ensure, for all $A \in \mathcal{N}_{d/2,r}$,

$$\lambda^F_{d/2,r} \cdot T(A) - f(A) \leq \lambda^F_{d/2,r} \cdot T(F) - f^{pc}_{d/2,r}(F). \tag{5.3}$$

Let $(\theta_A : A \in \mathcal{N}_{d/2,r})$ be feasible and optimal for $f^{pc}_{d/2,r}(F)$, and suppose that there is $A \in Z'$ such that $\theta_A > 0$. Then, (5.2) and (5.3) imply

$$\begin{aligned}
f^{pc}_{d/2,r}(F) &= \sum_{A \in \mathcal{N}_{d/2,r}} \theta_A f(A) \\
&> \sum_{A \in \mathcal{N}_{d/2,r}} \theta_A \big(\lambda^F_{d/2,r} \cdot T(A') - \lambda^F_{d/2,r} \cdot T(F) + f^{pc}_{d/2,r}(F)\big) = f^{pc}_{d/2,r}(F).
\end{aligned} \tag{5.4}$$

This is a contradiction and proves $\theta_A = 0$ for all $A \in Z'$. □

5.2. Prediction of the active set. Following an idea in [CR] for the approximation of scalar nonconvex variational problems, we can further remove nodes temporarily from a mesh with nodes $\mathcal{N}$; e.g., $\mathcal{N} \subseteq \mathcal{N}_{d,r}$ is a refinement of $\mathcal{N}_{2d,r}$, using an iterative method that we establish in the following lemma. The method consists of defining an appropriate subset $X \subseteq \mathcal{N}$ and seeking for a solution of a lower-dimensional subproblem. For a discrete set $\mathcal{N} \subseteq \mathbb{R}^{n\times m}$ we define

$$f^{pc}_{\mathcal{N}}(F) := \min\left\{\sum_{A\in\mathcal{N}} \theta_A f(A) : \forall A \in \mathcal{N},\, \theta_A \geq 0,\, \sum_{A\in\mathcal{N}} \theta_A = 1,\, \sum_{A\in\mathcal{N}} \theta_A T(A) = T(F)\right\}.$$

Optimality conditions for $f^{pc}_{\mathcal{N}}(F)$ guarantee the existence of some $\lambda^F_{\mathcal{N}} \in \mathbb{R}^\tau$ such that

$$\max_{A\in\mathcal{N}} \big(\lambda^F_{\mathcal{N}} \cdot T(A) - f(A)\big) \leq \lambda^F_{\mathcal{N}} \cdot T(F) - f^{pc}_{\mathcal{N}}(F). \tag{5.5}$$

Conversely, any $(\theta_A : A \in \mathcal{N})$ that is feasible in $f^{pc}_{\mathcal{N}}(F)$ is optimal if there exists $\lambda^F_{\mathcal{N}} \in \mathbb{R}^\tau$ such that (5.5) holds with $f^{pc}_{\mathcal{N}}(F)$ replaced by $\sum_{A\in\mathcal{N}} \theta_A f(A)$.

Given $X \subseteq \mathcal{N}$, we consider the following lower-dimensional subproblem of $f^{pc}_{\mathcal{N}}(F)$:

$$f^{pc}_{\mathcal{N},X}(F) := \min\left\{\sum_{A\in X} \theta_A f(A) : \forall A \in X,\, \theta_A \geq 0,\, \sum_{A\in X} \theta_A = 1,\, \sum_{A\in X} \theta_A T(A) = T(F)\right\}.$$

The next lemma states sufficient conditions on X such that $f^{pc}_{\mathcal{N},X}(F) = f^{pc}_{\mathcal{N}}(F)$ and directly leads to an iterative algorithm.

LEMMA 5.1. *Let $(\theta_A : A \in \mathcal{N})$ be feasible and optimal for $f^{pc}_{\mathcal{N}}(F)$ with multiplier $\lambda^F_{\mathcal{N}} \in \mathbb{R}^\tau$. Assume that $\varepsilon_{AS} > 0$ and $\tilde\lambda^F \in \mathbb{R}^\tau$ satisfy $\sup_{A\in\mathcal{N}} |(\tilde\lambda^F - \lambda^F_{\mathcal{N}}) \cdot T(A)| \leq \varepsilon_{AS}/2$. If*

$$X = \left\{A \in \mathcal{N} : \tilde\lambda^F \cdot T(A) - f(A) \geq \max_{A'\in\mathcal{N}} \big(\tilde\lambda^F \cdot T(A') - f(A')\big) - \varepsilon_{AS}\right\}$$

and if the optimization problem $f^{pc}_{\mathcal{N},X}(F)$ is feasible, then $f^{pc}_{\mathcal{N},X}(F) = f^{pc}_{\mathcal{N}}(F)$.

Proof. The optimality conditions (5.5) show (cf. (5.4) in the proof of Proposition 5.2), for all $A \in \mathcal{N}$,

$$\theta_A > 0 \implies A \in Y := \{A' \in \mathcal{N} : \lambda^F_{\mathcal{N}} \cdot T(A') - f(A') = \lambda^F_{\mathcal{N}} \cdot T(F) - f^{pc}_{\mathcal{N}}(F)\}.$$

Hence it suffices to show that $Y \subseteq X$. Let $A \in Y$. By assumption on ε_{AS}, the definitions of X and Y, and (5.5), there holds

$$\begin{aligned}
\tilde\lambda^F \cdot T(A) - f(A) &\geq \lambda^F_{\mathcal{N}} \cdot T(A) - f(A) - \frac{\varepsilon_{AS}}{2} \\
&= \lambda^F_{\mathcal{N}} \cdot T(F) - f^{pc}_{\mathcal{N}}(F) - \frac{\varepsilon_{AS}}{2} \\
&= \max_{A'\in\mathcal{N}} \big(\lambda^F_{\mathcal{N}} \cdot T(A') - f(A')\big) - \frac{\varepsilon_{AS}}{2} \\
&\geq \max_{A'\in\mathcal{N}} \big(\tilde\lambda^F \cdot T(A') - f(A')\big) - \varepsilon_{AS},
\end{aligned}$$

i.e., $A \in X$. □

Given some $\tilde\lambda^F$, we do in general not know ε_{AS} to define X as in the lemma. We may, however, enlarge ε_{AS} successively until the optimality conditions (5.5) are satisfied. Having computed a solution for some parameter d, we may then use the corresponding multiplier to define X on a finer mesh.

5.3. An iterative, adaptive algorithm. The ideas of the preceding propositions lead to the following algorithm that iteratively computes, for prescribed $d_0 > 0$, $r_0 > 0$, $M > 0$, and $J > 0$, a solution to $f^{pc}_{d_0/2^J,2^\ell r_0}(F)$, where for some $\ell \geq 0$, $r = 2^\ell r_0$ satisfies the conditions of Theorem A for $d = d_0/2^J$ if $p > n \wedge m$. If $p = n \wedge m$, we suppose that r_0 is large enough, i.e., $r_0 \geq r'$ for r' as in Theorem A. We assume that $F \in \omega_{d_0,r_0}$.

Algorithm $(A^{pc,\,adapt}_{r_0,d_0,J,F,M})$.

(a) Set $j := 0$, $d := d_0$, $r := r_0$, $\tilde{\lambda}^F := 0$, $\mathcal{N} := \mathcal{N}_{d,r}$, and $\varepsilon_{AS} := \infty$.

(b) Define

$$X := \Big\{A \in \mathcal{N} : \tilde{\lambda}^F \cdot T(A) - f(A) \geq \max_{A' \in \mathcal{N}} \big(\tilde{\lambda}^F \cdot T(A') - f(A')\big) - \varepsilon_{AS}\Big\} \\ \cup \big\{A \in \mathcal{N}_{d,r} : |F - A| \leq d\big\}.$$

(c) Compute $f^{pc}_{\mathcal{N},X}(F)$ and obtain a multiplier $\lambda^F_{\mathcal{N}} \in \mathbb{R}^\tau$.

(d) If for all $A \in \mathcal{N}$ there holds

$$\lambda^F_{\mathcal{N}} \cdot T(A) - f(A) \leq \lambda^F_{\mathcal{N}} \cdot T(F) - f^{pc}_{\mathcal{N}}(F),$$

go to (f).

(e) Set $\tilde{\lambda}^F := \lambda^F_{\mathcal{N}}$, $\varepsilon_{AS} := 2\varepsilon_{AS}$, and go to (b).

(f) If $p = n \wedge m$ or

$$c_T(n \wedge m)|\lambda^F_{\mathcal{N}}| \leq p c_f r^{p - n\wedge m} \quad \text{and} \\ c_T|\lambda^F_{\mathcal{N}}| r^{n \wedge m} - c_f r^p + c'_f \leq \lambda^F_{\mathcal{N}} \cdot T(F) - f^{pc}_{\mathcal{N}}(F),$$

go to (h).

(g) Set $r := 2r$, $\tilde{\lambda}^F := \lambda^F_{\mathcal{N}}$, and go to (b).

(h) If $j < J$, define

$$\mathcal{N} = \big\{A' \in \mathcal{N}_{d/2,r} : \exists A \in \mathcal{N},\ |A - A'| \leq d, \\ \lambda^F_{\mathcal{N}} \cdot T(A) - f(A) > \lambda^F_{\mathcal{N}} \cdot T(F) - f^{pc}_{\mathcal{N}}(F) - Md\big\},$$

set $\tilde{\lambda}^F := \lambda^F_{\mathcal{N}}$, $\varepsilon_{AS} := d$, $d := d/2$, $j := j + 1$, and go to (b).

(j) Stop.

Remark 5.1. (i) We set $\varepsilon_{AS} = d$ in step (h), since in the optimal case (if $\hat{f} \in C^{1,1}_{loc}(\mathbb{R}^\tau)$) Proposition 4.3 guarantees $|\lambda^F_{d,r} - \lambda^F_{d/2,r}| \leq \mathcal{O}(d)$ so that the conditions of Lemma 5.1 are satisfied up to some constant of order $\mathcal{O}(1)$.

(ii) Note that $\mathcal{N}_{d/2,r}$ need not be computed explicitly, since we can add nodes to and remove nodes from $\mathcal{N}$ locally.

(iii) Adding $\big\{A \in \mathcal{N}_{d,r} : |F - A| \leq d\big\}$ to X in step (b) guarantees the feasibility of $f^{pc}_{\mathcal{N},X}(F)$.

6. Numerical experiments I. In this section we report on the practical performance of Algorithm $(A^{pc,\,adapt}_{r_0,d_0,J,F,M})$ when applied to three choices of f for which explicit formulae for f^{pc} and f^{rc} are known.

Example 6.1 (see [Da2]). For $n = m = 2$ and $F \in \mathbb{R}^{2\times 2}$ let

$$f(F) := (|F|^2 - 1)^2.$$

Then (2.1) holds for $p = 4$ and $c_f = (c-2)/c$, $c'_f = 2c-1$ for all $c > 2$, and we choose $c = 3$. In this example $f^{pc} = f^{qc} = f^{**}$, where f^{**} is the convex envelope of f and for $F \in \mathbb{R}^{2\times 2}$ given by

$$f^{**}(F) = \begin{cases} (|F|^2-1)^2 & \text{for } |F| \geq 1, \\ 0 & \text{for } |F| \leq 1. \end{cases}$$

Example 6.2 (see [Ko, DW]). For $n = m = 2$,

$$A_1 := \begin{pmatrix} 5/4 & 0 \\ 0 & 3/4 \end{pmatrix} \quad \text{and} \quad A_2 := \begin{pmatrix} 3\sqrt{8}/8 & 3/8 \\ -5/8 & 5\sqrt{3}/8 \end{pmatrix},$$

and $F \in \mathbb{R}^{2\times 2}$ let

$$f(F) := \frac{1}{2}\min\bigl\{|F-A_1|^2, |F-A_2|^2\bigr\}.$$

Then (2.1) holds for $p = 2$, $c_f = 1/2^3$, and $c'_f = \max\{|A_1|^2, |A_2|^2\}/2 = 17/16$. Here, $f^{**} \neq f^{pc} = f^{qc}$ and f^{pc} is for $F \in \mathbb{R}^{2\times 2}$ given by

$$f^{pc}(F) = \begin{cases} f_1(F) & \text{for } f_1(F) - f_2(F) \leq -\lambda/2, \\ f_2(F) - \bigl(f_2(F) - f_1(F) + \lambda/2\bigr)/(2\lambda) & \text{for } |f_1(F) - f_2(F)| \leq \lambda/2, \\ f_2(F), & \text{for } f_1(F) - f_2(F) \geq \lambda/2, \end{cases}$$

where $f_j(F) = |F-A_j|^2/2$, $j = 1, 2$, and $\lambda = |A_1 - A_2|$.

Example 6.3 (see [KS, Do]). For $n = m = 2$ and $F \in \mathbb{R}^{2\times 2}$ a modification proposed in [Do] (to ensure continuity of f) of an energy density occurring in an optimal design problem in [KS] reads

$$f(F) := \begin{cases} 1 + |F|^2 & \text{for } |F| \geq \sqrt{2}-1, \\ 2\sqrt{2}|F| & \text{for } |F| \leq \sqrt{2}-1. \end{cases}$$

Then (2.1) holds for $p = 2$, $c_f = 1$, and $c'_f = 0$. Letting $\varrho(F) := \sqrt{|F|^2 + 2|\det F|}$ for $F \in \mathbb{R}^{2\times 2}$, there holds

$$f^{pc}(F) = f^{qc}(F) = \begin{cases} 1 + |F|^2 & \text{for } \varrho(F) \geq 1, \\ 2(\varrho(F) - |\det F|) & \text{for } \varrho(F) \leq 1. \end{cases}$$

Note that $f^{**} \neq f^{pc}$ in this example.

We tested Algorithm ($A^{pc,\,adapt}_{r_0,d_0,J,F,M}$) in Examples 6.1–6.3. The implementation of the algorithm was performed in Matlab with a generation of the adaptively refined grids in C. The experiments were performed on a node of a Compaq SC-Cluster with four Alpha-EV68 processors (1 GHz, 8 MB Cache/CPU) and 32 GB RAM.

We set $r_0 = 1$, $d_0 = 1$, $M = 1$, $J = 5$, and

$$F = \frac{1}{5}\begin{pmatrix} \pi & 1 \\ -1 & \pi \end{pmatrix}$$

to run Algorithm ($A^{pc,\,adapt}_{r_0,d_0,J,F,M}$) in Example 6.1. The a posteriori criterion of Proposition 4.2 enforced the algorithm to enlarge r_0 from 1 to 4 on the first level, i.e., for the largest d. Table 1 presents the errors

$$e = |f^{pc}_{d,r}(F) - f^{pc}(F)| \quad \text{and} \quad e' = |\lambda^F_{d,r} \cdot DT(F) - Df^{pc}(F)|,$$

Table 1

Discretization parameter d; errors $e = |f^{pc}_{d,r}(F) - f^{pc}(F)|$ and $e' = |\lambda^F_{d,r} \cdot DT(F) - Df^{pc}(F)|$; number of active, possible, and theoretical nodes; and CPU-time needed to compute $f^{pc}_{d,r}(F)$ in Example 6.1.

d	e	e'	$\#X$	$\#\mathcal{N}$	$\#\mathcal{N}_{d,r}$	CPU-time
1	0.656 637	2.000 000	266	6,561	6,561	0.1 s
1/2	0.037 543	0.325 927	132	160	83,521	0.2 s
1/4	0.009 873	0.206 980	237	768	1,185,921	0.2 s
1/8	0.000 177	0.001 392	2,137	3,920	17,850,625	0.7 s
1/16	0.000 010	0.000 058	14,360	33,920	276,922,881	3.7 s

Table 2

Discretization parameter d; errors $e = |f^{pc}_{d,r}(F) - f^{pc}(F)|$ and $e' = |\lambda^F_{d,r} \cdot DT(F) - Df^{pc}(F)|$; number of active, possible, and theoretical nodes; and CPU-time needed to compute $f^{pc}_{d,r}(F)$ in Example 6.2.

d	e	e'	$\#X$	$\#\mathcal{N}$	$\#\mathcal{N}_{d,r}$	CPU-time
1	0.165 961	0.454 711	6,561	6,561	6,561	0.9 s
1/2	0.072 351	0.115 344	794	43,568	83,521	1.1 s
1/4	0.014 273	0.042 726	1,311	167,136	1,185,921	1.4 s
1/8	0.004 031	0.063 871	1,766	715,504	17,850,625	2.4 s
1/16	0.001 139	0.017 684	4,889	3,082,192	276,922,881	6.5 s
1/32	0.000 204	0.011 032	12,140	13,694,256	4,362,470,401	21.8 s

the number of nodes in the set $\mathcal{N}$, the number of activated nodes in X, the theoretical number of nodes (i.e., the number of nodes in $\mathcal{N}_{d,r}$), and the CPU-time in seconds needed to compute $f^{pc}_{\mathcal{N}}(F)$.

We observe that e converges to 0 with experimental rate 4, and the experimental convergence rate for e' is better than linear. Due to the grid coarsening strategy and the active set strategy, the number of activated nodes in X, i.e., the size of each linear optimization problem, is remarkably small when compared to the possible and theoretical numbers of nodes, and the CPU-time needed to obtain an absolute error of about 10^{-5} is only 3.7 seconds. We obtained similar numbers e and e' for the more reliable choice $M = 100$, but the number of activated nodes and the CPU-time was significantly larger; e.g., 9.4 seconds were needed to achieve $e \leq 10^{-3}$.

To test Algorithm $(A^{pc,\,adapt}_{r_0,d_0,J,F,M})$ in Example 6.2 we set $J = 6$, $r_0 = 4$, $d_0 = 1$, $M = 10$, and

$$F = \frac{1}{10}\begin{pmatrix} \sqrt{2}/3 & 1/3 \\ \sqrt{5}/5 & \sqrt{2}/3 \end{pmatrix}.$$

Table 2 presents the errors and the numbers of nodes as in the previous example. The error e converges with an experimental convergence rate 1.8, i.e, $e \approx d^{1.8}$, and there seems to be approximately linear decay in e', but convergence cannot be deduced. The choice $M = 10$ is very optimistic in this example and leads to very small sets $\mathcal{N}$.

We ran Algorithm $(A^{pc,\,adapt}_{r_0,d_0,J,F,M})$ in Example 6.3 with $M = 100$, $J = 6$, $r_0 = 4$, $d_0 = 1$, and

$$F = \frac{1}{5}\begin{pmatrix} \pi & 0 \\ 0 & \pi \end{pmatrix}.$$

TABLE 3
Discretization parameter d; errors $e = |f^{pc}_{d,r}(F) - f^{pc}(F)|$ *and* $e' = |\lambda^F_{d,r} \cdot DT(F) - Df^{pc}(F)|$*; number of active, possible, and theoretical nodes; and CPU-time needed to compute* $f^{pc}_{d,r}(F)$ *in Example* 6.3.

d	e	e'	$\#X$	$\#\mathcal{N}$	$\#\mathcal{N}_{d,r}$	CPU-time
1	0.328 922	0.162 697	81	81	81	0.1 s
1/2	0.095 387	0.458 318	457	625	625	0.2 s
1/4	0.024 647	0.160 972	689	6,561	6,561	0.3 s
1/8	0.000 806	0.167 216	1,276	83,521	83,521	0.6 s
1/16	0.000 389	0.079 923	1,618	513,280	1,185,921	1.2 s
1/32	0.000 180	0.032 843	2,213	2,128,048	17,850,625	3.6 s

Table 3 presents the errors e and e'; the numbers of activated, possible, and theoretical nodes; and the CPU time as in the previous examples. We observe that e converges at least linearly to 0, while e' seems to converge linearly, at least for $d \leq 1/8$.

7. Numerical experiments II.

In this section we outline how our Algorithm $(A^{pc,\,adapt}_{r_0,d_0,J,F,M})$ may be used for the effective numerical simulation of nonconvex vectorial variational problems, and we report on two numerical experiments. The proposed algorithm aims to numerically relax and minimize variational problems of the form (M′), i.e.,

$$(\mathrm{M}') \quad \text{Minimize} \quad I(u) := \int_\Omega f(\nabla u)\,dx$$
$$\text{among } u \in \mathcal{A} := \{v \in W^{1,p}(\Omega;\mathbb{R}^m) : v|_{\Gamma_D} = u_D\},$$

where f is continuous and satisfies p-growth conditions, $\Omega \subseteq \mathbb{R}^n$ is a bounded Lipschitz domain, $\Gamma_D \subseteq \partial\Omega$ is closed and of positive surface measure, and $u_D = \tilde{u}|_{\Gamma_D}$ for some $\tilde{u}_D \in C(\overline{\Omega};\mathbb{R}^m)$. We further suppose that Ω is polyhedral and let $\mathcal{T}$ be a regular triangulation of Ω such that Γ_D is matched exactly by edges (respectively, faces) of elements in $\mathcal{T}$. We let $\mathcal{S}^1(\mathcal{T})^m$ denote the lowest order finite element space on $\mathcal{T}$ which consists of all globally continuous $\mathcal{T}$-elementwise affine functions in $W^{1,p}(\Omega;\mathbb{R}^m)$. Finally, we let $\tilde{u}_{D,h}$ be the nodal interpolant of $\tilde{u}_D$ on $\mathcal{T}$. The following algorithm is capable of finding an approximation of a weak limit of an infimizing sequence for the nonconvex vectorial variational problem (M′). The approximation scheme realizes a steepest descent approach and exploits the fact that Algorithm $(A^{pc,\,adapt}_{r_0,d_0,J,F,M})$ provides an approximation of $Df^{pc}(F)$.

ALGORITHM (A^{nvvp}). Input: $u_h^{(0)} \in \mathcal{S}^1(\mathcal{T})^m$ with $u_h^{(0)}|_{\Gamma_D} = \tilde{u}_{D,h}|_{\Gamma_D}$, parameters r_0, d_0, J, M for Algorithm $(A^{pc,\,adapt}_{r_0,d_0,J,\cdot,M})$, and a termination criterion $\delta > 0$.

(a) Set $j := 0$.

(b) Let $r_h \in \mathcal{S}^1(\mathcal{T})^m$ satisfy $r_h|_{\Gamma_D} = 0$ and

$$\int_\Omega \nabla r_h \cdot \nabla v_h\,dx = -\int_\Omega \big(\lambda^{\nabla u_h^{(j)}}_{d/4,r} \cdot DT(\nabla u_h^{(j)})\big) \cdot \nabla v_h\,dx$$

for all $v_h \in \mathcal{S}^1(\mathcal{T})^m$ with $v_h|_{\Gamma_D} = 0$.

(c) Compute $t^* \in [0,1]$, which is a local minimizer in $[0,1]$ for

$$t \mapsto \int_\Omega f^{pc}_{d,r}\big(\nabla(u_h^{(j)} + tr_h)\big)\,dx.$$

TABLE 4
Minimal energies for various approaches to the numerical solution of (M′) *in Example* 7.1.

h	f	f^{pc}	$f^{pc}_{d,r}$	f^{rc}_d
1/8	0.244 128	0.199 761	0.252 177 (0.313 327)	0.202 507
1/16	0.234 894	0.199 761	0.216 076 (0.274 192)	0.208 264
1/32	0.233 769	0.199 761	0.202 994 (0.261 244)	0.245 633

(d) Stop and set $u_h := u_h^{(j)}$ if $t^* < \delta$.

(e) Set $u_h^{(j+1)} := u_h^{(j)} + t^* r_h$, $j := j+1$, and go to (b).

Output: $u_h \in \mathcal{S}^1(\mathcal{T})^m$.

We specify f, Ω, $\tilde{u}_D$, and $\mathcal{T}$ in two examples. The first example involves affine boundary data on $\partial\Omega$.

Example 7.1 (see [DW]). Set $n = m := 2$, $\Omega := (0,1)^2$, $\tilde{u}_D(x) := Bx + c$ for

$$B = \begin{pmatrix} 1/2 & 1/4 \\ -1/4 & 15/32 \end{pmatrix}$$

and $c := (0, 1/4)$, and let f be as in Example 6.2. Given an integer $k > 0$, let $h_k := 1/k$ and $\mathcal{T}_k$ be the triangulation of Ω that consists of $2k^2$ triangles which are halved squares of side length h and with diagonals parallel to $(1,1)$.

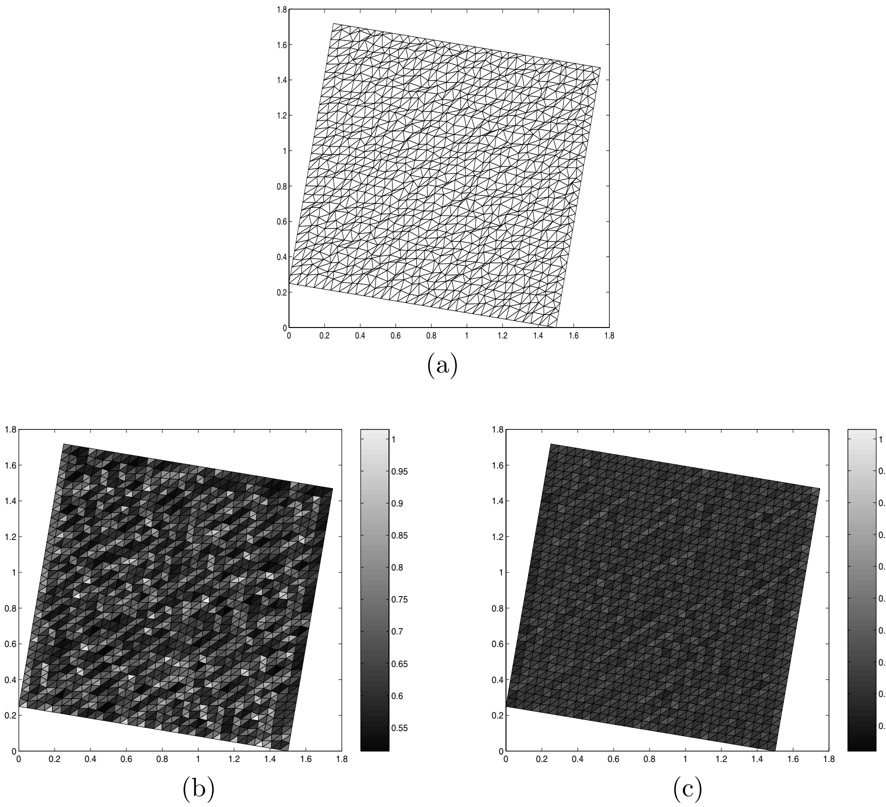
For $\ell = 1, 2, 3$ we ran Algorithm (A^{nvvp}) with $r_0 = 2$, $M = 10$, $J_\ell = 2 + \ell$, and $k_\ell = 2^{2+\ell}$ in Example 7.1. The initial $u_h^{(0)}$ was chosen as $u_h^{(0)} = \tilde{u}_{D,h} + \xi_h$, where $\xi_h \in \mathcal{S}^1(\mathcal{T})^2$ with $\xi_h|_{\Gamma_D} = 0$ is obtained from a linear interpolation of random values $\xi_h(z) \in [-h_\ell, h_\ell]^2$ in the free nodes z of $\mathcal{T}$. The termination criterion δ was set to $\delta = 0.03$, and Algorithm (A^{nvvp}) terminated after 175, 86, and 8 iterations for $\ell = 1$, 2, and 3, respectively. Table 4 displays the numerically relaxed energies for the output u_{h_ℓ} in Example 7.1 and compares them to values that we obtained when $f^{pc}_{d,r}(F)$ and $\lambda^F_{d,r} \cdot DT(F)$ are replaced by f and Df, f^{pc} and Df^{pc}, and a numerical approximation f^{rc}_d of f^{rc} in steps (b) and (c) of Algorithm (A^{nvvp}). (The numbers for f^{rc}_d are taken from [DW].) We observe that Algorithm (A^{nvvp}) significantly reduces the initial (numerically relaxed) energies shown in brackets in the fourth column of Table 4. Moreover, we may deduce from the numerical results that the (numerically relaxed) energies of the outputs of Algorithm (A^{nvvp}) converge to the optimal value $f^{pc}(B) = 0.199761$ for $(d, h) \to 0$. No experimental convergence can be deduced when the original function f or the numerically obtained approximation of the rank-1 convex envelope f^{rc}_d were employed.

Figure 2 displays the initial $u_h^{(0)}$, the numerical solution obtained from direct numerical minimization of (M′), and the numerical solution obtained from the numerical relaxation realized by Algorithm (A^{nvvp}) in Example 7.1. We observe mesh-dependent oscillations in the stress field defined by the numerical solution when no relaxation is used, while we observe a rather smooth stress field when the nonconvex vectorial variational problem is numerically polyconvexified.

The second example incorporates nonaffine boundary conditions on Γ_D.

Example 7.2 (see [DW]). Let n, m, Ω, Γ_D, f, and $\mathcal{T}$ be as in Example 7.1, and define for $x \in \overline{\Omega}$

$$\tilde{u}_D(x) := \frac{x - (1/2, 1/2)}{\sqrt{\left|x - (1/2,1/2)\right|^2 + 1/4}}.$$



(a)

(b)

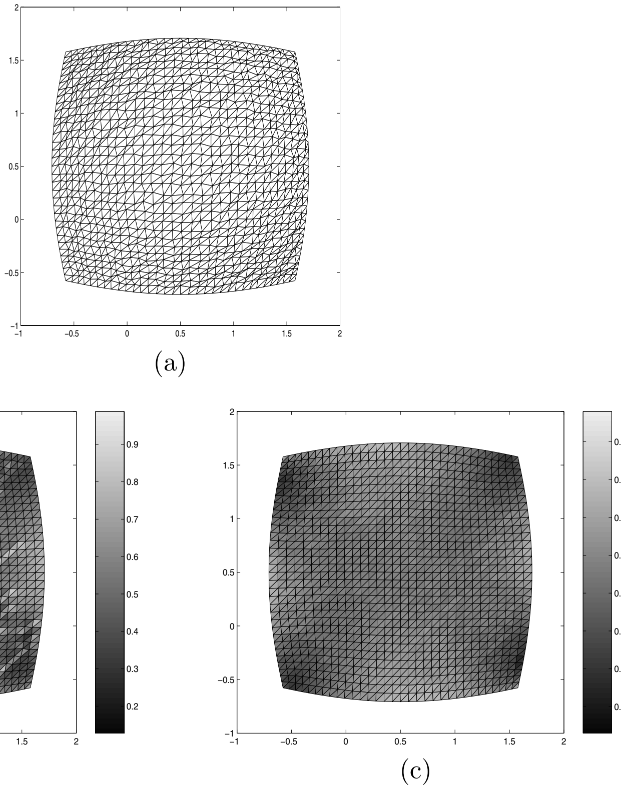
(c)

FIG. 2. (a) *Initial deformation* $u_h^{(0)}$, (b) *numerical solution for direct minimization employing* f *with modulus of the related stress field* $|Df(\nabla u_h)|$, *and* (c) *numerical solution* u_h *obtained from numerical relaxation with Algorithm* (A^{nvvp}) *with stress field* $|\lambda_{d/4,r}^{\nabla u_h} \cdot DT(\nabla u_h)|$ *in Example* 7.1.

TABLE 5
Minimal energies for various approaches to the numerical solution of (M′) *in Example* 7.2.

h	f	f^{pc}	$f_{d,r}^{pc}$	f_d^{rc}
1/8	0.186 939	0.170 023	0.231 421 (0.244 203)	0.175 216
1/16	0.189 875	0.168 945	0.186 094 (0.198 793)	0.173 716
1/32	0.188 556	0.168 583	0.173 779 (0.186 396)	0.183 599

Table 5 displays the minimal energies for various approaches to the numerical simulation of (M′) in Example 7.2. As in the previous example, we observe that the results obtained by Algorithm (A^{nvvp}) (with the same parameters as in the previous experiment) with the approximated polyconvex envelope of f approach the value that we obtained with the exact polyconvex envelope f^{pc}. The minimal energies obtained with the nonrelaxed functional and with the discrete approximation of the rank-1 convex envelope of f (numbers for f_d^{rc} are taken from [DW]) do not show a convergent behavior. As in the previous experiment, we observe in Figure 3 mesh-dependent oscillations in the numerical solution obtained from a direct minimization scheme. No significant oscillations can be found in the numerical solution computed with Algorithm (A^{nvvp}) and displayed in Figure 3(c).



(a)

(b)

(c)

FIG. 3. (a) *Initial deformation* $u_h^{(0)}$, (b) *numerical solution for direct minimization employing* f *with modulus of the related stress field* $|Df(\nabla u_h)|$, (c) *and numerical solution* u_h *obtained from numerical relaxation with Algorithm* (A^{nvvp}) *with stress field* $|\lambda_{d/4,r}^{\nabla u_h} \cdot DT(\nabla u_h)|$ *in Example* 7.2.

Remark 7.1. (i) A good initial $u_h^{(0)}$ may be obtained from solving (M′) with f replaced by a convex function, e.g., $F \mapsto |F|^2/2$, as a preprocessing step in Algorithm (A^{nvvp}).

(ii) A postprocessing procedure based on the algorithms in [Do, DW, Ba2] may be included in Algorithm (A^{nvvp}), which approximates the rank-1 convex envelope applied to ∇u_h almost everywhere in Ω. Then, if (up to numerical tolerances) there holds $f^{pc}(\nabla u_h) = f^{rc}(\nabla u_h)$ almost everywhere in Ω, one has that $f^{qc}(\nabla u_h) = f^{pc}(\nabla u_h)$ and (provided that $f^{pc} \le f^{qc} \le f^{rc}$ are smooth enough) that $Df^{qc}(\nabla u_h) = Df^{pc}(\nabla u_h)$ almost everywhere in Ω. In this case, u_h serves as an approximation of a stationary point of the quasi-convex relaxation of (M).

(iii) The computation of $f_{d,r}^{pc}(\nabla u_h^{(j)})$ and $\lambda_{d,r}^{\nabla u_h^{(j)}}$ in steps (b) and (c) of Algorithm (A^{nvvp}) has to be done on each element of the triangulation $\mathcal{T}$ (since ∇u_h is $\mathcal{T}$-elementwise constant). This may be time-consuming but can be parallelized without communication costs.

REFERENCES

[B1] J. M. BALL, *Convexity conditions and existence theorems in nonlinear elasticity*, Arch. Ration. Mech. Anal., 63 (1977), pp. 337–403.

[B2] J. M. BALL, *A version of the fundamental theorem for Young measures*, in Partial Differential Equations and Continuum Models of Phase Transitions, M. Rascle,

D. Serre, and M. Slemrod, eds., Lecture Notes in Phys. 344, Springer, Berlin, 1989, pp. 207–215.

[BJ] J. M. Ball and R. D. James, *Fine phase mixtures as minimizers of energy*, Arch. Rational Mech. Anal., 100 (1987), pp. 13–52.

[BKK] J. M. Ball, B. Kirchheim, and J. Kristensen, *Regularity of quasiconvex envelopes*, Calc. Var. Partial Differential Equations, 11 (2000), pp. 333–359.

[Ba1] S. Bartels, *Adaptive approximation of Young measure solutions in scalar non-convex variational problems*, SIAM J. Numer. Anal., 42 (2004), pp. 505–529.

[Ba2] S. Bartels, *Linear convergence in the approximation of rank-one convex envelopes*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 811–820.

[BP] S. Bartels and A. Prohl, *Multiscale resolution in the computation of crystalline microstructure*, Numer. Math., 96 (2004), pp. 641–660.

[CP1] C. Carstensen and P. Plecháč, *Numerical solution of the scalar double-well problem allowing microstructure*, Math. Comp., 66 (1997), pp. 997–1026.

[CP2] C. Carstensen and P. Plecháč, *Numerical analysis of compatible phase transitions in elastic solids*, SIAM J. Numer. Anal., 37 (2000), pp. 2061–2081.

[CR] C. Carstensen and T. Roubiček, *Numerical approximation of Young measures in non-convex variational problems*, Numer. Math., 84 (2000), pp. 395–414.

[CM] M. Chipot and S. Müller, *Sharp energy estimates for finite element approximations of non-convex problems*, in Variations of Domain and Free-Boundary Problems in Solid Mechanics, Solid Mech. Appl. 66, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1999, pp. 317–325.

[C] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics in Appl. Math. 5, SIAM, Philadelphia, 1990.

[Da1] B. Dacorogna, *A characterization of polyconvex, quasiconvex and rank one convex envelopes*, in Integral Functionals in Calculus of Variations, Proceedings of the International Workshop (Trieste, Italy 1985), Suppl. Rend. Circ. Mat. Palermo, II. Ser. 15, 1987, pp. 37–58.

[Da2] B. Dacorogna, *Direct methods in the calculus of variations*, Applied Math. Sci. 78, Springer-Verlag, Heidelberg, Germany, 1989.

[DH] B. Dacorogna and J.-P. Haeberly, *Some numerical methods for the study of the convexity notions arising in the calculus of variations*, M2AN Math. Model. Numer. Anal., 32 (1998), pp. 153–175.

[Do] G. Dolzmann, *Numerical computation of rank-one convex envelopes*, SIAM J. Numer. Anal., 36 (1999), pp. 1621–1635.

[DW] G. Dolzmann and N. J. Walkington, *Estimates for numerical approximations of rank one convex envelopes*, Numer. Math., 85 (2000), pp. 647–663.

[HH] K. Hackl and U. Hoppe, *On the calculation of microstructures for inelastic materials using relaxed energies*, Solid Mech. Appl., 108 (2003), pp. 77–86.

[Ko] R. V. Kohn, *The relaxation of a double-well energy*, Contin. Mech. Thermodyn., 3 (1991), pp. 193–236.

[KS] R. V. Kohn and G. Strang, *Optimal design and relaxation of variational problems.* I–III, Comm. Pure Appl. Math., 39 (1986), pp. 353–377.

[Kr] M. Kružík, *Numerical approach to double well problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1833–1849.

[L] M. Luskin, *On the computation of crystalline microstructure*, Acta Numer., 5 (1996), pp. 191–257.

[ML] C. Miehe and M. Lambrecht, *Analysis of microstructure development in shearbands by energy relaxation of incremental stress potentials: Large-strain theory for standard dissipative solids*, Internat. J. Numer. Methods Engrg., 58 (2003), pp. 1–41.

[M] S. Müller, *Variational models for microstructure and phase transitions*, Lecture Notes in Math. 1713, Springer, New York, 1999, pp. 85–210.

[NW1] R. A. Nicolaides and N. J. Walkington, *Computation of microstructure utilizing Young measure representations*, ARO Rep. 93-1, in Transactions of the Tenth Army Conference on Applied Mathematics and Computing, West Point, NY, 1992, U.S. Army Res. Office, 1993, pp. 57–68.

[NW2] R. A. Nicolaides and N. J. Walkington, *Strong convergence of numerical solutions to degenerate variational problems*, Math. Comp., 64 (1995), pp. 117–127.

[R1] T. Roubíček, *Numerical approximation of relaxed variational problems*, J. Convex Anal., 3 (1996), pp. 329–347.

[R2] T. Roubíček, *Relaxation in Optimization Theory and Variational Calculus*, De Gruyter Series in Nonlinear Anal. Appl. 4, De Gruyter, New York, 1997.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 386–408

ANALYSIS OF FIRST-ORDER SYSTEM LEAST SQUARES (FOSLS) FOR ELLIPTIC PROBLEMS WITH DISCONTINUOUS COEFFICIENTS: PART I[*]

MARKUS BERNDT[†], THOMAS A. MANTEUFFEL[‡], STEPHEN F. MCCORMICK[‡], AND GERHARD STARKE[§]

Abstract. First-order system least squares (FOSLS) is a recently developed methodology for solving partial differential equations. Among its advantages are that the finite element spaces are not restricted by the inf-sup condition imposed, for example, on mixed methods and that the least-squares functional itself serves as an appropriate error measure. This paper studies the FOSLS approach for scalar second-order elliptic boundary value problems with discontinuous coefficients, irregular boundaries, and mixed boundary conditions. A least-squares functional is defined, and ellipticity is established in a natural norm of an appropriately scaled least-squares bilinear form. For some geometries, this ellipticity is independent of the size of the jumps in the coefficients. The occurrence of singularities at interface corners, cross points, reentrant corners, and irregular boundary points is discussed, and a basis of singular functions with local support around singular points is established. A companion paper shows that the singular basis functions can be added at little extra cost and lead to optimal performance of standard finite element discretization and multilevel solver techniques, also independent of the size of coefficient jumps for some geometries.

Key words. least-squares discretization, second-order elliptic problems, finite elements, multilevel methods

AMS subject classifications. 65N55, 65N30, 65F10

DOI. 10.1137/S0036142903427688

1. Introduction.

The purpose of this paper is to apply first-order system least squares (FOSLS; cf. [11] and [12]) to scalar second-order elliptic boundary value problems in two dimensions with discontinuous coefficients, irregular boundaries, and mixed boundary conditions. Such problems arise in various applications, including flow in heterogeneous porous media [29], neutron transport [1], and biophysics [20]. In many physical applications, one is interested not only in accurate approximation of the physical quantity that satisfies the scalar equation, but also in certain of its derivatives. For example, fluid flow in a porous medium can be modeled by the equation

$$-\nabla \cdot (a\nabla p) = f \tag{1.1}$$

[*] Received by the editors May 7, 2003; accepted for publication (in revised form) August 31, 2004; published electronically June 14, 2005. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sinum/43-1/42768.html

[†] Los Alamos National Laboratory, T-7, Mail Stop B284, Los Alamos, NM 87545 (berndt@lanl.gov). The research of this author was supported by the Department of Energy, under contract W-7405-ENG-36, LA-UR-01-2711.

[‡] Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309-0526 (tmanteuf@boulder.colorado.edu, stevem@boulder.colorado.edu). The research of these authors was supported by the National Science Foundation under grant number DMS-8704169 and by the Department of Energy, Applied Math Program grant DE-FG03-94ER25217.

[§] Universität Hannover, Institut für Angewandte Mathematik, Welfengarten 1, 30167 Hannover, Germany (starke@ifam.uni-hannover.de).

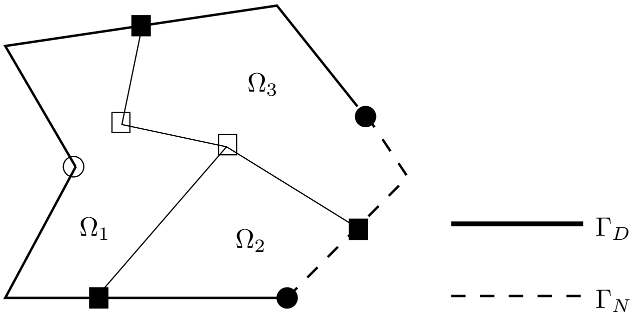


FIG. 1.1. *Polygonal domain Ω with subdomains Ω_i, $i = 1, 2, 3$, and two cross points.*

for the pressure p, where the scalar function a may have large jump discontinuities across interfaces. Of particular interest here is accurate approximation of the flux,

$$\mathbf{u} = a\nabla p. \tag{1.2}$$

For the purposes of discussion, consider problem (1.1) posed on a domain, Ω, composed of a union of polygonal subdomains, Ω_i, in which the coefficient a is constant on each subdomain (see Figure 1.1). In general, the flux, $\mathbf{u}$, will be infinite at certain points, which we will call singular points (see, for example, Strang and Fix [30, Chapter 8]). Singular points can be of several types:

Cross points: corner points of the boundary of Ω_i that lie in the interior of Ω (□ in Figure 1.1);

Boundary cross points: corner points of Ω_i on the boundary of Ω that touch another subdomain, Ω_j (■ in Figure 1.1);

Reentrant corners: reentrant corners of Ω (○ in Figure 1.1);

Irregular boundary points: points on the boundary of Ω that separate the Dirichlet boundary, Γ_D, from the Neumann boundary, Γ_N, for which the interior angle is greater than $\pi/2$ (● in Figure 1.1).

The solution, p, can be expressed as the sum of a finite number of singular functions plus a function that is locally smooth, that is, in $H^2(\Omega_i)$ for each i. Each singular function is associated with a singular point and, near the singular point, has the form $r^\alpha\Phi(\theta)$, where (r, θ) are polar coordinates about the singular point and $0 < \alpha < 1$. The character of a singular function depends only on local information near the singular point and is not difficult to compute (see section 5 and [3] for details).

There are many finite element methods for approximating the solution of (1.1). Some yield an approximate solution without specific knowledge of the singular functions, while others use the singular functions either implicitly or explicitly. Below we describe the major approaches.

Standard Galerkin method. The standard Galerkin method (cf. Strang and Fix [30]) establishes a weak form and seeks the approximation of p in $H^1(\Omega)$. Convergence deteriorates near the singular points. Early work using H^1 singular basis functions can be found in the monograph by Strang and Fix [30, section 8.2]. There, H^1 singular basis functions for p were introduced to eliminate the deteriorating finite element approximation near singular points. (See also Cox and Fix [16] and Grisvard [19, section 8.4.2].) A multilevel approach for simultaneously finding the approximate solution and determining the coefficients of the singular basis functions is developed by Brenner and Sung [9]. In [10], Cai and Kim describe a method that is equivalent to a Petrov/Galerkin method in which the singular basis functions are added to the trial space and the dual singular basis functions are added to the test space.

Mixed methods. In mixed finite element methods (see, e.g., [8, Chapter 10]), p and $\mathbf{u}$ are usually approximated by different finite element spaces, and, roughly speaking, a Galerkin condition is imposed on the first-order system resulting from (1.1) and (1.2). Normally, the pressure, p, is approximated in L^2 and the flux, $\mathbf{u}$, is approximated in $H(\mathrm{div})$. Only the integral of the flux is computed along edges of elements, and the pointwise resolution of singularities in the flux is poor.

The least squares methodology for systems of first order is by now several decades old and had its first application in continuum mechanics (see, for example, [21, 31, 22, 26, 15, 23]). Only fairly recently has it produced H^1 equivalent forms to which optimal multigrid solvers have been applied (see, for example, [12]). For a thorough review of the least-squares methodology, see [5] and the references therein. The following is an overview of specific least-squares methods and their applicability to the problem at hand.

Least-squares in $H(\mathrm{div})$. A similar approach is based on the FOSLS approach developed and analyzed, e.g., in [11, 12, 27, 28]. This methodology replaces the Galerkin condition by the minimization of a least-squares functional associated with a first-order system derived from (1.1) and (1.2). Assuming that $f \in L^2(\Omega)$, the least-squares functional can be defined using the $L^2(\Omega)$-norm. Even in the presence of discontinuities, this translates to ellipticity with respect to the H^1-norm for the pressure, p, and the $H(\mathrm{div})$-norm in the flux variable, $\mathbf{u}$. This approach, like the mixed method approach, computes only the integral flux and again does not resolve the singularity in the flux variable.

Weighted least-squares in $H(\mathrm{div})\cap H(\mathrm{curl})$. Augmenting the basic system with the curl-condition, $\nabla\times(\mathbf{u}/a) = 0$ (see [12, 27]), leads to ellipticity with respect to a scaled version of the $H(\mathrm{div}) \cap H(\mathrm{curl})$ norm in the flux variable. Standard finite element spaces, for example piecewise polynomials with the appropriate jump conditions across interfaces, are not dense in the scaled $H(\mathrm{div}) \cap H(\mathrm{curl})$ norm, and thus convergence cannot be obtained. However, the use of an appropriate weight function near each singular point yields ellipticity in a weighted (and scaled) $H(\mathrm{div})\cap H(\mathrm{curl})$ norm. The piecewise polynomial spaces are dense in this new space. The weighting effectively ignores the singularity while insulating the rest of the region from the presence of the singularity. For the case of reentrant corners, weighted least-squares approaches are presented and analyzed in [17, 16]. Specifically, the method presented in [17] for corner singularities does not rely on the explicit knowledge of the flux singularity at the corner. Its analytic part is computed implicitly. For a weighted least-squares approach in a more general setting, see [25].

Inverse norm functionals. Another potentially more general form of the least-squares approach is based on the $H^{-1}(\Omega)$-norm (see [6, 7, 13, 4]). Such schemes based on "inverse" norms can, in principle, be applied when $f \in H^{-1}(\Omega)$, although the theory has so far restricted f to $L^2(\Omega)$. Thus, both the $H^{-1}(\Omega)$ and $L^2(\Omega)$ versions of FOSLS have been developed under the same general assumptions that are usually in force for mixed methods. Standard finite element spaces are dense in L^2, and thus convergence is obtained, although only in an L^2 sense. This approach uses norms that do not generally take the coefficients of the equation into account and thus have performance that deteriorates for problems with large jumps in the coefficients.

FOSLL functionals.* A more recently developed approach, called FOSLL* [14], can be viewed as a least-squares method based on an inverse norm that involves the operator and thus has superior properties in the presence of large jumps in the coefficients. In addition, it handles the more general case, $f \in H^{-1}(\Omega)$.

Least-squares in $H(\mathrm{div}) \cap H(\mathrm{curl})$. The current paper is concerned with least-squares functionals using finite element spaces in $H(\mathrm{div}) \cap H(\mathrm{curl})$. This paper builds on the theory developed in [2]. Here, and in the companion paper [3], we describe a least-squares approach that includes a curl-condition, $\nabla \times (\mathbf{u}/a) = 0$. While the theory developed in [11] and [12] already allows for discontinuous coefficients, special care must be taken to prove ellipticity, in an appropriate norm, with constants that grow as slowly as possible with respect to the size of the jumps. For this purpose, an appropriate scaling of the least-squares functional that depends on the size of a in different parts of the domain is introduced.

The flux components will, in general, not be in $H^1(\Omega)$, nor will they be in $H^1(\Omega_i)$. Here, we construct singular basis functions for the flux, $\mathbf{u}$, that are in the scaled $H(\mathrm{div}) \cap H(\mathrm{curl})$ but not in $H^1(\Omega_i)$ and have support only near singular points. These are included in our finite element space. As a result, the flux can be computed very accurately near cross points. For standard mixed methods, it would be necessary to make sure that the Ladyzhenskaya–Babuška–Brezzi condition (cf. [8, section 10.5]) is satisfied for the finite element spaces that include the singular function. This is not the case for our first-order system least-squares approach.

In this paper and the companion paper [3], we show that one can add singular basis functions at little additional cost. A singular basis function is composed of a singular function multiplied by a cut-off function that takes the value one in a region around the singularity (the platform) and drops from one to zero in a narrow region around the platform (the fringe). The key is that the singular basis functions satisfy a homogeneous equation of type (1.1) in the platform. Thus, these singular basis functions are orthogonal to any standard basis function that is either supported completely inside the platform or supported completely outside the platform and fringe. Nonzero inner products arise only between singular basis functions and standard basis functions whose support intersects the fringe. As a result, the cost of adding a singular basis function is proportional to the number of grid points in the fringe. In our approach, the fringe has a width of one element, so this additional cost is $O(\sqrt{N})$, where N is the number of grid points.

In this paper, we introduce the problem in section 2; then, in section 3, we construct a scaled FOSLS functional for p and $\mathbf{u}$ and show that this functional is continuous and coercive in a scaled $H^1 \times H(\mathrm{div}) \cap H(\mathrm{curl})$-norm. The coercivity and continuity constants are shown to depend on the coefficient a in a complicated way that involves the geometry of the partition of Ω. We then introduce a *flux-only* functional for $\mathbf{u}$ alone and show that it is continuous and coercive in the scaled version of $H(\mathrm{div}) \cap H(\mathrm{curl})$. In section 4, we introduce the div-curl operator associated with the flux-only functional and discuss its properties. Then, in section 5, we show that the solution, $\mathbf{u}$, can be decomposed as

$$\mathbf{u} = \mathbf{u}_0 + \sum_{m=1}^{M} \sum_{n=1}^{N_m} b_{m,n} \mathbf{s}_{m,n},$$

where $\mathbf{s}_{m,n}$ are a finite number of singular basis functions associated with singular points $\mathbf{x}_m$, $m = 1, \ldots, M$, and $\mathbf{u}_0 \in H^1(\Omega_i)$ for every i. Thus, $\mathbf{u}_0$ can be approximated by standard finite elements within each domain, provided that they posses the proper jumps across domain interfaces.

In the companion paper [3], we show how to compute approximate singular basis functions, and then we construct a finite element basis using them. We develop error estimates by way of new results for nonconforming spaces in the FOSLS context. We prove that the accuracy of singular basis functions need only be $O(h^p)$, $p > 1/2$. Finally, we develop a multilevel algorithm that includes singular basis functions on all coarser levels and provide numerical results that illustrate its performance.

Our restriction to two-dimensional problems is mainly for the purpose of exposition. However, technical complications arise in higher dimensions. For example, two different types of singularities, associated with edges and with corners or cross points, arise in three dimensions. We do not consider these additional complications in the present paper.

2. Problem statement and preliminaries. Consider the following prototype problem on a bounded domain $\Omega \subset \Re^2$:

$$\begin{array}{rcll} -\nabla \cdot (a \nabla p) & = & f & \text{in } \Omega, \\ p & = & 0 & \text{on } \Gamma_D, \\ \mathbf{n} \cdot a \nabla p & = & 0 & \text{on } \Gamma_N, \end{array} \tag{2.1}$$

where $\mathbf{n}$ denotes the outward unit vector normal to the boundary, $f \in L^2(\Omega)$, and $a(x_1, x_2)$ is a scalar function that is uniformly positive and bounded in Ω a.e. but may have large jumps across interfaces. Suppose that Γ_D has positive measure, so that the Poincaré–Friedrichs inequality

$$\|p\|_{0,\Omega} \leq \gamma \|\nabla p\|_{0,\Omega} \tag{2.2}$$

holds for all functions satisfying the boundary conditions in (2.1). Then (2.1) has a unique solution in $H^1(\Omega)$.

Following [12], we rewrite (2.1) as a first-order system by introducing the flux variable, $\mathbf{u} = \sqrt{a} \nabla p$:

$$\begin{array}{rcll} \mathbf{u} - \sqrt{a} \nabla p & = & \mathbf{0} & \text{in } \Omega, \\ -\nabla \cdot \sqrt{a} \mathbf{u} & = & f & \text{in } \Omega, \\ p & = & 0 & \text{on } \Gamma_D, \\ \mathbf{n} \cdot \sqrt{a} \mathbf{u} & = & 0 & \text{on } \Gamma_N. \end{array} \tag{2.3}$$

Since $\mathbf{u}/\sqrt{a} = \nabla p$ with $p \in H^1(\Omega)$, we then have (cf. [18, Theorem 2.9])

$$\nabla \times \left(\frac{\mathbf{u}}{\sqrt{a}} \right) := \partial_1 \left(\frac{u_2}{\sqrt{a}} \right) - \partial_2 \left(\frac{u_1}{\sqrt{a}} \right) = 0 \quad \text{in } \Omega \, .$$

(By the term ∂_k, we mean $\partial / \partial x_k, k = 1, 2$.) Moreover, the homogeneous Dirichlet boundary condition on Γ_D implies the tangential flux condition

$$\boldsymbol{\tau} \cdot \left(\frac{\mathbf{u}}{\sqrt{a}} \right) := \frac{n_1 u_2 - n_2 u_1}{\sqrt{a}} = 0 \quad \text{on } \Gamma_D.$$

(Here, $\boldsymbol{\tau}$ is the counterclockwise unit tangent vector.)

Adding these equations to first-order system (2.3) yields the augmented, but consistent, system

$$
\begin{array}{rcll}
\mathbf{u} - \sqrt{a}\nabla p & = & \mathbf{0} & \text{in } \Omega, \\
-\nabla \cdot \sqrt{a}\mathbf{u} & = & f & \text{in } \Omega, \\
\nabla \times \left(\dfrac{\mathbf{u}}{\sqrt{a}} \right) & = & 0 & \text{in } \Omega, \\
p & = & 0 & \text{on } \Gamma_D, \\
\mathbf{n} \cdot \sqrt{a}\mathbf{u} & = & 0 & \text{on } \Gamma_N, \\
\boldsymbol{\tau} \cdot \left(\dfrac{\mathbf{u}}{\sqrt{a}} \right) & = & 0 & \text{on } \Gamma_D.
\end{array}
\tag{2.4}
$$

Problems (2.1) and (2.4) are equivalent in that their unique solutions are in correspondence (p solves (2.1) if and only if p and $\mathbf{u} = \sqrt{a}\nabla p$ solve (2.4)). If Γ_N is not connected, then we add the constraint

$$
\int_{\Gamma_{N_i}} \boldsymbol{\tau} \cdot \left(\frac{\mathbf{u}}{\sqrt{a}} \right) = 0 \tag{2.5}
$$

for every disjoint piece, Γ_{N_i}, of Γ_N. This constraint is necessary to ensure that the *flux-only* functional described below (see (3.17)) has a unique solution.

For both scalar and vector quantities, denote the standard Sobolev spaces as $L^2(\Omega)$ and $H^k(\Omega)$, with respective norms $\|\cdot\|_{0,\Omega}$ and $\|\cdot\|_{k,\Omega}$. We also define the spaces

$$
\begin{aligned}
H(\operatorname{div} a; \Omega) &:= \left\{ \mathbf{v} \in L^2(\Omega)^2 : \nabla \cdot \sqrt{a}\mathbf{v} \in L^2(\Omega) \right\}, \\
H(\operatorname{curl} a; \Omega) &:= \left\{ \mathbf{v} \in L^2(\Omega)^2 : \nabla \times \left(\frac{\mathbf{v}}{\sqrt{a}} \right) \in L^2(\Omega) \right\}, \\
V &:= \left\{ q \in H^1(\Omega) : q = 0 \text{ on } \Gamma_D \right\}, \\
\mathbf{W} &:= \Bigg\{ \mathbf{v} \in H(\operatorname{div} a; \Omega) \cap H(\operatorname{curl} a; \Omega) : \mathbf{n} \cdot \sqrt{a}\mathbf{v} = 0 \text{ on } \Gamma_N, \\
&\qquad\qquad \boldsymbol{\tau} \cdot \left(\frac{\mathbf{v}}{\sqrt{a}} \right) = 0 \text{ on } \Gamma_D, \int_{\Gamma_{N_i}} \boldsymbol{\tau} \cdot \left(\frac{\mathbf{u}}{\sqrt{a}} \right) = 0 \Bigg\}.
\end{aligned}
$$

Denote the respective seminorm and norm on $\mathbf{W}$ by

$$
\begin{aligned}
|\mathbf{v}|_{\mathbf{W}}^2 &:= \left\| \frac{1}{\sqrt{a}} \nabla \cdot \sqrt{a}\mathbf{v} \right\|_{0,\Omega}^2 + \left\| \sqrt{a}\nabla \times \frac{1}{\sqrt{a}}\mathbf{v} \right\|_{0,\Omega}^2, \\
\|\mathbf{v}\|_{\mathbf{W}}^2 &:= |\mathbf{v}|_{\mathbf{W}}^2 + \|\mathbf{v}\|_{0,\Omega}^2.
\end{aligned}
\tag{2.6}
$$

We show in Lemma 3.3 below that this seminorm is in fact a norm on $\mathbf{W}$ by establishing a Poincaré–Friedrichs-type inequality.

Note that $\mathbf{v} \in \mathbf{W}$ is characterized by the fact that, across any curve in Ω with normal $\mathbf{n}$ and tangent $\boldsymbol{\tau}$, both $\mathbf{n} \cdot \sqrt{a}\mathbf{v}$ and $\boldsymbol{\tau} \cdot \frac{1}{\sqrt{a}}\mathbf{v}$ are continuous (a.e.). (For the first condition see, for example, [32, Chapter 6.2]. The second condition can be derived analogously.) We refer to the continuity of these two terms at lines of discontinuity of a as *interface conditions* for $\mathbf{u} \in \mathbf{W}$. Clearly, for the solution of (2.1), we have $p \in V$ and $\mathbf{u} \in \mathbf{W}$, so it is appropriate to pose (2.4) on these spaces.

As mentioned above, our main interest is in the solution of (2.1) when $a(x_1, x_2)$ has large jumps. For this purpose, we assume that

$$\overline{\Omega} = \bigcup_{i=1}^{J} \overline{\Omega}_i, \tag{2.7}$$

where Ω_i are mutually disjoint, open, simply connected, polygonal regions (see Figure 1.1). Assume also that the restriction of $a(x_1, x_2)$ to Ω_i is in $C^{1,1}(\Omega_i)$ and that

$$c_1\omega_i \leq a(x_1, x_2) \leq c_2\omega_i \quad \text{for all } (x_1, x_2) \in \Omega_i, \tag{2.8}$$

with order one constants c_1, c_2 and arbitrary positive constants ω_i. In other words, $a(x_1, x_2)$ is assumed to be of approximate size ω_i throughout Ω_i for each i, but ω_i is allowed to have large variations over i. In the bounds derived below, we separate the dependence on the variation in $\{\omega_i\}$ from the variation within each Ω_i, that is, on c_1, c_2, and

$$c_3 := \max_{1\leq i\leq J} \|\nabla a\|_{0,\Omega_i} < \infty. \tag{2.9}$$

Given this decomposition of Ω, define the *split* seminorms and norms, respectively, as follows:

$$|\mathbf{v}|^2_{k,S} := \sum_{i=1}^{J} |\mathbf{v}|^2_{k,\Omega_i} \tag{2.10}$$

and

$$\|\mathbf{v}\|^2_{k,S} := \|\mathbf{v}\|^2_{0,\Omega} + \sum_{j=1}^{k} |\mathbf{v}|^2_{j,S}. \tag{2.11}$$

Let $H^k_S(\Omega)$ denote the closure of $C^\infty(\overline{\Omega})$ in the split norm, and define

$$\mathbf{W}^1_S := H^1_S(\Omega) \cap \mathbf{W}. \tag{2.12}$$

We now show that if a is piecewise constant ($c_1 = c_2$ in (2.8)) with respect to the decomposition, then

$$\|\mathbf{v}\|_{1,S} = \|\mathbf{v}\|_{\mathbf{W}} \quad \text{for every } \ \mathbf{v} \in H^1_S(\Omega). \tag{2.13}$$

We first need to establish two lemmas. For the first lemma, consider one polygonal, simply connected subdomain, Ω_i, of Ω, with vertices labeled $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K$ in counterclockwise order. Letting $\mathbf{x}_{K+1} = \mathbf{x}_1$, denote by Γ_j the side connecting $\mathbf{x}_j$ and $\mathbf{x}_{j+1}$. If Γ_j makes angle θ_j with the positive x_1-axis, then $\mathbf{n}_j = (\sin(\theta_j), -\cos(\theta_j))^t$ and $\boldsymbol{\tau}_j = (\cos(\theta_j), \sin(\theta_j))^t$ are the outward unit normal and counterclockwise unit tangent to Γ_j, respectively.

Lemma 2.1. *Assume that Ω_i is a polygonal domain and that $\mathbf{u} = (u_1, u_2)^t \in (H^2(\Omega_i))^2$; then*

$$\begin{aligned} \int\int_{\Omega_i} \partial_1 u_1 \partial_2 u_2 dz \quad &= \int\int_{\Omega_i} \partial_2 u_1 \partial_1 u_2 dz - \int_{\partial\Omega_i} (\boldsymbol{\tau}\cdot\mathbf{u}) d(\mathbf{n}\cdot\mathbf{u}) \\ &+ \frac{1}{2}\sum_{j=1}^{K} \left((\boldsymbol{\tau}_j\cdot\mathbf{u})(\mathbf{n}_j\cdot\mathbf{u})|_{\mathbf{x}_j} - (\boldsymbol{\tau}_{j-1}\cdot\mathbf{u})(\mathbf{n}_{j-1}\cdot\mathbf{u})|_{\mathbf{x}_j}\right). \end{aligned} \tag{2.14}$$

Proof. First, assume that Ω is simply connected. For $\mathbf{u} \in H^2(\Omega_i)$, Green's identity yields

$$\int\int_{\Omega_i} \partial_1 u_1 \partial_2 u_2 dz = \int\int_{\Omega_i} \partial_2 u_1 \partial_1 u_2 dz + \int_{\partial\Omega_i} u_1 du_2.$$

The definition of $\mathbf{n}_i$ and $\boldsymbol{\tau}_i$ and a bit of algebra yield

$$\int_{\Gamma_j} (\boldsymbol{\tau}_j \cdot \mathbf{u}) d(\mathbf{n}_j \cdot \mathbf{u}) = \frac{1}{2} \, (\boldsymbol{\tau}_j \cdot \mathbf{u})(\mathbf{n}_j \cdot \mathbf{u})|_{\mathbf{x}_j}^{\mathbf{x}_{j+1}} + \frac{1}{2} \, u_1 u_2|_{\mathbf{x}_j}^{\mathbf{x}_{j+1}} - \int_{\Gamma_j} u_1 du_2.$$

Summing over the edges yields the result. The result for a general connected polygonal domain is established by cutting Ω_i into simply connected polygonal subdomains and adding the result. □

Lemma 2.2. *For every* $\mathbf{u} \in \mathbf{W}_S^1$, *we have*

$$\int\int_{\Omega} \partial_1 u_1 \partial_2 u_2 dz = \int\int_{\Omega} \partial_2 u_1 \partial_1 u_2 dz. \tag{2.15}$$

Proof. First, let $\mathbf{u} \in H_S^2(\Omega) \cap \mathbf{W}$. The space $\mathbf{W}$ is characterized by the property that, for $\mathbf{u} \in \mathbf{W}$, both $\sqrt{a}\mathbf{n} \cdot \mathbf{u}$ and $\frac{1}{\sqrt{a}}\boldsymbol{\tau} \cdot \mathbf{u}$ are continuous (a.e.) across any curve in Ω. Thus, $(\mathbf{n} \cdot \mathbf{u})(\boldsymbol{\tau} \cdot \mathbf{u})$ is continuous (a.e.). In particular, this holds for the polygonal boundaries between the regions Ω_i. Let Γ_{ij} denote the edge joining Ω_i and Ω_j. Summing the boundary integrals in (2.14) over each Ω_i shows that Γ_{ij} is traversed once in each direction. Thus, only integrals on the boundary of Ω survive. This yields

$$\int\int_{\Omega} \partial_1 u_1 \partial_2 u_2 = \int\int_{\Omega} \partial_2 u_1 \partial_1 u_2 \tag{2.16}$$

$$+ \frac{1}{2} \sum_{j=1}^{\tilde{K}} \left((\tilde{\boldsymbol{\tau}}_j \cdot \mathbf{u})(\tilde{\mathbf{n}}_j \cdot \mathbf{u}) - (\tilde{\boldsymbol{\tau}}_{j-1} \cdot \mathbf{u})(\tilde{\mathbf{n}}_{j-1} \cdot \mathbf{u}) \right) |_{\tilde{\mathbf{x}}_j}, \tag{2.17}$$

where the $\tilde{\mathbf{x}}_j$ now denote the $\tilde{K}$ vertices $\tilde{\mathbf{x}}_j$ on the boundary of Ω, and the $\tilde{\mathbf{n}}_j$ and $\tilde{\boldsymbol{\tau}}_j$ are the corresponding standard normal and tangent vectors. The boundary conditions imposed on $\mathbf{W}$ now imply (2.15) for $\mathbf{u} \in H_S^2(\Omega) \cap \mathbf{W}$. The proof is completed by noting that Lemma 4.3.1.3 in [19] implies that $H_S^2(\Omega) \cap \mathbf{W}$ is dense in $\mathbf{W}_S^1 = H_S^1(\Omega) \cap \mathbf{W}$. □

The next result has important implications for the decomposition of $\mathbf{W}$.

Theorem 2.3. *Suppose* $a = \omega_i$ *(constant) on* Ω_i. *Then*

$$|\mathbf{u}|_{1,S} = |\mathbf{u}|_{\mathbf{W}} \quad \textit{for every} \;\; \mathbf{u} \in \mathbf{W}_S^1. \tag{2.18}$$

Proof. By definition,

$$\begin{aligned} |\mathbf{u}|_{\mathbf{W}}^2 &= \left\| \frac{1}{\sqrt{a}} \nabla \cdot \sqrt{a}\mathbf{u} \right\|_{0,\Omega}^2 + \left\| \sqrt{a} \nabla \times \frac{1}{\sqrt{a}} \mathbf{u} \right\|_{0,\Omega}^2 \\ &= \sum_{i=1}^{J} \left(\left\| \frac{1}{\sqrt{a}} \nabla \cdot \sqrt{a}\mathbf{u} \right\|_{0,\Omega_i}^2 + \left\| \sqrt{a} \nabla \times \frac{1}{\sqrt{a}} \mathbf{u} \right\|_{0,\Omega_i}^2 \right) \\ &= \sum_{i=1}^{J} (\|\nabla \cdot \mathbf{u}\|_{0,\Omega_i}^2 + \|\nabla \times \mathbf{u}\|_{0,\Omega_i}^2). \end{aligned}$$

The theorem now follows from Lemma 2.2 and the easily verified relation

$$\|\nabla\cdot\mathbf{u}\|_{0,\Omega_i}^2+\|\nabla\times\mathbf{u}\|_{0,\Omega_i}^2=|\mathbf{u}|_{1,\Omega_i}+2\langle\partial_1u_1,\,\partial_2u_2\rangle_{0,\Omega_i}-2\langle\partial_2u_1,\,\partial_1u_2\rangle_{0,\Omega_i}.\qquad\square$$

COROLLARY 2.4. *Suppose that $a(x,\,y)$ is now allowed to vary according to* (2.8) *and* (2.9). *Then,*

$$\frac{1}{\delta}\|\mathbf{u}\|_{\mathbf{W}}\le\|\mathbf{u}\|_{1,S}\le\delta\|\mathbf{u}\|_{\mathbf{W}}\quad\textit{for}\;\;\mathbf{u}\in\mathbf{W}_S^1,$$

where

$$\delta=\sqrt{1+c_3\left(\frac{c_3+\sqrt{c_3^2+8}}{4}\right)}$$

and c_3 is defined in (2.9).

Proof. Observe that

$$\left\|\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}\right\|_{0,\Omega_i}\le\|\nabla\cdot\mathbf{u}\|_{0,\Omega_i}+\left\|\frac{1}{2}(\nabla a)\cdot\mathbf{u}\right\|_{0,\Omega_i},$$
$$\left\|\sqrt{a}\nabla\times\frac{1}{\sqrt{a}}\mathbf{u}\right\|_{0,\Omega_i}\le\|\nabla\times\mathbf{u}\|_{0,\Omega_i}+\left\|\frac{1}{2}(\nabla^\perp a)\cdot\mathbf{u}\right\|_{0,\Omega_i}.$$

(Here, we use the notation $\nabla^\perp a:=(-\partial_2a,\,\partial_1a)^t$.) Using the ϵ-inequality twice now yields

$$\begin{aligned}|\mathbf{u}|_{\mathbf{W},\Omega_i}^2&\le\left(\|\nabla\cdot\mathbf{u}\|_{0,\Omega_i}+\frac{c_3}{2}\|\mathbf{u}\|_{0,\Omega_i}\right)^2+\left(\|\nabla\times\mathbf{u}\|_{0,\Omega_i}+\frac{c_3}{2}\|\mathbf{u}\|_{0,\Omega_i}\right)^2\\&\le(1+\epsilon)(\|\nabla\cdot\mathbf{u}\|_{0,\Omega_i}^2+\|\nabla\times\mathbf{u}\|_{0,\Omega_i}^2)+\left(1+\frac{1}{\epsilon}\right)\frac{c_3^2}{2}\|\mathbf{u}\|_{0,\Omega_i}^2\end{aligned}$$

for any $\epsilon>0$. Choosing $\epsilon=c_3\big(\frac{c_3+\sqrt{c_3^2+8}}{4}\big)$, summing over i, and appealing to Theorem 2.3 yields the lower bound. The upper bound is proved in a similar fashion. $\square$

Remark 1. Following the development in section 4.3 in [19], the above results can be extended to problem (2.1) with boundary conditions that involve both the conormal and tangential derivatives, as long as the coefficients remain constant on each edge. We believe that Theorem 2.3 also holds for regions Ω for which $\partial\Omega_i$ are piecewise $C^{1,1}$, but this remains an open question.

3. The least-squares functional. We now turn to the construction of the least-squares functional. An appropriate scaling of the equations in (2.4) leads to

$$G_\alpha(\mathbf{u},p;f):=\alpha\|\mathbf{u}-\sqrt{a}\nabla p\|_{0,\Omega}^2+\left\|\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}+\frac{1}{\sqrt{a}}f\right\|_{0,\Omega}^2+\left\|\sqrt{a}\,\nabla\times\frac{1}{\sqrt{a}}\mathbf{u}\right\|_{0,\Omega}^2\tag{3.1}$$

and associated bilinear form

$$\begin{aligned}\mathcal{F}_\alpha((\mathbf{u},p);(\mathbf{v},q))&=\alpha\langle\mathbf{u}-\sqrt{a}\nabla p,\mathbf{v}-\sqrt{a}\nabla q\rangle_{0,\Omega}\\&\quad+\left\langle\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u},\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{v}\right\rangle_{0,\Omega}+\left\langle\sqrt{a}\,\nabla\times\frac{1}{\sqrt{a}}\mathbf{u},\sqrt{a}\,\nabla\times\frac{1}{\sqrt{a}}\mathbf{v}\right\rangle_{0,\Omega},\end{aligned}\tag{3.2}$$

where $\alpha \geq 0$ will be determined later. Here, for the sake of notational simplicity, we agree that $\langle \cdot, \cdot \rangle_{0,\Omega}$ is meant componentwise for vector functions, e.g., if $\mathbf{w} = (w_1, w_2)$ and $\mathbf{z} = (z_1, z_2)$, then

$$\langle \mathbf{w}, \mathbf{z} \rangle_{0,\Omega} = \langle w_1, z_1 \rangle_{0,\Omega} + \langle w_2, z_2 \rangle_{0,\Omega} \,.$$

The solution of (2.4) also solves the minimization problem

$$G_\alpha(\mathbf{u}, p; f) = \min_{(\mathbf{v},q)\in \mathbf{W}\times V} G_\alpha(\mathbf{v}, q; f) \tag{3.3}$$

and, therefore, the variational problem

$$\mathcal{F}_\alpha((\mathbf{u}, p); (\mathbf{v}, q)) = -\left\langle \frac{1}{\sqrt{a}} f, \frac{1}{\sqrt{a}} \nabla \cdot \sqrt{a}\mathbf{v} \right\rangle_{0,\Omega} \quad \text{for all } (\mathbf{v}, q) \in \mathbf{W} \times V\,. \tag{3.4}$$

In Theorem 3.2, we will show that $(\mathcal{F}_\alpha((\mathbf{v}, q); (\mathbf{v}, q)))^{1/2}$ is uniformly equivalent to the scaled norm defined for $(\mathbf{v}, q) \in \mathbf{W} \times V$ by

$$\begin{aligned} &|||(\mathbf{v}, q)|||_\alpha \\ &:= \left(\left\| \frac{1}{\sqrt{a}} \nabla \cdot \sqrt{a}\mathbf{v} \right\|_{0,\Omega}^2 + \left\| \sqrt{a} \nabla \times \frac{1}{\sqrt{a}} \mathbf{v} \right\|_{0,\Omega}^2 + \alpha \|\mathbf{v}\|_{0,\Omega}^2 + \alpha \|\sqrt{a} \nabla q\|_{0,\Omega}^2 \right)^{1/2} . \end{aligned} \tag{3.5}$$

Note that, for sufficiently smooth a, we get

$$|||(\mathbf{v}, q)|||_\alpha \sim \left(\|\nabla \cdot \mathbf{v}\|_{0,\Omega}^2 + \|\nabla \times \mathbf{v}\|_{0,\Omega}^2 + \alpha \|\mathbf{v}\|_{0,\Omega}^2 + \alpha \|\sqrt{a} \nabla q\|_{0,\Omega}^2 \right)^{1/2} , \tag{3.6}$$

although our assumptions on a do not admit this equivalence in general.

Before we prove the main result, we must establish a scaled Poincaré–Friedrichs inequality. By assumption, Γ_D in (2.1) is a set of positive measures on $\partial\Omega$. Thus, a standard proof can be used to establish

$$\|p\|_{0,\Omega} \leq \gamma_0 \|\nabla p\|_{0,\Omega}, \tag{3.7}$$

for $p \in V$, where γ_0 depends only on Ω. In fact, we may choose γ_0 so that (3.7) holds on any subdomain composed of a union of the Ω_i whose closure is connected and intersects Γ_D in a set of positive measure. In this sense, γ_0 depends also on the partitioning (2.7).

Instead of (3.7), we seek scaled inequalities of the form

$$\|\sqrt{a} p\|_{0,\Omega} \leq c_4 \gamma_0 \|\sqrt{a} \nabla p\|_{0,\Omega} \quad \text{and} \quad \left\| \frac{1}{\sqrt{a}} p \right\|_{0,\Omega} \leq c_5 \gamma_0 \left\| \frac{1}{\sqrt{a}} \nabla^\perp p \right\|_{0,\Omega},$$

for $p \in V$. Of course, if each subdomain is such that $\Gamma_D \cap \overline{\Omega}_i$ is of positive measure, then we may choose, for example, $c_4 = \sqrt{c_2/c_1}$ (see (2.8)). In general, c_4 and c_5 depend on $a(x_1, x_2)$ in a more complicated way that we now characterize.

For each Ω_i, there is a connected path λ_i in Ω from Γ_D to Ω_i that passes through, say, $\overline{\Omega}_{j_1}, \overline{\Omega}_{j_2}, \ldots, \overline{\Omega}_{j_k} = \overline{\Omega}_i$ $(k \leq J)$ in turn, where $\Gamma_D \cap \overline{\Omega}_{j_1}$ and $\overline{\Omega}_{j_\ell} \cap \overline{\Omega}_{j_{\ell-1}}, \ell = 2, \ldots, k$, all have positive measure. We call such a path admissible. Now, let c_1, c_2, and ω_i be as in (2.8) and define

$$C_i = \min_{\lambda_i} \max_{\ell=1,\ldots,k} \frac{\omega_i}{\omega_{j_\ell}}, \quad D_i = \min_{\lambda_i} \max_{\ell=1,\ldots,k} \frac{\omega_{j_\ell}}{\omega_i}, \tag{3.8}$$

and

$$c_4 = \sqrt{\frac{c_2}{c_1}} \max_{i=1,\ldots,J} \sqrt{C_i}, \quad c_5 = \sqrt{\frac{c_2}{c_1}} \max_{i=1,\ldots,J} \sqrt{D_i}. \tag{3.9}$$

Note that, for certain geometries, c_4 or c_5 might depend on the maximum global variation in $a(x_1, x_2)$. However, for other geometries, c_4 or c_5 may be small even for arbitrary large global a-variations. We refer to this property by saying that c_4 and c_5 are P-uniform, meaning that c_4 and c_5 depend on a-variations along the best path to Γ_D, but are otherwise independent of the jumps in a.

LEMMA 3.1. *There exists a P-uniform constant, $\gamma \in (0, \sqrt{J}\gamma_0]$, such that*

$$\|\sqrt{a}p\|_{0,\Omega} \le c_4\gamma \|\sqrt{a}\nabla p\|_{0,\Omega} \quad \textit{for all} \;\; p \in V, \tag{3.10}$$

$$\left\|\frac{1}{\sqrt{a}}p\right\|_{0,\Omega} \le c_5\gamma \left\|\frac{1}{\sqrt{a}}\nabla^{\perp} p\right\|_{0,\Omega} \quad \textit{for all} \;\; p \in V, \tag{3.11}$$

where c_4 and c_5 are the P-uniform constants defined in (3.9).

Proof. Choose Ω_i and any of its admissible paths. By (3.7), we have

$$\sum_{\ell=1}^{k} \|p\|^2_{0,\Omega_{j_\ell}} \le \gamma_0^2 \sum_{\ell=1}^{k} \|\nabla p\|^2_{0,\Omega_{j_\ell}}.$$

In particular,

$$\|p\|^2_{0,\Omega_i} \le \gamma_0^2 \sum_{\ell=1}^{k} \|\nabla p\|^2_{0,\Omega_{j_\ell}}.$$

From (2.8), we have

$$\begin{aligned} \|\sqrt{a}p\|^2_{0,\Omega_i} &\le c_2\omega_i \|p\|^2_{0,\Omega_i} \le c_2\omega_i\gamma_0^2 \sum_{\ell=1}^{k} \|\nabla p\|^2_{0,\Omega_{j_\ell}} \\ &= c_2\gamma_0^2 \sum_{\ell=1}^{k} \frac{\omega_i}{\omega_{j_\ell}} \omega_{j_\ell} \|\nabla p\|^2_{0,\Omega_{j_\ell}} \le \frac{c_2}{c_1}\gamma_0^2 C_i \sum_{\ell=1}^{k} \|\sqrt{a}\nabla p\|^2_{0,\Omega_{j_\ell}}. \end{aligned}$$

Summation over i now yields (3.10) with $\gamma \le \sqrt{J}\gamma_0$. The proof of (3.11) is analogous. □

THEOREM 3.2. *If we choose $\alpha \le 1/c_4^2$, where c_4 is defined in* (3.9), *then there exist P-uniform constants γ_1 and γ_2 such that*

$$\mathcal{F}_\alpha((\mathbf{u}, p); (\mathbf{u}, p)) \ge \gamma_1 |||(\mathbf{u}, p)|||^2_\alpha \quad \textit{for all } (\mathbf{u}, p) \in \mathbf{W} \times V, \tag{3.12}$$

and

$$\mathcal{F}_\alpha((\mathbf{u}, p); (\mathbf{v}, q)) \le \gamma_2 |||(\mathbf{u}, p)|||_\alpha \; |||(\mathbf{v}, q)|||_\alpha \quad \textit{for all } (\mathbf{u}, p)\,, \, (\mathbf{v}, q) \in \mathbf{W} \times V. \tag{3.13}$$

Proof. The proof is similar to the proof of [11, Theorem 3.1] (see also [27, Theorems 2.1 and 2.2]). We include it here because we must confirm that the constants γ_1 and γ_2 are P-uniform. The main part of the proof consists of showing that the functionals

$$\hat{\mathcal{F}}_\alpha((\mathbf{u}, p); (\mathbf{v}, q)) := \alpha\langle \mathbf{u} - \sqrt{a}\nabla p, \mathbf{v} - \sqrt{a}\nabla q\rangle_{0,\Omega} + \left\langle \frac{1}{\sqrt{a}}\nabla \cdot \sqrt{a}\mathbf{u}, \frac{1}{\sqrt{a}}\nabla \cdot \sqrt{a}\mathbf{v} \right\rangle_{0,\Omega}$$

and

$$\hat{\mathcal{S}}_\alpha(\mathbf{u},p;\mathbf{v},q) := \alpha\langle \mathbf{u},\mathbf{v}\rangle_{0,\Omega} + \alpha\langle \sqrt{a}\nabla p, \sqrt{a}\nabla q\rangle_{0,\Omega} + \left\langle \frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}, \frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{v}\right\rangle_{0,\Omega}$$

satisfy

$$C_1\hat{\mathcal{S}}_\alpha(\mathbf{u},p;\mathbf{u},p) \le \hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{u},p)) \tag{3.14}$$

and

$$\hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{v},q)) \le C_2(\hat{\mathcal{S}}_\alpha(\mathbf{u},p;\mathbf{u},p))^{1/2}(\hat{\mathcal{S}}_\alpha(\mathbf{v},q;\mathbf{v},q))^{1/2}, \tag{3.15}$$

with constants C_1 and C_2 that are P-uniform. Since on $\partial\Omega$ we either have $p = 0$ or $\mathbf{n}\cdot\sqrt{a}\mathbf{u} = 0$, then integration by parts confirms that

$$\langle \mathbf{u}, \sqrt{a}\nabla p\rangle_{0,\Omega} + \langle \nabla\cdot\sqrt{a}\mathbf{u}, p\rangle_{0,\Omega} = 0.$$

For any $\beta > 0$, which we specify later, we have

$$\begin{aligned}
\hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{u},p)) &= \alpha\langle \mathbf{u},\mathbf{u}\rangle_{0,\Omega} + \alpha\langle \sqrt{a}\nabla p, \sqrt{a}\nabla p\rangle_{0,\Omega} - 2\alpha\langle \mathbf{u}, \sqrt{a}\nabla p\rangle_{0,\Omega}\\
&\quad + \left\langle \frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}, \frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}\right\rangle_{0,\Omega} + 2\alpha\beta\langle \nabla\cdot\sqrt{a}\mathbf{u}, p\rangle_{0,\Omega}\\
&\quad + 2\alpha\beta\langle \mathbf{u}, \sqrt{a}\nabla p\rangle_{0,\Omega} + \alpha^2\beta^2\langle \sqrt{a}p, \sqrt{a}p\rangle_{0,\Omega} - \alpha^2\beta^2\langle \sqrt{a}p, \sqrt{a}p\rangle_{0,\Omega}\\
&= \alpha\langle \mathbf{u} + (\beta-1)\sqrt{a}\nabla p, \mathbf{u} + (\beta-1)\sqrt{a}\nabla p\rangle_{0,\Omega}\\
&\quad + \left\langle \frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u} + \alpha\beta\sqrt{a}p, \frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u} + \alpha\beta\sqrt{a}p\right\rangle_{0,\Omega}\\
&\quad + \alpha(2\beta-\beta^2)\langle \sqrt{a}\nabla p, \sqrt{a}\nabla p\rangle_{0,\Omega} - \alpha^2\beta^2\langle \sqrt{a}p, \sqrt{a}p\rangle_{0,\Omega}\\
&\ge \alpha(2\beta-\beta^2)\langle \sqrt{a}\nabla p, \sqrt{a}\nabla p\rangle_{0,\Omega} - \alpha^2\beta^2\langle \sqrt{a}p, \sqrt{a}p\rangle_{0,\Omega}\\
&\ge \alpha(2\beta - (1+\gamma^2)\beta^2)\|\sqrt{a}\nabla p\|^2_{0,\Omega}\,,
\end{aligned}$$

where we used the assumption that $\alpha \le 1/c_4^2$ and where γ is from Lemma 3.1. Choosing $\beta = 1/(1+\gamma^2)$ leads to

$$\hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{u},p)) \ge \beta\alpha\|\sqrt{a}\nabla p\|^2_{0,\Omega}\,.$$

We then also have

$$\alpha\|\mathbf{u}\|^2_{0,\Omega} \le 2\alpha(\|\mathbf{u} - \sqrt{a}\nabla p\|^2_{0,\Omega} + \|\sqrt{a}\nabla p\|^2_{0,\Omega}) \le 2\left(1 + \frac{1}{\beta}\right)\hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{u},p))$$

and, clearly,

$$\left\|\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}\right\|^2_{0,\Omega} \le \hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{u},p)),$$

which completes the proof of (3.14). □

Upper bound (3.15) follows from

$$\hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{v},q)) \le 2(\hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{u},p)))^{1/2}(\hat{\mathcal{F}}_\alpha((\mathbf{v},q);(\mathbf{v},q)))^{1/2}$$

and

$$
\begin{aligned}
\hat{\mathcal{F}}_\alpha((\mathbf{u},p);(\mathbf{u},p)) &= \alpha\|\mathbf{u}-\sqrt{a}\nabla p\|_{0,\Omega}^2 + \left\|\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}\right\|_{0,\Omega}^2 \\
&\le 2\left(\alpha\|\mathbf{u}\|_{0,\Omega}^2 + \alpha\left\|\sqrt{a}\nabla p\right\|_{0,\Omega}^2 + \left\|\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}\right\|_{0,\Omega}^2\right) \qquad (3.16)\\
&= 2\hat{\mathcal{S}}_\alpha(\mathbf{u},p;\mathbf{u},p).
\end{aligned}
$$

The proof of Theorem 3.2 is completed by adding the term $\|\sqrt{a}\nabla\times(\mathbf{u}/\sqrt{a})\|_{0,\Omega}^2$ to both sides of inequalities (3.14) and (3.16). □

Theorem 3.2 establishes coercivity and continuity of the least-squares bilinear form $\mathcal{F}_\alpha((\cdot,\cdot);(\cdot,\cdot))$ in terms of the norm $|||(\cdot,\cdot)|||_\alpha$. This norm equivalence depends on the jumps in a along the best path to the Dirichlet boundary, but is otherwise independent of the jumps in a.

The scaling of the norm $|||(\cdot,\cdot)|||_\alpha$ has the following physical interpretation. Focusing first on p, imagine that the error q as measured by the term $\|\sqrt{a}\nabla q\|_{0,\Omega}^2$ is balanced over the domain; that is, $\sqrt{a}\nabla q$ is roughly constant. Then, in areas where $\sqrt{a}$ is relatively small, ∇q is correspondingly relatively large, and one has to expect a less accurate approximation (in the L^2 sense) there compared to areas where $\sqrt{a}$ is large and ∇q is therefore small. In contrast, approximation of the velocity $\mathbf{u}=\sqrt{a}\nabla p$ (assuming the error $\mathbf{v}$ is balanced in the sense of the term $|\mathbf{v}|_{1,\Omega}^2+\alpha\|\mathbf{v}\|_{0,\Omega}^2$; see (3.6)) can be expected to have balanced accuracy (in the L^2 sense) over Ω. Ellipticity with constants that are independent of the global jumps in a asserts that the scaling in $\mathcal{F}_\alpha((\cdot,\cdot);(\cdot,\cdot))$ correctly reflects these attributes.

Uniform coercivity and continuity of $\mathcal{F}$ in the norm $|||(\cdot,\cdot)|||_\alpha$ allows for effective computation of $\mathbf{u}$ and p together by finite element and multigrid techniques. Notice that the result is valid for all $\alpha\in[0,1/c_4^2]$. Proof of Theorem 3.2 for the case $\alpha=0$ is trivial, with $\gamma_1=\gamma_2=1$. Moreover, this choice reveals a perhaps simpler alternative: we can use a two-stage approach (cf. [13]) that first minimizes the flux-only functional,

$$
G_0(\mathbf{u};f) = \left\|\frac{1}{\sqrt{a}}(\nabla\cdot\sqrt{a}\mathbf{u}+f)\right\|_{0,\Omega}^2 + \left\|\sqrt{a}\,\nabla\times\left(\frac{\mathbf{u}}{\sqrt{a}}\right)\right\|_{0,\Omega}^2, \qquad (3.17)
$$

over $\mathbf{u}\in\mathbf{W}$, then fixes $\mathbf{u}/\sqrt{\mathbf{a}}$ and minimizes the Poisson functional,

$$
G_P\left(p;\frac{\mathbf{u}}{\sqrt{a}}\right) = \left\|\nabla p - \frac{\mathbf{u}}{\sqrt{a}}\right\|_{0,\Omega}^2,
$$

over $p\in V$. The efficacy of this two-stage approach is confirmed by the uniform coercivity and continuity of $G_P(p;0)$ in the $H^1(\Omega)$ seminorm $\|\nabla p\|_{0,\Omega}^2$, which by (3.7) is itself a norm on V, and of $G_1(\mathbf{u};0)$ in the $\mathbf{W}$ seminorm as defined in (2.6), which we now demonstrate is a norm on $\mathbf{W}$ by establishing a Poincaré–Friedrichs inequality.

LEMMA 3.3. *We have*

$$
\|\mathbf{u}\|_{0,\Omega} \le c_6\gamma|\mathbf{u}|_{\mathbf{W}} \quad \textit{for all} \quad \mathbf{u}\in\mathbf{W}, \qquad (3.18)
$$

where $c_6=\max\{c_4,c_5\}$ *(see* 3.9*) and* γ *is from Lemma* 3.1.

Proof. Consider a Helmholtz decomposition on $\mathbf{W}$: for $\mathbf{u}\in\mathbf{W}$, there exist p, $\psi\in H^1(\Omega)$ such that

$$
\mathbf{u} = \sqrt{a}\nabla p + \frac{1}{\sqrt{a}}\nabla^\perp\psi, \qquad (3.19)
$$

where p is unique the solution of (2.1) with $f = -\nabla \cdot \sqrt{a}\mathbf{u}$ and ψ is the unique (up to a constant) solution of

$$\begin{aligned} -\nabla \cdot \left(\frac{1}{a}\nabla\psi\right) &= -\nabla \times \frac{1}{\sqrt{a}}\mathbf{u} && \text{in } \Omega, \\ \psi &= C_i && \text{on } \Gamma_{N_i}, \\ \mathbf{n} \cdot \frac{1}{a}\nabla\psi &= 0 && \text{on } \Gamma_D, \end{aligned} \tag{3.20}$$

where C_i are arbitrary constants, one of which may be set to zero. Since $\mathbf{u} \in \mathbf{W}$, it satisfies the integral constraints

$$\int_{\Gamma_{N_i}} \boldsymbol{\tau} \cdot \frac{1}{\sqrt{a}}\mathbf{u} = 0$$

for each disjoint piece of Γ_N. Thus, we may set the constants $C_i = 0$, and (3.20) will have a unique solution.

Note that the decomposition is orthogonal in the L^2 sense:

$$\left\langle \sqrt{a}\nabla p,\, \frac{1}{\sqrt{a}}\nabla^\perp\psi \right\rangle_{0,\Omega} = 0. \tag{3.21}$$

We thus have

$$\|\mathbf{u}\|_{0,\Omega}^2 = \|\sqrt{a}\nabla p\|_{0,\Omega}^2 + \left\|\frac{1}{\sqrt{a}}\nabla^\perp\psi\right\|_{0,\Omega}^2. \tag{3.22}$$

Now,

$$-\nabla \cdot a\nabla p = -\nabla \cdot \sqrt{a}\mathbf{u},$$

so that, using (3.10),

$$\begin{aligned} \|\sqrt{a}\nabla p\|_{0,\Omega}^2 &= \langle -\nabla \cdot a\nabla p,\, p\rangle_{0,\Omega} \\ &= \langle -\nabla \cdot \sqrt{a}\mathbf{u},\, p\rangle_{0,\Omega} \\ &= \left\langle -\frac{1}{\sqrt{a}}\nabla \cdot \sqrt{a}\mathbf{u},\, \sqrt{a}p \right\rangle_{0,\Omega} \\ &\le \left\|\frac{1}{\sqrt{a}}\nabla \cdot \sqrt{a}\mathbf{u}\right\|_{0,\Omega} \|\sqrt{a}p\|_{0,\Omega} \\ &\le c_4\gamma \left\|\frac{1}{\sqrt{a}}\nabla \cdot \sqrt{a}\mathbf{u}\right\|_{0,\Omega} \|\sqrt{a}\nabla p\|_{0,\Omega}, \end{aligned}$$

which yields

$$\|\sqrt{a}\nabla p\|_{0,\Omega} \le c_4\gamma \left\|\frac{1}{\sqrt{a}}\nabla \cdot \sqrt{a}\mathbf{u}\right\|_{0,\Omega}. \tag{3.23}$$

Similarly, using (3.11),

$$\begin{aligned}\left\|\frac{1}{\sqrt{a}}\nabla^{\perp}\psi\right\|^2_{0,\Omega} &= \left\langle -\nabla\times\frac{1}{a}\nabla^{\perp}\psi,\, \psi\right\rangle_{0,\Omega}\\ &= \left\langle -\nabla\times\frac{1}{\sqrt{a}}\mathbf{u},\, \psi\right\rangle_{0,\Omega}\\ &= \left\langle -\sqrt{a}\nabla\times\frac{1}{\sqrt{a}}\mathbf{u},\, \frac{1}{\sqrt{a}}\psi\right\rangle_{0,\Omega}\\ &\le \left\|\sqrt{a}\nabla\times\frac{1}{\sqrt{a}}\mathbf{u}\right\|_{0,\Omega}\left\|\frac{1}{\sqrt{a}}\psi\right\|_{0,\Omega}\\ &\le c_5\gamma\left\|\sqrt{a}\nabla\times\frac{1}{\sqrt{a}}\mathbf{u}\right\|_{0,\Omega}\left\|\frac{1}{\sqrt{a}}\nabla^{\perp}\psi\right\|_{0,\Omega},\end{aligned}$$

which yields

$$\left\|\frac{1}{\sqrt{a}}\nabla^{\perp}\psi\right\|_{0,\Omega} \le c_5\gamma\left\|\sqrt{a}\nabla\times\frac{1}{\sqrt{a}}\mathbf{u}\right\|_{0,\Omega}. \tag{3.24}$$

The result now follows from (3.22)–(3.24), where $c_6 = \max\{c_4, c_5\}$. □

For simplicity of discussion, the following sections focus on the two-stage approach described above.

4. Scaled div-curl operator. We are now in a position to define the scaled div-curl operator and develop some tools that will aid in the proof of the decomposition of $\mathbf{W}$ in the next section. Define $\mathcal{L} : \mathbf{W} \to (L^2(\Omega))^2$ as follows:

$$\mathcal{L} := \begin{bmatrix} \frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a} \\ \\ \sqrt{a}\nabla\times\frac{1}{\sqrt{a}} \end{bmatrix}, \tag{4.1}$$

with domain $\mathcal{D}(\mathcal{L}) = \mathbf{W}$. It is straightforward to verify that the adjoint of $\mathcal{L}$ is given by

$$\mathcal{L}^* := -\left[\sqrt{a}\nabla\frac{1}{\sqrt{a}},\ \frac{1}{\sqrt{a}}\nabla^{\perp}\sqrt{a}\right], \tag{4.2}$$

with domain

(4.3)

$$\mathcal{D}(\mathcal{L}^*) := \left\{\mathbf{q}\ :\ \left(\frac{1}{\sqrt{a}}q_1,\, \sqrt{a}q_2\right)^t \in (H^1(\Omega))^2,\, q_1 = 0\ \text{ on }\ \Gamma_D,\, q_2 = C_i \text{ on } \Gamma_{N_i}\right\},$$

where C_i are arbitrary constants, one of which may be set to zero. We summarize properties of $\mathcal{L}$ and $\mathcal{L}^*$ in the following lemma.

LEMMA 4.1. *The operator $\mathcal{L}$ is continuous and coercive on $\mathbf{W}$, the range $\mathcal{R}(\mathcal{L})$ is closed in $(L^2(\Omega))^2$, and*

$$\mathcal{R}(\mathcal{L})^{\perp} = \mathcal{N}(\mathcal{L}^*) = \left\{\begin{pmatrix} 0 \\ \frac{1}{\sqrt{a}}\end{pmatrix}\right\}.$$

Proof. The first result follows directly from Lemma 3.3. For the second result, note for $\mathbf{u} \in \mathbf{W}$ we have

$$\|\mathbf{u}\|_{\mathbf{W}} \le (c_6\gamma + 1)|\mathbf{u}|_{\mathbf{W}} = \|\mathcal{L}\mathbf{u}\| \le \|\mathbf{u}\|_{\mathbf{W}}, \tag{4.4}$$

which implies that $\mathcal{R}(\mathcal{L})$ is closed in $(L^2(\Omega))^2$. For the last result, note for $\mathbf{u} \in \mathbf{W}$ that

$$\left\langle \frac{1}{\sqrt{a}}\nabla \cdot \sqrt{a}\mathbf{u},\, 0 \right\rangle + \left\langle \sqrt{a}\nabla \times \frac{1}{\sqrt{a}}\mathbf{u},\, \frac{1}{\sqrt{a}} \right\rangle = \int\int_{\Omega} \nabla \times \frac{1}{\sqrt{a}}\mathbf{u} = \oint \boldsymbol{\tau} \cdot \frac{1}{\sqrt{a}}\mathbf{u} = 0.$$

The last equality follows from the boundary conditions imposed on $\mathbf{u}$. Thus, $(0,\, \frac{1}{\sqrt{a}})^t \in \mathcal{R}(\mathcal{L})^{\perp} = \mathcal{N}(\mathcal{L}^*)$.

To show that this function spans $\mathcal{N}(\mathcal{L}^*)$, suppose that $\mathbf{q} \in \mathcal{D}(\mathcal{L}^*)$ satisfies

$$-\mathcal{L}^*\mathbf{q} = \sqrt{a}\nabla \frac{1}{\sqrt{a}} q_1 + \frac{1}{\sqrt{a}}\nabla^{\perp}\sqrt{a} q_2 = \mathbf{0}. \tag{4.5}$$

Let $p_1 = q_1/\sqrt{a}$, $p_2 = \sqrt{a}q_2$. From the boundary conditions on $\mathbf{q}$ and (4.5), we see that

$$\mathbf{n} \cdot \sqrt{a}\nabla p_1 = \mathbf{n} \cdot \left(\sqrt{a}\nabla p_1 + \frac{1}{\sqrt{a}}\nabla^{\perp} p_2 \right) = 0 \qquad \text{on } \Gamma_N. \tag{4.6}$$

Since $\frac{1}{\sqrt{a}}\nabla^{\perp} p_2 \in H(\operatorname{div} a; \Omega)$, then $\sqrt{a}\nabla p_1 \in H(\operatorname{div} a; \Omega)$. Thus, p_1 satisfies (2.1) with homogeneous data, which implies that $p_1 = 0$. This leaves $\nabla^{\perp} p_2 = 0$, which implies $p_2 = C$ and finally $q_2 = \frac{C}{\sqrt{a}}$ for some arbitrary constant C. Since this is the only solution of (4.5), the result is proved. □

Next, we define the restriction of $\mathcal{L}$ to $\mathbf{W}^1_S$:

$$\widehat{\mathcal{L}} := \mathcal{L}|_{\mathbf{W}^1_S}. \tag{4.7}$$

Since $\widehat{\mathcal{L}} \subseteq \mathcal{L}$, we know that $\mathcal{L}^* \subseteq \widehat{\mathcal{L}}^*$; that is,

$$\mathcal{D}(\widehat{\mathcal{L}}^*) = \left\{ \mathbf{q} \in (L^2(\Omega))^2 \ : \ \mathcal{L}^*\mathbf{q} \in (L^2(\Omega))^2,\, q_1 = 0 \text{ on } \Gamma_D,\, q_2 = C_i \text{ on } \Gamma_{N_i} \right\}. \tag{4.8}$$

This larger definition of $\mathcal{D}(\widehat{\mathcal{L}}^*)$ will be important in proving the decomposition in the next section. Finally, we have the following result.

LEMMA 4.2. *Subspace* $\mathbf{W}^1_S$ *is closed in* $\mathbf{W}$ *and* $\mathcal{R}(\widehat{\mathcal{L}}) \subseteq \mathcal{R}(\mathcal{L})$ *are both closed in* $(L^2(\Omega))^2$.

Proof. The result is an immediate consequence of Theorem 2.3, Corollary 2.4, and Lemma 4.1. □

5. Solution decomposition. Here, we introduce a splitting of the flux space $\mathbf{W}$ into a finite-dimensional space spanned by singular functions and locally smooth functions, that is, functions that are $H^1_S(\Omega)$. As a result, the flux $\mathbf{u}$ can be discretized as the sum of singular basis functions and standard basis functions that satisfy the interface conditions. This splitting provides the foundation for the finite element method that we present in [3]. For a detailed description of the finite element spaces, see also [2].

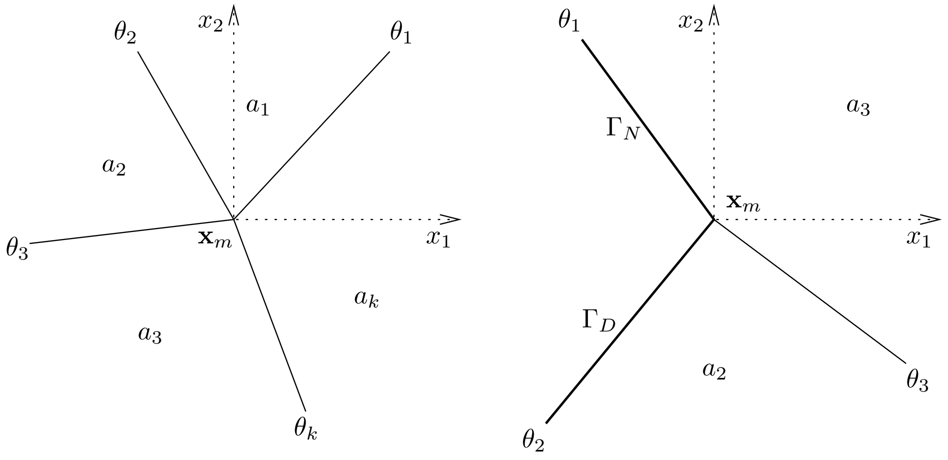


FIG. 5.1. *Cross point (on the left, $K = 4$), and boundary cross point (on the right, $K = 3$).*

In this context, a singular function is any function $\mathbf{u} \in \mathbf{W}$ such that $\mathbf{u} \notin \mathbf{W}^1_S(\Omega)$. This leads to a decomposition of any $\mathbf{u} \in \mathbf{W}$ as

$$\mathbf{u} = \mathbf{u}_0 + \sum_{m=1}^{M} \sum_{n=1}^{N_m} b_{m,n}\mathbf{s}_{m,n}, \tag{5.1}$$

where $\mathbf{u}_0 \in \mathbf{W}^1_S$, and $\mathbf{s}_{m,n}$, $n = 1, \ldots, N_m$, are singular functions associated with singular points $\mathbf{x}_m$, $m = 1, \ldots, M$.

This decomposition will be established, following the development in Kellogg [24] and Grisvard [19], by demonstrating a linearly independent set of functions $\mathbf{s}_{m,n} \in \mathbf{W}\backslash\mathbf{W}^1_S$ and then using a counting argument to show that they span all of $\mathbf{W}\backslash\mathbf{W}^1_S$. In fact, we will demonstrate two sets of functions, one associated with singular solutions of (2.1) and the other associated with singular solutions of (3.20), and show that they span the same space. The fact that they span the same space will be essential to the counting argument.

We first examine singular functions of the original equation (2.1). A singular function of (2.1) is a function $p \in H^1(\Omega) \setminus H^2_S(\Omega)$ for which $\nabla \cdot a\nabla p \in L^2(\Omega)$. As described in the introduction, singular points are associated with cross points, boundary cross points, reentrant corners, and irregular boundary points.

We begin with interior singular points. Boundary singular points are handled in a similar manner. First, we restrict our attention to the ball of radius R, call it $B_m(R)$, centered at the singular point $\mathbf{x}_m$ that contains no other singular points, and we establish a polar coordinate system (r, θ) centered at $\mathbf{x}_m$. For example, consider Figure 5.1. Denote the angle of the boundaries between segments to the positive x_1-axis by θ_i for $i = 1, \ldots, K$. In the following, we use the convention that $\theta_{-1} = \theta_K$ and $\theta_{K+1} = \theta_1$.

We seek solutions of the homogeneous equation

$$\nabla \cdot a\nabla p = \partial_r a\partial_r p + \frac{1}{r}a\partial_r p + \frac{1}{r^2}\partial_\theta a\partial_\theta p = 0 \tag{5.2}$$

in $B_m(R)$. Substituting $p = r^\alpha T(\theta)$ and dividing by $r^{\alpha-2}$ yields the problem

$$-(aT_\theta(\theta))_\theta = (a\alpha^2 + ra_r\alpha)T(\theta). \tag{5.3}$$

Here, we make the additional assumption on a that, within each segment, $\lim_{r\to 0} a_\theta = 0$. Since it was assumed above that $a \in C^{1,1}(\Omega_i)$ for each subdomain Ω_i, we also know that $\lim_{r\to 0} r a_r = 0$. Thus, we may substitute the value

$$\tilde{a}_i = \lim_{r\to 0} a(r,\theta) \qquad \text{in} \quad \Omega_i. \tag{5.4}$$

With this replacement, (5.3) now becomes the the Sturm–Liouville eigenvalue problem

$$-(\tilde{a}T')' = \tilde{a}\alpha^2 T \quad \text{on } [0, 2\pi). \tag{5.5}$$

Solutions of this equation are of the form

$$T_n(\theta) = A_{n,i}\cos(\alpha_n(\theta-\theta_i)) + B_{n,i}\sin(\alpha_n(\theta-\theta_i)), \tag{5.6}$$

for $\theta \in (\theta_i, \theta_{i+1})$, with corresponding eigenvalue

$$\lambda_n = \alpha_n^2. \tag{5.7}$$

The singular functions we seek are constructed by choosing only those $\alpha_n \in (0,1)$ for, say, $n = 1, \ldots, N_m$. Note that for any solution with $\alpha = \alpha_n \in (0,1)$, there is a solution with $\alpha = -\alpha_n \in (-1,0)$. These solutions will be important in the counting argument.

Now, let $\tilde{\delta}_m(r) \in H^2(0,R)$ be a smooth cut-off function that is equal to 1 for $r \in (0, R/2)$ and drops to 0 for $r \in (R/2, R)$. It is easy to see that

$$s_{m,n} := \tilde{\delta}_m(r) r^{\alpha_n} T_n(\theta) \tag{5.8}$$

is in the domain of boundary value problem (2.1). Moreover, for any cut-off function $\delta_m \in H^1(0,R)$, we see that

$$\mathbf{s}_{m,n} := \delta_m(r)\sqrt{a}\nabla r^{\alpha_n} T_n(\theta) \in \mathbf{W} \setminus \mathbf{W}_S^1. \tag{5.9}$$

The exponent α and the coefficients (A_i, B_i) can be determined by enforcing continuity of both $T(\theta)$ and $aT'(\theta)$ across interfaces. (We have dropped the first subscript for convenience.) This may be expressed as

$$\begin{bmatrix} 1 & 0 \\ 0 & -\tilde{a}_i \end{bmatrix} \begin{pmatrix} A_i \\ B_i \end{pmatrix} = \begin{bmatrix} \cos(\alpha(\theta_i-\theta_{i-1})) & \sin(\alpha(\theta_i-\theta_{i-1})) \\ \tilde{a}_{i-1}\sin(\alpha(\theta_i-\theta_{i-1})) & -\tilde{a}_{i-1}\cos(\alpha(\theta_i-\theta_{i-1})) \end{bmatrix} \begin{pmatrix} A_{i-1} \\ B_{i-1} \end{pmatrix}, \tag{5.10}$$

for $i = 1, \ldots, K$. Divide the second equation by $\tilde{a}_{i-1}$, define $\delta_i := \tilde{a}_i/\tilde{a}_{i-1}$ and

$$D_i := \begin{bmatrix} 1 & 0 \\ 0 & -\delta_i \end{bmatrix}, \qquad C_i := \begin{bmatrix} \cos(\alpha(\theta_i-\theta_{i-1})) & \sin(\alpha(\theta_i-\theta_{i-1})) \\ \sin(\alpha(\theta_i-\theta_{i-1})) & -\cos(\alpha(\theta_i-\theta_{i-1})) \end{bmatrix}, \tag{5.11}$$

and finally define $\underline{\beta}_i := (A_i, B_i)^t$. Then, the above constraints may be expressed by the homogeneous system

$$M\underline{b} = \begin{bmatrix} D_1 & 0 & \cdots & C_K \\ -C_1 & D_2 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & -C_{K-1} & D_K \end{bmatrix} \begin{pmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \\ \vdots \\ \underline{\beta}_K \end{pmatrix} = \begin{pmatrix} \underline{0} \\ \underline{0} \\ \vdots \\ \underline{0} \end{pmatrix}. \tag{5.12}$$

A nontrivial solution exists only when the determinant of M is zero. The corresponding null vector yields the coefficients.

We now turn our attention to singular solutions of the boundary value problem (3.20). In $B_m(R)$ we seek solutions to the homogeneous problem

$$\nabla \cdot \frac{1}{a} \nabla p = 0. \tag{5.13}$$

Following the same arguments, we are led to the Sturm–Liouville eigenvalue problem

$$-\left(\frac{1}{\tilde{a}}\hat{T}'\right)' = \frac{1}{\tilde{a}}\alpha^2 \hat{T} \quad \text{on } [0, 2\pi) \tag{5.14}$$

and solutions of the form

$$\hat{T}_n(\theta) = \hat{A}_{n,i}\cos(\alpha_n(\theta - \theta_i)) + \hat{B}_{n,i}\sin(\alpha_n(\theta - \theta_i)), \tag{5.15}$$

for $\theta \in (\theta_i, \theta i + 1)$.

Again, we choose only those $\alpha_n \in (0, 1)$. With $\tilde{\delta}(r) \in H^2(0, R)$, solutions of this Sturm–Liouville problem yield

$$\hat{s}_{m,n} = \tilde{\delta}_m(r) r^{\alpha_n} \hat{T}_n(\theta) \tag{5.16}$$

in the domain of boundary value problem (3.20) and, with $\delta_m \in H^1(0, R)$,

$$\hat{\mathbf{s}}_{m,n} = \delta_m(r)\frac{1}{\sqrt{a}}\nabla^{\perp} r^{\alpha_n}\hat{T}_n(\theta) \in \mathbf{W} \setminus \mathbf{W}^1_S. \tag{5.17}$$

It would appear that there are at least two families of singular function in $\mathbf{W}\backslash\mathbf{W}^1_S$. We now show that they are in fact the same family. To see this, first notice that the only change to the continuity constraints (5.10) is that $\tilde{a}_i$, $\tilde{a}_{i-1}$ are replaced by $1/\tilde{a}_i$ and $1/\tilde{a}_{i-1}$ respectively, which results in replacing D_i by D_i^{-1}. Thus, with the definition $\underline{\hat{\beta}}_i := (\hat{A}_i, \hat{B}_i)$ and similar notation for the other variables, the homogeneous system (5.12) becomes

$$\hat{M}\underline{\hat{b}} := \begin{bmatrix} D_1^{-1} & 0 & \cdots & C_K \\ -C_1 & D_2^{-1} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & -C_{K-1} & D_K^{-1} \end{bmatrix} \begin{pmatrix} \underline{\hat{\beta}}_1 \\ \underline{\hat{\beta}}_2 \\ \vdots \\ \underline{\hat{\beta}}_K \end{pmatrix} = \begin{pmatrix} \underline{0} \\ \underline{0} \\ \vdots \\ \underline{0} \end{pmatrix}. \tag{5.18}$$

We now show that $\det M = \det(\hat{M})$. Define the 2×2 rotation

$$Q_2 = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \tag{5.19}$$

and notice that $Q_2^t Q_2 = I_2, Q_2 C_i Q_2 = C_i$, and

$$Q_2 D_i Q_2 = \begin{bmatrix} \delta_i & 0 \\ 0 & -1 \end{bmatrix} = \delta_i D_i^{-1}. \tag{5.20}$$

Note that $\det(Q_2) = -1$ and define the $2K \times 2K$ block diagonal matrix $Q = \operatorname{diag}(Q_2, Q_2, \ldots, Q_2)$. This yields

$$QMQ = \begin{bmatrix} \delta_1 D_1^{-1} & 0 & \cdots & C_K \\ -C_1 & \delta_2 D_2^{-1} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & -C_{K-1} & \delta_K D_K^{-1} \end{bmatrix}. \tag{5.21}$$

Next, define the $2K \times 2K$ block matrices

$$\begin{aligned} \Delta_1 &:= \operatorname{diag}(\tilde{a}_1 I_2, \tilde{a}_2 I_2, \ldots, \tilde{a}_K I_2), \\ \Delta_2 &:= \operatorname{diag}(\tilde{a}_K I_2, \tilde{a}_1 I_2, \ldots, \tilde{a}_{K-1} I_2). \end{aligned}$$

We can now establish

$$\Delta_2 QMQ \Delta_1^{-1} = \hat{M}, \tag{5.22}$$

which yields

$$\det(\hat{M}) = \det(\Delta_1)\det(\Delta_2^{-1})\det(Q)^2\det(M) = \det(M). \tag{5.23}$$

Let $\alpha_n \in (0,1)$ be a root of $\det(M) = 0$, and consider the associated null vector $M\underline{b}_n = 0$. Using the above relationships, we have

$$0 = (\Delta_2 QM)\underline{b}_n = (\Delta_2 QMQ\Delta_1^{-1})(\Delta_1 Q^t \underline{b}_n) = \hat{M}(\Delta_1 Q^t \underline{b}_n). \tag{5.24}$$

Thus, $\hat{b}_n = (\Delta_1 Q^t \underline{b}_n)$ is the corresponding null vector of $\hat{M}$, which yields

$$\begin{pmatrix} \hat{A}_{n,i} \\ \hat{B}_{n,i} \end{pmatrix} = \tilde{a}_i \begin{pmatrix} -B_{n,i} \\ A_{n,i} \end{pmatrix}. \tag{5.25}$$

For convenience, define

$$\phi_n(r,\theta) = r^{\alpha_n}\left(A_{n,i}\cos(\alpha_n(\theta-\theta_i)) + B_{n,i}\sin(\alpha_n(\theta-\theta_i))\right), \tag{5.26}$$

$$\psi_n(r,\theta) = r^{\alpha_n}(\hat{A}_{n,i}\cos(\alpha_n(\theta-\theta_i)) + \hat{B}_{n,i}\sin(\alpha_n(\theta-\theta_i))), \tag{5.27}$$

for $\theta \in (\theta_i, \theta_{i+1})$. Recall that

$$\nabla = \begin{pmatrix} \partial_1 \\ \partial_2 \end{pmatrix} = \begin{bmatrix} \cos(\theta) & -\frac{1}{r}\sin(\theta) \\ \sin(\theta) & \frac{1}{r}\cos(\theta) \end{bmatrix} \begin{pmatrix} \partial_r \\ \partial_\theta \end{pmatrix} \tag{5.28}$$

and that $\nabla^\perp = Q_2^t \nabla$. Using (5.25), (5.26), and (5.28), it is a simple matter to confirm that

$$\sqrt{a}\nabla\phi_n = \frac{1}{\sqrt{a}}\nabla^\perp\psi_n. \tag{5.29}$$

Boundary singular points are handled in a similar fashion. Now, instead of periodic boundary conditions, the Sturm–Liouville problem (5.5) would require $T(\theta) = 0$ for θ corresponding to a boundary segment in Γ_D, and $T'(\theta) = 0$ for θ corresponding to Γ_N, while problem (5.14) would reverse the roles. It is straightforward to verify that the relationship (5.29) holds for these singular functions as well.

We summarize the above discussion and complete the proof of the decomposition (5.1) in the following theorem.

THEOREM 5.1. *Every* $\mathbf{u} \in \mathbf{W}$ *has a unique decomposition*

$$\mathbf{u} = \mathbf{u}_0 + \sum_{m=1}^{M} \sum_{n=1}^{N_m} b_{m,n} \mathbf{s}_{m,n},$$

where $\mathbf{u}_0 \in \mathbf{W}_S^1$ *and* $\mathbf{s}_{m,n}$, $n = 1, \ldots, N_m$, *are singular functions associated with singular points* $\mathbf{x}_m$, $m = 1, \ldots, M$.

Proof. From Lemma 4.1, we know that $\mathbf{W}_S^1$ is closed in $\mathbf{W}$, that $\mathcal{R}(\widehat{\mathcal{L}}) \subseteq \mathcal{R}(\mathcal{L})$ are both closed in $(L^2(\Omega))^2$, and that both $\mathcal{L}$ and $\widehat{\mathcal{L}}$ are injective. Thus, the codimension of $\mathbf{W}_S^1$ in $\mathbf{W}$ is the same as the codimension of $\mathcal{R}(\widehat{\mathcal{L}})$ in $\mathcal{R}(\mathcal{L})$. By Lemma 4.1, we know that the dimension of $\mathcal{R}(\mathcal{L})^\perp$ is one. We now seek $\mathcal{R}(\widehat{\mathcal{L}})^\perp = \mathcal{N}(\widehat{\mathcal{L}}^*)$. At each singular point $\mathbf{x}_m$, let $\hat{\delta} \in H^2(0, R)$ be a smooth cut-off function and, for each $\alpha_{m,n} \in (0, 1)$, construct functions similar to (5.8) and (5.16) as follows:

$$s_{m,n}^- := \delta_m(r) r^{-\alpha_{m,n}} T_{m,n}(\theta),$$
$$\hat{s}_{m,n}^- := \delta_m(r) r^{-\alpha_{m,n}} \hat{T}_{m,n}(\theta),$$

and define

$$\mathbf{s}_{m,n}^- := (s_{m,n}^-, -\hat{s}_{m,n}^-)^t. \tag{5.30}$$

From (5.29) we see that $\mathbf{s}_{m,n}^- \in \mathcal{D}(\widehat{\mathcal{L}}^*) \setminus \mathcal{D}(\mathcal{L}^*)$ and $\widehat{\mathcal{L}} \mathbf{s}_{m,n}^- \in (L^2(\Omega))^2$. Since $\mathcal{L}^*$ is surjective, we can find $\mathbf{q}_{m,n} \in \mathcal{D}(\mathcal{L}^*)$ such that

$$\mathcal{L}^* \mathbf{q}_{m,n} = -\widehat{\mathcal{L}}^* \mathbf{s}_{m,n}^- \tag{5.31}$$

and set

$$\mathbf{f}_{m,n} = \mathbf{q}_{m,n} + \mathbf{s}_{m,n}^-. \tag{5.32}$$

Clearly, $\mathbf{f}_{m,n} \in \mathcal{N}(\widehat{\mathcal{L}}^*)$.

It is straightforward to show that every element of $\mathcal{N}(\widehat{\mathcal{L}}^*)$ must be of this form, that is, must involve singular functions of both (2.1) and (3.20). Thus, the dimension of $\mathcal{N}(\widehat{\mathcal{L}}^*)$ is exactly equal to the number of such functions plus the one function in $\mathcal{N}(\mathcal{L}^*)$. We complete the proof by noting that the codimension of $\mathcal{N}(\mathcal{L}^*)$ in $\mathcal{N}(\widehat{\mathcal{L}}^*)$ is equal to the codimension of $\mathcal{R}(\widehat{\mathcal{L}})$ in $\mathcal{R}(\mathcal{L})$. □

This decomposition is the basis for the finite element discretization that is developed in the companion paper [3]. We only summarize the basic ideas here. Exponents and coefficients of singular basis functions $\mathbf{s}_{m,n}$ can be computed from the geometry of interfaces adjoining a singular point and the jumps in the coefficient a across these interfaces. Although our theoretical development employed cut-off functions independent of θ, any H^1 cut-off function may be used. We choose cut-off functions that equal one in a fixed region around the singular point and fall off to zero linearly in a small fringe region of width one grid cell.

The singular basis functions are included in the finite element space, together with standard elements, such as linear elements on triangles, that satisfy the interface conditions. Using functional G_0 to solve for the flux, inner products of standard elements with singular basis functions need only be calculated in the fringe region, thus saving a significant amount of work.

6. Conclusions. In this paper we have developed a FOSLS L^2 formulation for diffusion equations with discontinuous coefficients, irregular boundaries, and mixed boundary conditions. In Theorem 3.2, we showed the functional G_α in (3.1) to be coercive and continuous in $\mathbf{W} \times V$ with constants that are P-uniform. We then explored the flux-only functional, G_0 in (3.17), and in Lemma 3.3 and Lemma 4.1 showed that it is coercive and continuous in $\mathbf{W}$ with constants that are also P-uniform. Properties of the scaled div-curl operator (4.1) helped us to prove in Theorem 5.1 that $\mathbf{W}$ can be split into functions that are H^1 in each subdomain plus a finite number of singular basis functions with support in the neighborhood of the singular points.

These results form the theoretical basis for the finite element discretization of $\mathbf{W}$, a rigorous discretization error analysis, and a multilevel method, all of which are presented in the companion paper [3]. Our approach is different from others (see, for example, [9]) in that a rigorous discretization error analysis in the presence of approximate singular basis functions is possible, and a multilevel method can be devised that incorporates singular basis functions on all levels.

REFERENCES

[1] R. E. Alcouffe, A. Brandt, J. J. E. Dendy, Jr., and J. W. Painter, *The multi-grid method for the diffusion equation with strongly discontinuous coefficients*, SIAM J. Sci. Stat. Comput., 2 (1981), pp. 430–454.

[2] M. Berndt, *Adaptive Refinement and the Treatment of Discontinuous Coefficients for Multilevel First-Order System Least Squares (FOSLS)*, Ph.D. thesis, Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO, 1999.

[3] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients: Part* II, SIAM J. Numer. Anal., 43 (2005), pp. 409–436.

[4] P. Bochev, Z. Cai, T. A. Manteuffel, and S. F. McCormick, *Analysis of velocity-flux first-order system least-squares principles for the Navier–Stokes equations: Part* I, SIAM J. Numer. Anal, 35 (1998), pp. 990–1009.

[5] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev. 40 (1998), pp. 789–837.

[6] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *Least-squares methods for the Stokes equations based on a discrete minus one inner product*, J. Comp. Appl. Math., 74 (1996), pp. 155–173.

[7] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[9] S. C. Brenner and L. Y. Sung, *Multigrid methods for the computation of singular solutions and stress intensity factors* II, BIT, 37 (1997), pp. 623–643.

[10] Z. Cai and S. Kim, *A finite element method using singular functions for the Poisson equation: Corner singularities*, SIAM J. Numer. Anal., 39 (2001), pp. 286–299.

[11] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[12] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[13] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for the Stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[14] Z. Cai, T. A. Manteuffel, S. F. McCormick, and J. Ruge, *First-order system LL^* (FOSLL*): Scalar elliptic partial differential equations*, SIAM J. Numer. Anal., 39 (2001), pp. 1418–1445.

[15] T. F. Chen and G. J. Fix, *Least squares finite element simulation of transonic flows*, Appl. Numer. Math., 2 (1986), pp. 399–408.

[16] C. L. Cox and G. J. Fix, *On the accuracy of least squares methods in the presence of corner singularities*, Comput. Math. Appl., 10 (1984), pp. 463–475.

[17] G. Fix and E. Stephan, *Finite Element Methods of the Least Squares Type for Regions with Corners*, Tech. Report 81-41, Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA, 1981.

[18] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer, New York, 1986.

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

[20] M. J. Holst, *Multilevel Methods for the Poisson–Boltzmann Equation*, Ph.D. thesis, Numerical Computing Group, University of Illinois at Urbana-Champaign, Urbana, IL, 1993.

[21] D. C. Jespersen, *A least-square decomposition method for solving elliptic systems*, Math. Comp., 31 (1977), pp. 873–880.

[22] B. N. Jiang and J. Z. Chai, *Least-squares finite element analysis of steady high subsonic plane potential flows*, Acta Mech. Sinica, 1 (1980), pp. 90–93.

[23] B.-N. Jiang and C. L. Chang, *Least-squares finite elements for the Stokes problem*, Comput. Methods Appl. Mech. Engrg., 78 (1990), pp. 297–311.

[24] R. B. Kellogg, *Singularities in interface problems*, in Proceedings of the 2nd Annual Symposium on the Numerical Solution of Partial Differential Equations, B. Hubbard, ed., Academic Press, New York, 1971, pp. 351–400.

[25] T. A. Manteuffel, S. F. McCormick, and G. Starke, *First-order system least-squares for second-order elliptic problems with discontinuous coefficients*, in Proceedings of the Seventh Annual Copper Mountain Conference on Multigrid Methods, N. D. Melson, T. A. Manteuffel, and S. F. McCormick, eds., NASA, Hampton, VA, 1995, pp. 535–550.

[26] P. Neittaanmäki and J. Saranen, *On finite element approximation of the gradient for the solution to Poisson equation*, Numer. Math., 37 (1981), pp. 131–148.

[27] A. I. Pehlivanov and G. F. Carey, *Error estimates for least-squares mixed finite elements*, RAIRO Modél. Math. Anal. Numer., 28 (1994), pp. 499–516.

[28] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[29] T. F. Russell and M. F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media*, in The Mathematics of Reservoir Simulation, R. E. Ewing, ed., Frontiers in Appl. Math. 1, SIAM, Philadelphia, 1983, pp. 35–106.

[30] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[31] W. L. Wendland, *Elliptic Systems in the Plane*, Pitman, London, 1979.

[32] E. Zauderer, *Partial Differential Equations of Applied Mathematics*, 2nd ed., Ser. Pure Appl. Math., John Wiley & Sons, New York, 1988.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 409–436

ANALYSIS OF FIRST-ORDER SYSTEM LEAST SQUARES (FOSLS) FOR ELLIPTIC PROBLEMS WITH DISCONTINUOUS COEFFICIENTS: PART II[*]

MARKUS BERNDT[†], THOMAS A. MANTEUFFEL[‡], AND STEPHEN F. MCCORMICK[‡]

Abstract. First-order system least squares (FOSLS) is a methodology that offers an alternative to standard methods for solving partial differential equations. This paper studies the first-order system least-squares approach for scalar second-order elliptic boundary value problems with discontinuous coefficients. In a companion paper [M. Berndt, T. A. Manteuffel, S. F. McCormick, and G. Starke, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients: Part I*, SIAM J. Numer. Anal., 43 (2005), pp. 386–408], ellipticity of an appropriately scaled least-squares bilinear form is established in a natural norm. For some geometries this ellipticity is independent of the size of the jumps in the coefficients. The occurrence of singularities at interface corners, cross-points, reentrant corners, and irregular boundary points is discussed, and a basis of singular functions with local support around singular points is established. This paper describes a method for including discrete versions of the singular basis functions together with standard finite element spaces in a least-squares format at little additional computational cost. The singular basis functions are constructed to match the jump conditions that arise at interfaces between regions of continuity of the diffusion coefficient. Because these basis functions must be approximated in practice, the resulting discretization is by nature nonconforming. This necessitates the establishment here of a general error estimate for FOSLS L^2 minimization problems discretized by nonconforming finite elements. An advantage of the FOSLS formulation is that this estimate does not involve the consistency error term usually present in bounds for other methods. Based on this general estimate, error bounds are derived for the finite element space that includes singular basis functions. Numerical tests are included that confirm these discretization error bounds. Finally, a multilevel method is developed for solving the discrete system that uses singular basis functions on all levels, and its efficiency is demonstrated by the numerical results.

Key words. least-squares discretization, second-order elliptic problems, discontinuous coefficients, interface problems, singularities, finite elements, multilevel methods

AMS subject classifications. 65N55, 65N30, 65F10

DOI. 10.1137/S003614290342769X

1. Introduction.

In this paper we consider the application of first-order system least squares (FOSLS [11, 12]) to diffusion equations in the plane with jump-discontinuous coefficients:

$$\begin{aligned} -\nabla \cdot (a \nabla p) &= f && \text{in } \Omega, \\ p &= g_{D} && \text{on } \Gamma_D, \\ \mathbf{n} \cdot (a \nabla p) &= g_{N} && \text{on } \Gamma_N. \end{aligned} \tag{1.1}$$

[*] Received by the editors May 7, 2003; accepted for publication (in revised form) August 31, 2004; published electronically June 14, 2005. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sinum/43-1/42769.html

[†] Los Alamos National Laboratory, T-7, Mail Stop B284, Los Alamos, NM 87545 (berndt@lanl.gov). The research of this author was supported by the Department of Energy, under contract W-7405-ENG-36, LA-UR-02-3284.

[‡] Department of Applied Mathematics, Campus Box 526, University of Colorado at Boulder, Boulder, CO 80309-0526 (tmanteuf@boulder.colorado.edu, stevem@boulder.colorado.edu). The research of these authors was supported by the National Science Foundation under grant number DMS-8704169, and by the Department of Energy, Applied Math Program grant DE-FG03-94ER25217.

Here, $a > 0$ is a piecewise smooth function corresponding to some partition of domain $\Omega \subset \Re^2$, with boundary $\partial\Omega = \Gamma_D \cup \Gamma_N$ and outward unit normal $\mathbf{n}$, and the data $f, g_{_D}$, and $g_{_N}$ are appropriately smooth functions. Our focus is on a two-stage FOSLS scheme whose primary aim is to approximate the flux $a\nabla p$.

Studies of the problem of accurate approximation of p by inclusion of special basis functions (cf. [30, 23]) and adaptive refinement (cf. [28]) has been extensive, but the development of efficient multilevel algorithms for the calculation of stress intensity factors is lagging. The only example we are aware of is the full multigrid algorithm for interface problems stemming from cracks, introduced in [7, 9].

The least-squares methodology for systems of first order is by now several decades old and had its first applications in continuum mechanics (see, for example, [20, 32, 21, 25, 15, 22]). Only fairly recently has it produced H^1-equivalent forms to which optimal multigrid solvers have been applied (see, for example, [12]). For a thorough review of the least-squares methodology, see [4] and the references therein.

In the FOSLS formulations developed in [11, 12], the aim was to rewrite the original scalar equation as a first-order system in such a way that its associated least-squares functional has an H^1-equivalent homogeneous part. This equivalence enables simpler finite element discretization methods and ensures that the resulting discrete problem can be solved efficiently by a standard multigrid method. However, because we allow discontinuities in a here, the flux is discontinuous across interfaces and may be singular at some points in the domain. We are therefore led to the development of a special FOSLS L^2 approach for solving (1.1).

In this paper we develop a flux-only FOSLS functional that is continuous and coercive in a scaled space, $H(\operatorname{div} a, \Omega) \cap H(\operatorname{curl} a, \Omega)$, which we denote as $\mathbf{W}$ (see section 2). We denote the space of piecewise H^1 vector valued functions as $H^1_S(\Omega)$ (see section 4). In [2] it was shown that $H^1_S(\Omega) \cap \mathbf{W}$ has finite codimension in $\mathbf{W}$. The singular basis functions, together with $H^1_S(\Omega)\mathbf{W}$, span $\mathbf{W}$ (cf. [2]).

The basic idea behind our special FOSLS scheme is to include singular basis functions in the finite element space and thus accurately model the singular behavior of the flux. These basis functions are constructed so that their action in the weak form involves integration only inside a small fringe region around the singularity. Thus, the additional cost is minimal, yet optimal accuracy is retained.

Alternatives to the approach we develop here are described in detail in [2] and include adding H^1 singular basis functions in standard Galerkin methods to enhance the rate of convergence (cf. [30, 17, 7, 10]) and the use of $H(\operatorname{div})$ conforming finite element spaces with mixed formulations (see [8]) or with FOSLS functionals that are based on $H(\operatorname{div})$ (see [11, 26, 27]). Standard finite element spaces can be used with FOSLS functionals that are weighted to eliminate the overall impact on accuracy of the singular behavior of the flux [18, 17, 24]. Unfortunately, this weighting approach does not provide accurate resolution of the solution close to singularities of the flux, which is the main objective of the approach developed here.

Other alternatives use FOSLS based on inverse norms [6, 5, 13, 3] and FOSLS* [14]. The FOSLS L^2 approach developed here achieves accuracy in the stronger H^1-like norm, which may be preferred in many practical cases.

To estimate discretization accuracy for our special FOSLS scheme, we derive a general error bound for L^2-type FOSLS discretized by nonconforming finite elements. Similar nonconforming estimates for other methods typically involve consistency error terms (cf. [1]), but they are not needed in the FOSLS context. This special property of FOSLS is important because it means that error estimates for nonconforming finite elements may be derived solely from relatively simple interpolation error bounds.

The FOSLS reformulation of (1.1) is derived in section 2. In section 3, the calculation of exponents of singular basis functions is described, and, in section 4, we describe the finite element discretization scheme, complete with singular basis functions. In section 5, the general nonconforming error bound is derived and applied to estimating the accuracy of our augmented basis approach. These estimates are confirmed by the numerical results at the end of section 5. We introduce a multilevel solver in section 6 that is based on coarsening with singular basis functions on all levels. The W-cycle form of this algorithm exhibits typical multigrid convergence behavior, as the numerical results of section 7 also confirm.

2. Problem statement and FOSLS formulation. Assume that $\Omega \subset \Re^2$ is a simply connected polygonal region and that

$$\overline{\Omega} = \bigcup_{j=1}^{J} \overline{\Omega}_j, \tag{2.1}$$

where Ω_j are mutually disjoint open simply connected polygonal regions. Assume also that $\partial\Omega = \Gamma_D \cup \Gamma_N$, Γ_D has positive measure, and Γ_D and Γ_N both consist of a finite number of connected pieces. The case in which $\Gamma_D = \emptyset$ is a simple extension.

Consider the following div-curl first-order system for the scaled flux $\mathbf{u} := \sqrt{a}\nabla p$:

$$\begin{aligned} -\nabla \cdot (\sqrt{a}\mathbf{u}) &= f \quad \text{in } \Omega, \\ \nabla \times (\mathbf{u}/\sqrt{a}) &= 0 \quad \text{in } \Omega, \\ \mathbf{n} \cdot (\sqrt{a}\mathbf{u}) &= 0 \quad \text{on } \Gamma_N, \\ \mathbf{n} \times (\mathbf{u}/\sqrt{a}) &= 0 \quad \text{on } \Gamma_D. \end{aligned} \tag{2.2}$$

(We treat the homogeneous boundary condition case for simplicity. The general case of nonzero g_N and $g_D := \mathbf{n} \times \nabla p$ could be treated by standard lifting or superposition techniques.)

Under the additional smoothness assumptions $a \in C^{1,1}(\Omega)$, $f \in L^2(\Omega)$, and in the absence of reentrant corners and boundary points in which Γ_D and Γ_N meet with interior angle greater than $\pi/2$, we can assert the following [12]: scalar equation (1.1) has a unique solution $p \in H^2(\Omega)$; system (2.2) has a unique solution $\mathbf{u} \in H^1(\Omega)^2$; and the two problems are equivalent in the sense that their solutions correspond according to the relation $\mathbf{u} := \sqrt{a}\nabla p$.

We are interested here in the discontinuous coefficient case, where a is assumed only to be piecewise continuous. Theoretical properties of the first-order system and the corresponding FOSLS functional for this case are studied in the companion paper [2]. In the present paper, we focus on the discretization and multilevel solver for the discrete problem. Problems with reentrant corners and irregular boundary points can be handled in an analogous manner and are omitted for simplicity of presentation.

System (2.2) gives rise to the scaled least-squares functional

$$\mathcal{G}(\mathbf{u}; f) = \left\| (1/\sqrt{a})\nabla \cdot (\sqrt{a}\mathbf{u} + f) \right\|_0^2 + \left\| \sqrt{a}\nabla \times (\mathbf{u}/\sqrt{a}) \right\|_0^2 \tag{2.3}$$

and the associated FOSLS L^2 minimization problem

$$\mathbf{u} = \arg \min_{\mathbf{v} \in \mathbf{W}} \mathcal{G}(\mathbf{v}; f), \tag{2.4}$$

which is well posed on the space

$$\mathbf{W} = \{\mathbf{v} \in H(\operatorname{div} a, \Omega) \cap H(\operatorname{curl} a, \Omega) : \\ \mathbf{n} \cdot (\sqrt{a}\mathbf{v}) = 0 \,\text{on}\, \Gamma_N, \mathbf{n} \times (\mathbf{v}/\sqrt{a}) = 0 \text{ on}\, \Gamma_D\big\}, \tag{2.5}$$

where

$$H(\operatorname{div} a, \Omega) = \left\{\mathbf{v} \in L^2(\Omega)^2 : \nabla \cdot (\sqrt{a}\mathbf{v}) \in L^2(\Omega)\right\},$$
$$H(\operatorname{curl} a, \Omega) = \left\{\mathbf{v} \in L^2(\Omega)^2 : \nabla \times (\mathbf{v}/\sqrt{a}) \in L^2(\Omega)\right\}.$$

We equip $\mathbf{W}$ with the seminorm

$$|\mathbf{u}|_{\mathbf{W}} = \left\|(1/\sqrt{a})\nabla \cdot (\sqrt{a}\mathbf{u})\right\|_0^2 + \left\|\sqrt{a}\nabla \times (\mathbf{u}/\sqrt{a})\right\|_0^2.$$

Note that this is actually a norm because of the assumption that Γ_D has positive measure (see [2, Lemma 3.3]). Note also that $\mathcal{G}$ is trivially $\mathbf{W}$ elliptic in the sense that

$$\mathcal{G}(\mathbf{u}; 0) = |\mathbf{u}|_{\mathbf{W}}^2. \tag{2.6}$$

Minimization problem (2.4) leads to the following variational problem: find $\mathbf{u} \in \mathbf{W}$ such that

$$\mathcal{F}(\mathbf{u}, \mathbf{v}) = \left\langle f/a, \nabla \cdot (\sqrt{a}\mathbf{v})\right\rangle_{0,\Omega} \tag{2.7}$$

for all $\mathbf{v} \in \mathbf{W}$, with

$$\mathcal{F}(\mathbf{u}, \mathbf{v}) = \left\langle (1/a)\nabla \cdot (\sqrt{a}\mathbf{u}), \nabla \cdot (\sqrt{a}\mathbf{v})\right\rangle_{0,\Omega} + \left\langle a\nabla \times (\mathbf{u}/\sqrt{a}), \nabla \times (\mathbf{v}/\sqrt{a})\right\rangle_{0,\Omega}.$$

Suppose that a is piecewise continuous with respect to the partitioning (2.1) of Ω in the maximal sense; that is, a is continuous on Ω_j and no open set $\mathcal{O}_j \supset \Omega_j$ exists for which $a|_{\mathcal{O}_j}$ is continuous, $1 \leq j \leq J$. Under these assumptions, variational problem (2.7) has a unique solution in $\mathbf{W}$ that is also the unique solution of FOSLS minimization problem (2.4) (see [2]).

An edge that lies in the intersection of the closure of two subdomains is called an interface. Points where two interfaces meet are called cross-points. Cross-points, reentrant corners, and irregular boundary points are all potential singular points. Now, the solution $\mathbf{u} \in \mathbf{W}$ of problem (2.2) satisfies certain conditions across the interfaces. Denote by $\mathbf{n}_{\mathcal{I}}$ a unit vector normal to interface $\mathcal{I}$. Then

$$\mathbf{n}_{\mathcal{I}} \cdot (\sqrt{a}\mathbf{u}) \;\text{ and }\; \boldsymbol{\tau}_{\mathcal{I}} \cdot (\mathbf{u}/\sqrt{a}) \;\text{ are continuous a.e. across interfaces.} \tag{2.8}$$

These interface conditions must be true in order for the first two equations in (2.2) to make sense. For the first condition in (2.8) see, for example, [33, Chapter 6.2]. The second condition can be derived analogously.

3. Approximation of singularities.

In [2, section 5], a splitting of $\mathbf{W}$ into a finite-dimensional space spanned by singular functions and locally smooth functions is introduced. This leads to a decomposition of any $\mathbf{u} \in \mathbf{W}$ as

$$\mathbf{u} = \mathbf{u}_0 + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \omega_m \delta_m \mathbf{s}_{m,n}, \tag{3.1}$$

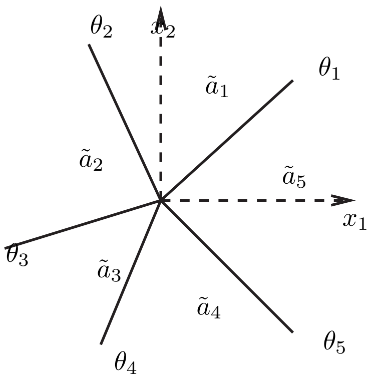


FIG. 3.1. *Cross-point,* $I = 5$.

where $\mathbf{u}_0|_{\Omega_i} \in H^1(\Omega_i)$, δ_m is a cut-off function at a singular point (see Figure 4.2 for an example), and $\mathbf{s}_{m,n}$, $n = 1, \ldots, N_m$, are the singular functions associated with singular point $\mathbf{x}_m$, $m = 1, \ldots, M$. In this paper, our focus is on singular points that are cross-points (see Figure 3.1 for an example). The other types of singular points, described in [2, section 5], can be treated in an analogous fashion.

The exact nature of a singularity at a cross-point can be calculated using a and the geometry of interfaces in the neighborhood of the cross-point. To obtain a simple representation of such singularities (see [23] and [30]), additional constraints on the behavior of a within Ω_i are necessary. We first summarize the results of [2, section 5] and then proceed to describe the numerical method that is used to calculate singular basis functions.

Given a polar coordinate system (r, θ), centered at a cross-point, we recall that $a \in C^{1,1}(\Omega_i)$, for each subdomain Ω_i, and assume that a satisfies

$$\lim_{r \to 0} a_\theta = 0, \qquad \lim_{r \to 0} r a_r = 0. \tag{3.2}$$

The task of finding a representation for the singularity at a cross-point reduces to finding solutions of the Sturm–Liouville eigenvalue problem (see [2, section 5] for a detailed derivation)

$$-(\tilde{a}T')' = \tilde{a}\alpha^2 T \ \text{ on } \ [0, 2\pi), \tag{3.3}$$

where $\tilde{a}_i = \lim_{r \to 0} a(r, \theta)$ in Ω_i.

Each interface that adjoins the cross-point is characterized by the angle of its tangent at the cross-point with the x_1-axis. Denote by I the number of interfaces adjoining at a given cross-point, and by θ_i, $i = 1, \ldots, I$, the angles their tangents make with the x_1-axis (see Figure 3.1 for an example with $I = 5$). Assume that these interface angles are ordered such that $\theta_i < \theta_{i+1}$, $i = 1, \ldots, I - 1$. Let $\theta_{I+1} = \theta_1$.

Eigenfunctions of (3.3) have the form

$$T_n(\theta) = \lambda_{n,i} \cos(\alpha_n \theta) + \mu_{n,i} \sin(\alpha_n \theta), \tag{3.4}$$

for $\theta \in (\theta_i, \theta_{i+1})$, where α_n^2 is the associated eigenvalue. According to [2, Theorem 5.1], we must calculate all eigenvalues $0 < \alpha_n^2 < 1$ and associated eigenfunctions of (3.3), to obtain all singular functions

$$\mathbf{s}_{m,n} = \sqrt{a}\nabla r^{\alpha_n} T_n(\theta) = \sqrt{a}\alpha_n r^{\alpha_n - 1} \begin{pmatrix} \lambda_{ni}\sin((\alpha_n-1)\theta) + \mu_{ni}\cos((\alpha_n-1)\theta) \\ \lambda_{ni}\cos((\alpha_n-1)\theta) - \mu_{ni}\sin((\alpha_n-1)\theta) \end{pmatrix}, \tag{3.5}$$

for $\theta \in (\theta_i,\, \theta_{i+1})$, where λ_{ni} and μ_{ni} are constant inside each Ω_i. Note also that $\mathbf{s}_{m,n} = \sqrt{a}\nabla\sigma_n$ with

$$\sigma_n(r,\theta) = r^{\alpha_n}(\lambda_{ni}\sin\alpha_n\theta + \mu_{ni}\cos\alpha_n\theta) \quad \text{for} \quad \theta \in (\theta_i,\, \theta_{i+1}). \tag{3.6}$$

Remark 1. In [2], a representation of the singular functions is used that differs slightly from (3.5). It is easy to show that the two representations are equivalent.

For convenience we will now drop the subscript n where the meaning is apparent. The exponent α and the coefficients (λ_i, μ_i) can be determined by enforcing continuity of both $T(\theta)$ and $\tilde{a}T'(\theta)$ across interfaces. To obtain first approximations to the eigenvalues α^2, we discretize eigenvalue problem (3.3) and solve the resulting algebraic generalized eigenvalue problem. Note that we are primarily interested in the smallest values of α^2, that is, $0 \leq \alpha^2 \leq 1$. The eigenvectors associated with these small eigenvalues are well approximated using a fairly coarse discretization. Values of α that are obtained in this way are used as starting values of a secant iteration that is based on the following idea.

Interface conditions (2.8) give rise to a $2I \times 2I$ nonlinear system of equations for α, λ_i, and μ_i, $i = 1, \ldots, I$, which can be written in the compact form

$$M(\alpha)(\lambda_1, \mu_1, \ldots, \lambda_I, \mu_I)^t = \underline{0}. \tag{3.7}$$

A nontrivial solution exists only when $M(\alpha)$ is singular, that is, when $\det M(\alpha) = 0$. To find roots of $\det M(\alpha)$, we use a secant iteration with starting values obtained from the solution of the discretized Sturm–Liouville eigenvalue discussed in above.

Suppose we compute an approximation $\tilde{\alpha} = \alpha + \eta$. To estimate η, we find $\underline{x}$ that has norm one and minimizes $\|M(\tilde{\alpha})\underline{x}\|$. Thus, $\underline{x}$ is a right singular vector of $M(\tilde{\alpha})$ and $\|M(\tilde{\alpha})\underline{x}\| = \sigma_n$, the smallest singular value of $M(\tilde{\alpha})$, which is easily computed because $M(\tilde{\alpha})$ is of small dimension. Moreover,

$$M(\tilde{\alpha})\underline{x} = \sigma_n \underline{r}, \tag{3.8}$$

where $\underline{r}$ is the left singular vector of $M(\tilde{\alpha})$ and $\|\underline{r}\| = 1$.

Let α be the exact value, and let the inexact $\tilde{\alpha} = \alpha + \eta$. Then we have the matrix expansion

$$M(\alpha + \eta) = M(\alpha) + \eta M'(\alpha + \hat{\eta}) \simeq M(\alpha) + \eta M'(\alpha + \eta),$$

where $\hat{\eta} \in (0, \eta)$. We can easily compute $M'(\alpha + \eta)$. Since $M(\alpha)$ is singular, we know that the distance from $M(\alpha + \eta)$ to $M(\alpha)$ in the Frobenius norm is larger than the smallest singular value of $M(\alpha + \eta)$; that is,

$$\sigma_n \leq \|M(\alpha + \eta) - M(\alpha)||_F \simeq \eta\|M'(\alpha + \eta)\|_F.$$

Thus, we get a lower bound on η. Conversely, we see that as $\eta \to 0$, we can get a bound on $\sigma_n \to 0$.

In conclusion, the coefficients become accurate at the same rate as alpha becomes accurate. Also, we can determine an approximate lower bound on the accuracy of alpha by computing the singular values of $M(\tilde{\alpha})$. If the error in alpha is too big, we do more computational work. (In [2, (5.12)], $M(\alpha)$ is scaled such that the dependence on a is lumped into 2×2 block diagonal terms D_i that have no dependence on α but depend on the ratio a_i/a_{i-1}. In this paper, we use a slightly different scaling, where $D_i = \text{diag}(a_{i-1}, -a_i)$.)

The calculations that are described in this section are not very costly, since typically, the number of interfaces adjoining a cross-point is very small, and the number of eigenvalues of the Sturm–Liouville problem (3.3) in which we are interested is of order $O(1)$ (see [23]).

4. Finite element discretization. For simplicity, we have assumed that domain Ω and its subdomains Ω_j are polygonal, which allows the geometry of the discontinuities of a to be resolved exactly using a triangular mesh. Let $\mathcal{T}_h$ be a quasiuniform triangulation (see, for example, [8, Definition 4.4.13]) constructed so that no element cuts across any interface (i.e., each element is contained in just one subdomain). Our discretization method is based on the decomposition introduced in [2] that isolates the singular functions from the piecewise H^1 functions. To this end, let $\delta_m \in H^1(\Omega)$ denote any given"cut-off" function that has value one in a small area about cross-point m and values that taper to 0 in a small outer "fringe," $1 \le m \le M$. (See Figure 4.2.) Let $\mathbf{s}_{m,n}$ be the nth singular basis function at cross-point m, $1 \le n \le N_m, 1 \le m \le M$. Then $\delta_m \mathbf{s}_{m,n} \in W$, provided that δ_m has support inside Ω and all interfaces inside the fringe and platform of δ_m are straight lines. We will be more specific about $\mathbf{s}_{m,n}$ below. Defining the "split" space of piecewise H^1 functions by

$$H^1_S(\Omega) := \left\{ \mathbf{u} \in (L^2(\Omega))^2 \ : \ \mathbf{u}|_{\Omega_j} \in (H^1(\Omega_j))^2, \ j = 1, \dots, J \right\} \tag{4.1}$$

and letting $\mathbf{W}^1_S := \mathbf{W} \cap H^1_S(\Omega)$, then our decomposition is given by

$$\mathbf{W} = \mathbf{W}^1_S \ \oplus \ \text{span}\left\{ \delta_m \mathbf{s}_{m,n} : 1 \le m \le M, 1 \le n \le N_m \right\}, \tag{4.2}$$

(cf. [2], Theorem 5.1). At interior nodes of the subdomains Ω_j, we use standard piecewise linear nodal basis functions, whose coefficients at the nodal values are the unknowns. (We will add certain quadratic basis functions shortly.) For vertices that lie on interfaces, we use piecewise linear basis functions that satisfy interface conditions (2.8) exactly and are scaled to have a maximum of one (see also Figure 4.1).

Singular components $\delta_m \mathbf{s}_{m,n}$ are discretized by choosing a discrete cut-off function $\delta_m = \delta^h_m$ and replacing $\mathbf{s}_{m,n}$ by a discrete approximation $\tilde{\mathbf{s}}_{m,n}$, as described in section 3. Each cross-point m is surrounded by the support of its cut-off function,

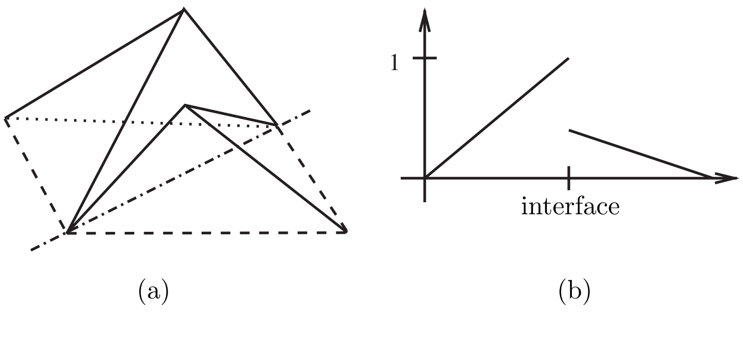


FIG. 4.1. *A discontinuous linear basis function in* 3D *view* (a) *and in side view* (b).

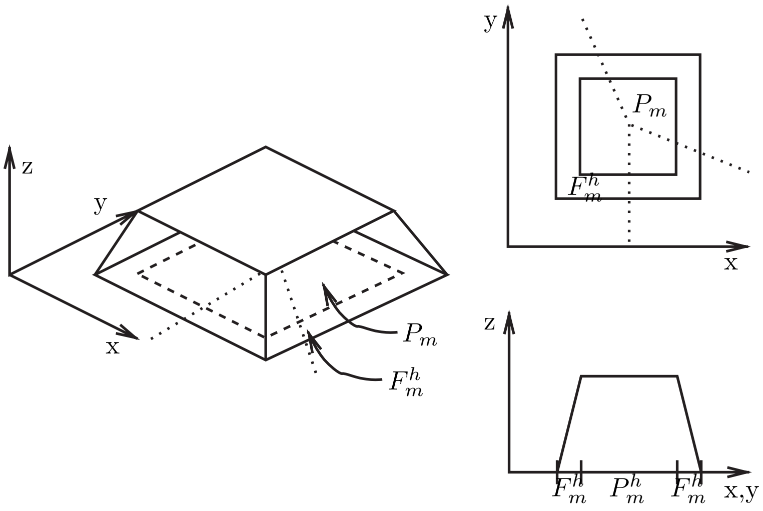


FIG. 4.2. *The cut-off function δ_m^h centered at a cross-point (P_m = platform, F_m^h = fringe, dotted lines are interfaces).*

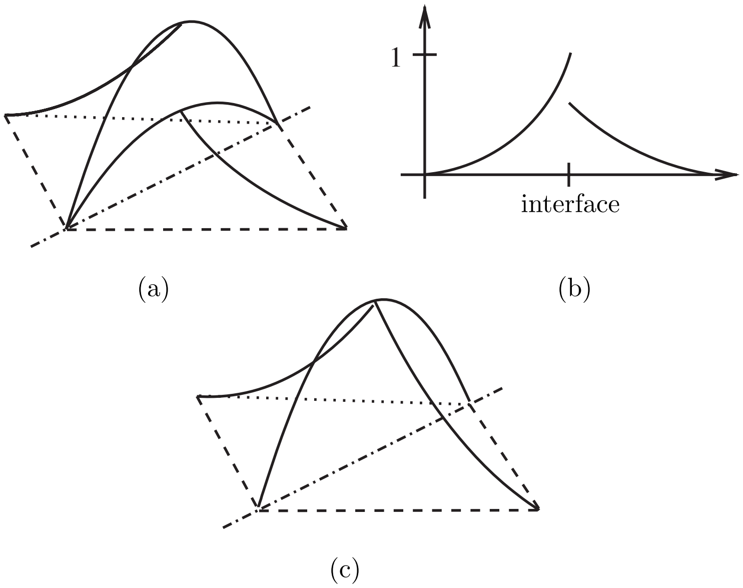
which consists of a platform P_m and outer fringe F_m^h consisting of one outer ring of level h triangles. The platform also consists of level h triangles, but it is otherwise fixed in size. The supports $P_m \cup F_m^h$ are constructed at each cross-point to be large enough to obtain a reasonable approximation to the singular functions but small enough to ensure that they do not intersect with each other. Cut-off function δ_m^h is then defined so that $\delta_m^h|_\tau$ is linear for all $\tau \subset F_m^h$ and has value 1 inside its platform. See Figure 4.2.

Denoting by $\mathcal{F}_\tau(\cdot,\cdot)$ the $\mathcal{F}$ inner product evaluated on the element τ, we have

$$\mathcal{F}_\tau(\delta_m^h \mathbf{s}_{m,n}, \mathbf{v}) = 0 \quad \text{for all} \;\; \tau \in P_m,$$

since $\delta_m^h = 1$ inside P_m. This implies that, for elements inside the platforms, entries in element stiffness matrices that involve singular basis functions are zero. Only elements in the fringes have element stiffness matrices that have contributions from integration of singular basis functions. To evaluate these fringe integral terms, we use two-dimensional Gaussian quadrature of order high enough to ensure that it does not corrupt the discretization error estimates we obtain in the following sections. (Recall that the singular functions are smooth in the fringe.) Outside platforms and fringes, there are no contributions from singular basis functions. In conclusion, each singular basis function need only be numerically integrated on the small number of elements that comprise the fringe of its cut-off function.

To control the computational work of integrating the singular basis functions, we have limited the fringes to width h, which reduces discretization accuracy. To avoid this loss, we introduce quadratic "bubble-like" basis functions in the fringes, with supports consisting of two triangles that share an edge within the fringe. Within each triangle, the quadratic function is defined to be the product of a linear function that is zero on one of the nonshared edges and another that is zero on the other nonshared edge. When the triangle pair is in a single Ω_j, the basis function is scaled to be 1 at the midpoint of the shared edge (see Figure 4.3(c)). If, instead, the edge coincides with an interface, the discontinuous basis function is such that it satisfies the interface



(a) (b)

(c)

Fig. 4.3. *A discontinuous quadratic basis function in* 3*D view* (a) *and in side view* (b); *a continuous quadratic basis function* (c).

conditions (2.8) exactly and has a maximum of one. See Figure 4.3(a) and (b) for a schematic.

In the next section, we derive an error estimate that illustrates the necessity for such an increase in discretization order inside the fringes (see the proof of Theorem 5.1).

Our discretization is, thus, defined by the space $\mathbf{W}^h$ of elements of the form

$$\mathbf{u}^h = \mathbf{u}_L^h + \mathbf{u}_Q^h + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \omega_{m,n} \delta_m^h \tilde{\mathbf{s}}_{m,n}, \tag{4.3}$$

where is $\mathbf{u}_L^h$ is piecewise linear (with respect to $\mathcal{T}^h$) and continuous in Ω_j, $\mathbf{u}_Q^h$ is piecewise quadratic and continuous in Ω_j but with support contained in the fringe, and $\tilde{\mathbf{s}}_{m,n}$ is an approximation to $\mathbf{s}_{m,n}$. The discrete problem corresponding to (2.7) is then as follows: Find $\mathbf{v}^h \in \mathbf{W}^h$ such that

$$\mathcal{F}(\mathbf{u}^h, \mathbf{v}^h) = \left\langle f/a, \nabla \cdot (\sqrt{a}\mathbf{v}^h) \right\rangle_{0,\Omega} \tag{4.4}$$

for all $\mathbf{v}^h \in \mathbf{W}^h$.

Both $\mathbf{u}_L^h$ and $\mathbf{u}_Q^h$ satisfy the interface conditions exactly, but the singular function approximations $\tilde{\mathbf{s}}_{m,n}$ do not, because $\tilde{\alpha}$ is not exact. This means that the discrete space is generally nonconforming: $\mathbf{W}^h \not\subset \mathbf{W}$. Thus, standard theory for discretization accuracy does not apply, and we are left to develop our own estimates.

5. Error estimates. We begin by establishing an error estimate for the case of a conforming subspace, where the singular basis functions are assumed to be known exactly. To cover the practical case, where approximate singular basis functions are used, we then derive a general error estimate for FOSLS L^2 formulations with nonconforming finite elements and apply it to the case of nonconforming singular basis functions.

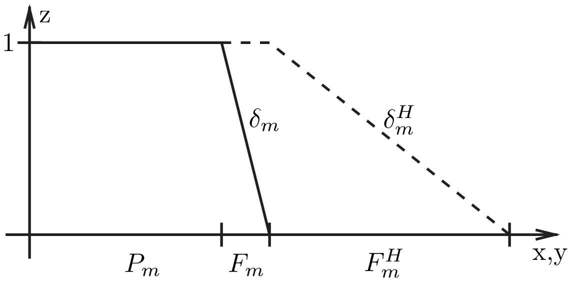


FIG. 5.1. *Side view of cut-off functions δ_m^h and δ_m^H.*

5.1. The conforming case: $\mathbf{W}^h \subset \mathbf{W}$. Let $2H_m$ be less than the shortest distance from P_m to the nearest other platform or boundary. Let F_m^H be a fringe of width H_m, and let δ_m^H be the associated cut-off function (see Figure 5.1). Using the decomposition of $\mathbf{W}$ given in (4.2), write the solution of variational problem (4.4) as

$$\mathbf{u} = \mathbf{u}_0 + \sum_{m=1}^{M}\sum_{n=1}^{N} \omega_{m,n}\delta_m^H \mathbf{s}_{m,n}, \tag{5.1}$$

where $\mathbf{u}_0 \in \mathbf{W}_S^1$. Now we state the error estimate for the conforming case.

THEOREM 5.1 (estimate for conforming $\mathbf{W}^h$). *Assume that $\mathbf{W}^h \subset \mathbf{W}$. Let $\mathbf{u} \in \mathbf{W}$ denote the solution of variational problem* (2.7) *and $\mathbf{u}^h \in \mathbf{W}^h$ the solution of discrete variational problem* (4.4). *Let $2H_m$ be less than the distance between P_m and the closest other platform or boundary, and assume that $0 < h < H_m$, for $m = 1, \ldots, M$. Then*

$$\left|\mathbf{u} - \mathbf{u}^h\right|_{\mathbf{W}} \le C_1 h^{\sigma-1} \left|\mathbf{u}_0\right|_{\sigma,S} + C_2 h \sup_{m,n} \left|\omega_{m,n}\right|, \tag{5.2}$$

where $\sigma \in (1, 2]$ depends on the smoothness of $\mathbf{u}_0$, $\omega_{m,n}$ are the coefficients in (5.1), *and the constants $C_1, C_2 > 0$ are independent of h.*

Proof. Since $\mathbf{W}^h \subset \mathbf{W}$, then by (4.4), $\mathbf{u}^h$ satisfies

$$\left|\mathbf{u} - \mathbf{u}^h\right|_{\mathbf{W}} = \inf_{\mathbf{v}^h \in \mathbf{W}^h} \left|\mathbf{u} - \mathbf{v}^h\right|_{\mathbf{W}} \le \left|\mathbf{u} - \mathbf{w}^h\right|_{\mathbf{W}}, \tag{5.3}$$

for any particular $\mathbf{w}^h \in \mathbf{W}^h$. As in (5.1), denote by δ_m^H the cut-off function that is one in $P_m \cup F_m$ and drops to zero linearly in the extended fringe F_m^H around $P_m \cup F_m^h$ of width H_m. See Figure 5.1 for a side view of this function. Note that all F_m^H, $m = 1, \ldots, M$, are mutually disjoint. Let $I^h : C(\Omega) \to \mathbf{W}$ denote a linear nodal interpolant operator outside $\bigcup_{m=1}^M F_m^h$ and a quadratic nodal interpolant operator in $\bigcup_{m=1}^M F_m^h$. Coefficients of linear basis functions are obtained by function evaluation at their vertex, whereas coefficients of quadratic basis functions are obtained by evaluating the difference of the function itself and the linear interpolant. We now write

$$\mathbf{u} = \mathbf{u}_0 + \sum_{m=1}^{M}\sum_{n=1}^{N_m} \omega_{m,n}(\delta_m^H - \delta_m^h)\mathbf{s}_{m,n} + \sum_{m=1}^{M}\sum_{n=1}^{N_m} \omega_{m,n}\delta_m^h \mathbf{s}_{m,n}, \tag{5.4}$$

where $\mathbf{u}_0 \in \mathbf{W}_S^1$ does not depend on h. Define

$$\mathbf{w}^h := I^h \left(\mathbf{u}_0 + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \omega_{m,n} \psi_m \mathbf{s}_{m,n} \right) + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \omega_{m,n} \delta_m^h \mathbf{s}_{m,n}, \tag{5.5}$$

where $\psi_m := \delta_m^H - \delta_m^h$. Substituting (5.4) and (5.5) into (5.3) and using the triangle inequality, we obtain

$$\left| \mathbf{u} - \mathbf{w}^h \right|_{\mathbf{W}} \leq \left| \mathbf{u}_0 - I^h \mathbf{u}_0 \right|_{\mathbf{W}} + \sum_{m=1}^{M} \sum_{n=1}^{N_m} |\omega_{m,n}| \left| \psi_m \mathbf{s}_{m,n} - I^h \psi_m \mathbf{s}_{m,n} \right|_{\mathbf{W}}. \tag{5.6}$$

Since $\mathbf{u}_0 \in \mathbf{W}_S^1$ does not depend on h, we can use [19, Theorem 4.4.20] to estimate the first term in (5.6):

$$\left| \mathbf{u}_0 - I^h \mathbf{u}_0 \right|_{\mathbf{W}} = \left| \mathbf{u}_0 - I^h \mathbf{u}_0 \right|_{1,S} \leq \tilde{C}_1 h^{\sigma - 1} \left| \mathbf{u}_0 \right|_{\sigma,S}. \tag{5.7}$$

Here, $1 < \sigma \leq 2$ depends on the smoothness of $\mathbf{u}_0$.

Since $\psi_m s_{m,n} \in \mathbf{W}_S^1$, we have

$$\left| (I - I^h) \psi_m s_{m,n} \right|_{\mathbf{W}}^2 \leq \sum_{\tau \in F_m^h} \left| (I - I^h) \psi_m s_{m,n} \right|_{1,\tau}^2 + \sum_{\tau \in F_m^H \backslash F_m^h} \left| (I - I^h) \psi_m s_{m,n} \right|_{1,\tau}^2.$$

Since the finite element space includes quadratics on F_m^h, the two terms on the right-hand side satisfy

$$\left| (I - I^h) \psi_m s_{m,n} \right|_{1,\tau} \leq \begin{cases} ch^2 \| \psi_m s_{m,n} \|_{3,\tau} & \text{for } \tau \in F_m^h, \\ ch \| \psi_m s_{m,n} \|_{2,\tau} & \text{for } \tau \in F_m^H \backslash F_m^h. \end{cases}$$

Using the inverse inequality and noting that ψ is linear, we have

$$\| \psi_m s_{m,n} \|_{3,\tau} \leq \frac{c}{h} \| s_{m,n} \|_{3,\tau} \qquad \text{for } \tau \in F_m^h$$

and

$$\| \psi_m s_{m,n} \|_{2,\tau} \leq \frac{c}{H_m} \| s_{m,n} \|_{2,\tau} \qquad \text{for } \tau \in F_m^H \backslash F_m^h.$$

Putting this all together, we have

$$\left| (I - I^h) \psi_m s_{m,n} \right|_{\mathbf{W}}^2 \leq ch^2 \| s_{m,n} \|_{3,F_m^h}^2 + \frac{c}{H_m} h^2 \| s_{m,n} \|_{2,F_m^H}^2. \tag{5.8}$$

Now, using (5.8) and (5.7) in (5.6), we obtain the estimate

$$\left| \mathbf{u} - \mathbf{w}^h \right|_{\mathbf{W}} \leq \tilde{C}_1 h^{\sigma-1} \left| \mathbf{u}_0 \right|_{\sigma,S} + h \sum_{m=1}^{M} C_m \sum_{n=1}^{N_m} |\omega_{m,n}| \, \| \mathbf{s}_{m,n} \|_{3,S,F_m^H} \tag{5.9}$$

$$\leq C_1 h^{\sigma-1} \left| \mathbf{u}_0 \right|_{\sigma,S} + C_2 h \sup_{m,n} |\omega_{m,n}|. \tag{5.10}$$

This completes the proof. □

Remark 2. If the right-hand side of (2.2) is sufficiently smooth, then adding all singular functions of the form (3.5) to the finite element space for which $\alpha_n \in (0, 2]$ yields a bound of $O(h)$ in (5.2) (see [7]).

In practice, subspace $\mathbf{W}^h$ contains only approximate singular basis functions, which implies $\mathbf{W}^h \not\subset \mathbf{W}$. In the next section, we derive an error estimate for a general nonconforming finite element space that is used in section 5.3.

5.2. A general error estimate for FOSLS L^2 formulations with nonconforming finite elements. In this section, we depart from the framework and notation that were introduced in the previous section. We introduce a general methodology for derivation of error estimates for FOSLS L^2 formulations that use nonconforming finite element spaces. First consider a general FOSLS L^2 functional

$$\mathcal{G}(\mathbf{u};\mathbf{f}) := \|\mathcal{L}\mathbf{u}-\mathbf{f}\|^2_{0,\Omega}\,, \tag{5.11}$$

where Ω is a bounded open domain, $\mathbf{u}$ an element of a Hilbert space $\boldsymbol{\mathcal{W}}$, $\mathbf{f} \in (L^2(\Omega))^k$, and $\mathcal{L}$ a first-order differential operator. This gives rise to the FOSLS L^2 minimization problem

$$\mathbf{u} = \arg\min_{\mathbf{v}\in\boldsymbol{\mathcal{W}}} \mathcal{G}(\mathbf{v};\mathbf{f}) \tag{5.12}$$

and its variational form: Find $\mathbf{u} \in \boldsymbol{\mathcal{W}}$ such that

$$\mathcal{F}(\mathbf{u},\mathbf{v}) = \ell(\mathbf{v}), \tag{5.13}$$

for all $\mathbf{v} \in \boldsymbol{\mathcal{W}}$, with

$$\begin{aligned}\mathcal{F}(\mathbf{u},\mathbf{v}) &= \langle\mathcal{L}\mathbf{u},\mathcal{L}\mathbf{v}\rangle_{0,\Omega}\,,\\ \ell(\mathbf{v}) &= \langle\mathbf{f},\mathcal{L}\mathbf{v}\rangle_{0,\Omega}\,.\end{aligned}$$

We assume that bilinear functional $\mathcal{F}$ is $\boldsymbol{\mathcal{W}}$-elliptic with respect to a norm $\|\cdot\|_{\boldsymbol{\mathcal{W}}}$ in the sense that respective continuity and coercivity constants C_{cont} and C_{coer} exist, for which

$$\begin{aligned}\mathcal{F}(\mathbf{u},\mathbf{v}) &\le C_{cont}\,\|\mathbf{u}\|_{\boldsymbol{\mathcal{W}}}\,\|\mathbf{v}\|_{\boldsymbol{\mathcal{W}}}\,,\\ C_{coer}\,\|\mathbf{u}\|^2_{\boldsymbol{\mathcal{W}}} &\le \mathcal{F}(\mathbf{u},\mathbf{u}),\end{aligned}$$

for all $\mathbf{u},\mathbf{v} \in \boldsymbol{\mathcal{W}}$.

Let $\{\Omega_j\}_{j=1,\dots,J}$ be an open partitioning of Ω such that all Ω_j are mutually disjoint and $\bigcup_{j=1}^J \overline{\Omega}_j = \overline{\Omega}$. Let $\boldsymbol{\mathcal{W}}^h$ be a finite element space for which the restriction of the operator $\mathcal{L}$ to the subdomain Ω_j is well defined. Define approximate bilinear form $\mathcal{F}^{nc}$ by

$$\mathcal{F}^{nc}(\mathbf{u}^h,\mathbf{v}^h) := \sum_{j=1}^J \left\langle\mathcal{L}\mathbf{u}^h,\mathcal{L}\mathbf{v}^h\right\rangle_{0,\Omega_j} \tag{5.14}$$

and approximate linear functional ℓ^{nc} by

$$\ell^{nc}(\mathbf{v}^h) := \sum_{j=1}^J \left\langle\mathbf{f},\mathcal{L}\mathbf{v}^h\right\rangle_{0,\Omega_j}\,, \tag{5.15}$$

for $\mathbf{u}^h,\mathbf{v}^h \in \boldsymbol{\mathcal{W}}^h$. Assume that $\mathcal{F}^{nc}$ is uniformly $\boldsymbol{\mathcal{W}}^h$-elliptic with respect to a norm $\|\cdot\|_{\boldsymbol{\mathcal{W}}^h}$, with respective continuity and coercivity constants $\tilde{C}_{cont}$ and $\tilde{C}_{coer}$. This ensures that the following approximate variational problem has a unique solution: Find $\mathbf{u}^h \in \boldsymbol{\mathcal{W}}^h$ such that

$$\mathcal{F}^{nc}(\mathbf{u}^h,\mathbf{v}^h) = \ell^{nc}(\mathbf{v}^h) \tag{5.16}$$

for all $\mathbf{v}^h \in \boldsymbol{\mathcal{W}}^h$.

THEOREM 5.2. *Consider a family of discrete problems that stem from a FOSLS L^2 minimization problem, whose associated approximate bilinear forms are uniformly $\boldsymbol{\mathcal{W}}^h$-elliptic. Then there exists a constant C, independent of the subspace $\boldsymbol{\mathcal{W}}^h$, such that*

$$\left\|\mathbf{u}-\mathbf{u}^h\right\|_{\boldsymbol{\mathcal{W}}^h} \leq C \inf_{\mathbf{v}^h \in \boldsymbol{\mathcal{W}}^h} \left\|\mathbf{u}-\mathbf{v}^h\right\|_{\boldsymbol{\mathcal{W}}^h} . \tag{5.17}$$

Proof. Let $\mathbf{v}^h \in \boldsymbol{\mathcal{W}}^h$ be arbitrary. Using $\boldsymbol{\mathcal{W}}^h$-ellipticity of $\mathcal{F}^{nc}$ and the definition of the approximate variational problem (5.16), we have

$$\begin{aligned} \tilde{C}_{coer} \left\|\mathbf{u}^h-\mathbf{v}^h\right\|^2_{\boldsymbol{\mathcal{W}}^h} &\leq \mathcal{F}^{nc}(\mathbf{u}^h-\mathbf{v}^h, \mathbf{u}^h-\mathbf{v}^h) \\ &= \mathcal{F}^{nc}(\mathbf{u}-\mathbf{v}^h, \mathbf{u}^h-\mathbf{v}^h) + \mathcal{F}^{nc}(\mathbf{u}^h-\mathbf{u}, \mathbf{u}^h-\mathbf{v}^h) \\ &= \mathcal{F}^{nc}(\mathbf{u}-\mathbf{v}^h, \mathbf{u}^h-\mathbf{v}^h) + \ell^{nc}(\mathbf{u}^h-\mathbf{v}^h) - \mathcal{F}^{nc}(\mathbf{u}, \mathbf{u}^h-\mathbf{v}^h). \end{aligned} \tag{5.18}$$

Using (5.14), (5.15), and the Cauchy–Schwarz inequality for any $\mathbf{w}^h \in \boldsymbol{\mathcal{W}}^h$, we have

$$\left|\ell^{nc}(\mathbf{w}^h) - \mathcal{F}^{nc}(\mathbf{u}, \mathbf{w}^h)\right| \leq \sum_{j=1}^{J} \left|\left\langle \mathbf{f}-\mathcal{L}\mathbf{u}, \mathcal{L}\mathbf{w}^h \right\rangle_{0,\Omega_j}\right| \leq \sum_{j=1}^{J} \|\mathbf{f}-\mathcal{L}\mathbf{u}\|_{0,\Omega_j} \left\|\mathcal{L}\mathbf{w}^h\right\|_{0,\Omega_j} .$$

Since $\mathbf{u} \in \boldsymbol{\mathcal{W}}$ is the solution of minimization problem (5.12), we deduce

$$\|\mathcal{L}\mathbf{u}-\mathbf{f}\|^2_{0,\Omega_j} \leq \mathcal{G}(\mathbf{u};\mathbf{f}) = 0$$

for all $j = 1, \ldots, J$, which implies

$$\ell^{nc}(\mathbf{w}^h) - \mathcal{F}^{nc}(\mathbf{u}, \mathbf{w}^h) = 0. \tag{5.19}$$

Choosing $\mathbf{w}^h = \mathbf{u}^h - \mathbf{v}^h$ in (5.18) and appealing to the continuity of $\mathcal{F}^{nc}$, we thus have

$$\tilde{C}_{coer} \left\|\mathbf{u}^h-\mathbf{v}^h\right\|_{\boldsymbol{\mathcal{W}}^h} \leq \tilde{C}_{cont} \left\|\mathbf{u}-\mathbf{v}^h\right\|_{\boldsymbol{\mathcal{W}}^h} . \tag{5.20}$$

The triangle inequality and (5.20) imply

$$\left\|\mathbf{u}-\mathbf{u}^h\right\|_{\boldsymbol{\mathcal{W}}^h} \leq \left(\frac{C_{cont}}{C_{coer}} + 1\right) \left\|\mathbf{u}-\mathbf{v}^h\right\|_{\boldsymbol{\mathcal{W}}^h} ,$$

which completes the proof. □

Remark 3. Uniform $\boldsymbol{\mathcal{W}}^h$ coercivity must be established before Theorem 5.2 can be applied.

Theorem 5.2 implies that, for a FOSLS L^2 formulation that is discretized using a nonconforming finite element space, an estimate analogous to Cea's lemma (cf. [16, Theorem 13.1]) holds. Inequality (5.17) does not involve a consistency error term, as in the fundamental estimate for nonconforming finite elements (see [1], commonly referred to as Strang's second lemma). In common use is a patch test (cf. [16, p. 221]), which determines whether this consistency error term approaches zero as $h \to 0$. The corollary shows that, in the FOSLS L^2 context, elements that do not satisfy such conditions can be used, provided that uniform $\boldsymbol{\mathcal{W}}^h$-ellipticity can be established for $\mathcal{F}^{nc}$ and an error estimate based solely on interpolation theory can be derived.

5.3. An error estimate in the nonconforming space $\mathbf{W}^h \not\subseteq \mathbf{W}$. The finite element space $\mathbf{W}^h$ contains singular functions that would ideally model the singular behavior of the solution at cross-points exactly. However, the exponents and coefficients of these singular functions can be calculated only approximately, which means generally that $\mathbf{W}^h \not\subseteq \mathbf{W}$.

We will use Theorem 5.2 to derive an error estimate for the nonconforming case. However, as noted in Remark 3, we first must establish uniform coercivity of the approximate bilinear form. Define the nonconforming functional by

$$G^{nc}(\mathbf{u}; f) = \sum_{i=1}^{J} \left\| \frac{1}{\sqrt{a}} (\nabla \cdot \sqrt{a}\mathbf{u} + f) \right\|_{0,\Omega_i}^2 + \left\| \frac{1}{\sqrt{a}} \nabla \times \left(\frac{\mathbf{u}}{\sqrt{a}} \right) \right\|_{0,\Omega_i}^2 , \tag{5.21}$$

the associated bilinear form

$$\mathcal{F}^{nc}(\mathbf{u}, \mathbf{v}) = \sum_{i=1}^{K} \left\langle \frac{1}{\sqrt{a}} \nabla \cdot \sqrt{a}\mathbf{u}, \frac{1}{\sqrt{a}} \nabla \cdot \sqrt{a}\mathbf{v} \right\rangle_{\Omega_i}^2 + \left\langle \sqrt{a}\nabla \times \frac{1}{\sqrt{a}\mathbf{u}}, \sqrt{a}\nabla \times \frac{1}{\sqrt{a}}\mathbf{v} \right\rangle_{\Omega_i}^2 , \tag{5.22}$$

and the seminorm

$$\mathcal{F}^{nc}(\mathbf{u}, \mathbf{u}) := |\mathbf{u}|_{\mathbf{W}^h}^2 . \tag{5.23}$$

We show in Appendix A.2 that $\mathcal{F}^{nc}$ is uniformly coercive in the norm $\|\mathbf{u}\|_{\mathbf{W}^h}^2 := \|\mathbf{u}\|_0^2 + |\mathbf{u}|_{\mathbf{W}^h}^2$. Thus, in what follows we use the seminorm. Now Theorem 5.2 implies that

$$\mathcal{F}^{nc}(\mathbf{u} - \mathbf{u}^h, \mathbf{u} - \mathbf{u}^h)^{1/2} = \left|\mathbf{u} - \mathbf{u}^h\right|_{\mathbf{W}^h} \leq C \inf_{\mathbf{v}^h \in \mathbf{W}^h} \left|\mathbf{u} - \mathbf{v}^h\right|_{\mathbf{W}^h} . \tag{5.24}$$

Denote by $\delta_m^h \tilde{\mathbf{s}}_{m,n}$ an approximation of the singular basis function $\delta_m^h \mathbf{s}_{m,n}$ such that $\tilde{\mathbf{s}}_{m,n}$ can be written in the general form (3.5) with approximate exponent $\tilde{\alpha}_{m,n}$ and approximate coefficient vectors $\tilde{\lambda}_{m,n}$ and $\tilde{\mu}_{m,n}$. The exact solution $\mathbf{u} \in \mathbf{W}$ of the FOSLS minimization problem (2.4) has the form

$$\mathbf{u} = \mathbf{u}_0 + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \omega_{m,n} \delta_m^h \mathbf{s}_{m,n} + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \omega_{m,n} \left(\delta_m^H - \delta_m^h \right) \mathbf{s}_{m,n}, \tag{5.25}$$

with $\mathbf{u}_0 \in \mathbf{W}_S^1$ independent of h. We choose

$$\mathbf{v}^h := I^h \mathbf{u}_0 + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \omega_{m,n} \delta_m^h \tilde{\mathbf{s}}_{m,n} + \sum_{m=1}^{M} \sum_{n=1}^{N_m} \omega_{m,n} \left(\delta_M^H - \delta_m^h \right) \tilde{\mathbf{s}}_{m,n},$$

where I^h is the interpolant operator that was introduced in (5.5). Using (5.24) and the triangle inequality, we have

$$\begin{aligned} \left|\mathbf{u} - \mathbf{u}^h\right|_{\mathbf{W}^h,\Omega_j} \leq C \left|\mathbf{u}_0 - \mathbf{v}^h\right|_{\mathbf{W}^h,\Omega_j} &+ \sum_{m=1}^{M} \sum_{n=1}^{N_m} |\omega_{m,n}| \left|\delta_m^h \left(\mathbf{s}_{m,n} - \tilde{\mathbf{s}}_{m,n}\right)\right|_{\mathbf{W}^h,\Omega_j} \\ &+ \sum_{m=1}^{M} \sum_{n=1}^{N_m} |\omega_{m,n}| \left|\left(\delta_m^H - \delta_m^h\right) \left(\mathbf{s}_{m,n} - \tilde{\mathbf{s}}_{m,n}\right)\right|_{\mathbf{W}^h,\Omega_j}, \end{aligned} \tag{5.26}$$

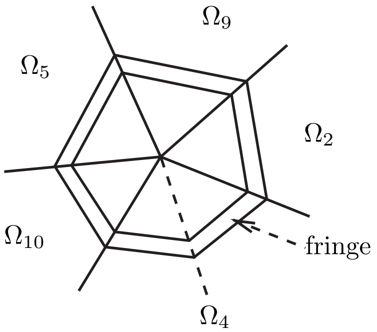


Fig. 5.2. *Geometry of interfaces meeting at a cross-point. Here, $I_m = 5$, $\{j_i\} = \{2, 9, 5, 10, 4\}$.*

for $j = 1, \ldots, J$. Standard interpolation error estimates can be used to estimate $|\mathbf{u}_0 - I^h\mathbf{u}_0|_{\mathbf{W}^h,\Omega_j}$, analogously to the proof of Theorem 5.1. We now derive estimates for the remaining terms, which involve the singular basis functions and their approximations.

Let $\Omega_{j_i} \subset \Omega$, $i = 1, \ldots, I_m$, be the set of subdomains that meet at cross-point m, ordered so that they appear consecutively with increasing i. (See Figure 5.2.) Let $\{\vartheta_{j_i}\}_{i=1,\ldots,I_m}$ be the set of angles at cross-point m. If the difference of two subsequent angles is greater than $\pi/2$, then we introduce artificial interfaces, equally spaced in the interval $(\vartheta_{j_i}, \vartheta_{j_{i+1}})$, such that the angles between subsequent interfaces are now smaller than $\pi/2$. (See the dashed interface line in Figure 5.2.) The platform and fringe are such that, for subdomains that do not require an artificial interface, $\Omega_{j_i} \cap (P_m \cup F_m)$ is an isosceles triangle. In the presence of artificial interfaces, platform and fringe are such that their intersection with Ω_{j_i} is the union of isosceles triangles whose sides are aligned with the artificial interfaces.

In the following calculations, we denote by $\tilde{I}_m$ the total number of actual and artificial interfaces, and by $\{\theta_{m,i}\}_{i=1,\ldots,\tilde{I}_m}$ the angles of these interfaces, ordered such that $\theta_{m,i} < \theta_{m,i+1}$, $i = 1, \ldots, \tilde{I}_m$ (define $\theta_{m,\tilde{I}_m+1} = 2\pi + \theta_{m,1}$). $P_{m,i}$ and $F_{m,i}$ are the parts of platform and fringe, respectively, that are enclosed by angles $\theta_{m,i}$ and $\theta_{m,i+1}$, $i = 1, \ldots, \tilde{I}_m$. Denote by $R_{m,i}$ the distance from $F_{m,i}$ to the cross-point, and by $h_{m,i}$ the radial width of $F_{m,i}$. We omit the subscript ${}_{m,n}$ for singular functions, when it is obvious to which singular functions we refer.

The cut-off function on $F_{m,i}$, for $\theta \in (\theta_{m,i}, \theta_{m,i+1})$, is

$$\delta^h(r,\theta) = \frac{\cos(\theta_{m,i+1/2} - \theta)}{h}\left(\frac{R_{m,i} + h_{m,i}}{\cos(\theta_{m,i+1/2} - \theta)} - r\right), \tag{5.27}$$

where $\theta_{m,i+1/2} = (\theta_{m,i+1} - \theta_{m,i})/2$, with partial derivatives

$$\delta^h_x(r,\theta) = -\frac{\cos\theta_{m,i+1/2}}{h}, \qquad \delta^h_y(r,\theta) = -\frac{\sin\theta_{m,i+1/2}}{h}. \tag{5.28}$$

In what follows, we assume the approximate singular basis functions, $\tilde{\mathbf{s}}_{m,n}$, are of the form (3.5), with known but inexact coefficients $\tilde{\alpha}_{m,n} = \alpha_{m,n} + \eta_m$, $\tilde{\lambda}_{m,n} = \lambda_{m,n} + O(\eta_m)$, and $\tilde{\mu}_{m,n} = \mu_{m,n} + O(\eta_m)$. We will drop subscripts where the meaning is clear.

Lemma 5.3. *Let $\eta_{m,n} = |\alpha_{m,n} - \tilde{\alpha}_{m,n}|$. The estimate*

$$\left|\delta^h_m \mathbf{s}_{m,n} - \delta^h_m \tilde{\mathbf{s}}_{m,n}\right|^2_{\mathbf{W}^h} \le \frac{C(a)\eta^2_{m,n}}{h} \tag{5.29}$$

holds for all cross-points and singular basis functions and some constant $C(a)$ independent of h and η.

Proof. Omitting the subscripts $_{m,n}$ for convenience, note that

$$\mathcal{F}^{nc}_{P_m}(\delta^h_m\mathbf{s}, \delta^h_m\mathbf{s}) = \mathcal{F}^{nc}_{P_m}(\delta^h_m\tilde{\mathbf{s}}, \delta^h_m\mathbf{s}) = \mathcal{F}^{nc}_{P_m}(\delta^h_m\tilde{\mathbf{s}}, \delta^h_m\tilde{\mathbf{s}}) = 0,$$

where P_m is the platform associated with singular function $\mathbf{s}$. Letting $\mathbf{s} = (s_1, s_2)^t$ and $\tilde{\mathbf{s}} = (\tilde{s}_1, \tilde{s}_2)^t$, (5.28) and some vector calculus implies

$$\begin{aligned}\left|\delta^h\mathbf{s} - \delta^h\tilde{\mathbf{s}}\right|_{\mathbf{W}^h,F_{m,i}} &= \left\|\nabla\cdot\left(\delta^h\mathbf{s} - \delta^h\tilde{\mathbf{s}}\right)\right\|^2_{0,F_{m,i}} + \left\|\nabla\times\left(\delta^h\mathbf{s} - \delta^h\tilde{\mathbf{s}}\right)\right\|^2_{0,F_{m,i}} \\ &= \left\|\nabla\delta^h\cdot(\mathbf{s} - \tilde{\mathbf{s}})\right\|^2_{0,F_{m,i}} + \left\|\nabla^\perp\delta^h\cdot(\mathbf{s} - \tilde{\mathbf{s}})\right\|^2_{0,F_{m,i}} \\ &= \frac{1}{h^2_{m,i}}\left(\|s_1 - \tilde{s}_1\|^2_{0,F_{m,i}} + \|s_2 - \tilde{s}_2\|^2_{0,F_{m,i}}\right).\end{aligned} \tag{5.30}$$

We now estimate only the last term in (5.30), since a similar estimate can be derived analogously for the other term. Let $\eta_\lambda = \lambda - \tilde{\lambda}$ and $\eta_\mu = \mu - \tilde{\mu}$, and note that η_λ and η_μ are of order $O(\eta)$. Then we get

$$\begin{aligned}\|s_2 - \tilde{s}_2\|^2_{0,F_{m,i}} &= a_i\left\|\alpha r^{\alpha-1}\left(\lambda\cos(\alpha-1)\theta - \mu\sin(\alpha-1)\theta\right)\right. \\ &\qquad \left.-\tilde{\alpha}r^{\tilde{\alpha}-1}(\tilde{\lambda}\cos(\alpha_h-1)\theta - \tilde{\mu}\sin(\tilde{\alpha}-1)\theta)\right\|^2_{0,F_{m,i}} \\ &= a_i\left\|(\alpha r^{\alpha-1} - \tilde{\alpha}r^{\tilde{\alpha}-1})(\lambda\cos(\alpha-1)\theta - \mu\sin(\alpha-1)\theta)\right. \\ &\qquad \left.-\tilde{\alpha}r^{\tilde{\alpha}-1}(\eta_\lambda\cos(\tilde{\alpha}-1)\theta - \eta_\mu\sin(\tilde{\alpha}-1)\theta)\right\|^2_{0,F_{m,i}} \\ &\le 2a_i\left\|(\alpha r^{\alpha-1} - \tilde{\alpha}r^{\tilde{\alpha}-1})(\lambda\cos(\alpha-1)\theta - \mu\sin(\alpha-1)\theta)\right\|^2_{0,F_{m,i}} \\ &\quad + 2a_i\left\|\tilde{\alpha}r^{\tilde{\alpha}-1}(\eta_\lambda\cos(\tilde{\alpha}-1)\theta - \eta_\mu\sin(\tilde{\alpha}-1)\theta)\right\|^2_{F_{m,i}} \\ &\le 4a_i(\lambda^2+\mu^2)\left\|\alpha r^{\alpha-1} - \tilde{\alpha}r^{\tilde{\alpha}-1}\right\|^2_{F_{m,i}} + 4a_i(\eta^2_\lambda + \eta^2_\mu)\left\|\tilde{\alpha}r^{\tilde{\alpha}-1}\right\|^2_{F_{m,i}}.\end{aligned} \tag{5.31}$$

Let $0 < r < R + h$, $0 < h < 1$, and $0 < \eta < 1$. Then we have

$$|1 - r^\eta| \le C_1\eta, \qquad \left\|\tilde{\alpha}r^{\tilde{\alpha}-1}\right\|^2_{0,F_{m,i}} \le C_2 h_{m,i}, \tag{5.32}$$

which, with $0 < \alpha, \tilde{\alpha} \le 1$, implies that

$$\begin{aligned}\left\|\alpha r^{\alpha-1} - \tilde{\alpha}r^{\tilde{\alpha}-1}\right\|^2_{0,F_{m,i}} &\le \left\|\alpha(r^{\alpha-1} - r^{\tilde{\alpha}-1}) - \eta r^{\tilde{\alpha}-1}\right\|^2_{0,F_{m,i}} \\ &\le 2\left\|r^{\alpha-1} - r^{\tilde{\alpha}-1}\right\|^2_{0,F_{m,i}} + 2\eta^2\left\|r^{\tilde{\alpha}-1}\right\|^2_{0,F_{m,i}} \\ &\le 2\left\|r^{\alpha-1}(1 - r^\eta)\right\|^2_{0,F_{m,i}} + 2\eta^2 C_2 h_{m,i} \\ &\le 2C_1^2\eta^2 C_2 h_{m,i} + 2\eta^2 C_2 h_{m,i} \\ &= C\eta^2 h_{m,i}.\end{aligned} \tag{5.33}$$

Combining estimates (5.31)–(5.33) with (5.30), we get

$$\left|\delta^h\mathbf{s} - \delta^h\tilde{\mathbf{s}}\right|_{\mathbf{W}^h,F_{m,i}} \le \frac{C(a_i)\eta^2}{h_{m,i}}. \tag{5.34}$$

TABLE 5.1
Diameter of the fringe is fixed $= 1/6$ *on all levels,* $\Omega = (0,1)^2$.

	With quadratics		Without quadratics	
h	$\mathcal{G}(\mathbf{u}^h; f)$	ratio	$\mathcal{G}(\mathbf{u}^h; f)$	ratio
1/24	4.29(-2)	1.81	9.66(-2)	1.11
1/48	2.37(-2)	2.49	8.68(-2)	1.20
1/96	9.51(-3)	3.14	7.22(-2)	1.34
1/192	3.02(-3)	3.56	5.36(-2)	1.52
1/384	8.48(-4)	3.78	3.52(-2)	1.69
1/768	2.24(-4)		2.08(-2)	

Summing over interfaces and artificial interfaces, and noting that $h_{m,i} = O(h)$, completes the proof. □

Using estimates (5.7) and (5.29) in (5.26) implies the main result of this section.

THEOREM 5.4. *Denote by* $\mathbf{u} \in \mathbf{W}$ *the solution of minimization problem* (2.4) *and by* $\mathbf{u}^h \in \mathbf{W}^h$ *the solution of the discretized variational problem* (2.7). *Also, let* $\eta > 0$ *be the maximum error in the exponent of the approximate singular basis function,* h *be the mesh size of triangulation* $\mathcal{T}^h$ *and the fringe width, and*

(5.35)
$$\kappa := \min_{\mathbf{s}}\{\alpha : \alpha \textit{ is the exponent of } \mathbf{s}, \alpha > 1, \nabla \cdot (\sqrt{a}\mathbf{s}) = 0, \textit{ and } \nabla \times (\mathbf{s}/\sqrt{a}) = 0\}.$$

Then

$$\left|\mathbf{u} - \mathbf{u}^h\right|_{\mathbf{W}^h} \le C(a)\left(h^{\kappa-1} + \eta/\sqrt{h}\right), \tag{5.36}$$

where the constant $C(a)$ *does not depend on* η, κ, *and* h.

To achieve a discretization error of order $O(h)$ in the nonconforming finite element space $\mathbf{W}^h$, we must ensure that $\kappa \ge 2$ and that the approximation error in the exponent η is of order $O(h^{3/2})$. The constraint on κ can be met by adding basis functions of the general form (3.5) that have exponents $1 < \alpha < 2$ and satisfy the first two equations in (2.2) with $f = 0$.

5.4. An example. Here, we present a numerical example to illustrate the theoretical results of the previous sections. We consider problem (2.2) on the unit square, with $f = 0$ and the Dirichlet boundary condition

$$\tau \cdot \mathbf{u} = \tau \cdot \mathbf{s} \quad \text{on } \partial(0,1)^2, \tag{5.37}$$

where $\mathbf{s}$ is the singular function associated with the coefficient

$$a(x,y) = \begin{cases} 1 & \text{for } 0 < x,y < 1/2 \text{ and } 1/2 < x,y < 1, \\ 100 & \text{elsewhere in } (0,1)^2. \end{cases} \tag{5.38}$$

For this checkerboard pattern, the exponent is approximately $\alpha \approx 0.126902069$.

Table 5.1 shows the value of the FOSLS functional at the solution for various values of h. The left two columns show results for the finite element space $\mathbf{W}^h$ consisting of linear, quadratic, and singular elements as described in section 4. The right two columns show results for a finite element space that contains only the linear and singular elements. The ratio refers to the quotient of two subsequent functional values: ratio $= \mathcal{G}(\mathbf{u}^h; f)/\mathcal{G}(\mathbf{u}^{h/2}; f)$. For the space $\mathbf{W}^h$, this ratio approaches 4 as h is decreased. In the space that does not contain quadratic elements in the fringe, the ratio approaches 2 for decreasing h. These results signal a lower approximation order when quadratic elements are not included in the finite element space.

6. A multilevel solver.

We now describe a multilevel solver for the linear system that arises from the finite element discretization using $\mathbf{W}^h$. Our goal is to use standard coarsening for linear elements and to include singular basis functions on every level. This results in a hierarchy of spaces $\mathbf{W}^{2^K h} \nsubseteq \mathbf{W}^{2^{K-1}h} \nsubseteq \cdots \nsubseteq \mathbf{W}^{2h} \nsubseteq \mathbf{W}^h$. The spaces are nested except for the singular basis functions, which are nonnested, since the fringes on different levels have different widths.

We coarsen such that the platform associated with a given cross-point has equal size on all levels. The choice of interpolation and restriction for linear and quadratic elements in fringes is driven by the interpolation for singular basis functions.

Consider interpolation of $\delta_m^{2h}\tilde{\mathbf{s}}_{m,n} \in \mathbf{W}^{2h}$ to $\mathbf{W}^h$. The coefficient $\omega_{m,n}^{2h}$ of $\delta_m^{2h}\tilde{\mathbf{s}}_{m,n}$ is transferred by injection: $\omega_{m,n}^h \leftarrow \omega_{m,n}^{2h}$. What is left is the difference $(\delta_m^{2h} - \delta_m^h)\tilde{\mathbf{s}}_{m,n}$ that is interpolated using standard linear and quadratic interpolation. Denote by N_L^h and N_Q^h the respective numbers of linear and quadratic basis functions in $\mathbf{W}^h$, and by $\boldsymbol{\psi}_l$, $l = 1, \ldots, N_L^h$, and $\boldsymbol{\phi}_k$, $k = 1, \ldots, N_Q^h$, the respective linear and quadratic basis functions in $\mathbf{W}^h$, such that

$$\begin{aligned}(6.1)\quad \mathbf{W}^h = \operatorname{span}\left\{\boldsymbol{\psi}_l \;:\; l = 1, \ldots, N_L^h\right\} \;\cup\; \operatorname{span}\left\{\boldsymbol{\phi}_k \;:\; k = 1, \ldots, N_Q^h\right\} \\ \cup\; \operatorname{span}\left\{\delta_m^h \tilde{\mathbf{s}}_{m,n} \;:\; m = 1, \ldots, M;\; n = 1, \ldots, N_m\right\}.\end{aligned}$$

Define the vectors $\boldsymbol{\beta}_{m,n}^h = \{\beta_{m,n}^{h,l}\}_{l=1,\ldots,N_L^h}$ and $\boldsymbol{\gamma}_{m,n}^h = \{\gamma_{m,n}^{h,k}\}_{k=1,\ldots,N_Q^h}$ so that their elements are the respective coefficients of linear and quadratic basis functions of the pointwise linear and quadratic interpolant of $(\delta_m^{2h} - \delta_m^h)\tilde{\mathbf{s}}_{m,n}$. We now have

$$(6.2)\qquad I^h(\delta_m^{2h} - \delta_m^h)\tilde{\mathbf{s}}_{m,n} := \sum_{l=1}^{N_L^h} \beta_l^h \boldsymbol{\psi}_{m,n}^{h,l} + \sum_{k=1}^{N_Q^h} \gamma_k^h \boldsymbol{\phi}_{m,n}^{h,k}.$$

Note that most of the $\beta_{m,n}^{h,l}$ and $\gamma_{m,n}^{h,k}$ are zero, since $\delta_m^{2h} - \delta_m^h$ is nonzero only in the fringe of level $2h$ associated with cross-point m.

Assume that the basis functions are ordered so that the first N_L^h are linear, the next N_Q^h are quadratic, and the last $N_S := \sum_{m=1}^M N_m$ are singular basis functions. Then the column in interpolation matrix $\mathcal{I}_{2h}^h$ corresponding to the singular basis function with index (m,n) is $(\beta_{m,n}^{h,1}, \ldots, \beta_{m,n}^{h,N_L^h}, \gamma_{m,n}^{h,1}, \ldots, \gamma_{m,n}^{h,N_Q^h}, 0, \ldots, 0, 1, 0, \ldots, 0)$. Fine-level linear basis functions centered at vertices inside the coarse-level fringe are interpolated using standard linear interpolation.

Figure 6.1 shows a section of the fringe in a triangular mesh of mesh size h (thick and thin lines), and a section of the fringe in the coarsened triangular mesh of mesh size $2h$ (thick lines). It illustrates the location of the quadratic nodes on both levels.

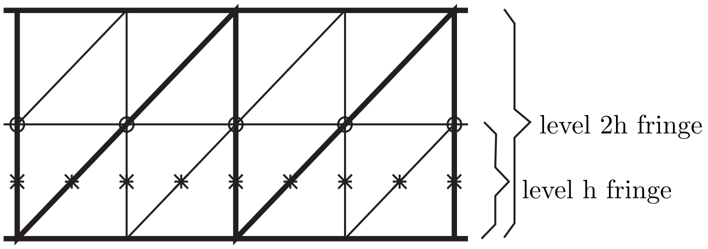


FIG. 6.1. *Location of quadratic points (level h: stars; level $2h$: circles).*

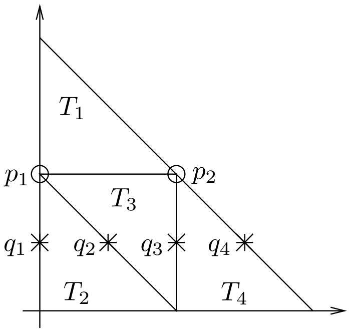


FIG. 6.2. *Coarse fringe triangle subdivided into four subtriangles.*

Note that quadratic nodes on level $2h$ coincide with vertices of triangles on level h. We now describe the interpolation formulas that are used for quadratic nodes on level h and linear nodes on level h that coincide with quadratic nodes on level $2h$. For the latter, we use linear interpolation from the two neighboring coarse points and add to that the value of the quadratic at that point.

Figure 6.2 shows a coarse fringe unit triangle in general (ξ, η) coordinates that is subdivided into four subtriangles $T_1, \ldots, T_4$. Triangles T_2, T_3, T_4 are in the fine fringe. Coarse-level quadratic basis functions are centered at points q_1^c and q_2^c, and fine-level quadratic basis functions are centered at points $q_1^f, \ldots, q_4^f$. (We denote by $q_1^c, q_2^c, q_1^f, \ldots, q_4^f$ the respective points, as well as the coefficients of quadratic basis functions at these points.) Assume that the linear part of the coarse-level function is zero. The quadratic part is

$$Q(\xi, \eta) = q_1^c Q_1^c(\xi, \eta) + q_2^c Q_2^c(\xi, \eta),$$

with

$$\begin{aligned} Q_1^c(\xi, \eta) &= 4\eta(1 - \xi - \eta), \\ Q_2^c(\xi, \eta) &= 4\xi\eta. \end{aligned}$$

First we interpolate the linear fine-level points that coincide with quadratic coarse-level points. We obtain the following linear functions:

$$\begin{aligned} L_2(\xi, \eta) &= 2\eta q_1^c && \text{in } T_2, \\ L_3(\xi, \eta) &= 2\eta(2(q_1^c - q_2^c)\xi - q_1^c) && \text{in } T_3, \\ L_4(\xi, \eta) &= 2\eta q_2^c && \text{in } T_4. \end{aligned}$$

We want the interpolant to be pointwise exact at fine-level quadratic points, so interpolation of fine-level quadratic points is determined by

$$\begin{array}{lcl cl} q_1^f &=& Q(0, 1/4) - L_2(0, 1/4) &=& 1/4\ q_1^c, \\ q_2^f &=& Q(1/4, 1/4) - L_2(1/4, 1/4) &=& 1/4\ q_2^c, \\ q_3^f &=& Q(1/2, 1/4) - L_4(1/2, 1/4) &=& 1/4\ q_1^c, \\ q_4^f &=& Q(3/4, 1/4) - L_4(3/4, 1/4) &=& 1/4\ q_2^c. \end{array} \tag{6.3}$$

Interpolation weights in (6.3) do not depend on ξ or η, so they are the same in (x, y) coordinates.

Denote by $\hat{I}_{2h}^{h} \in \Re^{(N_L+N_Q)\times(N_L+N_Q)}$ the matrix that interpolates linear and quadratic basis functions on level h from level $2h$, as defined above. Let $B^h := \{\boldsymbol{\beta}_{m,n}^h\}_{n=1,\ldots,N_m;m=1,\ldots,M}$ and $\Gamma^h := \{\boldsymbol{\gamma}_{m,n}^h\}_{n=1,\ldots,N_m;m=1,\ldots,M}$. Then we can write the interpolant operator $\mathcal{I}_{2h}^h : \mathbf{W}^{2h} \to \mathbf{W}^h$ in matrix form as

$$\mathcal{I}_{2h}^{h} = \left(\begin{array}{ccc} \multicolumn{2}{c}{\hat{I}_{2h}^{h}} & \begin{array}{c} B^h \\ \Gamma^h \end{array} \\ 0 & 0 & I \end{array} \right), \tag{6.4}$$

where I is the identity matrix in $\Re^{N_S \times N_S}$.

We extend this idea to all levels to obtain interpolation matrices $\mathcal{I}_{2^K h}^{2^{K-1}h}, \ldots, \mathcal{I}_{2h}^{h}$. The restriction operators are defined as the transpose of interpolation $\mathcal{I}_{2^k h}^{2^{k+1}h} = (\mathcal{I}_{2^{k+1}h}^{2^k h})^t$ for $k = 0, \ldots, K-1$. The coarse-grid stiffness matrix is determined from the Galerkin principle by fine-grid stiffness matrix S^h and interpolation matrix $\mathcal{I}_{2h}^h$:

$$S^{2h} = (\mathcal{I}_{2h}^{h})^t \; S^h \; \mathcal{I}_{2h}^{h}. \tag{6.5}$$

This definition is extended recursively to all levels.

We use Gauss–Seidel relaxation for pre- and postrelaxation in the multigrid iteration and solve the coarse grid problem approximately using algebraic multigrid (AMG) (see [29]). Numerical tests have shown that the standard AMG algorithm is not well suited to solving the resulting linear system. Therefore, we instead use a Schur complement approach that exploits the structure of the coarse-grid problem, with the subproblems treated by AMG.

To explain this Schur-AMG approach, note that the coarse-grid linear system has the general form

$$\begin{pmatrix} A & V \\ V^t & D \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}, \tag{6.6}$$

where submatrix A represents connections between linear and quadratic basis functions and D represents connections between singular basis functions. A is sparse and D is block diagonal. The off-diagonal submatrix V represents connections between the linear and quadratic basis functions and the singular basis functions. Linear system (6.6) can be reduced to

$$\mathbf{u} = A^{-1}\mathbf{f} - A^{-1}V\mathbf{w}, \tag{6.7}$$

$$\mathbf{w} = (D - V^t A^{-1} V)^{-1}(\mathbf{g} - V^t A^{-1}\mathbf{f}). \tag{6.8}$$

We calculate $A^{-1}\mathbf{f}$ and $A^{-1}V$ approximately using AMG. The inverse of $D - V^t A^{-1} V$ is then calculated directly, using Gaussian elimination, since the number of singular basis functions is assumed to be small.

Denote by N_C the number of coarse-grid linear and quadratic basis functions and by N_S the number of singular basis functions. Thus, $A \in \Re^{N_C \times N_C}$, $V \in \Re^{N_C \times N_S}$, and $D \in \Re^{N_S \times N_S}$. AMG has complexity of order $O(N_C)$, and the inverse of $D - V^t A^{-1} V$ can be calculated in $O(N_S^3)$ operations. Hence, the complexity of the coarse-grid solver is of order $O(N_S^3 + N_S N_C)$. Since N_S is assumed to be small in comparison to N_C, we can deduce that standard multilevel complexity analysis applies (see, for example, [31]).

Table 7.1
Influence of the coefficient $a(x,y)$ ($W(2,2)$ cycle).

	$a(x,y) \in \{1,100\}$				$a(x,y) \in \{1,10000\}$			
Levels =	3	4	5	6	3	4	5	6
1×1	.20	.15	.13	.13	.27	.23	.13	.13
2×2	.28	.23	.13	.13	.37	.35	.24	.13
3×3	.33	.29	.18	.13	.40	.42	.32	.17
4×4	.38	.36	.25	.13	.45	.48	.39	.23

Table 7.2
Effective convergence factors: influence of the number of relaxations ($a(x,y) \in [1,100]$).

	$W(1,1)$				$W(2,2)$				$W(4,4)$			
Levels =	3	4	5	6	3	4	5	6	3	4	5	6
1×1	.30	.24	.27	.27	.45	.39	.36	.36	.64	.59	.49	.49
2×2	.39	.34	.26	.26	.53	.48	.36	.36	.71	.68	.58	.49
3×3	.44	.36	.30	.26	.57	.54	.42	.36	.74	.72	.63	.49
4×4	.48	.47	.37	.26	.62	.60	.50	.36	.75	.72	.63	.49

7. Numerical results. To study the convergence properties of the multigrid algorithm described above let the domain Ω be a square partitioned in a checkerboard fashion into square subdomains of equal size, where the coefficient a is constant. In our examples, a takes on two values that are distributed over the square subdomains in a checkerboard fashion. We report asymptotic convergence factors of the functional value $(\mathcal{G}(\mathbf{u}^h; f))^{1/2}$ that were obtained by setting $f = 0$ and imposing homogeneous Dirichlet boundary conditions $\mathbf{n} \times \sqrt{a}\mathbf{u}^h = 0$ on $\partial\Omega$. Thus, the exact solution of the problem is $\mathbf{u}^h = 0$, which allows us to perform many iterations without encountering serious machine representation effects. To properly test convergence, we initialize all variables randomly.

Table 7.1 shows asymptotic convergence factors for the $W(2,2)$ cycle for two examples of a. Four checkerboard patterns are investigated, ranging from one singular basis function to 16. Each column shows asymptotic convergence factors for fixed mesh-size h. Hence, the domain Ω changes for varying numbers of singular basis functions. However, the width of fringes does not change.

We observe that, for a larger number of levels, typical multigrid convergence factors that are h-independent are attained. For a smaller number of levels, the convergence factors appear to grow with the number of singular basis functions. However, this dependency appears to weaken as more levels are added; it appears to be stronger for larger jumps in the coefficient a.

In Table 7.2, the influence of the number of pre- and postrelaxation steps in the W-cycle is shown. We display effective convergence factors relative to the $W(1,1)$ cycle. Since for integer $k > 1$ one $W(k,k)$ cycle is k times more costly than one $W(1,1)$ cycle, we have $\rho_{W(k,k),\text{effective}} = \rho_{W(k,k)}^{1/k}$. Increasing the number of pre- and postrelaxation steps increases effective convergence factors. It is, hence, most efficient to use $W(1,1)$ cycles.

Table 7.3 shows results for two convergence tolerances for the AMG iteration that is used to invert A approximately. For the larger tolerance of $1e-1$ convergence of the multilevel iteration is somewhat slower than for the smaller tolerance of $1e-9$. This difference is less pronounced when more levels are added.

Table 7.3
Influence of the coarse grid solver on the overall convergence ($W(2,2)$ cycles, $a(x,y) \in \{1,100\}$).

AMG tolerance =	$1e-9$				$1e-1$			
Geometry\levels =	3	4	5	6	3	4	5	6
1×1	.20	.15	.13	.13	.22	.15	.13	.13
2×2	.28	.23	.13	.13	.42	.31	.17	.13
3×3	.33	.29	.18	.13	.51	.40	.24	.13
4×4	.38	.36	.25	.13	.59	.51	.35	.17

8. Conclusions.

We introduced a finite element method for FOSLS L^2 formulation of the diffusion equation with discontinuous coefficients. Our approach uses singular basis functions to yield accurate approximation of the flux variable close to singular points in the domain at minimal additional computational cost. Stress intensity factors are also calculated. We developed a special discretization error analysis, since standard theory is not applicable. This led to a general error estimate for FOSLS L^2 discretizations with nonconforming finite elements. We also proposed a multilevel algorithm for the solution of the resulting linear system that uses nonstandard coarse spaces including coarse representations of singular basis functions. The performance of the algorithm is illustrated by numerical examples.

Appendix A. Uniform coercivity of $\mathcal{F}^{nc}$.

The purpose of this appendix is to establish the uniform coercivity of $\mathcal{F}^{nc}$. In [2] such a bound was established for the conforming functional. Here, we show that the coercivity constant will be independent of h and η, the error in the exponents of the singular basis function, only if η goes to zero at least as fast as the cosine of the angle between the mesh-dependent singular function $\delta_m^h \mathbf{s}_{m,n}^\eta$ and the subspace, $\mathbf{W}_S^1$, of piecewise H^1 functions. Unfortunately, a proof that the angle between the singular basis functions and $\mathbf{W}_S^1$ is $O(h)$ has not been found. In this next section we provide numerical proof.

A.1. Angle between $\delta_m^h \mathbf{s}_{m,n}$ and W_S^1.

In the example we present in Figure A.1, the domain Ω is divided into subdomains in a 2×2 checkerboard fashion, with $a = 1$ in the upper left and lower right subdomains, and $a = 100$ in the upper right and lower left subdomains. This configuration results in a singularity in the center of the domain with exponent $\alpha \approx 0.126902$. Figure A.1 depicts $1 - \cos\theta$ as a function of h, where θ is the angle between the singular basis function and the rest of

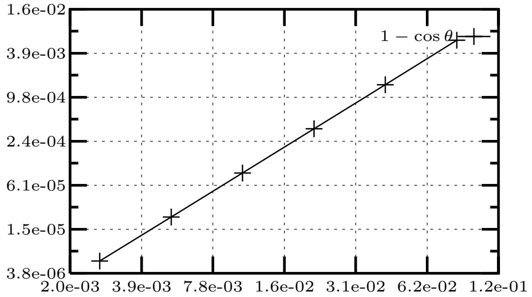


Fig. A.1. *A* $\log_2$*-*$\log_2$ *plot of* $1 - \cos\theta$ *(see* (A.1)*), where* h *is on the* x*-axis.* θ *is the angle between a singular basis function and the rest of the space.*

the space,

$$\cos\theta = \frac{\mathcal{F}^{nc}(\delta^h \mathbf{s}^h, I^h(\delta^h \mathbf{s}^h))}{\mathcal{F}^{nc}(\delta^h \mathbf{s}^h, \delta^h \mathbf{s}^h)^{1/2}\mathcal{F}^{nc}(I^h(\delta^h \mathbf{s}^h), I^h(\delta^h \mathbf{s}^h))^{1/2}}, \tag{A.1}$$

and I^h is the standard pointwise linear interpolation operator.

Both axes in the figure are on a $\log_2$ scale. All data points lie on a straight line, so we conjecture $1 - \cos\theta = O(h^2)$, and hence $\theta = O(h)$.

A.2. Uniform coercivity. We assume that the discrete space consists of the conforming linears and quadratics, with proper jumps across the interfaces, plus a finite number of singular basis functions. Here, as in section 5, we make the assumption that the approximate singular basis functions $\tilde{\mathbf{s}}_{m,n}$ are of the form (3.5), with known but inexact coefficients $\tilde{\alpha}_{m,n} = \alpha_{m,n} + \eta_{m,n}$, $\tilde{\lambda}_{m,n} = \lambda_{m,n} + O(\eta_{m,n})$, and $\tilde{\mu}_{m,n} = \mu_{m,n} + O(\eta_{m,n})$. We will drop subscripts where the meaning is clear.

We further assume that $\eta_{m,n}$ is sufficiently small to resolve the differences between exponents at a given singular point $\mathbf{x}_m$. That is, we assume that $\eta_{m,n} \leq \eta_{m,0}$ for $n = 1, \ldots, N_m$.

To emphasize the dependence on η, we denote the discrete subspace as

$$\mathbf{W}^{h,\eta} := \mathbf{W}_S^{1,h} + \operatorname{span}\left\{\delta_m^h \tilde{\mathbf{s}}_{m,n}\right\}_{n=1,m=1}^{N_m,M}. \tag{A.2}$$

Next we define a bound on the angle between the subspace spanned by the approximate singular basis functions at a given singular point $\mathbf{x}_m$ and the subspace $\mathbf{W}_S^1$. Let

$$\mathcal{S}_m(\eta, h) := \operatorname{span}\{\delta_m^h \tilde{\mathbf{s}}_{m,n}\}_{n=1}^{N_m}, \tag{A.3}$$

and let

$$\gamma_m(h) := \sup_{\eta \leq \eta_{m,0}} \sup_{\mathbf{s}\in\mathcal{S}_m, \mathbf{u}\in\mathbf{W}_S^1} \frac{\mathcal{F}^{nc}\langle \mathbf{s}, \mathbf{u}\rangle}{|\mathbf{s}|_{\mathbf{W}^h}|\mathbf{u}|_{\mathbf{W}^h}}. \tag{A.4}$$

This yields the following result.

LEMMA A.1 (strengthened Cauchy–Schwarz inequality). *For every singular point $\mathbf{x}_m$ there is a constant $\gamma_m(h) < 1.0$ such that*

$$\mathcal{F}^{nc}(\mathbf{s}, \mathbf{w}) \leq \gamma_m |\mathbf{s}|_{\mathbf{W}^h} |\mathbf{w}|_{\mathbf{W}^h} \tag{A.5}$$

for every $\mathbf{s} \in \mathcal{S}_m$ and $\mathbf{w} \in \mathbf{W}_S^1$.

Proof. For any fixed $h > 0, \eta > 0$, there is a positive angle between $\mathcal{S}_m$ and $\mathbf{W}_S^1$. Thus, $\gamma_m < 1.0$ □

This leads to the following bound.

LEMMA A.2. *Let $\mathbf{w} \in \mathbf{W}^{h,\eta}$ have the form*

$$\mathbf{w} = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \beta_{m,n} \delta^m \tilde{\mathbf{s}}_{m,n} + \mathbf{w}_0^h, \tag{A.6}$$

where $\mathbf{w}_0^h \in \mathbf{W}_S^{1,h}$. Then

$$\sum_{m=1}^{M} \sum_{n=1}^{N_m} \beta_{m,n}^2 |\delta_m^h \tilde{\mathbf{s}}_{m,n}|_{\mathbf{W}^h}^2 \leq \frac{C}{1-\gamma^2} |\mathbf{w}|_{\mathbf{W}^h}^2, \tag{A.7}$$

where $\gamma = \max_m \gamma_m$.

Proof. The result follows from the fact that the singular basis functions associated with different singular points are mutually orthogonal and from the strengthened Cauchy–Schwarz inequality. For this proof only, let $\mathcal{P}_m$ represent the platform and fringe around singular point m.

$$\begin{aligned}
|\mathbf{w}|^2_{\mathbf{W}^h} &\geq \left| \sum_{i=1}^{M} \sum_{n=1}^{N-m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right|^2_{\mathbf{W}^h} - 2\mathcal{F} \left\langle \sum_{i=1}^{M} \sum_{n=1}^{N_m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n},\, \mathbf{w}_0 \right\rangle + |\mathbf{w}_0|^2_{\mathbf{W}^h} \\
&\geq \sum_{i=1}^{M} \left| \sum_{n=1}^{N_m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right|^2_{\mathbf{W}^h} \\
&\quad - 2 \sum_{i=1}^{M} \gamma_m \left| \sum_{n=1}^{N_m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right|_{\mathbf{W}^h} |\mathbf{w}_0|_{\mathbf{W}^h, \mathcal{P}_m} + |\mathbf{w}_0|^2_{\mathbf{W}^h} \\
&\geq (1-\gamma^2) \sum_{i=1}^{M} \left| \sum_{n=1}^{N_m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right|^2_{\mathbf{W}^h} + |\mathbf{w}_0|^2_{\mathbf{W}^h} - \sum_{m=1}^{M} |\mathbf{w}_0|^2_{\mathbf{W}^h, \mathcal{P}_m} \\
&\geq (1-\gamma^2) \sum_{i=1}^{M} \left| \sum_{n=1}^{N_m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right|^2_{\mathbf{W}^h} .
\end{aligned}$$

Since the singular functions associated with any give singular point are linearly independent, there exists a constant C_m independent of h such that

$$\left| \sum_{n=1}^{N_m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right|_{\mathbf{W}^h} \geq C_m \sum_{n=1}^{N_m} |\beta_{m,n}|^2 \left\| \delta^h_m \tilde{\mathbf{s}}_{m,n} \right\|^2_{\mathbf{W}^h}. \tag{A.8}$$

This completes the proof. □

We now show a Poincaré–Friedrichs inequality for the bilinear form. The nonconformity is rooted in the fact that the singular basis functions, $\delta^h_m \tilde{\mathbf{s}}_{m,n}$, do not satisfy the jump conditions exactly. Suppose that Γ_{ij} is the interface between Ω_I and Ω_J. Let $[g]_{\Gamma_{ij}}$ denote the jump in g across Γ_{ij} . Let $\mathbf{w} \in \mathbf{W}^{h,\eta}$ be defined as in (A.6). We have

$$\begin{aligned}
\left[\mathbf{n} \cdot \sqrt{a}\mathbf{w}\right]_{\Gamma_{ij}} &= \left[\mathbf{n} \cdot \sqrt{a} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right]_{\Gamma_{ij}} \\
&= \sum_{m\ :\ \Gamma_{ij} \cap \mathcal{P}_k \neq \emptyset} \sum_{n=1}^{N_m} \beta_{m,n} \left[\mathbf{n} \cdot \sqrt{a} \delta^h_m \tilde{\mathbf{s}}_{m,n}\right]_{\Gamma_{ij}} , \\
\left[\boldsymbol{\tau} \cdot \frac{1}{\sqrt{a}} \mathbf{w} \right]_{\Gamma_{ij}} &= \left[\boldsymbol{\tau} \cdot \frac{1}{\sqrt{a}} \sum_{m=1}^{M} \sum_{n=1}^{N_m} \beta_{m,n} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right]_{\Gamma_{ij}} \\
&= \sum_{m\ :\ \Gamma_{ij} \cap \mathcal{P}_k \neq \emptyset} \sum_{n=1}^{N_m} \beta_{m,n} \left[\boldsymbol{\tau} \cdot \frac{1}{\sqrt{a}} \delta^h_m \tilde{\mathbf{s}}_{m,n} \right]_{\Gamma_{ij}} .
\end{aligned}$$

THEOREM A.3. *Let γ be defined as in Lemma* A.2. *Let η bound the maximum error in $\tilde{\alpha}_{m,n}$, $\tilde{\lambda}_{m,n}$, and $\tilde{\mu}_{m,n}$. Then, there exist constants C_1 and C_2 independent* *of h, η such that*

$$\|\mathbf{u}\|_{0,\Omega} \le C\frac{\eta^2}{1-\gamma^2}\,|\mathbf{u}|_{\mathbf{W}^h} \tag{A.9}$$

for all $\mathbf{w} \in \mathbf{W}^{h,\eta}$.

Proof. Consider a Helmholtz decomposition on $\mathbf{W}^h$: for $\mathbf{u} \in \mathbf{W}^h$, there exist p, $\psi \in H^1(\Omega)$ such that

$$\mathbf{u} = \sqrt{a}\nabla p + \frac{1}{\sqrt{a}}\nabla^\perp \psi \tag{A.10}$$

where p is the unique solution of the weak equation

$$\begin{array}{rcll} \langle a\nabla p,\, \nabla q\rangle & = & \langle \sqrt{a}\mathbf{u},\, \nabla q\rangle\,, & \\ p = q & = & 0 & \text{on } \Gamma_D, \\ \mathbf{n}\cdot a\nabla p & = & 0 & \text{on } \Gamma_N, \end{array} \tag{A.11}$$

and ψ is the unique (up to a constant) solution of

$$\begin{array}{ll} \left\langle \dfrac{1}{a}\nabla^\perp\psi,\, \nabla^\perp\phi\right\rangle = \left\langle \dfrac{1}{\sqrt{a}}\mathbf{u},\, \nabla^\perp\phi\right\rangle, & \\ \psi = C_i,\, \phi = 0 & \text{on } \Gamma_{N_i}, \\ \mathbf{n}\cdot\dfrac{1}{a}\nabla\psi = 0 & \text{on } \Gamma_D, \end{array} \tag{A.12}$$

where C_i are arbitrary constants, one of which may be set to zero.

Note that the decomposition is orthogonal in the L^2 sense:

$$\left\langle \sqrt{a}\nabla p,\, \frac{1}{\sqrt{a}}\nabla^\perp\psi\right\rangle_{0,\Omega} = 0. \tag{A.13}$$

We thus have

$$\|\mathbf{u}\|^2_{0,\Omega} = \left\|\sqrt{a}\nabla p\right\|^2_{0,\Omega} + \left\|\frac{1}{\sqrt{a}}\nabla^\perp\psi\right\|^2_{0,\Omega}. \tag{A.14}$$

This next step uses the fact that the jump conditions across boundaries are satisfied exactly except for the singular basis functions, which have support only on $\mathcal{P}_m$ for $m = 1, \dots, M$.

We assume that a is a constant on $\mathcal{P}_m \cap \Omega_i$ and that $\Gamma_{ij} \cap \mathcal{P}_k$ is a straight line starting from the singular point, $\mathbf{x}_m$. We have

$$\begin{aligned} \left\|\sqrt{a}\nabla p\right\|^2_{0,\Omega} &= \left\langle\sqrt{a}\nabla p,\, \sqrt{a}\nabla p\right\rangle_{0,\Omega} = \sum_{i=1}^{J}\langle a\nabla p,\, \nabla p\rangle_{0,\Omega_i} \\ &= \sum_{i=1}^{J}\left\langle -\nabla\cdot\sqrt{a}\mathbf{u},\, p\right\rangle_{0,\Omega_i} + \oint_{\partial\Omega_i}\left(\mathbf{n}\cdot\sqrt{a}\mathbf{u}\right)p \\ &= \sum_{i=1}^{J}\left\langle -\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u},\, \sqrt{a}p\right\rangle_{0,\Omega_i} + \sum_{ij}\int_{\Gamma_{ij}}\left[\mathbf{n}\cdot\sqrt{a}\mathbf{u}\right]p \\ &\le \sum_{i=1}^{J}\left\|\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\mathbf{u}\right\|_{0,\Omega_i}\left\|\sqrt{a}p\right\|_{0,\Omega_i} \\ &\quad + \left|\sum_{ij}\sum_{m:\Gamma_{ij}\cap\mathcal{P}_m\neq\emptyset}\int_{\Gamma_{ij}\cap\mathcal{P}_m}\sum_{n=1}^{N_m}\beta_{m,n}\left[\mathbf{n}\cdot\sqrt{a}\delta^h_m\tilde{\mathbf{s}}_{m,n}\right]p\right|. \end{aligned}$$

With our assumptions, we have

$$\begin{aligned}
&\int_{\Gamma_{ij}\cap\mathcal{P}_m} \sum_{n=1}^{N_m} \beta_{m,n} \left[\mathbf{n}\cdot\sqrt{a}\delta_m^h \tilde{\mathbf{s}}_{m,n}\right] p \\
&\leq C_m \eta \sum_{n=1}^{N_m} |\beta_{m,n}| \left| \int_0^{R_m} \delta_m^h(r) r^{(\tilde{\alpha}_n - 1)} p(r)dr \right| \\
&\leq C_m \eta \sum_{n=1}^{N_m} |\beta_{m,n}| \left(\left\|\sqrt{a}_i p\right\|_{1/2,\Gamma_{ij}\cap\mathcal{P}_k} + \left\|\sqrt{a}_j p\right\|_{1/2,\Gamma_{ij}\cap\mathcal{P}_m} \right),
\end{aligned}$$

where R_m is the radius of $\mathcal{P}_m$. Here C_m involves $\min_\Omega |a|$.

Plugging this into the expression above, after first using the ϵ-inequality twice, and using a trace inequality $\|\sqrt{a}p\|_{1/2,\partial(\Omega_i\cap\mathcal{P}_k)} \leq C\|\sqrt{a}\nabla p\|_{0,\Omega_i\cap\mathcal{P}_k}$, the fact that $\|\frac{1}{\sqrt{a}}\nabla\cdot\sqrt{a}\delta_k^h\tilde{\mathbf{s}}_k^\eta\|_{0,\mathcal{P}_k} \leq C$, where C is independent of h and η, and the Poincaré–Friedrichs inequality on p in [2, Lemma 3.1], we get

$$\begin{aligned}
&\left\|\sqrt{a}\nabla p\right\|_{0,\Omega}^2 \\
&\leq \sum_{i=1}^{J} \left\| \frac{1}{\sqrt{a}} \nabla\cdot\sqrt{a}\mathbf{u} \right\|_{0,\Omega_i} \left\|\sqrt{a}p\right\|_{0,\Omega_i} \\
&\quad + \sum_{ij} \sum_{m:\Gamma_{ij}\cap\mathcal{P}_m\neq\emptyset} C_m\eta \sum_{n=1}^{N_m} |\beta_{M,n}| \left(\left\|\sqrt{a}_i p\right\|_{1/2,\Gamma_{ij}\cap\mathcal{P}_m} + \left\|\sqrt{a}_j p\right\|_{1/2,\Gamma_{ij}\cap\mathcal{P}_m} \right) \\
&\leq \frac{1}{\epsilon_1} \sum_{i=1}^{J} \left\| \frac{1}{\sqrt{a}} \nabla\cdot\sqrt{a}\mathbf{u} \right\|_{0,\Omega_i}^2 + \epsilon_1 \left\|\sqrt{a}p\right\|_{0,\Omega_i}^2 + \frac{1}{\epsilon_2} \sum_m C_m^2 \eta^2 \sum_{n=1}^{N_m} |\beta_{m,n}|^2 \\
&\quad + 2\epsilon_2 \sum_{ij} \sum_{m:\Gamma_{ij}\cap\mathcal{P}_m\neq\emptyset} \left(\left\|\sqrt{a}_i p\right\|_{1/2,\Gamma_{ij}\cap\mathcal{P}_m}^2 + \left\|\sqrt{a}_j p\right\|_{1/2,\Gamma_{ij}\cap\mathcal{P}_m}^2 \right) \\
&\leq \frac{1}{\epsilon_1} \sum_{i=1}^{J} \left\| \frac{1}{\sqrt{a}} \nabla\cdot\sqrt{a}\mathbf{u} \right\|_{0,\Omega_i}^2 + \epsilon_1 c_4 \left\|\sqrt{a}\nabla p\right\|_{0,\Omega_i}^2 \\
&\quad + \frac{C\eta^2}{\epsilon_2} \sum_{i=1}^{J} \sum_{k:\Omega_i\cap\mathcal{P}_k\neq\emptyset} \sum_{n=1}^{N_m} |\beta_{m,n}|^2 \left\| \frac{1}{\sqrt{a}} \nabla\cdot\sqrt{a}\delta_m^h \tilde{\mathbf{s}}_{m,n} \right\|_{0,\Omega_i\cap\mathcal{P}_m}^2 + 2C_1\epsilon_2 \left\|\sqrt{a}\nabla p\right\|_{0,\Omega_i}^2 .
\end{aligned}$$

Choosing appropriate ϵ_1 and ϵ_2 yields

$$\begin{aligned}
\text{(A.15)}\quad \left\|\sqrt{a}\nabla p\right\|_{0,\Omega}^2 \leq C \Bigg(&\sum_{i=1}^{J} \left\| \frac{1}{\sqrt{a}} \nabla\cdot\sqrt{a}\mathbf{u} \right\|_{0,\Omega_i}^2 \\
&+ \eta^2 \sum_i \sum_{m:\Omega_i\cap\mathcal{P}_m\neq\emptyset} \sum_{n=1}^{N_m} |\beta_{m,n}|^2 \left\| \frac{1}{\sqrt{a}} \nabla\cdot\sqrt{a}\delta_m^h \tilde{\mathbf{s}}_{m,n} \right\|_{0,\Omega_i\cap\mathcal{P}_m}^2 \Bigg).
\end{aligned}$$

A similar result follows for the other term in the decomposition,

$$\text{(A.16)}\quad \left\|\frac{1}{\sqrt{a}}\nabla^{\perp}\psi\right\|_{0,\Omega}^2 \le C\left(\sum_{i=1}^{J}\left\|\sqrt{a}\nabla\times\frac{1}{\sqrt{a}}\mathbf{u}\right\|_{0,\Omega_i}^2 + \eta^2\sum_i\sum_{m:\Omega_i\cap\mathcal{P}_m\neq\emptyset}\sum_{n=1}^{N_m}|\beta_{m,n}|^2\left\|\sqrt{a}\nabla\times\frac{1}{\sqrt{a}}\delta_m^h\tilde{\mathbf{s}}_{m,n}\right\|_{0,\Omega_i\cap\mathcal{P}_m}^2\right).$$

Putting (A.15) and (A.16) together and applying Lemma A.2 yields

$$\text{(A.17)}\quad \begin{aligned}\|\mathbf{u}\|_{0,\Omega}^2 &= \left\|\sqrt{a}\nabla p\right\|_{0,\Omega}^2 + \left\|\frac{1}{\sqrt{a}}\nabla^{\perp}\psi\right\|_{0,\Omega}^2\\ &\le C|\mathbf{u}|_{\mathbf{W}^h}^2 + C\eta^2\sum_m\sum_{n=1}^{N_m}|\beta_{m,n}|^2|\delta_m^h\tilde{\mathbf{s}}_{m,n}|_{\mathbf{W}^h}^2\\ &\le\left(C_1+\frac{C_2\eta^2}{(1-\gamma^2)}\right)|\mathbf{u}|_{\mathbf{W}^h}^2.\end{aligned}$$

This completes the proof. □

COROLLARY A.4. *Under the assumption $\gamma = 1 - O(h^2)$, there are constant $C1$ and C_2, independent of h, such that*

$$\text{(A.18)}\quad \|\mathbf{u}\|_{0,\Omega}^2 \le \left(C_1 + C_2\frac{\eta^2}{h^2}\right)|\mathbf{u}|_{\mathbf{W}^h}.$$

Proof. The proof follows immediately from Theorem A.3. □

To eliminate h-dependence, we must ensure that $\eta = O(h)$. Recall that η is the error in the exponent of the singular basis function. In our numerical scheme for the calculation of the exponents, we have full control over their accuracy.

REFERENCES

[1] A. E. Berger, L. R. Scott, and G. Strang, *Approximate boundary conditions in the finite element method*, Sympos. Math., 10 (1972), pp. 295–313.

[2] M. Berndt, T. A. Manteuffel, S. F. McCormick, and G. Starke, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients: Part* I, SIAM J. Numer. Anal., 43 (2005), pp. 386–408.

[3] P. Bochev, Z. Cai, T. A. Manteuffel, and S. F. McCormick, *Analysis of velocity-flux first-order system least-squares principles for the Navier–Stokes equations: Part* I, SIAM J. Numer. Anal., 35 (1998), pp. 990–1009.

[4] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[5] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[6] J. H. Bramble and J. E. Pasciac, *Least-squares methods for the Stokes equations based on a discrete minus one inner product*, J. Comput. Appl. Math., 74 (1996), pp. 155–173.

[7] S. C. Brenner, *Multigrid methods for the computation of singular solutions and stress intensity factors* I: *Corner singularities*, Math. Comput., 68 (1999), pp. 559–583.

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Elements*, Springer-Verlag, New York, Berlin, 1994.

[9] S. C. Brenner and L. Y. Sung, *Multigrid methods for the computation of singular solutions and stress intensity factors* II: *Crack singularities*, BIT, 37 (1997), pp. 623–643.

[10] Z. Cai and S. Kim, *A finite element method using singular functions for the Poisson equation: Corner singularities*, SIAM J. Numer. Anal., 39 (2001), pp. 286–299.

[11] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[12] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[13] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for the Stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[14] Z. Cai, T. A. Manteuffel, S. F. McCormick, and J. Ruge, *First-order system LL* (FOSLL*): Scalar elliptic partial differential equations*, SIAM J. Numer. Anal., 39 (2001), pp. 1418–1445.

[15] T. F. Chen and G. J. Fix, *Least squares finite element simulation of transonic flows*, Appl. Numer. Math., 2 (1986), pp. 399–408.

[16] P. G. Ciarlet and J. L. Lions, *Finite Element Methods (Part* 1*)*, North–Holland, Amsterdam, 1990.

[17] C. L. Cox and G. J. Fix, *On the accuracy of least squares methods in the presence of corner singularities*, Comput. Math. Appl., 10 (1984), pp. 463–475.

[18] G. Fix and E. Stephan, *Finite Element Methods of the Least Squares Type for Regions with Corners*, Tech. Report 81-41, Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, VA, 1981.

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

[20] D. C. Jespersen, *A least-square decomposition method for solving elliptic systems*, Math. Comp., 31 (1977), pp. 873–880.

[21] B. N. Jiang and J. Z. Chai, *Least-squares finite element analysis of steady high subsonic plane potential flows*, Acta Mech. Sinica, 1 (1980), pp. 90–93 (in Chinese).

[22] B.-N. Jiang and C. L. Chang, *Least-squares finite elements for the Stokes problem*, Comput. Methods Appl. Mech. Engrg., 78 (1990), pp. 297–311.

[23] R. B. Kellogg, *Singularities in interface problems*, in Proceedings of the 2nd Symposium on Numerical Solution of Partial Differential Equations (SYNSPADE 1970), University of Maryland, 1971, pp. 351–400.

[24] T. A. Manteuffel, S. F. McCormick, and G. Starke, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients*, in Proceedings of the Seventh Copper Mountain Conference on Multigrid Methods, NASA Conference Publication 3339, NASA, Washington, DC, 1996, pp. 535–550.

[25] P. Neittaanmäki and J. Saranen, *On finite element approximation of the gradient for the solution to Poisson equation*, Numer. Math., 37 (1981), pp. 131–148.

[26] A. I. Pehlivanov and G. F. Carey, *Error estimates for least-squares mixed finite elements*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 499–516.

[27] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[28] U. Rüde, *On the accurate computation of singular solutions of Laplace's and Poisson's equation*, Multigrid Methods: Theory, Applications, and Supercomputing, 110 (1988), pp. 517–540.

[29] J. Ruge and K. Stüben, *Efficient solution of finite difference and finite element equations by algebraic multigrid (amg)*, in Multigrid Methods for Integral and Differential Equations, D. J. Paddon and H. Holstein, eds., Inst. Math. Appl. Conf. Ser. New Ser. 3, Oxford University Press, New York, 1985, pp. 169–212.

[30] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[31] K. Stüben and U. Trottenberg, *Multigrid methods: Fundamental algorithms, model problem analysis and applications*, in Multigrid Methods, W. Hackbusch and U. Trottenberg, eds., Lecture Notes in Math. 960, Springer–Verlag, Berlin, 1982, pp. 1–176.

[32] W. L. Wendland, *Elliptic Systems in the Plane*, Pitman, London, 1979.

[33] E. Zauderer, *Partial Differential Equations of Applied Mathematics*, 2nd ed., Ser. Pure Appl. Math., John Wiley & Sons, New York, 1988.

SIAM J. NUMER. ANAL.
Vol. 43, No. 1, pp. 437–454

ANALYSIS OF SOME PADÉ–CHEBYSHEV APPROXIMANTS*

SIDI MAHMOUD KABER† AND YVON MADAY†

Abstract. We present some rational approximations of the sign function and analyze their convergence. The rate of convergence is shown to increase with the degree of the denominator of the rational approximation. Several numerical tests are presented.

Key words. Padé approximation, approximation by rational functions

AMS subject classifications. 41A21, 41A20

DOI. 10.1137/S0036142903436708

1. Introduction.

The motivation of this work comes primarily from the spectral discretization of partial differential equations (consult the monographs [1] and [12]). For regular solutions, the numerical spectral approximation is highly accurate. On the opposite, for discontinuous solutions the Gibbs phenomenon prevents the high convergence of the numerical approximations. However, the numerical solution u_N is close to the projection of the solution u on $\mathbb{P}_N$, the space of algebraic polynomials of degree $\leq N$. In other terms, the coefficients of u_N are close to those of the solution u. The way to design an accurate solution from the knowledge of the Fourier (or general orthogonal expansions) coefficients of the solution is called filtering in the numerical analysis literature. We refer to [11] for the state-of-the-art on the spectral approximation of discontinuous solutions and the filtering problem.

What we present here is a filtering procedure (or acceleration of convergence) based on rational approximation. The problem can be stated as follows: given the first K Chebyshev coefficients $\hat{u}_k$ of a function u, design a polynomial $\mathcal{P}$ of degree N and a polynomial $\mathcal{Q}$ of degree M in such a way that the rational function $\mathcal{P}/\mathcal{Q}$ is a better approximation of u (in a sense to be made more precise) than the best approximation of u by K-degree polynomials, namely, the finite expansion $\sum_{k=0}^{K} \hat{u}_k T_k$.

The idea of approximating by rational functions comes from the theory of Padé approximants [18], [2]. There exist linear and nonlinear Padé approximants of a function u with power series

$$u(x) = \sum_{k\geq 0} u_k x^k.$$

- The linear Padé approximant of u of order (n, m) is a rational function $r = p/q$, with polynomials p (of degree n) and q (of degree m) defined by

  $$q(x)u(x) - p(x) = \mathcal{O}(x^{n+m+1}) \tag{1.1}$$

  as $x \to 0$. Plugging $p = \sum_{i=1}^{n} p_i x^i$ and $q = \sum_{j=1}^{m} q_j x^j$ into (1.1) and equating the coefficients of the powers x^k, we get the linear system to be solved (with unknowns p_i and q_j) in order to compute q and p.

*Received by the editors October 23, 2003; accepted for publication (in revised form) September 10, 2004; published electronically June 27, 2005. This work was supported by DOE grant DE.FG02.-98ER25346.

http://www.siam.org/journals/sinum/43-1/43670.html

†Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, Boite Courrier 187, 75252 Paris, Cedex 05, France (maday@ann.jussieu.fr). Part of this work was done while the first author was visiting the Department of Applied Mathematics, Brown University, Providence, RI.

- The nonlinear Padé approximants of u of order (n, m) are solutions $r = p/q$ of the following nonlinear problem:

$$\frac{p(x)}{q(x)} - u(x) = \mathcal{O}(x^{n+m+1}). \tag{1.2}$$

  Problem (1.1) always has a nontrivial solution, while problem (1.2) may have no solution. If a solution q of the linear Padé problem satisfies $q(0) \neq 0$, then $r = p/q$ is a solution of the nonlinear Padé problem.

Padé approximations can be generalized to polynomial expansions other than the powers x^k, namely, expansions in terms of orthogonal polynomials. For Fourier and Chebyshev expansions, consult [3], [13], [20] and the more recent reference [4].

In order to improve (by reducing the Gibbs phenomenon) the Fourier spectral method, some Padé–Fourier approximants were used in [10]. One can use the transformation $x = \cos\theta$ (then $T_k(x) = \cos(k\theta)$) to refer to the framework of [10]. The present work does not exploit this specific transformation and thus could be generalized to other orthogonal polynomial expansions. Let us specify that instead of the analyticity of the function u (required in the Padé approximations), we require here only that u belong to an L^2 space. General Padé–Jacobi expansions are presented in [6], where the case of quadratic denominators is analyzed. Numerical tests for the Legendre case are also presented in this reference. Recursive algorithms to compute some Padé–Legendre approximants are given in [14]. The first references on the Padé–Legendre approximants we are aware of are [8] and [9]. Several properties of the Padé–Legendre approximants can be found in [5], as well as various numerical tests showing that the rate of convergence of these approximations increases with the degree of their denominators. This is precisely what we prove in this work for Padé–Chebyshev approximants.

The paper is organized as follows: The remainder of this section is devoted to some notation and properties of some special functions. In section 2, we recall the limits of the polynomial approximation of discontinuous functions. Section 3 introduces the rational approximations, which are analyzed in section 4. Some numerical tests that confirm the analysis are also given. The paper ends with a list of some directions of work.

Notation.

- $\mathbb{P}_n$ is the set of algebraic polynomials of degree $\leq n$.
- $\mathbb{R}_{n,m}$ is the set of rational functions $r = p/q$, with $p \in \mathbb{P}_n$ and $q \in \mathbb{P}_m$.
- Γ denotes Euler's gamma function

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt.$$

  Note that $\Gamma(z+1) = z\Gamma(z) = z!$. We will make use of the formula (see [7, p. 3, eq. (6)])[1]

$$\Gamma(1+z)\Gamma(1-z) = \frac{\pi z}{\sin \pi z},$$

  with $z = n + 1/2$, $n \in \mathbb{N}$:

$$\Gamma\left(\frac{3}{2}+n\right)\Gamma\left(\frac{1}{2}-n\right) = (-1)^n\left(n+\frac{1}{2}\right)\pi \quad \forall n \in \mathbb{N}. \tag{1.3}$$

[1] We refer to [7] for all the properties of the special functions used in this work.

- For $z \in \mathbb{C}$ and $n \in \mathbb{N}$, $(z)_n$ is the Pochhammer symbol defined by

  $$(z)_n = \frac{\Gamma(z+n)}{\Gamma(z)} = \begin{cases} 1 & \text{if } n = 0, \\ z(z+1)\dots(z+n-1) & \text{if } n \geq 1. \end{cases} \tag{1.4}$$

  Note that $(1)_n = n!$ and for all $z \in \mathbb{N}$

  $$(-z)_n = 0 \quad \forall n \geq z+1.$$

- ${}_pF_q(x_1, \dots, x_p; y_1, \dots, y_q; .)$ is the hypergeometric function defined for $z \in \mathbb{C}$ by

  $${}_pF_q(x_1, \dots, x_p; y_1, \dots, y_q; z) = \sum_{k \geq 0} \frac{(x_1)_k \dots (x_p)_k}{(y_1)_k \dots (y_q)_k} \frac{z^k}{k!}. \tag{1.5}$$

  The sum in (1.5) is finite if one of the argument $x_i \in -\mathbb{N}$. In that case ${}_pF_q$ is an element of $\mathbb{P}_{-x_i}$. The particular case ($n \in \mathbb{N}$)

  $${}_3F_2(a, b, -n; d, e; z) = \sum_{k=0}^{n} \frac{(a)_k (b)_k (-n)_k}{(d)_k (e)_k} \frac{z^k}{k!} \quad (\in \mathbb{P}_n)$$

  with $e + d + n = 1 + a + b$ leads to the Saalchütz formula (see [7, p. 66, eq. (30)]),

  $${}_3F_2(a, b, -n; d, e; 1) = \frac{(d-a)_n (d-b)_n}{(d)_n (d-a-b)_n}. \tag{1.6}$$

2. Polynomial approximation of a discontinuous function. Let $\omega(x) = 1/\sqrt{1-x^2}$ be the standard Chebyshev weight. On the space L^2_ω of the functions $u : I =]-1, 1[\to \mathbb{R}$ such that $\int_I u^2(x)\omega(x)dx$ is finite, we define the scalar product

$$\langle u, v \rangle_\omega = \int_I u(x)v(x)\omega(x)dx \qquad \forall (u, v) \in L^2_\omega \times L^2_\omega \tag{2.1}$$

and the norm

$$\|u\|_\omega = \sqrt{\langle u, u \rangle_\omega} = \left(\int_I u(x)^2 \omega(x) \right)^{1/2} \qquad \forall u \in L^2_\omega.$$

The Chebyshev expansion of a function $u \in L^2_\omega$ is

$$u = \sum_{k=0}^{\infty} \hat{u}_k T_k, \quad \hat{u}_k = \frac{1}{\|T_k\|^2_\omega} \langle u, T_k \rangle_\omega,$$

with T_k the Chebyshev polynomial of degree k and

$$\|T_k\|_\omega = \begin{cases} \sqrt{\pi} & \text{if } k = 0, \\ \sqrt{\pi/2} & \text{if } k \geq 1. \end{cases}$$

The truncated series $\pi^n_\omega(u) = \sum_{k=0}^n \hat{u}_k T_k$ is the best approximation of u in $\mathbb{P}_n$ in the sense that

$$\|\pi^n_\omega(u) - u\|_\omega \ \leq \ \|q - u\|_\omega \quad \forall q \in \mathbb{P}_n.$$

It is also the projection of u onto $\mathbb{P}_n$:

$$\langle \pi_\omega^n(u), \varphi \rangle_\omega = \langle u, \varphi \rangle_\omega \quad \forall \varphi \in \mathbb{P}_n. \tag{2.2}$$

The approximation error is

$$e_{\omega,n}(u) = \|\pi_\omega^n(u) - u\|_\omega = \left(\frac{2}{\pi} \sum_{k>n} |\hat{u}_k|^2 \right)^{1/2} .$$

If not only u belongs to L^2_ω but also all the derivatives of u up to s, then the following decrease of the error holds:

$$e_{\omega,n}(u) \leq \mathrm{Const}(s) n^{-s} \left(\sum_{k=0}^{s} \|u^{(k)}\|_\omega^2 \right)^{1/2} .$$

This shows the rapid decay of the error if the function u is regular enough. On the contrary, let us consider the sign function

$$S(x) = \begin{cases} -1 & \text{for } x \in [-1, 0[, \\ \;\;1 & \text{for } x \in]0, 1], \end{cases}$$

with Chebyshev expansion $S = \sum_{k=0}^{\infty} \hat{s}_k T_k$. Straightforward computations give

$$\hat{s}_{2k} = 0, \quad \hat{s}_{2k+1} = \frac{4}{\pi} \frac{(-1)^k}{2k+1}. \tag{2.3}$$

Hence

$$S = \sum_{k=0}^{\infty} \hat{s}_{2k+1} T_{2k+1}, \tag{2.4}$$

and the approximation error is

$$e_{\omega,2n+1}(S) = \sqrt{\frac{8}{\pi}} \left(\sum_{k>n} \frac{1}{(2k+1)^2} \right)^{1/2} ,$$

which gives rise to the Gibbs phenomenon.

Gibbs phenomenon. There exists a constant C such that the equivalence

$$e_{\omega,2n+1}(S) \simeq \frac{C}{\sqrt{n}} \quad \text{as } n \to +\infty \tag{2.5}$$

holds. This shows the slow convergence of $\pi_\omega^n(S)$ toward S. On Figure 2.1 several approximations $\pi_\omega^n(S)$ are displayed. In particular, one can notice oscillations of order $\mathcal{O}(1)$ near the singularity $x = 0$. This is the Gibbs phenomenon. More precisely there exists a sequence $(x_n)_n \to 0^+$ such that [12]

$$\pi_\omega^n(S)(x_n) \to \gamma > S(0) = 0.$$

Filtering. The filtering procedures try to cancel the oscillations. Such oscillations (inherent to the Gibbs phenomenon) appear when discontinuous functions are approximated by polynomials (or other globally defined smooth functions, e.g., wavelets). This is the typical situation when using spectral methods to compute solutions of hyperbolic equations. We refer again to [11] for this problem.

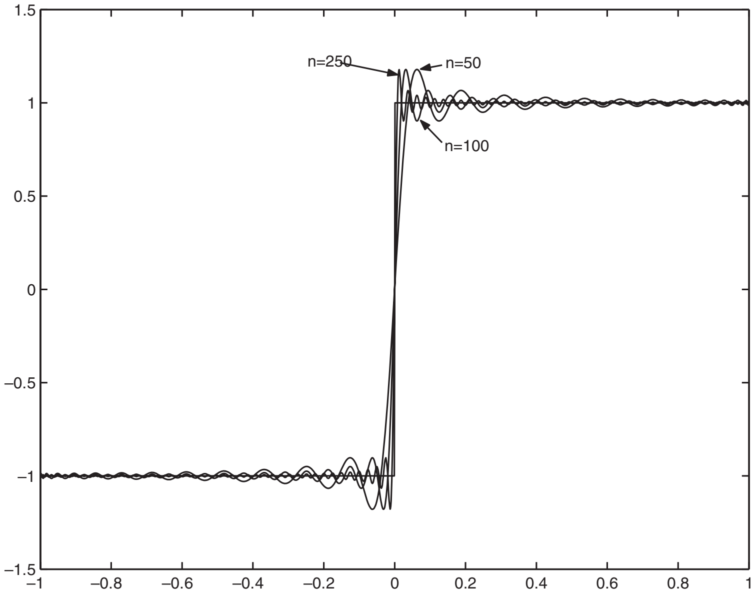


FIG. 2.1. *The sign function S and $\pi_\omega^n(S)$ for $n = 50, 100, 250$.*

Rational versus polynomial approximation. Let us recall the following result proved by Newman in [16]: there exists a rational approximation $r \in \mathbb{P}_{n,n}$ such that

$$\max_{-1\le x\le 1} |\, r(x) - |x|\, | \le 3e^{-\sqrt{n}}.$$

What is remarkable in this result is that polynomial approximation of $|x|$ is only of first order. Namely there exists $p \in \mathbb{P}_n$ such that

$$\forall q \in \mathbb{P}_n, \qquad \max_{-1\le x\le 1} |\, q(x) - |x|\, | \ge \max_{-1\le x\le 1} |\, p(x) - |x|\, | \simeq \frac{\text{Const}}{n}.$$

Newman's result motivates the use of rational approximations, especially when polynomial approximation fails to provide highly accurate approximation.

The main problem is related to the fact that $\mathbb{R}_{n,m}$ is not a linear space, unlike $\mathbb{P}_n$.

3. A rational approximation.

We explain how to define a suitable rational approximation $\mathcal{R}_{N,M}(u)$ of a function u having in mind the special case of the sign function.

$$\mathcal{R}_{N,M}(u) = \frac{\mathcal{P}_{N,M}}{\mathcal{Q}_{N,M}} \in \mathbb{R}_{N,M}.$$

Since $\mathbb{P}_N = \mathbb{R}_{N,0}$, rational approximation obviously contains polynomial approximation. The goal here is to improve the convergence rate, i.e.,

$$\|\mathcal{R}_{N,M}(u) - u\|_\omega \ll \|\pi_\omega^{N+M}(u) - u\|_\omega. \tag{3.1}$$

For the sake of simplicity and when this does not cause any confusion, we will drop the subscripts ${}_{N,M}$. Just like in the classical Padé setting, there are two ways to define the rational approximation $\mathcal{P}/\mathcal{Q}$.

3.1. Definition of the Padé–Chebyshev approximants.

Nonlinear Padé–Chebyshev approximation. The rational approximation $\mathcal{R}(u) = \mathcal{P}/\mathcal{Q}$ is defined by the orthogonality conditions

$$\left\langle \frac{\mathcal{P}}{\mathcal{Q}} - u, \varphi \right\rangle_{\omega} = 0 \qquad \forall \varphi \in \mathbb{P}_K, \tag{3.2}$$

with K as large as possible. That is to say, $\mathcal{R}(u) - u$ is orthogonal to $\mathbb{P}_K$ with respect to the scalar product (2.1). Equations (3.2) form a nonlinear system of $K+1$ nonlinear equations and $N+M+2$ unknowns. This system could have no solution, but if it admits a solution and if $\mathcal{Q}$ never vanishes on I, then this solution is unique [19]. The nonlinear system is hard to handle, especially for general orthogonal series (Legendre, for instance). Actually, this problem was solved in [10] for trigonometric polynomials. Using the transformation $x = \cos\theta$ a Chebyshev expansion becomes a trigonometric expansion, and the tools in [10] can be used. But such a transformation does not exist for other orthogonal polynomials. The present work does not exploit this specific transformation in order to be generalized to other orthogonal polynomial expansions.

Note that, for the sign function, explicit expression for the coefficients of the Padé–Chebyshev is given in [17], but the method cannot be generalized to other functions.

Note that from (3.2), the first Chebyshev coefficients of $\mathcal{R}(u)$ and that of the function u coincide. Note also that only a finite portion of the spectrum of u is required to define $\mathcal{R}(u)$. The complete knowledge of u (i.e., of $\hat{u}_k$ for all $k \in \mathbb{Z}$) is not necessary.

Linear Padé–Chebyshev approximation. Here one imposes the orthogonality relations

$$\langle \mathcal{Q}u - \mathcal{P}, \varphi \rangle_{\omega} = 0 \qquad \forall \varphi \in \mathbb{P}_K. \tag{3.3}$$

This means that $\mathcal{Q}u - \mathcal{P}$ is orthogonal to $\mathbb{P}_K$, or, in other terms, the function $\mathcal{Q}u - \mathcal{P}$ has only high frequencies,

$$\mathcal{Q}u - \mathcal{P} = \sum_{k>K} \gamma'_k T_k. \tag{3.4}$$

Writing (3.3) for $k = N+1, \dots, K$ gives the linear system to be solved by $\mathcal{Q}$:

$$\widehat{(\mathcal{Q}u)}_k = 0 \quad \forall k = N+1, \dots, K. \tag{3.5}$$

This homogeneous linear system has $M+1$ unknowns and $K-N$ equations. For $K \leq N+M$ the system admits a nonpotentially trivial solution. Note that the highest mode of u involved in system (3.5) is $\hat{u}_{K+M}$. Here also the complete knowledge of u is not required.

From now on, we consider only linear Padé approximation. Some properties of this rational approximation follow. A basic one is the reproduction of polynomials and rational functions if K is large enough.

3.2. Some general properties. The problem of existence and uniqueness of the rational approximation $\mathcal{R}$ is that of the existence and uniqueness of the denominator $\mathcal{Q}$. In the general case, uniqueness is not guaranteed. Let us consider the example $n \geq 1$, $K = N = n+1$, and $M = 1$ in (3.3). From the recurrence formula

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x),$$

we deduce that $2xT_n(x) - T_{n-1}(x)$ is orthogonal to $\mathbb{P}_n$ with respect to the weight ω. Hence $\frac{T_{n-1}}{2x} \in \mathbb{P}_{n,1}$ is a linear Padé approximant of T_n which is of course its own Padé approximant in $\mathbb{P}_{n,1}$.

One can insure uniqueness using the appropriate parameters N, M, and K.

Proposition 3.1 (uniqueness). *If $(\mathcal{Q}_1, \mathcal{P}_1)$ and $(\mathcal{Q}_2, \mathcal{P}_2)$ are two solutions of* (3.3)*, then for K large enough ($K \geq M + 2N$), they determine the same Padé–Chebyshev solution:*

$$\frac{\mathcal{P}_1}{\mathcal{Q}_1} = \frac{\mathcal{P}_2}{\mathcal{Q}_2}.$$

Proof. We first suppose $u > 0$ and prove that the polynomial $\mathcal{P}_1\mathcal{Q}_2 - \mathcal{P}_2\mathcal{Q}_1 \in \mathbb{P}_{M+N}$ is orthogonal to $\mathbb{P}_{K-N}$ with respect to the weight function $u(x)\omega(x)$. Hence, for $K - N \geq N + M$, $\mathcal{P}_1\mathcal{Q}_2 - \mathcal{P}_2\mathcal{Q}_1$ is the null polynomial. Indeed

$$\begin{aligned}\langle(\mathcal{P}_1\mathcal{Q}_2 - \mathcal{P}_2\mathcal{Q}_1)u, \varphi\rangle_\omega &= \langle u\mathcal{Q}_2, \mathcal{P}_1\varphi\rangle_\omega - \langle u\mathcal{Q}_1, \mathcal{P}_2\varphi\rangle_\omega,\\ &= \langle \mathcal{P}_2, \mathcal{P}_1\varphi\rangle_\omega - \langle \mathcal{P}_1, \mathcal{P}_2\varphi\rangle_\omega = 0 \qquad \forall\varphi \in \mathbb{P}_{K-N}.\end{aligned}$$

In the general case, consider the strictly positive function $v = u + M$ with $M > \|u\|_\infty$. □

Proposition 3.2. *If $u \in \mathbb{P}_N$ or $u \in \mathbb{R}_{N,M}$ with nonvanishing denominator, then $\mathcal{R}_{N,M}(u) = u$, provided $K \geq M + N$ in* (3.3).

Proof. If $u \in \mathbb{P}_N$, then the function $\mathcal{Q}u$ belongs to $\mathbb{P}_K$ and relations (3.4) imply $\mathcal{P} = \mathcal{Q}u$, that is, $\mathcal{R}_{N,M}(u) = u$. Now let $u = p/q$ with $p \in \mathbb{P}_N$, $q \in \mathbb{P}_M$, $q(x) \neq 0$, hence assumed $\mathcal{Q} > 0$ on I. The orthogonality relations (3.3) read, for all $\varphi \in \mathbb{P}_K$,

$$0 = \left\langle \mathcal{P} - \mathcal{Q}\frac{p}{q}, \varphi \right\rangle_\omega = \int_I (q\mathcal{P} - p\mathcal{Q})\,\frac{\varphi}{q}\omega(x)dx.$$

In other terms, the polynomial $q\mathcal{P} - p\mathcal{Q}$ is orthogonal to $\mathbb{P}_K$ with respect to the inner product defined by the function ω/q. But this polynomial has degree less than K, and hence it is the null polynomial and $\mathcal{R}_{N,M}(u) = u$. □

Proposition 3.3 (parity). *For $K \geq M + 2N$, and assuming $\mathcal{Q}(0) \neq 0$,*

$$\begin{cases} u \text{ is odd} & \Longrightarrow \quad \mathcal{Q} \text{ is even and } \mathcal{P} \text{ is odd},\\ u \text{ is even} & \Longrightarrow \quad \mathcal{Q} \text{ is even and } \mathcal{P} \text{ is even.}\end{cases}$$

Proof. Consider an odd function u. For all $\varphi \in \mathbb{P}_K$ the orthogonality relations give

$$\int_{-1}^{1} [\mathcal{Q}(-x)u(-x) - \mathcal{P}(-x)]\,\varphi(-x)\omega(-x)dx = 0,$$

or, equivalently, for all $\psi \in \mathbb{P}_K$,

$$\int_{-1}^{1} [\mathcal{Q}(-x)u(x) + \mathcal{P}(-x)]\,\psi(x)\omega(x)dx.$$

Hence $(-\mathcal{P}(-x), \mathcal{Q}(-x))$ is a couple of Padé approximants of u of order (N, M). Assuming $K \geq M+2N$, we know from Proposition 3.1 that $\mathcal{P}(x)\mathcal{Q}(-x) = -\mathcal{Q}(x)\mathcal{P}(-x)$. Hence the polynomial $\mathcal{P}(x)$ is a divisor of the polynomial $\mathcal{P}(-x)\mathcal{Q}(x)$. If $\mathcal{P}(x)/\mathcal{Q}(x)$ is irreducible, then $\mathcal{P}(x)$ is a divisor of $\mathcal{P}(-x)$, which means that $\mathcal{P}$ is an odd or an even polynomial, in which case $\mathcal{Q}$ is an even or an odd polynomial, respectively. Assuming that $\mathcal{Q}(0) \neq 0$, then $\mathcal{P}$ is even. The conclusion holds true if $\mathcal{P}(x)/\mathcal{Q}(x)$ is not irreducible by considering a common polynomial factor S such that $\mathcal{P} = S\mathcal{P}_1$ and $\mathcal{Q} = S\mathcal{Q}_1$ with $\mathcal{P}_1/\mathcal{Q}_1$ irreducible. □

Proposition 3.4. *Let $\mathcal{Q}$ be a solution of the system* (3.5) *with $K \geq N + M$. Then the numerator of the (linear) Padé approximation is*

$$\mathcal{P} = \pi_\omega^N\left(\mathcal{Q}u\right) = \pi_\omega^K\left(\mathcal{Q}u\right) = \pi_\omega^N\left(\mathcal{Q}\pi_\omega^K(u)\right). \tag{3.6}$$

Proof. The orthogonality relations (3.3) directly imply the first two equalities. For the last one, write, for all $k \leq N$,

$$\begin{aligned}\langle \pi_\omega^N\left(\mathcal{Q}\pi_\omega^K(u)\right), T_k\rangle_\omega &= \langle \mathcal{Q}\pi_\omega^K(u), T_k\rangle_\omega = \langle \pi_\omega^K(u), \mathcal{Q}T_k\rangle_\omega \\ &= \langle u, \mathcal{Q}T_k\rangle_\omega = \langle \mathcal{Q}u, T_k\rangle_\omega \\ &= \langle \pi_\omega^N\left(\mathcal{Q}(u)\right), T_k\rangle_\omega = \langle \mathcal{P}, T_k\rangle_\omega,\end{aligned}$$

which yields the statement. □

Before concentrating on the special case of the sign function, we list the different steps of our goal.

- First of all, compute the denominator $\mathcal{Q}$. This is the main problem.
- Define the numerator $\mathcal{P}$ by relations (3.6):

$$\mathcal{P} = \pi_\omega^N\left(\mathcal{Q}\pi_\omega^K u\right). \tag{3.7}$$

- Define the rational approximation $\mathcal{R}$ by

$$\mathcal{R}(u) = \frac{\mathcal{P}}{\mathcal{Q}}. \tag{3.8}$$

Once the rational approximation is defined, we have to

- prove the convergence

$$\lim_{N\to+\infty} \mathcal{R}(u) = u, \tag{3.9}$$

  in a sense to be specified later;
- evaluate the error $\|\mathcal{R}(u) - u\|_\omega$ and show that $\mathcal{R}$ behaves like a Chebyshev series whose coefficients decrease faster than the coefficients $\hat{u}_k$ of u.

From now on we consider only the sign function.

4. The special case of the sign function. Taking into account the parity of the sign function (see Proposition 3.3), we search for even $\mathcal{Q}$ and odd $\mathcal{P}$. We expand $\mathcal{P}$ in the Chebyshev basis,

$$\mathcal{P} = \sum_{n=0}^{\mathcal{N}} \hat{p}_{2n+1}T_{2n+1},$$

and $\mathcal{Q}$ in either the Chebyshev basis or the usual canonical basis:

$$\mathcal{Q} = \sum_{m=0}^{\mathcal{M}} q_{2m}T_{2m} \quad \text{or} \quad \mathcal{Q}(x) = \sum_{m=0}^{\mathcal{M}} q_{2m}x^{2m}.$$

4.1. How to compute the denominator. Let us develop the denominator $\mathcal{Q}$ in a basis $(\varphi_{_{2m}})_{m=0}^{\mathcal{M}}$ of even polynomials of $\mathbb{P}_{2\mathcal{M}}$:

$$\mathcal{Q} = \sum_{m=0}^{\mathcal{M}} q_{2m}\varphi_{_{2m}}.$$

The expansion of $\varphi_{_{2m}} S$ in the Chebyshev basis is

$$\varphi_{_{2m}} S = \sum_{k\geq 0} c_{2k+1}^{m} T_{2k+1},$$

with

$$c_{2k+1}^{m} = \frac{4}{\pi}\int_0^1 \varphi_{_{2m}}(x) T_{2k+1}(x)\omega(x)dx. \tag{4.1}$$

Hence the expansion of $\mathcal{Q}S$ in the Chebyshev basis is

$$\mathcal{Q}S = \sum_{k\geq 0} \Lambda_{2k+1}^{(\mathcal{M})} T_{2k+1},$$

with

$$\Lambda_{2k+1}^{(\mathcal{M})} = \sum_{m=0}^{\mathcal{M}} c_{2k+1}^{m} q_{2m}. \tag{4.2}$$

In order to satisfy (3.3), the idea is to split $\mathcal{Q}S$ into three terms,

$$\mathcal{Q}S = \sum_{k=0}^{\mathcal{N}} \Lambda_{2k+1}^{(\mathcal{M})} T_{2k+1} + \sum_{k=\mathcal{N}+1}^{K} \Lambda_{2k+1}^{(\mathcal{M})} T_{2k+1} + \sum_{k>K} \Lambda_{2k+1}^{(\mathcal{M})} T_{2k+1}. \tag{4.3}$$

The first term in the right-hand side will be the numerator of the rational approximation, the second term must vanish, and the last one is the remainder.

Determining the denominator is equivalent to finding $(q_{2m})_{m=0}^{\mathcal{M}}$ such that

$$\Lambda_{2k+1}^{(\mathcal{M})} = 0 \quad \forall k = \mathcal{N}+1, \ldots, K. \tag{4.4}$$

Remark 4.1. Problem (4.4) is a linear system of $K - \mathcal{N}$ equations and $\mathcal{M}+1$ unknowns. For $K \leq \mathcal{N} + \mathcal{M}$, the system always has a nontrivial solution. In what follows we fix

$$K = \mathcal{N} + \mathcal{M} \tag{4.5}$$

in order to ensure the existence of a nonzero solution. But we don't know if such a solution is unique since uniqueness is guaranteed only for $K \geq \mathcal{M} + 2\mathcal{N}$ (see Proposition 3.1).

4.1.1. First method. Let us first expand $\mathcal{Q}$ in the Chebyshev basis:

$$\mathcal{Q} = \sum_{m=0}^{\mathcal{M}} q_{2m} T_{2m}. \tag{4.6}$$

In this case

$$\begin{aligned} c_{2k+1}^m &= \frac{4}{\pi}\int_0^1 T_{2m}(x)T_{2k+1}(x)\omega(x)dx \\ &= \frac{4}{\pi}\int_0^{\pi/2} \cos(2m\theta)\cos[(2k+1)\theta]d\theta \\ &= (-1)^{m+k+1}\frac{4}{\pi}\frac{2k+1}{(2m+2k+1)(2m-2k-1)} \end{aligned}$$

and

$$\Lambda_{2k+1}^{(\mathcal{M})} = (-1)^{k+1}\frac{4}{\pi}(2k+1)\sum_{m=0}^{\mathcal{M}} \frac{(-1)^m}{(2m+2k+1)(2m-2k-1)}q_{2m}. \tag{4.7}$$

The next Proposition gives a solution of (4.4) expressed in the Chebyshev basis.

PROPOSITION 4.2. *The coefficients* q_{2m} *defined by*

$$q_{2m} = \frac{(-1)^m}{m!}\frac{(\mathcal{N}+\mathcal{M}+3/2)_m(-\mathcal{M})_m}{(\mathcal{N}+1/2)_m} \tag{4.8}$$

are solutions of the linear system (4.4)–(4.2)–(4.5).

Proof. We want to prove that for $k = \mathcal{N}+1, \dots, \mathcal{N}+\mathcal{M}$,

$$\sum_{m=0}^{\mathcal{M}} (-1)^m \frac{1}{(2m+2k+1)(2m-2k-1)}q_{2m} = 0, \tag{4.9}$$

or, equivalently,

$$\sum_{m=0}^{\mathcal{M}} (-1)^m \frac{(k-1/2)_m}{(k+3/2)_m}q_{2m} = 0,$$

by using the identities

$$2m+2k+1 = 2\left(k+\frac{1}{2}\right)\frac{(k+3/2)_m}{(k+1/2)_m}, \quad 2m-2k-1 = 2\left(k-\frac{1}{2}\right)\frac{(k+1/2)_m}{(k-1/2)_m}.$$

With the given values of q_{2m}, we have

$$\begin{aligned} \sum_{m=0}^{\mathcal{M}}(-1)^m\frac{(k-1/2)_m}{(k+3/2)_m}q_{2m} &= \sum_{m=0}^{\mathcal{M}}\frac{1}{m!}\frac{(k-1/2)_m(\mathcal{N}+\mathcal{M}+3/2)_m(-\mathcal{M})_m}{(k+3/2)_m(\mathcal{N}+1/2)_m} \\ &= {}_3F_2\left(k-\frac{1}{2}, \mathcal{N}+\mathcal{M}+\frac{3}{2}, -\mathcal{M}; k+\frac{3}{2}, \mathcal{N}+\frac{1}{2}; 1\right) \\ &= \frac{(2)_{\mathcal{M}}(k-\mathcal{N}-\mathcal{M})_{\mathcal{M}}}{(k+3/2)_{\mathcal{M}}(1/2-\mathcal{N}-\mathcal{M})_{\mathcal{M}}}. \end{aligned}$$

The last equality uses the Saalchütz formula (1.6). The factor $(k-\mathcal{N}-\mathcal{M})_{\mathcal{M}}$ vanishes for $k = \mathcal{N}+1, \dots, \mathcal{N}+\mathcal{M}$, and hence (4.9) holds. $\square$

Let us prove that asymptotically (i.e., when $\mathcal{N} \to \infty$) the denominators with degrees $\mathcal{M} = 2(2\mathcal{M}' + 1)$ vanish at $x = \pm 1/\sqrt{2}$. For fixed $\mathcal{M}$ and x, we define

$$\begin{aligned}\mathcal{Q}_{\infty,\mathcal{M}}(x) &= \lim_{\mathcal{N}\to+\infty} \mathcal{Q}_{\mathcal{N},\mathcal{M}}(x) = \sum_{m=0}^{\mathcal{M}} \left[\lim_{\mathcal{N}\to+\infty} q_{2m}\right] T_{2m}(x)\\ &= \sum_{m=0}^{\mathcal{M}} \frac{\mathcal{M}!}{m!(\mathcal{M}-m)!} T_{2m}(x) = \sum_{m=0}^{\mathcal{M}} \left(\begin{array}{c}\mathcal{M}\\ m\end{array}\right) T_{2m}(x).\end{aligned}$$

Hence

$$\begin{aligned}\mathcal{Q}_{\infty,\mathcal{M}}(\pm 1/\sqrt{2}) &= \sum_{m=0}^{\mathcal{M}} \left(\begin{array}{c}\mathcal{M}\\ m\end{array}\right) T_{2m}(1/\sqrt{2}) = \sum_{m=0}^{\mathcal{M}} \left(\begin{array}{c}\mathcal{M}\\ m\end{array}\right) \cos\left(m\frac{\pi}{2}\right)\\ &= \text{Real}\left[\sum_{m=0}^{\mathcal{M}} \left(\begin{array}{c}\mathcal{M}\\ m\end{array}\right) e^{im\frac{\pi}{2}}\right] = \text{Real}\left[(1+e^{i\frac{\pi}{2}})^{\mathcal{M}}\right]\\ &= \text{Real}\left[(1+i)^{\mathcal{M}}\right] = \text{Real}\left[\left((1+i)^2\right)^{2\mathcal{M}'+1}\right] = 0.\end{aligned}$$

Remark 4.3. The denominator computed by the first method may have zeros inside I. On the other hand $\mathcal{Q}(x)$ is the sum of term with different signs. These two properties are unsafe for numerical purposes.

4.1.2. Second method. We now expand $\mathcal{Q}$ in the canonical basis:

$$\mathcal{Q}(x) = \sum_{m=0}^{\mathcal{M}} q_{2m} x^{2m}. \tag{4.10}$$

In that case, using equation (30) of [7, p. 12], we get

$$\begin{aligned}c_{2k+1}^{m} &= \frac{4}{\pi}\int_0^1 x^{2m} T_{2k+1}(x)\omega(x)dx\\ &= \frac{4}{\pi}\int_0^{\pi/2} (\cos\theta)^{2m} \cos[(2k+1)\theta]d\theta\\ &= \frac{\Gamma(2m+1)}{2^{2m-1}} \frac{1}{\Gamma(m+k+3/2)\Gamma(m-k+1/2)}.\end{aligned}$$

Straightforward computations give

$$\left(\frac{1}{2}\right)_m = \prod_{j=1}^{m} \frac{2j-1}{2} = \frac{(2m)!}{m!\,2^{2m}} = \frac{\Gamma(2m+1)}{m!\,2^{2m}}.$$

From this, we deduce

$$c_{2k+1}^{m} = 2\frac{m!\left(\frac{1}{2}\right)_m}{\Gamma(m+k+3/2)\Gamma(m-k+1/2)}$$

and

$$\Lambda_{2k+1}^{(\mathcal{M})} = 2\sum_{m=0}^{\mathcal{M}} \frac{m!\;\left(\frac{1}{2}\right)_m}{\Gamma(m+k+3/2)\Gamma(m-k+1/2)} q_{2m}. \tag{4.11}$$

Remark 4.4. Using (1.3), we get

$$\Gamma(m+k+3/2)\Gamma(m-k+1/2) = (3/2+k)_m(1/2-k)_m(-1)^k(k+1/2)\pi$$

and an expression of $\Lambda_{2k+1}^{(\mathcal{M})}$ in terms of the coefficients of the sign function given in (2.3),

$$\Lambda_{2k+1}^{(\mathcal{M})} = \hat{s}_{2k+1} \underbrace{\sum_{m=0}^{\mathcal{M}} \frac{m!\,\left(\frac{1}{2}\right)_m}{(3/2+k)_m(1/2-k)_m} q_{2m}}_{=\sigma_k^{(\mathcal{M})}}. \tag{4.12}$$

The next proposition (taken from [15]) gives a solution of (4.4) expressed in the canonical basis.

Proposition 4.5. *The coefficients* q_{2m} *defined by*

$$q_{2m} = \frac{(-\mathcal{M})_m(-\mathcal{N}-1/2)_m(\mathcal{N}+\mathcal{M}+3/2)_m}{(m!)^2(1/2)_m}, \qquad m = 0,\dots,\mathcal{M}, \tag{4.13}$$

are solutions of the linear system (4.4)–(4.2)–(4.5).

Proof. Using (4.12) it is sufficient to compute $\sigma_k^{(\mathcal{M})}$ for $k = \mathcal{N}+1,\dots,\mathcal{N}+\mathcal{M}$:

$$\begin{aligned}
\sigma_k^{(\mathcal{M})} &= \sum_{m=0}^{\mathcal{M}} \frac{m!\,\left(\frac{1}{2}\right)_m}{(3/2+k)_m(1/2-k)_m} q_{2m} \\
&= \sum_{m=0}^{\mathcal{M}} \frac{(-\mathcal{M})_m(-\mathcal{N}-1/2)_m(\mathcal{N}+\mathcal{M}+3/2)_m}{(3/2+k)_m(1/2-k)_m} \frac{1}{m!} \\
&= {}_3F_2(-\mathcal{M}, -\mathcal{N}-1/2, \mathcal{N}+\mathcal{M}+3/2; 3/2+k, 1/2-k; 1).
\end{aligned}$$

Using the Saalchütz formula (1.6), we obtain

$$\sigma_k^{(\mathcal{M})} = \frac{(k+\mathcal{N}+2)_{\mathcal{M}}(k-\mathcal{N}-\mathcal{M})_{\mathcal{M}}}{(3/2+k)_{\mathcal{M}}(k-\mathcal{M}+1/2)_{\mathcal{M}}}. \tag{4.14}$$

For $k = \mathcal{N}+1,\dots,\mathcal{N}+\mathcal{M}$, the three factors $(k+\mathcal{N}+2)_{\mathcal{M}}$, $(3/2+k)_{\mathcal{M}}$, and $(k-\mathcal{M}+1/2)_{\mathcal{M}}$ never vanish while $(k-\mathcal{N}-\mathcal{M})_{\mathcal{M}} = 0$. □

We get an explicit representation of $\mathcal{Q}$:

$$\mathcal{Q}(x) = \sum_{m=0}^{\mathcal{M}} \frac{(-\mathcal{M})_m(-\mathcal{N}-1/2)_m(\mathcal{N}+\mathcal{M}+3/2)_m}{(m!)^2(1/2)_m} x^{2m}.$$

Let us now give some properties of $\mathcal{Q}$.

Remark 4.6 (sign of the denominator). For $\mathcal{N} \geq \mathcal{M}-1$ and $m \in [0,\mathcal{M}]$, $(-\mathcal{M})_m(-\mathcal{N}-1/2)_m > 0$. Hence $\mathcal{Q}(x)$ is the sum of positive terms, which is nice for the stability of the numerical computations. This was not the case in the previous method to compute the denominator, as mentioned in Remark 4.3. For $x > 0$, $\mathcal{Q}$ is a monotone increasing function, as it is a positive linear combination of monotone functions. Hence

$$\mathcal{Q}(x) > \mathcal{Q}(0) = 1.$$

For $\mathcal{M} = 1$, the roots of $\mathcal{Q}_{\mathcal{N},1}(x) = 1 + (2\mathcal{N} + 1)(\mathcal{N} + 5/2)x^2$ are imaginary $\simeq \pm\frac{i}{\sqrt{2\mathcal{N}}}$ and approach zero as $\mathcal{N} \to +\infty$. In the general case $\mathcal{M} > 1$, the zeros of $\mathcal{Q}_{\mathcal{N},\mathcal{M}}$ are closer and closer to 0 as $\mathcal{N} \to +\infty$. In order to prove this we first establish a bound of the form

$$\frac{q_{2m}}{q_{2(m+1)}} \leq s^2 \qquad \forall m = 0, \dots, \mathcal{M} - 1, \tag{4.15}$$

with s positive real number depending only on $\mathcal{M}$ and $\mathcal{N}$. It is easy to see that

$$s = \left(\frac{\mathcal{M}^3}{\mathcal{N}^2 - \mathcal{M}^2}\right)^{1/2} \tag{4.16}$$

satisfy (4.15).

PROPOSITION 4.7. *Let $z \in \mathbb{C}$ be a root of $\mathcal{Q}_{\mathcal{N},\mathcal{M}}$ defined by Proposition* 4.5 *with $\mathcal{N} > \mathcal{M}$; then*

$$|z|^2 \leq \frac{\mathcal{M}^3}{\mathcal{N}^2 - \mathcal{M}^2}. \tag{4.17}$$

Proof. For s given by (4.16), we define the scalars $s_{2m} = s^{2m}q_{2m}$ and the polynomial $\mathcal{Q}_s$:

$$\mathcal{Q}_s(x) = (1 - x^2)x^{2\mathcal{M}}\mathcal{Q}(s/x) = s_{2\mathcal{M}} - s_0 x^{2(\mathcal{M}+1)} + \sum_{m=0}^{\mathcal{M}-1}(s_{2m} - s_{2(m+1)})x^{2(\mathcal{M}-m)}.$$

Using the fact that (4.16) implies $0 < s_0 \leq s_2 \leq \cdots \leq s_{2\mathcal{M}}$, we lower bound $\mathcal{Q}_s(z)$ for all $z \in \mathbb{C}$,

$$|\mathcal{Q}_s(z)| \geq |s_{2\mathcal{M}}| - \left|s_0 z^{2(\mathcal{M}+1)}\right| - \sum_{m=0}^{\mathcal{M}-1}\left|(s_{2m} - s_{2(m+1)})z^{2(\mathcal{M}-m)}\right|.$$

For a z such that $|z| < 1$, we get

$$|\mathcal{Q}_s(z)| > |s_{2\mathcal{M}}| - |s_0| - \sum_{m=0}^{\mathcal{M}-1}\left|s_{2m} - s_{2(m+1)}\right| = 0.$$

Hence all the roots of $\mathcal{Q}_s$ have modulus larger than 1. Noticing that $\mathcal{Q}(z) = 0 \Longrightarrow \mathcal{Q}_s(s/z) = 0$, we get (4.17). □

Proposition 4.7 gives a rate of convergence of the poles of the rational approximation $\mathcal{R}_{\mathcal{N},\mathcal{M}}(\mathcal{S})$ toward the singularity of $\mathcal{S}$:

$$\forall z \in \mathbb{C}, \quad \mathcal{Q}_{\mathcal{N},\mathcal{M}}(z) = 0 \Longrightarrow \forall \varepsilon > 0, \lim_{\mathcal{N}\to+\infty} |\mathcal{N}^{1-\varepsilon}z| = 0.$$

From now on we consider only the second method to compute the denominator, namely, the one given by Proposition 4.5. Note that the expansion of the denominator in terms of x^k instead of the orthogonal polynomials T_k was already suggested in [5] for other purposes, not related to the locations of the poles of the rational approximation.

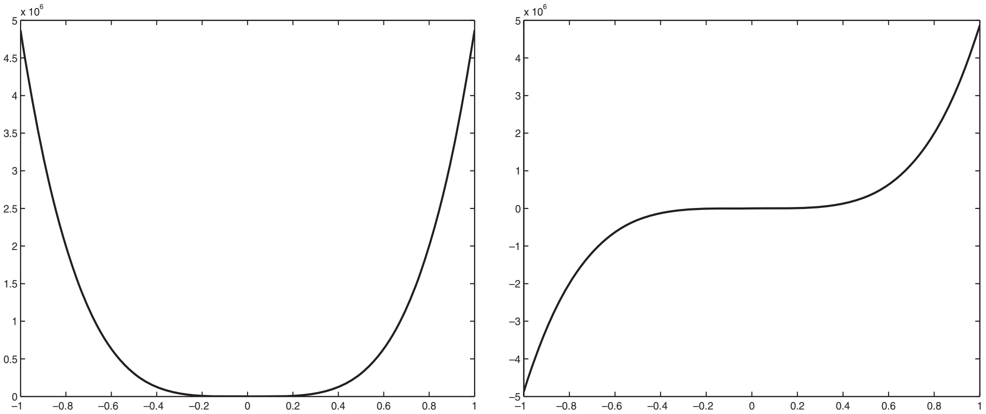


FIG. 4.1. *Typical shapes of $\mathcal{Q}$ and $\mathcal{P}$, $\mathcal{N} = 50$, $\mathcal{M} = 2$.*

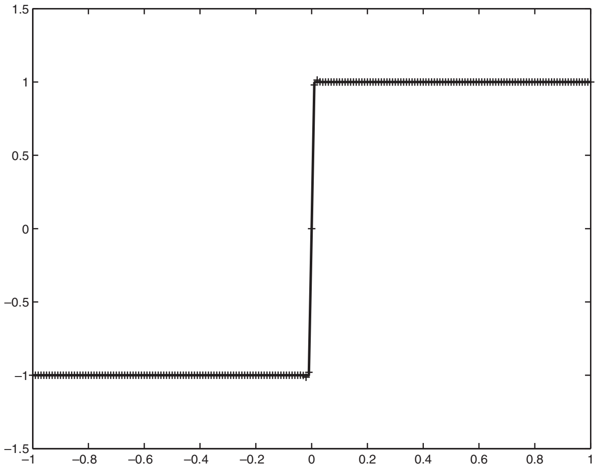


FIG. 4.2. *Rational approximation* (+) *of the sign function (solid line),* $\mathcal{N} = 50$, $\mathcal{M} = 2$.

4.2. Determination of the numerator. With the $q_{2\mathcal{M}}$ defined by Proposition 4.5 and the $\Lambda_{2k+1}^{(\mathcal{M})}$ given by (4.12)–(4.13), the expansion (4.3) leads to defining the numerator $\mathcal{P}$ by

$$\mathcal{P} = \sum_{n=0}^{\mathcal{N}} \hat{p}_{2n+1} T_{2n+1}, \quad \hat{p}_{2n+1} = \Lambda_{2n+1}^{(\mathcal{M})} \quad \forall n = 0, \dots, \mathcal{N}. \tag{4.18}$$

The typical shape of $\mathcal{P}$ is displayed on Figure 4.1.

4.3. Analysis of the rational approximation. We consider here the rational approximation defined by the second method. The approximation $\mathcal{R} = \mathcal{P}/\mathcal{Q}$ ($\mathcal{N} = 50$, $\mathcal{M} = 2$) plotted in Figure 4.2 is very accurate: the Gibbs phenomenon has almost been eliminated. The reader should consult [15] for a comparison of several methods aimed at eliminating this Gibbs phenomenon, in terms of reducing the Gibbs constant (the overshoot near the discontinuity) and increasing the steepness (the value of the derivative of the approximation at the discontinuity). In Figure 4.3 we displayed

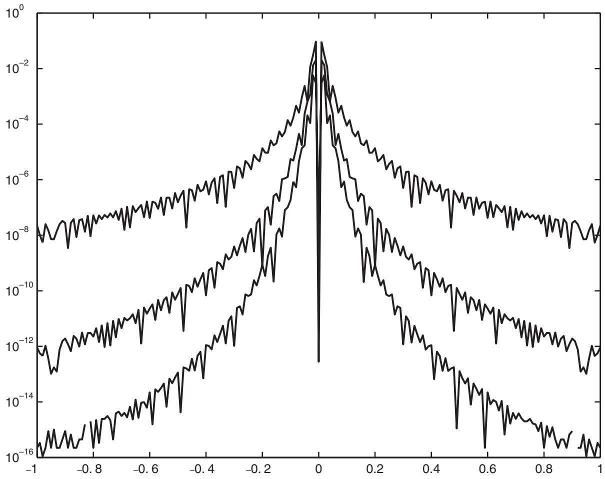


FIG. 4.3. *Rational approximation: log (decimal) of the error* $|\mathcal{R}(S)(x) - Sign(x)|$*,* $\mathcal{N} = 50$*,* $\mathcal{M} = 1, 2, 3$*.*

the pointwise error, $|\mathcal{R}(S)(x) - S(x)|$ for $\mathcal{M} = 1$, 2, and 3. This representation shows the efficiency of the method used as a filter for the spectral approximations of discontinuous solutions.

In this section, we analyze these spectacular results. Let the parameter $\mathcal{M}$ be fixed and the parameter $\mathcal{N} = K - \mathcal{M}$ go to infinity. We know from Proposition 4.5 that $\sigma_k^{(\mathcal{M},\mathcal{N})} = 0$ for $k = \mathcal{N}+1, \dots, K = \mathcal{N}+\mathcal{M}$. The next lemma makes precise the behavior of $\sigma_k^{(\mathcal{M},\mathcal{N})}$ for large values of k.

LEMMA 4.8. *The factors* $\sigma_k^{(\mathcal{M})}$ *introduced in* (4.12) *with* q_{2m} *defined by Proposition* 4.5 *satisfy, for fixed* $\mathcal{M}$*,*

$$\sigma_k^{(\mathcal{M})} \simeq \left(1 - \frac{\mathcal{N}^2}{k^2}\right)^{\mathcal{M}} \qquad \text{as } k \to +\infty. \tag{4.19}$$

Proof. We fix $\mathcal{M}$, and let the q_{2m}'s be as defined by Proposition 4.5. Using (1.4) in (4.14), we get

$$\begin{aligned}\sigma_k^{(\mathcal{M})} &= \frac{\Gamma(k+\mathcal{N}+\mathcal{M}+2)\Gamma(k-\mathcal{N})\Gamma(3/2+k)\Gamma(k+1/2-\mathcal{M})}{\Gamma(k+\mathcal{N}+2)\Gamma(k-\mathcal{N}-\mathcal{M})\Gamma(3/2+k+\mathcal{M})\Gamma(k+1/2)}\\ &= \gamma_1\gamma_2\gamma_3\gamma_4,\end{aligned}$$

with

$$\gamma_1 = \frac{\Gamma(k+\mathcal{N}+2+\mathcal{M})}{\Gamma(k+\mathcal{N}+2)} \simeq (k+\mathcal{N})^{\mathcal{M}} \qquad \text{for large values of } k,$$

$$\gamma_2 = \frac{\Gamma(k-\mathcal{N})}{\Gamma(k-\mathcal{N}-\mathcal{M})} \simeq (k-\mathcal{N})^{\mathcal{M}},$$

$$\gamma_3 = \frac{\Gamma(k+1/2-\mathcal{M})}{\Gamma(k+3/2+\mathcal{M})} = \frac{1}{(k+1/2-\mathcal{M})\cdots(k+1/2+\mathcal{M})} \simeq \frac{1}{k^{2\mathcal{M}+1}},$$

$$\gamma_4 = \frac{\Gamma(3/2+k)}{\Gamma(k+1/2)} = k+1/2.$$

For $k \to +\infty$, we get (4.19). □

THEOREM 4.9 (convergence). *The rational approximation $\mathcal{R}(S)$ with denominator's coefficients given by Proposition 4.5 and fixed $\mathcal{M}$ converges to S as $\mathcal{N} \to +\infty$:*

$$\lim_{\mathcal{N}\to+\infty} \mathcal{R}(S) = S \qquad \text{in } L^2_\omega. \tag{4.20}$$

Proof. The denominator $\mathcal{Q}$ never vanishes on I and is always ≥ 1; it follows that

$$\begin{aligned}
\|\mathcal{R}(S) - S\|^2_\omega &= \int_I \frac{(\mathcal{Q}S - \mathcal{P})^2}{\mathcal{Q}^2}\omega(x)dx \\
&\leq \int_I (\mathcal{Q}S - \mathcal{P})^2\omega(x)dx \qquad (\mathcal{Q}(x) \geq 1) \\
&\leq \frac{2}{\pi}\sum_{k>K} \left|(\widehat{\mathcal{Q}S - \mathcal{P}})_{2k+1}\right|^2 \\
&\leq \frac{2}{\pi}\sum_{k>K} |\Lambda^{(\mathcal{M})}_{2k+1}|^2 \qquad \text{(see (4.3))} \\
&\leq \frac{2}{\pi}\sum_{k>K} |\hat{s}_{2k+1}|^2|\sigma^{(\mathcal{M})}_k|^2 \qquad \text{(see (4.12))}.
\end{aligned}$$

The boundedness of $\sigma^{(\mathcal{M})}_k$ and $S \in L^2_\omega$ end the proof. □

The next theorem makes precise the rate of convergence of the rational approximation. If we consider the norm of the error $\mathcal{R}(S) - S$ in a region excluding a small vicinity of the singularity, we get an acceleration of the convergence. For $\varepsilon > 0$, let us define $\eta = 1/\mathcal{N}^\varepsilon$, $I_\eta =]-1, -\eta[\cup]\eta, 1[$ and $\|\mathcal{R}(S) - S\|_{L^2_\omega(I_\eta)} := (\int_{I_\eta} |\frac{\mathcal{P}}{\mathcal{Q}} - S|^2\omega(x)dx)^{1/2}$.

THEOREM 4.10 (acceleration). *There exists a constant $C_\mathcal{M} \in \mathbb{R}$ depending solely on $\mathcal{M}$ such that*

$$\|\mathcal{R}(S) - S\|_{L^2_\omega(I_\eta)} \leq \frac{C_\mathcal{M}}{\mathcal{N}^{(1-\varepsilon)2\mathcal{M}}}\|\pi^{2K+1}_\omega(S) - S\|_{L^2_\omega(I)}. \tag{4.21}$$

Proof. We derive from the monotonicity of $\mathcal{Q}$ and the asymptotic formula $q_{2\mathcal{M}} \simeq \text{Const}_\mathcal{M}\, \mathcal{N}^{2\mathcal{M}}$ a lower bound of $\mathcal{Q}(x)$:

$$|x| > \eta \Longrightarrow \mathcal{Q}(x) > \mathcal{Q}(\eta) > q_{2\mathcal{M}}\eta^{2\mathcal{M}} > \text{Const}_\mathcal{M}\, (\eta\mathcal{N})^{2\mathcal{M}}.$$

Using this bound, we write

$$\begin{aligned}
\|\mathcal{R}(S) - S\|^2_{L^2_\omega(I_\eta)} &= \int_{I_\eta} \frac{(\mathcal{Q}S - \mathcal{P})^2}{\mathcal{Q}^2}\omega(x)dx \\
&\leq \frac{\text{Const}}{(\eta\mathcal{N})^{4\mathcal{M}}}\int_{I_\eta} (\mathcal{Q}S - \mathcal{P})^2\omega(x)dx \\
&\leq \frac{\text{Const}}{(\eta\mathcal{N})^{4\mathcal{M}}}\int_{I} (\mathcal{Q}S - \mathcal{P})^2\omega(x)dx \\
&\leq \frac{\text{Const}'}{(\eta\mathcal{N})^{4\mathcal{M}}}\sum_{k>\mathcal{K}} |\hat{s}_{2k+1}\sigma^{(\mathcal{M})}_k|^2.
\end{aligned}$$

The boundedness of $\sigma^{(\mathcal{M})}_k$ stated in Lemma 4.8 ends the proof. □

For all $x \in I$, we have

$$\mathcal{R}(S)(x) = \sum_{n=0}^{\mathcal{N}} \frac{\hat{p}_{2n+1}}{\mathcal{Q}(x)} T_{2n+1}(x) = \sum_{n=0}^{\mathcal{N}} \frac{\hat{s}_{2n+1}\sigma_n^{(\mathcal{M})}}{\mathcal{Q}(x)} T_{2n+1}(x).$$

Hence the rational approximation is like a Chebyshev series with variable coefficients

$$\hat{r}_{2n+1}(x) = \frac{\hat{s}_{2n+1}\sigma_n^{(\mathcal{M})}}{\mathcal{Q}(x)}$$

decreasing faster than the coefficients of the original function. Taking $\eta = \frac{1}{\mathcal{N}^{\varepsilon}}$ $(\varepsilon > 0)$, we get

$$|x| > \frac{1}{\mathcal{N}^{\varepsilon}} \Longrightarrow |\hat{r}_{2n+1}(x)| < \frac{\text{Const}}{\mathcal{N}^{2\mathcal{M}-2\varepsilon\mathcal{M}}} |\hat{s}_{2n+1}|.$$

This indicates a local acceleration factor of up to $1/\mathcal{N}^{2\mathcal{M}}$.

5. Conclusions. The analysis of some rational approximations of the sign function have been performed. These approximations are built from a finite Chebyshev expansion of the solution. The rate of convergence of these approximations increases with the degree of the denominators. Numerical tests support the theoretical results. Several generalizations of this work should be done.

- Analysis of the Padé–Chebyshev approximations for general discontinuous functions
- Analysis of the Padé–Legendre approximations of the sign function. As far as we know, this has never been done. Quadratic denominators ($\mathcal{M} = 1$) are considered in [6], but the remarkable numerical results of [5] (obtained with $\mathcal{M} \geq 1$) have not yet been fully analyzed.

These two problems are currently under investigation.

Acknowledgment. The authors would like to thank the anonymous referees for valuable suggestions about the bibliography.

REFERENCES

[1] Ch. Bernardi and Y. Maday, *Spectral methods,* in Handbook of Numerical Analysis, V, North-Holland, Amsterdam, 1997.

[2] C. Brezinski, *Padé-type approximation and general orthogonal polynomials*, Internat. Ser. Numer. Math. 50, Birkhäuser-Verlag, Basel, 1980.

[3] C. W. Clenshaw and K. Lord, *Rational approximations from Chebyshev series*, in Studies in Numerical Analysis, B. K. P. Scaife, ed., Academic Press, London, 1974, pp. 95–113.

[4] T. Driscoll and B. Fornberg, *A Padé-based algorithm for overcoming the Gibbs phenomenon*, Numer. Algorithms, 26 (2001), pp. 77–92.

[5] L. Emmel, *Méthode Spectrale Multidomaine de Viscosité Évanescente pour des Problèmes Hyperboliques non Linéaires*, Ph.D. Thesis, Université Pierre et Marie Curie, Paris, 1998.

[6] L. Emmel, S. M. Kaber and Y. Maday, *Padé-Jacobi filtering for spectral approximations of discontinuous solutions*, Numer. Algorithms, 33 (2003), pp. 251–264.

[7] A. Erdélyi, *Higher Transcendental Functions*, Vol. 1, McGraw-Hill, New York, 1953.

[8] J. Fleicher, *Analytic continuation of scattering amplitudes and Padé approximants*, Nucl. Phys. B, 37 (1972), pp. 59–76; erratum: Nucl. Phys. B, 44 (1972), p. 641.

[9] J. Fleicher, *Nonlinear Padé approximants for Legendre series*, J. Math. Phys., 14 (1973), pp. 246–248.

[10] J. F. Geer, *Rational trigonometric approximations using Fourier series partial sums*, J. Sci. Comput., 10 (1995), pp. 325–356.

[11] D. Gottlieb and J. S. Hesthaven, *Spectral methods for hyperbolic problems*, J. Comput. Appl. Math., 128 (2001), pp. 83–131.

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

[13] W. B. Gragg, *Laurent, Fourier, and Chebyshev-Padé tables*, in Padé and Rational Approximation. Theory and Application, E. B. Saff and R. S. Varga, eds., Academic Press, New York, 1977, pp. 61–72.

[14] A. C. Matos, *Recursive computation of Padé-Legendre approximants and some acceleration properties*, Numer. Math., 89 (2001), pp. 535–560.

[15] G. Németh and G. Páris, *The Gibbs phenomenon in generalized Padé approximation*, J. Math. Phys., 26 (1985), pp. 1175–1178.

[16] D. J. Newman, *Rational approximation to* $|x|$, Michigan Math. J., 11 (1964), pp. 11–14.

[17] S. Paszkowski, *Approximation uniforme des fonctions continues par des fonctions rationnelles*, Zastos. Mat., 6 (1963), pp. 441–458.

[18] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions*, Encyclopedia Math. Appl. 28, Cambridge University Press, Cambridge, UK, 1987.

[19] A. Sidi, *Uniqueness of Padé approximants from series of orthogonal polynomials*, Math. Comp., 31 (1977), pp. 738–739.

[20] R. D. Small and R. J. Charon, *Continuous and discrete nonlinear approximations based on Fourier Series*, IMA J. Numer. Anal., 8 (1988), pp. 281–293.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 455–473

FOURTH-ORDER NONOSCILLATORY UPWIND AND CENTRAL SCHEMES FOR HYPERBOLIC CONSERVATION LAWS*

ÁNGEL BALAGUER† AND CARLOS CONDE‡

Abstract. The aim of this work is to solve hyperbolic conservation laws by means of a finite volume method for both spatial and time discretization. We extend the ideas developed in [X.-D. Liu and S. Osher, *SIAM J. Numer. Anal.,* 33 (1996), pp. 760–779; X.-D. Liu and E. Tadmor, *Numer. Math.,* 79 (1998), pp. 397–425] to fourth-order upwind and central schemes. In order to do this, once we know the cell-averages of the solution, $\overline{u}_j^n$, in cells I_j at time $T = t^n$, we define a new three-degree reconstruction polynomial that in each cell, I_j, presents the same shape as the cell-averages $\{\overline{u}_{j-1}^n, \overline{u}_j^n, \overline{u}_{j+1}^n\}$. By combining this reconstruction with the nonoscillatory property and the maximum principle requirement described in [X.-D. Liu and S. Osher, *SIAM J. Numer. Anal.,* 33 (1996), pp. 760–779] we obtain a fourth-order scheme that satisfies the total variation bounded (TVB) property. Extension to systems is carried out by componentwise application of the scalar framework. Numerical experiments confirm the order of the schemes presented in this paper and their nonoscillatory behavior in different test problems.

Key words. central schemes, upwind schemes, high order, nonoscillatory, hyperbolic conservation laws

AMS subject classification. 65M06

DOI. 10.1137/S0036142903437106

1. Introduction.

In this paper we present three fourth-order numerical schemes in order to solve one-dimensional hyperbolic conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad u(x,0) = u_0(x), \tag{1.1}$$

where $u_0(x)$ is a known bounded function.

Many of the high-order methods used to solve this problem employ an interpolating polynomial that reconstructs the pointvalues of the solution in terms of the cell-averages. There are two main types of schemes: upwind schemes and central schemes. Godunov-type schemes [7] are the forerunner for upwind schemes, which compute the cell-averages of the solution in the same spatial cells at all time steps. Similarly, Van Leer [25] presented a scheme with second-order accuracy in space and time. Later, Colella and Woodward [5] used two-degree polynomials, although their scheme satisfies the total variation diminishing (TVD) property, and, hence, it is limited to second order of accuracy in the L^1 norm. Harten et al. [8] introduced the essentially nonoscillatory (ENO) schemes, with an order of accuracy higher than two and able to capture sharp shocks without introducing oscillations. Similarly, different high-order numerical schemes have been developed, such as the weighted ENO (WENO) schemes (see Liu, Osher, and Chan [18], Jiang and Shu [10], or Balsara and Shu [2]). Extensions to multidimensional systems can also be found in Casper

*Received by the editors October 3, 2003; accepted for publication (in revised form) September 7, 2004; published electronically June 30, 2005.

http://www.siam.org/journals/sinum/43-2/43710.html

†Universidad Politécnica de Valencia, E.T.S.I. Geodésica, Cartográfica y Topográfica, Camino de Vera s/n, 46022 Valencia, Spain (abalague@mat.upv.es). The research of this author was supported by the Ministerio de Ciencia y Tecnología of Spain grant BTE2002-04552 and grant REN2003-04998.

‡Universidad Politécnica de Madrid, E.T.S.I. Minas, C/. Ríos Rosas, 21, 28003 Madrid, Spain (cconde@dmami.upm.es).

and Atkins [4] or Balaguer et al. [1]. The schemes found in the latter references use the high-order Runge–Kutta schemes developed in Shu and Osher [23] for time integration, which maintain the spatial operator stability properties.

Although the first-order Lax–Friedrich scheme (see [6]) is probably the forerunner for central schemes, the central scheme of Nessyahu and Tadmor [21] has generated a significant number of works on high-resolution schemes that maintain the simplicity of the Riemann solver-free approach. The scheme developed in Nessyahu and Tadmor [21] has been extended to accuracy orders higher than 2 (see Liu and Tadmor [20], Jiang et al. [11], or Qiu and Shu [22]) and to several spatial dimensions (see Levy and Tadmor [15] and Jiang and Tadmor [12]). High-order central WENO schemes are described in Levy, Puppo, and Russo [16], [17].

We have focused our attention on the upwind scheme developed in Liu and Osher [19] and the central scheme described in Liu and Tadmor [20], which are third-order schemes, in the sense of local truncation error in regions without discontinuities. The algorithm developed in Liu and Osher [19] leads to a conservative scheme that satisfies the local maximum principle and guarantees that the number of extrema in the solution does not exceed the number of extrema of the initial condition $u_0(x)$. These properties allow achieving the total variation bounded (TVB) property. The approach used in that reference uses a simple centered stencil with quadratic reconstruction.

Liu and Tadmor [20] apply the procedure described in Liu and Osher [19] to central schemes and show the results obtained when solving differential equation systems. The resulting scheme is third-order accurate in space and time. In both references ([19] and [20]), time integration is performed using a finite volume method, approximating the resulting integrals with respect to time by a Gauss [19] or a Simpson [20] quadrature rule. The values of the solution at half time steps are approximated using a Taylor expansion. Jiang et al. [11] present a procedure to convert schemes which are based on staggered spatial grids into nonstaggered schemes, which are simpler to implement in frameworks which involve complex geometries and boundary conditions. However, it has been in some cases superseded by the semidiscrete central schemes (see Kurganov and Tadmor [13]) and their high-order extensions.

This paper presents an extension of the schemes developed in Liu and Osher [19] and Liu and Tadmor [20]. In contrast to them, our scheme is a fourth-order scheme in the sense of local truncation error. To this end, we will use a finite volume method with a conservative degree-three polynomial reconstruction that calculates the pointvalues of the solution from the cell-averages, by avoiding the increase in the number of solution extrema at the interior of each cell. This condition, together with the nonoscillatory property and the maximum principle requirement described in Liu and Osher [19], avoids spurious numerical oscillations in the computed solution.

The integrals respecting the two variables, space and time, are evaluated by means of a two-point Gauss quadrature. The values of the solution at half-time steps are calculated using a Taylor expansion with a fourth-order error, using the local Cauchy–Kowalewski procedure (see [8]) to approximate the time derivatives of the solution as a function of the derivatives with respect to x. We also present an extension to systems of equations, where the computed solution at quadrature nodes is obtained by the so called natural continuous extension of Runge–Kutta schemes (see Zennaro [26], Bianco, Puppo, and Russo [3], or Levy, Puppo, and Russo [16]).

In this paper, first, we present the equations that define the upwind and central schemes for solving the problem (1.1). Next, the fourth-order nonoscillatory reconstruction procedure is described. It guarantees that the resulting numerical scheme satisfies the properties that generate its nonoscillatory behavior. Finally, some problems with known analytical solution are solved to verify the order of the schemes presented here and to compare their behavior with the schemes developed in Liu and Osher [19] and Liu and Tadmor [20].

2. Upwind and central schemes. Let us suppose that the time interval is discretized uniformly into the values $t^n = n \cdot \Delta t,\ n = 0, 1, 2, \ldots, NT$. We assume that the grid points $\{x_j\}$ are distributed uniformly at the spatial domain at which (1.1) will be defined, verifying that $x_j = x_{j-1} + \Delta x\, \forall j = 1, \ldots, NX$, where Δx is a known constant. Given a point (x_j, t^n), we consider the control volume defined by $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \times \left[t^n, t^{n+1}\right]$, where $x_{j\pm\frac{1}{2}} = x_j \pm \Delta x/2$. By integrating (1.1) over this control volume, we obtain

$$\text{(2.1)}\ \overline{u}_j^{n+1} = \overline{u}_j^n - \frac{1}{\Delta x}\left[\int_{t^n}^{t^{n+1}} f\left(u\left(x_{j+\frac{1}{2}}, \tau\right)\right) d\tau - \int_{t^n}^{t^{n+1}} f\left(u\left(x_{j-\frac{1}{2}}, \tau\right)\right) d\tau\right],$$

where the cell average $\overline{u}_j^n$ is defined as

$$\text{(2.2)}\qquad \overline{u}_j^n = \frac{1}{\Delta x}\int_{I_j} u\left(\varphi, t^n\right) d\varphi, \quad I_j = \left\{\varphi, |\varphi - x| \le \frac{\Delta x}{2}\right\}.$$

In (2.1) there is a relationship between the average values of the solution at the limit of the time interval, $\overline{u}_j^n$, $\overline{u}_j^{n+1}$, and its pointvalues at the boundary of the space interval, $u(x_{j\pm\frac{1}{2}}, \tau)$. The steps to follow in the implementation of numerical schemes can be described as follows.

(1) For each time value $t^n,\ n \in \{0, 1, \ldots, NT - 1\}$, we have an approximation of the cell-averages of the solution $\overline{w}_j^n \cong \overline{u}_j^n\, \forall j \in \{0, 1, \ldots, NX\}$, at the nodes x_j. The approximation will be of order $O((\Delta x)^4)$.

(2) The pointvalues of $w(x, t^n)\, \forall x \in \{x_0 - \Delta x/2, \ldots, x_{NX} + \Delta x/2\}$ are reconstructed using a piecewise polynomial interpolation,

$$\text{(2.3)}\qquad w(x, t^n) \equiv \sum_{j=0}^{NX} R_j(x; \overline{w}^n)\chi_j(x), \quad \chi_j(x) = \begin{cases} 1 & \text{if} \quad x \in I_j, \\ 0 & \text{if} \quad x \notin I_j, \end{cases}$$

where $R_j(x; \overline{w}^n)$ is a polynomial that reconstructs the pointvalues of the solution using the discrete values $\overline{w}_i^n$, $i \in \{0, 1, \ldots, NX\}$, verifying

$$\text{(2.4)}\quad \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} R_j(x; \overline{w}^n) = \overline{w}_j^n, \quad R_j(x; \overline{w}^n) = w(x, t^n) + O((\Delta x)^4)\ \ \forall x \in I_j.$$

(3) In the case of the central schemes, the average values $\overline{w}_{j+\frac{1}{2}}^n$ are calculated using the approximation given in (2.3):

$$\text{(2.5)}\qquad \overline{w}_{j+\frac{1}{2}}^n \equiv \frac{1}{\Delta x}\left[\int_{x_j}^{x_j + \Delta x/2} R_j(x; \overline{w}^n)\, dx + \int_{x_j + \Delta x/2}^{x_{j+1}} R_{j+1}(x; \overline{w}^n)\, dx\right].$$

The integrals in (2.1) are evaluated in an exact way taking into account that $R_j(x; \overline{w}^n)$ and $R_{j+1}(x; \overline{w}^n)$ are degree-three polynomials.

(4) The integrals with respect to the time variable are approximated using a two-point Gauss quadrature. Thus,

$$\int_{t^n}^{t^{n+1}} f(w(x_{j\pm\frac{1}{2}},\tau))d\tau \approx \frac{\Delta t}{2}\Big(f(w(x_{j\pm\frac{1}{2}},t^n+\beta_0))+f(w(x_{j\pm\frac{1}{2}},t^n+\beta_1))\Big) \tag{2.6}$$

$$\beta_0=\Delta t\left(\frac{1-1/\sqrt{3}}{2}\right),\quad \beta_1=\Delta t\left(\frac{1+1/\sqrt{3}}{2}\right). \tag{2.7}$$

In order to approximate the pointvalues of w at the time steps that appear in (2.6), we may use a Taylor expansion with an error $O((\Delta x^4))$. This technique is used, for example, in Liu and Osher [19] and Liu and Tadmor [20]. Another efficient method would be the natural continuous extension of Runge–Kutta methods advocated by Bianco, Puppo, and Russo [3] and Levy, Puppo, and Russo [16]. We will use this method in the resolution of systems of equations which achieve the same accuracy with much lower computational effort.

(5) In order to calculate the cell-averages of w at t^{n+1}, we distinguish two cases.

(5a) *Upwind schemes.* The cells are intervals centered at each x_j (equation (2.1), after replacing the function u—in that equation—for the function w). In order to calculate the value of $f(w(x_{j\pm\frac{1}{2}},t^n+\beta_k))$ in expression (2.6), we will use the Roe flux with entropy fix, although other fluxes can also be used, as those described in Liu and Osher [19].

(5b) *Central schemes.* The cells are intervals centered at each $x_{j+\frac{1}{2}}$.

$$\overline{w}^{n+1}_{j+\frac{1}{2}}=\overline{w}^{n}_{j+\frac{1}{2}}-\frac{1}{\Delta x}\left[\int_{t^n}^{t^{n+1}} f\left(w\left(x_{j+1},\tau\right)\right)d\tau-\int_{t^n}^{t^{n+1}} f\left(w\left(x_{j},\tau\right)\right)d\tau\right]. \tag{2.8}$$

The terms on the right-hand side in (2.1) and (2.8) are calculated using the approximations (2.5)–(2.7).

(6) We go back to step (1) and restart the procedure until calculating $\overline{w}_i^{NT}\cong\overline{u}(x_i,t^{NT})$, $i\in\{0,1,\dots,NX\}$. Then, we use formula (2.3) to obtain the pointvalues with $O((\Delta x)^4)$.

3. Fourth-order nonoscillatory reconstruction.

This section presents the reconstruction procedure used to obtain each $R_j(x;\overline{w}^n)$ from the cell averages $\overline{w}^n_k$, $k\in\{j-2,j-1,j,j+1,j+2\}$.

3.1. Fourth order and conservation.

Initially, we will consider the degree-three polynomial that verifies these conditions:

$$p_j(x_j;\overline{w}^n)=\overline{w}^n_j,\quad p_j(x_{j-1};\overline{w}^n)=\overline{w}^n_{j-1},\quad p_j(x_{j+1};\overline{w}^n)=\overline{w}^n_{j+1}, \tag{3.1}$$

$$\Delta x\frac{dp_j}{dx}(x_j;\overline{w}^n)=\Delta x\frac{\partial\overline{w}}{\partial x}(x_j,t^n)\equiv d^n_j,\quad\text{where }\overline{w}(x,t^n)=\frac{1}{\Delta x}\int_{x-\Delta x/2}^{x+\Delta x/2} w(\varphi,t^n)\,d\varphi.$$

This polynomial can be expressed as

$$\begin{aligned}p_j\left(x;\overline{w}^n\right)=\overline{w}^n_j+d^n_j\cdot\left(\frac{x-x_j}{\Delta x}\right)+\left(\frac{\overline{w}^n_{j-1}-2\overline{w}^n_j+\overline{w}^n_{j+1}}{2}\right)\cdot\left(\frac{x-x_j}{\Delta x}\right)^2\\ +\left(\frac{-\overline{w}^n_{j-1}+\overline{w}^n_{j+1}-2d^n_j}{2}\right)\cdot\left(\frac{x-x_j}{\Delta x}\right)^3.\end{aligned} \tag{3.2}$$

Since

$$w(x,t^n)=\overline{w}(x,t^n)-\frac{1}{24}\left(\Delta x\right)^2\frac{\partial^2\overline{w}(x,t^n)}{\partial x^2}+O\left(\Delta x\right)^4, \tag{3.3}$$

the conservative polynomial, $q_j(x;\overline{w}^n)$ that verifies the conditions in (2.4) can be defined as

$$q_j(x;\overline{w}^n)=p_j(x;\overline{w}^n)-\frac{1}{24}\left(\Delta x\right)^2\frac{d^2p_j(x;\overline{w}^n)}{dx^2}. \tag{3.4}$$

Therefore,

$$\begin{aligned} q_j(x;\overline{w}^n)&=\overline{w}^n_j-\frac{1}{24}\left(\overline{w}^n_{j-1}-2\overline{w}^n_j+\overline{w}^n_{j+1}\right)+\left(\frac{\overline{w}^n_{j-1}-\overline{w}^n_{j+1}+10d^n_j}{8}\right)\left(\frac{x-x_j}{\Delta x}\right)\\ &+\left(\frac{\overline{w}^n_{j-1}-2\overline{w}^n_j+\overline{w}^n_{j+1}}{2}\right)\left(\frac{x-x_j}{\Delta x}\right)^2+\left(\frac{-\overline{w}^n_{j-1}+\overline{w}^n_{j+1}-2d^n_j}{2}\right)\left(\frac{x-x_j}{\Delta x}\right)^3. \end{aligned} \tag{3.5}$$

In case that

$$d^n_j=ds^n_j\equiv\frac{2}{3}\overline{w}^n_{j+1}-\frac{2}{3}\overline{w}^n_{j-1}-\frac{1}{12}\overline{w}^n_{j+2}+\frac{1}{12}\overline{w}^n_{j-2}, \tag{3.6}$$

$q_j(x;\overline{w}^n)$ coincides with the centered polynomial, defined as the average value between two conservative piecewise polynomials: the conservative polynomial which uses $\left\{\overline{w}^n_{j-1},\overline{w}^n_j,\overline{w}^n_{j+1},\overline{w}^n_{j+2}\right\}$ and the polynomial based on $\left\{\overline{w}^n_{j-2},\overline{w}^n_{j-1},\overline{w}^n_j,\overline{w}^n_{j+1}\right\}$. Then, by replacing the value of d^n_j given in (3.6) in expression (3.5), we obtain the following conservative polynomial that verifies conditions (2.4):

$$q^*_j(x;\overline{w}^n)=C^n_{o,j}+C^n_{1,j}\left(\frac{x-x_j}{\Delta x}\right)+C^n_{2,j}\left(\frac{x-x_j}{\Delta x}\right)^2+C^n_{3,j}\left(\frac{x-x_j}{\Delta x}\right)^3, \tag{3.7}$$

$$C^n_{o,j}=\overline{w}^n_j-\frac{1}{24}\left(\overline{w}^n_{j+1}-2\overline{w}^n_j+\overline{w}^n_{j-1}\right), C^n_{1,j}=\frac{-5\overline{w}^n_{j+2}+34\overline{w}^n_{j+1}-34\overline{w}^n_{j-1}+5\overline{w}^n_{j-2}}{48},$$

$$C^n_{2,j}=\frac{1}{2}\left(\overline{w}^n_{j+1}-2\overline{w}^n_j+\overline{w}^n_{j-1}\right),\ C^n_{3,j}=\frac{1}{12}\left(\overline{w}^n_{j+2}-2\overline{w}^n_{j+1}+2\overline{w}^n_{j-1}-\overline{w}^n_{j-2}\right).$$

3.2. Shape-preserving when the cell-averages form a monotone sequence. We will define d^n_j in (3.5) so that if the cell-averages $\{\overline{w}^n_{j-1},\overline{w}^n_j,\overline{w}^n_{j+1}\}$ form a monotone sequence, then $q_j(x;\overline{w}^n)$ is monotone on I_j. We will denote as shape-preserving properties the following:

(I) $q_j(x;\overline{w}^n)$ is monotonically increasing in I_j if $\overline{w}^n_{j-1}\le\overline{w}^n_j\le\overline{w}^n_{j+1}$.

(II) $q_j(x;\overline{w}^n)$ is monotonically decreasing in I_j if $\overline{w}^n_{j-1}\ge\overline{w}^n_j\ge\overline{w}^n_{j+1}$.

To simplify the notation, first we define

$$\overline{W}C^n_j=\overline{w}^n_{j+1}-\overline{w}^n_{j-1},\ \overline{W}R^n_j=\overline{w}^n_{j+1}-\overline{w}^n_j,\ \overline{W}C2^n_j=\overline{w}^n_{j+2}-\overline{w}^n_{j-2}. \tag{3.8}$$

Observation 1. If $d^n_j=\frac{\overline{W}C^n_j}{2}$, then according to (3.5) $q_j(x;\overline{w}^n)$ coincides with a quadratic polynomial. In this case, $\frac{dq_j(x_j\pm\Delta x/2;\overline{w}^n)}{dx}=\pm\frac{\overline{w}^n_{j\pm1}-\overline{w}^n_j}{\Delta x}$, and therefore the shape-preserving properties are verified.

Observation 2. In the case at which $d^n_j=ds^n_j$ (defined in (3.6)), the following hold.

1. If $\left(2\cdot\overline{W}C^n_j=\overline{W}C2^n_j\right)$, then $d^n_j=\frac{\overline{W}C^n_j}{2}$ (see Observation 1).

2. If $\left(2\cdot\overline{W}C_j^n > \overline{W}C2_j^n\right)$ and $\overline{w}_{j-1}^n \le \overline{w}_j^n \le \overline{w}_{j+1}^n$, then, according to (3.7),

$$\frac{d^3 q_j^*(x;\overline{w}^n)}{dx^3} = \frac{1}{(\Delta x)^3}\left(6\cdot C_{3,j}^n\right) = \frac{1}{2\cdot(\Delta x)^3}\left(\overline{W}C2_j^n - 2\cdot\overline{W}C_j^n\right) < 0,$$

and thus $\frac{d(q_j^*(x;\overline{w}^n))}{dx}$ achieves the minimum value at the endpoints of the interval under consideration. Since

$$\begin{aligned}\frac{dq_j^*(x_j+\Delta x/2;\overline{w}^n)}{dx} &= \frac{1}{24\Delta x}\left(-\overline{W}C2_j^n + 2\cdot\overline{W}C_j^n + 24\cdot\overline{W}R_j^n\right) > 0,\\ \frac{dq_j^*(x_j-\Delta x/2;\overline{w}^n)}{dx} &= \frac{1}{24\Delta x}\left(-\overline{W}C2_j^n + 26\cdot\overline{W}C_j^n - 24\cdot\overline{W}R_j^n\right) > 0,\end{aligned} \tag{3.9}$$

$\mathrm{Min}\{\frac{d(q_j^*(x;\overline{w}^n))}{dx}\ \forall x\in I_j\} > 0$ and $q_j^*(x;\overline{w}^n)$ is monotonically increasing in I_j.

3. If $\left(2\cdot\overline{W}C_j^n < \overline{W}C2_j^n\right)$ and $\overline{w}_{j-1}^n \le \overline{w}_j^n \le \overline{w}_{j+1}^n$, then the derivative of $q_j^*(x;\overline{w}^n)$, defined in (3.7), has a minimum at point

$$x_{MI} = x_j + \frac{\Delta x}{3}\left(\frac{2\cdot\overline{W}R_j^n - \overline{W}C_j^n}{(1/6)\left(2\cdot\overline{W}C_j^n - \overline{W}C2_j^n\right)}\right) = x_j - 4\cdot\Delta x\left(\frac{\overline{W}R_j^n - (1/2)\overline{W}C_j^n}{\left(\overline{W}C2_j^n - 2\cdot\overline{W}C_j^n\right)}\right).$$

In this way, if $\left|\overline{W}R_j^n - \frac{1}{2}\overline{W}C_j^n\right| \ge \frac{1}{8}\left|\overline{W}C2_j^n - 2\cdot\overline{W}C_j^n\right|$, then

$$\left(\overline{W}R_j^n > \frac{1}{2}\overline{W}C_j^n \Rightarrow x_{MI} \le x_j - \frac{\Delta x}{2}\right) \text{ and } \left(\overline{W}R_j^n < \frac{1}{2}\overline{W}C_j^n \Rightarrow x_{MI} \ge x_j + \frac{\Delta x}{2}\right).$$

Thus, the minimum $\mathrm{Min}\{\frac{d(q_j^*(x;\overline{w}^n))}{dx}\ \forall x\in I_j\}$ is achieved at one of these boundary points, $x = x_j \pm \frac{\Delta x}{2}$. However, in this case we cannot ensure that the inequalities given in (3.9) are always verified.

4. If $\left(2\cdot\overline{W}C_j^n < \overline{W}C2_j^n\right)$ and $\overline{w}_{j-1}^n \ge \overline{w}_j^n \ge \overline{w}_{j+1}^n$, then we can prove that $q_j^*(x;\overline{w}^n)$ is monotonically decreasing in I_j.

5. If $\left(2\cdot\overline{W}C_j^n > \overline{W}C2_j^n\right)$ and $\overline{w}_{j-1}^n \ge \overline{w}_j^n \ge \overline{w}_{j+1}^n$, then we can prove that $|x_{MI} - x_j| \ge \Delta x/2$ when $\left|\overline{W}R_j^n - \frac{1}{2}\overline{W}C_j^n\right| \ge \frac{1}{8}\left|\overline{W}C2_j^n - 2\cdot\overline{W}C_j^n\right|$, but we cannot ensure that $q_j^*(x;\overline{w}^n)$ is always monotonically decreasing in I_j.

Observation 3. Supposing that $2\cdot d_j^n < \overline{W}C_j^n$, then the derivative of $q_j(x;\overline{w}^n)$, defined in (3.5), has a minimum at point

$$x_{MI} = x_j + \frac{\Delta x}{3}\left(\frac{2\cdot\overline{W}R_j^n - \overline{W}C_j^n}{2\cdot d_j^n - \overline{W}C_j^n}\right).$$

In addition,

$$\frac{dq_j(x_{MI};\overline{w}^n)}{dx} = q_{xj}^n(d_j^n) \equiv \frac{1}{8\Delta x}\left(10\cdot d_j^n - \overline{W}C_j^n\right) + \frac{1}{6\Delta x}\left(\frac{\left(2\cdot\overline{W}R_j^n - \overline{W}C_j^n\right)^2}{\left(2\cdot d_j^n - \overline{W}C_j^n\right)}\right).$$

This is a function that depends on d_j^n and coincides with a hyperbola. In it, the value of $d_j^n = \frac{\overline{W}C_j^n}{2} - S_j^n\frac{\sqrt{15}}{15}\left|2\cdot\overline{W}R_j^n - \overline{W}C_j^n\right|$ is a local maximum of $q_{xj}^n(d_j^n)$ when $S_j^n > 0$ and a local minimum of $q_{xj}^n(d_j^n)$ when $S_j^n < 0$ being $S_j^n = \mathrm{Sign}(\overline{W}C_j^n)$.

3.2.1. Definition of d_j^n. The polynomial $q_j^*(x;\overline{w}^n)$ defined in (3.7) does not fulfill the shape-preserving properties defined in this subsection for any sequence of values $\left\{\overline{w}_{j-2}^n, \overline{w}_{j-1}^n, \overline{w}_j^n, \overline{w}_{j+1}^n, \overline{w}_{j+2}^n\right\}$. Therefore, we have to define a procedure that adequately defines the slopes d_j^n.

We will consider the value of $d_j^n = ds_j^n$ (given in (3.6)) except when this shape-preserving property is not fulfilled. For this, we use the notation

$$ds1_j^n = \frac{\overline{W}C_j^n}{10},\ ds2_j^n = \frac{1}{2}\left(\overline{W}C_j^n - 4\cdot \overline{W}R_j^n\right),\ ds3_j^n = \frac{1}{2}\left(4\cdot \overline{W}R_j^n - 3\cdot \overline{W}C_j^n\right),$$

$$(3.10)\quad S_j^n = \text{Sign}(\overline{W}C_j^n),\ C1 = \frac{\sqrt{15}}{15},\ C2 = \frac{15-\sqrt{15}}{28}$$

and define d_j^n in the following way:

(A1) If $S_j^n = 0$, then $d_j^n = 0$.

(A2) If $S_j^n \neq 0$ and $\left(2\cdot S_j^n \cdot \overline{W}C_j^n \geq S_j^n \cdot \overline{W}C2_j^n\right)$, then $d_j^n = ds_j^n$.

(A3) If $S_j^n \neq 0$ and $\left(2\cdot S_j^n \cdot \overline{W}C_j^n < S_j^n \cdot \overline{W}C2_j^n\right)$, then the following hold:

(A3.1) If $\overline{w}_j^n = \frac{\overline{w}_{j+1}^n + \overline{w}_{j-1}^n}{2}$, we define

$$d_j^n = \begin{cases} \text{Max}\left\{ds1_j^n, ds_j^n\right\} & \text{if } S_j^n > 0, \\ \text{Min}\left\{ds1_j^n, ds_j^n\right\} & \text{if } S_j^n < 0. \end{cases}$$

(A3.2) If $\overline{w}_j^n \neq \frac{\overline{w}_{j+1}^n + \overline{w}_{j-1}^n}{2}$, then the following hold:

(A3.2.1) If $\left|\overline{W}R_j^n - \frac{1}{2}\overline{W}C_j^n\right| \geq \frac{1}{8}\left|\overline{W}C2_j^n - 2\cdot \overline{W}C_j^n\right|$, then

$$d_j^n = \begin{cases} \text{Max}\left\{ds2_j^n,\ ds3_j^n,\ ds_j^n\right\} & \text{if } S_j^n > 0, \\ \text{Min}\left\{ds2_j^n,\ ds3_j^n,\ ds_j^n\right\} & \text{if } S_j^n < 0. \end{cases}$$

(A3.2.2) If $\left|\overline{W}R_j^n - \frac{1}{2}\overline{W}C_j^n\right| < \frac{1}{8}\left|\overline{W}C2_j^n - 2\cdot \overline{W}C_j^n\right|$, then

$$d_j^n = \begin{cases} \frac{\overline{W}C_j^n}{2} - S_j^n \cdot C1 \cdot \left|2\cdot \overline{W}R_j^n - \overline{W}C_j^n\right| & \text{if } \left|\frac{\overline{W}R_j^n}{\overline{W}C_j^n} - \frac{1}{2}\right| \leq C2, \\ \frac{\overline{W}C_j^n}{2} & \text{if } \left|\frac{\overline{W}R_j^n}{\overline{W}C_j^n} - \frac{1}{2}\right| > C2. \end{cases}$$

THEOREM 3.1. *With this definition of d_j^n the polynomial $q_j(x;\overline{w}^n)$ defined by means of* (3.5) *verifies the following shape-preserving properties:*

(I) *$q_j(x;\overline{w}^n)$ is monotonically increasing in I_j if $\overline{w}_{j-1}^n \leq \overline{w}_j^n \leq \overline{w}_{j+1}^n$.*

(II) *$q_j(x;\overline{w}^n)$ is monotonically decreasing in I_j if $\overline{w}_{j-1}^n \geq \overline{w}_j^n \geq \overline{w}_{j+1}^n$.*

Proof. Let us suppose that $\overline{w}_{j-1}^n \leq \overline{w}_j^n \leq \overline{w}_{j+1}^n$. We have to prove that $q_j(x;\overline{w}^n)$ is monotonically increasing in $I_j = \left[x_j - \frac{\Delta x}{2}, x_j + \frac{\Delta x}{2}\right]$. A similar argument allows us to prove that $q_j(x;\overline{w}^n)$ is monotonically decreasing in I_j when $\overline{w}_{j-1}^n \geq \overline{w}_j^n \geq \overline{w}_{j+1}^n$. We will consider all the possible cases.

Case (A1). $S_j^n = 0$, that is, $\overline{w}_{j-1}^n = \overline{w}_j^n = \overline{w}_{j+1}^n$. Then $d_j^n = 0$ and $q_j(x;\overline{w}^n) = \overline{w}_j^n$. This polynomial has the same degree of monotonicity as the cell averages $\left\{\overline{w}_{j-1}^n, \overline{w}_j^n, \overline{w}_{j+1}^n\right\}$.

Case (A2). $S_j^n = 1$ and $\left(2\cdot \overline{W}C_j^n \geq \overline{W}C2_j^n\right)$. In this case, $d_j^n = ds_j^n$ and, after observation 2, $q_j(x;\overline{w}^n)$ is monotonically increasing in I_j.

Case (A3.1). $S_j^n = 1$, $\left(2 \cdot \overline{W}C_j^n < \overline{W}C2_j^n\right)$, and $\overline{w}_j^n = \frac{\overline{w}_{j+1}^n + \overline{w}_{j-1}^n}{2}$. Then $(2 \cdot \overline{W}R_j^n = \overline{W}C_j^n)$ and according to Observation 3,

$$x_{MI} = x_j, \;\; \frac{dq_j(x_{MI}; \overline{w}^n)}{dx} = \frac{1}{8\Delta x}\left(10 \cdot d_j^n - \overline{W}C_j^n\right).$$

Since in this case $d_j^n = \text{Max}\{\frac{\overline{W}C_j^n}{10}, ds_j^n\}$, then $\frac{dq_j(x_{MI};\overline{w}^n)}{dx} \geq 0$ and $q_j(x;\overline{w}^n)$ is monotonically increasing in I_j.

Case (A3.2.1). $S_j^n = 1$, $\left(2 \cdot \overline{W}C_j^n < \overline{W}C2_j^n\right)$, $\overline{w}_j^n \neq \frac{\overline{w}_{j+1}^n + \overline{w}_{j-1}^n}{2}$, and $\left|\overline{W}R_j^n - \frac{1}{2}\overline{W}C_j^n\right| \geq \frac{1}{8}\left|\overline{W}C2_j^n - 2 \cdot \overline{W}C_j^n\right|$. In this case, $d_j^n = \text{Max}\{\frac{1}{2}(\overline{W}C_j^n - 4 \cdot \overline{W}R_j^n), \frac{1}{2}(4 \cdot \overline{W}R_j^n - 3 \cdot \overline{W}C_j^n),\, ds_j^n\}$. Since $\left(2 \cdot \overline{W}C_j^n < \overline{W}C2_j^n\right)$, then $2 \cdot ds_j^n < \overline{W}C_j^n$. In addition,

$$\overline{W}R_j^n = 0 \Rightarrow d_j^n = \frac{\overline{W}C_j^n}{2}, \;\; \overline{W}R_j^n = \overline{W}C_j^n \Rightarrow d_j^n = \frac{\overline{W}C_j^n}{2}.$$

If $d_j^n = \frac{\overline{W}C_j^n}{2}$, it follows that $q_j(x;\overline{w}^n)$ coincides with a quadratic polynomial, which is monotonically increasing in I_j, as we have seen in Observation 1. Therefore, we can suppose that $0 < \overline{W}R_j^n < \overline{W}C_j^n$. Hence

$$\left(\overline{W}C_j^n - 4 \cdot \overline{W}R_j^n\right) < \overline{W}C_j^n, \;\; \overline{W}C_j^n - \left(4 \cdot \overline{W}R_j^n - 3 \cdot \overline{W}C_j^n\right) = 4 \cdot \left(\overline{W}C_j^n - \overline{W}R_j^n\right) > 0$$

so that it follows that $(2 \cdot d_j^n < \overline{W}C_j^n)$. On the other hand, applying Observations 2 and 3,

$$\left.\begin{array}{l} d_j^n \geq ds_j^n \\ \overline{W}R_j^n > \frac{1}{2}\overline{W}C_j^n \end{array}\right\} \Rightarrow x_{MI} = x_j + \tfrac{\Delta x}{3}\left(\tfrac{2 \cdot \overline{W}R_j^n - \overline{W}C_j^n}{2 \cdot d_j^n - \overline{W}C_j^n}\right) \leq x_j + \tfrac{\Delta x}{3}\left(\tfrac{2 \cdot \overline{W}R_j^n - \overline{W}C_j^n}{2 \cdot ds_j^n - \overline{W}C_j^n}\right)$$
$$\leq x_j - \tfrac{\Delta x}{2},$$
$$\left.\begin{array}{l} d_j^n \geq ds_j^n \\ \overline{W}R_j^n < \frac{1}{2}\overline{W}C_j^n \end{array}\right\} \Rightarrow x_{MI} = x_j - \tfrac{\Delta x}{3}\left(\tfrac{\overline{W}C_j^n - 2 \cdot \overline{W}R_j^n}{2 \cdot d_j^n - \overline{W}C_j^n}\right) \geq x_j - \tfrac{\Delta x}{3}\left(\tfrac{\overline{W}C_j^n - 2 \cdot \overline{W}R_j^n}{2 \cdot ds_j^n - \overline{W}C_j^n}\right)$$
$$\geq x_j + \tfrac{\Delta x}{2}.$$

In this way, because of $d_j^n \geq ds_j^n$ we deduce the following:

$$x_{MI} \notin \left]x_j - \frac{\Delta x}{2}, x_j + \frac{\Delta x}{2}\right[.$$

Thus, the minimum, $\text{Min}\{\frac{d(q_j(x;\overline{w}^n))}{dx}\, \forall x \in I_j\}$, is achieved at one of the boundary points $x = x_j \pm \frac{\Delta x}{2}$. By derivating in formula (3.5), we get that

$$\frac{dq_j(x_j + \Delta x/2; \overline{w}^n)}{dx} \geq 0 \Longleftrightarrow d_j^n \geq \frac{1}{2}\left(\overline{W}C_j^n - 4 \cdot \overline{W}R_j^n\right),$$
$$\frac{dq_j(x_j - \Delta x/2; \overline{w}^n)}{dx} \geq 0 \Longleftrightarrow d_j^n \geq \frac{1}{2}\left(4 \cdot \overline{W}R_j^n - 3 \cdot \overline{W}C_j^n\right).$$

The definition of d_j^n allows us to state that $\text{Min}\{\frac{d(q_j(x;\overline{w}^n))}{dx}\, \forall x \in I_j\} \geq 0$, and thus we conclude that $q_j(x;\overline{w}^n)$ is monotonically increasing in I_j.

Case (A3.2.2). $S_j^n = 1$, $\left(2 \cdot \overline{W}C_j^n < \overline{W}C2_j^n\right)$, $\overline{w}_j^n \neq \frac{\overline{w}_{j+1}^n + \overline{w}_{j-1}^n}{2}$, and $\left|\,\overline{W}R_j^n - \frac{1}{2}\overline{W}C_j^n\,\right| < \frac{1}{8}\left|\,\overline{W}C2_j^n - 2 \cdot \overline{W}C_j^n\,\right|$. In this case,

$$d_j^n = \begin{cases} \frac{\overline{W}C_j^n}{2} - \frac{\sqrt{15}}{15}\left|\,2 \cdot \overline{W}R_j^n - \overline{W}C_j^n\,\right| & \text{if } \left|\frac{\overline{W}R_j^n}{\overline{W}C_j^n} - \frac{1}{2}\right| \leq \frac{15-\sqrt{15}}{28}, \\ \frac{\overline{W}C_j^n}{2} & \text{if } \left|\frac{\overline{W}R_j^n}{\overline{W}C_j^n} - \frac{1}{2}\right| > \frac{15-\sqrt{15}}{28} \end{cases}$$

so that it follows that $2\cdot d^n_j \leq \overline{W}C^n_j$. In the case in which $2\cdot d^n_j = \overline{W}C^n_j$ it follows that $q_j(x;\overline{w}^n)$ coincides with a two-degree polynomial, which is monotonically increasing in I_j, as we have seen in Observation 1. Therefore, we can suppose that $2\cdot d^n_j < \overline{W}C^n_j$ and $|\frac{\overline{W}R^n_j}{\overline{W}C^n_j} - \frac{1}{2}| \leq \frac{15-\sqrt{15}}{28}$. In this situation, $\frac{dq_j(x;\overline{w}^n)}{dx}$ reaches a minimum at point $x_{MI} = x_j + \frac{\Delta x}{3}(\frac{2\cdot\overline{W}R^n_j - \overline{W}C^n_j}{2\cdot d^n_j - \overline{W}C^n_j})$ (see Observation 3). Given that

$$\overline{W}R^n_j > \frac{1}{2}\overline{W}C^n_j \Rightarrow \left\{\left(\frac{5}{6}\,\overline{W}C^n_j - \frac{2}{3}\overline{W}R^n_j\right) < \frac{\overline{W}C^n_j}{2} - \frac{\sqrt{15}}{15}\left(2\cdot\overline{W}R^n_j - \overline{W}C^n_j\right) = d^n_j\right\},$$

$$\overline{W}R^n_j < \frac{1}{2}\overline{W}C^n_j \Rightarrow \left\{\left(\frac{1}{6}\,\overline{W}C^n_j + \frac{2}{3}\overline{W}R^n_j\right) < \frac{\overline{W}C^n_j}{2} - \frac{\sqrt{15}}{15}\left(\overline{W}C^n_j - 2\cdot\overline{W}R^n_j\right) = d^n_j\right\},$$

then

$$\overline{W}R^n_j > \tfrac{1}{2}\overline{W}C^n_j \Rightarrow \left\{x_{MI} = x_j + \tfrac{\Delta x}{3}\left(\tfrac{2\cdot\overline{W}R^n_j - \overline{W}C^n_j}{2\cdot d^n_j - \overline{W}C^n_j}\right) < x_j - \tfrac{\Delta x}{2}\right\},$$

$$\overline{W}R^n_j < \tfrac{1}{2}\overline{W}C^n_j \Rightarrow \left\{x_{MI} = x_j + \tfrac{\Delta x}{3}\left(\tfrac{2\cdot\overline{W}R^n_j - \overline{W}C^n_j}{2\cdot d^n_j - \overline{W}C^n_j}\right) > x_j + \tfrac{\Delta x}{2}\right\}.$$

On the other hand,

$$\overline{W}R^n_j > \tfrac{1}{2}\overline{W}C^n_j \Rightarrow \left\{\tfrac{dq_j(x_j \pm \Delta x/2;\overline{w}^n)}{dx} \geq 0 \Longleftrightarrow d^n_j \geq \tfrac{1}{2}\left(4\cdot\overline{W}R^n_j - 3\cdot\overline{W}C^n_j\right)\right\},$$

$$\overline{W}R^n_j < \tfrac{1}{2}\overline{W}C^n_j \Rightarrow \left\{\tfrac{dq_j(x_j \pm \Delta x/2;\overline{w}^n)}{dx} \geq 0 \Longleftrightarrow d^n_j \geq \tfrac{1}{2}\left(\overline{W}C^n_j - 4\cdot\overline{W}R^n_j\right)\right\}.$$

Given that we have supposed that $|\frac{\overline{W}R^n_j}{\overline{W}C^n_j} - \frac{1}{2}| \leq \frac{15-\sqrt{15}}{28}$, then d^n_j verifies the latter inequalities, and thus $q_j(x;\overline{w}^n)$ is monotonically increasing in I_j. □

3.3. Conditions in cells with extrema points. In order to guarantee that $q_j(x;\overline{w}^n)$ has the same shape as the cell-averages $\overline{w}^n_j$ in the domain I_j, we add these requirements to those used in the previous section:

1. $q_j(x;\overline{w}^n)$ has a maximum in I_j if and only if $\overline{w}^n_{j-1} < \overline{w}^n_j > \overline{w}^n_{j+1}$.
2. $q_j(x;\overline{w}^n)$ has a minimum in I_j if and only if $\overline{w}^n_{j-1} > \overline{w}^n_j < \overline{w}^n_{j+1}$.

On the other hand, the definition of θ^n_j that we will use later in (3.13) requires that the following properties are satisfied:

1. If $\overline{w}^n_{j-1} < \overline{w}^n_j > \overline{w}^n_{j+1}$, then $q_j\left(x_j - \frac{\Delta x}{2};\overline{w}^n\right) \geq \frac{1}{2}\left(\overline{w}^n_{j-1} + \overline{w}^n_j\right)$ and $q_j(x_j + \frac{\Delta x}{2};\overline{w}^n) \geq \frac{1}{2}\left(\overline{w}^n_j + \overline{w}^n_{j+1}\right)$.
2. If $\overline{w}^n_{j-1} > \overline{w}^n_j < \overline{w}^n_{j+1}$, then $q_j\left(x_j - \frac{\Delta x}{2};\overline{w}^n\right) \leq \frac{1}{2}\left(\overline{w}^n_{j-1} + \overline{w}^n_j\right)$ and $q_j(x_j + \frac{\Delta x}{2};\overline{w}^n) \leq \frac{1}{2}\left(\overline{w}^n_j + \overline{w}^n_{j+1}\right)$.

According to the notation given in (3.8) and (3.10), supposing that

$$ds4^n_j = \frac{1}{6}\left(5\cdot\overline{W}C^n_j - 4\cdot\overline{W}R^n_j\right),\ ds5^n_j = \frac{1}{6}\left(4\cdot\overline{W}R^n_j + \overline{W}C^n_j\right), \tag{3.11}$$

we define d^n_j in cells with extrema points in the following way:

(B) If $\overline{w}^n_{j-1} < \overline{w}^n_j > \overline{w}^n_{j+1}$ (the cell averages have a maximum), then the following hold:

(B1) If $\overline{W}C2^n_j = 2\cdot\overline{W}C^n_j$, then $d^n_j = ds^n_j \equiv \frac{2}{3}\overline{W}C^n_j - \frac{1}{12}\overline{W}C2^n_j = \frac{1}{2}\overline{W}C^n_j$.

(B2) If $\overline{W}C2_j^n < 2 \cdot \overline{W}C_j^n$, then $d_j^n = \mathrm{Min}\left\{ds2_j^n,\ ds4_j^n,\ ds_j^n\right\}$.

(B3) If $\overline{W}C2_j^n > 2 \cdot \overline{W}C_j^n$, then $d_j^n = \mathrm{Max}\left\{ds3_j^n,\ ds5_j^n,\ ds_j^n\right\}$.

(C) If $\overline{w}_{j-1}^n > \overline{w}_j^n < \overline{w}_{j+1}^n$ (the cell averages have a minimum), then the following hold:

(C1) If $\overline{W}C2_j^n = 2 \cdot \overline{W}C_j^n$, then $d_j^n = ds_j^n = \frac{2}{3}\overline{W}C_j^n - \frac{1}{12}\overline{W}C2_j^n = \frac{1}{2}\overline{W}C_j^n$.

(C2) If $\overline{W}C2_j^n < 2 \cdot \overline{W}C_j^n$, then $d_j^n = \mathrm{Min}\left\{ds3_j^n,\ ds5_j^n,\ ds_j^n\right\}$.

(C3) If $\overline{W}C2_j^n > 2 \cdot \overline{W}C_j^n$, then $d_j^n = \mathrm{Max}\left\{ds2_j^n,\ ds4_j^n,\ ds_j^n\right\}$.

Observation 4. Similar reasoning to that used in the proof of Theorem 3.1 allows us to prove that with this definition of d_j^n, if $\overline{w}_{j-1}^n < \overline{w}_j^n > \overline{w}_{j+1}^n$, then $q_j(x;\overline{w}^n)$ has a maximum at $\left[x_j - \frac{\Delta x}{2}, x_j + \frac{\Delta x}{2}\right]$, verifying these two relations:

$$q_j\left(x_j - \frac{\Delta x}{2};\overline{w}^n\right) \geq \frac{1}{2}\left(\overline{w}_{j-1}^n + \overline{w}_j^n\right), \quad q_j\left(x_j + \frac{\Delta x}{2};\overline{w}^n\right) \geq \frac{1}{2}\left(\overline{w}_j^n + \overline{w}_{j+1}^n\right).$$

Similarly, if we suppose that $\overline{w}_{j-1}^n > \overline{w}_j^n < \overline{w}_{j+1}^n$, then we can verify that $q_j(x;\overline{w}^n)$ has a minimum at $\left[x_j - \frac{\Delta x}{2}, x_j + \frac{\Delta x}{2}\right]$, verifying the following conditions:

$$q_j\left(x_j - \frac{\Delta x}{2};\overline{w}^n\right) \leq \frac{1}{2}\left(\overline{w}_{j-1}^n + \overline{w}_j^n\right), \quad q_j\left(x_j + \frac{\Delta x}{2};\overline{w}^n\right) \leq \frac{1}{2}\left(\overline{w}_j^n + \overline{w}_{j+1}^n\right).$$

3.4. Removing the spurious extrema of $w(x, t^n)$ at points $x_j + \Delta x/2$.

To obtain a nonoscillatory reconstruction we will add some additional requirements for the calculation of $R_j(x, \overline{w}^n)$:

$$\begin{aligned}
&\text{(a)}\, \overline{w}(x_j, t^n) > \overline{w}(x_{j+1}, t^n) \Rightarrow \left(R_j(x_j + \Delta x/2;\overline{w}^n) \geq R_{j+1}(x_j + \Delta x/2;\overline{w}^n)\right),\\
&\text{(b)}\, \overline{w}(x_j, t^n) < \overline{w}(x_{j+1}, t^n) \Rightarrow \left(R_j(x_j + \Delta x/2;\overline{w}^n) \leq R_{j+1}(x_j + \Delta x/2;\overline{w}^n)\right),\\
&\text{(c)}\, \overline{w}(x_j, t^n) = \overline{w}(x_{j+1}, t^n) \Rightarrow \left(R_j(x_j + \Delta x/2;\overline{w}^n) = R_{j+1}(x_j + \Delta x/2;\overline{w}^n)\right).
\end{aligned} \tag{3.12}$$

These properties together to those viewed in sections 3.2 and 3.3 have been defined so that the resulting reconstruction polynomial $w(x, t^n)$, defined in (2.3), presents a nonoscillatory nature in the sense that the number of extrema of $w(x, t^n)$ does not exceed the number shown in the function $\sum_{j=1}^{NX} \overline{w}_j^n \chi_j(x)$. The nonincreasing number of extrema implies convergence along the lines of Liu and Tadmor [20].

To verify (3.12), Liu and Osher [19] consider the modification of the form

$$R_j(x;\overline{w}^n) \equiv \theta_j^n q_j(x;\overline{w}^n) + (1 - \theta_j^n)\overline{w}_j^n, \tag{3.13}$$

where $\theta_j^n \in [0, 1]$. The algorithm that allows us to obtain θ_j^n is described in detail in Liu and Osher [19], although in that reference it is only used when $q_j(x;\overline{w}^n)$ is a conservative parabola. Notice that the value of θ_j^n that appears in formula (3.13) takes a value equal to 1 in all the cells with extrema points (see Liu and Osher [19]). Conditions given in section 3.3 avoid the development of spurious extrema of $R_j(x;\overline{w}^n)$ in the endpoints of an interval with a local maximum or a local minimum.

Remark 1. Parameter θ_j^n used in (3.13) is defined in Liu and Osher [19] so that $(1 - \theta_j^n)$ is proportional to the interface jump $q_{j+1}\left(x_j + \frac{\Delta x}{2};\overline{w}^n\right) - q_j\left(x_j + \frac{\Delta x}{2};\overline{w}^n\right)$. If $d_j^n = ds_j^n$ (given by (3.6)), then the reconstruction is fourth-order accurate. As a consequence of the fourth-order accuracy in polynomials $q_j(x;\overline{w}^n)$ and $q_{j+1}(x;\overline{w}^n)$, the size of the interface jump, and consequently of $(1 - \theta_j^n)$, is of order $O((\Delta x)^4)$. In this way, the definition of $R_j(x;\overline{w}^n)$ given in (3.13) still verifies the properties in (2.4).

However, the definition of d_j^n introduced in cases (A3.1), (A3.2.1), (A3.2.2), (B2), (B3), (C2), and (C3) may cause the value of $(1-\theta_j^n)$ to not be of order $O((\Delta x)^4)$ in a small number of cells, especially near a local maximum, a local minimum, or a discontinuity. For example, if $\overline{W}R_j^n \approx \overline{W}C_j^n$ or $\overline{W}R_j^n \approx 0$ and $S_j^n > 0$, then d_j^n can be closer to $\overline{W}C_j^n/2$ (which is the slope of the parabola) than to ds_j^n. Moreover, in case (A3.2.2) d_j^n can be equal to $\overline{W}C_j^n/2$. In order to achieve the experimental fourth-order of accuracy in some experiments with smooth solutions we will give a special treatment of the cells that are near extrema (see conditions (4.2)). In fact, one should require that $d_j^n/ds_j^n = 1 + O((\Delta x)^3)$ on smooth solutions.

Remark 2. $R_j(x;\overline{w}^n)$ defined in (3.13), with $q_j(x:\overline{w}^n)$ described in subsections 3.1, 3.2, and 3.3, has the same shape as the cell-averages $\{\overline{w}_{j-1}^n, \overline{w}_j^n, \overline{w}_{j+1}^n\}$.

3.5. Definition of slopes d_j^n in (3.2) so that $p_j(x;f^n)$ fulfills conditions given in sections 3.2 and 3.3. A few modifications are needed to compute the nonoscillatory reconstruction from pointvalues for the flux $f_j^n = f(w_j^n)$ which is needed in the Runge–Kutta method with natural continuous extension described in Levy, Puppo, and Russo [16]. According to (3.2) the degree-three polynomial from the pointvalues $f_k^n,\ k \in \{j-2, j-1, j, j+1, j+2\}$, is given by

$$\begin{aligned} p_j\left(x;f^n\right) = f_j^n + d_j^n \cdot \left(\frac{x-x_j}{\Delta x}\right) + \left(\frac{f_{j-1}^n - 2f_j^n + f_{j+1}^n}{2}\right) \cdot \left(\frac{x-x_j}{\Delta x}\right)^2 \\ + \left(\frac{-f_{j-1}^n + f_{j+1}^n - 2d_j^n}{2}\right) \cdot \left(\frac{x-x_j}{\Delta x}\right)^3 . \end{aligned} \tag{3.14}$$

To ensure that $p_j(x;f^n)$ fulfills the requirements of sections 3.2 and 3.3, we define d_j^n in the same way as in those sections with the exceptions that

$$ds1_j^n = 0,\ ds2_j^n = \frac{1}{2}\left(WC_j^n - 8 \cdot WR_j^n\right),\ ds3_j^n = \frac{1}{2}\left(8 \cdot WR_j^n - 7 \cdot WC_j^n\right), \tag{3.15}$$

$$C1 = \frac{\sqrt{3}}{6},\ C2 = \frac{6}{12+\sqrt{3}},\ WC_j^n = f_{j+1}^n - f_{j-1}^n,\ WR_j^n = f_{j+1}^n - f_j^n, \tag{3.16}$$

and cell-averages $\overline{w}_j^n$ are substituted by pointvalues f_j^n. The evaluation of $\partial f/\partial x$ in the Runge–Kutta step is performed by $\theta_j^n \frac{dp_j(x;f^n)}{dx}$, where θ_j^n is defined as in Liu and Osher [19]. Thus, we maintain high accuracy and control over oscillations.

4. Numerical experiments. In order to verify the behavior and accuracy of the numerical schemes that are presented in this paper, several test-type problems with known analytical solution are solved next. Time integrals are performed by a Taylor expansion (Taylor-upwind and Taylor-central schemes) or by the fourth-order Runge–Kutta method with natural continuous extensions developed in Levy, Puppo, and Russo [16] (RK-NCE-central scheme). In this last case the reconstruction defined in section 3.5 will also be used.

Problem 1. We solve the linear transport equation

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial u(x,t)}{\partial x} = 0,\ \ -1 \le x \le 1, \tag{4.1}$$

subject to 2-periodic initial data, $u(x,0) = u_0(x)$. To verify the accuracy of the numerical schemes, different $u_0(x)$ functions have been used.

Table 4.1

Linear transport equation (4.1) *with* $u_0(x) = \sin(\pi\, x)$. *Errors at* $T = 10$.

(a) Taylor-upwind scheme, with $\Delta t = 0.8\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
40	$3.071427\ 10^{-5}$	4.14	$2.498590\ 10^{-5}$	4.16
80	$1.746260\ 10^{-6}$	4.05	$1.396909\ 10^{-6}$	4.06
160	$1.056129\ 10^{-7}$	4.02	$8.374969\ 10^{-8}$	4.02
320	$6.529029\ 10^{-9}$		$5.152456\ 10^{-9}$	

(b) Taylor-central scheme, with $\Delta t = 0.4\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
40	$5.558861\ 10^{-5}$	4.10	$7.473019\ 10^{-5}$	4.02
80	$3.231451\ 10^{-6}$	4.14	$4.594014\ 10^{-6}$	3.86
160	$1.836719\ 10^{-7}$	4.09	$3.153211\ 10^{-7}$	3.87
320	$1.076991\ 10^{-8}$		$2.155580\ 10^{-8}$	

(c) RK-NCE-central scheme, with $\Delta t = 0.25\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
40	$1.531422\ 10^{-4}$	4.18	$1.689413\ 10^{-4}$	4.35
80	$8.423959\ 10^{-6}$	4.06	$8.305782\ 10^{-6}$	4.02
160	$5.053152\ 10^{-7}$	4.06	$5.130312\ 10^{-7}$	3.97
320	$3.034160\ 10^{-8}$		$3.280898\ 10^{-8}$	

The first function is $u_0(x) = \sin(\pi\, x)$. Table 4.1 shows the errors and the experimental order of accuracy in L^1 and L^∞ norms at time $T = 10$. NX indicates the total number of cells so that the step size $\Delta x = 2/NX$. Using a Taylor expansion for the time evolution, we have selected a time step so that $\Delta t = 0.8\Delta x$ in the upwind scheme, whereas in the central scheme $\Delta t = 0.4\Delta x$. When we use a RK-NCE-central scheme, $\Delta t = 0.25\Delta x$ as in Levy, Puppo, and Russo [16]. Table 4.1 shows that numerical schemes described in this paper are about fourth-order accuracy in L^1 and L^∞ norms, which is an improvement over the schemes described in Liu and Osher [19] and Liu and Tadmor [20], which are third-order schemes.

The second initial condition chosen is $u_0(x) = \sin^4(\pi\, x)$. Table 4.2 shows the errors in L^1 and L^∞ norms at time $T = 10$. The schemes presented here maintain the fourth-order accuracy, even with finer grids, without the need of satisfying the local maximum principle described in Liu and Osher [19]. The nonconsideration of that local maximum principle implies that the θ_j that appear in formula (3.13) take a value equal to 1 in all the cells with extrema points (see Liu and Osher [19]). To improve the accuracy of the numerical schemes presented in this paper, in the results shown in Table 4.2 we have added two additional requirements (see Remark 1):

$$\begin{aligned} &\text{If } \overline{w}^n_{j-1} < \overline{w}^n_j > \overline{w}^n_{j+1} \Rightarrow d^n_{j-1} = ds^n_{j-1},\ d^n_{j+1} = ds^n_{j+1}.\\ &\text{If } \overline{w}^n_{j-1} > \overline{w}^n_j < \overline{w}^n_{j+1} \Rightarrow d^n_{j-1} = ds^n_{j-1},\ d^n_{j+1} = ds^n_{j+1}. \end{aligned} \tag{4.2}$$

Thus, we avoid the slope d^n_j taking a value close to $\overline{W}C^n_j/2$ in the neighboring cells to those containing the extrema points of the solution. Remember that with such a slope, $q_j(x; \overline{w}^n)$ coincides with a conservative quadratic polynomial. As mentioned in Remark 1, parameter θ^n_j is defined in such a way that $(1 - \theta^n_j)$ is proportional to the interface jump of the cell centered in x_j. Conditions (4.2) cause the interface jump to be lower at the boundary of cells with extrema points.

The third initial condition is a discontinuous 2-periodic function that was used in Balsara and Shu [2]. This is a severe problem since it consists of a combination of functions that are not smooth, with other ones, which are smooth, but with a high

TABLE 4.2

Linear transport equation (4.1) *with* $u_0(x) = \sin^4(\pi x)$. *Errors at* $T = 10$.

(a) Taylor-upwind scheme, with $\Delta t = 0.8\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
80	2.430672 10^{-4}	4.16	2.937589 10^{-4}	4.17
160	1.362647 10^{-5}	4.04	1.636374 10^{-5}	3.93
320	8.262462 10^{-7}	4.01	1.075533 10^{-6}	4.06
640	5.110279 10^{-8}		6.461631 10^{-8}	

(b) Taylor-central scheme, with $\Delta t = 0.4\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
80	3.499147 10^{-4}	4.14	5.062172 10^{-4}	4.49
160	1.977973 10^{-5}	4.06	2.258267 10^{-5}	3.96
320	1.189089 10^{-6}	3.99	1.454815 10^{-6}	4.06
640	7.460379 10^{-8}		8.713768 10^{-8}	

(c) RK-NCE-central scheme, with $\Delta t = 0.25\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
80	1.052742 10^{-3}	4.15	1.637417 10^{-3}	4.60
160	5.930366 10^{-5}	4.07	6.747230 10^{-5}	4.02
320	3.535521 10^{-6}	3.98	4.154436 10^{-6}	4.03
640	2.237979 10^{-7}		2.539025 10^{-7}	

gradient in zones close to the peaks. The initial condition is given by

$$u_0(x) = \begin{cases} \frac{1}{6}\left(G(x,\beta,z-\delta) + G(x,\beta,z+\delta) + 4G(x,\beta,z)\right), & -0.8 \le x \le -0.6, \\ 1, & -0.4 \le x \le -0.2, \\ 1 - |10\,(x-0.1)|, & 0 \le x \le 0.2, \\ \frac{1}{6}\left(F(x,\alpha,a-\delta) + F(x,\alpha,a+\delta) + 4F(x,\alpha,a)\right), & 0.4 \le x \le 0.6, \\ 0 & \text{otherwise} \end{cases} \tag{4.3}$$

defined as

$$G(x,\beta,z) = e^{-\beta(x-z)^2}, \quad F(x,\alpha,a) = \sqrt{\operatorname{Max}\left(1-\alpha^2(x-a)^2, 0\right)}. \tag{4.4}$$

The constants that appear in (4.3) and (4.4) are given by

$$a = 0.5; \;\; z = -0.7; \;\; \delta = 0.005; \;\; \alpha = 10; \;\; \beta = \frac{\log(2)}{36\delta^2}. \tag{4.5}$$

Figure 4.1 shows the numerical results obtained at time $T = 20$, with the Taylor-central scheme developed in this paper, comparing the numerical solution with the analytical solution which is represented by a continuous line. Unlike the solutions presented in Balsara and Shu [2], here we have considered a coarser grid with $NX = 500$. This shows the greater accuracy of our scheme, in comparison with the conservative quadratic polynomial developed in Liu and Osher [19], in particular at the peaks of the Gaussian curve and in the triangle. In addition, the profiles are more symmetrical than those computed in Levy, Puppo, and Russo [16], and the values of the numerical solution are bounded by the maximum and minimum of the initial condition despite not using the maximum principle property given in Liu and Osher [19]. This is a condition that is not fulfilled when $q_j(x;\overline{w}^n)$ is replaced by the two-degree polynomial used in this reference. Previous remarks for the Taylor-central scheme are also valid for the Taylor-upwind and RK-NCE-central schemes.

The last initial condition is given by

$$u_0(x+0.5) = \begin{cases} -x\sin\left(\frac{3}{2}\pi\,x^2\right) & \text{if } -1 < x < -\frac{1}{3}, \\ |\sin(2\pi\,x)| & \text{if } |x| \le \frac{1}{3}, \\ 2x - 1 - \sin(3\pi\,x)/6, & \text{if } \frac{1}{3} < x \le 1, \end{cases} \tag{4.6}$$

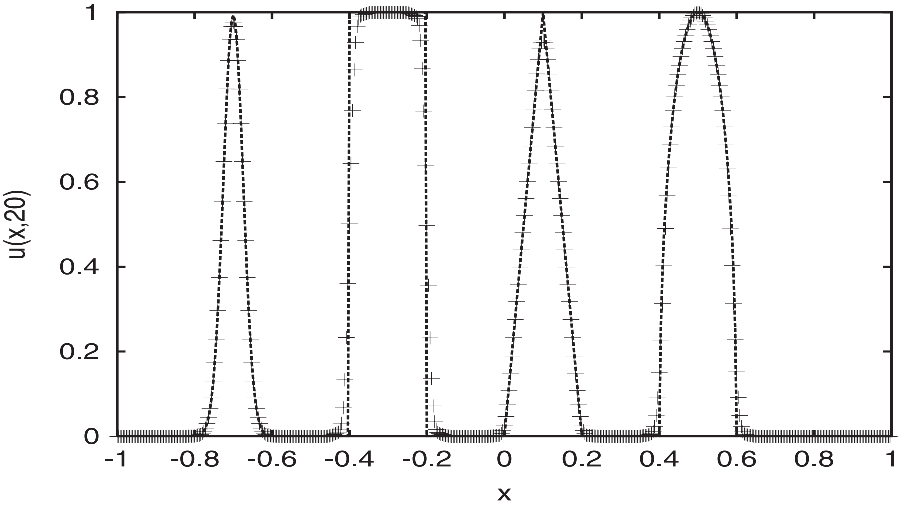


FIG. 4.1. *Numerical and analytical solutions of problem* 1 *at* $T = 20$ *with* $u_0(x)$ *defined by* (4.3)–(4.5)*, considering a grid with* $NX = 500$*. We have used the Taylor-central scheme, with* $\Delta t = 0.45\Delta x$*.*

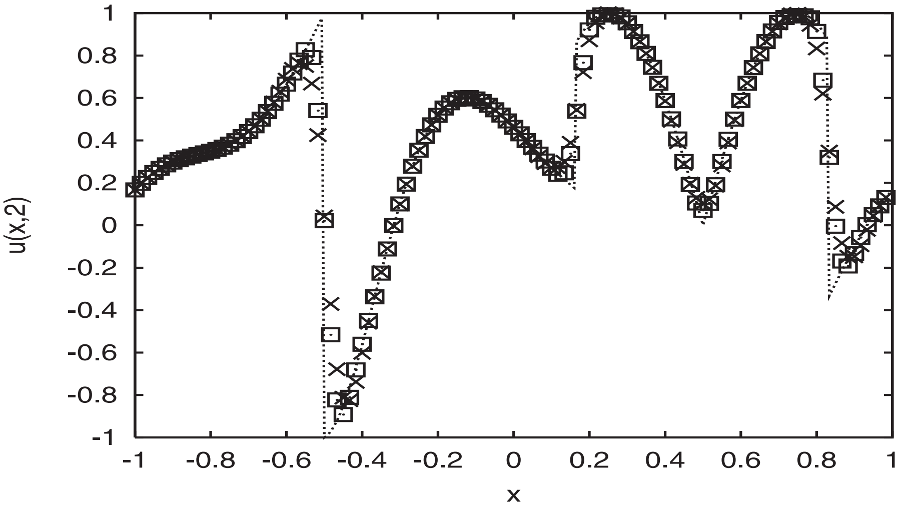


FIG. 4.2. *Numerical and analytical solutions of Problem* 1 *at* $T = 2$ *with* $u_0(x)$ *defined by* (4.6)*, considering* $NX = 120$ *and* $\Delta t = 0.4\Delta x$*. The solution marked with* ⊡ *has been computed by our Taylor-central scheme. The solution marked with* × *is the solution obtained considering that* $q_j(x; \overline{w}^n)$ *coincides with the quadratic polynomial of Liu and Osher* [19] *and Liu and Tadmor* [20].

supposing that it extends to the entire 2-period domain.

Condition (4.6) consists of a function highly discontinuous, used in the numerical experiments developed in Harten [9]. Figure 4.2 shows the results obtained at $T = 2$, with $NX = 120$, comparing the results obtained with our Taylor-central scheme and those in which $q_j(x; \overline{w}^n)$ is the conservative two-degree polynomial used in Liu and Osher [19] and Liu and Tadmor [20]. It can be observed that the greater accuracy of our schemes is especially noted around the discontinuities. On the other hand, the

Table 4.3

Errors in the resolution of Burgers's equation with conditions (4.7) *at* $T = 0.3$.

(a) Taylor-upwind scheme, with $\Delta t = 0.6\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
80	$2.551210\ 10^{-6}$	4.02	$1.018463\ 10^{-5}$	4.06
160	$1.567119\ 10^{-7}$	4.01	$6.108933\ 10^{-7}$	3.94
320	$9.734252\ 10^{-9}$	4.00	$3.988465\ 10^{-8}$	4.07
640	$6.090892\ 10^{-10}$	3.95	$2.370593\ 10^{-9}$	3.37
1280	$3.947035\ 10^{-11}$		$2.291545\ 10^{-10}$	

(b) Taylor-central scheme, with $\Delta t = 0.33\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
80	$1.536227\ 10^{-6}$	4.18	$8.083824\ 10^{-6}$	4.22
160	$8.454566\ 10^{-8}$	4.10	$4.340597\ 10^{-7}$	4.18
320	$4.916180\ 10^{-9}$	4.07	$2.392437\ 10^{-8}$	4.10
640	$2.935978\ 10^{-10}$	3.99	$1.393063\ 10^{-9}$	4.13
1280	$1.846273\ 10^{-11}$		$7.967693\ 10^{-11}$	

(c) RK-NCE-central scheme, with $\Delta t = 0.18\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
80	$2.703482\ 10^{-6}$	4.18	$1.875872\ 10^{-5}$	4.22
160	$1.496377\ 10^{-7}$	4.21	$1.006314\ 10^{-6}$	4.24
320	$8.089214\ 10^{-9}$	4.17	$5.322066\ 10^{-8}$	4.33
640	$4.495486\ 10^{-10}$	4.12	$2.641527\ 10^{-9}$	4.13
1280	$2.589858\ 10^{-11}$		$1.506755\ 10^{-10}$	

numerical solution is delimited by the maximum and minimum of the initial condition without the need of satisfying the local maximum principle of Liu and Osher [19].

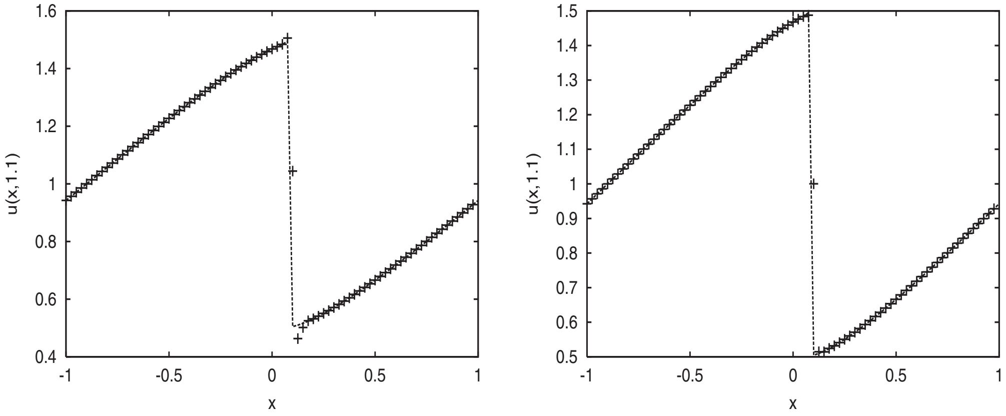
Problem 2. Burgers' equation is solved with 2-periodic initial data:

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial\left(\frac{1}{2}u^2(x,t)\right)}{\partial x} = 0, \quad -1 \le x \le 1, \quad u\left(x,0\right) = 1 + \frac{1}{2}\sin(\pi x). \tag{4.7}$$

Recall that the analytical solution of this problem is smooth up to the critical time $T = 2/\pi$. Liu and Osher [19] and Liu and Tadmor [20] show the results obtained using a parabolic reconstruction at $T = 0.3$. Table 4.3 presents the numerical errors obtained with our schemes, together with the experimental order of accuracy at $T = 0.3$. Our schemes (upwind and central) have an order of accuracy which is about 4 in both L^1 and L^∞ norms. However, the maximum order obtained with the schemes described by Liu and Osher [19] (Taylor-upwind scheme, $\Delta t = 0.6\Delta x$) and Liu and Tadmor [20] (Taylor-central scheme, $\Delta t = 0.33\Delta x$) is lower than 2.3 in the L^∞ norm and lower than 2.87 in the L^1 norm. In the RK-NCE-central scheme we have chosen $\Delta t = 0.18\Delta x$ as in Levy, Puppo, and Russo [16].

At $T = 1.1$ the analytical solution of problem (4.7) develops a discontinuity. Our numerical scheme maintains an order of accuracy of about 4 when the errors are calculated at a distance equal to 0.1 away from the discontinuity. Figure 4.3(a) shows the result obtained with our Taylor-central scheme. Like in the scheme developed in Liu and Tadmor [20], the numerical solution is not bounded by the maximum and minimum of the analytical solution. In order to obtain this property it is necessary to add the maximum principle requirement described in Liu and Osher [19], as we can see in Figure 4.3(b). However, without the condition of maximum principle, the numerical solution retains results of the same quality as the analytical solution. Previous remarks are also valid for the Taylor-upwind and RK-NCE-central schemes.

Problem 3. Here we apply the schemes developed in this paper to Buckley–



(a) (b)

FIG. 4.3. *Numerical and analytical solutions of Problem* 2 *at* $T = 1.1$*, with our Taylor-central scheme, considering a grid with* $NX = 80$ *and* $\Delta t = 0.33\Delta x$*.* (a) *Not using the local maximum principle.* (b) *Using the local maximum principle.*

Leverett's problem, whose flux is nonconvex:

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \quad -1 \leq x \leq 1, \quad f(u) = \frac{4u^2}{4u^2 + (1-u)^2} \tag{4.8}$$

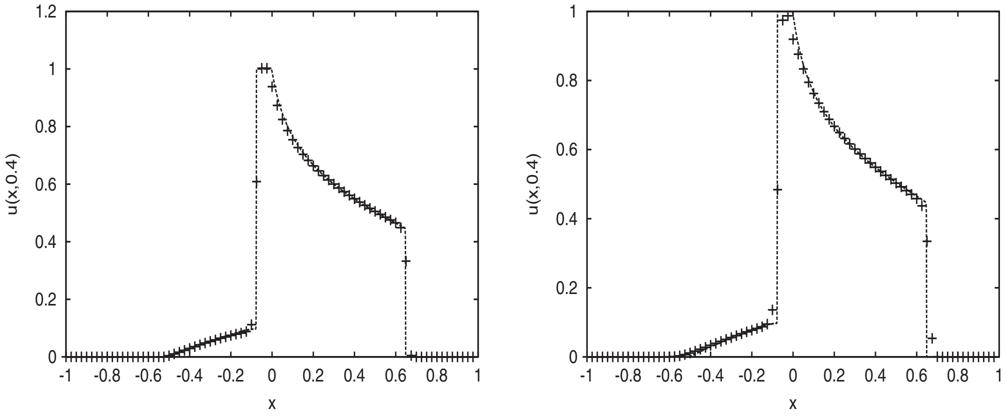
subject to the initial condition

$$u_0(x) = \begin{cases} 1 & x \in [-0.5, 0], \\ 0 & \text{otherwise.} \end{cases} \tag{4.9}$$

Similarly to Liu and Osher [19] and Jiang et al. [11], we have computed the solution at $T = 0.4$ with our Taylor-upwind and central schemes. Figure 4.4 shows the results obtained with $NX = 80$. In contrast to the scheme described in Liu and Osher [19], our Taylor-upwind scheme presents instabilities in the solution of the problem (4.8)–(4.9) for $\Delta t = 0.3\Delta x$. However, it presents very accurate solutions when $\Delta t = 0.25\Delta x$. Moreover, the condition of the local maximum principle described in Liu and Osher [19] has not been necessary, as shown in Figure 4.4(a). The central schemes described in this paper provide smoother solutions than the upwind scheme for the resolution of the problem under study (Figures 4.4(a)–4.4(b)), although the three schemes present a similar behavior.

Euler equations of gas dynamics. We test our schemes on the system of Euler equations of gas dynamics for a gas with $\gamma = 1.4$. We consider a problem with smooth analytical solution and the two Riemann problems studied in Liu and Tadmor [20]. The variables ρ, m, E denote the density, momentum, and total energy per unit volume, respectively. Moreover, p denotes the pressure and v denotes the velocity.

Problem 4. The initial condition is set to be $\rho(x,0) = 1 + 0.2\sin(\pi x)$, $v(x,0) = 1$, $p(x,0) = 1$, with 2-periodic boundary conditions, $-1 \leq x \leq 1$. The exact solution is $\rho(x,t) = 1 + 0.2\sin(\pi(x-t))$, $v = 1$, $p = 1$. We compute the solution at $T = 2$ as in Qiu and Shu [22], using our RK-NCE-central scheme with the componentwise reconstruction described in this paper. Table 4.4 shows the results obtained considering $\Delta t = 0.1\Delta x$. We can see that our scheme achieves its designed order of accuracy.



(a) (b)

Fig. 4.4. *Numerical and analytical solutions of Problem* 3 *at* $T = 0.4$*, using a grid with* $NX = 80$*.* (a) *Taylor-upwind scheme with* $\Delta t = 0.25\Delta x$*.* (b) *Taylor-central scheme with* $\Delta t = 0.1\Delta x$*.*

Table 4.4
Errors of density in the resolution of Problem 4 *at* $T = 2$*.*

RK-NCE-central scheme, with $\Delta t = 0.1\Delta x$.

NX	L^1 error	L^1 order	L^∞ error	L^∞ order
40	$9.910941\ 10^{-6}$	4.53	$7.147074\ 10^{-6}$	4.53
80	$3.870207\ 10^{-7}$	4.26	$3.083984\ 10^{-7}$	4.27
160	$2.020669\ 10^{-8}$	4.10	$1.601040\ 10^{-8}$	4.11
320	$1.178203\ 10^{-9}$	4.04	$9.296230\ 10^{-10}$	4.04
640	$7.181752\ 10^{-11}$		$5.656586\ 10^{-11}$	

Problem 5. Shock tube problem with Sod's initial data [24]:

$$\begin{cases} (\rho_l, m_l, E_l) = (1, 0, 2.5), & x < 0.5, \\ (\rho_r, m_r, E_r) = (0.125, 0, 0.25), & x > 0.5. \end{cases}$$

Problem 6. Shock tube problem with the Lax's initial data [14]:

$$\begin{cases} (\rho_l, m_l, E_l) = (0.445, 0.311, 8.928), & x < 0.5, \\ (\rho_r, m_r, E_r) = (0.5, 0, 1.4275), & x > 0.5. \end{cases}$$

In Problems 5 and 6 the computational domain is $[0, 1]$. We integrate the equations to $T = 0.16$, i.e., before the perturbations reach the boundary of the computational region (free flow boundary conditions). We compute the numerical solution with our RK-NCE-central scheme, using the componentwise reconstruction described in this paper. In Figure 4.5 we plot the computed solution with $NX = 200$ grid points as in Liu and Tadmor [20]. We observe the improved resolution in comparison to the corresponding third-order central results of that reference. However, our solutions present more oscillations, which is in agreement with what is commented on in [22]. Qiu and Shu [22] conclude that the componentwise central WENO scheme will become more oscillatory when the order of accuracy increases. Qiu and Shu [22] also observe that the oscillations disappear when the reconstruction is performed on characteristic variables. It is conceivable to expect that the same thing will happen with the new reconstruction proposed in this paper. This will be explored in some future work.

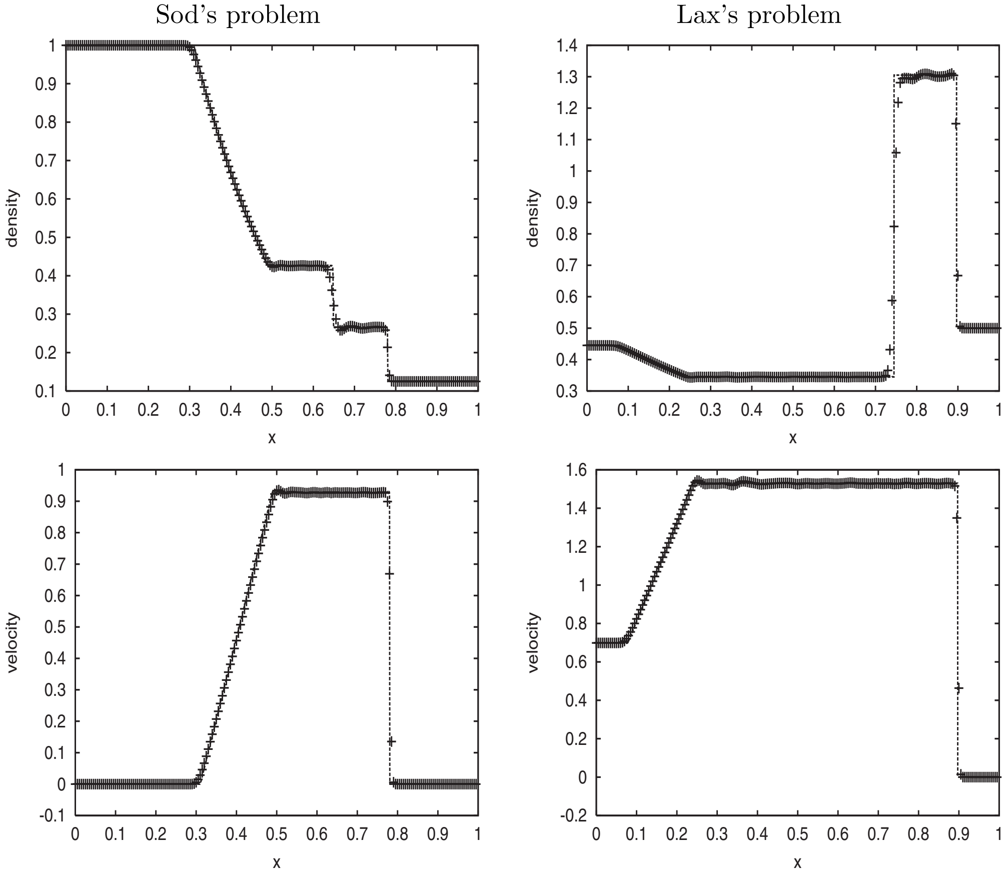


FIG. 4.5. *Numerical and analytical solutions of Euler equations for Problems* 5 *and* 6 *at time* $T = 0.16$*, with our RK-NCE-central scheme, considering a grid with* $NX = 200$*. We have considered that* $\Delta t = 0.1\Delta x$ *in the Sod problem and* $\Delta t = 0.09\Delta x$ *in the Lax problem.*

5. Conclusions. This paper presents a new fourth-order nonoscillatory reconstruction procedure for upwind and central schemes that solves hyperbolic conservation laws in one spatial dimension, improving the accuracy of the schemes developed in Liu and Osher [19] and Liu and Tadmor [20]. We have proved that our schemes are *number of extrema decreasing* and this implies convergence along the lines of Liu and Tadmor [20]. Numerical experiments have shown that our schemes are fourth-order accurate, conservative, and nonoscillatory, presenting good behavior without the need of satisfying the local maximum principle described in Liu and Osher [19]. Future research will extend these schemes to several spatial variables. We also may study the linear stability of these schemes by a procedure similar to that developed in Bianco, Puppo, and Russo [3].

Acknowledgments. We express our gratitude to the four anonymous reviewers for their helpful comments. We thank Dr. Arturo Hidalgo López for his suggestions. We would like to thank the Foreign Language Coordination Office at the Polytechnic University of Valencia for their help in translating this paper.

REFERENCES

[1] A. Balaguer, C. Conde, J. A. López, and V. Martínez, *A finite volume method with a modified ENO scheme using a Hermite interpolation to solve advection-diffusion equations*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 2339–2371.

[2] D. S. Balsara and C. W. Shu, *Monotonicity preserving weighted essentially non-oscillatory schemes with increasingly high order of accuracy,* J. Comput. Phys., 160 (2000), pp. 405–452.

[3] F. Bianco, G. Puppo, and G. Russo, *High-order central schemes for hyperbolic systems of conservation laws*, SIAM J. Sci. Comput., 21 (1999), pp. 294–322.

[4] J. Casper and H. L. Atkins, *A finite volume high-order ENO scheme for two-dimensional hyperbolic systems*, J. Comput. Phys., 106 (1993), pp. 62–76.

[5] P. Colella and P. Woodward, *The piecewise parabolic method (PPM) for gas-dynamical simulations*, J. Comput. Phys., 54 (1984), pp. 174–201.

[6] K. O. Friedrichs and P. D. Lax, *Systems of conservation equations with a convex extension*, Proc. Nat. Acad. Sci. U.S.A., 68 (1971), pp. 1686–1688.

[7] S. K. Godunov, *A finite difference method for the numerical computation of discontinuous solutions of the equations of fluid dynamics*, Mat. Sb., 47 (1959), pp. 271–290.

[8] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, *Uniformly high order accurate essentially non-oscillatory schemes* III, J. Comput. Phys., 71 (1987), pp. 231–303.

[9] A. Harten, *ENO schemes with subcell resolution*, J. Comput. Phys., 83 (1989), pp. 148–184.

[10] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.

[11] G.-S. Jiang, D. Levy, C.-T. Lin, S. Osher, and E. Tadmor, *High-resolution nonoscillatory central schemes with nonstaggered grids for hyperbolic conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2147–2168.

[12] G.-S. Jiang and E. Tadmor, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.

[13] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 214–282.

[14] P. D. Lax, *Weak solutions of non-linear hyperbolic equations and their numerical computation*, Comm. Pure Appl. Math., 7 (1954), pp. 159–193.

[15] D. Levy and E. Tadmor, *Non-oscillatory central schemes for the incompressible* 2*D euler equations*, Math. Res. Lett., 4 (1997), pp. 321–340.

[16] D. Levy, G. Puppo, and G. Russo, *Central WENO schemes for hyperbolic systems of conservation laws*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 547–571.

[17] D. Levy, G. Puppo, and G. Russo, *A fourth-order central WENO scheme for multidimensional hyperbolic systems of conservation laws*, SIAM J. Sci. Comput., 24 (2002), pp. 480–506.

[18] X.-D. Liu, S. Osher, and T. Chan, *Weighted essentially non-oscillatory schemes*, J. Comput. Phys., 115 (1994), pp. 200–212.

[19] X.-D. Liu and S. Osher, *Nonoscillatory high order accurate self-similar maximum principle satisfying shock capturing schemes* I, SIAM J. Numer. Anal., 33 (1996), pp. 760–779.

[20] X.-D. Liu and E. Tadmor, *Third order nonoscillatory central scheme for hyperbolic conservation laws*, Numer. Math., 79 (1998), pp. 397–425.

[21] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.

[22] J. Qiu and C.-W. Shu, *On the construction, comparison, and local characteristic decomposition for high-order central WENO schemes*, J. Comput. Phys., 183 (2002), pp. 187–209.

[23] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes* I, J. Comput. Phys., 83 (1988), pp. 32–78.

[24] G. Sod, *A survey of several finite difference methods for systems of nonlinear hyperbolic conservation laws*, J. Comput. Phys., 22 (1978), pp. 1–31.

[25] B. Van Leer, *Towards the ultimate conservative difference scheme* V. *A second order sequel to Godunov's method*, J. Comput. Phys., 32 (1979), pp. 101–136

[26] M. Zennaro, *Natural continuous extensions of Runge–Kutta methods*, Math. Comp., 46 (1986) pp. 119–133.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 474–501

CONVERGENCE OF A NUMERICAL SCHEME FOR STRATIGRAPHIC MODELING*

R. EYMARD†, T. GALLOUËT‡, V. GERVAIS§, AND R. MASSON§

Abstract. In this paper, we consider a multilithology diffusion model used in the field of stratigraphic basin simulations to simulate large scale depositional transport processes of sediments described as a mixture of L lithologies. This model is a simplified one for which the surficial fluxes are proportional to the slope of the topography and to a lithology fraction with unitary diffusion coefficients.

The main variables of the system are the sediment thickness h, the L surface concentrations c_i^s in lithology i of the sediments at the top of the basin, and the L concentrations c_i in lithology i of the sediments inside the basin. For this simplified model, the sediment thickness decouples from the other unknowns and satisfies a linear parabolic equation. The remaining equations account for the mass conservation of the lithologies, and couple, for each lithology, a first order linear equation for c_i^s with a linear advection equation for c_i for which c_i^s appears as an input boundary condition. For this coupled system, a weak formulation is introduced.

The system is discretized by an implicit time integration and a cell centered finite volume method. This numerical scheme is shown to satisfy stability estimates and to converge, up to a subsequence, to a weak solution of the problem.

Key words. finite volume method, stratigraphic modeling, linear first order equations, convergence analysis, weak formulation

AMS subject classifications. 35M10, 35Q99, 65M12

DOI. 10.1137/S0036142903426208

1. Introduction.

Recent progress in geosciences, and more especially in seismic- and sequence-stratigraphy, have improved the understanding of sedimentary basins infill. Indeed, the sediment's architecture is the response to complex interactions between the available space created in the basin by sea level variations, tectonic, compaction, the sediment supply (boundary fluxes, sediment production), and the transport of the sediments at the surface of the basin. In order to have a quantified view of this response and to determine the relative influence of each involved process, stratigraphic models have been developed.

Among basin infill models considering the dynamics of sediment transport, authors usually distinguish between fluid-flow and dynamic-slope models (see [14], [15]). The first ones use fluid-flow equations and empirical algorithms to simulate the transport of sediments in the hydrodynamic flow field (see, e.g., [16]). They provide an accurate description of depositional processes for small scales in time and space, but, at larger scale's such as basin scales, they are computationally too expensive.

Dynamic-slope models use mass conservation equations of sediments combined with diffusive transport laws. These laws do not describe each geological process in detail but average over these processes (river transport, creep, slumps, and small

*Received by the editors April 15, 2003; accepted for publication (in revised form) August 20, 2004; published electronically June 30, 2005.
http://www.siam.org/journals/sinum/43-2/42620.html

†Département de Mathématiques, Université de Marne La Vallée, 5 boulevard Descartes, Champs sur Marne, F-77454, Marne La Vallée, Cedex 2, France (eymard@math.univ-mlv.fr).

‡LATP, Université de Provence, 39 rue Frédéric Joliot Curie, 13453 Marseille Cedex 13, France (Thierry.Gallouet@cmi.univ-mrs).

§Institut Français du Pétrole, 1 et 4 av. de Bois Préau, 92852 Rueil Malmaison Cedex, France (veronique.gervais@ifp.fr, roland.masson@ifp.fr).

slides). One can refer to [1], [7], [8], [10], [14], and [17] for a detailed description of these models. The dynamic-slope models have been shown to offer a good description of sedimentation and erosion processes for large time scales (greater than 10^4 y) and basin space scales (greater than 1 km).

We consider here a dynamic-slope model simulating the evolution of a sedimentary basin in which sediments are modeled as a mixture of several lithologies $i = 1, \dots, L$ characterized by different grain size populations. The surficial transport process is a multilithology diffusive model introduced in [14], for which the fluxes are proportional to the slope of the topography and to a lithology fraction c_i^s of the sediments at the surface of the basin (see also [9] and [5]). In what follows, a simplified model is considered for which the diffusion coefficients are taken equal to one. It results that the sediment thickness variable h is decoupled from the other unknowns of the system (i.e., for each lithology, the surface concentration c_i^s and the concentration c_i in lithology i of the sediments in the basin) and satisfies a linear parabolic equation.

The remaining equations accounting for the mass conservation of the lithologies couple, for all $i = 1, \dots, L$, a first order linear equation for the surface concentration variable c_i^s and a linear advection equation for the basin concentration variable c_i for which c_i^s appears as an input boundary condition at the top of the basin. In order to cope with the difficulty of defining the trace of the basin concentration c_i at the top of the basin, an original weak formulation is introduced for this coupled problem.

The system is discretized by an implicit integration in time and a cell centered finite volume scheme in space. The objective of this article is to prove, under Hypothesis 1, the convergence of the approximate solutions for the sediment thickness variable h and for the concentration variables c_i^s, c_i, $i = 1, \dots, L$, up to a subsequence, to a weak solution of problem (2.7) in the sense of Definition 2.1 as the mesh size and time step tend to 0. We state this result in Theorem 3.3 in section 3, after presenting the mathematical model, the weak formulation, and the finite volume scheme.

Regarding the coupling between the parabolic equation for h and the first order linear equations for the variables c_i^s, $i = 1, \dots, L$, our model shares some common features with two phase Darcy flows for which such coupling between an elliptic or parabolic equation and a hyperbolic equation also comes in. The convergence of various numerical schemes for such models have been the subject of several studies. For example, one can refer to [12] for finite differences, to [2] and [3] for mixed and hybrid finite element methods, to [4] for the control volume finite element discretization, and to [19], [18], and [6] for the cell centered finite volume scheme.

The main originality of this work is rather concerned with the coupling between the surface and the basin concentration variables.

The remaining of the paper outlines as follows. The mathematical model and its weak formulation are defined in section 2, and the fully implicit finite volume discretization is derived in section 3. In section 4, stability and error estimates on the discrete solution for the sediment thickness and its time derivative are obtained. Finally, the convergence of the approximate solutions to a weak solution of the problem is proved in section 5.

2. Mathematical model and weak formulation.

A basin model specifies the geometry defined by the basin horizontal extension, the position of its base due to vertical tectonics displacements, and the sea level variations. It provides a description of the sediments considered as a mixture of different lithologies such as sand or shale. Finally, it specifies the sediment transport laws and their coupling, as well as the sediment fluxes at the boundary of the basin (boundary conditions).

In this paper, the multilithology diffusion model described in [14], [9], and [5] is studied in a simplified case for which the diffusion coefficients of the lithologies are equal (to one to fix ideas). Also, for the sake of simplicity, the tectonics displacements as well as the sea level variations are not considered in what follows.

The projection of the basin on a reference horizontal plane is considered as a fixed domain $\Omega \subset \mathbb{R}^d$, defining the horizontal extension of the basin, with $d = 1$ for two dimensional basin models and $d = 2$ for three dimensional models.

We denote by h the sediment thickness variable defined on the domain $\mathcal{D} = \Omega \times \mathbb{R}_+^*$ and by $\mathcal{B}$ the domain $\{(x, z, t)$ such that $(x, t) \in \mathcal{D}, z < h(x, t)\}$.

The sediments are modeled as a mixture of L lithologies characterized by their grain size population. Each lithology, $i = 1, \ldots, L$, is considered as an uncompressible material of constant grain density and null porosity. On each point of the basin, the mixture is described by its composition given by the concentrations c_i, defined on $\mathcal{B}$, and such that $c_i \geq 0$ for $i = 1, \ldots, L$, and $\sum_{i=1}^{L} c_i = 1$.

The model assumes that the sediment fluxes are nonzero only at the surface of the basin (i.e., for $z = h$). The sediments transported by these surficial fluxes, i.e., which are deposited at the surface of the basin in case of sedimentation, or which pass through the surface in case of erosion, are characterized by their concentrations denoted by c_i^s, defined on $\mathcal{D}$, and such that $c_i^s \geq 0$ for $i = 1, \ldots, L$, and $\sum_{i=1}^{L} c_i^s = 1$.

Since the compaction is not considered, no change in time of the concentration c_i can occur inside the basin. It results that $\partial_t c_i = 0$ on $\mathcal{B}$. The evolution of c_i is governed by the boundary condition at the top of the basin stating that $c_i|_{z=h} = c_i^s$ in the case of sedimentation $\partial_t h > 0$. Let $\mathcal{D}^+$ denote the domain $\{(x, t) \in \mathcal{D}$ such that $\partial_t h(x, t) > 0\}$; then c_i satisfies the conservation equation;

$$(2.1) \qquad \begin{cases} \partial_t c_i = 0 & \text{on } \mathcal{B}, \\ c_i|_{z=h} = c_i^s & \text{on } \mathcal{D}^+. \end{cases}$$

The conservation of the thickness fraction in lithology i

$$(2.2) \qquad \mathcal{M}_i(x, t) = \int_0^{h(x,t)} c_i(x, z, t) dz, \ (x, t) \in \mathcal{D},$$

with $\sum_{i=1}^{L} \mathcal{M}_i = h$, states that for all $i = 1, \ldots, L$

$$(2.3) \qquad \begin{cases} \partial_t \mathcal{M}_i + \operatorname{div} \mathbf{f}_i = 0 & \text{on } \mathcal{D}, \\ \sum_{i=1}^{L} c_i^s = 1 & \text{on } \mathcal{D}. \end{cases}$$

In the multilithology diffusive model described in [14], the flux $\mathbf{f}_i$ is proportional to the gradient of the topography h and to the concentration c_i^s, with a diffusion coefficient k_i. In what follows, we shall restrict ourselves to the simplified case $k_i = 1$ for all $i = 1, \ldots, L$, i.e., $\mathbf{f}_i := -c_i^s \nabla h$, so that the sediment thickness variable h decouples from the concentrations and satisfies a linear parabolic equation (see (2.6)).

Neumann boundary conditions are imposed to h on $\partial\Omega \times \mathbb{R}_+^*$,

$$\nabla h \cdot \vec{n} = g \text{ on } \partial\Omega \times \mathbb{R}_+^*,$$

with $\vec{n}$ the unit normal vector to $\partial\Omega$, outward to Ω, and Dirichlet boundary conditions are prescribed to the surface concentrations

$$c_i^s = \tilde{c}_i \text{ on } \Sigma^+,$$

with $\Sigma^+ = \{(x, t) \in \partial\Omega \times \mathbb{R}_+^*, g(x, t) > 0\}$, $\tilde{c}_i \geq 0$ for all $i = 1, \ldots, L$, and $\sum_{i=1}^{L} \tilde{c}_i = 1$.

Initial conditions are prescribed to the sediment thickness such that $h|_{t=0} = h^0$ on Ω, and to the basin concentrations such that $c_i|_{t=0} = c_i^0$ on the domain $\{(x,z), x \in \Omega, z < h^0(x)\}$, with $c_i^0 \geq 0$ for all $i = 1,\dots,L$, and $\sum_{i=1}^L c_i^0 = 1$.

In the following, we shall consider the new coordinate system for which the vertical position of a point in the basin is measured downward from the top of the basin, i.e., given by the change of variable $(x,\xi,t) = (x', h(x',t') - z, t')$. In this coordinate system, let $u_i(x,\xi,t) = c_i(x, h(x,t) - \xi, t)$ on $\Omega \times \mathbb{R}_+^* \times \mathbb{R}_+^*$ and $u_i^0(x,\xi) = c_i^0(x, h^0(x) - \xi, t)$ on $\Omega \times \mathbb{R}_+^*$. Gathering all the equations, we obtain the following multilithology diffusive model:

$$\text{(2.4)} \qquad \text{surface conservations: } \begin{cases} u_i|_{\xi=0}\,\partial_t h + \operatorname{div}(-c_i^s \nabla h) = 0 & \text{on } \mathcal{D}, \\ \sum_{i=1}^L c_i^s = 1 & \text{on } \mathcal{D}, \\ \nabla h \cdot \vec{n}|_{\partial\Omega\times\mathbb{R}_+^*} = g & \text{on } \partial\Omega \times \mathbb{R}_+^*, \\ c_i^s|_{\Sigma^+} = \tilde{c}_i & \text{on } \Sigma^+, \\ h|_{t=0} = h^0 & \text{on } \Omega, \end{cases}$$

$$\text{(2.5)} \qquad \text{column conservations: } \begin{cases} \partial_t u_i + \partial_t h\, \partial_\xi u_i = 0 & \text{on } \Omega \times \mathbb{R}_+^* \times \mathbb{R}_+^*, \\ u_i|_{\xi=0} = c_i^s & \text{on } \mathcal{D}^+, \\ u_i^0|_{t=0} = u_i^0 & \text{on } \Omega \times \mathbb{R}_+^*, \end{cases}$$

where we have taken into account the equality $\partial_t \mathcal{M}_i = u_i|_{\xi=0}\,\partial_t h$ on $\mathcal{D}$ which derives formally from the definition (2.2) and the equation $\partial_t c_i = 0$ on $\mathcal{B}$.

For this simplified model, summing (2.4) over $i = 1,\dots,L$, it appears that the variable h satisfies the parabolic equation

$$\text{(2.6)} \qquad \begin{cases} \partial_t h - \Delta h = 0 & \text{on } \Omega \times \mathbb{R}_+^*, \\ \nabla h \cdot \vec{n}|_{\partial\Omega\times\mathbb{R}_+^*} = g & \text{on } \partial\Omega \times \mathbb{R}_+^*, \\ h_{|t=0} = h^0 & \text{on } \Omega, \end{cases}$$

while the remaining concentration variables (c_i^s, u_i) verify, for each $i = 1,\dots,L$, the system of equations

$$\text{(2.7)} \qquad \begin{cases} u_i|_{\xi=0}\,\partial_t h + \operatorname{div}(-c_i^s \nabla h) = 0 & \text{on } \mathcal{D}, \\ c_i^s|_{\Sigma_+} = \tilde{c}_i & \text{on } \Sigma^+, \\ \partial_t u_i + \partial_t h\, \partial_\xi u_i = 0 & \text{on } \Omega \times \mathbb{R}_+^* \times \mathbb{R}_+^*, \\ u_i|_{\xi=0} = c_i^s & \text{on } \mathcal{D}^+, \\ u_i|_{t=0} = u_i^0 & \text{on } \Omega \times \mathbb{R}_+^*. \end{cases}$$

The sediment thickness variable is decoupled from the concentrations variables and satisfies the linear system (2.6). The solution of this system is then used in problem (2.7), which is linear with respect to the variables c_i^s and u_i.

In what follows, the following assumptions are made on the data.

HYPOTHESIS 1.

(i) *Ω is an open bounded subset of $\mathbb{R}^d$, of class $\mathcal{C}^\infty$,*

(ii) *$h^0 \in \mathcal{C}^2(\bar{\Omega})$,*

(iii) *$g \in \mathcal{C}^1(\partial\Omega \times \mathbb{R}_+) \cap L^2(\partial\Omega \times \mathbb{R}_+)$,*

(iv) *g and h^0 are chosen according to the assumptions of Theorem* 5.3 *of* [11, p. 320] *so that the unique solution h of* (2.6) *is in $C^2(\bar{\Omega} \times [0,T])$ for all $T > 0$,*

(v) *$\tilde{c}_i \in L^\infty(\Sigma^+)$ with $\tilde{c}_i \geq 0$ for $i = 1,\dots,L$, and $\sum_{i=1}^L \tilde{c}_i = 1$,*

(vi) *$u_i^0 \in L^\infty(\Omega \times \mathbb{R}_+^*)$, $u_i^0 \geq 0$ for $i = 1,\dots,L$, and $\sum_{i=1}^L u_i^0 = 1$.*

In the following, we shall denote by $\mathcal{C}_c^\infty(\mathbb{R}^n)$ the space of real valued functions

$$\{\varphi \in \mathcal{C}^\infty(\mathbb{R}^n) \,|\, \mathrm{supp}(\varphi) \text{ bounded in } \mathbb{R}^n\}.$$

To obtain a rigorous mathematical formulation of (2.7), we are looking for weak solutions defined as follows for all $i = 1, \dots, L$.

DEFINITION 2.1. *Let us assume that Hypothesis* 1 *holds and let h denote the solution of problem* (2.6)*. Then $(c_i^s, u_i) \in L^\infty(\Omega \times \mathbb{R}_+^*) \times L^\infty(\Omega \times \mathbb{R}_+^* \times \mathbb{R}_+^*)$ is said to be a weak solution of* (2.7) *if it satisfies* (i) *for all $\varphi \in \mathcal{A} = \{v \in \mathcal{C}_c^\infty(\mathbb{R}^{d+2}) \,|\, v(.,0,.) = 0$ on $\mathcal{D} \setminus S^+\}$*

$$\begin{aligned}(2.8)\qquad &\int_\Omega \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} \left[\partial_t \varphi(x,\xi,t) + \partial_t h(x,t)\, \partial_\xi \varphi(x,\xi,t)\right] u_i(x,\xi,t)\, dt\, d\xi\, dx \\ &+ \int_\Omega \int_{\mathbb{R}_+} u_i^0(x,\xi)\varphi(x,\xi,0)\, d\xi\, dx + \int_\Omega \int_{\mathbb{R}_+} \partial_t h(x,t) c_i^s(x,t)\varphi(x,0,t)\, dt\, dx = 0,\end{aligned}$$

(ii) *for all $\psi \in \mathcal{A}_0 = \{v \in \mathcal{C}_c^\infty(\mathbb{R}^{d+2}) \,|\, v(.,0,.) = 0$ on $\partial\Omega \times \mathbb{R}_+^* \setminus \Sigma^+\}$*

$$\begin{aligned}(2.9)\qquad &-\int_\Omega \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} \left[\partial_t \psi(x,\xi,t) + \partial_t h(x,t)\, \partial_\xi \psi(x,\xi,t)\right] u_i(x,\xi,t)\, dt\, d\xi\, dx \\ &- \int_\Omega \int_{\mathbb{R}_+} u_i^0(x,\xi)\psi(x,\xi,0)\, d\xi\, dx + \int_{\mathbb{R}_+} \left(\int_\Omega c_i^s(x,t)\, \nabla h(x,t) \cdot \nabla \psi(x,0,t)\, dx \right. \\ &\left. - \int_{\partial\Omega} \tilde{c}_i(x,t) g(x,t) \psi(x,0,t) d\gamma(x) \right) dt = 0.\end{aligned}$$

3. Finite volume scheme. The system (2.4)–(2.5) is discretized by a fully implicit time integration and a finite volume method with cell centered variables. We shall consider in what follows admissible meshes according to the following definition.

DEFINITION 3.1 (admissible meshes). *Let Ω be a bounded domain of $\mathbb{R}^d$, $d = 1$ or* 2*. In the following, $m(.)$ will be used to denote a measure on R^d equal to the Lebesgue measure if $d \geq 1$, and, if $d = 0$, the measure of a point is set to one and the measure of the empty set to zero. An admissible finite volume mesh of Ω for the discretization of problem* (2.4)–(2.5) *is given by a family of "control volumes," denoted by $\mathcal{K}$, which are open disjoint subsets of Ω, and a family of points of Ω, denoted by $\mathcal{P}$, satisfying the following properties:*

(i) *The closure of the union of all the control volumes of $\mathcal{K}$ is $\bar{\Omega}$.*

(ii) *For any κ, $\kappa' \in \mathcal{K}$ with $\kappa \neq \kappa'$, either the $(d-1)$-dimensional measure $m(\bar{\kappa} \cap \bar{\kappa}')$ is null, or it is strictly positive and $\bar{\kappa} \cap \bar{\kappa}'$ is included in a hyperplane of $\mathbb{R}^d$. In the following, we will denote by Σ_{int} the family of subsets σ of Ω contained in hyperplanes of $\mathbb{R}^d$ with strictly positive measures, and such that there exist $\kappa, \kappa' \in \mathcal{K}$ with $m(\bar{\kappa} \cap \bar{\kappa}') > 0$ and $\bar{\sigma} = \bar{\kappa} \cap \bar{\kappa}'$. We shall also denote by $\kappa|\kappa' \in \Sigma_{int}$ the edge between the cells κ and κ'.*

(iii) *The family $\mathcal{P} = (x_\kappa)_{\kappa \in \mathcal{K}}$ is such that $x_\kappa \in \bar{\kappa}$ (for any $\kappa \in \mathcal{K}$), and, if $\sigma = \kappa|\kappa'$, it is assumed that $x_\kappa \neq x_{\kappa'}$ and that the straight line going through x_κ and $x_{\kappa'}$ is orthogonal to the edge $\kappa|\kappa'$.*

(iv) *For any $\kappa \in \mathcal{K}$, there exists a subset Σ_κ of Σ_{int} such that $\partial\kappa \setminus \partial\Omega = \bar{\kappa} \setminus (\kappa \cup \partial\Omega) = \cup_{\sigma \in \Sigma_\kappa} \bar{\sigma}$.*

We shall denote by $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ this admissible mesh.

Let $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ be an admissible mesh of Ω in the sense of Definition 3.1. In what follows, $\delta\mathcal{K} = \sup\{\mathrm{diam}(\kappa), \kappa \in \mathcal{K}\}$ will denote the mesh size of $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$, $|\kappa|$ (resp., $|\sigma|$, $|\partial\kappa \cap \partial\Omega|$) is the d-dimensional measure of the cell $m(\kappa)$ (resp., the $(d-1)$-dimensional measure $m(\sigma)$, $m(\partial\kappa \cap \partial\Omega)$), $\mathcal{K}_\kappa$ the set of neighboring cells of κ (excluding κ), $T_{\kappa\kappa'} = T_\sigma$ the transmissibility of the edge $\sigma = \kappa|\kappa'$, defined by $T_{\kappa\kappa'} := \frac{|\sigma|}{d(\kappa,\kappa')}$ with $d(\kappa,\kappa')$ the distance between the points x_κ and $x_{\kappa'}$, $\mathrm{reg}(\mathcal{K})$ the geometrical factor defined by $\mathrm{reg}(\mathcal{K}) = \max_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} \frac{\delta\mathcal{K}}{d(\kappa,\kappa')}$, and $\vec{n}_{\kappa\kappa'}$ the unit normal vector to $\sigma = \kappa|\kappa'$ outward to κ.

We shall also denote by $X(\mathcal{K})$ the set of real valued functions on Ω which are constant over each control volume of the mesh and, for any subset $\mathcal{O}$ of $\mathbb{R}^d$, by $\chi_{\mathcal{O}}$ the function on $\mathbb{R}^d$ equal to one on $\mathcal{O}$ and null elsewhere. Finally, for any function f, let us define $f^+ = \max(f,0) \geq 0$, $f^- = -\min(f,0) \geq 0$, such that $f = f^+ - f^-$, and $|f| = f^+ + f^-$.

Following [6], we shall use the discrete seminorm defined as follows.

Definition 3.2 (discrete $\mathcal{H}^1$ seminorm). *Let Ω be an open bounded subset of $\mathbb{R}^d$, $d=1$ or 2, and $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ be an admissible finite volume mesh of Ω in the sense of Definition* 3.1. *For $u \in X(\mathcal{K})$, the discrete $\mathcal{H}^1$ seminorm of u is defined by*

$$|u|_{1,\mathcal{K}} = \left(\sum_{\sigma\in\Sigma_{int}} T_\sigma (D_\sigma u)^2 \right)^{\frac{1}{2}},$$

where u_κ is the value of u in the control volume κ and $D_\sigma u = |u_\kappa - u_{\kappa'}|$ with $\sigma = \kappa|\kappa'$.

Remark 1. Let $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ be an admissible mesh of Ω in the sense of Definition 3.1 and $|\Omega|$ denote the d-dimensional measure of the domain Ω. Considering the d-dimensional measure of the set of cones of vertex x_κ and base $\sigma \in \Sigma_{int} \cap \partial\kappa$ for all $\kappa \in \mathcal{K}$ and $\sigma \in \Sigma_{int}$, one can prove that

$$\sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} |\sigma|\, d(\kappa,\kappa') \leq d\,|\Omega|. \tag{3.1}$$

The time discretization is denoted by $t^n, n \in \mathbb{N}$, such that $t^0 = 0$ and $\Delta t^{n+1} = t^{n+1} - t^n > 0$. In the following, the superscript n, $n \in \mathbb{N}$, will be used to denote that the variables are considered at time t^n. Assuming that the set $\{\Delta t^n \,|\, n \in \mathbb{N}\}$ is bounded, let Δt denote $\sup\{\Delta t^n \,|\, n \in \mathbb{N}\}$, and, for a given $T > 0$, let $N_{\Delta t}$ be the integer such that $t^{N_{\Delta t}} < T \leq t^{N_{\Delta t}+1}$.

Let us now recall the discretization of (2.4)–(2.5) already introduced in [5]. For all control volumes $\kappa \in \mathcal{K}$, the following initial values are defined:

1. h^0_κ is the initial approximation of h in κ defined by $h^0_\kappa = h^0(x_\kappa)$.
2. $u^0_{i,\kappa}$, for all species i, is the approximation of u^0_i on the cell κ, defined by $u^0_{i,\kappa}(\xi) = \frac{1}{|\kappa|}\int_\kappa u^0_i(x,\xi)\,dx$ for $\xi \in \mathbb{R}^*_+$, and let $c^0_{i,\kappa}$ be defined on $(-\infty, h^0_\kappa)$ by $c^0_{i,\kappa}(z) = u^0_{i,\kappa}(h^0_\kappa - z)$.

We now give a discretization of (2.4)–(2.5) within a given control volume $\kappa \in \mathcal{K}$ between times t^n and t^{n+1}.

Conservation of surface sediments:

$$\begin{aligned} &\frac{\Delta\mathcal{M}^{n+1}_{i,\kappa}}{\Delta t^{n+1}}\, |\kappa| \;+ \sum_{\kappa'\in\mathcal{K}_\kappa} c^{s,n+1}_{i,\kappa\kappa'}\, T_{\kappa\kappa'}(h^{n+1}_\kappa - h^{n+1}_{\kappa'}) \\ &\quad -|\partial\kappa\cap\partial\Omega|\,\tilde{c}^{n+1}_{i,\kappa}\, g^{(+),n+1}_\kappa + |\partial\kappa\cap\partial\Omega|\, c^{s,n+1}_{i,\kappa}\, g^{(-),n+1}_\kappa = 0, \end{aligned} \tag{3.2}$$

$$\sum_{i=1}^{L} c_{i,\kappa}^{s,n+1} = 1. \tag{3.3}$$

Conservation of column sediments:

$$\text{if } h_\kappa^{n+1} \geq h_\kappa^n \begin{cases} \Delta\mathcal{M}_{i,\kappa}^{n+1} = c_{i,\kappa}^{s,n+1}(h_\kappa^{n+1} - h_\kappa^n), \\ c_{i,\kappa}^{n+1}(z) = c_{i,\kappa}^n(z),\ z < h_\kappa^n, \\ c_{i,\kappa}^{n+1}(z) = c_{i,\kappa}^{s,n+1},\ z \in (h_\kappa^n, h_\kappa^{n+1}), \end{cases} \tag{3.4}$$

$$\text{else} \quad \begin{cases} \Delta\mathcal{M}_{i,\kappa}^{n+1} = \int_{h_\kappa^n}^{h_\kappa^{n+1}} c_{i,\kappa}^n(z)dz, \\ c_{i,\kappa}^{n+1}(z) = c_{i,\kappa}^n(z),\ z < h_\kappa^{n+1}. \end{cases} \tag{3.5}$$

In (3.2)–(3.5), the following notation is used.

1. h_κ^n is the approximation of the sediment thickness h at time t^n in κ.

2. $c_{i,\kappa}^{s,n+1}$ is the approximation of the surface sediment concentration i at time t^{n+1} in κ.

3. The function $c_{i,\kappa}^n$, defined on the column $(-\infty, h_\kappa^n)$. is the approximation of the sediment concentration in lithology i in the column $\{(x,z), x \in \kappa, z < h(x,t^n)\}$ at time t^n.

4. $c_{i,\kappa\kappa'}^{s,n+1}$ is the upstream weighted evaluation of the surface sediment concentration in lithology i at the edge σ between the cells κ and κ' with respect to the sign of $h_\kappa^{n+1} - h_{\kappa'}^{n+1}$:

$$c_{i,\kappa\kappa'}^{s,n+1} = \begin{cases} c_{i,\kappa}^{s,n+1} \text{ if } h_\kappa^{n+1} > h_{\kappa'}^{n+1}, \\ c_{i,\kappa'}^{s,n+1} \text{ otherwise.} \end{cases}$$

5. $g_\kappa^{(+),n+1}$ and $g_\kappa^{(-),n+1}$ are the following approximations of the boundary fluxes g^+ and g^-:

$$g_\kappa^{(+),n+1} = \begin{cases} \frac{1}{\Delta t^{n+1}} \frac{1}{|\partial\kappa\cap\partial\Omega|} \int_{t^n}^{t^{n+1}} \int_{\partial\kappa\cap\partial\Omega} g^+(x,t)\, d\gamma(x)dt & \text{if } |\partial\kappa \cap \partial\Omega| \neq 0, \\ 0 & \text{else,} \end{cases}$$

$$g_\kappa^{(-),n+1} = \begin{cases} \frac{1}{\Delta t^{n+1}} \frac{1}{|\partial\kappa\cap\partial\Omega|} \int_{t^n}^{t^{n+1}} \int_{\partial\kappa\cap\partial\Omega} g^-(x,t)\, d\gamma(x)dt & \text{if } |\partial\kappa \cap \partial\Omega| \neq 0, \\ 0 & \text{else,} \end{cases}$$

and consequently for all $\kappa \in \mathcal{K}$,

$$g_\kappa^{n+1} = \frac{1}{\Delta t^{n+1}} \frac{1}{|\partial\kappa \cap \partial\Omega|} \int_{t^n}^{t^{n+1}} \int_{\partial\kappa\cap\partial\Omega} g(x,t)\, d\gamma(x)dt = g_\kappa^{(+),n+1} - g_\kappa^{(-),n+1}.$$

6. $\tilde{c}_{i,\kappa}^{n+1}$ is the approximation of $\tilde{c}_i$ extended by 0 on $(\partial\Omega \times \mathbb{R}_+^*) \setminus \Sigma^+$:

$$\tilde{c}_{i,\kappa}^{n+1} = \begin{cases} \frac{1}{\Delta t^{n+1}} \frac{1}{|\partial\kappa\cap\partial\Omega|} \int_{t^n}^{t^{n+1}} \int_{\partial\kappa\cap\partial\Omega} \tilde{c}_i(x,t) d\gamma(x)dt & \text{if } |\partial\kappa \cap \partial\Omega| \neq 0, \\ 0 & \text{else,} \end{cases}$$

and it results that $\tilde{c}_{i,\kappa}^{n+1} \in [0,1]$.

Considering the coordinate system $\xi = h_\kappa^n - z$, the function $u_{i,\kappa}^n$ is defined for all $\kappa \in \mathcal{K}$, $n \geq 0$, and $i = 1, \ldots, L$ by

$$u_{i,\kappa}^n(\xi) = c_{i,\kappa}^n(h_\kappa^n - \xi) \text{ for all } \xi \in \mathbb{R}_+^*. \tag{3.6}$$

Let us note that, to obtain a fully discrete scheme, the initial condition $u^0_{i,\kappa}(\xi)$ is projected for each κ on a piecewise constant finite element subspace of $L^\infty(\mathbb{R}^*_+)$. Then, the scheme (3.4)–(3.5) generates a piecewise constant approximation of $u^n_{i,\kappa}(\xi)$ on each cell κ for all $i = 1, \dots, L$, with time-dependent mesh sizes in the direction ξ.

For the sake of simplicity, it is assumed in the remainder of this article that $\Delta t = \Delta t^n$ for all $n \geq 1$, although all the results presented in what follows readily extend to variable time steps.

In sections 4 and 5, we shall prove, for all $n \geq 0$, the existence of solutions $(h^n_\kappa)_{\kappa\in\mathcal{K}}$, $(c^{s,n+1}_{i,\kappa})_{\kappa\in\mathcal{K}}$, $(c^n_{i,\kappa})_{\kappa\in\mathcal{K}}$, and $(u^n_{i,\kappa})_{\kappa\in\mathcal{K}}$, $i = 1, \dots, L$, to problem (3.2)–(3.6). These solutions are unique except for the surface concentration $c^{s,n+1}_{i,\kappa}$ which is arbitrary (such that $\sum_{j=1}^L c^{s,n+1}_{j,\kappa} = 1$) at some degenerate points $(\kappa, n+1)$ for which it is chosen according to Lemma 5.1.

For any admissible mesh $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ of Ω in the sense of Definition 3.1, any time step $\Delta t > 0$, and $i = 1, \dots, L$, let $h_{\mathcal{K},\Delta t}$, $c^s_{i,\mathcal{K},\Delta t}$ defined on $\Omega \times \mathbb{R}^*_+$ and $u_{i,\mathcal{K},\Delta t}$ defined on $\Omega \times \mathbb{R}^*_+ \times \mathbb{R}^*_+$ denote the functions such that

$$\left\{ \begin{aligned} h_{\mathcal{K},\Delta t}(x,t) &= h^{n+1}_\kappa, \\ u_{i,\mathcal{K},\Delta t}(x,\xi,t) &= u^{n+1}_{i,\kappa}(\xi), \\ c^s_{i,\mathcal{K},\Delta t}(x,t) &= c^{s,n+1}_{i,\kappa} \end{aligned} \right. \tag{3.7}$$

for all $x \in \kappa$, $\kappa \in \mathcal{K}$, $t \in (t^n, t^{n+1}]$, $\xi \in \mathbb{R}^*_+$, $n \geq 0$, where h^n_κ, $c^{s,n+1}_{i,\kappa}$, $c^n_{i,\kappa}$ are any given solution of (3.2)–(3.6) chosen according to Lemma 5.1. From Lemma 5.1, the functions $h_{\mathcal{K},\Delta t}$ and $u_{i,\mathcal{K},\Delta t}$ do not depend on the choice of the solution of (3.2)–(3.6).

The aim of this article is then to prove the following theorem.

THEOREM 3.3. *Hypothesis* 1 *is assumed to hold. For all* $m \in \mathbb{N}$, *let* $(\mathcal{K}_m, \Sigma^m_{int}, \mathcal{P}_m)$ *be an admissible mesh of* Ω *in the sense of Definition* 3.1 *and* $\Delta t_m > 0$. *Let us assume that there exists* $\alpha > 0$ *such that* $reg(\mathcal{K}_m) \leq \alpha$ *for all* $m \in \mathbb{N}$, *and that* $\Delta t_m \to 0$, $\frac{\delta \mathcal{K}_m}{\sqrt{\Delta t_m}} \to 0$ *as* $m \to \infty$.

For all $m \in \mathbb{N}$ *and* $i = 1, \dots, L$, *let* $h_{\mathcal{K}_m,\Delta t_m}$, $u_{i,\mathcal{K}_m,\Delta t_m}$ *denote the unique functions defined by* (3.7) *and* $c^s_{i,\mathcal{K}_m,\Delta t_m}$ *be a function defined by* (3.7), *from any solution of* (3.2)–(3.6) *chosen according to Lemma* 5.1 *with* $\mathcal{K} = \mathcal{K}_m$, $\Delta t = \Delta t_m$.

Then, the sequence $(h_{\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ *converges to the solution* h *of problem* (2.6) *in* $L^\infty(0,T;L^2(\Omega))$ *for all* $T > 0$, *and there exists a subsequence of* $(\mathcal{K}_m, \Delta t_m)_{m\in\mathbb{N}}$, *still denoted by* $(\mathcal{K}_m, \Delta t_m)_{m\in\mathbb{N}}$, *such that, for all* $i \in \{1, \dots, L\}$, *the subsequence* $(c^s_{i,\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ (*resp.*, $(u_{i,\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$) *converges to a function* c^s_i *in* $L^\infty(\Omega \times \mathbb{R}^*_+)$ (*resp.*, u_i *in* $L^\infty(\Omega \times \mathbb{R}^*_+ \times \mathbb{R}^*_+)$) *for the weak-$\star$ topology. Furthermore, for all* $i \in \{1, \dots, L\}$, *the limit* (c^s_i, u_i) *is a weak solution of problem* (2.7) *in the sense of Definition* 2.1.

This convergence result will be obtained in section 4 for the approximate solution for the sediment thickness and in section 5 for the approximate concentrations.

4. Stability and convergence for the approximate sediment thickness and its time derivative. Summing (3.2) over $i = 1, \dots, L$ yields that for all $n \in \mathbb{N}$, the solution $(h^{n+1}_\kappa)_{\kappa\in\mathcal{K}}$ satisfies the following implicit finite volume discretization of (2.6):

$$|\kappa| \frac{h^{n+1}_\kappa - h^n_\kappa}{\Delta t} + \sum_{\kappa' \in \mathcal{K}_\kappa} T_{\kappa\kappa'}(h^{n+1}_\kappa - h^{n+1}_{\kappa'}) - |\partial\kappa \cap \partial\Omega|\, g^{n+1}_\kappa = 0, \tag{4.1}$$

with $h^0_\kappa = h^0(x_\kappa)$. The proof of existence and uniqueness of the solution $(h^n_\kappa)_{\kappa\in\mathcal{K}}$ for all $n \geq 0$ is classical and can be found, e.g., in [6] for any admissible mesh $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ of Ω.

The following proposition provides estimates of the error on h and its time derivative. The error estimates on h have already been proved in [6].

PROPOSITION 4.1. *Let us assume that Hypothesis* 1 *holds and let h denote the solution of problem* (2.6). *Let $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ be an admissible mesh of Ω in the sense of Definition* 3.1, *$T > 0$, and $\Delta t \in (0, T)$. For all $n \in \{0, \dots, N_{\Delta t} + 1\}$, let $(h^n_\kappa)_{\kappa\in\mathcal{K}}$ be the solution of* (4.1) *and $e^n_{\mathcal{K}} \in X(\mathcal{K})$ be defined by $e^n_{\mathcal{K}}(x) = e^n_\kappa = h(x_\kappa, t^n) - h^n_\kappa$ for all $x \in \kappa$, $\kappa \in \mathcal{K}$. Then, there exist D_1, D_2, D_3, and $D_4 > 0$ depending only on $\|\nabla\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, T, and Ω such that*

$$\|e^n_{\mathcal{K}}\|^2_{L^2(\Omega)} \leq D_1(\Delta t + \delta\mathcal{K})^2 \quad \textit{for all } n \in \{1, \dots, N_{\Delta t} + 1\}, \tag{4.2}$$

$$\sum_{n=0}^{N_{\Delta t}} \Delta t\, |e^{n+1}_{\mathcal{K}}|^2_{1,\mathcal{K}} \leq D_2\, (\Delta t + \delta\mathcal{K})^2, \tag{4.3}$$

$$\sum_{n=0}^{N_{\Delta t}} \Delta t \left\| \frac{e^{n+1}_{\mathcal{K}} - e^n_{\mathcal{K}}}{\Delta t} \right\|^2_{L^2(\Omega)} \leq D_3 \frac{(\delta\mathcal{K} + \Delta t)^2}{\Delta t}, \tag{4.4}$$

$$\begin{aligned} &\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int} \\ \sigma=\kappa|\kappa'}} |\sigma|\, d(\kappa,\kappa') \bigg(\frac{h^{n+1}_{\kappa'} - h^{n+1}_\kappa}{d(\kappa,\kappa')} \\ &- \frac{1}{\Delta t}\frac{1}{|\sigma|} \int_{t^n}^{t^{n+1}} \int_\sigma \nabla h(x,t)\cdot \vec{n}_{\kappa\kappa'} d\gamma(x)\, dt \bigg)^2 \leq D_4(\Delta t + \delta\mathcal{K})^2. \end{aligned} \tag{4.5}$$

Proof. Integrating (2.6) over the control volume $\kappa \in \mathcal{K}$ and time interval (t^n, t^{n+1}) for all $n \in \{0, \dots, N_{\Delta t}\}$, one obtains

$$\int_{t^n}^{t^{n+1}} \int_\kappa \partial_t h(x,t) dx dt - \int_{t^n}^{t^{n+1}} \int_{\partial\kappa} \nabla h(x,t)\cdot \vec{n}_\kappa d\gamma(x) dt = 0, \tag{4.6}$$

where $\vec{n}_\kappa$ is the normal unit vector to $\partial\kappa$ outward to κ. Subtracting (4.1) from (4.6)/Δt and using the definition of g^{n+1}_κ yield the following equation for the error e^{n+1}_k:

$$|\kappa| \frac{e^{n+1}_\kappa - e^n_\kappa}{\Delta t} + \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}(e^{n+1}_\kappa - e^{n+1}_{\kappa'}) = -|\kappa| P^n_\kappa - \sum_{\sigma\in\Sigma_\kappa} |\sigma| R^n_{\kappa,\sigma} \tag{4.7}$$

with the consistency residuals

$$R^n_{\kappa,\sigma} = \frac{1}{\Delta t}\frac{1}{|\sigma|} \int_{t^n}^{t^{n+1}} \int_\sigma \left[\frac{h(x_{\kappa'}, t^{n+1}) - h(x_\kappa, t^{n+1})}{d(\kappa,\kappa')} - \nabla h(x,t)\cdot\vec{n}_{\kappa\kappa'} \right] d\gamma(x) dt$$

for all $\kappa \in \mathcal{K}$ and $\sigma \in \Sigma_\kappa \cap \Sigma_{\kappa'}$, and

$$P^n_\kappa = \frac{1}{\Delta t}\frac{1}{|\kappa|} \int_{t^n}^{t^{n+1}} \int_\kappa (\partial_t h(x,t) - \partial_t h(x_\kappa, t))\, dx dt \qquad \text{for all } \kappa \in \mathcal{K}.$$

Thanks to the regularity of h, there exists $C_1 > 0$ depending on $\|\nabla\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$ only such that

$$|P^n_\kappa| \leq C_1\, \delta\mathcal{K}, \tag{4.8}$$

and $C_2 > 0$ depending only on $\|\partial_t \nabla h\|_{L^\infty(\Omega\times(0,2T))}$, and $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$ such that

$$|R^n_{\kappa,\sigma}| \le C_2\,(\delta\mathcal{K} + \Delta t). \tag{4.9}$$

Then, multiplying (4.7) by e^{n+1}_κ and summing over the cells $\kappa \in \mathcal{K}$ yield the estimate

$$\begin{aligned}\sum_{\kappa\in\mathcal{K}} |\kappa|(e^{n+1}_\kappa - e^n_\kappa)e^{n+1}_\kappa + \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}(e^{n+1}_\kappa - e^{n+1}_{\kappa'})^2 \\ = -\Delta t \sum_{\kappa\in\mathcal{K}} |\kappa| P^n_\kappa\, e^{n+1}_\kappa - \Delta t \sum_{\kappa\in\mathcal{K}}\sum_{\sigma\in\Sigma_\kappa} |\sigma| R^n_{\kappa,\sigma}\, e^{n+1}_\kappa.\end{aligned} \tag{4.10}$$

Let us note that $R_{\kappa,\sigma} = -R_{\kappa',\sigma}$ for all $\sigma = \kappa|\kappa' \in \Sigma_{int}$ so that $R_\sigma = |R_{\kappa,\sigma}|$ for $\sigma \in \Sigma_\kappa$ can be defined for all $\sigma \in \Sigma_{int}$. Then, using in (4.10) the equality $(e^{n+1}_\kappa - e^n_\kappa)\, e^{n+1}_\kappa = \frac{1}{2}\big[(e^{n+1}_\kappa)^2 - (e^n_\kappa)^2 + (e^{n+1}_\kappa - e^n_\kappa)^2\big]$, Young's inequality, (3.1), (4.8), and (4.9), we obtain

$$\begin{aligned}\|e^{n+1}_{\mathcal{K}}\|^2_{L^2(\Omega)} + \Delta t\, |e^{n+1}_{\mathcal{K}}|^2_{1,\mathcal{K}} \le \|e^n_{\mathcal{K}}\|^2_{L^2(\Omega)} \\ +\Delta t\, C_3\,(\Delta t + \delta\mathcal{K})\, \|e^{n+1}_{\mathcal{K}}\|^2_{L^2(\Omega)} + \Delta t\, C_4\,(\delta\mathcal{K} + \Delta t)^2,\end{aligned} \tag{4.11}$$

with C_3 and C_4 depending only on $\|\nabla\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, and Ω. Using the same arguments as in [6], the estimate (4.2) derives from (4.11). Summing (4.11) over $n \in \{0,\dots,N_{\Delta t}\}$ and using inequality (4.2) and the property $e^0_\kappa = 0$ for all $\kappa \in \mathcal{K}$, we obtain inequality (4.3).

Then, (4.3) is equivalent to

$$\begin{aligned}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} |\sigma|\, d(\kappa,\kappa')\left(\frac{h^{n+1}_{\kappa'} - h^{n+1}_\kappa}{d(\kappa,\kappa')} - \frac{h(x_{\kappa'},t^{n+1}) - h(x_\kappa,t^{n+1})}{d(\kappa,\kappa')}\right)^2 \\ \le D_2\,(\Delta t + \delta\mathcal{K})^2.\end{aligned} \tag{4.12}$$

Furthermore,

$$\begin{aligned}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} |\sigma|\, d(\kappa,\kappa')\bigg(\frac{h(x_{\kappa'},t^{n+1}) - h(x_\kappa,t^{n+1})}{d(\kappa,\kappa')} \\ -\frac{1}{\Delta t}\frac{1}{|\sigma|}\int_{t^n}^{t^{n+1}}\int_\sigma \nabla h(x,t)\cdot\vec{n}_{\kappa\kappa'}\,d\gamma(x)dt\bigg)^2 \\ = \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} |\sigma|\, d(\kappa,\kappa')(R^n_\sigma)^2 \;\le\; C_5\,(\delta\mathcal{K} + \Delta t)^2,\end{aligned} \tag{4.13}$$

with C_5 depending on $\|\nabla\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, T, and Ω. The estimate (4.5) derives from (4.12) and (4.13).

To prove (4.4), let us multiply (4.7) by $(e^{n+1}_\kappa - e^n_\kappa)/\Delta t$ and sum over $\kappa \in \mathcal{K}$:

$$\begin{aligned}\Delta t \sum_{\kappa\in\mathcal{K}} |\kappa| \left(\frac{e^{n+1}_\kappa - e^n_\kappa}{\Delta t}\right)^2 + \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}(e^{n+1}_\kappa - e^{n+1}_{\kappa'})(e^{n+1}_\kappa - e^{n+1}_{\kappa'} - e^n_\kappa + e^n_{\kappa'}) \\ = -\Delta t \sum_{\kappa\in\mathcal{K}} |\kappa|\, P^n_\kappa\, \frac{e^{n+1}_\kappa - e^n_\kappa}{\Delta t} - \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} |\sigma| R^n_{\kappa,\sigma}(e^{n+1}_\kappa - e^{n+1}_{\kappa'} - e^n_\kappa + e^n_{\kappa'}).\end{aligned}$$

From $(e_\kappa^{n+1}-e_{\kappa'}^{n+1})(e_\kappa^{n+1}-e_{\kappa'}^{n+1}-e_\kappa^n+e_{\kappa'}^n) = \frac{1}{2}\big[\,(e_\kappa^{n+1}-e_{\kappa'}^{n+1})^2-(e_\kappa^n-e_{\kappa'}^n)^2+(e_\kappa^{n+1}-e_{\kappa'}^{n+1}-e_\kappa^n+e_{\kappa'}^n)^2\,\big]$ and Young's inequality, it results that

$$\begin{aligned}
&2\Delta t\sum_{\kappa\in\mathcal{K}}|\kappa|\left(\frac{e_\kappa^{n+1}-e_\kappa^n}{\Delta t}\right)^2+\sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}}T_{\kappa\kappa'}(e_\kappa^{n+1}-e_{\kappa'}^{n+1}-e_\kappa^n+e_{\kappa'}^n)^2\\
&+\sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}}T_{\kappa\kappa'}(e_\kappa^{n+1}-e_{\kappa'}^{n+1})^2\le\sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}}T_{\kappa\kappa'}(e_\kappa^n-e_{\kappa'}^n)^2\\
&+\Delta t\sum_{\kappa\in\mathcal{K}}|\kappa|\,(P_\kappa^n)^2+\Delta t\sum_{\kappa\in\mathcal{K}}|\kappa|\left(\frac{e_\kappa^{n+1}-e_\kappa^n}{\Delta t}\right)^2\\
&+\sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}}d(\kappa,\kappa')\,|\sigma|\,(R_\sigma^n)^2+\sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}}T_{\kappa\kappa'}(e_\kappa^{n+1}-e_{\kappa'}^{n+1}-e_\kappa^n+e_{\kappa'}^n)^2.
\end{aligned}\tag{4.14}$$

Summing (4.14) for all $n\in\{0,\dots,N_{\Delta t}\}$ and using (4.8), (4.9), (3.1), and the property $e_\kappa^0=0$ for all $\kappa\in\mathcal{K}$, we get

$$\sum_{n=0}^{N_{\Delta t}}\Delta t\sum_{\kappa\in\mathcal{K}}|\kappa|\left(\frac{e_\kappa^{n+1}-e_\kappa^n}{\Delta t}\right)^2\le C_6\,(\delta\mathcal{K})^2+C_7\,\frac{(\Delta t+\delta\mathcal{K})^2}{\Delta t}$$

with C_6 and $C_7>0$ depending only on $\|\nabla\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, Ω, and T, which proves (4.4). ☐

Remark 2. According to (4.4) given in Proposition 4.1, the discrete time derivative of the error tends to zero with the mesh size and time step under an inverse CFL condition. This condition is due to the fact that the finite volume scheme is implicit in time and that few assumptions have been made on the regularity of h. However, it is possible to get rid of this inverse CFL condition by assuming h much more regular. Such a result can be found in [13].

COROLLARY 1. *Let us assume that Hypothesis* 1 *holds, and let h denote the solution of problem* (2.6). *Let $(\mathcal{K},\Sigma_{int},\mathcal{P})$ be an admissible mesh of Ω in the sense of Definition* 3.1, *$T>0$, $\Delta t\in(0,T)$, and let $\beta>0$ be such that $\delta\mathcal{K}\le\beta\sqrt{\Delta t}$. For all $n\in\{0,\dots,N_{\Delta t}+1\}$, let $(h_\kappa^n)_{\kappa\in\mathcal{K}}$ be the solution of* (4.1), *and let us define $h_\mathcal{K}^n\in X(\mathcal{K})$ (resp., $\delta_t h_\mathcal{K}^n\in X(\mathcal{K})$) by $h_\mathcal{K}^n(x)=h_\kappa^n$ (resp., $\delta_t h_\mathcal{K}^n(x)=\frac{h_\kappa^{n+1}-h_\kappa^n}{\Delta t}$) for $x\in\kappa$, $\kappa\in\mathcal{K}$. Then, there exist $D_5>0$ depending only on $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, $\|\nabla\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, Ω, and T and D_6, D_6', $D_6''>0$ depending on $\|\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, $\|\nabla\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, Ω, and T, with D_6 also depending on β, such that*

$$\sum_{n=0}^{N_{\Delta t}}\Delta t\,|h_\mathcal{K}^{n+1}|_{1,\mathcal{K}}^2\le D_5,\tag{4.15}$$

and

$$\sum_{n=0}^{N_{\Delta t}}\Delta t\,\|\delta_t h_\mathcal{K}^n\|_{L^2(\Omega)}^2\le D_6'+D_6''\,\frac{(\delta\mathcal{K}+\Delta t)^2}{\Delta t}\le D_6.\tag{4.16}$$

Proof. The proof is straightforward, using the error estimates (4.3) and (4.4), the regularity of h, and the estimate (3.1). ☐

For any admissible mesh $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ of Ω in the sense of Definition 3.1 and any time step $\Delta t > 0$, let $(h_\kappa^n)_{\kappa\in\mathcal{K}}$ for all $n \geq 0$ be the solution of (4.1), and let $\delta_t h_{\mathcal{K},\Delta t}$ denote the function defined on $\Omega \times \mathbb{R}_+^*$, such that for all $x \in \kappa$, $\kappa \in \mathcal{K}$, $t \in (t^n, t^{n+1}]$, $n \geq 0$,

$$\delta_t h_{\mathcal{K},\Delta t}(x,t) = \frac{h_\kappa^{n+1} - h_\kappa^n}{\Delta t}. \tag{4.17}$$

PROPOSITION 4.2. *Let us assume that Hypothesis* 1 *holds, and let h denote the solution of problem* (2.6)*. Let us consider a family of admissible discretizations $(\mathcal{K}, \Sigma_{int}, \mathcal{P}, \Delta t)$ of $\Omega \times \mathbb{R}_+^*$, with $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ an admissible mesh of Ω in the sense of Definition* 3.1 *and $\Delta t > 0$ a time step. For a given discretization $(\mathcal{K}, \Sigma_{int}, \mathcal{P}, \Delta t)$ of this family, let $h_{\mathcal{K},\Delta t}$ (resp., $\delta_t h_{\mathcal{K},\Delta t}$) be the function defined by* (3.7) *(resp., by* (4.17)*) from the solution of* (4.1)*. Then, for all $T > 0$, $h_{\mathcal{K},\Delta t}$ converges to h in $L^\infty(0,T;L^2(\Omega))$ as Δt and $\delta\mathcal{K}$ tend to* 0*, and $\delta_t h_{\mathcal{K},\Delta t}$ converges to $\partial_t h$ in $L^2(\Omega \times (0,T))$ as Δt, $\delta\mathcal{K}$ and $\frac{\delta\mathcal{K}}{\sqrt{\Delta t}}$ tend to* 0*.*

Proof. Let $T > 0$, and let $(\mathcal{K}, \Sigma_{int}, \mathcal{P}, \Delta t)$ be an admissible discretization of $\Omega \times \mathbb{R}_+^*$ with $\Delta t < T$. For all $x \in \kappa$, $\kappa \in \mathcal{K}$, and $t \in (t^n, t^{n+1}]$, $n \in \{0, \ldots, N_{\Delta t}\}$, one has

$$\begin{aligned} h(x,t) - h_{\mathcal{K},\Delta t}(x,t) &= (h(x,t) - h(x_\kappa, t^{n+1})) + (h(x_\kappa, t^{n+1}) - h_\kappa^{n+1}) \\ &= (h(x,t) - h(x_\kappa, t^{n+1})) + e_\kappa^{n+1}. \end{aligned}$$

Thus, for all $t \in (t^n, t^{n+1}]$, $n \in \{0, \ldots, N_{\Delta t}\}$,

$$\begin{aligned} &\int_\Omega |h(x,t) - h_{\mathcal{K},\Delta t}(x,t)|^2 dx \\ &\quad \leq 2 \sum_{\kappa\in\mathcal{K}} \left[\int_\kappa |h(x,t) - h(x_\kappa, t^{n+1})|^2 dx + |\kappa| (e_\kappa^{n+1})^2 \right]. \end{aligned} \tag{4.18}$$

Thanks to Proposition 4.1, there exists $C_1 > 0$ depending only on $\|\nabla \partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, and Ω such that

$$\sum_{\kappa\in\mathcal{K}} |\kappa| (e_\kappa^{n+1})^2 \leq C_1 \big(\Delta t + \delta\mathcal{K}\big)^2 \quad \text{for all } n \in \{0, \ldots, N_{\Delta t}\}. \tag{4.19}$$

Furthermore, thanks to the regularity of h, there exists $C_2 > 0$ depending only on $\|\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$ and $\|\nabla h\|_{L^\infty(\Omega\times(0,2T))}$ such that, for all $x \in \kappa$ and $t \in (t^n, t^{n+1}]$,

$$|h(x,t) - h(x_\kappa, t^{n+1})| \leq C_2 \big(\delta\mathcal{K} + \Delta t\big). \tag{4.20}$$

Then, using (4.19) and (4.20) in (4.18) yields, for all $t \in (0,T)$,

$$\|h(.,t) - h_{\mathcal{K},\Delta t}(.,t)\|_{L^2(\Omega)}^2 \leq C_3 \big(\delta\mathcal{K} + \Delta t\big)^2,$$

and consequently, $\|h - h_{\mathcal{K},\Delta t}\|_{L^\infty(0,T;L^2(\Omega))} \leq C_3' \big(\delta\mathcal{K} + \Delta t\big)$, where C_3, C_3' depend on $\|\nabla \partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, $\|\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, and Ω, so that the convergence holds.

Furthermore, for all $x \in \kappa$, $\kappa \in \mathcal{K}$, and $t \in (t^n, t^{n+1}]$, $n \in \{0, \ldots, N_{\Delta t}\}$, one has

$$\partial_t h(x,t) - \delta_t h_{\mathcal{K},\Delta t}(x,t) = \left(\partial_t h(x,t) - \frac{h(x_\kappa, t^{n+1}) - h(x_\kappa, t^n)}{\Delta t} \right) + \frac{e_\kappa^{n+1} - e_\kappa^n}{\Delta t}.$$

Thanks to the regularity of h, there exists a constant $C_4 > 0$ depending only on $\|\partial_t^2 h\|_{L^\infty(\Omega\times(0,2T))}$ and $\|\nabla\partial_t h\|_{L^\infty(\Omega\times(0,2T))}$, such that

$$\left|\frac{h(x_\kappa, t^{n+1}) - h(x_\kappa, t^n)}{\Delta t} - \partial_t h(x,t)\right| \le C_4\big(\delta\mathcal{K} + \Delta t\big),$$

from which, together with (4.4), results

$$\|\partial_t h - \delta_t h_{\mathcal{K},\Delta t}\|^2_{L^2(\Omega\times(0,T))} \le C_5\big(\delta\mathcal{K} + \Delta t\big)^2 + C_6\frac{\big(\delta\mathcal{K} + \Delta t\big)^2}{\Delta t},$$

with C_5 and C_6 depending only on Ω, T, $\|h\|_{W^{2,\infty}(\Omega\times(0,2T))}$. Thus, the convergence of $\delta_t h_{\mathcal{K},\Delta t}$ to $\partial_t h$ in $L^2(\Omega\times(0,T))$ as Δt, $\delta\mathcal{K}$, and $\frac{\delta\mathcal{K}}{\sqrt{\Delta t}} \to 0$ is proved. □

5. Convergence of sequences of approximate concentrations toward a weak solution. We shall first prove the existence of a solution for the concentrations satisfying stability estimates from which the weak-$\star$ convergence, up to a subsequence, of the concentrations in L^∞ is deduced.

Existence, stability, and weak-$\star$ convergence.

LEMMA 5.1. *Let $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ be an admissible mesh of Ω in the sense of Definition 3.1, $\Delta t > 0$, and, for all $n \in \mathbb{N}$, let $(h_\kappa^n)_{\kappa\in\mathcal{K}}$ be the solution of* (4.1)*. For $i \in \{1, \dots, L\}$ and $n \in \mathbb{N}$, there exists a unique solution $(c_{i,\kappa}^n)_{\kappa\in\mathcal{K}}$, and there exists at least one solution $(c_{i,\kappa}^{s,n+1})_{\kappa\in\mathcal{K}}$ to the set of equations* (3.2)*–*(3.5) *such that*

$$c_{i,\kappa}^{s,n+1} \in [0,1] \text{ for all } \kappa \in \mathcal{K} \text{ and } n \in \mathbb{N}. \tag{5.1}$$

Furthermore, one has

$$c_{i,\kappa}^n(z) \in [0,1] \text{ for all } \kappa \in \mathcal{K},\ z < h_\kappa^n, \text{ and } n \in \mathbb{N}.$$

Proof. The complete proof can be found in [5]. It is done by induction over $n \in \mathbb{N}^*$ and over the cells $\kappa \in \mathcal{K}$ sorted by decreasing topographical order. For the highest topographical point(s) κ, the fluxes at the edges of the cell κ are either input boundary fluxes or ouput fluxes. Let us consider a control volume $\kappa \in \mathcal{K}$ and a time $n \in \mathbb{N}^*$, and let us assume that the proposition holds for all the previous times t^{l+1}, $0 \le l < n$, and all the lower cells at time t^{n+1}. It results from the induction hypothesis and the upwinding of c_i^s that $c_{i,\kappa}^{s,n+1}$ can be computed explicitly from the lower cell concentrations c_i^s using (3.2), and that the inequality

$$\sum_{\kappa'\in\mathcal{K}_\kappa,\, h_\kappa^{n+1} < h_{\kappa'}^{n+1}} T_{\kappa\kappa'}\, c_{i,\kappa\kappa'}^{s,n+1}\,(h_\kappa^{n+1} - h_{\kappa'}^{n+1}) \le 0 \tag{5.2}$$

holds for all $i = 1, \dots, L$. Let us first assume that $h_\kappa^{n+1} - h_\kappa^n \le 0$ (erosion). It results from the induction hypothesis that

$$c_{i,\kappa}^{s,n+1}\left(\sum_{\kappa'\in\mathcal{K}_\kappa,\, h_\kappa^{n+1} \ge h_{\kappa'}^{n+1}} T_{\kappa\kappa'}\,(h_\kappa^{n+1} - h_{\kappa'}^{n+1}) + |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(-),n+1}\right) \ge 0$$

for all i. In this equation, either the term into brackets is strictly positive for all $i = 1, \dots, L$ and then $c_{i,\kappa}^{s,n+1} \ge 0$, or it vanishes for all i and the point $(\kappa, n+1)$ is a degenerate point in the sense that all the fluxes at the edges of the control volume κ vanish and $h_\kappa^{n+1} = h_\kappa^n$. The concentrations can in that case be chosen arbitrarily such that $\sum_{i=1}^L c_{i,\kappa}^{s,n+1} = 1$. Let us now consider the sedimentation case for which $h_\kappa^{n+1} - h_\kappa^n > 0$. It results from (3.2) and the induction hypothesis that

$$c_{i,\kappa}^{s,n+1}\left(\frac{h_\kappa^{n+1}-h_\kappa^n}{\Delta t^{n+1}}\,|\kappa| + \sum_{\kappa'\in\mathcal{K}_\kappa,\, h_\kappa^{n+1}\geq h_{\kappa'}^{n+1}} T_{\kappa\kappa'}\,(h_\kappa^{n+1}-h_{\kappa'}^{n+1}) + |\partial\kappa\cap\partial\Omega|\, g_\kappa^{(-),n+1}\right) \geq 0,$$

and hence $c_{i,\kappa}^{s,n+1} \geq 0$ for all $i = 1,\dots,L$. Since $h_\kappa^{n+1} = h_\kappa^n$ for any degenerate point $(\kappa, n+1)$, there exists a unique column concentration $c_{i,\kappa}^{n+1}$ solution of the set of equations (3.2)–(3.5) for each lithology. □

Let us define for all $\kappa \in \mathcal{K}$, $n \in \mathbb{N}$, and $t \in (t^n, t^{n+1}]$ the following interpolation of the discrete sediment thickness:

$$h_\kappa(t) = h_\kappa^n + (t - t^n)\,\frac{h_\kappa^{n+1} - h_\kappa^n}{\Delta t}. \tag{5.3}$$

Then, the discrete solutions $(c_{i,\kappa}^n)_{n\in\mathbb{N}}$, $(u_{i,\kappa}^n)_{n\in\mathbb{N}}$, and $(c_{i,\kappa}^{s,n+1})_{n\in\mathbb{N}}$, given by Lemma 5.1, are extended to $t \in \mathbb{R}_+$ for all $\kappa \in \mathcal{K}$ as follows:

$$\begin{cases} c_{i,\kappa}(z,t) = \begin{cases} c_{i,\kappa}^n(z)\,\chi_{(-\infty,h_\kappa^n]} + c_{i,\kappa}^{s,n+1}\,\chi_{(h_\kappa^n,h_\kappa(t))} & \text{if } h_\kappa^{n+1} \geq h_\kappa^n, \\ c_{i,\kappa}^n(z)\,\chi_{(-\infty,h_\kappa(t))} & \text{otherwise} \end{cases} \\ \qquad\qquad \text{for all } t \in (t^n, t^{n+1}] \text{ and } z < h_\kappa(t), \\ c_{i,\kappa}(z,0) = c_{i,\kappa}^0(z) \text{ for all } z < h_\kappa^0, \end{cases} \tag{5.4}$$

$$u_{i,\kappa}(\xi,t) = c_{i,\kappa}(h_\kappa(t)-\xi,t) \text{ for all } t \geq 0 \text{ and } \xi \in \mathbb{R}_+^*, \tag{5.5}$$

$$c_{i,\kappa}^s(t) = c_{i,\kappa}^{s,n+1} \text{ for all } t \in (t^n, t^{n+1}]. \tag{5.6}$$

For any admissible mesh $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ of Ω in the sense of Definition 3.1 and any time step $\Delta t > 0$, let $\bar{u}_{i,\mathcal{K},\Delta t}$ be defined on $\Omega \times \mathbb{R}_+^* \times \mathbb{R}_+$, and let $c_{i,\mathcal{K},\Delta t}$ be defined on $\{(z,t),\, t \geq 0,\, z < h_\kappa(t)\}$, such that

$$\begin{cases} \bar{u}_{i,\mathcal{K},\Delta t}(x,\xi,t) = u_{i,\kappa}(\xi,t), \\ c_{i,\mathcal{K},\Delta t}(x,z,t) = c_{i,\kappa}(z,t) \end{cases} \tag{5.7}$$

for all $x \in \kappa$, $\kappa \in \mathcal{K}$, $t \geq 0$, $\xi \in \mathbb{R}_+^*$, $z < h_\kappa(t)$.

From Lemma 5.1, the unique functions $c_{i,\mathcal{K},\Delta t}$, $\bar{u}_{i,\mathcal{K},\Delta t}$, $u_{i,\mathcal{K},\Delta t}$ defined by (5.7) and (3.7) and any function $c_{i,\mathcal{K},\Delta t}^s$ defined by (3.7) from any solution of (3.2)–(3.6) chosen according to Lemma 5.1 take their values into the interval $[0,1]$. We deduce the following result.

Proposition 5.2. *For all $m \in \mathbb{N}$, let $(\mathcal{K}_m, \Sigma_{int}^m, \mathcal{P}_m)$ be an admissible mesh of Ω in the sense of Definition* 3.1*, and let $\Delta t_m > 0$. Let us assume that $\Delta t_m \to 0$ and $\delta\mathcal{K}_m \to 0$ as $m \to \infty$.*

For all $m \in \mathbb{N}$ and $i = 1,\dots,L$, let $u_{i,\mathcal{K}_m,\Delta t_m}$ (resp., $\bar{u}_{i,\mathcal{K}_m,\Delta t_m}$) denote the unique function defined by (3.7) *(resp., by* (5.7)*) and $c_{i,\mathcal{K}_m,\Delta t_m}^s$ be a function defined by* (3.7)*, from any solution of* (3.2)*–*(3.6) *chosen according to Lemma* 5.1 *with $\mathcal{K} = \mathcal{K}_m$, $\Delta t = \Delta t_m$.*

Then, under Hypothesis 1*, there exists a subsequence of $(\mathcal{K}_m, \Delta t_m)_{m\in\mathbb{N}}$, still denoted by $(\mathcal{K}_m, \Delta t_m)_{m\in\mathbb{N}}$, such that for all $i \in \{1,\dots,L\}$*

(i) *the subsequence* $(c^s_{i,\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ *converges to a function* c^s_i *in* $L^\infty(\Omega\times\mathbb{R}^*_+)$ *for the weak-*$\star$ *topology, and*

(ii) *the subsequences* $(u_{i,\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ *and* $(\bar{u}_{i,\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ *converge to a function* u_i *in* $L^\infty(\Omega\times\mathbb{R}^*_+\times\mathbb{R}^*_+)$ *for the weak-*$\star$ *topology.*

Proof. For the sake of simplicity, the subscript i is dropped. Thanks to Lemma 5.1, the sequence $(c^s_{\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ (resp., $(u_{\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ and $(\bar{u}_{\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$) is bounded in $L^\infty(\Omega\times\mathbb{R}^*_+)$ (resp., in $L^\infty(\Omega\times\mathbb{R}^*_+\times\mathbb{R}^*_+)$). Then, there exists a subsequence of $(\mathcal{K}_m,\Delta t_m)_{m\in\mathbb{N}}$, still denoted by $(\mathcal{K}_m,\Delta t_m)_{m\in\mathbb{N}}$, such that $(c^s_{\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ (resp., $(u_{\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ and $(\bar{u}_{\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$) converges to c^s (resp., u and u') in $L^\infty(\Omega\times\mathbb{R}^*_+)$ (resp., in $L^\infty(\Omega\times\mathbb{R}^*_+\times\mathbb{R}^*_+)$) for the weak-$\star$ topology. It remains to prove that $u=u'$ in $L^\infty(\Omega\times\mathbb{R}^*_+\times\mathbb{R}^*_+)$.

Using definitions (3.6) and (5.5), for $x\in\kappa$, $\kappa\in\mathcal{K}_m$, and $t\in(t^n,t^{n+1}]$, the functions $\bar{u}_{\mathcal{K}_m,\Delta t_m}$ and $u_{\mathcal{K}_m,\Delta t_m}$ are related as follows:

$$u^{n+1}_\kappa(\xi)=\begin{cases}u_\kappa(\xi-(h_\kappa(t)-h^n_\kappa),t) & \text{for all } \xi\geq h_\kappa(t)-h^n_\kappa \quad \text{if } h^{n+1}_\kappa\geq h^n_\kappa,\\ u_\kappa(\xi+(h_\kappa(t)-h^{n+1}_\kappa),t) & \text{for all } \xi\geq 0 \quad \text{if } h^{n+1}_\kappa< h^n_\kappa.\end{cases}$$

Let $\varphi\in\mathcal{C}^\infty_c(\Omega\times\mathbb{R}^*_+\times\mathbb{R}^*_+)$ and $T>0$ be such that $\varphi(.,.,t)=0$ for all $t\geq T$. Since the concentrations are bounded in $[0,1]$, it can be shown that

$$\left|\int_\Omega\int_{\mathbb{R}^*_+}\int_{\mathbb{R}^*_+}(\bar{u}_{\mathcal{K}_m,\Delta t_m}-u_{\mathcal{K}_m,\Delta t_m})\,\varphi(x,\xi,t)\,dx\,d\xi\,dt\right|$$
$$\leq C_1\sum_{n=0}^{N_{\Delta t_m}}\Delta t_m\sum_{\kappa\in\mathcal{K}_m}|\kappa||h^{n+1}_\kappa-h^n_\kappa|,$$

with C_1 depending only on φ, Ω, and T. From the estimate (4.16) it results that

$$\left|\int_\Omega\int_{\mathbb{R}^*_+}\int_{\mathbb{R}^*_+}(\bar{u}_{\mathcal{K}_m,\Delta t_m}-u_{\mathcal{K}_m,\Delta t_m})\,\varphi(x,\xi,t)\,dx\,d\xi\,dt\right|\to 0 \text{ as } m\to\infty,$$

and $u=u'$ in the space of distributions on $\Omega\times\mathbb{R}^*_+\times\mathbb{R}^*_+$, and hence in $L^\infty(\Omega\times\mathbb{R}^*_+\times\mathbb{R}^*_+)$. □

Flux term. The following proposition provides a result of convergence for the flux term appearing in the discretization of the surface conservation equation. It will be used to show that (c^s_i,u_i) satisfies the second equation (2.9) of the weak formulation. The proof of this proposition is an adaptation to the coupling of a parabolic and a hyperbolic equation of the result proved in [6] for the coupling of an elliptic and a hyperbolic equation in the case of a two phase Darcy flow.

Proposition 5.3. *Let us assume that Hypothesis* 1 *holds and let* h *denote the solution of problem* (2.6). *Let us consider a family of admissible discretizations* $(\mathcal{K},\Sigma_{int},\mathcal{P},\Delta t)$ *of* $\Omega\times\mathbb{R}^*_+$, *with* $(\mathcal{K},\Sigma_{int},\mathcal{P})$ *an admissible mesh of* Ω *in the sense of Definition* 3.1 *and* $\Delta t>0$ *a time step. Let us also assume that there exist* α *and* $\beta>0$ *such that, for all discretizations* $(\mathcal{K},\Sigma_{int},\mathcal{P},\Delta t)$ *of this family,* $\delta\mathcal{K}\leq\beta\sqrt{\Delta t}$ *and* $reg(\mathcal{K})\leq\alpha$. *For any admissible discretization* $(\mathcal{K},\Sigma_{int},\mathcal{P},\Delta t)$, *let* $h_{\mathcal{K},\Delta t}$ *denote the function defined by* (3.7) *from the solution of* (4.1), *and let* $(c^{s,n+1}_{i,\kappa})_{\kappa\in\mathcal{K}_m,n\geq 0}$ *be any solution of* (3.2)–(3.5) *chosen according to Lemma* 5.1. *Let* $T>0$, *then, for all* $\varphi\in\mathcal{A}^s_0=\{v\in\mathcal{C}^\infty_c(\mathbb{R}^{d+1})\,|\,v(x,t)=0$ *on* $\partial\Omega\times\mathbb{R}^*_+\setminus\Sigma^+\}$, *and for all* $i=1,\ldots,L$,

$$T_{i,\mathcal{K},\Delta t}\to\int_0^T\left(\int_\Omega c^s_i(x,t)\,\nabla h(x,t)\cdot\nabla\varphi(x,t)\,dx-\int_{\partial\Omega}\tilde{c}_i(x,t)\,g(x,t)\,\varphi(x,t)\,d\gamma(x)\right)dt$$

as $\Delta t \to 0$, *with*

$$
\begin{aligned}
T_{i,\mathcal{K},\Delta t} = &\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c^{s,n+1}_{i,\kappa\kappa'} \left(h^{n+1}_\kappa - h^{n+1}_{\kappa'}\right) \varphi(x_\kappa, t^{n+1})\\
&- \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega| \left(g^{(+),n+1}_\kappa \tilde{c}^{n+1}_{i,\kappa} - g^{(-),n+1}_\kappa c^{s,n+1}_{i,\kappa}\right) \varphi(x_\kappa, t^{n+1}).
\end{aligned}
$$

Columns property. The following proposition states that the column concentrations interpolated in time $\bar{u}_{i,\mathcal{K},\Delta t}$, $i = 1, \ldots, L$, satisfy in the weak sense a linear advection equation. This property is used in the proof of Theorem 3.3 to show the convergence, up to a subsequence, of the approximate solutions to a solution of the weak formulation (2.7).

PROPOSITION 5.4. *Let us assume that Hypothesis* 1 *holds and let* h *denote the solution of problem* (2.6). *Let* $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ *be an admissible mesh of* Ω *in the sense of Definition* 3.1, $T > 0$, *and* $\Delta t \in (0, T)$.

Let $h_{\mathcal{K},\Delta t}$, $u_{i,\mathcal{K},\Delta t}$, $i = 1, \ldots, L$ *(resp.,* $\delta_t h_{\mathcal{K},\Delta t}$ *and* $\bar{u}_{i,\mathcal{K},\Delta t}$, $i = 1, \ldots, L$*), denote the unique functions defined by* (3.7) *(resp., by* (4.17) *and* (5.7)*) and* $c^s_{i,\mathcal{K},\Delta t}$, $i = 1, \ldots, L$, *be a function defined by* (3.7), *from any solution of* (3.2)–(3.6) *chosen according to Lemma* 5.1.

Then, for any $\kappa \in \mathcal{K}$ *and* $i \in \{1, \ldots, L\}$, *the following hold.*

(i) *For all* $\varphi \in W_T = \{v \in \mathcal{C}^\infty_c(\mathbb{R}^2) \,|\, v(.,T) = 0$ *on* $\mathbb{R}\}$,

$$
\begin{aligned}
&\int_0^T \int_{\mathbb{R}_+} \left[\partial_t\varphi(\xi,t) + \partial_t h_\kappa(t)\, \partial_\xi\varphi(\xi,t)\right] u_{i,\kappa}(\xi,t)\, d\xi\, dt\\
&\quad + \int_{\mathbb{R}_+} u^0_{i,\kappa}(\xi)\varphi(\xi,0)\, d\xi + \int_0^T \partial_t h_\kappa(t) u_{i,\kappa}(0,t)\varphi(0,t)\, dt = 0.
\end{aligned}
\tag{5.8}
$$

(ii) *For all* $\varphi \in \mathcal{A}^s_{T,\kappa} = \{v \in \mathcal{C}^\infty_c(\mathbb{R}^2) \,|\, v(.,T) = 0$ *on* $\mathbb{R}$ *and* $v(0,t) = 0$ *for all* $t \geq 0$ *such that* $\partial_t h_\kappa(t) \leq 0\}$,

$$
\begin{aligned}
&\int_0^T \int_{\mathbb{R}_+} \left[\partial_t\varphi(\xi,t) + \partial_t h_\kappa(t)\, \partial_\xi\varphi(\xi,t)\right] u_{i,\kappa}(\xi,t)\, d\xi\, dt\\
&\quad + \int_{\mathbb{R}_+} u^0_{i,\kappa}(\xi)\varphi(\xi,0)\, d\xi + \int_0^T \partial_t h_\kappa(t) c^s_{i,\kappa}(t)\varphi(0,t)\, dt = 0.
\end{aligned}
\tag{5.9}
$$

Proof. Thanks to definition (5.4), $\partial_t c_{i,\kappa}(z,t) = 0$ for all $z \in (-\infty, h_\kappa(t))$ and $t \in (0,T)$. It results that for all $\psi \in W^{1,\infty}(\mathbb{R} \times \mathbb{R}_+)$, compactly supported, one has

$$
\begin{aligned}
0 = &\int_0^T \int_{-\infty}^{h_\kappa(t)} \partial_t c_{i,\kappa}(z,t)\psi(z,t) dz\, dt = \int_0^T \partial_t \left(\int_{-\infty}^{h_\kappa(t)} c_{i,\kappa}(z,t)\psi(z,t) dz \right) dt\\
&- \int_0^T \int_{-\infty}^{h_\kappa(t)} c_{i,\kappa}(z,t)\partial_t\psi(z,t) dz\, dt - \int_0^T \partial_t h_\kappa(t) c_{i,\kappa}(h_\kappa(t),t)\psi(h_\kappa(t),t)\, dt,
\end{aligned}
$$

and consequently

$$
\begin{aligned}
&\int_0^T \int_{-\infty}^{h_\kappa(t)} c_{i,\kappa}(z,t)\partial_t\psi(z,t) dz\, dt + \int_0^T \partial_t h_\kappa(t) c_{i,\kappa}(h_\kappa(t),t)\psi(h_\kappa(t),t)\, dt\\
&\quad - \int_{-\infty}^{h_\kappa(T)} c_{i,\kappa}(z,T)\psi(z,T)\, dz + \int_{-\infty}^{h_\kappa(0)} c^0_{i,\kappa}(z)\psi(z,0)\, dz = 0.
\end{aligned}
\tag{5.10}
$$

Let φ be in W_T, and let $\psi \in W^{1,\infty}(\mathbb{R}\times\mathbb{R}_+)$ be such that

$$\psi(z,t) = \varphi(h_\kappa(t) - z, t) \quad \forall\,(z,t) \in \mathbb{R}\times\mathbb{R}_+.$$

Considering (5.10) in the new coordinate system $\xi = h_\kappa(t) - z$ and using the property $\varphi(.,T) = 0$, (5.8) is derived. Finally, thanks to the definition of $\mathcal{A}^s_{T,\kappa}$ and since $u_{i,\kappa}(0,t) = c^s_{i,\kappa}(t)$ if $\partial_t h_\kappa(t) > 0$, we obtain (5.9). □

Convergence. We will now prove that the limits $(c^s_i, u_i)_{i=1,\dots,L}$ are solutions of the weak formulation given in Definition 2.1.

Lemma 5.5. *Let $\mathcal{O}$ be an open bounded subset of $\mathbb{R}^d$, and let $(f_n)_{n\in\mathbb{N}}$ be a sequence of $L^1(\mathcal{O})$ which converges to f in $L^1(\mathcal{O})$. Let us define, for any $g \in L^1(\mathcal{O})$, $\mathcal{S}^+_g = \{x \in \mathcal{O}\,|\,g(x) > 0\}$ and $\mathcal{S}^-_g = \{x \in \mathcal{O}\,|\,g(x) \le 0\}$; then*

$$I_n = \int_{\mathcal{O}} f_n \chi_{\mathcal{S}^+_{f_n}\cap\mathcal{S}^-_f} \to 0 \text{ as } n\to\infty, \text{ and } J_n = \int_{\mathcal{O}} f_n \chi_{\mathcal{S}^-_{f_n}\cap\mathcal{S}^+_f} \to 0 \text{ as } n\to\infty.$$

Proof. Note that if $f \in L^1(\mathcal{O})$, then f^+ and f^- belong to $L^1(\mathcal{O})$; thus

$$I_n = \int_{\mathcal{O}} f_n \chi_{\mathcal{S}^+_{f_n}} \chi_{\mathcal{S}^-_f} = \int_{\mathcal{O}} f^+_n \chi_{\mathcal{S}^-_f} = \int_{\mathcal{O}} (f^+_n - f^+)\chi_{\mathcal{S}^-_f} + \int_{\mathcal{O}} f^+ \chi_{\mathcal{S}^-_f}.$$

Since $\int_{\mathcal{O}} f^+\chi_{\mathcal{S}^-_f} = 0$ and $|f^+_n - f^+| \le |f_n - f|$ on $\mathcal{O}$, we deduce that $I_n \to 0$ as $n\to\infty$. The proof is similar for J_n. □

Let us now prove the convergence result given by Theorem 3.3.

Proof of Theorem 3.3. The convergence of the approximate solutions for the sediment thickness toward the solution of problem (2.6) has already been proved in Proposition 4.2. Let us now show that the limits $(c^s_i, u_i)_{i=1,\dots,L}$ given by Proposition 5.2 satisfy the weak formulation of problem (2.7) in the sense of Definition 2.1.

Let i belong to $\{1,\dots,L\}$ and $\varphi \in \mathcal{A}$. Since $\varphi \in \mathcal{C}^\infty_c(\mathbb{R}^{d+2})$, there exists $T > 0$ such that, for all $t \ge T$, $\varphi(.,.,t) = 0$. Let $m_0 \in \mathbb{N}$ be such that $\Delta t_{m_0} < T$. For the sake of simplicity, we shall drop the subscript i.

For all $\kappa \in \mathcal{K}_m$, $m \in \mathbb{N}$, note that $\varphi(x_\kappa,.,.) \in W_T$. Applying (5.8) to the test function $\varphi(x_\kappa,.,.)$ and summing this equation over $\kappa \in \mathcal{K}_m$, we get, for any $m \ge m_0$,

$$\begin{aligned}
&\underbrace{\sum_{\kappa\in\mathcal{K}_m} |\kappa| \int_{\mathbb{R}_+}\int_0^T \left[\partial_t\varphi(x_\kappa,\xi,t) + \partial_t h_\kappa(t)\,\partial_\xi\varphi(x_\kappa,\xi,t)\right] u_\kappa(\xi,t)\,dt\,d\xi}_{(A_m)}\\
&\qquad + \underbrace{\sum_{\kappa\in\mathcal{K}_m} |\kappa| \int_{\mathbb{R}_+} u^0_\kappa(\xi)\varphi(x_\kappa,\xi,0)\,d\xi}_{(B_m)}\\
&\qquad + \underbrace{\sum_{\kappa\in\mathcal{K}_m} |\kappa| \int_0^T \partial_t h_\kappa(t) u_\kappa(0,t)\varphi(x_\kappa,0,t)\,dt}_{(C_m)} = 0.
\end{aligned}\tag{5.11}$$

In this equation, (A_m) is equal to

$$\int_\Omega\int_{\mathbb{R}_+}\int_0^T \left[\partial_t\varphi_{\mathcal{K}_m}(x,\xi,t) + \delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\,\partial_\xi\varphi_{\mathcal{K}_m}(x,\xi,t)\right] \bar{u}_{\mathcal{K}_m,\Delta t_m}(x,\xi,t)\,dt\,d\xi\,dx,$$

where $\varphi_{\mathcal{K}_m}(x,\xi,t) = \varphi(x_\kappa,\xi,t)$ for all $x \in \kappa$. Thanks to Proposition 4.2, the sequence of functions $(\delta_t h_{\mathcal{K}_m,\Delta t_m})$ converges strongly to $\partial_t h$ in $L^2(\Omega\times(0,T))$ as $m\to\infty$. Since $\varphi\in\mathcal{A}$, we deduce that the sequence $(\partial_\xi\varphi_{\mathcal{K}_m}\cdot\delta_t h_{\mathcal{K}_m,\Delta t_m})$ converges to $\partial_\xi\varphi\cdot\partial_t h$ in $L^1(\Omega\times\mathbb{R}_+^*\times\mathbb{R}_+^*)$. Since the sequence $(\bar{u}_{\mathcal{K}_m,\Delta t_m})$ converges to u in $L^\infty(\Omega\times\mathbb{R}_+^*\times\mathbb{R}_+^*)$ for the weak-$\star$ topology, we conclude that

$$(5.12)\quad (A_m)\to\int_\Omega\int_{\mathbb{R}_+}\int_{\mathbb{R}_+}\big[\partial_t\varphi(x,\xi,t)+\partial_t h(x,t)\,\partial_\xi\varphi(x,\xi,t)\big]u(x,\xi,t)\,dt\,d\xi\,dx$$
$$\text{as } m\to\infty.$$

Let us define $u^0_{\mathcal{K}_m}$ by $u^0_{\mathcal{K}_m}(x,\xi)=u^0_\kappa(\xi)$ for all $x\in\kappa$, $\kappa\in\mathcal{K}_m$, and $\xi\in\mathbb{R}_+^*$. From Hypothesis 1 on u^0, it results that $u^0_{\mathcal{K}_m}$ converges to u^0 in $L^1(\Omega\times(0,T))$ for all $T>0$, and consequently

$$(5.13)\qquad (B_m)\to\int_\Omega\int_{\mathbb{R}_+}u^0(x,\xi)\varphi(x,\xi,0)\,d\xi\,dx \text{ as } m\to\infty.$$

In (5.11), (C_m) is equal to

$$(C_m)=\int_\Omega\int_0^T\delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\bar{u}_{\mathcal{K}_m,\Delta t_m}(x,0,t)\varphi_{\mathcal{K}_m}(x,0,t)\,dt\,dx.$$

Let us introduce the following notation:

$$\begin{aligned}
\mathcal{P}^+_{\mathcal{K}_m} &= \{(x,t)\in\Omega\times(0,T)\,|\,\delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)>0\},\\
\mathcal{P}^-_{\mathcal{K}_m} &= \{(x,t)\in\Omega\times(0,T)\,|\,\delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\le 0\},\\
\mathcal{P}^+ &= \{(x,t)\in\Omega\times(0,T)\,|\,\partial_t h(x,t)>0\},\\
\mathcal{P}^- &= \{(x,t)\in\Omega\times(0,T)\,|\,\partial_t h(x,t)\le 0\}.
\end{aligned}$$

Noticing that $\mathcal{P}^+_{\mathcal{K}_m}=(\mathcal{P}^+\setminus(\mathcal{P}^+\cap\mathcal{P}^-_{\mathcal{K}_m}))\cup(\mathcal{P}^+_{\mathcal{K}_m}\cap\mathcal{P}^-)$ and $\mathcal{P}^-_{\mathcal{K}_m}=(\mathcal{P}^-\setminus(\mathcal{P}^-\cap\mathcal{P}^+_{\mathcal{K}_m}))\cup(\mathcal{P}^-_{\mathcal{K}_m}\cap\mathcal{P}^+)$, one has

$$\begin{aligned}
(C_m)=&\int_\Omega\int_0^T\delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\,c^s_{\mathcal{K}_m,\Delta t_m}(x,t)\,\varphi_{\mathcal{K}_m}(x,0,t)\\
&\qquad\cdot[\chi_{\mathcal{P}^+}-\chi_{\mathcal{P}^+\cap\mathcal{P}^-_{\mathcal{K}_m}}+\chi_{\mathcal{P}^+_{\mathcal{K}_m}\cap\mathcal{P}^-}]\,dt\,dx\\
&+\int_\Omega\int_0^T\delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\,\bar{u}_{\mathcal{K}_m,\Delta t_m}(x,0,t)\,\varphi_{\mathcal{K}_m}(x,0,t)\\
&\qquad\cdot[\chi_{\mathcal{P}^-}-\chi_{\mathcal{P}^-\cap\mathcal{P}^+_{\mathcal{K}_m}}+\chi_{\mathcal{P}^-_{\mathcal{K}_m}\cap\mathcal{P}^+}]\,dt\,dx.
\end{aligned}$$

Since the functions $c^s_{\mathcal{K}_m,\Delta t_m}(x,t)$, $\bar{u}_{\mathcal{K}_m,\Delta t_m}(x,0,t)$, and $\varphi_{\mathcal{K}_m}(x,0,t)$ are bounded on $\Omega\times(0,T)$ and $(\delta_t h_{\mathcal{K}_m,\Delta t_m})$ converges to $\partial_t h$ in $L^2(\Omega\times(0,T))$, Lemma 5.5 applied to the sequence $(\delta_t h_{\mathcal{K}_m,\Delta t_m})_{m\in\mathbb{N}}$ yields

$$\int_\Omega\int_0^T\delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\,c^s_{\mathcal{K}_m,\Delta t_m}(x,t)\,\varphi_{\mathcal{K}_m}(x,0,t)[-\chi_{\mathcal{P}^+\cap\mathcal{P}^-_{\mathcal{K}_m}}+\chi_{\mathcal{P}^+_{\mathcal{K}_m}\cap\mathcal{P}^-}]\,dt\,dx\to 0,$$

$$\int_\Omega\int_0^T\delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\,\bar{u}_{\mathcal{K}_m,\Delta t_m}(x,0,t)\varphi_{\mathcal{K}_m}(x,0,t)[-\chi_{\mathcal{P}^-\cap\mathcal{P}^+_{\mathcal{K}_m}}+\chi_{\mathcal{P}^-_{\mathcal{K}_m}\cap\mathcal{P}^+}]\,dt\,dx\to 0$$

as $m\to\infty$. Furthermore, $\varphi\in\mathcal{A}$, so that the sequence $(\varphi_{\mathcal{K}_m}(.,0,.)\,\delta_t h_{\mathcal{K}_m,\Delta t_m})$ converges to $\varphi(.,0,.)\,\partial_t h$ in $L^1(\Omega\times(0,T))$. As the sequence $(c^s_{\mathcal{K}_m,\Delta t_m})$ converges to

c^s in $L^\infty(\Omega \times \mathbb{R}_+^*)$ for the weak-$\star$ topology, we conclude that

$$\int_\Omega \int_0^T \delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t) c^s_{\mathcal{K}_m,\Delta t_m}(x,t) \varphi_{\mathcal{K}_m}(x,0,t)\, \chi_{\mathcal{P}_+}\, dt\, dx \to \int_\Omega \int_0^T \partial_t h(x,t) c^s(x,t) \varphi(x,0,t)\, \chi_{\mathcal{P}_+}\, dt\, dx \text{ as } m \to \infty.$$

On $\chi_{\mathcal{P}_-}$, by definition, one has $\varphi(x,0,t) = 0$. Since $\bar{u}_{\mathcal{K}_m,\Delta t_m}(x,0,t)$ is bounded and the sequence $(\varphi_{\mathcal{K}_m}(.,0,.)\, \delta_t h_{\mathcal{K}_m,\Delta t_m})$ converges to $\varphi(.,0,.)\, \partial_t h$ in $L^1(\Omega \times (0,T))$, we obtain

$$\int_\Omega \int_0^T \delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\, \bar{u}_{\mathcal{K}_m,\Delta t_m}(x,0,t)\, \varphi_{\mathcal{K}_m}(x,0,t)\, \chi_{\mathcal{P}_-}\, dt\, dx \to 0 \text{ as } m \to \infty,$$

and finally

$$(C_m) \to \int_\Omega \int_0^T \partial_t h(x,t) c^s(x,t) \varphi(x,0,t) dt\, dx = \int_\Omega \int_{\mathbb{R}_+} \partial_t h(x,t) c^s(x,t) \varphi(x,0,t)\, dt\, dx$$

as $m \to \infty$. Then (c_i^s, u_i) satisfy the first part (2.8) of the weak formulation.

Let $\varphi \in \mathcal{A}_0$. Since $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{d+2})$, there exists $T > 0$ such that $\varphi(.,.,t) = 0$ for all $t \geq T$. Let $m_0 \in \mathbb{N}$ be such that $\Delta t_{m_0} < T$.

Multiplying the scheme (3.2) by $\varphi(x_\kappa, 0, t^{n+1})$ and summing over $\kappa \in \mathcal{K}_m$ and $n \in \{0, \dots, N_{\Delta t_m}\}$, one obtains, for any $m \geq m_0$,

$$\underbrace{\sum_{n=0}^{N_{\Delta t_m}} \sum_{\kappa \in \mathcal{K}_m} |\kappa| \Delta \mathcal{M}_\kappa^{n+1} \varphi(x_\kappa, 0, t^{n+1})}_{(1_m)}$$
$$+ \underbrace{\sum_{n=0}^{N_{\Delta t_m}} \Delta t_m \sum_{\kappa \in \mathcal{K}_m} \sum_{\kappa' \in \mathcal{K}_\kappa} c^{s,n+1}_{\kappa\kappa'} T_{\kappa\kappa'} (h_\kappa^{n+1} - h_{\kappa'}^{n+1}) \varphi(x_\kappa, 0, t^{n+1})}_{(2_m)}$$
$$\underbrace{- \sum_{n=0}^{N_{\Delta t_m}} \Delta t_m \sum_{\kappa \in \mathcal{K}_m} |\partial\kappa \cap \partial\Omega| \left(\tilde{c}_\kappa^{n+1} g_\kappa^{(+),n+1} - c_\kappa^{s,n+1} g_\kappa^{(-),n+1} \right) \varphi(x_\kappa, 0, t^{n+1})}_{(3_m)} = 0.$$

Since $\varphi(.,0,.) \in \mathcal{A}_0^s$, Proposition 5.3 with $\mathcal{K} = \mathcal{K}_m$ and $\Delta t = \Delta t_m$ states that (2_m)+(3_m) converges to

$$\begin{aligned} A &= \int_0^T \left(\int_\Omega c^s(x,t)\, \nabla h(x,t) \cdot \nabla \varphi(x,0,t)\, dx - \int_{\partial\Omega} \tilde{c}(x,t) g(x,t) \varphi(x,0,t) d\gamma(x) \right) dt \\ &= \int_{\mathbb{R}_+} \left(\int_\Omega c^s(x,t)\, \nabla h(x,t) \cdot \nabla \varphi(x,0,t)\, dx - \int_{\partial\Omega} \tilde{c}(x,t) g(x,t) \varphi(x,0,t) d\gamma(x) \right) dt, \end{aligned}$$

as $m \to \infty$. Let us now prove the convergence of

$$A'_m = -(1_m) = -\sum_{n=0}^{N_{\Delta t_m}} \sum_{\kappa \in \mathcal{K}_m} |\kappa| \Delta \mathcal{M}_\kappa^{n+1} \varphi(x_\kappa, 0, t^{n+1})$$

toward

$$B = \int_\Omega \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} \left[\partial_t \varphi(x,\xi,t) + \partial_t h(x,t)\, \partial_\xi \varphi(x,\xi,t)\right] u(x,\xi,t)\, dt\, d\xi\, dx \\ + \int_\Omega \int_{\mathbb{R}_+} u^0(x,\xi)\varphi(x,\xi,0)\, d\xi\, dx$$

as $m \to \infty$. From (5.12) and (5.13), we have, for any $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}^{d+2}) \supset \mathcal{A}_0$,

$$B'_m = \sum_{\kappa \in \mathcal{K}_m} |\kappa| \int_{\mathbb{R}_+} \int_0^T \left[\partial_t \varphi(x_\kappa,\xi,t) + \partial_t h_\kappa(t)\, \partial_\xi \varphi(x_\kappa,\xi,t)\right] u_\kappa(\xi,t)\, dt\, d\xi \\ + \sum_{\kappa \in \mathcal{K}_m} |\kappa| \int_{\mathbb{R}_+} u_\kappa^0(\xi)\varphi(x_\kappa,\xi,0)\, d\xi \to B \text{ as } m \to \infty,$$

and, from (5.11), $B'_m = -\sum_{\kappa\in\mathcal{K}_m} |\kappa| \int_0^T \partial_t h_\kappa(t) u_\kappa(0,t)\varphi(x_\kappa,0,t)\, dt$. Hence, it will suffice to show that $|A'_m - B'_m| \to 0$ as $m \to \infty$.

For given $\kappa \in \mathcal{K}_m$ and $n \in \{0,\dots,N_{\Delta t_m}\}$, let us recall that

$$\Delta\mathcal{M}_\kappa^{n+1} = \begin{cases} \int_{h_\kappa^n}^{h_\kappa^{n+1}} c_\kappa^{n+1}(z) dz & \text{if } h_\kappa^{n+1} \geq h_\kappa^n, \\ \int_{h_\kappa^n}^{h_\kappa^{n+1}} c_\kappa^n(z)\, dz & \text{if } h_\kappa^{n+1} < h_\kappa^n. \end{cases}$$

Considering the change of coordinates $z = h_\kappa(t)$ in these integrals, one can show that, in both the sedimentation ($h_\kappa^{n+1} \geq h_\kappa^n$) and erosion ($h_\kappa^{n+1} < h_\kappa^n$) cases, one has

$$\Delta\mathcal{M}_\kappa^{n+1} = \int_{t^n}^{t^{n+1}} c_\kappa(h_\kappa(t),t)\partial_t h_\kappa(t)\, dt = \int_{t^n}^{t^{n+1}} u_\kappa(0,t)\partial_t h_\kappa(t)\, dt.$$

Substituting this equality in the definition of A'_m leads to

$$B'_m - A'_m = \sum_{\kappa\in\mathcal{K}_m} |\kappa| \int_T^{t^{N_{\Delta t_m}+1}} \bar{u}_{\mathcal{K}_m,\Delta t_m}(x,0,t)\, \delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t)\varphi(x_\kappa,0,t) dt \\ - \sum_{\kappa\in\mathcal{K}_m} |\kappa| \sum_{n=0}^{N_{\Delta t_m}} \int_{t^n}^{t^{n+1}} \bar{u}_{\mathcal{K}_m,\Delta t_m}(x,0,t) \delta_t h_{\mathcal{K}_m,\Delta t_m}(x,t) [\varphi(x_\kappa,0,t^{n+1}) - \varphi(x_\kappa,0,t)]\, dt.$$

Thanks to the regularity of φ, there exists $C_1 > 0$, depending only on φ, such that $|\varphi(x_\kappa,0,t^{n+1}) - \varphi(x_\kappa,0,t)| \leq C_1\, \Delta t_m$ for all $t \in [t^n, t^{n+1}]$. Since the function $\delta_t h_{\mathcal{K}_m,\Delta t_m}$ is uniformly bounded in $L^2(\Omega \times (0, t^{N_{\Delta t_m}+1}))$, and $\bar{u}_{\mathcal{K}_m,\Delta t_m} \in [0,1]$, and $|t^{N_{\Delta t_m}+1} - T| < \Delta t_m$, the convergence of $|A'_m - B'_m|$ to 0 as $m \to \infty$ is obtained, which ends the proof of the theorem. □

6. Conclusion.

In this article, a fully implicit finite volume discretization of the multilithology stratigraphic model is considered in the simplified case for which the diffusion coefficients of all the lithologies are equal.

In such a case, the sediment thickness variable decouples from the other variables and satisfies a parabolic equation. A weak formulation has been defined for the remaining surface and basin concentration variables in order to cope with the difficulty to define the trace of the basin concentrations at the top of the basin. Then, the main result of this article is the convergence, up to a subsequence, of the discrete sediment thickness in $L^\infty(0,T;L^2(\Omega))$ and of the discrete concentrations in the L^∞ weak-$\star$ topology to a weak solution.

In particular, this proves the existence of at least one solution to the weak formulation for the coupled problem. The uniqueness of such a solution, and hence the full convergence of the discrete solutions, will be obtained in a forthcoming paper.

Appendix. Proof of Proposition 5.3.

To prove Proposition 5.3, the following weak-BV estimate will be used. It is an extension to the coupling of a parabolic and a hyperbolic equation of the result proved in [6] for the coupling of an elliptic and a hyperbolic equation in the case of a two phase Darcy flow.

LEMMA A.1. *Let us assume that Hypothesis* 1 *holds, and let* h *denote the solution of problem* (2.6). *Let* $i \in \{1,\dots,L\}$ $(\mathcal{K}, \Sigma_{int}, \mathcal{P})$ *be an admissible mesh of* Ω *in the sense of Definition* 3.1, $T > 0$, *and* $\Delta t \in (0,T)$. *Let* $\alpha > 0$ *be such that* $reg(\mathcal{K}) \leq \alpha$ *and* $\beta > 0$ *be such that* $\delta\mathcal{K} \leq \beta\sqrt{\Delta t}$. *Then, there exists* $H > 0$, *depending only on* T, Ω, $\|h\|_{W^{2,\infty}(\Omega\times(0,2T))}$, $\|g\|_{L^2(\partial\Omega\times\mathbb{R}_+)}$, β, *and* α, *such that the following inequality holds:*

$$\begin{aligned}(\text{A.1})\qquad &\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}\,|h_\kappa^{n+1} - h_{\kappa'}^{n+1}|\,|c_{i,\kappa}^{s,n+1} - c_{i,\kappa'}^{s,n+1}| \\ &\quad + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\,|c_{i,\kappa}^{s,n+1} - \tilde{c}_{i,\kappa}^{n+1}|\,g_\kappa^{(+),n+1} \leq \frac{H}{\sqrt{\delta\mathcal{K}}}.\end{aligned}$$

Proof. Let i belong to the set $\{1,\dots,L\}$. Again, the subscript i will be dropped in the proof, and c_i^s will be denoted by c. Multiplying (3.2) by c_κ^{n+1} and summing over $\kappa \in \mathcal{K}$ and $n \in \{0,\dots,N_{\Delta t}\}$ yield that

$$\begin{aligned}(\text{A.2})\qquad &\sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} |\kappa| c_\kappa^{*,n+1}\, c_\kappa^{n+1}(h_\kappa^{n+1} - h_\kappa^n) \\ &\quad + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c_{\kappa\kappa'}^{n+1}\, c_\kappa^{n+1}(h_\kappa^{n+1} - h_{\kappa'}^{n+1}) \\ &\quad - \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa\cap\partial\Omega|\, g_\kappa^{(+),n+1} \tilde{c}_\kappa^{n+1} c_\kappa^{n+1} \\ &\quad + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa\cap\partial\Omega|\, g_\kappa^{(-),n+1} (c_\kappa^{n+1})^2 = 0,\end{aligned}$$

where $c_\kappa^{*,n+1}$ is defined by $c_\kappa^{*,n+1}(h_\kappa^{n+1} - h_\kappa^n) = \Delta\mathcal{M}_\kappa^{n+1}$, such that $c_\kappa^{*,n+1} \in [0,1]$.

The upstream evaluation of the surface concentrations at the edges of the control volumes implies that, for all $\kappa \in \mathcal{K}$,

$$\begin{aligned}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c_{\kappa\kappa'}^{n+1}\, c_\kappa^{n+1}(h_\kappa^{n+1} - h_{\kappa'}^{n+1}) &= \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} (c_\kappa^{n+1})^2 (h_\kappa^{n+1} - h_{\kappa'}^{n+1})^+ \\ &\quad - \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c_{\kappa'}^{n+1}\, c_\kappa^{n+1}(h_\kappa^{n+1} - h_{\kappa'}^{n+1})^-.\end{aligned}$$

Therefore, since $(h_\kappa - h_{\kappa'})^+ = (h_{\kappa'} - h_\kappa)^-$, one has

$$\begin{aligned}&\sum_{\kappa\in\mathcal{K}} \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c_{\kappa\kappa'}^{n+1}\, c_\kappa^{n+1}(h_\kappa^{n+1} - h_{\kappa'}^{n+1}) \\ &\quad = \sum_{\kappa\in\mathcal{K}} \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}((c_\kappa^{n+1})^2 - c_{\kappa'}^{n+1}\, c_\kappa^{n+1})(h_\kappa^{n+1} - h_{\kappa'}^{n+1})^+.\end{aligned}$$

Then, using the equalities $(c_\kappa)^2 - c_\kappa\, c_{\kappa'} = \frac{1}{2}(c_\kappa - c_{\kappa'})^2 + \frac{1}{2}\big((c_\kappa)^2 - (c_{\kappa'})^2\big)$, $(h_\kappa - h_{\kappa'})^+ = (h_{\kappa'} - h_\kappa)^-$, and $(h_\kappa - h_{\kappa'}) = (h_\kappa - h_{\kappa'})^+ - (h_\kappa - h_{\kappa'})^-$ leads to the following successive equalities:

$$\begin{aligned}
&\sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c_{\kappa\kappa'}^{n+1}\, c_\kappa^{n+1}(h_\kappa^{n+1} - h_{\kappa'}^{n+1})\\
&\quad = \frac{1}{2}\sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}(c_\kappa^{n+1} - c_{\kappa'}^{n+1})^2(h_\kappa^{n+1} - h_{\kappa'}^{n+1})^+\\
&\quad +\frac{1}{2}\sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}(c_\kappa^{n+1})^2(h_\kappa^{n+1} - h_{\kappa'}^{n+1})^+\\
&\quad -\frac{1}{2}\sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}(c_{\kappa'}^{n+1})^2(h_\kappa^{n+1} - h_{\kappa'}^{n+1})^+\\
&\quad = \frac{1}{2}\sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}(c_\kappa^{n+1} - c_{\kappa'}^{n+1})^2(h_\kappa^{n+1} - h_{\kappa'}^{n+1})^+\\
&\quad +\frac{1}{2}\sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}(c_\kappa^{n+1})^2(h_\kappa^{n+1} - h_{\kappa'}^{n+1}).
\end{aligned}\tag{A.3}$$

Furthermore, summing (3.2) over $i \in \{1, \dots, L\}$, we obtain, for all $\kappa \in \mathcal{K}$ and $n \in \{0, \dots, N_{\Delta t}\}$,

$$|\kappa|(h_\kappa^{n+1} - h_\kappa^n) + \Delta t \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}(h_\kappa^{n+1} - h_{\kappa'}^{n+1}) - \Delta t\, |\partial\kappa \cap \partial\Omega|\, g_\kappa^{n+1} = 0. \tag{A.4}$$

Multiplying (A.4) by $(c_\kappa^{n+1})^2$ and summing over $\kappa \in \mathcal{K}$ gives in (A.3)

$$\begin{aligned}
&\sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c_{\kappa\kappa'}^{n+1}\, c_\kappa^{n+1}(h_\kappa^{n+1} - h_{\kappa'}^{n+1}) = -\frac{1}{2}\sum_{\kappa\in\mathcal{K}} |\kappa|(c_\kappa^{n+1})^2 \frac{h_\kappa^{n+1} - h_\kappa^n}{\Delta t}\\
&+\frac{1}{2}\sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\,(c_\kappa^{n+1})^2 g_\kappa^{n+1} + \frac{1}{2}\sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'}(c_\kappa^{n+1} - c_{\kappa'}^{n+1})^2(h_\kappa^{n+1} - h_{\kappa'}^{n+1})^+,
\end{aligned}$$

which finally results in the equality

$$\begin{aligned}
&\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}}\sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c_{\kappa\kappa'}^{n+1}\, c_\kappa^{n+1}(h_\kappa^{n+1} - h_{\kappa'}^{n+1})\\
&\qquad - \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1} \tilde{c}_\kappa^{n+1} c_\kappa^{n+1}\\
&\qquad + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(-),n+1} (c_\kappa^{n+1})^2\\
&\quad = \frac{1}{2}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}(c_\kappa^{n+1} - c_{\kappa'}^{n+1})^2 |h_\kappa^{n+1} - h_{\kappa'}^{n+1}|\\
&\qquad + \frac{1}{2}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1}(c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1})^2 - \frac{1}{2}\sum_{n=0}^{N_{\Delta t}}\sum_{\kappa\in\mathcal{K}} |\kappa|(c_\kappa^{n+1})^2(h_\kappa^{n+1} - h_\kappa^n)\\
&\qquad + \frac{1}{2}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} (|\partial\kappa \cap \partial\Omega|\, g_\kappa^{(-),n+1}(c_\kappa^{n+1})^2 - |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1}(\tilde{c}_\kappa^{n+1})^2).
\end{aligned}$$

Using this last result in (A.2), together with $g_\kappa^{(-),n+1} \geq 0$ for all $\kappa \in \mathcal{K}$ and $n \in \{0, \dots, N_{\Delta t}\}$, one obtains the estimate

$$\begin{aligned}
&\frac{1}{2}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}(c_\kappa^{n+1} - c_{\kappa'}^{n+1})^2 |h_\kappa^{n+1} - h_{\kappa'}^{n+1}| \\
&\quad + \frac{1}{2}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1} (c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1})^2 \\
&\quad \leq \frac{1}{2}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\kappa| \big[(c_\kappa^{n+1})^2 - 2\, c_\kappa^{*,n+1}\, c_\kappa^{n+1}\big] \frac{h_\kappa^{n+1} - h_\kappa^n}{\Delta t} \\
&\quad + \frac{1}{2}\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega| g_\kappa^{(+),n+1} (\tilde{c}_\kappa^{n+1})^2.
\end{aligned} \tag{A.5}$$

Noticing that, according to Corollary 1,

$$\begin{aligned}
&\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\kappa| \big[(c_\kappa^{n+1})^2 - 2\, c_\kappa^{*,n+1}\, c_\kappa^{n+1}\big] \frac{h_\kappa^{n+1} - h_\kappa^n}{\Delta t} \\
&\quad \leq C_1(T,\Omega) \left(\sum_{n=0}^{N_{\Delta t}} \Delta t \|\delta_t h_{\mathcal{K}}^n\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}} \leq C_1(T,\Omega) D_6
\end{aligned} \tag{A.6}$$

and

$$\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1}\, (\tilde{c}_\kappa^{n+1})^2 \leq C_2(\Omega, T)\, \|g^+\|_{L^2(\partial\Omega\times\mathbb{R}_+)}, \tag{A.7}$$

we deduce from (A.5), (A.6), and (A.7) the estimate

$$\begin{aligned}
&\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}(c_\kappa^{n+1} - c_{\kappa'}^{n+1})^2 |h_\kappa^{n+1} - h_{\kappa'}^{n+1}| \\
&\quad + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1} (c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1})^2 \leq C_1 \sqrt{D_6} + C_2\, \|g^+\|_{L^2(\partial\Omega\times\mathbb{R}_+)}.
\end{aligned}$$

Finally, the Cauchy–Schwarz inequality yields

$$\begin{aligned}
&\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'} |c_\kappa^{n+1} - c_{\kappa'}^{n+1}| |h_\kappa^{n+1} - h_{\kappa'}^{n+1}| \\
&\quad + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega| |c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1}|\, g_\kappa^{(+),n+1} \\
&\quad \leq \left(\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}(c_\kappa^{n+1} - c_{\kappa'}^{n+1})^2 |h_\kappa^{n+1} - h_{\kappa'}^{n+1}| \right. \\
&\quad \left. + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega| (c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1})^2 g_\kappa^{(+),n+1} \right)^{\frac{1}{2}} \\
&\quad \left(\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'} |h_\kappa^{n+1} - h_{\kappa'}^{n+1}| + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1} \right)^{\frac{1}{2}}.
\end{aligned}$$

The term

$$\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'} |h_{\kappa}^{n+1} - h_{\kappa'}^{n+1}| \leq \left(\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'} \right)^{\frac{1}{2}} \left(\sum_{n=0}^{N_{\Delta t}} \Delta t |h_{\mathcal{K}}^{n+1}|_{1,\mathcal{K}}^2 \right)^{\frac{1}{2}}$$

is estimated by Corollary 1 and the following bound from (3.1):

$$\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'} \leq \left(\sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} |\sigma|\, d(\kappa,\kappa') \right) \frac{2T\,\alpha^2}{\delta\mathcal{K}^2} \leq \frac{2Td\,\alpha^2\,|\Omega|}{\delta\mathcal{K}^2}. \tag{A.8}$$

We conclude from estimates similar to (A.7) that the inequality (A.1) holds. □

Proof of Proposition 5.3. Let i belong to the set $\{1,\dots,L\}$, and let $(\mathcal{K},\Sigma_{int},\mathcal{P},\Delta t)$ be an admissible discretization of $\Omega\times\mathbb{R}_+^*$ with $\Delta t < T$. For all $\kappa\in\mathcal{K}$, $x\in\partial\kappa\cap\partial\Omega$, $t\in\ (t^n,t^{n+1}]$, $n\geq 0$, let us define

$$\tilde{c}_{i,\mathcal{K},\Delta t}(x,t) = \tilde{c}_{i,\kappa}^{n+1}.$$

Throughout this proof we shall now drop the subscript i and use the simplified notation $c_i^s = c$.

Let us define the auxiliary expression E_3 by

$$E_3 = \sum_{n=0}^{N_{\Delta t}} \int_{t^n}^{t^{n+1}} \left(\int_{\Omega} c_{\mathcal{K},\Delta t}(x,t)\,\nabla h(x,t)\cdot\nabla\varphi(x,t^{n+1})\,dx \right. \\ \left. - \int_{\partial\Omega} \tilde{c}_{\mathcal{K},\Delta t}(x,t)\,g(x,t)\,\varphi(x,t^{n+1})\,d\gamma(x) \right) dt.$$

From the L^∞ weak-$\star$ convergence of $c_{\mathcal{K},\Delta t}$ to c and $\tilde{c}_{\mathcal{K},\Delta t}$ to $\tilde{c}$ as Δt and $\delta\mathcal{K}\to 0$, and their boundedness, it results that

$$E_3 \to \int_0^T \left(\int_{\Omega} c(x,t)\,\nabla h(x,t)\cdot\nabla\varphi(x,t)dx - \int_{\partial\Omega} \tilde{c}(x,t)g(x,t)\varphi(x,t)d\gamma(x) \right) dt$$

as $\Delta t\to 0$.

Multiplying (2.6) by $\varphi(x,t^{n+1})$ and integrating it over the time interval (t^n,t^{n+1}) and cell κ yield

$$\int_{t^n}^{t^{n+1}} \int_{\kappa} \partial_t h(x,t)\,\varphi(x,t^{n+1})\,dx\,dt - \int_{t^n}^{t^{n+1}} \int_{\kappa} \Delta h(x,t)\,\varphi(x,t^{n+1})\,dx\,dt = 0.$$

Since $\varphi\in\mathcal{C}_c^\infty(\mathbb{R}^{d+1})$, one obtains

$$\int_{t^n}^{t^{n+1}} \int_{\kappa} \nabla h(x,t)\cdot\nabla\varphi(x,t^{n+1})\,dx\,dt = -\int_{t^n}^{t^{n+1}} \int_{\kappa} \partial_t h(x,t)\,\varphi(x,t^{n+1})\,dx\,dt \\ + \int_{t^n}^{t^{n+1}} \int_{\partial\kappa} \nabla h(x,t)\cdot\vec{n}_\kappa\,\varphi(x,t^{n+1})d\gamma(x)\,dt,$$

where $\vec{n}_\kappa$ is the normal unit vector to $\partial\kappa$ outward to κ. Thus, one has

$$\begin{aligned} E_3 \quad &= \sum_{n=0}^{N_{\Delta t}} \sum_{\substack{\sigma\in\Sigma_{int} \\ \sigma=\kappa|\kappa'}} (c_\kappa^{n+1} - c_{\kappa'}^{n+1}) \int_{t^n}^{t^{n+1}} \int_\sigma \nabla h(x,t)\cdot \vec{n}_{\kappa\kappa'}\, \varphi(x,t^{n+1})\, d\gamma(x)\, dt \\ &+ \sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} \int_{t^n}^{t^{n+1}} \int_{\partial\kappa\cap\partial\Omega} (c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1})\, g(x,t)\, \varphi(x,t^{n+1})\, d\gamma(x)\, dt \\ &- \sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} \int_{t^n}^{t^{n+1}} \int_\kappa c_\kappa^{n+1}\, \partial_t h(x,t)\, \varphi(x,t^{n+1})\, dx\, dt. \end{aligned}$$

Defining the second auxiliary expression E_2 by

$$\begin{aligned} E_2 \quad &= -\sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} |\kappa| c_\kappa^{n+1} (h_\kappa^{n+1} - h_\kappa^n)\, \varphi(x_\kappa, t^{n+1}) \\ &+ \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int} \\ \sigma=\kappa|\kappa'}} (c_\kappa^{n+1} - c_{\kappa'}^{n+1}) T_{\kappa\kappa'} (h_{\kappa'}^{n+1} - h_\kappa^{n+1}) \frac{1}{|\sigma|} \int_\sigma \varphi(x,t^{n+1})\, d\gamma(x) \\ &+ \sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} \int_{t^n}^{t^{n+1}} \int_{\partial\kappa\cap\partial\Omega} (c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1})\, g(x,t)\, \varphi(x,t^{n+1})\, d\gamma(x)\, dt, \end{aligned}$$

we have

$$\begin{aligned} E_3 - E_2 = &-\sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} c_\kappa^{n+1} \int_{t^n}^{t^{n+1}} \int_\kappa \Big[\partial_t h(x,t)\, \varphi(x,t^{n+1}) \\ &- \frac{(h_\kappa^{n+1} - h_\kappa^n)}{\Delta t}\, \varphi(x_\kappa, t^{n+1}) \Big]\, dx\, dt + \sum_{n=0}^{N_{\Delta t}} \sum_{\substack{\sigma\in\Sigma_{int} \\ \sigma=\kappa|\kappa'}} (c_\kappa^{n+1} - c_{\kappa'}^{n+1}) \\ &\cdot \int_{t^n}^{t^{n+1}} \int_\sigma \left[\nabla h(x,t)\cdot \vec{n}_{\kappa\kappa'} - \frac{h_{\kappa'}^{n+1} - h_\kappa^{n+1}}{d(\kappa,\kappa')} \right] \varphi(x,t^{n+1}) d\gamma(x)\, dt. \end{aligned}$$

Multiplying (2.6) by $\varphi(x_\kappa, t^{n+1})$ and c_κ^{n+1} and integrating it over the time interval (t^n, t^{n+1}) and cell κ yield

$$\text{(A.9)} \qquad \begin{aligned} &\sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} \int_{t^n}^{t^{n+1}} \int_\kappa c_\kappa^{n+1}\, \partial_t h(x,t)\, \varphi(x_\kappa, t^{n+1})\, dx\, dt \\ &\quad - \sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} \int_{t^n}^{t^{n+1}} \int_{\partial\kappa} c_\kappa^{n+1}\, \nabla h(x,t)\cdot \vec{n}_\kappa\, \varphi(x_\kappa, t^{n+1})\, d\gamma(x)\, dt = 0. \end{aligned}$$

Similarly, multiplying (4.1) by c_κ^{n+1} and $\varphi(x_\kappa, t^{n+1})$ and summing the result over $\kappa\in\mathcal{K}$ and $n\in\{0,\dots,N_{\Delta t}\}$, we obtain

$$\text{(A.10)} \qquad \begin{aligned} &\sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} |\kappa| c_\kappa^{n+1} (h_\kappa^{n+1} - h_\kappa^n) \varphi(x_\kappa, t^{n+1}) \\ &\quad + \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} \sum_{\kappa'\in\mathcal{K}_\kappa} T_{\kappa\kappa'} c_\kappa^{n+1} (h_\kappa^{n+1} - h_{\kappa'}^{n+1}) \varphi(x_\kappa, t^{n+1}) \\ &\quad - \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa\cap\partial\Omega|\, g_\kappa^{n+1} c_\kappa^{n+1} \varphi(x_\kappa, t^{n+1}) = 0. \end{aligned}$$

Then, $E_3 - E_2 +$ (A.9) $-$ (A.10) yields the equality

$$E_3 - E_2 = \sum_{n=0}^{N_{\Delta t}} \sum_{\kappa \in \mathcal{K}} c_\kappa^{n+1} \int_{t^n}^{t^{n+1}} \int_\kappa \partial_t h(x,t) \left[\varphi(x_\kappa, t^{n+1}) - \varphi(x, t^{n+1})\right] dx\, dt$$
$$+ \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa \in \mathcal{K}} c_\kappa^{n+1} \sum_{\sigma \in \Sigma_\kappa \cap \Sigma_{\kappa'}} \int_\sigma \left[\frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \nabla h(x,t) \cdot \vec{n}_{\kappa\kappa'}\, dt - \frac{h_{\kappa'}^{n+1} - h_\kappa^{n+1}}{d(\kappa, \kappa')}\right]$$
$$\cdot \left[\varphi(x, t^{n+1}) - \varphi(x_\kappa, t^{n+1})\right] d\gamma(x).$$

Since φ is regular, there exists $C_1 > 0$ depending only on φ such that, for all $\kappa \in \mathcal{K}$ and $x \in \kappa$,

$$|\varphi(x, t^{n+1}) - \varphi(x_\kappa, t^{n+1})| \le C_1\, \delta\mathcal{K}. \tag{A.11}$$

Thanks to the regularity of h, there exists $C_2 > 0$ depending only on $\|h\|_{L^\infty(0,2T;W^{2,\infty}(\Omega))}$, such that, for all $\kappa \in \mathcal{K}$, $\sigma \in \Sigma_\kappa$, $x \in \sigma$, and $t \in (0, 2T)$,

$$\left| \frac{1}{|\sigma|} \int_\sigma \nabla h(u,t) \cdot \vec{n}_\kappa\, d\gamma(u) - \nabla h(x,t) \cdot \vec{n}_\kappa \right| \le C_2\, \delta\mathcal{K}.$$

Thus, the following estimate is derived:

$$|E_3 - E_2| \le C_3\, \delta\mathcal{K}\, \|\partial_t h\|_{L^\infty(\Omega \times [0,2T])} + 2C_1\, \delta\mathcal{K} \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma \in \Sigma_{int} \\ \sigma = \kappa|\kappa'}} \Bigg(|\sigma| C_2 \delta\mathcal{K}$$
$$+ \left| T_{\kappa\kappa'}(h_{\kappa'}^{n+1} - h_\kappa^{n+1}) - \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \int_\sigma \nabla h(u,t) \cdot \vec{n}_{\kappa\kappa'}\, d\gamma(u)\, dt \right| \Bigg).$$

The last term in this estimate is bounded using Cauchy–Schwarz inequality as follows:

$$\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma \in \Sigma_{int} \\ \sigma = \kappa|\kappa'}} \left| T_{\kappa\kappa'}(h_{\kappa'}^{n+1} - h_\kappa^{n+1}) - \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \int_\sigma \nabla h(u,t) \cdot \vec{n}_{\kappa\kappa'}\, d\gamma(u)\, dt \right|$$
$$\le \left[\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma \in \Sigma_{int} \\ \sigma = \kappa|\kappa'}} T_{\kappa\kappa'} \right]^{\frac{1}{2}} \left[\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma \in \Sigma_{int} \\ \sigma = \kappa|\kappa'}} d(\kappa, \kappa')\, |\sigma| \right.$$
$$\left. \left(\frac{h_{\kappa'}^{n+1} - h_\kappa^{n+1}}{d(\kappa, \kappa')} - \frac{1}{\Delta t} \frac{1}{|\sigma|} \int_{t^n}^{t^{n+1}} \int_\sigma \nabla h(u,t) \cdot \vec{n}_{\kappa\kappa'}\, d\gamma(u)\, dt \right)^2 \right]^{\frac{1}{2}}.$$

Finally, using (A.8), (4.5), and the bound $\sum_{\sigma \in \Sigma_{int}} |\sigma| \le \frac{d\, \alpha\, |\Omega|}{\delta\mathcal{K}}$, we obtain that

$$E_3 - E_2 \to 0 \text{ as } \Delta t \to 0.$$

It remains only to prove that $E_2 - E \to 0$ as $\Delta t \to 0$. Removing (A.10) from E yields

$$\begin{aligned} E = &\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} \sum_{\kappa'\in\mathcal{K}_\kappa} \underbrace{T_{\kappa\kappa'}(c_{\kappa\kappa'}^{n+1} - c_\kappa^{n+1})\,(h_\kappa^{n+1} - h_{\kappa'}^{n+1})\,\varphi(x_\kappa, t^{n+1})}_{(F)} \\ &+ \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1}(c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1})\,\varphi(x_\kappa, t^{n+1}) \\ &- \sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} |\kappa| c_\kappa^{n+1}(h_\kappa^{n+1} - h_\kappa^n)\,\varphi(x_\kappa, t^{n+1}). \end{aligned}$$

Thanks to the upstream evaluation of the concentrations at the edges, (F) vanishes if $h_\kappa^{n+1} \geq h_{\kappa'}^{n+1}$. In the opposite case, it is equal to $T_{\kappa\kappa'}(c_{k'}^{n+1} - c_\kappa^{n+1})\,(h_\kappa^{n+1} - h_{\kappa'}^{n+1})\,\varphi(x_\kappa, t^{n+1})$, and thus

$$\begin{aligned} E = &\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}(c_{\kappa'}^{n+1} - c_\kappa^{n+1})\,(h_\kappa^{n+1} - h_{\kappa'}^{n+1})\,\varphi(x_{\kappa\kappa'}, t^{n+1}) \\ &+ \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega|\, g_\kappa^{(+),n+1}(c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1})\,\varphi(x_\kappa, t^{n+1}) \\ &- \sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} |\kappa| c_\kappa^{n+1}(h_\kappa^{n+1} - h_\kappa^n)\,\varphi(x_\kappa, t^{n+1}), \end{aligned}$$

with

$$x_{\kappa\kappa'} = \begin{cases} x_\kappa & \text{if } h_\kappa \leq h_{\kappa'}, \\ x_{\kappa'} & \text{otherwise.} \end{cases}$$

Therefore, $E_2 - E$ writes

$$\begin{aligned} E_2 - E = &\sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'}(c_{\kappa'}^{n+1} - c_\kappa^{n+1})\,(h_\kappa^{n+1} - h_{\kappa'}^{n+1}) \\ &\left[\frac{1}{|\sigma|}\int_\sigma \varphi(x, t^{n+1})\,d\gamma(x) - \varphi(x_{\kappa\kappa'}, t^{n+1})\right] + \sum_{n=0}^{N_{\Delta t}} \sum_{\kappa\in\mathcal{K}} (c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1}) \\ &\int_{t^n}^{t^{n+1}} \int_{\partial\kappa\cap\partial\Omega} \Big[g(x,t)\,\varphi(x, t^{n+1}) - g_\kappa^{(+),n+1}\,\varphi(x_\kappa, t^{n+1})\Big] d\gamma(x)\,dt. \end{aligned}$$

Thanks to the regularity of φ, there exists $C_3 > 0$ depending only on φ such that

$$\left|\frac{1}{|\sigma|}\int_\sigma \varphi(x, t^{n+1})\,d\gamma(x) - \varphi(x_{\kappa\kappa'}, t^{n+1})\right| \leq C_3\,\delta\mathcal{K}. \tag{A.12}$$

Furthermore, since $\varphi \in \mathcal{A}_0^s$, one has

$$\int_{\partial\kappa\cap\partial\Omega} g(x, t^{n+1})\,\varphi(x, t^{n+1}) d\gamma(x) = \int_{\partial\kappa\cap\partial\Omega} g^+(x, t^{n+1})\,\varphi(x, t^{n+1}) d\gamma(x).$$

Finally, inequalities (A.11) and (A.12) and the definition of $g_\kappa^{(+),n+1}$ give the estimate

$$|E_2 - E| \leq C_3\,\delta\mathcal{K} \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\substack{\sigma\in\Sigma_{int}\\ \sigma=\kappa|\kappa'}} T_{\kappa\kappa'} |c_{\kappa'}^{n+1} - c_\kappa^{n+1}|\,|h_\kappa^{n+1} - h_{\kappa'}^{n+1}|$$

$$+ C_1\,\delta\mathcal{K} \sum_{n=0}^{N_{\Delta t}} \Delta t \sum_{\kappa\in\mathcal{K}} |\partial\kappa \cap \partial\Omega||c_\kappa^{n+1} - \tilde{c}_\kappa^{n+1}| g_\kappa^{(+),n+1}.$$

It results from Lemma A.1 that $|E_2 - E| \leq C_4\,\delta\mathcal{K}\,\frac{H}{\sqrt{\delta\mathcal{K}}}$, which ends the proof. □

REFERENCES

[1] R. Anderson and N. Humphrey, *Interaction of weathering and transport processes in the evolution of arid landscapes*, in Quantitative Dynamics Stratigraphy, T. Cross, ed., Prentice–Hall, 1989, Englewood Cliffs, NJ, pp. 349–361.

[2] T. Arbogast, M. F. Wheeler, and N.-Y. Zhang, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media*, SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[3] G. Chavent and J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation*, North–Holland, Amsterdam, 1986.

[4] R. Eymard and T. Gallouët, *Convergence d'un schéma de type éléments finis—volumes finis pour un système couplé elliptique-hyperbolique*, M2AN Math. Model. Numer. Anal., 27 (1993), pp. 843–891.

[5] R. Eymard, T. Gallouët, D. Granjeon, R. Masson, and Q. Tran, *Multi-lithology stratigraphic model under maximum erosion rate constraint*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 527–548.

[6] R. Eymard, T. Gallouët, and R. Herbin, *The Finite Volume Method*, in Handbook of Numerical Analysis, P. Ciarlet and J. Lions, eds., Handb. Numer. Anal. 7, North–Holland, Amsterdam, 2000, pp. 715–1022.

[7] P. Flemings and T. Jordan, *A synthetic stratigraphic model of foreland basin development*, J. Geophysical Research, 94 (1989), pp. 3851–3866.

[8] D. Granjeon, *Modélisation Stratigraphique Déterministe: Conception et Application d'un Modèle Diffusif 3D Multilithologique*, Ph.D. thesis, Géosciences Rennes, Rennes, France, 1997.

[9] D. Granjeon and P. Joseph, *Concepts and applications of a 3D multiple lithology, diffusive model in stratigraphic modelling*, in Numerical Experiments in Stratigraphy, J. Harbaugh and al., eds., Society for Sedimentary Geology Special Publication 62, R. W. Dalrymple, Editor of Special Publications, Tulsa, OK, 1999, pp. 197–210.

[10] P. Kenyon and D. Turcotte, *Morphology of a delta prograding by bulk sediment transport*, Geological Society of America Bulletin, 96 (1985), pp. 1457–1465.

[11] O. Ladyzenskaja, V. Solonnikov, and N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, Transl. Math. Monogr. 23, AMS, Providence, RI, 1968.

[12] D. Peaceman, *Fundamentals of Numerical Reservoirs Simulation*, Elsevier Science, Amsterdam, 1991.

[13] N. Ramarosy, *Application de la Méthode des Volumes Finis à des Problèmes d'Environnement et de Traitement d'Image*, Ph.D. thesis, Université de Paris-Sud, Paris, France, 1999.

[14] J. Rivenaes, *Application of a dual-lithology, depth-dependent diffusion equation in stratigraphic simulation*, Basin Research, 4 (1992), pp. 133–146.

[15] J. Rivenaes, *Impact of sediment transport efficiency on large-scale sequence architecture: Results from stratigraphic computer simulation*, Basin Research, 9 (1997), pp. 91–105.

[16] D. Tetzlaff and J. Harbaugh, *Simulating Clastic Sedimentation*, Van Norstrand Reinhold, New York, 1989.

[17] G. Tucker and R. Slingerland, *Erosional dynamics, flexural isostasy, and long-lived escarpments: A numerical modeling study*, J. Geophysical Research, 99 (1994), pp. 12,229–12,243.

[18] S. Verdière and M. Vignal, *Numerical and theoretical study of a dual mesh method using finite volume schemes for two phase flow problems in porous media*, Numer. Math., 80 (1998), pp. 601–639.

[19] M. Vignal, *Convergence of finite volume scheme for a system of an elliptic equation and a hyperbolic equation*, M2AN Math. Model. Numer., 30 (1996), pp. 841–872.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 502–520

DOMAIN DECOMPOSITION SPECTRAL APPROXIMATIONS FOR AN EIGENVALUE PROBLEM WITH A PIECEWISE CONSTANT COEFFICIENT*

M. S. MIN† AND D. GOTTLIEB†

Abstract. Consider a model eigenvalue problem with a piecewise constant coefficient. We split the domain at the discontinuity of the coefficient function and define the multidomain variational formulation for the eigenproblem. The discrete multidomain variational formulations are defined for Legendre–Galerkin and Legendre-collocation methods. The spectral rate of convergence of the approximate eigensolutions is proven for the Legendre–Galerkin method. The minmax principle is used for the convergence analysis.

The Legendre-collocation, Chebyshev-collocation, Legendre-collocation penalty, and Chebyshev-collocation penalty methods are also defined by using the multidomain approach, and their numerical results applied to the eigenproblem are demonstrated. The spectral convergence for the eigenvalues and eigenfunctions is confirmed for all the multidomain spectral techniques presented here.

Key words. discontinuous problems, multidomain variational formulation, minmax principle, domain decomposition, Legendre–Galerkin method, Legendre-collocation method, Legendre-collocation penalty method, Chebyshev-collocation method, Chebyshev-collocation penalty method

AMS subject classifications. 41A10, 41A25

DOI. 10.1137/S0036142903423836

1. Introduction.

We consider Maxwell's equations governing the electromagnetic wave propagation in periodically structured dielectric arrays cast as an eigenvalue problem. The dielectric function corresponding to the periodic arrays is represented by a periodically piecewise constant function.

The electromagnetic wave propagation in a periodic dielectric medium was first studied by Rayleigh in 1887, identifying the fact that there exists a narrow frequency gap prohibiting light propagation through one-dimensional periodic twinning planes. A hundred years later, the concepts of omnidirectional forbidden frequency gaps in two and three dimensions were introduced, leading to many subsequent developments in the fabrication, theory, and application of electromagnetic wave propagation to optical fibers [14], [3], [15], [23]. Computation has become a primary tool for carrying out frequency gap calculation for various periodic dielectric structures.

Numerous numerical studies have focused on predicting the forbidden eigenfrequencies accurately by solving Maxwell's equations in the frequency domain [1], [6], [17], [16]. However, numerical analysis has been lacking, and high-order methods have not been applied to such problems yet.

In [18], Fourier–Galerkin and Fourier-collocation methods are applied to a single domain, and their theoretical and numerical convergence studies for the eigensolutions are demonstrated. As a result of the presence of the discontinuity in the coefficient function in a single domain, the solution is only in H_p^2, and the rates of convergence of the eigensolutions by Fourier methods are between *second order* and *third order*.

*Received by the editors February 26, 2003; accepted for publication (in revised form) July 29, 2004; published electronically June 30, 2005. This work was supported by AFOSR grant F49620-02-1-0113, NSF grant DMS-020745, and HPC/F33615-01-C-1866.

http://www.siam.org/journals/sinum/43-2/42383.html

†Division of Applied Mathematics, Brown University, Providence, RI (msmin@cfm.brown.edu, dig@cfm.brown.edu).

In this paper, we apply domain decomposition techniques for spectral methods. We obtain spectrally accurate eigensolutions by using multidomain Legendre and Chebyshev approximations. Implementations in two dimensions are extended in [19].

In the multidomain approach, we split the domain into subdomains in order for the discontinuous coefficient function to be smooth in each subdomain, so that the solutions are infinitely smooth in each subdomain. Then we reformulate the problem in multidomain variational form. The finite-dimensional space for each subdomain is defined by the Legendre polynomials of finite degrees, and boundary and interface conditions are imposed strongly for the Legendre–Galerkin, Legendre-collocation, and Chebyshev-collocation methods and weakly for the Legendre-collocation penalty and Chebyshev-collocation penalty methods.

We restrict the penalty parameter to a specific one in this paper and leave the study for the proper range of the parameters to future work. Convergence analysis for the eigenvalues and eigenfunctions is carried out for the Legendre–Galerkin method. For the collocation cases, two different methods are introduced by choosing two different test spaces for the same trial space.

The numerical results for Legendre–Galerkin, Legendre-collocation, Legendre-collocation penalty, Chebyshev-collocation, and Chebyshev-collocation penalty methods presented here show a spectral rate of convergence for the eigensolutions. In terms of accuracy, the results of the Legendre–Galerkin and the Legendre-collocation methods, which use the same space for the test and trial spaces, are comparable and more accurate than the results of the Legendre-collocation penalty and the Legendre-collocation methods, which use different spaces for the trial and test space. The penalty method is favorable because of the simplicity in implementation for the same magnitude of accuracy.

We organize this paper as follows. In section 2 we reformulate the eigenproblem into a multidomain variational formulation. We recall the minmax principle to characterize the lth eigenvalue by minimizing the maximum of the Rayleigh quotient over l-dimensional subspaces. In section 3 we present the finite-dimensional space used for the approximate solution. The procedure to find the basis for the finite approximant space is shown. In section 4 we define the multidomain variational formulation for the Legendre–Galerkin method. We provide a convergence analysis for the eigenvalues and eigenfunctions. The theory is confirmed by numerical results. In section 5 two different Legendre-collocation methods are defined, based on the test space chosen. The numerical results for eigensolutions by those methods show a spectral rate of convergence. In section 6, we discuss the Legendre-collocation penalty method by defining a multidomain variational formulation with the penalty approach [7] for the boundary and interface constraints. Chebyshev-collocation approximations are also tested and their numerical results presented. Section 7 discusses the asymptotic behavior of the largest approximate eigenvalues. Section 8 gives a brief conclusion.

2. The multidomain variational formulation.

The source-free Maxwell equations describing the transverse-magnetic mode in one-dimensional periodic media can be cast as the following generalized eigenvalue problem: find λ and u in $H_p^2(-\pi,\pi)$ (where p stands for periodic), such that

$$-u'' = \lambda \epsilon(x) u, \tag{2.1}$$

where $\epsilon(x) = 1$ in $(-\pi, 0)$ and $\epsilon(x) = \omega^2$ in $[0, \pi)$, $\omega \neq 1$. The function u represents the electric field pattern, and the dielectric function $\epsilon(x)$ describes a unit cell from a multilayer structure with 2π-periodicity. This problem was considered in [18].

Recall the variational formulation of (2.1) from [18]: Find λ and $u \in H^1_p(-\pi,\pi)$ such that

$$a(u,v) = \lambda(u,v) \quad \text{for } v \in H^1_p(-\pi,\pi), \tag{2.2}$$

where

$$a(u,v) = \int_{-\pi}^{\pi} u'v'dx \quad \text{and} \quad (u,v) = \int_{-\pi}^{\pi} u\,v\,\epsilon\,dx. \tag{2.3}$$

Since $a(u,v)$ is Hermitian, the eigenvalue can be characterized by the following two statements from [9], [18], [20], [21].

THEOREM 2.1. *Let λ_l denote the eigenvalues of* (2.1)*, and let S_l be any l-dimensional subspace of $H^1_p(\Omega)$. Then, for $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_l \cdots$,*

$$\lambda_l = \min_{S_l \subset H^1_p(\Omega)} \max_{v \in S_l} \frac{a(v,v)}{(v,v)}. \tag{2.4}$$

We also recall the following lemma from [9] and [18].

LEMMA 2.2. *Let λ_i be arranged in ascending order, and define*

$$E_{i,j} = span\{u_i, \ldots, u_j\}, \tag{2.5}$$

where u_i is the eigenfunction corresponding to the eigenvalue λ_i. Then

$$\lambda_l = \max_{v \in E_{k,l}} \frac{a(v,v)}{(v,v)}, \quad k \leq l, \tag{2.6}$$

$$\lambda_l = \min_{v \in E_{l,m}} \frac{a(v,v)}{(v,v)}, \quad l \leq m. \tag{2.7}$$

It is natural here to consider splitting the domain. Denote the domain by $\Omega = (-\pi,\pi)$ and divide it into two subdomains, say, $\Omega_- = (-\pi,0)$ and $\Omega_+ = (0,\pi)$, so that $\epsilon(x)$ is smooth in each subdomain. Denote the restrictions by $u_- = u\big|_{\Omega_-}$ and $u_+ = u\big|_{\Omega_+}$, which are distributional solutions to the given equation (2.1). Integrate by parts in Ω_1 and Ω_2, respectively, and define the following bilinear forms:

$$a(u,v)_- = \int_{-\pi}^{0} u'_-v'_-dx + u'_-v_-\Big|_0^{-\pi}, \qquad (u,v)_- = \int_{-\pi}^{0} u_-v_-\epsilon_-dx, \tag{2.8}$$

$$a(u,v)_+ = \int_{0}^{\pi} u'_+v'_+dx + u'_+v_+\Big|_\pi^{0}, \qquad (u,v)_+ = \int_{0}^{\pi} u_+v_+\epsilon_+dx. \tag{2.9}$$

Remark 1. It is clear to see that, for $u, v \in H^1_p(\Omega)$,

$$a(u,v) = a(u,v)_- + a(u,v)_+, \tag{2.10}$$

$$(u,v) = (u,v)_- + (u,v)_+. \tag{2.11}$$

3. Finite-dimensional subspace. In this section we present the finite-dimensional space used in our approximation. Denote by $\mathcal{P}_N = \text{span}\{L_k(\xi), \xi \in [-1,1]\}$ the space of Legendre polynomials of degree at most N. We define the local variables x_- and x_+ by

$$x_-(\xi) = \frac{\pi}{2}(\xi - 1) \quad \text{and} \quad x_+(\xi) = \frac{\pi}{2}(\xi + 1). \tag{3.1}$$

The approximation space V_{2N-2} is the $(2N-2)$-dimensional space defined by

$$V_{2N-2} = \{\phi \in H_p^2(\Omega) : \phi_-(x_-(\xi)) \in \mathcal{P}_N \ \text{ and } \ \phi_+(x_+(\xi)) \in \mathcal{P}_N\}. \tag{3.2}$$

To apply the Galerkin approximation, we need to find a basis of V_{2N-2}. This is done as follows.

Let ϕ in V_{2N-2} be expressed by

$$\phi_- = \phi\big|_{x_-(\xi)} = \sum_{k=0}^{N} c_k L_k(\xi), \tag{3.3}$$

$$\phi_+ = \phi\big|_{x_+(\xi)} = \sum_{k=0}^{N} d_k L_k(\xi). \tag{3.4}$$

Since ϕ and ϕ' are continuous at $x = 0$ and 2π-periodic, we apply the following conditions:

$$\phi_-(0) = \phi_+(0), \ \ \phi_-(-\pi) = \phi_+(\pi), \tag{3.5}$$

$$\phi'_-(0) = \phi'_+(0), \ \ \phi'_-(-\pi) = \phi'_+(\pi). \tag{3.6}$$

Letting $\alpha_k = L'_k(1) = \frac{k(k+1)}{2}$, and applying the boundary and the interface conditions (3.5) and (3.6), we obtain the following relations for the coefficients:

$$c_N = \sum_{k=1}^{\frac{N-2}{2}} \left[-\frac{(\alpha_N + \alpha_{2k})}{2\alpha_N} c_{2k} + \frac{(\alpha_N - \alpha_{2k})}{2\alpha_N} d_{2k} \right], \tag{3.7}$$

$$d_N = \sum_{k=1}^{\frac{N-2}{2}} \left[\ \frac{(\alpha_N - \alpha_{2k})}{2\alpha_N} c_{2k} - \frac{(\alpha_N + \alpha_{2k})}{2\alpha_N} d_{2k} \right], \tag{3.8}$$

$$c_{N-1} = \sum_{k=0}^{\frac{N-4}{2}} \left[-\frac{(\alpha_{N-1} + \alpha_{2k+1})}{2\alpha_{N-1}} c_{2k+1} - \frac{(\alpha_{N-1} - \alpha_{2k+1})}{2\alpha_{N-1}} d_{2k+1} \right], \tag{3.9}$$

$$d_{N-1} = \sum_{k=0}^{\frac{N-4}{2}} \left[-\frac{(\alpha_{N-1} - \alpha_{2k+1})}{2\alpha_{N-1}} c_{2k+1} - \frac{(\alpha_{N-1} + \alpha_{2k+1})}{2\alpha_{N-1}} d_{2k+1} \right]. \tag{3.10}$$

For simplicity, here we use the following notation:

$$\beta_k = \frac{\alpha_{N_k} + \alpha_k}{2\alpha_{N_k}} \qquad \text{and} \qquad \gamma_k = (-1)^k \frac{\alpha_{N_k} - \alpha_k}{2\alpha_{N_k}}, \tag{3.11}$$

$$N_k = \begin{cases} \ \ N & \text{for even } k, \\ N-1 & \text{for odd } k. \end{cases} \tag{3.12}$$

Then, substituting (3.7)–(3.10) into (3.3)–(3.4), we get

$$\phi_- = \phi\big|_{x_-(\xi)} = \sum_{k=0}^{N-2} c_k \{L_k(\xi) - \beta_k L_{N_k}(\xi)\} + \sum_{k=0}^{N-2} d_k \gamma_k L_{N_k}(\xi), \tag{3.13}$$

$$\phi_+ = \phi\big|_{x_+(\xi)} = \sum_{k=0}^{N-2} c_k \gamma_k L_{N_k}(\xi) + \sum_{k=0}^{N-2} d_k \{L_k(\xi) - \beta_k L_{N_k}(\xi)\}. \tag{3.14}$$

Thus, one can easily see the basis for V_{2N-2} given by

$$\{\phi_k, \psi_k\} \quad \text{for } 0 \le k \le N-2, \tag{3.15}$$

where the basis functions are defined by

$$(\phi_k)_- = L_k(\xi) - \beta_k L_{N_k}(\xi), \quad (\phi_k)_+ = \gamma_k L_{N_k}(\xi), \tag{3.16}$$

$$(\psi_k)_- = \gamma_k L_{N_k}(\xi), \quad (\psi_k)_+ = L_k(\xi) - \beta_k L_{N_k}(\xi). \tag{3.17}$$

Now, we are ready to construct the multidomain Legendre–Galerkin scheme in the following section.

4. Legendre–Galerkin method.

Find λ^N, $u^N \in V_{2N-2}$ such that

$$a(u^N, v^N) = \lambda^N (u^N, v^N) \quad \text{for all } v^N \in V_{2N-2}. \tag{4.1}$$

From the relations (2.10) and (2.11), the two inner products can be expressed by

$$a(u^N, v^N) = \int_{-\pi}^{0} (u_-^N)'(v_-^N)' dx + \int_0^{\pi} (u_+^N)'(v_+^N)' dx = \int_{-\pi}^{\pi} (u^N)'(v^N)' dx, \tag{4.2}$$

$$(u^N, v^N) = \int_{-\pi}^{0} (u_-^N)(v_-^N)(\epsilon_-) dx + \int_0^{\pi} (u_+^N)(v_+^N)(\epsilon_+) dx = \int_{-\pi}^{\pi} u^N v^N \epsilon\, dx. \tag{4.3}$$

4.1. Numerical scheme and its results.

The approximate eigenfunction $u^N \in V_{2N-2}$ can be expanded by the basis found in section 3 with an unknown set of $(2N-2)$ coefficients:

$$u^N = \sum_{k=0}^{N-2} [(\hat{u}_\phi^N)_k \phi_k + (\hat{u}_\psi^N)_k \psi_k]. \tag{4.4}$$

Take $v^N = \phi_n$ for $0 \le n \le N-2$, and substitute u^N in the form (4.4) to the variational formulation (4.2)–(4.3). We obtain the Legendre–Galerkin scheme as follows:

$$K \hat{\mathbf{u}}^N = \lambda^N M \hat{\mathbf{u}}^N. \tag{4.5}$$

The following are defined for the notation in scheme (4.5):

$$K = \begin{bmatrix} K_-^- & K_+^- \\ K_-^+ & K_+^+ \end{bmatrix}, \quad M = \begin{bmatrix} M_-^- & M_+^- \\ M_-^+ & M_+^+ \end{bmatrix}, \quad \text{and} \quad \hat{\mathbf{u}}^N = \begin{bmatrix} \hat{u}_\phi^N \\ \hat{u}_\psi^N \end{bmatrix}, \tag{4.6}$$

where

$$(K_-^-)_{k,n} = \int_{-\pi}^{\pi} (\phi_k)'(\phi_n)' dx, \quad (K_+^-)_{k,n} = \int_{-\pi}^{\pi} (\phi_k)'(\psi_n)' dx,$$
$$(K_-^+)_{k,n} = \int_{-\pi}^{\pi} (\psi_k)'(\phi_n)' dx, \quad (K_+^+)_{k,n} = \int_{-\pi}^{\pi} (\psi_k)'(\psi_n)' dx,$$

$$(M_-^-)_{k,n} = \int_{-\pi}^{\pi} \phi_k \phi_n \epsilon(x) dx, \quad (M_+^-)_{k,n} = \int_{-\pi}^{\pi} \phi_k \psi_n \epsilon(x) dx,$$
$$(M_-^+)_{k,n} = \int_{-\pi}^{\pi} \psi_k \phi_n \epsilon(x) dx, \quad (M_+^+)_{k,n} = \int_{-\pi}^{\pi} \psi_k \psi_n \epsilon(x) dx,$$

Table 1

Relative errors of eigenvalues for $\omega = 2$ and the discrete l_2-errors of $u_i - u_i^N$ for the multidomain Legendre–Galerkin method.

λ_i	N	$(\lambda_i^N - \lambda_i)/\lambda_i$	$\|u_i - u_i^N\|$
0.369875	4	2.94(-03)	2.61(-03)
	8	7.04(-10)	4.38(-07)
	16	6.15(-15)	1.03(-15)
	32	7.38(-14)	3.49(-15)
	64	6.40(-13)	2.55(-14)
0.536233	4	4.39(-04)	8.77(-03)
	8	1.63(-10)	2.38(-06)
	16	2.89(-14)	2.45(-14)
	32	3.43(-14)	3.60(-14)
	64	9.31(-14)	1.42(-13)
1.607115	4	3.55(-02)	1.71(-01)
	8	1.75(-06)	4.16(-04)
	16	2.76(-16)	5.43(-11)
	32	5.94(-15)	4.82(-14)
	64	3.96(-14)	2.85(-13)
1.937181	4	3.28(-01)	1.05(-01)
	8	1.45(-04)	4.71(-04)
	16	6.07(-15)	2.31(-10)
	32	2.64(-14)	8.45(-15)
	64	2.84(-13)	8.09(-14)

and

$$\hat{u}_\phi^N = [(\hat{u}_\phi^N)_0, (\hat{u}_\phi^N)_1, \ldots, (\hat{u}_\phi^N)_{N-2}]^T \quad \text{and} \quad \hat{u}_\psi^N = [(\hat{u}_\psi^N)_0, (\hat{u}_\psi^N)_1, \ldots, (\hat{u}_\psi^N)_{N-2}]^T.$$

Now, we solve the generalized matrix eigenproblem (4.5) numerically and obtain the approximate $l(\leq 2N-2)$th eigenvalues, λ_l^N, and the set of orthogonal vectors $\hat{\mathbf{u}}^N$, which approximates the lth eigenfunction u_l as the coefficients in the expansion of the basis of V_{2N-2}. In Table 1, the relative errors for $\lambda_l^N - \lambda_l$ and the discrete l_2-errors of $u_l - u_l^N$ as N increases are provided for the first few eigenvalues in an ascending order and the associated eigenfunctions. The numerical results demonstrate that the errors decay exponentially as N increases.

4.2. Error estimates for eigenvalues and eigenfunctions. We show the error estimates for the approximate eigenvalues and eigenfunctions for the multidomain Legendre–Galerkin method.

We first treat the approximate eigenvalues. Let $P_N u$ be defined by

$$P_N u = \sum_{k=0}^{N-2} [(\hat{u}_\phi)_k \phi_k + (\hat{u}_\psi)_k \psi_k], \tag{4.7}$$

where the expansion coefficients $(\hat{u}_\phi)_k$ and $(\hat{u}_\psi)_k$ will be defined later in this section.

From the minmax principle [9], [21], we can characterize the eigenvalue for the multidomain Legendre–Galerkin procedure by

$$\lambda_l^N = \min_{S_l \subset V_{2N-2}} \max_{v \in S_l} \frac{a(v,v)}{(v,v)}. \tag{4.8}$$

Lemma 4.1. *Let λ_l^N be the approximation to λ_l as obtained by the Legendre–Galerkin procedure* (4.5)*, and let $P_N u$ be defined as in* (4.7)*. Then*

$$\lambda_l \leq \lambda_l^N \leq \lambda_l \max_{v\in E_{1,l}} \frac{a(P_N v, P_N v)}{a(v,v)} \max_{v\in E_{1,l}} \frac{(v,v)}{(P_N v, P_N v)}. \tag{4.9}$$

Proof. Since V_{2N-2} is a subspace of $H_p^1(\Omega)$, it is true that $\lambda_l \leq \lambda_l^N$. Now, let $PE_{1,l}$ be spanned by $P_N u_1, \ldots, P_N u_l$. For simplicity, we denote $Pu = P_N u$. Clearly $PE_{1,l}$ is the l-dimensional subspace of V_{2N-2}. Using the minmax principle, we have

$$\begin{aligned}\lambda_l^N &\leq \max_{v\in PE_{1,l}} \frac{a(v,v)}{(v,v)}\\ &= \max_{v\in E_{1,l}} \frac{a(Pv,Pv)}{(Pv,Pv)}\\ &= \max_{v\in E_{1,l}} \frac{a(v,v)}{(v,v)}\frac{a(Pv,Pv)}{a(v,v)}\frac{(v,v)}{(Pv,Pv)}.\end{aligned}$$

From Lemma 2.2, the proof follows. □

Lemma 4.2. *For $u_{i=1,\ldots,l} \in H_p^1(\Omega)$, where $(u_i)_- \in H^m(\Omega_-)$ and $(u_i)_+ \in H^m(\Omega_+)$,*

$$\max_{v\in E_{1,l}} \frac{(v,v)}{(Pv,Pv)} \leq 1 + C(l)N^{-m}, \tag{4.10}$$

where the constant $C(l)$ is independent of N.

Proof. We follow the procedure in [18]. For $v = \sum_{i=1}^l \mu_i u_i$ in $E_{1,l}$, we have

$$\begin{aligned}\frac{(v,v)-(Pv,Pv)}{(v,v)} &\leq \frac{2|(v, v-Pv)|}{(v,v)}\\ &\leq \frac{2\sum_{i,j=1}^l |\mu_i||\mu_j||(u_i - Pu_i, u_j)|}{\sum_{i=1}^l |\mu_i|^2}\\ &= 2l \max_{i,j=1,\ldots,l} |(u_i - Pu_i, u_j)|.\end{aligned}$$

For the last term above, we have

$$|(u_i - Pu_i, u_j)| \leq |(u_i - Pu_i, u_j)_-| + |(u_i - Pu_i, u_j)_+|. \tag{4.11}$$

Now consider an eigenfunction $u = u_i$ and its projection Pu onto the space V_{2N-2}. Since

$$Pu_- = \sum_{k=0}^{N-2} \left[(\hat{u}_\phi)_k(\phi_k)_- + (\hat{u}_\psi)_k(\psi_k)_-\right], \tag{4.12}$$

$$Pu_+ = \sum_{k=0}^{N-2} \left[(\hat{u}_\phi)_k(\phi_k)_+ + (\hat{u}_\psi)_k(\psi_k)_+\right], \tag{4.13}$$

we can rewrite them in terms of Legendre polynomials as follows:

$$Pu_- = \sum_{k=0}^{N-2} (\hat{u}_\phi)_k L_k + c_{N-1}L_{N-1} + c_N L_N, \tag{4.14}$$

$$Pu_+ = \sum_{k=0}^{N-2} (\hat{u}_\psi)_k L_k + d_{N-1}L_{N-1} + d_N L_N, \tag{4.15}$$

where

$$c_{N-1} = \sum_{k=0}^{\frac{N-4}{2}} \left[-\beta_{2k+1}(\hat{u}_\phi)_{2k+1} + \gamma_{2k+1}(\hat{u}_\psi)_{2k+1}\right],$$

$$c_N = \sum_{k=1}^{\frac{N-2}{2}} \left[-\beta_{2k}(\hat{u}_\phi)_{2k} + \gamma_{2k}(\hat{u}_\psi)_{2k}\right],$$

$$d_{N-1} = \sum_{k=0}^{\frac{N-4}{2}} \left[\gamma_{2k+1}(\hat{u}_\phi)_{2k+1} - \beta_{2k+1}(\hat{u}_\psi)_{2k+1}\right],$$

$$d_N = \sum_{k=1}^{\frac{N-2}{2}} \left[\gamma_{2k}(\hat{u}_\phi)_{2k} - \beta_{2k}(\hat{u}_\psi)_{2k}\right].$$

Now we identify

$$(\hat{u}_\phi)_k = \frac{2k+1}{2}\int_{-1}^{1} u_-(x_-(\xi))L_k(\xi)d\xi, \tag{4.16}$$

$$(\hat{u}_\psi)_k = \frac{2k+1}{2}\int_{-1}^{1} u_+(x_+(\xi))L_k(\xi)d\xi, \tag{4.17}$$

which are exactly the Legendre coefficients for u_- and u_+, respectively. For clarity, we replace the notation $(\hat{u}_\phi)_k$ by $(\hat{u}_-)_k$, and similarly $(\hat{u}_\psi)_k$ by $(\hat{u}_+)_k$. Then, considering an eigenfunction u_i, the expansion coefficients of Pu_i are denoted by $(\hat{u}_{i_-})_k$ and $(\hat{u}_{i_+})_k$. Then we have

$$\begin{aligned}
&|(u_i - Pu_i, u_j)_-| \\
&= \left| \left(\sum_{k\geq N-1}^{\infty} (\hat{u}_{i_-})_k L_k - c_{N-1}L_{N-1} - c_N L_N, \sum_{n=0}^{\infty} (\hat{u}_{j_-})_n L_n \right)_- \right| \\
&\leq \frac{\pi}{2} \sum_{k\geq N-1}^{\infty} |(\hat{u}_{i_-})_k||(\hat{u}_{j_-})_k| \int_{-1}^{1} L_k^2 d\xi \\
&+ \frac{\pi}{2}|c_{N-1}||(\hat{u}_{j_-})_{N-1}| \int_{-1}^{1} L_{N-1}^2 d\xi + \frac{\pi}{2}|c_N||(\hat{u}_{j_-})_N| \int_{-1}^{1} L_N^2 d\xi = RHS(1).
\end{aligned}$$

We examine the two terms $|c_{N-1}|$ and $|c_N|$. From the Cauchy–Schwarz inequality, we have

$$|c_{N-1}| \leq \left\{ \sum_{k=0}^{\frac{N-4}{2}} (|\beta_{2k+1}|^2 + |\gamma_{2k+1}|^2) \right\}^{1/2} \left\{ \sum_{k=0}^{\frac{N-4}{2}} (|(\hat{u}_{i_-})_{2k+1}|^2 + |(\hat{u}_{i_+})_{2k+1}|^2) \right\}^{1/2},$$

$$|c_N| \leq \left\{ \sum_{k=1}^{\frac{N-2}{2}} (|\beta_{2k}|^2 + |\gamma_{2k}|^2) \right\}^{1/2} \left\{ \sum_{k=1}^{\frac{N-2}{2}} (|(\hat{u}_{i_-})_{2k}|^2 + |(\hat{u}_{i_+})_{2k}|^2) \right\}^{1/2}.$$

Since $|\beta_k|$, $|\gamma_k| < 1$, and $|(\hat{u}_{j_-})_k|$, $|(\hat{u}_{j_+})_k|$ decay like $O(k^{-m})$ in [4] and [5], it is clear that $|c_{N-1}|$ and $|c_N|$ are bounded by $O(N)$. Since $\int_{-1}^{1} L_k^2(\xi)d\xi = \frac{2}{2k+1}$, the second term of $RHS(1)$ which is the leading term decays like $O(N^{-m})$. Therefore we have

$$|(u_i - Pu_i, u_j)_-| \leq CN^{-m}. \tag{4.18}$$

Similarly, we get

$$|(u_i - Pu_i, u_j)_+| \leq CN^{-m}. \tag{4.19}$$

This completes the proof. □

LEMMA 4.3. *For $u_{i=1,\dots,l} \in H^1_p(\Omega)$, where $(u_i)_- \in H^m(\Omega_-)$ and $(u_i)_+ \in H^m(\Omega_+)$,*

$$\max_{v \in E1,l} \frac{a(Pv,Pv)}{a(v,v)} \leq 1 + C(l)N^{-m}, \tag{4.20}$$

where the constant $C(l)$ is independent of N.

Proof. Since

$$\frac{a(Pv,Pv)}{a(v,v)} = 1 - \frac{a(v,v) - a(Pv,Pv)}{a(v,v)}, \tag{4.21}$$

we examine the convergency of the last term of (4.21). Following the similar procedure as in Lemma 4.2, we obtain

$$\begin{aligned} |a(u_i - Pu_i, u_j)_-| &= \left| a\left(\sum_{k \geq N-1}^{\infty} (\hat{u}_{i_-})_k L_k - c_{N-1}L_{N-1} - c_N L_N, \sum_{n=0}^{\infty} (\hat{u}_{j_-})_n L_n \right)_- \right| \\ &\leq \frac{\pi}{2} \sum_{k \geq N-1}^{\infty} \sum_{n=0}^{\infty} |(\hat{u}_{i_-})_k||(\hat{u}_{j_-})_k| \int_{-1}^{1} L'_k L'_n d\xi \\ &+ \frac{\pi}{2} |c_{N-1}||(\hat{u}_{j_-})_{N-1}| \sum_{n=0}^{\infty} \int_{-1}^{1} L'_{N-1} L'_n d\xi \\ &+ \frac{\pi}{2} |c_N||(\hat{u}_{j_-})_N| \sum_{n=0}^{\infty} \int_{-1}^{1} L'_N L'_n d\xi = RHS(2). \end{aligned}$$

Since the leading term of $RHS(2)$ decays like $O(N^{-m})$, we have

$$|a(u_i - Pu_i, u_j)_-| \leq CN^{-m}. \tag{4.22}$$

Similarly, we get

$$|a(u_i - Pu_i, u_j)_+| \leq CN^{-m}. \tag{4.23}$$

This completes the proof. □

As consequences of Lemmas 4.2 and 4.3, we have the following theorems.

THEOREM 4.4. *For $u_{i=1,\dots,l} \in H^1_p(\Omega)$, where $(u_i)_- \in H^m(\Omega_-)$ and $(u_i)_+ \in H^m(\Omega_+)$, let λ^N_l be the lth eigenvalue obtained by the multidomain Legendre–Galerkin approximation from (4.1) to the eigenvalue λ_l. Then*

$$|\lambda^N_l - \lambda_l| \leq C(l)N^{-m}, \tag{4.24}$$

where $C(l)$ is independent of N.

For the approximate eigenvectors, we can state the following.

THEOREM 4.5. *For $u_{i=1,\dots,l} \in H^1_p(\Omega)$, where $(u_i)_- \in H^m(\Omega_-)$ and $(u_i)_+ \in H^m(\Omega_+)$, let u_l^N be the lth eigenfunction of the multidomain Legendre–Galerkin approximation* (4.1) *to the eigenfunction u_l. Then*

$$||u_l - u_l^N|| \le C(l)N^{-m}, \tag{4.25}$$

where $C(l)$ is independent of N.

The proof follows the same way as in [21].

5. Legendre-collocation methods. The Legendre–Gauss–Lobatto points ξ_i are defined by

$$\xi_0 = -1,\ \ \xi_N = 1,\ \ \xi_i (i = 1, \dots, N-1) \text{ zeros of } L'_N, \tag{5.1}$$

and the Legendre–Gauss–Lobatto weights are

$$w_i = \frac{2}{N(N+1)} \frac{1}{[L_N(\xi_i)]^2}. \tag{5.2}$$

Denoting

$$(u_-)_i = u\big|_{(x_-)_i} \qquad \text{for } (x_-)_i = \frac{\pi}{2}(\xi_i - 1), \tag{5.3}$$

$$(u_+)_i = u\big|_{(x_+)_i} \qquad \text{for } (x_+)_i = \frac{\pi}{2}(\xi_i + 1), \tag{5.4}$$

we define two discrete bilinear forms that approximate $a(u,v)_-$ and $a(u,v)_+$:

$$a(u,v)_{h-} = \sum_{i=0}^{N} (u_-)'_i (v_-)'_i w_i + (u_-)'_0 (v_-)_0 - (u_-)'_N (v_-)_N, \tag{5.5}$$

$$a(u,v)_{h+} = \sum_{i=0}^{N} (u_+)'_i (v_+)'_i w_i + (u_+)'_0 (v_+)_0 - (u_+)'_N (v_+)_N. \tag{5.6}$$

To approximate $(u,v)_-$ and $(u,v)_+$, define

$$(u,v)_{h-} = \sum_{i=0}^{N} (u_-)_i (v_-)_i (\epsilon_-)_i w_i, \tag{5.7}$$

$$(u,v)_{h+} = \sum_{i=0}^{N} (u_+)_i (v_+)_i (\epsilon_+)_i w_i. \tag{5.8}$$

It is natural to define the following discrete bilinear forms approximating the continuous bilinear forms $a(u,v)$ and (u,v):

$$a(u,v)_h = a(u,v)_{h-} + a(u,v)_{h+}, \tag{5.9}$$

$$(u,v)_h = (u,v)_{h-} + (u,v)_{h+}. \tag{5.10}$$

Now, we state the multidomain discrete variational formulation of (2.2): Find λ^c and u^c in V_{2N-2} such that

$$a(u^c, v)_h = \lambda^c (u^c, v)_h \qquad \text{for } \ v \in V, \tag{5.11}$$

where V is a suitable space that will be specified later. In the following subsections, we introduce two different Legendre-collocation methods by taking the space V differently.

5.1. Legendre-collocation method 1. Our first Legendre-collocation method takes the space V_{2N-2} as a test space. Find λ^c and u^c in V_{2N-2} such that

$$a(u^c, v^c)_h = \lambda^c (u^c, v^c)_h \text{ for } v^c \in V_{2N-2}. \tag{5.12}$$

To construct the scheme, we expand

$$(u_-^c)_i = \sum_{j=0}^{N} (u_-^c)_j l_j(\xi_i), \tag{5.13}$$

$$(u_+^c)_i = \sum_{j=0}^{N} (u_+^c)_j l_j(\xi_i), \tag{5.14}$$

where the Lagrange interpolation polynomials of degree N based on the Legendre–Gauss–Lobatto points [8], [10] are

$$l_j(\xi) = -\frac{1}{N(N+1)} \frac{(1-\xi^2)L_N'(\xi)}{(\xi-\xi_j)L_N(\xi_j)}. \tag{5.15}$$

Take $v^c = \phi_n(x)(0 \le n \le N-2)$, which is the basis for V_{2N-2}, and substitute u_-^c, u_+^c, and v^c in (5.5) and (5.6). Applying the continuity and the periodicity for u^c, that is, $(u_-^c)_0 = (u_+^c)_N$ and $(u_-^c)_N = (u_+^c)_0$, we get

$$\begin{aligned} a(u^c, v^c)_h = {} & (u_-^c)_0 \left(\sum_{i=0}^{N} [D_{i0} (\phi_{n_-})_i' w_0 + D_{iN} (\phi_{n_-})_i' w_N] \right) \\ & + (u_+^c)_0 \left(\sum_{i=0}^{N} [D_{i0} (\phi_{n_+})_i' w_0 + D_{iN} (\phi_{n_+})_i' w_N] \right) \\ & + \sum_{j=1}^{N-1} (u_-^c)_j \sum_{i=0}^{N} D_{ij} (\phi_{n_-})_i' w_i + \sum_{j=1}^{N-1} (u_+^c)_j \sum_{i=0}^{N} D_{ij} (\phi_{n_+})_i' w_i, \end{aligned}$$

where $D_{ij} = l_j'(\xi_i)$ is the differentiation matrix of Lagrange polynomials based on Legendre–Gauss–Lobatto points [10], [13]. Similarly, we have

$$\begin{aligned} (u^c, v^c)_h = {} & (u_-^c)_0 [(\phi_{0_-})_0 (\epsilon_-)_0 w_0 + (\phi_{0_+})_N (\epsilon_+)_N w_N] \\ & + (u_+^c)_0 [(\phi_{0_+})_0 (\epsilon_+)_0 w_0 + (\phi_{0_-})_N (\epsilon_-)_N w_N] \\ & + \sum_{j=1}^{N-1} (u_-^c)_j (\phi_{n_-})_j (\epsilon_-)_j w_j + \sum_{j=1}^{N-1} (u_+^c)_j (\phi_{n_+})_j (\epsilon_+)_j w_j. \end{aligned}$$

Applying the same procedure for $v^c = \psi_n(x)(0 \le n \le N-2)$, we have a system of $2N-2$ equations with the unknown vector

$$\mathbf{u}^c = [(u_-^c)_0, (u_-^c)_1, \ldots, (u_-^c)_{N-1}, (u_+^c)_0, (u_+^c)_1, \ldots, (u_+^c)_{N-1}]^T.$$

From the remaining boundary and the interface conditions, that is, $(u_-^c)_0' = (u_+^c)_N'$ and $(u_-^c)_N' = (u_+^c)_0'$, we get two more equations:

$$(u_-^c)_0 [D_{0N} - D_{N0}] + \sum_{j=1}^{N-1} (u_-^c)_j D_{jN} + (u_+^c)_0 [D_{NN} - D_{00}] - \sum_{j=1}^{N-1} (u_+^c)_j D_{j0} = 0,$$

$$(u_-^c)_0 [D_{0N} - D_{N0}] + \sum_{j=1}^{N-1} (u_-^c)_j D_{j0} + (u_+^c)_0 [D_{NN} - D_{00}] - \sum_{j=1}^{N-1} (u_+^c)_j D_{jN} = 0.$$

Table 2

Domain decomposition Legendre-collocation methods (LC1 = method 1, LC2 = method 2) for the relative errors of eigenvalues for $\omega = 2$ and the l_2-discrete errors of eigenfunctions.

Methods		LC1		LC2	
λ_i	N	$(\lambda_i^c - \lambda_i)/\lambda_i$	$\|u_i - u_i^c\|$	$(\lambda_i^{cc} - \lambda_i)/\lambda_i$	$\|u_i - u_i^{cc}\|$
0.369875	4	2.94(-03)	2.91(-03)	3.81(-01)	1.97(-01)
	8	7.04(-10)	4.65(-07)	-5.74(-05)	2.95(-06)
	16	-1.75(-14)	2.33(-15)	-2.61(-13)	1.35(-14)
	32	-2.23(-13)	3.19(-14)	2.30(-12)	1.40(-13)
	64	-6.66(-13)	3.32(-13)	9.42(-12)	4.81(-13)
0.536233	4	-6.14(-03)	2.17(-02)	1.39(+00)	1.06(+00)
	8	-8.61(-09)	4.22(-06)	-3.38(-05)	2.99(-05)
	16	-2.48(-15)	3.10(-14)	-1.49(-13)	3.10(-13)
	32	0.00(+00)	1.59(-13)	5.81(-14)	4.88(-13)
	64	-2.05(-13)	2.46(-13)	-5.60(-13)	2.18(-12)
1.607115	4	-7.80(-02)	1.29(-01)	8.07(-01)	8.79(-01)
	8	-2.89(-05)	6.33(-04)	-2.10(-03)	5.43(-03)
	16	6.90(-16)	7.33(-11)	-8.33(-10)	1.59(-09)
	32	1.28(-14)	2.26(-13)	2.99(-14)	4.62(-13)
	64	-4.42(-15)	6.55(-13)	-1.85(-13)	1.35(-12)
1.937181	4	3.28(-01)	1.17(-01)	1.62(+00)	2.49(-01)
	8	1.45(-04)	4.99(-04)	1.13(-02)	1.89(-03)
	16	1.83(-15)	2.38(-10)	2.20(-08)	3.83(-09)
	32	-6.49(-14)	3.86(-14)	4.56(-13)	8.42(-14)
	64	-2.68(-13)	1.24(-13)	1.85(-12)	3.32(-13)

Finally, we can represent the Legendre-collocation (method 1) scheme (5.11) in matrix form:

$$K\mathbf{u}^c = \lambda^c M\mathbf{u}^c, \tag{5.16}$$

where the dimension of the matrices K and M is $2N \times 2N$.

The numerical results are presented in Table 2 for the first few eigenvalues in ascending order and the corresponding eigenfunctions, showing that the relative errors for eigenvalues and l_2-errors of the eigenfunctions decay exponentially as N increases.

5.2. Legendre-collocation method 2.

Let us first define the $(N-1)$-dimensional space

$$\bar{\mathcal{L}}_{N-1} = \mathrm{span}\{l_j(\xi) | 1 \leq j \leq N-1, \ \ \xi \in [-1,1]\}, \tag{5.17}$$

where the Lagrange interpolation polynomials of degree N based on the Legendre–Gauss–Lobatto points are

$$l_j(\xi) = -\frac{1}{N(N+1)} \frac{(1-\xi^2)L_N'(\xi)}{(\xi - \xi_j)L_N(\xi_j)}. \tag{5.18}$$

Then we define

$$W_{2N-2} = \{\varphi \in C_p^0(\Omega) | \varphi_-(x_-(\xi)) \in \bar{\mathcal{L}}_{N-1} \ \text{ and } \ \varphi_+(x_+(\xi)) \in \bar{\mathcal{L}}_{N-1}\}, \tag{5.19}$$

the basis of which is given by $\{\varphi_n, \ \zeta_n\}_{n=1}^{N-1}$, where

$$\varphi_n = \begin{cases} l_n(\xi) & \text{in } [-\pi, 0], \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \zeta_n = \begin{cases} 0 & \text{in } [-\pi, 0], \\ l_n(\xi) & \text{otherwise.} \end{cases}$$

The multidomain discrete variational formulation of (2.2) for this method is to find λ^{cc} and u^{cc} in V_{2N-2} such that

$$a(u^{cc}, w^c)_h = \lambda^{cc}(u^{cc}, w^c)_h \text{ for } w^c \in W_{2N-2}. \tag{5.20}$$

Setting

$$(u_-^{cc})_i = \sum_{j=0}^{N} (u_-^{cc})_j l_j(\xi_i),$$
$$(u_+^{cc})_i = \sum_{j=0}^{N} (u_+^{cc})_j l_j(\xi_i)$$

and plugging them into (5.20) with $w^c = \varphi_n(x)$, $\zeta_n(x)$ $(1 \le n \le N-1)$, we get

$$K\mathbf{u}^{cc} = \lambda^{cc} M\mathbf{u}^{cc}, \tag{5.21}$$

where, denoting the second derivative matrix of Lagrange interpolation polynomials of degree N based on Legendre–Gauss–Lobatto points by D^2, the entries of $(2N \times 2N)$ matrices K and M are expressed by

$$K_{nj} = \begin{cases} D^2_{nj} \begin{cases} 1 \le n \le N-1, & 0 \le j \le N, \\ N+1 \le n \le 2N, & N \le j \le 2N, \end{cases} \\ 0 \quad \text{otherwise}, \end{cases} \tag{5.22}$$

and

$$M_{nj} = \operatorname{diag}\{(\epsilon_-)_1, \ldots, (\epsilon_-)_{N-1}, (\epsilon_+)_1, \ldots, (\epsilon_+)_{N-1}\}.$$

Additionally, two more equations are incorporated into the first and Nth rows of the matrices K and M, which are from the boundary and interface constraints:

$$(u_-^{cc})_0[D_{0N} - D_{N0}] + \sum_{j=1}^{N-1} (u_-^{cc})_j D_{jN} + (u_+^{cc})_0[D_{NN} - D_{00}] - \sum_{j=1}^{N-1} (u_+^{cc})_j D_{j0} = 0,$$
$$(u_-^{cc})_0[D_{0N} - D_{N0}] + \sum_{j=1}^{N-1} (u_-^{cc})_j D_{j0} + (u_+^{cc})_0[D_{NN} - D_{00}] - \sum_{j=1}^{N-1} (u_+^{cc})_j D_{jN} = 0,$$

where D is defined as in section 5.1.

Solving (5.21) numerically, one obtains the eigenvalues λ_l^{cc} $(l \le 2N)$ and the associated eigenvector

$$\mathbf{u}^{cc} = [(u_-^{cc})_0, (u_-^{cc})_1, \ldots, (u_-^{cc})_{N-1}, (u_+^{cc})_0, (u_+^{cc})_1, \ldots, (u_+^{cc})_{N-1}]^T.$$

The numerical results are presented in Table 2 for the first few eigenvalues in ascending order and the corresponding eigenfunctions, showing the exponential rate of convergence. Simply replacing the set of points and weights by the Chebyshev–Gauss–Lobatto points and weights, one can construct the multidomain Chebyshev-collocation method, the results of which are provided in Table 3.

TABLE 3

Relative errors of eigenvalues $\omega = 2$ and the l_2-discrete errors of eigenfunctions for multidomain Chebyshev-collocation method 2.

λ_i	N	$(\lambda_i^{cc} - \lambda_i)/\lambda_i$	$\|u_i - u_i^{cc}\|$
0.369875	4	-2.75(-02)	2.81(-03)
	8	-2.20(-05)	1.22(-06)
	16	8.55(-15)	1.99(-15)
	32	-1.76(-13)	2.80(-14)
	64	-4.58(-12)	1.15(-12)
0.536233	4	4.21(-02)	2.51(-02)
	8	-1.39(-05)	1.18(-05)
	16	-5.77(-14)	5.06(-14)
	32	-5.38(-14)	1.72(-13)
	64	-3.55(-13)	9.13(-13)
1.607115	4	1.55(-01)	4.70(-01)
	8	-8.48(-04)	1.82(-03)
	16	-2.28(-10)	4.31(-10)
	32	4.98(-14)	8.25(-14)
	64	-1.41(-13)	9.66(-13)
1.937181	4	6.97(-01)	4.74(-02)
	8	4.52(-03)	6.72(-04)
	16	6.13(-09)	1.05(-09)
	32	-2.88(-14)	1.56(-14)
	64	-8.99(-13)	4.72(-13)

6. Legendre-collocation penalty method. In this section, the notation used in section 4 represents the same definition. Let $\mathcal{L}_{N+1}$ be the $(N+1)$-dimensional space of Legendre–Lagrange interpolation polynomials of degree N defined by

$$\mathcal{L}_{N+1} = \text{span}\{l_j(\xi) | 0 \le j \le N, \quad \xi \in [-1, 1]\}, \tag{6.1}$$

where

$$l_j(\xi) = -\frac{1}{N(N+1)} \frac{(1-\xi^2) L_N'(\xi)}{(\xi - \xi_j) L_N(\xi_j)}. \tag{6.2}$$

Let Y_{2N+2} be the $(2N+2)$-dimensional space of piecewise continuous interpolation polynomials defined as

$$Y_{2N+2} = \{\eta \in L^2(\Omega) | \eta_-(x_-(\xi)) \in \mathcal{L}_{N+1} \text{ and } \quad \eta_+(x_+(\xi)) \in \mathcal{L}_{N+1}\},$$

the basis of which is given by

$$\{\eta_n, \varsigma_n\} \quad \text{for} \quad 0 \le n \le N, \tag{6.3}$$

where

$$\eta_n = \begin{cases} l_n(\xi) & \text{in } [-\pi, 0], \\ 0 & \text{otherwise} \end{cases} \quad \text{and} \quad \varsigma_n = \begin{cases} 0 & \text{in } (-\pi, 0), \\ l_n(\xi) & \text{otherwise.} \end{cases}$$

Now, we define the discrete bilinear form approximating $a(u, v)$ with penalty boundary constraints:

$$\begin{aligned} a(u, v)_\tau = {} & a(u, v)_{h_-} + a(u, v)_{h_+} \\ & + \tau_1\{(u_+)_N - (u_-)_0\} \\ & + \tau_2\{(u_+)_0 - (u_-)_N\} \\ & + \tau_3\{(u_+)'_N - (u_-)'_0\} \\ & + \tau_4\{(u_-)'_N - (u_+)'_0\}, \end{aligned}$$

where τ_i $(1 \le i \le 4)$ are suitable constants depending on N, to be chosen later.

Now, we state the multidomain discrete variational formulation of (2.2) by penalty approach. Find λ^τ and u^τ in Y_{2N+2} such that

$$a(u^\tau, v^\tau)_\tau = \lambda^\tau (u^\tau, v^\tau)_h \ \text{ for } v^\tau \in Y_{2N+2}. \tag{6.4}$$

To construct the scheme, we expand

$$(u^\tau_-)_i = \sum_{j=0}^{N} (u^\tau_-)_j l_j(\xi_i), \tag{6.5}$$

$$(u^\tau_+)_i = \sum_{j=0}^{N} (u^\tau_+)_j l_j(\xi_i). \tag{6.6}$$

Take $v^\tau = \eta_n(x)$ and $v^\tau = \varsigma_n(x)$, and choose $\tau_1 = \tau_2 = \sigma_1^\tau$ and $\tau_3 = \tau_4 = \sigma_2^\tau$. Then we define the following matrix K, whose dimensions are $2N + 2$:

$$K = \begin{bmatrix} D^2 & O \\ O & D^2 \end{bmatrix}, \tag{6.7}$$

where the $(N + 1) \times (N + 1)$ matrix D^2 is the second derivative matrix of Lagrange interpolation polynomials based on Legendre–Gauss–Lobatto points. Also, defining the matrices

$$M = \operatorname{diag}\{(\epsilon_-)_0, \ldots, (\epsilon_-)_N, (\epsilon_+)_0, \ldots, (\epsilon_+)_N\},$$

$$B_1 = \begin{bmatrix} -1 & \ldots & 0 & 0 & \ldots & 1 \\ 0 & \ldots & 0 & 0 & \ldots & 0 \\ \vdots & \ldots & 0 & 0 & \ldots & \vdots \\ 0 & \ldots & -1 & 1 & \ldots & 0 \\ 0 & \ldots & -1 & 1 & \ldots & 0 \\ \vdots & \ldots & 0 & 0 & \ldots & \vdots \\ 0 & \ldots & 0 & 0 & \ldots & 0 \\ -1 & \ldots & 0 & 0 & \ldots & 1 \end{bmatrix} \tag{6.8}$$

Table 4

Relative errors of eigenvalues for $\omega = 2$ and the L_2-discrete errors of eigenfunctions for the Legendre-collocation penalty (LCP) method and the Chebyshev-collocation penalty (CCP) method.

Methods		LCP		CCP	
λ_i	N	$(\lambda_i^\tau - \lambda_i)/\lambda_i$	$\Vert u_i - u_i^\tau \Vert$	$(\lambda_i^\tau - \lambda_i)/\lambda_i$	$\Vert u_i - u_i^\tau \Vert$
0.369875	4	-6.35(-02)	2.91(-03)	-2.76(-02)	2.64(-03)
	8	-5.73(-05)	5.09(-06)	-2.21(-05)	1.18(-06)
	16	-2.18(-13)	2.53(-14)	-1.79(-12)	4.17(-13)
	32	5.10(-12)	4.75(-13)	-5.86(-12)	1.37(-12)
	64	-8.03(-12)	3.88(-12)	-1.28(-10)	2.83(-11)
0.536233	4	-4.81(-02)	4.12(-02)	-4.30(-02)	2.43(-02)
	8	-3.40(-05)	4.01(-05)	-1.41(-05)	1.13(-05)
	16	-6.40(-13)	3.75(-12)	4.47(-13)	3.07(-12)
	32	2.05(-13)	1.72(-12)	2.44(-12)	1.73(-11)
	64	3.90(-11)	3.47(-10)	4.10(-11)	2.93(-10)
1.607115	4	-1.96(-01)	4.42(-01)	-1.52(-01)	4.53(-01)
	8	-2.10(-03)	4.06(-03)	-8.34(-04)	1.70(-03)
	16	-8.32(-10)	1.50(-09)	-2.26(-10)	4.18(-10)
	32	1.70(-12)	4.20(-11)	6.39(-13)	6.93(-12)
	64	-2.21(-12)	8.62(-11)	1.20(-11)	1.36(-10)
1.937181	4	5.11(-01)	5.28(-02)	7.13(-01)	5.81(-02)
	8	1.13(-02)	1.82(-03)	4.57(-03)	6.38(-04)
	16	2.20(-08)	3.71(-09)	6.15(-09)	1.02(-09)
	32	4.61(-13)	6.62(-13)	-1.34(-12)	7.64(-13)
	64	1.01(-11)	4.35(-12)	-2.51(-11)	1.61(-11)

and

$$B_2 = \begin{bmatrix} -D_{01} & \dots & -D_{0N} & D_{N1} & \dots & D_{NN} \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ D_{N1} & \dots & D_{NN} & -D_{01} & \dots & -D_{0N} \\ -D_{01} & \dots & -D_{0N} & D_{N1} & \dots & D_{NN} \\ 0 & \dots & 0 & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & 0 & \dots & 0 \\ D_{N1} & \dots & D_{NN} & -D_{01} & \dots & -D_{0N} \end{bmatrix}, \tag{6.9}$$

and letting

$$K_\tau = K + \sigma_1^\tau B_1 + \sigma_2^\tau B_2, \tag{6.10}$$

we can represent the Legendre-collocation penalty scheme for (6.4):

$$K_\tau \mathbf{u}^\tau = \lambda^\tau M \mathbf{u}^\tau, \tag{6.11}$$

where $\mathbf{u}^\tau = [(u_-^\tau)_0, \dots, (u_-^\tau)_N, (u_+^\tau)_0, \dots, (u_+^\tau)_N]^T$.

The numerical computations are carried out for the case $\sigma_1^\tau = \sigma_2^\tau = \left\{\frac{2}{\pi} N(N+1)\right\}^2$ [11], [12], which is chosen for the matrix K_τ to be symmetric positive definite [9], [20]. The results, shown in Table 4, demonstrate the exponential rate of convergence.

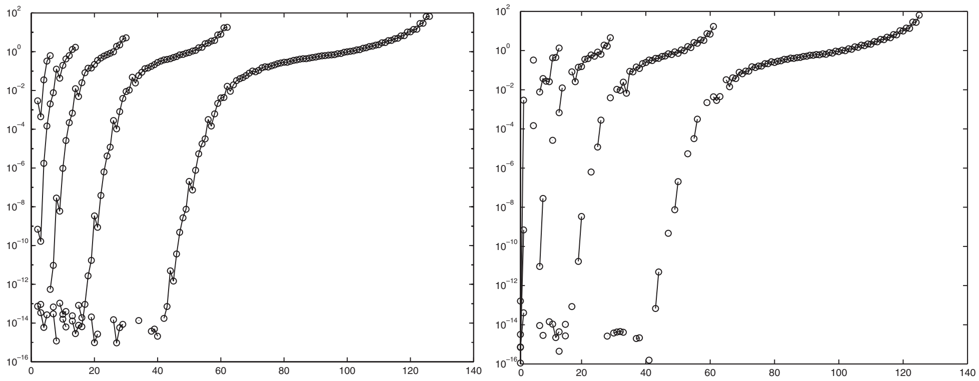


FIG. 1. *Legendre–Galerkin method (left) and Legendre-collocation method* 1 *(right): the relative errors of all the eigenvalues for* $N = 4, 8, 16, 32, 64$.

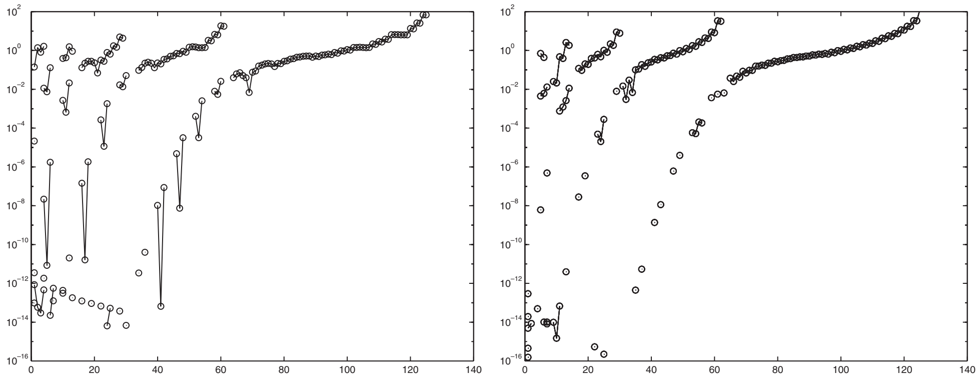


FIG. 2. *Legendre-collocation method* 2 *(left) and Legendre-collocation penalty method (right): the relative errors of all the eigenvalues for* $N = 4, 8, 16, 32, 64$.

Simply replacing the set of points and the weights by the Chebyshev–Gauss–Lobatto points and weights, one can construct the multidomain Chebyshev-collocation penalty method, whose results also are provided in Table 4.

The theoretical analysis of the convergence for the multidomain spectral penalty method is left for future study, as is the analysis for optimizing the parameter τ_i for this eigenvalue problem.

7. Discussion. In this section we discuss the asymptotic behavior of the largest approximate eigenvalues obtained by the multidomain spectral techniques for the eigenproblem with a discontinuous coefficient. Figures 1–2 demonstrate the relative errors of the eigenvalues with fixed $N = 4, 8, 16, 32, 64$ for each different method. The figures show that for the approximations with degree N, the fraction $\frac{2}{\pi}$ of the approximate eigenvalues converges to the analytic eigenvalues exponentially. Bernardi and Maday [2] and Vandeven [22] give rigorous proofs for finding the fraction of the approximate eigenvalues that approximate the eigenvalues of the second-order spectral differentiation operator.

We present the relative errors for the first 29 eigenvalues for a fixed $N = 16$ in

Table 5

Relative errors of eigenvalues for $\omega = 2$ for the domain decomposition Legendre–Galerkin (LG), Legendre-collocation method 1 (LC1), Legendre-collocation method 2 (LC2), Chebyshev-collocation method 2 (CC2), Legendre-collocation penalty method (LCP), Chebyshev-collocation penalty method (CCP): λ_l^* is the lth approximate eigenvalue and $N = 16$.

	$(\lambda_l^* - \lambda_l)/\lambda_l$					
l	LG	LC1	LC2	CC2	LCP	CCP
1	-6.15(-15)	-1.68(-14)	-9.67(-14)	8.55(-15)	-1.05(-12)	-1.79(-12)
2	-2.89(-14)	-8.28(-16)	-1.49(-13)	-5.77(-14)	-2.58(-14)	4.47(-13)
3	-2.76(-16)	4.69(-15)	-8.33(-10)	-2.28(-10)	-8.31(-10)	-2.26(-10)
4	-6.07(-15)	1.14(-16)	2.20(-08)	6.13(-09)	2.20(-08)	6.15(-09)
5	5.52(-13)	-1.42(-11)	8.41(-12)	-1.69(-08)	-1.31(-08)	-2.37(-08)
6	9.48(-12)	9.48(-12)	1.74(-06)	4.80(-07)	1.74(-06)	4.81(-07)
7	2.83(-08)	2.83(-08)	-1.19(-04)	-3.16(-05)	-1.19(-04)	-3.16(-05)
8	5.94(-09)	-7.70(-08)	-7.80(-05)	-2.20(-05)	-7.87(-05)	-2.24(-05)
9	9.59(-07)	-8.19(-06)	-8.51(-04)	-2.36(-04)	-8.44(-04)	-2.33(-04)
10	2.62(-05)	2.62(-05)	2.70(-03)	7.56(-04)	2.71(-03)	7.61(-04)
11	2.15(-04)	-1.11(-03)	6.59(-04)	1.20(-03)	5.59(-04)	1.16(-03)
12	6.74(-04)	6.74(-04)	2.11(-02)	2.61(-03)	2.11(-02)	2.62(-03)
13	1.25(-02)	1.25(-02)	-2.07(-02)	1.12(-02)	-2.06(-02)	1.12(-02)
14	4.88(-03)	-1.50(-02)	-6.98(-02)	-3.63(-02)	-7.02(-02)	-3.65(-02)
15	2.53(-02)	-2.96(-02)	-6.49(-02)	-4.14(-02)	-6.42(-02)	-4.10(-02)
16	8.25(-02)	8.25(-02)	1.29(-01)	1.17(-01)	1.30(-01)	1.18(-01)
17	1.40(-01)	2.55(-02)	2.22(-01)	9.57(-02)	2.25(-01)	9.61(-02)
18	1.44(-01)	1.40(-01)	2.76(-01)	2.05(-01)	2.76(-01)	2.04(-01)
19	2.23(-01)	1.53(-01)	2.76(-01)	1.92(-01)	2.74(-01)	1.91(-01)
20	3.57(-01)	3.57(-01)	2.30(-01)	3.88(-01)	2.27(-01)	3.87(-01)
21	4.77(-01)	3.89(-01)	6.88(-02)	4.15(-01)	6.65(-02)	4.15(-01)
22	5.89(-01)	5.89(-01)	3.17(-01)	6.21(-01)	3.18(-01)	6.20(-01)
23	6.97(-01)	5.20(-01)	2.76(-01)	4.69(-01)	2.77(-01)	4.69(-01)
24	8.38(-01)	8.38(-01)	7.77(-01)	1.00(+00)	7.78(-01)	1.00(+00)
25	9.78(-01)	6.44(-01)	6.49(-01)	8.09(-01)	6.49(-01)	8.09(-01)
26	1.89(+00)	1.89(+00)	1.74(+00)	2.19(+00)	1.75(+00)	2.21(+00)
27	2.22(+00)	1.67(+00)	1.46(+00)	1.86(+00)	1.48(+00)	1.88(+00)
28	4.59(+00)	4.58(+00)	4.96(+00)	8.90(+00)	4.99(+00)	9.00(+00)
29	5.22(+00)	3.93(+00)	4.31(+00)	7.82(+00)	4.33(+00)	7.91(+00)

Table 5. One can see that $\frac{1}{\pi}$ of the eigenvalues approximate the analytic eigenvalues of the problem very accurately. One also can see that the Legendre–Galerkin method and the Legendre-collocation method 1 are more accurate than the other collocation methods. However, the Legendre-collocation method 2 and Legendre-collocation penalty method are relatively easier to implement because of their simplicity in dealing with the basis of the space used in the approximation.

8. Conclusion. In this paper, we have mainly discussed the Legendre–Galerkin, Legendre-collocation, and Legendre-collocation penalty methods with a domain decomposition approach in order to get exponentially accurate eigensolutions for a model eigenvalue problem with a piecewise continuous coefficient.

Acknowledgments. Special thanks to Wai-Sun Don, Jan Hesthaven, and Yvon Maday for their discussions and several suggestions regarding the topics in this paper.

REFERENCES

[1] W. Axmann and P. Kuchment, *An efficient finite element method for computing spectra of photonic and acoustic band-gap materials:* I. *Scalar case,* J. Comput. Phys., 150 (1999), pp. 468–481.

[2] C. Bernardi and Y. Maday, *Spectral Methods,* Handb. Numer. Anal. 5, North-Holland, Amsterdam, 1997.

[3] J. Broeng, D. Mogilevtsev, S. E. Barkou, and A. Bjarklev, *Photonic crystal fibres: A new class of optical waveguides*, Opt. Fiber Technol., 5 (1999), pp. 305–330.

[4] C. Canuto and A. Quarteroni, *Error estimates for spectral and pseudospectral approximations of hyperbolic equations,* SIAM J. Numer. Anal., 19 (1982), pp. 629–642.

[5] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics,* Springer Ser. Comput. Phys., Springer-Verlag, New York, 1988.

[6] D. C. Dobson, *An efficient method for band structure calculations in* 2*D photonic crystals,* J. Comput. Phys., 149 (1999), pp. 363–376.

[7] D. Funaro and D. Gottlieb, *A new method of imposing boundary conditions in pseudospectral approximations of hyperbolic equations,* Math. Comp., 57 (1988), pp. 599–613.

[8] D. Funaro, *Polynomial Approximation of Differential Equations,* Springer-Verlag, New York, 1991.

[9] G. H. Golub and C. Van Loan, *Matrix Computations,* Johns Hopkins University Press, Baltimore, MD, 1996.

[10] D. Gottlieb, M. Y. Hussaini, and S. A. Orszag, *Theory and applications of spectral methods*, in Spectral Methods for Partial Differential Equations, R. G. Voigt, D. Gottlieb, and M. Y. Hussaini, eds., SIAM, Philadelphia, 1984, pp. 1–54.

[11] D. Gottlieb and L. Lustman, *The spectrum of the Chebyshev collocation operator for the heat equation*, SIAM J. Numer. Anal., 20 (1983), pp. 909–921.

[12] J. S. Hesthaven, *Integration preconditioning of pseudospectral operators.* I. *Basic linear operators,* SIAM J. Numer. Anal., 35 (1998), pp. 1571–1593.

[13] J. S. Hesthaven and D. Gottlieb, *Spectral Approximation of Partial Differential Equations: Numerical Analysis and Applications,* Lecture Notes, Brown University, Providence, RI, 1996.

[14] J. D. Joannopoulos, R. D. Meade, and J. N. Winn, *Photonic Crystals: Molding the Flow of Light,* Princeton University Press, Princeton, NJ, 1995.

[15] S. John, *Strong localization of photons in certain disordered dielectric superlattices*, Phys. Rev. Lett., 58 (1987), pp. 2486–2489.

[16] S. G. Johnson and J. D. Joannopoulos, *Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis,* Optics Express, 8 (2001), pp. 173–190.

[17] C. Mias, J. P. Webb, and R. L. Ferrari, *Finite element modelling of electromagnetic waves in doubly and triply periodic structures,* IEEE Proc.-Optoelectron, 145 (1999), pp. 111–118.

[18] M. S. Min and D. Gottlieb, *On the convergence of the Fourier approximation for eigenvalues and eigenfunctions of discontinuous problems,* SIAM J. Numer. Anal., 40 (2003), pp. 2254–2269.

[19] M. S. Min, Q. Y. Chen, and Y. Maday, *Spectral method for* 2*D photonic band structures,* in Proceedings of SPIE, Photonic Crystal Materials and Devices II, Proc. SPIE 5360, SPIE, Bellingham, WA, 2004, pp. 44–51.

[20] Y. Saad, *Iterative Methods for Sparse Linear Systems,* 2nd ed., SIAM, Philadelphia, 2003.

[21] G. Strang and G. Fix, *An Analysis of the Finite Element Method,* Prentice-Hall, Englewood Cliffs, NJ, 1973.

[22] H. Vandeven, *On the eigenvalues of second-order spectral differentiation operators,* in Spectral and High Order Methods for Partial Differential Equations, in Proceedings of the ICOSAHOM '89 Conference, 1989, pp. 313–318.

[23] E. Yablonovitch, *Inhibited spontaneous emission in solid-state physics and electronics*, Phys. Rev. Lett., 58 (1987), pp. 2059–2062.

SIAM J. Numer. Anal.
Vol. 43, No. 2, pp. 521–539

CONVERGENCE ANALYSIS OF WAVELET SCHEMES FOR CONVECTION-REACTION EQUATIONS UNDER MINIMAL REGULARITY ASSUMPTIONS*

JIANGGUO LIU[†], BOJAN POPOV[‡], HONG WANG[§], AND RICHARD E. EWING[†]

Abstract. In this paper, we analyze convergence rates of wavelet schemes for time-dependent convection-reaction equations within the framework of the Eulerian–Lagrangian localized adjoint method (ELLAM). Under certain minimal assumptions that guarantee H^1-regularity of exact solutions, we show that a generic ELLAM scheme has a convergence rate $\mathcal{O}(h/\sqrt{\Delta t} + \Delta t)$ in L^2-norm. Then, applying the theory of operator interpolation, we obtain error estimates for initial data with even lower regularity. Namely, it is shown that the error of such a scheme is $\mathcal{O}((h/\sqrt{\Delta t})^\theta + (\Delta t)^\theta)$ for initial data in a Besov space $B^\theta_{2,q}(0 < \theta < 1, 0 < q \leq \infty)$. The error estimates are a priori and optimal in some cases. Numerical experiments using orthogonal wavelets are presented to illustrate the theoretical estimates.

Key words. characteristic method, convection-reaction equation, convergence rate, Eulerian–Lagrangian method, wavelet method

AMS subject classifications. 65M12, 65M25, 65M60, 76M25, 76S05

DOI. 10.1137/S0036142903433832

1. Introduction.

This paper is concerned with convergence rates of the wavelet schemes established in [24] for an initial value problem (IVP) to the following multidimensional linear convection-reaction equation

$$\begin{cases} u_t + \nabla \cdot (\mathbf{V}u) + Ru = f(\mathbf{x}, t), & (\mathbf{x}, t) \in \mathbb{R}^d \times (0, T], \\ u(\mathbf{x}, 0) = u_0(\mathbf{x}), & \mathbf{x} \in \mathbb{R}^d, \end{cases} \tag{1.1}$$

where $u(\mathbf{x}, t)$ is the unknown concentration function, $\mathbf{V}(\mathbf{x}, t)$ is a fluid velocity field, $R(\mathbf{x}, t)$ is a first order reaction coefficient, $f(\mathbf{x}, t)$ is a source/sink term, and $u_0(\mathbf{x})$ is a prescribed initial condition. It is assumed that $u_0(\mathbf{x})$ and $f(\mathbf{x}, t)$ are compactly supported, and hence so is the exact solution $u(\mathbf{x}, t)$ for any finite time.

Convection-dominated reactive transport equations arise from remediation of subsurface contamination, nuclear waste disposal, biodegradation, numerical simulation of petroleum reservoir, and many other applications. The solutions to these types of problems usually are not smooth and raise serious challenges to numerical methods. Standard finite difference or finite element methods (FEMs) produce either excessively oscillatory or smeared solutions. Therefore, many special schemes have been developed to overcome these difficulties.

Characteristics-based methods were developed in the late 1970s and the early 1980s to solve convection-dominated problems. Systematic study including error estimates for the FEM in the case of nondegenerate diffusion was given in [17]. Improved estimates for such problems were given in [12] with a special discussion for

*Received by the editors August 26, 2003; accepted for publication (in revised form) September 22, 2004; published electronically June 30, 2005.

http://www.siam.org/journals/sinum/43-2/43383.html

[†]Institute for Scientific Computation, Texas A&M University, College Station, TX 77843-3404 (jliu@isc.tamu.edu, richard-ewing@tamu.edu).

[‡]Department of Mathematics, Texas A&M University, College Station, TX 77843-3368 (popov@math.tamu.edu).

[§]Department of Mathematics, University of South Carolina, Columbia, SC 29208-0001 (hwang@math.sc.edu).

degenerate diffusion. In the case of pure advection, [11] gives optimal rates of convergence for the piecewise linear FEM but with the restrictions of very small time steps ($\Delta t = \mathcal{O}(h^2)$) and sufficiently smooth exact solutions. There are many other results for convection-dominated diffusion problems, but in most cases the derived estimates adversely depend on the size of the diffusion parameter. In a recent paper [2], Bause and Knabner derived error estimates for Lagrangian–Galerkin methods, which are uniform with respect to the diffusion parameter, and therefore they remark that their results can carry over to the limit case of pure convection.

Among the existing characteristic methods, the Eulerian–Lagrangian localized adjoint method (ELLAM) [5] holds some advantages. It symmetrizes the governing equation, naturally incorporates boundary conditions, and conserves mass. However, ELLAM introduces further difficulties to the already fairly complicated analyses of characteristic methods. It was shown in [19, 25] that within the ELLAM framework, the piecewise linear FEM with uniform spatial partition has an optimal error estimate; in other words, it has a convergence rate $\mathcal{O}(h^2 + \Delta t)$, under the assumptions that $u \in L^\infty(0, T; H^3(\Omega))$ and $u_t \in L^2(0, T; H^2(\Omega))$. Many other papers also assume the exact solution is at least in H^2 in space. However, it is clear that the requirement $u \in L^\infty(0, T; H^s(\Omega))$ for any initial condition $u_0 \in H^s(\Omega)$ will imply that the fluid velocity is, roughly speaking, s times differentiable.

In this paper, requirements on regularity of the solution and the velocity field will be significantly reduced. We shall prove that under certain minimal assumptions that guarantee H^1-smoothness of the exact solution, ELLAM schemes including the orthogonal wavelet schemes satisfy

$$\max_{0\le n\le N} \|u(\mathbf{x}, t_n) - U_h^n(\mathbf{x})\|_{L^2(\mathbb{R}^d)} \le C(h/\sqrt{\Delta t} + \Delta t), \tag{1.2}$$

where $u(\mathbf{x}, t_n)$ is the exact solution at time t_n, and $U_h^n(\mathbf{x})$ is the corresponding numerical solution with spatial step size h and temporal step size Δt. The constant C depends only on the norms of the velocity, reaction, source, and initial data, but not on the norm of the exact solution $u(\mathbf{x}, t)$ itself.

Applying the theory of operator interpolation, we could obtain the error estimates

$$\max_{0\le n\le N} \|u(\mathbf{x}, t_n) - U_h^n(\mathbf{x})\|_{L^2(\mathbb{R}^d)} \le C\left[(h/\sqrt{\Delta t})^\theta + (\Delta t)^\theta\right] \tag{1.3}$$

for initial data u_0 in a Besov space $B^\theta_{2,q}(0 < \theta < 1, 0 < q \le \infty)$, where C could be additionally dependent on θ. This extends our results to a wide class of data, including discontinuous initial conditions, moving sharp fronts, or shocks. Generally speaking, Besov spaces provide subtler characterization on regularity of functions than Sobolev spaces do. Sometimes the exact order of approximation accuracy can be described only when Besov spaces are used, as illustrated by our Example 2 in section 8. Efforts in applying Besov spaces to other problems in numerical analysis can also be observed in literature, e.g., a recent work by Bacuta, Bramble, and Xu [1].

Errors in the wavelet schemes come from truncation, characteristic tracking, quadrature, and round-off. In this paper, we disregard quadrature rule error in the computations of wavelet coefficients. That is, we assume numerical integration is exact, following the common practices of most researchers [12, 17, 25]. The classic book [6] (see sections 4.1 and 4.4 of that work) has a full discussion on quadrature errors in the finite element method for elliptic problems. For time-dependent problems, some discussions can be found in [21].

The rest of this paper is organized as follows: Section 2 outlines the main ideas in the ELLAM weak formulation and establishes our numerical schemes, including the orthogonal wavelet schemes, within the ELLAM framework. Section 3 presents error estimates for the numerical schemes under some minimal regularity assumptions on the given data that guarantee H^1-stability of the exact solutions (Theorem 1). The proof of Theorem 1 is presented in section 4. In section 5, we state and prove the stability lemmas about the exact solutions used in section 4. Section 6 extends the results on convergence rates of the numerical schemes to initial data in Besov spaces. In section 7, we discuss optimality of our error estimates. Numerical experiments are presented in section 8 to illustrate the theoretical results. Finally, section 9 concludes the paper with some remarks.

2. ELLAM schemes.

In this section, we establish the ELLAM weak formulation for problem (1.1). Based on this weak formulation, we derive a generic (abstract) ELLAM scheme. Then we present the wavelet-ELLAM schemes using orthogonal wavelets.

2.1. ELLAM formulation.

Let $[0, T]$ be the time period, N a positive integer, $\Delta t := T/N$, and $t_n = n\Delta t$ $(n = 0, \dots, N)$ be a uniform partition of $[0, T]$. To establish a weak formulation for (1.1), we choose test function $w(\mathbf{x}, t)$ in such a way that it vanishes outside the space-time strip $\mathbb{R}^d \times (t_{n-1}, t_n]$ and is discontinuous in time at time t_{n-1}. Then integration by parts gives us

$$\begin{aligned}&\int_{\mathbb{R}^d} u(\mathbf{x}, t_n) w(\mathbf{x}, t_n) d\mathbf{x} - \int_{t_{n-1}}^{t_n} \int_{\mathbb{R}^d} (u(w_t + \mathbf{V}\cdot\nabla w - Rw))(\mathbf{x}, t) d\mathbf{x}dt \\ &\quad = \int_{\mathbb{R}^d} u(\mathbf{x}, t_{n-1}) w(\mathbf{x}, t_{n-1}^+) d\mathbf{x} + \int_{t_{n-1}}^{t_n} \int_{\mathbb{R}^d} f(\mathbf{x}, t) w(\mathbf{x}, t) d\mathbf{x}dt,\end{aligned} \tag{2.1}$$

where $w(\mathbf{x}, t_{n-1}^+) := \lim_{t\to t_{n-1}^+} w(\mathbf{x}, t)$ takes into account the fact that $w(\mathbf{x}, t)$ is discontinuous in time at time t_{n-1}.

To cancel the second term on the left side of (2.1), we require the test function to satisfy the adjoint equation

$$w_t + \mathbf{V}\cdot\nabla w - Rw = 0. \tag{2.2}$$

Solving the above problem yields an explicit expression for the test function:

$$w(\mathbf{y}(s; \mathbf{x}, t_n), s) = w(\mathbf{x}, t_n)\, e^{\int_{t_n}^{s} R(\mathbf{y}(r;\mathbf{x},t_n),r)dr}, \quad s \in (t_{n-1}, t_n], \tag{2.3}$$

where the characteristic $\mathbf{y}(s; \mathbf{x}, t_n)$ passing through $(\mathbf{x}, t_n)$ is determined by

$$\begin{cases} \dfrac{d\mathbf{y}}{ds} = \mathbf{V}(\mathbf{y}, s), \\ \mathbf{y}(s; \mathbf{x}, t_n)|_{s=t_n} = \mathbf{x}. \end{cases} \tag{2.4}$$

Then we are led to the reference equation

$$\begin{aligned}\int_{\mathbb{R}^d} u(\mathbf{x}, t_n) w(\mathbf{x}, t_n) d\mathbf{x} &= \int_{\mathbb{R}^d} u(\mathbf{x}, t_{n-1}) w(\mathbf{x}, t_{n-1}^+) d\mathbf{x} \\ &\quad + \int_{t_{n-1}}^{t_n} \int_{\mathbb{R}^d} f(\mathbf{x}, t) w(\mathbf{x}, t) d\mathbf{x}dt.\end{aligned} \tag{2.5}$$

Let $\mathbf{x}^* = \mathbf{y}(t_{n-1}; \mathbf{x}, t_n)$. Applying (2.3), we can rewrite the first term on the right side of (2.5) as

$$\begin{aligned}&\int_{\mathbb{R}^d} u(\mathbf{x}, t_{n-1}) w(\mathbf{x}, t_{n-1}^{+}) d\mathbf{x} \\ &\quad = \int_{\mathbb{R}^d} u(\mathbf{x}^*, t_{n-1}) w(\mathbf{x}, t_n) e^{\int_{t_n}^{t_{n-1}} R(\mathbf{y}(s;\mathbf{x},t_n),s)ds} \mathbf{J}(\mathbf{x}^*, \mathbf{x}) d\mathbf{x},\end{aligned} \tag{2.6}$$

where $\mathbf{J}(\mathbf{x}^*, \mathbf{x})$ is the Jacobian of $\mathbf{x}^*$ with respect to $\mathbf{x}$.

2.2. A generic ELLAM scheme. Based on (2.3), we first approximate the test function $w(\mathbf{y}(s; \mathbf{x}, t_n), s)$ in the space-time strip $\mathbb{R}^d \times (t_{n-1}, t_n]$ by

$$w(\mathbf{x}, t_n)\ e^{\int_{t_n}^{s} R(\mathbf{x},t_n)dr} \equiv w(\mathbf{x}, t_n)\ e^{R(\mathbf{x},t_n)(s-t_n)},$$

and then use it to approximate the second (source) term on the right side of (2.5),

$$\begin{aligned}&\int_{t_{n-1}}^{t_n} \int_{\mathbb{R}^d} f(\mathbf{y}, s) w(\mathbf{y}, s) d\mathbf{y} ds \\ &\quad = \int_{\mathbb{R}^d} \int_{t_{n-1}}^{t_n} f(\mathbf{y}, s) w(\mathbf{y}, s) \mathbf{J}(\mathbf{y}, \mathbf{x}) ds d\mathbf{x} \\ &\quad = \int_{\mathbb{R}^d} \int_{t_{n-1}}^{t_n} f(\mathbf{x}, t_n) w(\mathbf{x}, t_n) e^{R(\mathbf{x},t_n)(s-t_n)} ds d\mathbf{x} + E(f, w) \\ &\quad = \int_{\mathbb{R}^d} f(\mathbf{x}, t_n) w(\mathbf{x}, t_n) G_n(\mathbf{x}) d\mathbf{x} + E(f, w),\end{aligned} \tag{2.7}$$

where $\mathbf{J}(\mathbf{y}, \mathbf{x})$ is the Jacobian of $\mathbf{y}$ with respect to $\mathbf{x}$,

$$G_n(\mathbf{x}) := \int_{t_{n-1}}^{t_n} e^{R(\mathbf{x},t_n)(s-t_n)} ds = \begin{cases} \dfrac{1 - e^{-R(\mathbf{x},t_n)\Delta t}}{R(\mathbf{x}, t_n)} & \text{if } R(\mathbf{x}, t_n) \neq 0, \\ \Delta t & \text{otherwise,} \end{cases} \tag{2.8}$$

and $E(f, w)$ is the error term

$$E(f, w) := \int_{t_{n-1}}^{t_n} \int_{\mathbb{R}^d} f(\mathbf{y}, s) w(\mathbf{y}, s) d\mathbf{y} ds - \int_{\mathbb{R}^d} f(\mathbf{x}, t_n) w(\mathbf{x}, t_n) G_n(\mathbf{x}) d\mathbf{x}. \tag{2.9}$$

In practice, exact tracking of characteristics is usually unavailable, and we have to resort to numerical means. All commonly used numerical methods, e.g., Euler and Runge–Kutta methods, can be employed to solve (2.4). Let $(\mathbf{x}^{**}, t_{n-1})$ be the numerical back-tracking image of $(\mathbf{x}, t_n)$; then $w(\mathbf{x}^{**}, t_{n-1}^{+})$ can be approximated by $w(\mathbf{x}, t_n) e^{-R(\mathbf{x},t_n)\Delta t}$.

Now a generic ELLAM scheme can be established as follows: Let $V_h \subset L^2(\mathbb{R}^d)$ be an approximation subspace. Find $U_h^n(\mathbf{x}) \in V_h$ such that for any test function w with $w(\mathbf{x}, t_n) \in V_h$,

$$\begin{aligned}\int_{\mathbb{R}^d} U_h^n(\mathbf{x}) w(\mathbf{x}, t_n) d\mathbf{x} &= \int_{\mathbb{R}^d} U_h^{n-1}(\mathbf{x}^{**}) w(\mathbf{x}, t_n) e^{-R(\mathbf{x},t_n)\Delta t} \mathbf{J}(\mathbf{x}^{**}, \mathbf{x}) d\mathbf{x} \\ &\quad + \int_{\mathbb{R}^d} f(\mathbf{x}, t_n) w(\mathbf{x}, t_n) G_n(\mathbf{x}) d\mathbf{x},\end{aligned} \tag{2.10}$$

where $U_h^0(\mathbf{x})$ is an approximation to $u_0(\mathbf{x})$ from V_h obtained by some means. For example, it can be taken as the L^2-orthogonal projection of u_0 into V_h.

There are many choices for the approximation subspace V_h. It could be constructed through finite elements or wavelets.

2.3. Wavelet schemes. Let

$$\cdots \subset \mathcal{V}_{-1} \subset \mathcal{V}_0 \subset \mathcal{V}_1 \subset \cdots \subset \mathcal{V}_j \subset \mathcal{V}_{j+1} \subset \cdots \subset L^2(\mathbb{R})$$

be a multiresolution analysis in $L^2(\mathbb{R})$ generated by an orthogonal scaling function $\phi(x)$ and let $\psi(x)$ be the associated orthogonal wavelet [10]. We construct d-dimensional scaling function and wavelets through tensor products. Let $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d, \mathbf{k} = (k_1, \ldots, k_d) \in \mathbb{Z}^d, \mathbf{e} = (e_1, \ldots, e_d) \in \{0, 1\}^d =: \hat{E}$, and $E := \hat{E} \setminus \{\mathbf{0}\}$. Then we define scaling function $\mathbf{\Phi}_{j,\mathbf{k}}(\mathbf{x}) := \prod_{i=1}^d \phi_{j,k_i}(x_i)$ and the wavelets

$$\Psi^{\mathbf{e}}_{j,\mathbf{k}}(\mathbf{x}) := \prod_{i=1}^{d} \Big(\phi_{j,k_i}(x_i)\Big)^{1-e_i} \Big(\psi_{j,k_i}(x_i)\Big)^{e_i}, \qquad \mathbf{e} \in E.$$

Furthermore, we define $V_j := \bigotimes_{i=1}^d \mathcal{V}_j$ as the closed linear span in $L^2(\mathbb{R}^d)$ of all functions of the form $f_1(x_1) \cdots f_d(x_d)$, where $f_i \in \mathcal{V}_j$ for $i = 1, \ldots, d$. Then $\langle V_j \rangle_{j \in \mathbb{Z}}$ forms a multiresolution analysis in $L^2(\mathbb{R}^d)$.

Clearly, we can use any subspace V_j in the above multiresolution analysis as an approximation subspace V_h in the ELLAM scheme (2.10). This gives us an orthogonal wavelet-ELLAM scheme.

For numerical implementations, we assume $\Omega \subset \mathbb{R}^d$ to be a rectangular domain such that the support of the solution $u(\mathbf{x}, t)$ to problem (1.1) is contained in Ω for all $t \in [0, T]$. Let $J_c < J_f$ be the chosen coarsest and finest resolution levels. For all j with $J_c \leq j \leq J_f$, we define

$$\Lambda_j := \{\mathbf{k} : \text{supp}\Phi_{j,\mathbf{k}} \cap \Omega \neq \emptyset\}, \tag{2.11}$$

$$\Lambda_{j,\mathbf{e}} := \{\mathbf{k} : \text{supp}\Psi^{\mathbf{e}}_{j,\mathbf{k}} \cap \Omega \neq \emptyset\}, \tag{2.12}$$

$$\mathcal{S}_j(\Omega) := \text{Span}\{\Phi_{j,\mathbf{k}}(\mathbf{x}) \mid \mathbf{k} \in \Lambda_j\}. \tag{2.13}$$

Then we set $V_h = \mathcal{S}_{J_f}(\Omega)$ with $h = 1/2^{J_f}$. There are two equivalent choices for the basis functions of V_h:

- only the scaling functions at the finest level J_f;
- the scaling functions at the coarsest level J_c plus all wavelets from level J_c to level $J_f - 1$.

When only the scaling functions at the finest level are used as the basis functions, we are seeking $U^n_h(\mathbf{x}) \in V_h$ with

$$U^n_h(\mathbf{x}) = \sum_{\mathbf{k} \in \Lambda_{J_f}} c^n_{J_f,\mathbf{k}} \Phi_{J_f,\mathbf{k}}(\mathbf{x}). \tag{2.14}$$

The orthogonality of scaling functions implies that we have an explicit scheme and the coefficients are given by

$$\begin{aligned} c^n_{J_f,\mathbf{k}} = &\int_\Omega U^{n-1}_h(\mathbf{x}^{**}) \Phi_{J_f,\mathbf{k}}(\mathbf{x}) e^{-R(\mathbf{x},t_n)\Delta t} \mathbf{J}(\mathbf{x}^{**}, \mathbf{x}) d\mathbf{x} \\ &+ \int_\Omega f(\mathbf{x}, t_n) \Phi_{J_f,\mathbf{k}}(\mathbf{x}) G_n(\mathbf{x}) d\mathbf{x}, \qquad \mathbf{k} \in \Lambda_{J_f}. \end{aligned} \tag{2.15}$$

This is the Scheme I discussed in [24].

When both the scaling functions at the coarsest level and the wavelets at fine levels are used as the basis functions, we get Scheme II with

$$U_h^n(\mathbf{x}) = \sum_{\mathbf{k}\in\Lambda_{J_c}} c_{J_c,\mathbf{k}}^n \Phi_{J_c,\mathbf{k}}(\mathbf{x}) + \sum_{j=J_c}^{J_f-1} \sum_{\mathbf{k}\in\Lambda_{j,\mathbf{e}}} d_{j,\mathbf{k}}^{n,\mathbf{e}} \Psi_{j,\mathbf{k}}^{\mathbf{e}}(\mathbf{x}). \tag{2.16}$$

Scheme II is also an explicit scheme. By choosing $w(\mathbf{x}, t_n) = \Phi_{J_c,\mathbf{k}}(\mathbf{x})$ or $\Psi_{j,\mathbf{k}}^{\mathbf{e}}(\mathbf{x})$, respectively, we obtain, again through the orthogonality,

$$\begin{aligned} c_{J_c,\mathbf{k}}^n &= \int_\Omega U_h^{n-1}(\mathbf{x}^{**})\Phi_{J_c,\mathbf{k}}(\mathbf{x})e^{-R(\mathbf{x},t_n)\Delta t}\mathbf{J}(\mathbf{x}^{**},\mathbf{x})d\mathbf{x} \\ &\quad + \int_\Omega f(\mathbf{x},t_n)\Phi_{J_c,\mathbf{k}}(\mathbf{x})G_n(\mathbf{x})d\mathbf{x}, \qquad \mathbf{k}\in\Lambda_{J_c}, \\ d_{j,\mathbf{k}}^{n,\mathbf{e}} &= \int_\Omega U_h^{n-1}(\mathbf{x}^{**})\Psi_{j,\mathbf{k}}^{\mathbf{e}}(\mathbf{x})e^{-R(\mathbf{x},t_n)\Delta t}\mathbf{J}(\mathbf{x}^{**},\mathbf{x})d\mathbf{x} \\ &\quad + \int_\Omega f(\mathbf{x},t_n)\Psi_{j,\mathbf{k}}^{\mathbf{e}}(\mathbf{x})G_n(\mathbf{x})d\mathbf{x}, \qquad \mathbf{k}\in\Lambda_{j,\mathbf{e}}, \ \ J_c \le j \le J_f - 1. \end{aligned} \tag{2.17}$$

In Scheme II, the first part on the right side of (2.16) provides a basic approximation. As more fine terms in the second part come in, we obtain better approximations.

As we know, solutions to convection-dominated transport equations often admit steep fronts and even jump discontinuities within very small regions but are smooth outside these regions. On the other hand, one prominent feature of wavelets is their localization capability. The terms in the wavelet expansion with noticeable coefficients correspond to the rough regions of the solution around those local singularities. We can drop the terms with small coefficients that correspond to the smooth regions of the solution. Therefore, the number of unknowns to be solved will be reduced. In other words, an adaptive multilevel scheme with mass-conservative compression can be constructed. This is the Scheme III presented in [24], which is, in some sense, equivalent to the traditional FEM with local refinement.

Due to compression, the wavelet basis functions (or elements) used in Scheme III vary at different time steps but are adapted to the solution we are looking for. This is a typical case of nonlinear approximation [14]. Of course, Schemes I and II are still in the category of linear approximation and are the main target of this paper. Theoretical analysis on convergence rates of Scheme III is much harder and will be addressed in our future work.

As proven in [24], all three wavelet schemes are explicit and unconditionally stable. In other words, they are not subject to the severe restriction of the Courant–Friedrichs–Lewy (CFL) condition. This allows us to use relatively large time steps.

3. Error estimate for solutions with H^1-regularity.

In this section, we derive an error estimate for the generic ELLAM scheme for exact solutions with only H^1-regularity.

Throughout this paper, we use L^p $(1 \le p \le \infty)$ to denote the standard Lebesgue spaces. Accordingly, W_p^k are the standard Sobolev spaces. When $p = 2$, we use H^k for W_2^k. For $1 \le q \le \infty$, we define

$$L^q(a,b;W_p^k) := \{u(\mathbf{x},t) | u(\cdot,t) : (a,b) \mapsto W_p^k, \|u(\cdot,t)\|_{W_p^k} \in L^q(a,b)\}.$$

In addition, $(W_\infty^1(\mathbb{R}^d \times [0,T]))^d$ is the space of vector-valued functions whose components are in the space $W_\infty^1(\mathbb{R}^d \times [0,T])$.

Assumption A. The approximation subspace V_h used in the generic ELLAM scheme (2.10) has the following approximation property:

$$\|u - P_h u\|_{L^2} \leq C_0 h \|u\|_{H^1} \qquad \forall u \in H^1(\mathbb{R}^d), \tag{3.1}$$

where C_0 is a constant independent of h and u. The above inequality is also called a Jackson-type inequality in literature.

Remark A. The above Jackson-type inequality is satisfied by commonly used wavelets, e.g., Daubechies' wavelets; see [14].

Assumption B.

(i) $\mathbf{V} \in (W^1_\infty(\mathbb{R}^d \times [0,T]))^d$.

(ii) $\mathrm{div}\mathbf{V} \in W^1_\infty(\mathbb{R}^d \times [0,T])$.

(iii) $R \in W^1_\infty(\mathbb{R}^d \times [0,T])$.

Now we state our first theorem on the error estimate for exact solutions with only H^1-regularity.

THEOREM 1. *Let $u(\mathbf{x},t)$ be the exact solution of problem* (1.1) *and let $U_h^n(\mathbf{x})$ be the numerical solution generated by the generic ELLAM scheme* (2.10)*. Then under Assumptions* A *and* B*, the following error estimate in L^2-norm holds:*

$$\begin{aligned} &\max_{0\leq n\leq N} \|u(\mathbf{x},t_n) - U_h^n(\mathbf{x})\|_{L^2(\mathbb{R}^d)} \\ &\leq C\Bigg[\frac{h}{\sqrt{\Delta t}}\left(\|u_0\|_{H^1(\mathbb{R}^d)} + \|f\|_{L^2(0,T;H^1(\mathbb{R}^d))}\right) \\ &\qquad +\Delta t\left(\|u_0\|_{L^2(\mathbb{R}^d)} + \|f\|_{L^2(0,T;L^2(\mathbb{R}^d))} + \|f_\tau\|_{L^2(0,T;L^2(\mathbb{R}^d))}\right)\Bigg], \end{aligned} \tag{3.2}$$

where Δt and h are the temporal and spatial step sizes, respectively, and f_τ is the total derivative, i.e., the derivative along the characteristic direction. The constant C depends only on the final time T and the norms of $\mathbf{V}, \mathrm{div}\mathbf{V}, R$ in the corresponding spaces in Assumption B*. When orthogonal wavelet schemes are used, $h = 1/2^J$ with J being the finest spatial resolution used in the wavelet schemes.*

4. Proof of Theorem 1. In this section, we first estimate the error $E_n(w)$ defined in (4.5). Then we estimate the error involving the source term defined in (2.9). Applying a discrete Gronwall inequality, we derive the final error estimate stated in Theorem 1.

Let $u_n(\mathbf{x}) := u(\mathbf{x},t_n)$ and let $P_h u_n$ be the L^2-orthogonal projection of u_n into the subspace V_h, that is,

$$(P_h u_n, v_h) = (u_n, v_h) \qquad \forall\, v_h \in V_h. \tag{4.1}$$

Now with $P_h u_n, U_h^n \in V_h$, we have the orthogonality

$$u_n - U_h^n = (u_n - P_h u_n) \oplus (P_h u_n - U_h^n), \tag{4.2}$$

$$\|u_n - U_h^n\|^2_{L^2(\mathbb{R}^d)} = \|u_n - P_h u_n\|^2_{L^2(\mathbb{R}^d)} + \|P_h u_n - U_h^n\|^2_{L^2(\mathbb{R}^d)}. \tag{4.3}$$

Here $u_n - P_h u_n$ is the approximation error, which is completely determined by u_n and the chosen approximation subspace V_h. It is independent of the numerical scheme being used: a wavelet method or a traditional finite element method. Therefore we only need to bound the second term on the right side of (4.3).

Let w be a test function such that $w(\mathbf{x}, t_n) \in V_h$. Subtracting (2.10) from (2.5) and then applying (2.6) and (2.9), we obtain

$$(P_h u_n - U_h^n, w(\mathbf{x}, t_n)) = E_n(w) + E(f, w), \tag{4.4}$$

where

$$\begin{aligned} E_n(w) := &\int_{\mathbb{R}^d} u(\mathbf{x}^*, t_{n-1}) w(\mathbf{x}, t_n) e^{\int_{t_n}^{t_{n-1}} R(\mathbf{y},s)ds} \mathbf{J}(\mathbf{x}^*, \mathbf{x}) d\mathbf{x} \\ &- \int_{\mathbb{R}^d} U_h^{n-1}(\mathbf{x}^{**}) w(\mathbf{x}, t_n) e^{-R(\mathbf{x},t_n)\Delta t} \mathbf{J}(\mathbf{x}^{**}, \mathbf{x}) d\mathbf{x}. \end{aligned} \tag{4.5}$$

From now on, we shall use C to denote a constant that is independent of the spatial and temporal mesh sizes but may depend on the final time T and the norms of the velocity field and the reaction in the corresponding spaces in Assumption B, whereas C_0 will be used for an absolute constant that does not depend on any of the aforementioned terms. All these constants may take different values in different occurrences.

The following basic estimates are easy to verify and shall be repeatedly used in this section. When Δt is small enough, we have

(i) $\mathbf{J}(\mathbf{x}^*, \mathbf{x}) \le 1 + C\Delta t$,

(ii) $\mathbf{J}(\mathbf{x}^{**}, \mathbf{x}) \le 1 + C\Delta t$,

(iii) $e^{\int_{t_n}^{t_{n-1}} R(\mathbf{y},s)ds} \le 1 + C\Delta t$,

(iv) $e^{-R(\mathbf{x},t_n)\Delta t} \le 1 + C\Delta t$.

Recall that Assumption B, part (i), implies that the velocity field is Lipschitz in both space and time. When the Euler method is used for tracking characteristics, we have

$$\mathbf{x}^{**} = \mathbf{x} + \mathbf{V}(\mathbf{x}, t_n)(t_{n-1} - t_n), \tag{4.6}$$

and hence

$$|\mathbf{x}^* - \mathbf{x}^{**}| \le C(\Delta t)^2, \tag{4.7}$$

where $C = C_0 \|\mathbf{V}\|_{(W_\infty^1(\mathbb{R}^d \times [0,T]))^d}$. Furthermore, the following estimate holds under Assumption B, parts (i) and (ii):

$$|\mathbf{J}(\mathbf{x}^*, \mathbf{x}) - \mathbf{J}(\mathbf{x}^{**}, \mathbf{x})| \le C(\Delta t)^2. \tag{4.8}$$

4.1. Estimate on error $E_n(w)$. For the error $E_n(w)$, we split it into 4 terms:

$$E_n(w) = I_n^{(1)} + I_n^{(2)} + I_n^{(3)} + I_n^{(4)}, \tag{4.9}$$

where

$$\begin{aligned} I_n^{(1)} &:= \int_{\mathbb{R}^d} \Big[u(\mathbf{x}^*, t_{n-1}) - u(\mathbf{x}^{**}, t_{n-1})\Big] w(\mathbf{x}, t_n) e^{\int_{t_n}^{t_{n-1}} R(\mathbf{y},s)ds} \mathbf{J}(\mathbf{x}^*, \mathbf{x}) d\mathbf{x}, \\ I_n^{(2)} &:= \int_{\mathbb{R}^d} u(\mathbf{x}^{**}, t_{n-1}) w(\mathbf{x}, t_n) e^{\int_{t_n}^{t_{n-1}} R(\mathbf{y},s)ds} \Big[\mathbf{J}(\mathbf{x}^*, \mathbf{x}) - \mathbf{J}(\mathbf{x}^{**}, \mathbf{x})\Big] d\mathbf{x}, \\ I_n^{(3)} &:= \int_{\mathbb{R}^d} u(\mathbf{x}^{**}, t_{n-1}) w(\mathbf{x}, t_n) \Big[e^{\int_{t_n}^{t_{n-1}} R(\mathbf{y},s)ds} - e^{-R(\mathbf{x},t_n)\Delta t}\Big] \mathbf{J}(\mathbf{x}^{**}, \mathbf{x}) d\mathbf{x}, \\ I_n^{(4)} &:= \int_{\mathbb{R}^d} \Big[u(\mathbf{x}^{**}, t_{n-1}) - U_h^{n-1}(\mathbf{x}^{**})\Big] w(\mathbf{x}, t_n) e^{-R(\mathbf{x},t_n)\Delta t} \mathbf{J}(\mathbf{x}^{**}, \mathbf{x}) d\mathbf{x}. \end{aligned} \tag{4.10}$$

Note that $I_n^{(1)}$ and $I_n^{(2)}$ in (4.10) reflect the error from inexact tracking of characteristics. These two terms will vanish when exact tracking of characteristics is available.

Estimate on $\boldsymbol{I_n^{(1)}}$ in (4.10). The H^1-stability of the exact solution (Lemma 4 in section 5) and estimate (4.7) yield

$$\left(\int_{\mathbb{R}^d} |u(\mathbf{x}^*, t_n) - u(\mathbf{x}^{**}, t_n)|^2 d\mathbf{x} \right)^{1/2} \leq C\|u_n\|_{H^1(\mathbb{R}^d)}(\Delta t)^2.$$

Therefore,

$$|I_n^{(1)}| \leq C(\Delta t)^2 \|u_n\|_{H^1(\mathbb{R}^d)} \|w(\mathbf{x}, t_n)\|_{L^2(\mathbb{R}^d)}. \tag{4.11}$$

Estimate on $\boldsymbol{I_n^{(2)}}$ in (4.10). Based on estimate (4.8), we have

$$|I_n^{(2)}| \leq C(\Delta t)^2 \|u_{n-1}\|_{L^2(\mathbb{R}^d)} \|w(\mathbf{x}, t_n)\|_{L^2(\mathbb{R}^d)}. \tag{4.12}$$

Estimate on $\boldsymbol{I_n^{(3)}}$ in (4.10). Let R_τ be the total derivative of R (along characteristic) and $\mathbf{z} = \mathbf{y}(r; \mathbf{x}, t_n)$ for $r \in [s, t_n]$; then

$$\left| e^{\int_{t_n}^{t_{n-1}} R(\mathbf{y},s)ds} - e^{-R(\mathbf{x},t_n)\Delta t} \right| \leq 2 \int_{t_{n-1}}^{t_n} \int_s^{t_n} |R_\tau(\mathbf{z}, r)| dr ds \leq \|R_\tau\|_\infty (\Delta t)^2.$$

But $R_\tau = \nabla R \cdot \mathbf{V} + R_t$, so we have

$$|I_n^{(3)}| \leq C(\Delta t)^2 \|u_{n-1}\|_{L^2(\mathbb{R}^d)} \|w(\mathbf{x}, t_n)\|_{L^2(\mathbb{R}^d)}. \tag{4.13}$$

Estimate on $\boldsymbol{I_n^{(4)}}$ in (4.10). It is easy to derive the following estimate:

$$|I_n^{(4)}| \leq (1 + C\Delta t) \|u_{n-1} - U_h^{n-1}\|_{L^2(\mathbb{R}^d)} \|w(\mathbf{x}, t_n)\|_{L^2(\mathbb{R}^d)}. \tag{4.14}$$

Substitution of estimates (4.11), (4.12), (4.13), and (4.14) into (4.9) gives us an estimate on the error $E_n(w)$ defined in (4.5):

$$\begin{aligned} |E_n(w)| \leq & \left(\frac{1}{2} + C\Delta t \right) \left(\|u_{n-1} - U_h^{n-1}\|^2_{L^2(\mathbb{R}^d)} + \|w(\mathbf{x}, t_n)\|^2_{L^2(\mathbb{R}^d)} \right) \\ & + C(\Delta t)^3 \|u_{n-1}\|^2_{L^2(\mathbb{R}^d)} + C\Delta t \|w(\mathbf{x}, t_n)\|^2_{L^2(\mathbb{R}^d)}, \end{aligned} \tag{4.15}$$

where C depends on the final time T and the norms of $\mathbf{V}, \mathrm{div}\mathbf{V}, R$ in the corresponding spaces in Assumption B.

4.2. Estimate on source term. Now we estimate the error in the approximation to the source term. Let $\mathbf{y} = \mathbf{y}(s; \mathbf{x}, t_n), s \in [t_{n-1}, t_n]$, and $\mathbf{z} = \mathbf{y}(r; \mathbf{x}, t_n), r \in [s, t_n]$. According to (2.3), (2.7), and (2.9), we have

$$E(f, w) = I_f^{(1)} + I_f^{(2)} + I_f^{(3)}, \tag{4.16}$$

where

$$\begin{aligned} I_f^{(1)} &:= \int_{t_{n-1}}^{t_n} \int_{\mathbb{R}^d} f(\mathbf{y}, s) e^{\int_{t_n}^s R(\mathbf{z},r)dr} w(\mathbf{x}, t_n) [\mathbf{J}(\mathbf{y}, \mathbf{x}) - 1] d\mathbf{x} ds, \\ I_f^{(2)} &:= \int_{t_{n-1}}^{t_n} \int_{\mathbb{R}^d} f(\mathbf{y}, s) \left[e^{\int_{t_n}^s R(\mathbf{z},r)dr} - e^{R(\mathbf{x},t_n)(s-t_n)} \right] w(\mathbf{x}, t_n) d\mathbf{x} ds, \\ I_f^{(3)} &:= \int_{t_{n-1}}^{t_n} \int_{\mathbb{R}^d} [f(\mathbf{y}, s) - f(\mathbf{x}, t_n)] \, e^{R(\mathbf{x},t_n)(s-t_n)} w(\mathbf{x}, t_n) d\mathbf{x} ds. \end{aligned} \tag{4.17}$$

For convenience, let $J_n := [t_{n-1}, t_n]$.

Estimate on $I_f^{(1)}$ in (4.17). Applying the Cauchy–Schwarz inequality first in space and then in time gives

$$\begin{aligned}\left|\int_{t_{n-1}}^{t_n}\int_{\mathbb{R}^d} f(\mathbf{y},s)e^{\int_{t_n}^{s} R(\mathbf{z},r)dr} w(\mathbf{x},t_n)[\mathbf{J}(\mathbf{y},\mathbf{x})-1]d\mathbf{x}ds\right| \\ \leq \int_{t_{n-1}}^{t_n} 4\sqrt{2}\|\mathrm{div}\mathbf{V}\|_\infty (t_n-s)\|f(\cdot,s)\|_{L^2(\mathbb{R}^d)}\|w(\mathbf{x},t_n)\|_{L^2(\mathbb{R}^d)}ds \\ \leq \frac{4\sqrt{2}}{\sqrt{3}}\|\mathrm{div}\mathbf{V}\|_\infty(\Delta t)^{\frac{3}{2}}\|f\|_{L^2(J_n;L^2(\mathbb{R}^d))}\|w(\mathbf{x},t_n)\|_{L^2(\mathbb{R}^d)}.\end{aligned}$$

Therefore, we have

$$|I_f^{(1)}| \leq \frac{2\sqrt{2}}{\sqrt{3}}\|\mathrm{div}\mathbf{V}\|_\infty\left[(\Delta t)^2\|f\|^2_{L^2(J_n,L^2(\mathbb{R}^d))}+\Delta t\|w(\mathbf{x},t_n)\|^2_{L^2(\mathbb{R}^d)}\right]. \tag{4.18}$$

Estimate on $I_f^{(2)}$ in (4.17). Similar to the above, we have

$$|I_f^{(2)}| \leq \frac{2\sqrt{2}}{\sqrt{3}}\|R\|_\infty\left[(\Delta t)^2\|f\|^2_{L^2(J_n;L^2(\mathbb{R}^d))}+\Delta t\|w(\mathbf{x},t_n)\|^2_{L^2(\mathbb{R}^d)}\right]. \tag{4.19}$$

Estimate on $I_f^{(3)}$ in (4.17). Let f_τ be the total derivative of f; then

$$\begin{aligned}\left|\int_{\mathbb{R}^d}[f(\mathbf{y},s)-f(\mathbf{x},t_n)]e^{R(\mathbf{x},t_n)(s-t_n)}w(\mathbf{x},t_n)d\mathbf{x}\right| \\ \leq 2\int_s^{t_n}\int_{\mathbb{R}^d}|f_\tau(\mathbf{z},r)|\,|w(\mathbf{x},t_n)|d\mathbf{x}dr \\ \leq 2\sqrt{2}(t_n-s)^{1/2}\|f_\tau\|_{L^2([s,t_n];L^2(\mathbb{R}^d))}\|w(\mathbf{x},t_n)\|_{L^2(\mathbb{R}^d)}.\end{aligned}$$

Therefore,

$$|I_f^{(3)}| \leq \frac{2\sqrt{2}}{3}\left[(\Delta t)^2\|f_\tau\|^2_{L^2(J_n;L^2(\mathbb{R}^d))}+\Delta t\|w(\mathbf{x},t_n)\|^2_{L^2(\mathbb{R}^d)}\right]. \tag{4.20}$$

Now we piece together the above estimates (4.18), (4.19), and (4.20) to obtain an estimate on the source term

$$\begin{aligned}|E(f,w)| \leq C\Delta t\|w(\mathbf{x},t_n)\|^2_{L^2(\mathbb{R}^d)} \\ +C(\Delta t)^2\left[\|f\|^2_{L^2(J_n;L^2(\mathbb{R}^d))}+\|f_\tau\|^2_{L^2(J_n;L^2(\mathbb{R}^d))}\right],\end{aligned} \tag{4.21}$$

where the constant C can be taken as $C=\frac{2\sqrt{2}}{\sqrt{3}}[\|\mathrm{div}\mathbf{V}\|_\infty+\|R\|_\infty+1]$.

4.3. Final estimate. Combining estimates (4.15) and (4.21) with (4.4) and taking $w(\mathbf{x},t_n)$ as $P_hu_n-U_h^n$, we get

$$\begin{aligned}\|P_hu_n-U_h^n\|^2_{L^2(\mathbb{R}^d)} \leq \left(\frac{1}{2}+C\Delta t\right)\left(\|P_hu_n-U_h^n\|^2_{L^2(\mathbb{R}^d)}+\|u_{n-1}-U_h^{n-1}\|^2_{L^2(\mathbb{R}^d)}\right) \\ +C(\Delta t)^3\|u_{n-1}\|^2_{L^2(\mathbb{R}^d)}+C(\Delta t)^2\left[\|f\|^2_{L^2(J_n;L^2(\mathbb{R}^d))}+\|f_\tau\|^2_{L^2(J_n;L^2(\mathbb{R}^d))}\right].\end{aligned}$$

Note that $\|u_{n-1}-U_h^{n-1}\|^2 = \|u_{n-1}-P_hu_{n-1}\|^2 + \|P_hu_{n-1}-U_h^{n-1}\|^2$. Taking Δt small enough such that $C\Delta t \le 1/2$, we obtain

$$\begin{aligned}\|P_hu_n-U_h^n\|^2 \le{}& \left(\frac{1}{2}+C\Delta t\right)\left(\|P_hu_n-U_h^n\|^2+\|P_hu_{n-1}-U_h^{n-1}\|^2\right)\\ &+\|u_{n-1}-P_hu_{n-1}\|^2+C(\Delta t)^3\|u_{n-1}\|^2\\ &+C(\Delta t)^2\left[\|f\|^2_{L^2(J_n;L^2(\mathbb{R}^d))}+\|f_\tau\|^2_{L^2(J_n;L^2(\mathbb{R}^d))}\right].\end{aligned}$$

Summing both sides for $n=1,2,\ldots,m$ ($m\le N$) and canceling identical terms (also $U_h^0=P_hu_0$), we get

$$\begin{aligned}\|P_hu_m-U_h^m\|^2 \le{}& \left(\frac{1}{2}+C\Delta t\right)\|P_hu_m-U_h^m\|^2+C\Delta t\sum_{n=1}^{m-1}\|P_hu_n-U_h^n\|^2\\ &+\sum_{n=0}^{m-1}\|u_n-P_hu_n\|^2+C(\Delta t)^3\sum_{n=0}^{m-1}\|u_n\|^2\\ &+C(\Delta t)^2\left[\|f\|^2_{L^2(0,T;L^2(\mathbb{R}^d))}+\|f_\tau\|^2_{L^2(0,T;L^2(\mathbb{R}^d))}\right].\end{aligned}$$

Taking Δt small enough such that $C\Delta t\le 1/4$, we have

$$\begin{aligned}\|P_hu_m-U_h^m\|^2 \le{}& C\Delta t\sum_{n=1}^{m-1}\|P_hu_n-U_h^n\|^2\\ &+C_0\sum_{n=0}^{m-1}\|u_n-P_hu_n\|^2+C(\Delta t)^3\sum_{n=0}^{m-1}\|u_n\|^2\\ &+C(\Delta t)^2\left[\|f\|^2_{L^2(0,T;L^2(\mathbb{R}^d))}+\|f_\tau\|^2_{L^2(0,T;L^2(\mathbb{R}^d))}\right].\end{aligned}$$

Applying the discrete Gronwall inequality, we obtain

$$\begin{aligned}\|P_hu_m-U_h^m\|^2_{L^2(\mathbb{R}^d)} \le{}& C\sum_{n=0}^{m-1}\|u_n-P_hu_n\|^2_{L^2(\mathbb{R}^d)}+C(\Delta t)^3\sum_{n=0}^{m-1}\|u_n\|^2_{L^2(\mathbb{R}^d)}\\ &+C(\Delta t)^2\left[\|f\|^2_{L^2(0,T;L^2(\mathbb{R}^d))}+\|f_\tau\|^2_{L^2(0,T;L^2(\mathbb{R}^d))}\right].\end{aligned} \tag{4.22}$$

The first two terms on the right side of the above estimate reflect the error buildup during the iterative process in numerical scheme (2.10). However, this can be controlled through the stability of the exact solution.

For the IVP to the linear convection-reaction equation in conservative form (1.1), Lemmas 3 and 4 in section 5 hold under the conditions in Assumption B and yield

$$\max_{0\le n\le N}\|u_n\|_{L^2(\mathbb{R}^d)} \le C\left(\|u_0\|_{L^2(\mathbb{R}^d)}+\|f\|_{L^2(0,T;L^2(\mathbb{R}^d))}\right), \tag{4.23}$$

$$\max_{0\le n\le N}\|u_n\|_{H^1(\mathbb{R}^d)} \le C\left(\|u_0\|_{H^1(\mathbb{R}^d)}+\|f\|_{L^2(0,T;H^1(\mathbb{R}^d))}\right), \tag{4.24}$$

where C depends only on the final time T and the norms of $\mathbf{V}, \mathrm{div}\mathbf{V}, R$ in the corresponding spaces in Assumption B.

Applying (4.23), we get

$$\sum_{n=0}^{m-1}\|u_n\|^2_{L^2(\mathbb{R}^d)} \le \frac{C}{\Delta t}\left(\|u_0\|^2_{L^2(\mathbb{R}^d)}+\|f\|^2_{L^2(0,T;L^2(\mathbb{R}^d))}\right). \tag{4.25}$$

Combining Assumption A with (4.24), we obtain

$$\sum_{n=0}^{m-1} \|u_n - P_h u_n\|_{L^2(\mathbb{R}^d)}^2 \le C\frac{h^2}{\Delta t}\left(\|u_0\|_{H^1(\mathbb{R}^d)}^2 + \|f\|_{L^2(0,T;H^1(\mathbb{R}^d))}^2\right). \tag{4.26}$$

Combining (3.1), (4.3), (4.22), and the above two estimates, we finish the proof of Theorem 1. □

5. Stability of exact solutions.

In this section, we prove the stability lemmas used in the last section about the exact solution to a linear convection-reaction equation. All results are first established for the solution to an IVP to the linear transport equation in nonconservative form:

$$\begin{cases} u_t + \mathbf{V}\cdot\nabla u = cu + f, & (\mathbf{x},t)\in\mathbb{R}^d\times(0,T],\\ u(\mathbf{x},0) = u_0(\mathbf{x}), & \mathbf{x}\in\mathbb{R}^d, \end{cases} \tag{5.1}$$

where $u_0 \in L^2(\mathbb{R}^d)$. However, the results can be easily passed to the linear convection-reaction equation in the conservative form (1.1) by simply setting $c = -(\mathrm{div}\mathbf{V} + R)$.

In this section, Assumption B, part (i), can be relaxed to $\mathbf{V}\in L^1(0,T;(W^1_\infty(\mathbb{R}^d))^d)$; that is, $\mathbf{V}$ is Lipschitz with respect to only spatial variables. Although it is possible to prove the existence, uniqueness, and regularity results below under even weaker assumptions, we restrict our attention to the case $\mathbf{V}\in L^1(0,T;(W^1_\infty(\mathbb{R}^d))^d)$ to keep the assumptions simple.

LEMMA 2 (existence and uniqueness). *Suppose* $\mathbf{V}\in L^1(0,T;(W^1_\infty(\mathbb{R}^d))^d)$, $c\in L^1(0,T;L^\infty(\mathbb{R}^d))$, *and* $f\in L^1(0,T;L^2(\mathbb{R}^d))$. *If* $u_0\in L^2(\mathbb{R}^d)$, *then there exists a unique solution to* (5.1) *in* $L^\infty(0,T;L^2(\mathbb{R}^d))$ *and the solution can be explicitly expressed as*

$$u(\mathbf{x},t) = u_0(\mathbf{y}(0;\mathbf{x},t))e^{\int_0^t c(\mathbf{y}(s;\mathbf{x},t),s)ds} + \int_0^t f(\mathbf{y}(s;\mathbf{x},t),s)e^{\int_s^t c(\mathbf{y}(r;\mathbf{x},t),r)dr}ds. \tag{5.2}$$

The uniqueness of the solution u in Lemma 2 is a corollary of the results in [16]. The solution formula (5.2) can be derived using the results in [16], the techniques in [22], and standard density arguments.

LEMMA 3 (L^2-stability). *Assume that* $\mathbf{V}, c,$ *and* f *satisfy the conditions of Lemma 2. If* $u_0\in L^2(\mathbb{R}^d)$, *then the unique solution satisfies*

$$\|u(\cdot,t)\|_{L^2(\mathbb{R}^d)} \le C\left[\|u_0\|_{L^2(\mathbb{R}^d)} + \|f\|_{L^1(0,t;L^2(\mathbb{R}^d))}\right] \qquad \forall t\in[0,T], \tag{5.3}$$

where C *can be taken as* $C = e^{\int_0^t[\|c(\cdot,r)\|_\infty + \frac{1}{2}\|\mathrm{div}\mathbf{V}(\cdot,r)\|_\infty]dr}$.

Proof. Applying the triangle and Minkowski inequalities to the representation formula (5.2), we obtain

$$\begin{aligned} \|u(\cdot,t)\|_2 \le e^{\int_0^t \|c(\cdot,s)\|_\infty ds}\Bigg[\left(\int_{\mathbb{R}^d}|u_0(\mathbf{y}(0;\mathbf{x},t))|^2 d\mathbf{x}\right)^{\frac{1}{2}} \\ + \int_0^t\left(\int_{\mathbb{R}^d}|f(\mathbf{y}(s;\mathbf{x},t),s)|^2 d\mathbf{x}\right)^{\frac{1}{2}} ds\Bigg]. \end{aligned} \tag{5.4}$$

Let $\mathbf{J}(\mathbf{y},\mathbf{x})$ be the Jacobian of mapping $\mathbf{x}\mapsto\mathbf{y} := \mathbf{y}(s;\mathbf{x},t)$. It is known [13] that

$$\mathbf{J}(\mathbf{y},\mathbf{x}) = e^{\int_t^s \mathrm{div}\mathbf{V}(\mathbf{y}(r;\mathbf{x},t),r)dr}. \tag{5.5}$$

Hence, for any $s \leq t \in [0,T]$, we have

$$e^{-\int_s^t \|\mathrm{div}\mathbf{V}(\cdot,r)\|_\infty dr} \leq \mathbf{J}(\mathbf{y},\mathbf{x}) \leq e^{\int_s^t \|\mathrm{div}\mathbf{V}(\cdot,r)\|_\infty dr}. \tag{5.6}$$

Of course, we can take $\mathbf{y} := \mathbf{y}(s;\mathbf{x},t)$ as the starting point. Then the reversibility of flow ensures that the Jacobian $\mathbf{J}(\mathbf{x},\mathbf{y})$ of the inverse mapping $\mathbf{y}(s;\mathbf{x},t) \mapsto \mathbf{x}$ satisfies the same estimate. A change of variables in (5.4) and then an application of the above estimates on Jacobians conclude the proof. ☐

Next we shall impose some additional conditions on c and f so that the solution $u \in L^\infty(0,T;H^1(\mathbb{R}^d))$, provided $u_0 \in H^1(\mathbb{R}^d)$. In other words, under these sufficient conditions on $\mathbf{V}$, c, and f, the solution operator $E_t : u_0 \mapsto u(\cdot,t)$ is a bounded operator from $H^1(\mathbb{R}^d)$ to $H^1(\mathbb{R}^d)$.

LEMMA 4 (H^1-stability). *Assume that* $\mathbf{V} \in L^1(0,T;(W^1_\infty(\mathbb{R}^d))^d)$*,* $c \in L^1(0,T;W^1_\infty(\mathbb{R}^d))$*, and* $f \in L^1(0,T;H^1(\mathbb{R}^d))$*. If* $u_0 \in H^1(\mathbb{R}^d)$*, then the unique weak solution satisfies*

$$\|u(\cdot,t)\|_{H^1(\mathbb{R}^d)} \leq C\big(\|u_0\|_{H^1(\mathbb{R}^d)} + \|f\|_{L^1(0,t;H^1(\mathbb{R}^d))}\big) \quad \forall t \in [0,T], \tag{5.7}$$

where the constant C *depends only on* $\|\mathbf{V}\|_{L^1(0,T;(W^1_\infty(\mathbb{R}^d))^d)}$ *and* $\|c\|_{L^1(0,T;W^1_\infty(\mathbb{R}^d))}$.

Proof. Using the chain rule to the solution formula (5.2), we can derive expressions for all partial derivatives $\frac{\partial u}{\partial x_i}, 1 \leq i \leq d$, which involve spatial partial derivatives of the flow $\mathbf{y}(s;\mathbf{x},t), c$, and f. Let $\mathbf{z}(s;\mathbf{x},t)$ be any partial derivative of $\mathbf{y} := \mathbf{y}(s;\mathbf{x},t)$ with respect to one of the variables $x_1,\ldots,x_d$. We shall derive a uniform bound for $\mathbf{z}(s;\cdot,t)$. Let $|\mathbf{z}(s;\mathbf{x},t)|$ be the usual Euclidean norm for a vector in $\mathbb{R}^d$. The following assertion was proven in [8, 22]: Let $\mathbf{V}$ satisfy

$$|(\mathbf{V}(\mathbf{x},t) - \mathbf{V}(\mathbf{y},t)) \cdot (\mathbf{x} - \mathbf{y})| \leq K(t)|\mathbf{x} - \mathbf{y}|^2;$$

then

$$|\mathbf{z}(s;\mathbf{x},t)| \leq |\mathbf{z}(t;\mathbf{x},t)|e^{\int_0^T K(r)dr} = e^{\int_0^T K(r)dr}.$$

Here we have used the fact $\mathbf{y}(t;\mathbf{x},t) = \mathbf{x}$ to get $|\mathbf{z}(t;\mathbf{x},t)| = 1$. It is not difficult to verify that if $\mathbf{V} \in L^1(0,T;(W^1_\infty(\mathbb{R}^d))^d)$, then the above inequality holds with $K(t) = \|\mathbf{V}(\cdot,t)\|_{W^1_\infty(\mathbb{R}^d)}$. Therefore,

$$\|\mathbf{z}(s,\cdot,t)\|_\infty := \text{ess sup}_{\mathbf{x}\in\mathbb{R}^d}|\mathbf{z}(s,\mathbf{x},t)| \leq e^{\|\mathbf{V}\|_{L^1(0,T;(W^1_\infty(\mathbb{R}^d))^d)}}. \tag{5.8}$$

This is the standard estimate for a Lipschitz flow; see (1.4) in [7]. Based on the uniform bound (5.8), we can estimate all partial derivatives as follows:

$$\Big\|\frac{\partial u}{\partial x_i}(\cdot,t)\Big\|_{L^2(\mathbb{R}^d)} \leq C\big(\|u_0\|_{H^1(\mathbb{R}^d)} + \|f\|_{L^1(0,T;H^1(\mathbb{R}^d))}\big),$$

where the constant C depends only on the norms of $\mathbf{V}$ in $L^1(0,T;(W^1_\infty(\mathbb{R}^d))^d)$ and c in $L^1(0,T;W^1_\infty(\mathbb{R}^d))$. We finish the proof by combining the above estimate and the result of Lemma 3. ☐

6. Extension to Besov spaces. Besov spaces provide subtle characterization of regularity of functions. Interpolation of spaces and operators is a classical topic in harmonic analysis and has many interesting applications. In this section, we cite only the minimal requisite for our discussion. Readers are referred to [3, 15] for complete accounts of these two topics.

Besov spaces involve moduli of smoothness rather than the distributional derivatives used for Sobolev spaces. Let $1 \leq p \leq \infty$, $f \in L^p(\mathbb{R}^d)$, $\mathbf{h} \in \mathbb{R}^d$, and let $\Delta_{\mathbf{h}}$ be the usual difference operator: $\Delta_{\mathbf{h}} f(\mathbf{x}) := f(\mathbf{x}+\mathbf{h}) - f(\mathbf{x})$. Let k be a positive integer and $t > 0$. The kth modulus of smoothness of function f is defined as

$$\omega_k(f,t)_p := \sup_{|\mathbf{h}| \leq t} \|\Delta^k_{\mathbf{h}}(f,\mathbf{x})\|_{L^p(\mathbb{R}^d)}. \tag{6.1}$$

Suppose $\alpha > 0$ and $0 < q \leq \infty$. Let k be a positive integer such that $k > \alpha$. The Besov space $B^\alpha_{p,q}(\mathbb{R}^d)$ consists of functions $f \in L^p(\mathbb{R}^d)$ (if $p < \infty$) or $C(\mathbb{R}^d)$ (if $p = \infty$) such that

$$\|f\|_{B^\alpha_{p,q}} := \begin{cases} \|f\|_{L^p} + \left\{ \int_0^\infty \left[t^{-\alpha}\omega_k(f,t)_p\right]^q \frac{dt}{t} \right\}^{1/q}, & 0 < q < \infty, \\ \|f\|_{L^p} + \sup\limits_{t>0} t^{-\alpha}\omega_k(f,t)_p, & q = \infty, \end{cases} \tag{6.2}$$

is finite.

It is known that $B^\alpha_{p,q_1} \subset B^\alpha_{p,q_2}$ if $q_1 < q_2$. When $p = 2$, one has $B^\alpha_{2,2} = H^\alpha$ with equivalent norms.

Interpolation of spaces can be defined by K-functionals. Let $X_1 \subset X_0$ be Banach spaces. For any $f \in X_0$ and $t > 0$,

$$K(f,t) := K(f,t;X_0,X_1) := \inf_{g \in X_1} \{\|f-g\|_{X_0} + t\|g\|_{X_1}\}. \tag{6.3}$$

Here t is viewed as a penalty factor. The intermediate space $[X_0, X_1]_{\theta,q}$ consists of all $f \in X_0$ for which

$$\|f\|_{\theta,q} := \begin{cases} \left\{ \int_0^\infty \left[t^{-\theta}K(f,t)\right]^q \frac{dt}{t} \right\}^{1/q}, & 0 < q < \infty, \\ \sup\limits_{t>0} t^{-\theta}K(f,t), & q = \infty, \end{cases} \tag{6.4}$$

is finite. Obviously, $X_1 \subset [X_0, X_1]_{\theta,q} \subset X_0$.

Amazingly, the K-functional for the pair (L^p, W^k_p) is equivalent to the modulus of smoothness, and hence we have the following (see [3, 15]).

Lemma 5 (interpolation of spaces). *Let $k > 0$ be an integer and let $1 \leq p \leq \infty$. For any $0 < \theta < 1, 0 < q \leq \infty$, we have*

$$\left[L^p(\mathbb{R}^d), W^k_p(\mathbb{R}^d)\right]_{\theta,q} = B^{\theta k}_{p,q}(\mathbb{R}^d). \tag{6.5}$$

Especially,

$$\left[L^2(\mathbb{R}^d), H^1(\mathbb{R}^d)\right]_{\theta,q} = B^{\theta}_{2,q}(\mathbb{R}^d). \tag{6.6}$$

Lemma 6 (interpolation of operators). *Suppose that $X_1 \subset X_0$ and Y are Banach spaces. If T is a linear operator from X_i to Y with norm M_i $(i = 0,1)$, then T is also a linear operator from $[X_0, X_1]_{\theta,q}$ to Y with norm not exceeding $M_0^{1-\theta}M_1^\theta$ for any $0 < \theta < 1, 0 < q \leq \infty$.*

THEOREM 7. *Suppose that Assumptions* A *and* B *are satisfied. Then for any* $0 < \theta < 1, 0 < q \leq \infty$, *the following error estimate holds:*

$$\begin{aligned}&\max_{0\leq n\leq N} \|u(\mathbf{x}, t_n) - U_h^n(\mathbf{x})\|_{L^2(\mathbb{R}^d)} \\ &\quad\leq C\Big\{ \Big[(h/\sqrt{\Delta t})^\theta + (\Delta t)^\theta\Big] \|u_0\|_{B^\theta_{2,q}(\mathbb{R}^d)} \\ &\qquad +(h/\sqrt{\Delta t} + \Delta t) \left[\|f\|_{L^2(0,T;H^1(\mathbb{R}^d))} + \|f_\tau\|_{L^2(0,T;L^2(\mathbb{R}^d))}\right] \Big\},\end{aligned} \tag{6.7}$$

where Δt *and* h *bear the same meaning as that in Theorem* 1, *but* C *may additionally depend on* θ.

Proof. We split the error as

$$u_n - U_h^n = [u_n^{(1)} - U_h^{(1),n}] + [u_n^{(2)} - U_h^{(2),n}],$$

where $u_n^{(1)}$ and $U_h^{(1),n}$ are the exact and numerical solutions for problem (1.1) without source term (i.e., $f \equiv 0$), whereas $u_n^{(2)}$ and $U_h^{(2),n}$ are the exact and numerical solutions for the problem with no initial data ($u_0 \equiv 0$).

Recall that in [24] we proved the numerical solution is L^2-stable. Here (4.23) is the L^2-stability of the exact solution. Combined, they imply

$$\|u_n^{(1)} - U_h^{(1),n}\|_{L^2(\mathbb{R}^d)} \leq C\|u_0\|_{L^2(\mathbb{R}^d)}$$

for $u_0 \in L^2(\mathbb{R}^d)$. If $u_0 \in H^1(\mathbb{R}^d)$, then by Theorem 1 we have

$$\|u_n^{(1)} - U_h^{(1),n}\|_{L^2(\mathbb{R}^d)} \leq C(h/\sqrt{\Delta t} + \Delta t)\|u_0\|_{H^1(\mathbb{R}^d)}.$$

Applying Lemmas 5 and 6 to the linear operator $E_n : u_0 \mapsto u_n^{(1)} - U_h^{(1),n}$, we obtain

$$\begin{aligned}\|u_n^{(1)} - U_h^{(1),n}\|_{L^2(\mathbb{R}^d)} &\leq C(h/\sqrt{\Delta t} + \Delta t)^\theta \|u_0\|_{B^\theta_{2,q}(\mathbb{R}^d)} \\ &\leq C\Big[(h/\sqrt{\Delta t})^\theta + (\Delta t)^\theta\Big] \|u_0\|_{B^\theta_{2,q}(\mathbb{R}^d)}.\end{aligned}$$

By Theorem 1 again, we have

$$\|u_n^{(2)} - U_h^{(2),n}\|_{L^2(\mathbb{R}^d)} \leq C(h/\sqrt{\Delta t} + \Delta t)\big(\|f\|_{L^2(0,T;H^1(\mathbb{R}^d))} + \|f_\tau\|_{L^2(0,T;L^2(\mathbb{R}^d))}\big).$$

Then the conclusion of Theorem 7 follows from a triangle inequality. □

7. Optimality of our error estimates. For solutions with H^1- or even lower regularity, we can use Haar wavelets to carry out our numerical approximations. Haar wavelets are the simplest wavelets and have only one vanishing moment. We know that the order of approximation accuracy is usually the minimum of the order of smoothness of the function being approximated and the order of the method, which is the number of vanishing moments for orthogonal wavelets. For a solution $u \in H^1$, Assumption A in section 3 is satisfied for the approximation subspace V_h generated by Haar wavelets.

However, in the original formulation of ELLAM discussed in subsection 2.1, the test function $w(\mathbf{x}, t)$ is assumed to be in $H^1(\mathbb{R}^d)$ for any $t \in (t_{n-1}, t_n]$ and required to satisfy the adjoint equation (2.2). But Haar scaling functions and wavelets are not in $H^1(\mathbb{R})$. However, mollifications with any cut-off function can be applied to Haar scaling functions or wavelets [18], so that (2.2) is satisfied for the mollifications. Then, based on the L^2-stability of the exact solution and density arguments, we can still get the reference equation (2.5) when the test function $w(x,t)$ is such that $w(x,t_n)$ is taken as a Haar scaling function or wavelet. Numerical scheme (2.10) can be established for Haar scaling functions and wavelets without difficulty. Moreover, when Haar scaling functions are used in the ELLAM formulation, the coefficients are exactly the cell averages of the unknown function on dyadic cells, thus we have local conservation of mass.

Let us consider a simple convection equation: $u_t - u_x = 0$, with a time step $\Delta t = \alpha h$ $(0 < \alpha < 1)$. For this case, the ELLAM scheme with Haar scaling functions becomes a monotone scheme, which is widely used for hyperbolic conservation laws. Monotone schemes are a special case of linear formal first order schemes; see [4] for details. According to Theorem 4.4 in [4], there exists a constant C_1 such that

$$\sup_{\|u_0\|_{H^1(\mathbb{R})}\leq 1} \|u(\mathbf{x},T) - U_h^N(\mathbf{x})\|_{L^2(\mathbb{R})} \geq C_1 h^{1/2}. \tag{7.1}$$

An explicit value for C_1 can be derived using a modification of the argument in [23].

On the other hand, for this equation with any initial condition $u_0 \in H^1(\mathbb{R})$ and $\Delta t = \alpha h$ $(0 < \alpha < 1)$, our Theorem 1 implies an upper bound for the error in the numerical solution at the final time T:

$$\|u(\mathbf{x},T) - U_h^N(\mathbf{x})\|_{L^2(\mathbb{R})} \leq C_2 h^{1/2}. \tag{7.2}$$

The above lower and upper bounds imply that our error estimate is optimal for this case for the class of initial conditions $u_0 \in H^1$. We also want to point out that it is possible to give more examples with an optimal rate 1/2: for a different equation or a different subspace generated by smooth basis functions.

Regarding the term $h/\sqrt{\Delta t}$ in the error estimates, our understanding is that it reflects the behavior of numerical schemes when only H^1-smoothness is assumed for exact solutions. From approximation theory, we know that only first order $\mathcal{O}(h)$ approximation accuracy can be achieved at each fixed time step. For time-marching schemes, error will accumulate. However, the buildup of error can be controlled by the stability of solutions. In our theoretical estimates, we derive an upper bound for the error in the form $h/\sqrt{\Delta t}$. A similar error estimate with adverse dependence on Δt in the form

$$\|u - u_h\|_{L^\infty(0,T;L^1(\mathbb{R}))} \leq \|u_0\|_{TV}\left(h + \frac{2\sqrt{T}}{\sqrt{3}}\frac{h}{\sqrt{\Delta t}}\right)$$

can also be found in [20] for one-dimensional nonlinear conservation law. Here u_h is the numerical solution and $\|\cdot\|_{TV}$ denotes the total variation.

To balance the two parts in our error estimates, the optimal choice for Δt is $\Delta t = Ch^{2/3}$. No lower bound for the error for the case $\Delta t = Ch^\beta$ $(\beta \neq 1)$ is covered in [4, 23]. It is also almost impossible to derive a general error estimate that is optimal for all schemes and all $0 < \beta < \infty$. But the rates observed in our numerical experiments in the next section are close to our theoretical rates for the case $\Delta t = Ch^{2/3}$.

It might not be exciting if small time steps Δt have to be used, e.g., for the traditional finite difference methods that are subject to the CFL condition. However, our ELLAM schemes are CFL-free [24]. So we are allowed to use relatively large time steps. This saves computations while retaining accuracy since information from characteristics are exploited.

8. Numerical experiments. In this section, we present one-dimensional numerical experiments to illustrate the theoretical results proven in previous sections.

We consider two examples with $V(x,t) = 1$, $R(x,t) = 0.2\sin t$, $f(x,t) = 0$, $\Omega = [0,2]$, and $T = 1$. The exact solution is then given by $u(x,t) = u_0(\xi)e^{0.2(\cos t-1)}$, where ξ is obtained by back-tracking the characteristic from (x,t) to $(\xi,0)$.

Example 1. The initial condition is specified as a cusp function:

$$u_0(x) = \begin{cases} A\Big(1 - \Big|\dfrac{x-\alpha}{\beta}\Big|^{\frac{1}{2}+\gamma}\Big) & \text{if } |x-\alpha| \le \beta, \\ 0 & \text{otherwise,} \end{cases} \tag{8.1}$$

where $\alpha \in \mathbb{R}, \beta > 0$, and $0 < \gamma \le 1/2$. It can be verified that $u_0 \in H^{1+\delta}(\mathbb{R})$ for any $0 < \delta < \gamma$.

In the numerical experiment, we take $A = 1, \alpha = 0.5, \beta = 0.25, \gamma = 0.01$. So the initial data is barely in $H^1(\mathbb{R})$, or, precisely, $u_0 \in H^{1+\delta}$ for $0 < \delta < 0.01$. The second order Daubechies scaling function and wavelet are used in the wavelet schemes.

We can attain only first order approximation in space for the initial data since it is barely in $H^1(\mathbb{R})$. The adverse dependence of the error estimate on Δt in section 3 indicates that time discretization has to be done carefully. More time steps do not necessarily mean better approximations. The best result can be attained when $\Delta t = h/\sqrt{\Delta t}$, that is, $\Delta t = h^{2/3}$, if the constants in the estimate are ignored. In other words, the wavelet schemes have convergence rate $h^{2/3}$. In Table 8.1, the numerical solution at the final time step has a convergence rate 0.74 in h, which is just about 10% better than the theoretical estimate 2/3.

TABLE 8.1
Convergence rates in h for Example 1.

h	$\Delta t = Ch^{2/3}$	$\|u_0 - U_0\|_{L^2}$	$\|u_T - U_T\|_{L^2}$
$1/2^6$	1/16	7.502E-3	5.134E-3
$1/2^9$	1/64	9.181E-4	1.114E-4
$1/2^{12}$	1/256	1.124E-4	2.457E-4
$1/2^{15}$	1/1024	1.376E-5	4.880E-5
Convergence rates		1.01	0.74

Example 2. The initial condition is the indicator function $\chi_{[\alpha,\beta]}$ of the interval $[\alpha,\beta]$:

$$u_0(x) = \chi_{[\alpha,\beta]} = \begin{cases} 1 & x \in [\alpha,\beta], \\ 0 & \text{otherwise.} \end{cases} \tag{8.2}$$

It is known that, for any interval I containing $[\alpha,\beta]$ in its interior, we have $\chi_{[\alpha,\beta]} \in H^{\frac{1}{2}-\delta}(I)$ for any $0 < \delta \le \frac{1}{2}$ but $\notin H^{\frac{1}{2}}(I)$. Direct calculations indicate that

$$\|\chi_{[\alpha,\beta]}\|_{H^{\frac{1}{2}-\delta}(I)} = \mathcal{O}(\delta^{-\frac{1}{2}}) \to \infty \quad \text{as } \delta \to 0. \tag{8.3}$$

Hence, the approximation order could not be characterized well using norms in Sobolev spaces. This is one place where we should use Besov spaces. It can be verified that $\chi_{[\alpha,\beta]} \in B^{\frac{1}{2}}_{2,\infty}(I)$ but $\notin B^{\frac{1}{2}}_{2,q}(I)$ for $q < \infty$.

The numerical results for $[\alpha,\beta] = [0.25, 0.75]$ are shown in Table 8.2. The order of approximation to the initial data (u_0) is exactly 1/2 because $u_0 \in B^{\frac{1}{2}}_{2,\infty}$. After the

Table 8.2
Convergence rates in h for Example 2.

h	$\Delta t = Ch^{2/3}$	$\|u_0 - U_0\|_{L^2}$	$\|u_T - U_T\|_{L^2}$
$1/2^7$	1/20	3.232E-2	5.375E-2
$1/2^{10}$	1/80	1.142E-2	2.248E-2
$1/2^{13}$	1/320	4.040E-3	1.151E-2
$1/2^{16}$	1/1280	1.428E-3	5.053E-3
Convergence rates		0.50	0.38

time-marching procedure, the approximation error at the final time step is of order 0.38, close to the theoretical estimate 1/3.

In these two numerical examples, convergence rates are a little better than the theoretically proven rates because the velocity field is nice (but we can easily compute the exact solutions in these cases). If the velocity field is exactly Lipschitz, i.e., it has minimal regularity, we expect numerical results to be even closer to the theoretical estimates.

9. Concluding remarks. Some similar results concerning convergence rates of Lagrangian–Galerkin methods for convection-dominated diffusion problems are presented in [2]. Their estimates are uniform in the small diffusion parameter and can be carried over to the hyperbolic limit case—linear convection equations without a reaction term. Their results are consistent with ours but require a smoother velocity field ($\mathbf{V} \in C(0, T; C^2(\overline{\Omega}))$).

The assumption $R \in W^1_\infty(\mathbb{R}^d \times [0, T])$ in our paper means some smoothness in the reaction term is required. Note that the test functions in ELLAM rely on the properties of the reaction along characteristics. Generally speaking, this type of smoothness is needed for time-marching schemes; otherwise we could not use well the information of the solutions at previous time steps.

In section 6, we did not discuss interpolations on the source term. The limitation is due to the way the source term is truncated in numerical scheme (2.10). Recall that in (2.7) we approximate the double integral for the source term by a single integral. This assumes some smoothness of the source term along the temporal direction or the characteristic direction. In return, the computational cost for the source term is reduced. Of course, the numerical scheme can be modified to allow even lower regularity for the source term, but accordingly the computational cost will increase.

For the wavelet scheme with adaptive compression, i.e., Scheme III in [24], computational cost will be significantly reduced through thresholding wavelet coefficients in the smooth regions. But the approximation will deteriorate as the threshold parameter is increased. A good choice of the threshold, in other words, a quantitative description of the trade-off between computational cost and approximation accuracy, is a delicate issue of nonlinear approximation and is already under our investigation.

REFERENCES

[1] C. Bacuta, J. H. Bramble, and J. Xu, *Regularity estimates for elliptic boundary value problems in Besov spaces*, Math. Comp., 72 (2003), pp. 1577–1595.

[2] M. Bause and P. Knabner, *Uniform error analysis for Lagrangian–Galerkin approximations of convection-dominated problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1954–1984.

[3] C. Bennett and R. Sharpley, *Interpolation of Operators*, Academic Press, San Diego, CA, 1988.

[4] P. Brenner, V. Thomée, and, L. B. Wahlbin, *Besov Spaces and Applications to Difference Methods for Initial Value Problems*, Lecture Notes in Math. 434, Springer-Verlag, Berlin, 1975.

[5] M. A. Celia, T. F. Russell, I. Herrera, and R. E. Ewing, *An Eulerian-Lagrangian localized adjoint method for the advection-diffusion equation*, Adv. Water Resour., 13 (1990), pp. 187–206.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[7] F. Colombini and N. Lerner, *Uniqueness of continuous solutions for BV vector fields*, Duke Math. J., 111 (2002), pp. 357–384.

[8] E. Conway, *Generalized solutions of linear differential equations with discontinuous coefficients and the uniqueness question for multidimensional quasilinear conservation laws*, J. Math. Anal. Appl., 18 (1967), pp. 238–251.

[9] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228.

[10] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[11] C. N. Dawson, T. F. Dupont, and M. F. Wheeler, *The rate of convergence of the modified method of characteristics for linear advection equation in one dimension*, in Mathematics for Large Scale Computing, J. C. Diaz, ed., Marcel Dekker, New York, 1989, pp. 115–126.

[12] C. N. Dawson, T. F. Russell, and M. F. Wheeler, *Some improved error estimates for the modified method of characteristics*, SIAM J. Numer. Anal., 26 (1989), pp. 1487–1512.

[13] B. Desjardins, *A few remarks on ordinary differential equations*, Comm. Partial Differential Equations, 21 (1996), pp. 1667–1703.

[14] R. A. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.

[15] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.

[16] R. DiPerna and P. Lions, *Ordinary differential equations, transport theory, and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[17] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problem based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[18] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[19] R. E. Ewing and H. Wang, *An optimal-order estimate for Eulerian–Lagrangian localized adjoint methods for variable-coefficient advection-reaction problems*, SIAM J. Numer. Anal., 33 (1996), pp. 318–348.

[20] B. J. Lucier, *Error bounds for the methods of Glimm, Godunov, and LeVeque*, SIAM J. Numer. Anal., 22 (1985), pp. 1074–1081.

[21] K. W. Morton, A. Priestley, and E. Süli, *Stability of the Lagrange–Galerkin method with nonexact integration*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 625–653.

[22] G. Petrova and B. Popov, *Linear transport equations with μ-monotone coefficients*, J. Math. Anal. Appl., 260 (2001), pp. 307–324.

[23] T. Tang and Z. H. Teng, *The sharpness of Kuznetsov's $O(\sqrt{\Delta x})L^1$-error estimate for monotone difference schemes*, Math. Comput., 64 (1995), pp. 581–589.

[24] H. Wang and J. Liu, *Development of CFL-free, explicit schemes for multidimensional advection-reaction equations*, SIAM J. Sci. Comput., 23 (2001), pp. 1418–1438.

[25] H. Wang, X. Shi, and R. E. Ewing, *An ELLAM scheme for multidimensional advection-reaction equations and its optimal-order error estimate*, SIAM J. Numer. Anal., 38 (2001), pp. 1846–1885.

SIAM J. Numer. Anal.
Vol. 43, No. 2, pp. 540–558

ERROR BOUNDS FOR MONOTONE APPROXIMATION SCHEMES FOR HAMILTON–JACOBI–BELLMAN EQUATIONS[*]

GUY BARLES[†] and ESPEN R. JAKOBSEN[‡]

Abstract. We obtain error bounds for monotone approximation schemes of Hamilton–Jacobi–Bellman equations. These bounds improve previous results of Krylov and the authors. The key step in the proof of these new estimates is the introduction of a switching system which allows the construction of approximate, (almost) smooth supersolutions for the Hamilton–Jacobi–Bellman equation.

Key words. Hamilton–Jacobi–Bellman equations, switching system, viscosity solution, approximation schemes, finite difference methods, convergence rate, error bound

AMS subject classifications. 65N15, 65N06, 35J60, 35J70, 49L25

DOI. 10.1137/S003614290343815X

1. Introduction.

This paper is a continuation of a work started in [2] (see also Jakobsen [21]), whose aim is to prove results on the rate of convergence of monotone approximation schemes for possibly degenerate Hamilton–Jacobi–Bellman (HJB) equations by purely analytical methods. Krylov [26, 27] obtained such results in a rather general framework but by using a combination of PDE arguments and rather deep probabilistic estimates, which we want to avoid.

The strategy we used in [2] is based on the idea that the HJB equation and the approximation scheme should play symmetrical roles. Unfortunately, this leads to unnatural restrictions on the data when the scheme in consideration is a finite difference method. These restrictions do not appear in [27]. In the present paper, we use a more classical strategy, in which the HJB equation plays the central role. Our approach yields results in the full generality, improving those of [26, 27] and [2].

In order to be more specific, we introduce the HJB equation, which is written in the form

$$F(x,u,Du,D^2u)=0 \quad \text{in} \quad \mathbb{R}^N, \tag{1.1}$$

with

$$\begin{aligned} F(x,t,p,X) &= \sup_{\alpha\in\mathcal{A}} \mathcal{L}^\alpha(x,t,p,X),\\ \mathcal{L}^\alpha(x,t,p,X) &= -\mathrm{tr}[a^\alpha(x)X]-b^\alpha(x)p+c^\alpha(x)t-f^\alpha(x), \end{aligned}$$

where tr denotes the trace. The coefficients a, b, c, f are, at least, continuous functions defined on $\mathbb{R}^N\times\mathcal{A}$ with values, respectively, in the space $S(N)$ of symmetric $N\times N$ matrices, $\mathbb{R}^N$ and $\mathbb{R}$. The space of controls, $\mathcal{A}$, is assumed to be a compact metric space. Precise assumptions on the data will be given later on. Under classical

[*] Received by the editors December 3, 2003; accepted for publication (in revised form) October 17, 2004; published electronically June 30, 2005.

http://www.siam.org/journals/sinum/43-2/43815.html

[†] Laboratoire de Mathématiques et Physique Théorique, University of Tours, Parc de Grandmont, 37200 Tours, France (guy.barles@univ-tours.fr).

[‡] Department of Mathematical Sciences, Norwegian University of Science and Technology, 7491 Trondheim, Norway (erj@math.ntnu.no). The research of this author was supported by the Research Council of Norway, grant 151608/432.

assumptions, it is well known that this equation is associated to a stochastic optimal control problem, and that the value function of this problem is the unique viscosity solution of the equation. Moreover, the value function is typically bounded and Hölder continuous, and the regularity depends on the properties of a, b, c, and f.

The monotone approximation schemes we consider are of the type

$$S(h, x, u_h(x), [u_h]_x) = 0 \quad \text{in} \quad \mathbb{R}^N, \tag{1.2}$$

where S is, loosely speaking, a consistent, monotone, and uniformly continuous approximation of F in (1.1). The approximate solution is u_h, $[u_h]_x$ is a function defined from u_h, and the approximation parameter is h. This abstract notation was introduced by Barles and Souganidis [3] to display clearly the monotonicity of the scheme: S is nondecreasing in u_h and nonincreasing in $[u_h]_x$ with the classical ordering for functions. Typical approximation schemes which we have in mind are finite difference methods (FDMs) and control schemes based on the dynamic programming principle. We refer to Dupuis and Kushner [11] and Camilli and Falcone [5] for more information about such schemes.

In the viscosity solutions setting the first results on convergence rates for monotone schemes were obtained by Crandall and Lions [10]. Later, the first-order case was studied by many authors considering different schemes and assumptions [7, 36, 37, 13, 1, 25, 29, 35, 28, 24]. Only recently did Krylov [26, 27] obtain the first results for second-order equations (for HJB equations), and these results were then partially extended by Barles and Jakobsen [2] and Jakobsen [21]. These results concern only HJB equations, or, equivalently, equations with convex/concave Lipschitz continuous nonlinearity F. In the nonconvex (or nonconcave) case, to the best of our knowledge, there are no general results. There exist results only in particular cases like, for example, in one space dimension [20] and for obstacle problems [19].

From a technical point of view, the upper estimate on $u - u_h$ is much easier to obtain than the lower estimate. Roughly speaking, a regularization of the solution u by convolution provides approximate smooth subsolutions of the equation because of the convexity of the equation. By inserting this smooth subsolution in the scheme and using consistency, one is led to the upper bound after choosing an optimal parameter of regularization. It is worth pointing out that a nontrivial difficulty in performing this argument is the x-dependence in the equation. This difficulty was solved by a very clever argument of Krylov [27] which is used extensively in [2] and in the present paper.

Unfortunately, this is clearly a one-sided argument working only for convex equations. In general, there is no simple way to build approximate smooth supersolutions which would lead to the lower estimate on $u - u_h$. It is precisely this difficulty that we overcome here. In fact, we do not really build a sequence of approximate smooth supersolutions, but rather a sequence of supersolutions which behave as if they were smooth. The key step here is to introduce switching system approximations of the HJB equation and study their rates of convergence. This admittedly strange idea leads us roughly speaking to consider equations that are linear (convex and concave) and from there to the solution of the above problem. Krylov [27] uses piecewise constant control approximations, which is more or less the interpretation of switching systems. Despite this similarity, the connection between his arguments and ours is not so clear. Our approach is inspired by Evans and Friedman [12] (see also [6]), and the rates of convergence are obtained by combining the above-mentioned clever argument of Krylov and an approach suggested by Lions [33]. Even if we do not make a point of proving general results in this direction, this part has an independent interest. It seems to be the first time that the rate of convergence is obtained for such switching system approximations in the case of second-order equations.

In order to give a flavor of our results, for HJB equation satisfying natural assumptions and with bounded Lipschitz continuous solutions, we prove a lower estimate of the form $h^{1/5}$ for a standard finite difference method. The corresponding result in Krylov [27] was $h^{1/27}$.

For control schemes, the results of this paper do not give the best error bounds available. They can be found in [2], where the richer structure of such schemes is fully exploited.

The paper is organized as follows: In section 2 we introduce the switching system and prove the rate of convergence. This result is then used in section 3 for obtaining the rate of convergence of the approximation scheme (1.2). In section 4, we apply the result of section 3 to a typical finite difference method for the HJB equation taken from Dupuis and Kushner [11]. In order to simplify the exposure, the proofs in the paper are presented in a context where all the solutions are Lipschitz continuous. In section 5, we provide without proofs extensions to the case of $C^{0,\delta}(\mathbb{R}^N)$-solutions. We also discuss the fact that our approach is rather close to provide results for the nonconvex (nonconcave) case. Finally, the appendix collects several results for switching systems (well-posedness, regularity, and continuous dependence) which are used throughout the paper.

We conclude this introduction by explaining the notation we will use throughout this paper. By $|\cdot|$ we mean the standard Euclidian norm in any $\mathbb{R}^p$-type space (including the space of $N \times P$ matrices). In particular, if $X \in S(N)$, then $|X|^2 = \mathrm{tr}(XX^T)$, where X^T denotes the transpose of X. Now if w is a bounded function from $\mathbb{R}^N$ into either $\mathbb{R}$, $\mathbb{R}^M$, or the space of $N \times P$ matrices, we set

$$|w|_0 = \sup_{y \in \mathbb{R}^N} |w(y)|.$$

If w is also Lipschitz continuous, we set

$$[w]_1 = \sup_{x \neq y} \frac{|w(x) - w(y)|}{|x-y|} \quad \text{and} \quad |w|_1 = |w|_0 + [w]_1.$$

We denote by $\leq$ the component-by-component ordering in $\mathbb{R}^M$ and the ordering in the sense of positive semidefinite matrices in $S(N)$. For the rest of this paper we let ρ denote the same, fixed, positive smooth function with support in $\{|x| < 1\}$ and mass 1. From this function ρ, we define the sequence of mollifiers $\{\rho_\varepsilon\}_{\varepsilon>0}$ as follows:

$$\rho_\varepsilon(x) = \frac{1}{\varepsilon^N}\rho\left(\frac{x}{\varepsilon}\right) \quad \text{in} \quad \mathbb{R}^N.$$

We also use the following spaces: $C_b(\mathbb{R}^N)$ and $C^{0,\delta}(\mathbb{R}^N)$, $\delta \in (0,1]$, denoting, respectively, the space of bounded continuous functions on $\mathbb{R}^N$ and the space of bounded δ-Hölder continuous functions on $\mathbb{R}^N$.

2. Convergence rate for a switching system.

In this section, we obtain the rate of convergence for certain switching system approximations to the HJB equation (1.1). Such approximations have been studied in [12, 6], and a viscosity solutions theory of switching systems can be found in [38, 18, 17]. We consider the following type of switching systems:

$$F_i(x, v, Dv_i, D^2 v_i) = 0 \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I} := \{1, \ldots, M\}, \tag{2.1}$$

where the solution $v = (v_1, \dots, v_M)$ is in $\mathbb{R}^M$, and for $i \in \mathcal{I}$, $x \in \mathbb{R}^N$, $r = (r_1, \dots, r_M) \in \mathbb{R}^M$, $p \in \mathbb{R}^N$, and $X \in \mathcal{S}^N$, F_i is given by

$$F_i(x, r, p, X) = \max \left\{ \sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha(x, r_i, p, X);\ r_i - \mathcal{M}_i r \right\},$$

where $\mathcal{A}_i \subset \mathcal{A}$, $\mathcal{L}^\alpha$ is defined below (1.1), and for $k > 0$,

$$\mathcal{M}_i r = \min_{j \neq i} \{r_j + k\}.$$

Under suitable assumptions on the data, we have the existence and uniqueness of a solution v of this system. Moreover, it is not difficult to see that, as $k \to 0$, every component of v converges locally uniformly to the solution of the HJB equation

$$\sup_{\alpha \in \overline{\mathcal{A}}} \mathcal{L}^\alpha(x, u, Du, D^2 u) = 0 \quad \text{in} \quad \mathbb{R}^N\ , \tag{2.2}$$

where $\overline{\mathcal{A}} = \cup_i \mathcal{A}_i$.

The objective of this section is to obtain an error bound for this convergence. For the sake of simplicity, we restrict ourselves to the situation where the solutions are Lipschitz continuous. However, it is not difficult to adapt our approach to more general situations, and we give results in this direction in section 5.

We will use the following assumptions:

(A1) For any $\alpha \in \mathcal{A}$, $a^\alpha = \frac{1}{2}\sigma^\alpha \sigma^{\alpha T}$ for some $N \times P$ matrix σ^α. Furthermore, there are constants λ, K independent of α such that

$$c \geq \lambda > 0 \quad \text{and} \quad |\sigma^\alpha|_1 + |b^\alpha|_1 + |c^\alpha|_1 + |f^\alpha|_1 \leq K.$$

(A2) The constant λ in (A1) satisfies $\lambda > \sup_\alpha \left\{ [\sigma^\alpha]_1^2 + [b^\alpha]_1 \right\}$.

As the reader will see below and in the following sections, assumption (A1) ensures the well-posedness of all the equations and systems of equations we consider in this paper. If we assume in addition (A2), all solutions will belong to $C^{0,1}(\mathbb{R}^N)$. We refer to the appendix for a precise justification of these claims. In the present situation, we have the following well-posedness and regularity result.

PROPOSITION 2.1. (i) *Assume* (A1). *If w_1 and w_2 are sub- and supersolutions of* (2.1) *or* (2.2), *then $w_1 \leq w_2$.*

(ii) *Assume* (A1) *and* (A2). *Then there exist unique solutions v and u of* (2.1) *and* (2.2), *respectively, satisfying*

$$|v|_1 + |u|_1 \leq C,$$

where the constant C depends only on K, λ from (A1).

In order to obtain the rate of convergence for the switching approximation, we use the before-mentioned regularization procedure of Krylov [27, 2]. This procedure requires the introduction of the following auxiliary system:

$$F_i^\varepsilon(x, v^\varepsilon, Dv_i^\varepsilon, D^2 v_i^\varepsilon) = 0 \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I}, \tag{2.3}$$

where $v^\varepsilon = (v_1^\varepsilon, \dots, v_M^\varepsilon)$,

$$F_i^\varepsilon(x, r, p, M) = \max \left\{ \sup_{\alpha \in \mathcal{A}_i, |e| \leq \varepsilon} \mathcal{L}^\alpha(x + e, r_i, p, X);\ r_i - \mathcal{M}_i r \right\},$$

and $\mathcal{L}$ and $\mathcal{M}$ are defined below (1.1) and (2.1), respectively. By Theorems A.1 and A.3 in the appendix, we have the following result.

Proposition 2.2. (i) *Assume* (A1). *If w_1 and w_2 are sub- and supersolutions of* (2.3)*, then $w_1 \leq w_2$.*

(ii) *Assume* (A1) *and* (A2). *Then there exists a unique solution v^ε of* (2.3) *satisfying*

$$|v^\varepsilon|_1 + \frac{1}{\varepsilon}|v^\varepsilon - v|_0 \leq C,$$

where v solves (2.1) *and the constant C depends only on K, λ from* (A1).

We are now in position to state and prove the main result of this section.

Theorem 2.3. *Assume* (A1) *and* (A2). *If u and v are the solutions of* (2.2) *and* (2.1)*, respectively, then for k small enough,*

$$0 \leq v_i - u \leq Ck^{1/3} \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I},$$

where C depends only on λ, K from (A1).

Remark 2.1. This seems to be the first time the rate of convergence is obtained for switching system approximations of second-order equations.

Proof of Theorem 2.3. Since $w = (u, \dots, u)$ is a subsolution of (2.1), comparison for (2.1) (Proposition 2.1(i)) yields $u \leq v_i$ for $i \in \mathcal{I}$.

To get the other bound, we use an argument suggested by Lions [33] together with the regularization procedure of Krylov [27]. Consider first system (2.3). It follows that, for every $|e| \leq \varepsilon$,

$$\sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha(x + e, v_i^\varepsilon(x), Dv_i^\varepsilon, D^2 v_i^\varepsilon) \leq 0 \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I}.$$

After a change of variables, we see that for every $|e| \leq \varepsilon$, $v^\varepsilon(x - e)$ is a subsolution of the following system of uncoupled equations:

$$\sup_{\alpha \in \mathcal{A}_i} \mathcal{L}^\alpha(x, w_i, Dw_i, D^2 w_i) = 0 \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I}. \tag{2.4}$$

Define $v_\varepsilon := v^\varepsilon * \rho_\varepsilon$, where $\{\rho_\varepsilon\}_\varepsilon$ is the sequence of mollifiers defined at the end of the introduction. A Riemann-sum approximation shows that $v_\varepsilon(x)$ can be viewed as the limit of convex combinations of $v^\varepsilon(x - e)$'s for $|e| < \varepsilon$. Since the $v^\varepsilon(x - e)$'s are subsolutions of the convex(!) equation (2.4), so are the convex combinations. By the stability result for viscosity subsolutions we can now conclude that v_ε is itself a subsolution of (2.4). We refer to the appendix in [2] for more details.

On the other hand, since v^ε is a continuous subsolution of (2.3), we have

$$v_i^\varepsilon \leq \min_{j \neq i} v_j^\varepsilon + k \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I}.$$

It follows that $\max_i v_i^\varepsilon(x) - \min_i v_i^\varepsilon(x) \leq k$, and hence

$$|v_i^\varepsilon - v_j^\varepsilon|_0 \leq k, \quad i, j \in \mathcal{I}.$$

Then, by the definition and properties of v_ε, we have

$$|D^n v_{\varepsilon i} - D^n v_{\varepsilon j}|_0 \leq C\frac{k}{\varepsilon^n}, \quad n \in \mathbb{N}, \quad i, j \in \mathcal{I},$$

where C depends only on ρ. Furthermore, from these bounds, we see that for $\varepsilon < 1$,

$$\left|\sup_{\alpha\in\mathcal{A}_i} \mathcal{L}^\alpha[v_{\varepsilon j}] - \sup_{\alpha\in\mathcal{A}_i} \mathcal{L}^\alpha[v_{\varepsilon i}]\right| \leq C\frac{k}{\varepsilon^2} \quad \text{in} \quad \mathbb{R}^N, \quad i,j\in\mathcal{I}.$$

Here C depends only on $|\sigma|_0, |b|_0, |c|_0$, and ρ. Since v_ε is a subsolution of (2.4), this means that

$$\sup_{\alpha\in\mathcal{A}} \mathcal{L}^\alpha(x, v_{\varepsilon i}, Dv_{\varepsilon i}, D^2 v_{\varepsilon i}) \leq C\frac{k}{\varepsilon^2} \quad \text{in} \quad \mathbb{R}^N, \quad i\in\mathcal{I}.$$

So by (A1) and the definition of $\mathcal{L}$, we see that $v_{\varepsilon i} - \frac{1}{\lambda}C\frac{k}{\varepsilon^2}$ is a subsolution of (2.2).

Comparison for (2.2) (Proposition 2.1(i)) yields

$$v_{\varepsilon i} - u \leq \frac{1}{\lambda}C\frac{k}{\varepsilon^2} \quad \text{in} \quad \mathbb{R}^N, \quad i\in\mathcal{I}.$$

Hence, by Proposition 2.2(ii) and properties of mollifiers, we have

$$v_i - u \leq v_i - v_{\varepsilon i} + v_{\varepsilon i} - u \leq C\varepsilon + \frac{1}{\lambda}C\frac{k}{\varepsilon^2} \quad \text{in} \quad \mathbb{R}^N, \quad i\in\mathcal{I}.$$

Minimizing with respect to ε yields the result. □

3. Convergence rate for the HJB equation. In this section we derive an error bound for the convergence of the solution of the scheme (1.2) to the solution of the HJB equation (1.1). This result is general and derived using only PDE methods, and it extends and improves earlier results by Krylov [26, 27], Barles and Jakobsen [2], and Jakobsen [21].

We assume that assumptions (A1) and (A2) of section 2 hold. As a special case of Proposition 2.1, we have the following well-posedness and regularity result for (1.1).

PROPOSITION 3.1. (i) *Assume* (A1). *If w_1 and w_2 are sub- and supersolutions of* (1.1)*, then $w_1 \leq w_2$.*

(ii) *Assume* (A1) *and* (A2). *Then there exists a unique solution u of* (1.1) *satisfying*

$$|u|_1 \leq C,$$

where the constant C depends only on K, λ from (A1).

For the scheme (1.2) we assume the following:

(S1) (Monotonicity.) For every $h > 0$, $x \in \mathbb{R}^N$, $r \in \mathbb{R}$, $m \geq 0$, and bounded continuous functions u, v such that $u \leq v$ in $\mathbb{R}^N$, the following holds:

$$S(h, x, r+m, [u+m]_x) \geq \lambda m + S(h, x, r, [v]_x).$$

(S2) (Regularity.) For every $h > 0$ and $\phi \in C_b(\mathbb{R}^N)$, $x \mapsto S(h, x, \phi(x), [\phi]_x)$ is bounded and continuous in $\mathbb{R}^N$ and the function $r \mapsto S(h, x, r, [\phi]_x)$ is uniformly continuous for bounded r, uniformly in $x \in \mathbb{R}^N$.

(S3) (Consistency.) There exist integers $n, k_i \geq 0$, $i = 1, 2, \ldots, n$, and a constant K_c such that for every $h \geq 0$, $x \in \mathbb{R}^N$, and smooth function ϕ,

$$\left|F(x, \phi(x), D\phi(x), D^2\phi(x)) - S(h, x, \phi(x), [\phi]_x)\right| \leq K_c \sum_{k_i\neq 0} |D^i\phi|_0 h^{k_i}.$$

Remark 3.1. Conditions (S1) and (S2) imply a comparison result for bounded continuous solutions of (1.2); see [2].

Before we continue, we mention that the upper bound on the error $u - u_h$ is known from [2]; see also [27, 21]. Let us state the result here.

PROPOSITION 3.2. *Assume* (A1), (A2), (S1)–(S3), *and that* (1.2) *has a unique solution* $u_h \in C_b(\mathbb{R}^N)$. *If* u *is the solution of* (1.1), *then, for sufficiently small* $h > 0$, *we have*

$$u - u_h \leq Ch^\gamma \quad in \quad \mathbb{R}^N,$$

where $\gamma := \min_{k_i \neq 0} \left\{ \frac{k_i}{i} \right\}$ *and* C *depends only on* λ, K, K_c *from* (A1), (S3).

Remark 3.2. Existence of $u_h \in C_b(\mathbb{R}^N)$ must be proved for each particular scheme S. We refer to [26, 27, 2, 21] for examples of such arguments.

As mentioned in the introduction, the proof of this proposition relies on the regularization procedure of Krylov, which was also used in section 2. The idea is to obtain a smooth subsolution of (1.1) which is close to the solution of this equation. This then yields the upper bound after classical computations. This approach, however, does not yield the lower bound unless you require much stronger assumptions on the scheme (1.2); see [2, 21, 26].

To avoid such restrictive assumptions, we use a different technique here. The key point is to obtain approximate "almost smooth" supersolutions by considering the following switching system approximation of (1.1):

$$F_i^\varepsilon(x, v^\varepsilon, Dv_i^\varepsilon, D^2 v_i^\varepsilon) = 0 \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I} := \{1, \ldots, M\}, \tag{3.1}$$

where $v^\varepsilon = (v_1^\varepsilon, \ldots, v_M^\varepsilon)$,

$$F_i^\varepsilon(x, r, p, X) = \max \left\{ \min_{|e| \leq \varepsilon} \mathcal{L}^{\alpha_i}(x + e, r_i, p, X);\ r_i - \mathcal{M}_i r \right\},$$

and $\mathcal{L}$ and $\mathcal{M}$ are defined below (1.1) and (2.1), respectively. The solution of this system is expected to be close to the solution of (1.1) if k and ε are small and $\{\alpha_i\}_{i \in \mathcal{I}} \subset \mathcal{A}$ is a sufficiently refined grid for $\mathcal{A}$. In fact, for this to be true we need to assume that the coefficients $\sigma^\alpha, b^\alpha, c^\alpha, f^\alpha$ can be approximated uniformly in x by $\sigma^{\alpha_i}, b^{\alpha_i}, c^{\alpha_i}, f^{\alpha_i}$. The precise assumption is as follows:

(A3) For every $\delta > 0$, there are $M \in \mathbb{N}$ and $\{\alpha_i\}_{i=1}^M \subset \mathcal{A}$, such that for any $\alpha \in \mathcal{A}$,

$$\inf_{1 \leq i \leq M} (|\sigma^\alpha - \sigma^{\alpha_i}|_0 + |b^\alpha - b^{\alpha_i}|_0 + |c^\alpha - c^{\alpha_i}|_0 + |f^\alpha - f^{\alpha_i}|_0) < \delta.$$

Remark 3.3. The typical cases where (A3) is satisfied are (i) when $\mathcal{A}$ is a finite set and (ii) when all coefficients are uniformly continuous in α, uniformly in x.

For (3.1), we have the following result.

LEMMA 3.3. *Assume* (A1) *and* (A2).

(a) *There exists a unique solution* v^ε *of* (3.1) *satisfying* $|v^\varepsilon|_1 \leq C$, *where* C *depends only on* λ, K *from* (A1).

(b) *Assume in addition* (A3), *and let* u *denote the solution of* (1.1). *Then for any* $\delta > 0$ *there are* $M \in \mathbb{N}$ *and* $\{\alpha_i\}_{i=1}^M \subset \mathcal{A}$ *such that the solution* v_ε *of* (3.1) *satisfy*

$$\max_i |u - v_i^\varepsilon|_0 \leq C(\varepsilon + k^{1/3} + \delta),$$

where C *depends only on* λ, K *from* (A1).

The (almost) smooth supersolutions of (1.1) we are looking for are built out of the v_i^ε's by mollification. Before giving the next lemma, we remind the reader that the sequence of mollifiers $\{\rho_\varepsilon\}_\varepsilon$ is defined at the end of the introduction.

LEMMA 3.4. *Assume* (A1), (A2), *and define* $v_{\varepsilon i} := \rho_\varepsilon * v_i^\varepsilon$ *for* $i \in \mathcal{I}$.

(a) *There is a constant C depending only on λ, K from* (A1), *such that*

$$|v_{\varepsilon j} - v_i^\varepsilon|_0 \le C(k+\varepsilon) \quad \textit{for} \quad i, j \in \mathcal{I}.$$

(b) *Assume in addition that $\varepsilon \le (4\sup_i [v_i^\varepsilon]_1)^{-1}k$. For every $x \in \mathbb{R}^N$, if $j := \mathrm{argmin}_{i\in\mathcal{I}} v_{\varepsilon i}(x)$, then*

$$\mathcal{L}^{\alpha_j}(x, v_{\varepsilon j}(x), Dv_{\varepsilon j}(x), D^2 v_{\varepsilon j}(x)) \ge 0\ .$$

Lemma 3.4(b) implies that $w := \min_{i\in\mathcal{I}} v_{\varepsilon i}$ is a viscosity supersolution of (1.1) in all of $\mathbb{R}^N$ (at least this follows from the proof). This function is an “almost smooth” supersolution in the sense that, at any point x, it is only the smooth function $v_{\varepsilon j}$ of Lemma 3.4(b) (which is a supersolution at this point) which is really playing a role. This can be seen from the proof of the rate of convergence below.

We will prove these two lemmas after having stated and proved the main result of this paper—the result giving the lower bound on the error $u - u_h$ for the scheme (1.2).

THEOREM 3.5. *Assume* (A1)–(A3), (S1), (S3) *and that* (1.2) *has a unique solution $u_h \in C_b(\mathbb{R}^N)$. If u is the solution of* (1.1), *then, for sufficiently small $h > 0$, we have*

$$-Ch^{\bar\gamma} \le u - u_h \quad \textit{in} \quad \mathbb{R}^N,$$

where $\bar\gamma := \min_{k_i \ne 0}\{\frac{k_i}{3i-2}\}$ and C depends only on λ, K, K_c from (A1), (S3).

Proof. We fix a $\delta > 0$ and pick the corresponding $\{\alpha_i\}_\mathcal{I}$ according to (A3). Then we consider the solution v^ε of (3.1) corresponding to this choice of $\{\alpha_i\}_\mathcal{I}$. Lemma 3.3 yields existence and properties of v^ε. Furthermore, we mollify this function to obtain v_ε as in Lemma 3.4.

We proceed to obtain an estimate for

$$m := \sup_{y\in\mathbb{R}^N}\{u_h(y) - w(y)\},$$

where $w := \min_{i\in\mathcal{I}} v_{\varepsilon i}$. In order to have a “max” instead of a “sup,” we approximate m by

$$m_\kappa := \sup_{y\in\mathbb{R}^N}\{u_h(y) - w(y) - \kappa\phi(y)\}, \tag{3.2}$$

where $\kappa > 0$ is a small constant and $\phi(y) = (1+|y|^2)^{1/2}$. Since u_h and w are continuous, it is clear that the supremum (3.2) is attained at some point $x \in \mathbb{R}^N$. Because of the definition of w, it is easy to see that x is also a maximum point of

$$\sup_{y\in\mathbb{R}^N}\{u_h(y) - v_{\varepsilon i}(y) - \kappa\phi(y)\} \tag{3.3}$$

when $i = \mathrm{argmin}_{j\in\mathcal{I}} v_{\varepsilon j}(x)$. Notice that this supremum is still m_κ.

Now take $\varepsilon = (4\sup_i [v_i^\varepsilon]_1)^{-1}k$. From Lemma 3.4(b), the properties of ϕ, and (A1), we see that

$$\sup_{\alpha\in\mathcal{A}} \mathcal{L}^\alpha(x, (v_{\varepsilon i} + \kappa\phi)(x), D(v_{\varepsilon i} + \kappa\phi)(x), D^2(v_{\varepsilon i} + \kappa\phi)(x)) \ge -C\kappa, \tag{3.4}$$

where C depends only on K from (A1) ($C = \sup_{\alpha,x}\{|\sigma^\alpha|_0^2 + |b^\alpha|_0\}$).

Let us estimate m_κ. By (3.4) and (S3) we have

$$-C\kappa \leq S(h, x, (v_{\varepsilon i} + \kappa\phi)(x), [v_{\varepsilon i} + \kappa\phi]_x) + K_c \sum_{k_i \neq 0} |D^i(v_{\varepsilon i} + \kappa\phi)|_0 h^{k_i}.$$

By the definitions of $v_{\varepsilon i}$ and ϕ, we can conclude that

$$-C \sum_{k_i \neq 0} \varepsilon^{1-i} h^{k_i} + \mathcal{O}(\kappa) \leq S(h, x, (v_{\varepsilon i} + \kappa\phi)(x), [v_{\varepsilon i} + \kappa\phi]_x), \tag{3.5}$$

where C depends only on the mollifier ρ and λ, K from (A1). On the other hand, using (S1), (3.3), and the definition of m_κ, we see that

$$\begin{aligned} S(h, x, (v_{\varepsilon i} + \kappa\phi)(x), [v_{\varepsilon i} + \kappa\phi]_x) &\leq S(h, x, u_h(x) - m_\kappa, [u_h - m_\kappa]_x) \\ &\leq -\lambda m_\kappa + S(h, x, u_h(x), [u_h]_x) = -\lambda m_\kappa, \end{aligned}$$

where the last equality follows since u_h is the solution of (1.2). From this inequality and (3.5), we have

$$\lambda m_\kappa \leq C \sum_{k_i \neq 0} \varepsilon^{1-i} h^{k_i} + \mathcal{O}(\kappa).$$

From this estimate, we obtain the estimate for m by sending $\kappa \to 0$ and noting that $m_\kappa \to m$.

Using the estimate for m, we now derive the lower bound on the error. Fix an arbitrary $y \in \mathbb{R}^N$. From the definition of m, we see that

$$\begin{aligned} u_h(y) - u(y) &\leq u_h(y) - v_{\varepsilon i}(y) + v_{\varepsilon i}(y) - u(y) \\ &\leq m + v_{\varepsilon i}(y) - u(y). \end{aligned}$$

Using the bound on m, and Lemmas 3.4(a) and 3.3(b), we obtain

$$u_h(y) - u(y) \leq C \left(\sum_{k_i \neq 0} \varepsilon^{1-i} h^{k_i} + \varepsilon + k + k^{1/3} + \delta \right).$$

The constant C does not depend on y, and therefore the right-hand side is a uniform in y upper bound for $u_h - u$.

The conclusion follows by choosing

$$\varepsilon = \max_{k_i \neq 0} h^{\frac{3k_i}{3i-2}} \quad \text{and} \quad k = 4 \sup_i [v_i^\varepsilon]_1 \varepsilon$$

and sending $\delta \to 0$ (since all constants are independent of the size of $\mathcal{I}$). □

Now we give the proofs of Lemmas 3.3 and 3.4.

Proof of Lemma 3.3.

1. First we approximate (1.1) by the following equation:

$$\sup_{i \in \mathcal{I}} \mathcal{L}^{\alpha_i}(x, v, Dv, D^2 v) = 0 \quad \text{in} \quad \mathbb{R}^N.$$

From assumption (A3) and Theorems A.1 and A.3 in the appendix, we have the following result: There exist a unique solution v of the above equation satisfying

$$|v - u|_0 \leq C\delta,$$

where C depends only on λ, K from (A1).

2. We continue by approximating the above equation by the following switching system:

$$\max\Big\{\mathcal{L}^{\alpha_i}(x, v_i, Dv_i, D^2 v_i);\ v_i - \mathcal{M}_i v\Big\} = 0 \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I},$$

where $\mathcal{M}$ is defined below (2.1). From Proposition 2.1 and Theorem 2.3, we have existence and uniqueness of a solution $\bar{v}$ of the above system satisfying

$$|\bar{v}_i - v|_0 \leq Ck^{1/3}, \quad i \in \mathcal{I},$$

where C depends only on the mollifier ρ and λ, K from (A1).

3. The switching system defined in the previous step is nothing but (3.1) with $\varepsilon = 0$ or (2.3) with the $\mathcal{A}_i$'s being singletons. Proposition 2.2 yields the existence and uniqueness of a solution v^ε of (3.1) satisfying

$$|v^\varepsilon|_1 + \frac{1}{\varepsilon}|v^\varepsilon - \bar{v}|_0 \leq C,$$

where C depends only on λ, K from (A1).

4. The proof is complete by combining the estimates in steps 1–3 and noting that (A3) is only needed in step 1. □

Proof of Lemma 3.4. We start with (a). From the properties of mollifiers and the Lipschitz continuity of v^ε, it is immediate that

$$|v_{\varepsilon i} - v_i^\varepsilon|_0 \leq C\varepsilon, \quad i \in \mathcal{I}, \tag{3.6}$$

where $C = \max_i [v_i^\varepsilon]_1$ depends only on K, λ from (A1). Furthermore we saw in the proof of Theorem 2.3 that

$$0 \leq \max_i v_i^\varepsilon - \min_i v_i^\varepsilon \leq k \quad \text{in} \quad \mathbb{R}^N.$$

From these two estimates, (a) follows.

Now consider (b). We consider an arbitrary point $x \in \mathbb{R}^N$ and set

$$j = \text{argmin}_{i \in \mathcal{I}} v_{\varepsilon i}(x).$$

Then, by definition of $\mathcal{M}$ and j, we have

$$v_{\varepsilon j}(x) - \mathcal{M}_j v_\varepsilon(x) = \max_{i \neq j}\big\{v_{\varepsilon j}(x) - v_{\varepsilon i}(x) - k\big\} \leq -k.$$

The bound (3.6) then leads to

$$v_j^\varepsilon(x) - \mathcal{M}_j v^\varepsilon(x) \leq -k + 2\max_i [v_i^\varepsilon]_1 \varepsilon,$$

and by using the Lipschitz continuity of v^ε (Lemma 3.3),

$$v_j^\varepsilon(y) - \mathcal{M}_j v^\varepsilon(y) \leq -k + 2\max_i [v_i^\varepsilon]_1(\varepsilon + |x - y|).$$

From this we conclude that if $|x - y| < \varepsilon$ and $\varepsilon \leq (4\max_i [v_i^\varepsilon]_1)^{-1}k$, then

$$v_j^\varepsilon(y) - \mathcal{M}_j v^\varepsilon(y) < 0.$$

Equation (3.1) then implies

$$\inf_{|e|\le\varepsilon} \mathcal{L}^{\alpha_j}(y+e, v_j^\varepsilon(y), Dv_j^\varepsilon(y), D^2 v_j^\varepsilon(y)) = 0.$$

After a change of variables we see that for every $|e| \le \varepsilon$,

$$\mathcal{L}^{\alpha_j}(x, v_j^\varepsilon(x-e), Dv_j^\varepsilon(x-e), D^2 v_j^\varepsilon(x-e)) \ge 0. \tag{3.7}$$

In other words, for every $|e| \le \varepsilon$, $v_j^\varepsilon(x-e)$ is a supersolution at x of

$$\mathcal{L}^{\alpha_j}(x, w, Dw, D^2 w) = 0. \tag{3.8}$$

By mollifying (3.7) we see formally that $v_{\varepsilon j}$ is also a supersolution of (3.8) at x and hence a (viscosity) supersolution of the HJB equation (1.1) at x. This is correct since $v_{\varepsilon j}$ can be viewed as the limit of convex combinations of supersolutions $v_j^\varepsilon(x-e)$ of the linear and hence concave equation (3.8); we refer to the proof of Theorem 2.3 and to the appendix in [2] for the details. We conclude the proof by noting that since $v_{\varepsilon j}$ is smooth, it is in fact a classical supersolution of (1.1) at x. □

4. Monotone finite difference methods. As an application of the results in the previous section we derive here the rate of convergence for a finite difference scheme proposed by Kushner [11, 14] for the N-dimensional HJB equation (1.1). The notation for these schemes is taken from [11, 14]. We start by naming the difference operators we need. Let $\{e_i\}_{i=1}^N$ denote the standard basis in $\mathbb{R}^N$ and define

$$\begin{aligned}
\Delta_{x_i}^{\pm} w(x) &= \pm\frac{1}{h}\{w(x \pm e_i h) - w(x)\},\\
\Delta_{x_i}^{2} w(x) &= \frac{1}{h^2}\{w(x + e_i h) - 2w(x) + w(x - e_i h)\},\\
\Delta_{x_i x_j}^{+} w(x) &= \frac{1}{2h^2}\{2w(x) + w(x + e_i h + e_j h) + w(x - e_i h - e_j h)\}\\
&\quad - \frac{1}{2h^2}\{w(x + e_i h) + w(x - e_i h) + w(x + e_j h) + w(x - e_j h)\},\\
\Delta_{x_i x_j}^{-} w(x) &= \frac{1}{2h^2}\{w(x + e_i h) + w(x - e_i h) + w(x + e_j h) + w(x - e_j h)\}\\
&\quad - \frac{1}{2h^2}\{2w(x) + w(x + e_i h - e_j h) + w(x - e_i h + e_j h)\}.
\end{aligned}$$

Now we define the schemes as follows:

$$\tilde{F}(x, u_h(x), \Delta_{x_i}^{\pm} u_h(x), \Delta_{x_i}^{2} u_h(x), \Delta_{x_i x_j}^{\pm} u_h(x)) = 0, \tag{4.1}$$

where

$$\begin{aligned}
\tilde{F}(x, t, p_i^{\pm}, A_{ii}, A_{ij}^{\pm}) = \sup_{\alpha\in\mathcal{A}} \Bigg\{ \sum_{i=1}^{N} \Bigg[-\frac{a_{ii}^{\alpha}}{2} A_{ii} + \sum_{j\neq i} \Bigg(-\frac{a_{ij}^{\alpha+}}{2} A_{ij}^{+} + \frac{a_{ij}^{\alpha-}}{2} A_{ij}^{-} \Bigg)\\
- b_i^{\alpha+}(x) p_i^{+} + b_i^{\alpha-}(x) p_i^{-} \Bigg] + c^{\alpha}(x) t - f^{\alpha}(x) \Bigg\},
\end{aligned}$$

and $b^+ = \max\{b, 0\}$ and $b^- = (-b)^+$ $(b = b^+ - b^-)$.

Assume that (A1) holds. In order to obtain the required monotonicity of these schemes, we need to assume in addition that the matrix a is diagonally dominant:

$$a_{ii}^{\alpha}(x) - \sum_{j\neq i} |a_{ij}^{\alpha}(x)| \geq 0 \quad \text{in} \quad \mathbb{R}^N, \quad i = 1, \dots, N. \tag{4.2}$$

We also assume that the coefficients are normalized so that

$$\sum_{i=1}^{N} \left\{ a_{ii}^{\alpha}(x) - \sum_{j\neq i} |a_{ij}^{\alpha}(x)| + |b_i^{\alpha}(x)| \right\} \leq 1 \quad \text{in} \quad \mathbb{R}^N. \tag{4.3}$$

Assumption (4.2) is standard in numerical analysis; see [11, 14]. We also refer to Lions and Mercier [34] and to Bonnans and Zidani [4] for a discussion of this condition. Assumption (4.3) is always satisfied after a multiplication in (1.1) by an appropriate positive constant.

From the results in section 3, we have the following bound on $u - u_h$.

THEOREM 4.1. *Assume* (A1)–(A3), (4.2), *and* (4.3) *hold. If u and $u_h \in C_b(\mathbb{R}^N)$ are solutions of* (1.1) *and* (4.1), *respectively, then for $h > 0$ sufficiently small,*

$$|u - u_h|_0 \leq C h^{1/5}.$$

Remark 4.1. Krylov [27] obtains the rate 1/27 using probabilistic methods. One contribution of this paper is to improve this rate to 1/5.

By Proposition 3.2 and Theorem 3.5 the above result holds if we can define S in (1.2), check that assumptions (S1)–(S3) hold with $k_2 = 1$, $k_4 = 2$, and $k_i = 0$ otherwise, and prove existence of $u_h \in C_b(\mathbb{R}^N)$. Let us proceed to write down S. In order to better see the monotonicity of the scheme and to fix some more notation, we are going to rewrite (4.1) as a discrete dynamical programming principle. We refer to [11] for the probabilistic interpretation. Define the following one-step transition probabilities:

$$\begin{aligned}
p^{\alpha}(x,x) &= 1 - \sum_{i=1}^{N} \left\{ a_{ii}^{\alpha}(x) - \sum_{j\neq i} |a_{ij}^{\alpha}(x)| + h|b_i^{\alpha}(x)| \right\}, \\
p^{\alpha}(x, x \pm e_i h) &= \frac{a_{ii}^{\alpha}(x)}{2} - \sum_{j\neq i} \frac{|a_{ij}^{\alpha}(x)|}{2} + h b_i^{\alpha\pm}(x), \\
p^{\alpha}(x, x + e_i h \pm e_j h) &= \frac{a_{ij}^{\alpha\pm}(x)}{2}, \\
p^{\alpha}(x, x - e_i h \pm e_j h) &= \frac{a_{ij}^{\alpha\mp}(x)}{2},
\end{aligned}$$

and $p^{\alpha}(x,y) = 0$ for all other y. Note that by (4.2) and (4.3), $0 \leq p^{\alpha}(x,y) \leq 1$ for all α, x, y if $h \leq 1$. Furthermore $\sum_{z\in h\mathbb{Z}^N} p^{\alpha}(x, x+z) = 1$ for all α, x. Tedious but straightforward computations show that the following equation is equivalent to (4.1):

$$u_h(x) = \inf_{\alpha\in\mathcal{A}} \left\{ \frac{1}{1 + h^2 c^{\alpha}(x)} \left(\sum_{z\in h\mathbb{Z}^N} p^{\alpha}(x, x+z) u_h(x+z) + h^2 f^{\alpha}(x) \right) \right\}.$$

This is the discrete dynamical programming principle. From this equation we define S. For $\phi \in C_b(\mathbb{R}^N)$, set $[\phi]_x^h(\cdot) := \phi(x+\cdot)$ and

$$S(h,y,t,[\phi]_x^h) := \sup_{\alpha\in\mathcal{A}} \left\{ -\frac{1}{h^2}\left[\sum_{z\in h\mathbb{Z}^N} p^\alpha(y,y+z)[\phi]_x^h(z) - t\right] + c^\alpha(x)t - f^\alpha(y) \right\}.$$

Using this definition of S, it is easy to check (S1)–(S3); see the lemma below (see also [2]). Existence of solutions $u_h \in C_b(\mathbb{R}^N)$ of (4.1) can be proved using the contraction mapping theorem; we refer to [26, 27, 2, 21] for such arguments. Thus, we may conclude that Theorem 4.1 holds.

Lemma 4.2. *Assume* (A1), (A2), (4.2), (4.3), *and* $0 < h < 1$. *Then the scheme* (4.1) *satisfies conditions* (S1)–(S3), *where* (S3) *takes the form*

$$|F(x,v,Dv,D^2v) - S(h,x,v(x),[v]_x)| \le \sup_\alpha |b^\alpha|_0 |D^2v|_0 h + \sup_\alpha |\sigma^\alpha|_0^2 |D^4v|_0 h^2.$$

5. Extensions and remarks. Let us first consider the case when (A2) is not satisfied. Then the solutions of the different equations are only Hölder continuous; e.g., for the HJB equation (1.1) we have the following result.

Lemma 5.1. *Assume* (A1) *and define* $\lambda_0 := \sup_{\mathcal{A}}\{[\sigma]_1^2 + [b]_1\}$. *If* $\lambda < \lambda_0$, *then there exists a unique solution* $u \in C^{0,\delta}(\mathbb{R}^N)$ *of* (1.1), *where* $\delta = \lambda/\lambda_0$.

This result was proved in [30]. We claim that under (A1), we have the same regularity (the same δ) for all equations considered in this paper. We skip the tedious proof of this claim. In the rest of this section, the solutions of the different equations are assumed to belong to $C^{0,\delta}(\mathbb{R}^N)$ with the same fixed $\delta \in (0,1]$.

Lower than Lipschitz regularity of solutions implies lower convergence rates than obtained in sections 2–4. We will now state the Hölder versions of these results without proofs. The proofs are not much different from the proofs given above, and, moreover, the Hölder case was extensively studied in [2]. We start with the convergence rate for the switching system approximation of section 2.

Proposition 5.2. *Assume* (A1). *If* $\bar{u}$ *and* v *are the solutions of* (2.2) *and* (2.1) *belonging to* $C^{0,\delta}(\mathbb{R}^N)$, *then for* k *small enough,*

$$0 \le v_i - \bar{u} \le Ck^{\frac{\delta}{2+\delta}} \quad in \quad \mathbb{R}^N, \quad i \in \mathcal{I},$$

where C *depends only on* λ, K *from* (A1).

The upper bound on the error for monotone approximation schemes (1.2) for the HJB equation (1.1) is given by the following result.

Proposition 5.3. *Assume* (A1), (S1)–(S3) *and that* (1.2) *has a unique solution* $u_h \in C_b(\mathbb{R}^N)$. *If* $u \in C^{0,\delta}(\mathbb{R}^N)$ *is the solution of* (1.1), *then for sufficiently small* $h > 0$, *we have*

$$u - u_h \le Ch^{\delta\gamma} \quad in \quad \mathbb{R}^N,$$

where γ *and* C *are defined in Proposition* 3.2.

This proposition was essentially proved in [2]; see [21] for this form of the result. Finally, we have come to the Hölder version of the main result of this paper.

Proposition 5.4. *Assume* (A1), (A3), (S1), (S3) *and that* (1.2) *has a unique solution* $u_h \in C_b(\mathbb{R}^N)$. *If* $u \in C^{0,\delta}(\mathbb{R}^N)$ *is the solution of* (1.1), *then for sufficiently small* $h > 0$, *we have*

$$-Ch^{\bar{\gamma}} \le u - u_h \quad in \quad \mathbb{R}^N,$$

where $\bar{\gamma} := \min_{k_i \neq 0}\{\frac{\delta^2 k_i}{(2+\delta)i - 2\delta}\}$ *and* C *depends only on* λ, K, K_c *from* (A1), (S3).

Remark 5.1. Above we removed assumption (A2). It is also possible to weaken assumption (A1) by assuming that c, f are only Hölder continuous. This would then lead to Hölder continuous solutions with lower Hölder exponents than above. The above results would continue to hold, however, but now with a different δ. We refer to [2] for results in this direction.

Next, we comment on a possible extension to the nonconvex/nonconcave case. We are interested in the Isaacs equations coming from stochastic differential games,

$$F(x, u, Du, D^2u) = 0 \quad \text{in} \quad \mathbb{R}^N, \tag{5.1}$$

where

$$\begin{aligned} F(x, t, p, X) &= \sup_{\alpha \in \mathcal{A}} \inf_{\beta \in \mathcal{B}} \mathcal{L}^{\alpha,\beta}(x, t, p, X), \\ \mathcal{L}^{\alpha,\beta}(x, t, p, X) &= -\text{tr}[a^{\alpha,\beta}(x)X] - b^{\alpha,\beta}(x)p + c^{\alpha,\beta}(x)t - f^{\alpha,\beta}(x), \end{aligned}$$

and $\mathcal{A}, \mathcal{B}$ are compact metric spaces. Assume that assumptions like (A1)–(A3) are satisfied for this problem. In this case we have well-posedness and Lipschitz regularity results for (5.1) (see the appendix).

Let $\{\alpha_i\}_{i=1}^M \subset \mathcal{A}$ be a suitable refined grid for $\mathcal{A}$, and consider the question of finding the rate of convergence for the following switching system approximation of (5.1):

$$F_i(x, v, Dv_i, D^2v_i) = 0 \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I} := \{1, \dots, M\}, \tag{5.2}$$

where $v = (v_1, \dots, v_M)$,

$$F_i(x, r, p, M) = \max\Big\{ \inf_{\beta \in \mathcal{B}} \mathcal{L}^{\alpha_i,\beta}(x, r_i, p, X);\ r_i - \mathcal{M}_i r \Big\},$$

and $\mathcal{M}$ is defined just below (2.1) in section 2. To the best of our knowledge, this question is still an open problem, and clearly the method used in section 2 cannot be extended to this case.

However, if we assume that this question has been resolved, then the proof of Theorem 3.5 can be extended to give a lower bound for the error of approximation schemes for (5.1). The only problem we face here is to extend the proof of Lemma 3.4(b). But this is trivial because of the concavity of the function $\inf_{\beta \in \mathcal{B}} \mathcal{L}^{\alpha_i,\beta}(x, t, p, X)$.

To get the upper bound on the error, we only need to assume that the Isaacs condition is satisfied, i.e.,

$$\sup_{\alpha \in \mathcal{A}} \inf_{\beta \in \mathcal{B}} \mathcal{L}^{\alpha,\beta}(x, t, p, X) = \inf_{\beta \in \mathcal{B}} \sup_{\alpha \in \mathcal{A}} \mathcal{L}^{\alpha,\beta}(x, t, p, X)$$

for any $x \in \mathbb{R}^N$, $t \in \mathbb{R}$, $p \in \mathbb{R}^N$, and $X \in \mathcal{S}^N$. The upper bound can then be obtained by a symmetric argument, changing "sup" to "inf," "max" to "min," and conversely.

Thus, the rate of convergence of approximation schemes for Isaacs equations would follow from our method if the rate of convergence of the corresponding switching system can be obtained.

Appendix. Well-posedness, regularity, and continuous dependence for switching systems. In this section we give well-posedness, regularity, and continuous dependence results for solutions of a very general switching system that has as special cases the scalar HJB and Isaacs equations (1.1) and (5.1), and the switching systems (2.1), (2.3), (3.1), (5.2).

We consider the following system:

$$F_i(x, u, Du_i, D^2u_i) = 0 \quad \text{in} \quad \mathbb{R}^N, \quad i \in \mathcal{I} := \{1, \dots, M\}, \tag{A.1}$$

with

$$F_i(x, r, p, X) = \max\Big\{ \sup_{\alpha\in\mathcal{A}} \inf_{\beta\in\mathcal{B}} \mathcal{L}_i^{\alpha,\beta}(x, r_i, p, X);\ r_i - \mathcal{M}_i r \Big\},$$
$$\mathcal{L}_i^{\alpha,\beta}(x, t, p, X) = -\mathrm{tr}[a_i^{\alpha,\beta}(x)X] - b_i^{\alpha,\beta}(x)p + c_i^{\alpha,\beta}(x)t - f_i^{\alpha,\beta}(x),$$

where $\mathcal{M}$ is defined below (2.1), $\mathcal{A}, \mathcal{B}$ are compact metric spaces, r is a vector $r = (r_1, \dots, r_M)$, and $k > 0$ is a constant (the switching cost). See [12, 6, 38, 18, 17] for more information about such systems.

We make the following assumptions:

(A1) For any α, β, i, $a_i^{\alpha,\beta} = \frac{1}{2}\sigma_i^{\alpha,\beta}\sigma_i^{\alpha,\beta T}$ for some $N \times P$ matrix $\sigma_i^{\alpha,\beta}$. Furthermore, there are constants λ, C independent of i, α, β, such that

$$c \geq \lambda > 0 \quad \text{and} \quad [\sigma_i^{\alpha,\beta}]_1 + [b_i^{\alpha,\beta}]_1 + [c_i^{\alpha,\beta}]_1 + |f_i^{\alpha,\beta}|_1 \leq C.$$

(A2) The constant λ in (A1) satisfies $\lambda > \sup_{i,\alpha,\beta}\{[\sigma_i^{\alpha,\beta}]_1^2 + [b_i^{\alpha,\beta}]_1\}$.

We start with comparison, existence, uniqueness, and L^∞ bounds on the solution and its gradient. Before stating the results, we first define $USC(\mathbb{R}^N; \mathbb{R}^M)$ and $LSC(\mathbb{R}^N; \mathbb{R}^M)$ to be the spaces of upper and lower semicontinuous functions from $\mathbb{R}^N$ into $\mathbb{R}^M$, respectively.

THEOREM A.1. *Assume* (A1) *holds.*

(i) *If $u \in USC(\mathbb{R}^N; \mathbb{R}^M)$ is a subsolution of* (A.1) *bounded above and $v \in LSC(\mathbb{R}^N; \mathbb{R}^M)$ is a supersolution of* (A.1) *bounded below, then $u \leq v$ in $\mathbb{R}^N$.*

(ii) *There exists a unique bounded continuous solution u of* (A.1) *satisfying*

$$\max_i |u_i|_0 \leq \sup_{i,\alpha,\beta} \frac{|f_i^{\alpha,\beta}|_0}{\lambda}.$$

(iii) *If in addition* (A2) *holds, then u is Lipschitz continuous and*

$$\max_i [u_i]_1 \leq \sup_{i,\alpha,\beta} \frac{|u^i|_0[c_i^{\alpha,\beta}]_1 + [f_i^{\alpha,\beta}]_1}{\lambda - [\sigma_i^{\alpha,\beta}]_1^2 - [b_i^{\alpha,\beta}]_1}.$$

Remark A.1. These bounds have the same form as those for linear equations [15] and HJB equations [30].

Before giving the proof, we state and prove a key technical lemma.

LEMMA A.2. *Let $u \in USC(\mathbb{R}^N; \mathbb{R}^M)$ be a bounded-above subsolution of* (A.1) *and $\bar{u} \in LSC(\mathbb{R}^N; \mathbb{R}^M)$ be a bounded-below supersolution of another equation* (A.1) *where the functions $\mathcal{L}_i^{\alpha,\beta}$ are replaced by functions $\bar{\mathcal{L}}_i^{\alpha,\beta}$ satisfying the same assumptions. Let $\phi \in C^2(\mathbb{R}^{2N})$ be a function bounded from below. We denote*

$$\psi_i(x, y) = u_i(x) - \bar{u}_i(y) - \phi(x, y)$$

and $M = \sup_{i,x,y} \psi_i(x, y)$. If there exists a maximum point for M, i.e., a point (i', x_0, y_0) such that $\psi_{i'}(x_0, y_0) = M$, then there exists $i_0 \in \mathcal{I}$ such that (i_0, x_0, y_0) is also a maximum point for M, and, in addition, $\bar{u}_{i_0}(y_0) < \mathcal{M}_{i_0}\bar{u}(y_0)$.

Loosely speaking this lemma means that whenever we do doubling of variables for systems of type (A.1), we can ignore the $u_i - \mathcal{M}_i u$ part of the equations. So we are more or less back in the scalar case with equations $\sup_\alpha \inf_\beta \mathcal{L}_{i_0}^{\alpha,\beta}[u^{i_0}] \le 0$ and $\sup_\alpha \inf_\beta \bar{\mathcal{L}}_{i_0}^{\alpha,\beta}[\bar{u}^{i_0}] \ge 0$.

Proof of Lemma A.2. The proof is a "no-loop" argument taken from Ishii and Koike [18]. We assume by contradiction that $\bar{u}_j(y_0) \ge \mathcal{M}_j \bar{u}(y_0)$ for every $j \in A$, where A is the set of j's such that (j, x_0, y_0) is a maximum point for ψ.

We pick a $j \in A$. By the definition of $\mathcal{M}_j$, there is $l \in \mathcal{I}$ such that

$$\mathcal{M}_j \bar{u}(y_0) = \bar{u}_l(y_0) + k.$$

By assumption, we have $\bar{u}_j(y_0) \ge \bar{u}_l(y_0) + k$. On the other hand, since u is a subsolution of (A.1), it follows that

$$u_j(x_0) \le \mathcal{M}_j u(x_0) \le u_l(x_0) + k.$$

Combining these inequalities yields

$$u_j(x_0) - u_l(x_0) \le k \le \bar{u}_j(y_0) - \bar{u}_l(y_0).$$

These inequalities first imply that $l \in A$, and therefore the last inequality is an equality. This, again, implies $\bar{u}_j(y_0) = \bar{u}_l(y_0) + k$.

Since A is finite we may find $j_1, \dots, j_K \in A$ such that $\bar{u}_{j_i}(y_0) = \bar{u}_{j_{i+1}}(y_0) + k$ for $i = 1, \dots, K-1$ and (importantly!) $j_1 = j_K$. But now

$$0 = \sum_{i=1}^{K-1} \big(\bar{u}_{j_i}(y_0) - \bar{u}_{j_{i+1}}(y_0)\big) = (K-1)k > 0,$$

which is a contradiction. The proof is complete. □

Proof of Theorem A.1. Comparison, uniqueness, and existence is proved in [18] for the Dirichlet problem for (1.1) on a bounded domain under assumptions that are satisfied for our problem. The key point here is the comparison principle. To extend this result to an unbounded domain, we only need to modify the test function used in [18] in the standard way. The proof remains practically unchanged.

Let

$$M := \sup_{i,\alpha,\beta} \frac{|f_i^{\alpha,\beta}|_0}{\lambda}.$$

Then the bound on $|u|_0$ follows from the comparison principle after checking that M $(-M)$ is a supersolution (subsolution) of (A.1). To get the bound on the gradient of u, consider

$$m := \sup_{i,x,y \in \mathbb{R}^N} \{u_i(x) - u_i(y) - L|x-y|\}.$$

If, by setting

$$L := \sup_{i,\alpha,\beta} \frac{|u_i|_0 [c_i^{\alpha,\beta}]_1 + [f_i^{\alpha,\beta}]_1}{\lambda - [\sigma_i^{\alpha,\beta}]_1^2 - [b_i^{\alpha,\beta}]_1},$$

we can conclude that $m \le 0$, then we are done. Assume for simplicity that the maximum is attained in $\bar{x}, \bar{y}$. If $\bar{x} = \bar{y}$, then $m = 0$ and we are done. If not, then

$L|x-y|$ is smooth at $\bar{x}, \bar{y}$ and a doubling of variables argument leads to $m \leq 0$. This argument is standard after an application of Lemma A.2 which reduces the problem to a scalar problem (see also the proof of Theorem A.3). We refer to the appendix of [15] for details in the (linear) scalar case. Since the maximum need not be attained, we must modify the test function in the standard way. We skip the details. □

Now we proceed to obtain continuous dependence on the coefficients.

THEOREM A.3. *Let u and $\bar{u}$ be solutions of* (A.1) *with coefficients σ, b, c, f and $\bar{\sigma}, \bar{b}, \bar{c}, \bar{f}$, respectively. If both sets of coefficients satisfy* (A1) *with the same λ, and $|u|_1 + |\bar{u}|_1 \leq M < \infty$, then*

$$\lambda \max_i |u_i - \bar{u}_i|_0 \leq K \sup_{i,\alpha,\beta} |\sigma - \bar{\sigma}|_0 + \sup_{i,\alpha,\beta} \Big\{ 2M|b-\bar{b}|_0 + M|c-\bar{c}|_0 + |f-\bar{f}|_0 \Big\},$$

where

$$K^2 \leq 8M \sup_{i,\alpha,\beta} \Big\{ 2M[\sigma]_1^2 \wedge [\bar{\sigma}]_1^2 + 2M[b]_1 \wedge [\bar{b}]_1 + M[c]_1 \vee [\bar{c}]_1 + [f]_1 \wedge [\bar{f}]_1 \Big\}.$$

Outline of proof. Define

$$m := \sup_{i,x,y} \psi^i(x,y) := \sup_{i,x,y} \left\{ u_i(x) - \bar{u}_i(y) - \frac{1}{\delta}|x-y|^2 - \varepsilon(|x|^2 + |y|^2) \right\}.$$

By the assumptions the supremum is attained at some point (i_0, x_0, y_0). By Lemma A.2, the index i_0 may be chosen so that $\bar{u}_{i_0}(y_0) < \mathcal{M}_{i_0}\bar{u}(y_0)$. With this in mind, the maximum principle for semicontinuous functions [8, 9] and the definition of viscosity solutions imply the following inequality:

$$\sup_\alpha \inf_\beta \mathcal{L}_{i_0}^{\alpha,\beta}(x_0, u_{i_0}, p_x, X) - \sup_\alpha \inf_\beta \bar{\mathcal{L}}_{i_0}^{\alpha,\beta}(y_0, \bar{u}_{i_0}, p_y, Y) \leq 0,$$

where $(p_x, X) \in \overline{D}^{2,+} u_{i_0}(x_0)$ and $(p_y, Y) \in \overline{D}^{2,-} \bar{u}_{i_0}(y_0)$ (see [8, 9] for the notation). Furthermore $p_x = \frac{2}{\delta}(x_0 - y_0) + 2\varepsilon x_0$, $p_y = \frac{2}{\delta}(x_0 - y_0) - 2\varepsilon y_0$, and

$$\begin{pmatrix} X & 0 \\ 0 & Y \end{pmatrix} \leq \frac{2}{\delta} \begin{pmatrix} I & -I \\ -I & I \end{pmatrix} + 2\varepsilon \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} + \mathcal{O}(\kappa)$$

for some $\kappa > 0$. In the end we will fix δ and ε and send $\kappa \to 0$, so we simply ignore the $\mathcal{O}(\kappa)$-term in the following. The first inequality implies

$$0 \leq \sup_{i,\alpha,\beta} \Big\{ -\operatorname{tr}[\bar{a}(y_0)Y] + \operatorname{tr}[a(x_0)X] + \bar{b}(y_0)p_x - b(x_0)p_y + \bar{c}(y_0)\bar{u}(y_0) - c(x)u(x_0) + \bar{f}(y_0) + f(x_0) \Big\}.$$

Note that Lipschitz regularity of the solutions and a standard argument yields

$$|x_0 - y_0| \leq \delta M.$$

So using Ishii's trick on the second-order terms [16, pp. 33, 34], and a few other manipulations, we get

$$0 \leq \sup_{i,\alpha,\beta} \left\{ \frac{2}{\delta}|\sigma(x_0) - \bar{\sigma}(y_0)|^2 + 2M|b(x_0) - \bar{b}(y_0)| + C\varepsilon(1 + |x_0|^2 + |y_0|^2) + M|c(x_0) - \bar{c}(y_0)| - \lambda m + |f(x_0) - \bar{f}(y_0)| \right\}.$$

Some more work leads to an estimate for m depending on δ and ε, and using the definition of m, we obtain an upper bound for $u - \bar{u}$. We finish the proof of the upper bound on $u - \bar{u}$ by minimizing this expression with respect to δ and sending $\varepsilon \to 0$. The lower bound follows in a similar fashion. □

Remark A.2. For more details on such manipulations, we refer to [22, 23].

REFERENCES

[1] R. Abgrall, *Numerical discretization of the first-order Hamilton–Jacobi equation on triangular meshes*, Comm. Pure Appl. Math., 49 (1996), pp. 1339–1373.

[2] G. Barles and E. R. Jakobsen, *On the convergence rate of approximation schemes for Hamilton–Jacobi–Bellman equations*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 33–54.

[3] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptot. Anal., 4 (1991), pp. 271–283.

[4] J. F. Bonnans and H. Zidani, *Consistency of generalized finite difference schemes for the stochastic HJB equation*, SIAM J. Numer. Anal., 41 (2003), pp. 1008–1021.

[5] F. Camilli and M. Falcone, *An approximation scheme for the optimal control of diffusion processes*, RAIRO Modél. Math. Anal. Numér., 29 (1995), pp. 97–122.

[6] I. Capuzzo-Dolcetta and L. C. Evans, *Optimal switching for ordinary differential equations*, SIAM J. Control Optim., 22 (1984), pp. 143–161.

[7] I. Capuzzo-Dolcetta and H. Ishii, *Approximate solutions of the Bellman equation of deterministic control theory*, Appl. Math. Optim., 11 (1984), pp. 161–181.

[8] M. G. Crandall and H. Ishii, *The maximum principle for semicontinuous functions*, Differential Integral Equations, 3 (1990), pp. 1001–1014.

[9] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[10] M. G. Crandall and P.-L. Lions, *Two approximations of solutions of Hamilton–Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[11] P. Dupuis and H. J. Kushner, *Numerical Methods for Stochastic Control Problems in Continuous Time*, Springer-Verlag, New York, 2001.

[12] L. C. Evans and A. Friedman, *Optimal stochastic switching and the Dirichlet problem for the Bellman equation*, Trans. Amer. Math. Soc., 253 (1979), pp. 365–389.

[13] M. Falcone, *A numerical approach to the infinite horizon problem of deterministic control theory*, Appl. Math. Optim., 15 (1987), pp. 1–13.

[14] W. H. Fleming and H. M. Soner, *Controlled Markov Processes and Viscosity Solutions*, Springer-Verlag, New York, 1993.

[15] H. Ishii, *On the equivalence of two notions of weak solutions, viscosity solutions and distribution solutions*, Funkcial. Ekvac., 38 (1995), pp. 101–120.

[16] H. Ishii, *On uniqueness and existence of viscosity solutions of fully nonlinear second-order elliptic PDEs*, Comm. Pure Appl. Math., 42 (1989), pp. 15–45.

[17] H. Ishii and S. Koike, *Viscosity solutions for monotone systems of second-order elliptic PDEs*, Comm. Partial Differential Equations, 16 (1991), pp. 1095–1128.

[18] H. Ishii and S. Koike, *Viscosity solutions of a system of nonlinear second-order elliptic PDEs arising in switching games*, Funkcial. Ekvac., 34 (1991), pp. 143–155.

[19] E. R. Jakobsen, *On error bounds for monotone approximation schemes for multi-dimensional Isaacs equations*, submitted.

[20] E. R. Jakobsen, *Error bounds for monotone approximation schemes for non-convex degenerate elliptic equations in $\mathbb{R}^1$*, BIT, 44 (2004), pp. 269–285.

[21] E. R. Jakobsen, *On the rate of convergence of approximation schemes for Bellman equations associated with optimal stopping time problems*, Math. Models Methods Appl. Sci., 13 (2003), pp. 613–644.

[22] E. R. Jakobsen and K. H. Karlsen, *Continuous dependence estimates for viscosity solutions of fully nonlinear degenerate parabolic equations*, J. Differential Equations, 183 (2002), pp. 497–525.

[23] E. R. Jakobsen and K. H. Karlsen, *Continuous dependence estimates for viscosity solutions of fully nonlinear degenerate elliptic equations*, Electron. J. Differential Equations, 2002 (2002), pp. 1–10.

[24] E. R. Jakobsen, K. H. Karlsen, and N. H. Risebro, *On the convergence rate of operator splitting for Hamilton–Jacobi equations with source terms*, SIAM J. Numer. Anal., 39 (2001), pp. 499–518.

[25] G. Kossioris, C. Makridakis, and P. Souganidis, *Finite volume schemes for Hamilton–Jacobi equations*, Numer. Math., 83 (1999), pp. 427–442.

[26] N. V. Krylov, *On the rate of convergence of finite-difference approximations for Bellman's equations*, St. Petersburg Math. J., 9 (1997), pp. 639–650.

[27] N. V. Krylov, *On the rate of convergence of finite-difference approximations for Bellman's equations with variable coefficients*, Probab. Theory Related Fields, 117 (2000), pp. 1–16.

[28] O. Lepsky, *Spectral viscosity approximations to Hamilton–Jacobi solutions*, SIAM J. Numer. Anal., 38 (2000), pp. 1439–1453.

[29] C.-T. Lin and E. Tadmor, *L^1-stability and error estimates for approximate Hamilton–Jacobi solutions*, Numer. Math., 87 (2001), pp. 701–735.

[30] P.-L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations. Part* I: *The dynamic programming principle and applications*, Comm. Partial Differential Equations, 8 (1983), pp. 1101–1174.

[31] P.-L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations. Part* II: *Viscosity solutions and uniqueness*, Comm. Partial Differential Equations, 8 (1983), pp. 1229–1276.

[32] P.-L. Lions, *Optimal control of diffusion processes and Hamilton–Jacobi–Bellman equations. Part* III: *Regularity of the optimal cost function*, in Nonlinear Partial Differential Equations and Their Applications, Séminaire du Collège de France, Vol. 5 (Paris, 1981/1982), Pitman, Boston, MA, 1983, pp. 95–205.

[33] P.-L. Lions, *Personal communication*.

[34] P.-L. Lions and B. Mercier, *Approximation numérique des équations de Hamilton–Jacobi–Bellman*, RAIRO Anal. Numér., 14 (1980), pp. 369–393.

[35] P. Soravia, *Estimates of convergence of fully discrete schemes for the Isaacs equation of pursuit-evasion differential games via maximum principle*, SIAM J. Control Optim., 36 (1998), pp. 1–11.

[36] P. E. Souganidis, *Approximation schemes for viscosity solutions of Hamilton–Jacobi equations*, J. Differential Equations, 59 (1985), pp. 1–43.

[37] P. E. Souganidis, *Max-min representations and product formulas for the viscosity solutions of Hamilton–Jacobi equations with applications to differential games*, Nonlinear Anal., 9 (1985), pp. 217–257.

[38] N. Yamada, *Viscosity solutions for a system of elliptic inequalities with bilateral obstacles*, Funkcial. Ekvac., 30 (1987), pp. 417–425.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 559–577

CONVERGENCE OF UPWIND FINITE DIFFERENCE SCHEMES FOR A SCALAR CONSERVATION LAW WITH INDEFINITE DISCONTINUITIES IN THE FLUX FUNCTION*

SIDDHARTHA MISHRA[†]

Abstract. We consider the scalar conservation law with flux function discontinuous in the space variable, i.e.,

$$\begin{aligned} u_t + (H(x)f(u) + (1 - H(x))g(u))_x = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}_+, \\ u(0, x) = u_0(x) \quad \text{in } \mathbb{R}, \end{aligned} \tag{0.1}$$

where H is the Heaviside function and f and g are smooth with the assumptions that either f is convex and g is concave or f is concave and g is convex. The existence of a weak solution of (0.1) is proved by showing that upwind finite difference schemes of Godunov and Enquist–Osher type converge to a weak solution. Uniqueness follows from a Kruzkhov-type entropy condition. We also provide explicit solutions to the Riemann problem for (0.1). At the level of numerics, we give easy-to-implement numerical schemes of Godunov and Enquist–Osher type. The central feature of this paper is the modification of the singular mapping technique (the main analytical tool for these types of equations) which allows us to show that the numerical schemes converge. Equations of type (0.1) with the above hypothesis on the flux may occur when considering the following scalar conservation law with discontinuous flux:

$$\begin{aligned} u_t + (k(x)f(u))_x = 0, \\ u(0, x) = u_0(x), \end{aligned} \tag{0.2}$$

with f convex and k of indefinite sign.

Key words. conservation laws, discontinuous fluxes, finite differences, singular mapping

AMS subject classifications. 35F25, 35L65, 65M06, 65M12, 76S05, 76M12, 76M20

DOI. 10.1137/030602745

1. Introduction.

We are interested in the following scalar conservation law:

$$\begin{aligned} u_t + (f(k(x), u))_x = 0 \quad \text{in } \mathbb{R} \times \mathbb{R}_+, \\ u(0, x) = u_0(x) \quad \text{in } \mathbb{R}. \end{aligned} \tag{1.1}$$

Here, the flux f is smooth in k and u but may depend discontinuously on the space variable through a coefficient k. Note that (0.2) and (0.1) are special cases of (1.1) and are referred to as the multiplicative case and the 2-flux case, respectively.

Equations of type (1.1) are of great practical interest in several areas of physics and engineering, e.g., in the water flooding model in the petroleum industry, in modeling continuous sedimentation in the ideal clarifier thickener units used in waste water treatment plants, in traffic flows on highways with changing road conditions, and in ion etching in the semiconductor industry. For more details on some of these applications, see [26].

As in the continuous case, i.e., when k is Lipschitz, weak solutions are sought for (1.1). For definitions, see section 2. As in the continuous case, we need to impose

*Received by the editors August 30, 2004; accepted for publication (in revised form) September 29, 2004; published electronically June 30, 2005. Part of this work appeared in the author's master's thesis as the chapter "Scalar Conservation Laws with Discontinuous Flux," IISc, 2003.
http://www.siam.org/journals/sinum/43-2/60274.html

[†] Department of Mathematics, Indian Institute of Science, Bangalore, Karuataka 560012, India and TIFR Center, IISc Campus, Bangalore, Karuataka 560012, India (sid@math.tifrbng.res.in).

additional admissibility criteria or entropy conditions in order to obtain stability and uniqueness for the weak solutions. We note that at present, there is no complete theory for the entropy solutions of (1.1).

Equation (0.1) with both f and g convex was investigated by Adimurthi and Gowda in [1] by considering the corresponding Hamilton–Jacobi equation. They obtained an explicit Hopf–Lax-type formula for the solution and used it to prescribe a proper characteristic entropy condition at the interface ($x = 0$) which entailed the omission of undercompressive waves ($f' > 0, g' < 0$) at the interface. Coupled with a Kruzkhov-type entropy condition in the interior, the authors were able to show stability and uniqueness of weak solutions for (0.1). In [3], the authors developed a Godunov-type algorithm for approximating solutions of (0.1). They worked under the hypothesis that f and g have one minima (maxima) and no maxima (minima) in the phase space. It must be noted that the interface entropy condition of [3] is different in some cases from other prevailing entropy theories. In view of this, we are motivated to investigate (1.1) in detail by using [1], [3] as a starting point; the following questions have to be considered:

1. Can we extend the algorithm of [3] and the entropy theory of [1] to the more general equations of the form (1.1)?
2. Can we extend the algorithm of [3] and the entropy theory of [1] to a larger class of fluxes, including nonconvex and sign changing fluxes?
3. Can we develop numerical schemes for (1.1) based on fluxes other than Godunov, such as the Enquist–Osher and Lax–Friedrichs schemes and other higher order versions which are consistent with interface entropy conditions like that of [1]?
4. How does the entropy theory of [1], [3] compare with other entropy theories, and what is the correct notion of the entropy solution of (1.1)?

These questions are being investigated in a series of papers by Adimurthi, Gowda, and the author. This is the first paper in the series and looks at some aspects of questions 2 and 3. Other papers in this series are [4], [5], and [6].

In this paper, we consider the case when one of the fluxes is convex and the other is concave (see section 2 for precise hypotheses on the fluxes). We propose a notion of entropy solutions for (0.1) based on a Kruzkhov-type condition away from the interface. A special feature of the flux geometry considered in this paper is that no extra condition (except the usual Rankine–Hugoniot condition) is required at the interface. We report explicit solutions for the associated Riemann problem and use them to build up an easy-to-implement Godunov scheme. An Enquist–Osher scheme is also proposed and both schemes are shown to converge to the entropy solution of (0.1). It should be noted that a particular case of (0.2) is the case when $f = k_1 h$ and $g = k_2 h$ with h convex and k_1, k_2 of different signs. Thus, this case will serve as a building block for developing numerical methods for (1.1) when k is of indefinite sign. This will be done in a forthcoming paper [6].

Equations of type (1.1) have been dealt with extensively in the literature both from a theoretical as well as a numerical point of view. In [14], [15], Gimse and Risebro obtained solutions for the Riemann problem under the assumptions of convexity and used the solutions to develop a front tracking algorithm to show existence of a weak solution. Uniqueness was obtained by minimizing $|u_+ - u_-|$, the jump at the interface. Diehl investigated equations of type (0.1) with applications in the clarifier thickener unit in [11], [12]. He obtained solutions of the Riemann problem under extremely general hypotheses on the fluxes and showed uniqueness for the Riemann problem by using a variation condition which he termed the Γ condition. Results for the front tracking algorithm were obtained in [23], [24], [22], and [10] (including a time dependent discontinuous coefficient).

The first results for explicit finite difference schemes for (0.2) were obtained by Towers in [28], [29]. In [28], Towers developed a staggered grid scheme for (0.2) which just used the solution of a scalar Riemann problem for continuous flux. In that paper, he developed staggered versions of both the Godunov and the Enquist–Osher schemes for (0.2) when f is convex and k is assumed to be strictly positive. In [29], the author considered (0.2) with nonconvex fluxes and proved convergence for a staggered version of the Enquist–Osher scheme. These works are a motivation for this and other forthcoming papers where we develop and analyze numerical schemes for (1.1) based on exact or approximate Riemann solvers. In this paper, we tackle the question of the indefinite sign of the coefficient, which was left open in [28].

More recently, there has been a series of papers on (1.1). In [19], Karlsen, Risebro, and Towers studied (0.1) with an added degenerate parabolic term by using an Enquist–Osher-type scheme. In [20], the authors considered the vanishing viscosity limit of (0.2) and showed that it exists by using compensated compactness. They also included a degenerate parabolic term. A general entropy theory for (1.1) with degenerate parabolic terms was developed in [21], where well-posedness was shown for fluxes satisfying a certain "crossing condition" and a modified Kruzkhov-type entropy condition which agrees with that of [3] for (0.1) except in the undercompressive intersections case. Among other works are those of Burger et al. on the clarifier thickener model in [9] and of Karlsen, Klingenberg, and Risebro in [18], where a relaxation scheme for (0.2) was proposed and shown to converge.

A very recent paper of Karlsen and Towers [17] deals with (1.1) (including a time dependent discontinuity in the flux) by proposing a Lax–Friedrichs scheme and showing that the approximations converge the entropy solution. They were able to handle very general fluxes and sign-changing coefficients using compensated compactness. Another very recent work that has come to the notice of the author after this paper was completed is that of Audusse and Perthame in [7], in which they proposed an alternative concept of entropy solution for (1.1) by using adapted Kruzkhov entropies and showing uniqueness.

The concept of entropy solution proposed in this paper does not require extra assumptions at the interface on account of the special concave-convex flux geometry. Concepts of entropy solutions that do not require interface entropy conditions were noticed by Bagnerini and Rascle in [8] (with monotone fluxes) and more recently by Audusse and Perthame in [7]. The author wishes to clarify that the hypotheses on the fluxes considered in [7] are very different from the one in this paper. For example, in the special case of (0.1), the authors of [7] required that either both fluxes are monotone or both are convex (with same minimum value), which is different from the concave-convex flux geometry considered here. Similarly the results of [21] do not apply, as the "crossing condition" is not satisfied by the fluxes of this paper.

It is well known in the literature that it is difficult to handle the case of sign-changing coefficients. For instance, we quote the authors in [18]: "Also, sign changes in k are usually ruled out with the singular mapping due to the additional analytical difficulties." One way to overcome this difficulty is the use of compensated compactness such as with the vanishing viscosity and relaxation schemes for (0.2) in [20] and [18], respectively, and for the Lax–Friedrichs scheme in [17]. In this paper, we obtain the first convergence for Godunov and Enquist–Osher-type schemes (for sign-changing coefficients), which have their own place in the hierarchy of numerical methods. It should be emphasized that although the compensated compactness approach makes it easier to handle nonconvex fluxes and sign-changing coefficients, it is the singular mapping approach (with its derivative-type estimates) that gives more regular solutions. In particular, the singular mapping leads to solutions with traces that are required for the entropy theory, such as the BV_t solutions of [21], [9]. Hence, it is the belief of the author that each approach has its own utility and the singular mapping may be more useful keeping in mind the existing entropy theories.

We have organized this paper in the following way. In section 2, we will describe the continuous problem in detail, i.e., a precise description of the hypothesis on fluxes f and g and the initial data. We also define the entropy conditions and show that the entropy solutions of (0.1), if they exist, are unique. In section 3, we give explicit solutions of the Riemann problem for (0.1), in this case satisfying the entropy condition. In section 4, we describe our finite difference schemes of both Godunov and Enquist–Osher type and investigate some of their properties. Section 5 deals with the convergence of the schemes and is the core of this paper. Section 6 describes various numerical experiments comparing our schemes with those of Towers in [28]. In section 7, we derive certain conclusions from this paper.

2. The continuous problem.

As noted earlier, in this section we will describe the continuous problem for (0.1) in some detail. We begin with a precise description of the various hypotheses in this paper. First we give the hypotheses on the fluxes f and g. Let $s < S \in \mathbb{R}$, such that $[s, S]$ is the domain of definition of the fluxes (the phase space). The fluxes satisfy the following hypotheses:

(H_1) $f, g : [s, S] \to \mathbb{R}$ are Lipschitz continuous.

(H_2) $f(s) = g(s), f(S) = g(S)$.

(H_3) Either f is convex and g is concave on $[s, S]$ or f is concave and g is convex on $[s, S]$

As in [3], we can easily extend the convexity hypothesis to the following:

($\overline{H_3}$) Either f has one minima and no maxima and g has one maxima and no minima on $[s, S]$ or f has one maxima and no minima and g has one minima and no minima on $[s, S]$.

We remark that the hypothesis (H_2) is mostly to ensure that the solutions are bounded in L^∞, but this is a sufficient condition and is by no means necessary for boundedness and can be relaxed. The key hypotheses on the fluxes are (H_3) or ($\overline{H_3}$), which make the flux geometry of concave-convex or mixed type. We also need the following constant:

$$M = \max\{Lip(f), Lip(g)\}. \tag{2.1}$$

For any fixed $s < S$, we can have four possible cases of the fluxes, which we enumerate as follows.

Case A_1. f is convex and g is concave with $g(s) \leq g(S)$.

Case A_2. f is convex and g is concave with $g(s) > g(S)$.

Case B_1. f is concave and g is convex with $g(s) \leq g(S)$.

Case B_2. f is concave and g is convex with $g(s) > g(S)$.

Note that the hypothesis on the fluxes implies that the fluxes f and g only intersect at the endpoints of the domain of definition, i.e., s and S. Figure 2.1 depicts the various cases as given above.

Now we state our assumptions on the initial data. As in [3], we will need an estimator of the variation of the initial data, which we denote as $\overline{N}(f, g, u_0)$ and

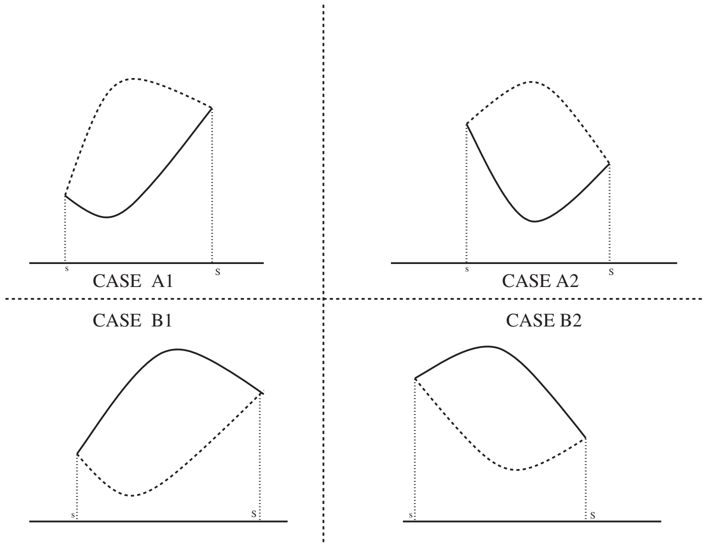


FIG. 2.1. *Various cases of fluxes f and g satisfying the hypotheses of this paper, with the thick line representing f and the dashed line g.*

which we define in section 4. The hypotheses on the initial data are as follows:

(IN_1) $s \leq u_0(x) \leq S \quad \forall x \in \mathbb{R}$.

(IN_2) $\overline{N}(f, g, u_0) \leq C < +\infty$.

We now come to the definition of the entropy solution of (0.1). We start by defining a weak solution of (0.1) as a function $u \in L^\infty_{\text{loc}}(\mathbb{R} \times \mathbb{R}_+)$ such that $\forall \varphi \in C_0^\infty(\overline{\mathbb{R} \times \mathbb{R}_+})$, the following holds:

$$\int_{\mathbb{R}} \int_{\mathbb{R}+} u\varphi_t + (H(x)f(u) + (1 - H(x))g(u))\varphi_x dxdt + \int_{\mathbb{R}} u_0(x)\varphi(0, x)dx = 0\,. \tag{2.2}$$

It is easy to check that u is a weak solution of (0.1) iff it satisfies in the weak sense

$$\begin{aligned} u_t + (g(u))_x &= 0, \quad x < 0, t > 0,\\ u_t + (f(u))_x &= 0, \quad x > 0, t > 0,\\ u(0, x) &= u_0(x) \end{aligned} \tag{2.3}$$

and the following interface Rankine–Hugoniot condition:

$$\begin{gathered} f(u^+(t)) \;=\; g(u^-(t)) \qquad \text{for a.e. } t,\\ u^+(t) = \lim_{x \to 0+} u(x, t), \qquad u^-(t) = \lim_{x \to 0-} u(x, t). \end{gathered} \tag{2.4}$$

We also need to specify a Kruzkhov-type interior entropy condition; for that we define the following:

Entropy pairs. → $\{\varphi_i, \psi_i\}_{i=1,2}$ is said to be a entropy pair for (0.1) if φ_i is convex in $[s, S]$ and $\psi_1'(\theta) = \phi_1'(\theta)f'(\theta), \psi_2'(\theta) = \phi_2'(\theta)g'(\theta)$.

Let $u_0 \in L^\infty(\mathbb{R})$ be the initial data with $s \leq u_0(x) \leq S$ $\forall x \in \mathbb{R}$; then u is said to satisfy the interior entropy condition if, for any entropy pairs $(\varphi_i, \psi_i)_{i=1,2}, u$ satisfies in the sense of distributions the following inequalities:

$$
\begin{aligned}
&\frac{\partial \varphi_1(u)}{\partial t} + \frac{\partial \psi_1(u)}{\partial x} \le 0 \quad \forall x > 0, t > 0,\\
&\frac{\partial \varphi_2(u)}{\partial t} + \frac{\partial \psi_2(u)}{\partial x} \le 0 \quad \forall x < 0, t > 0.
\end{aligned}
\tag{2.5}
$$

Now we define the entropy solution of (0.1) as follows: $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ is said to be an entropy solution for (0.1) if the following hold:

1. u is a weak solution of (0.1).
2. For all $t > 0$, $u^+(x,t)$ and $u^-(x,t)$ exist for a.e. x.
3. u satisfies the interior entropy condition (2.5).

Equipped with the above notation, we state the main existence and uniqueness theorem of this paper.

Theorem 2.1. *Let u_0 satisfy $(IN_1), (IN_2)$ and let the fluxes f and g satisfy H_1, H_2 and $\overline{H_3}$, respectively; then there exists a function $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ such that u is a weak solution of* (0.1) *and the following hold:*

1. *For almost all t and $x \in \mathbb{R}, u(x-,t)$ and $u(x+,t)$ exist.*
2. *u satisfies the interior entropy condition* (2.5) *and is hence unique.*

Remark. Note that unlike in [1], [3], we do not require any interface entropy condition at $x = 0$ for the entropy solutions. This is an interesting outcome of the flux geometry that we are considering, which leads to the fact that any solution that satisfies the interior entropy condition and the Rankine–Hugoniot condition at the interface is unique.

We will show the existence of a weak solution by showing that upwind finite difference schemes of Godunov and Enquist–Osher type converge to it. Currently, we show the uniqueness of the entropy solutions for (0.1). For that as in [1], [3], we need to define the interface entropy functional as

$$
\begin{aligned}
I(u,v,t) = {} & sgn(u^-(t) - v^-(t))(g(u^-(t)) - g(v^-(t)))\\
& - sgn(u^+(t) - v^+(t))(f(u^+(t)) - f(v^+(t))).
\end{aligned}
$$

We have the following lemma regarding the sign of I.

Lemma 2.2. *Let u and v be two entropy solutions of* (0.1)*; then for almost all $t > 0$, we have that $I(u,v,t) \equiv 0$.*

Proof. The proof follows easily from the Rankine–Hugoniot condition (2.4) and from the flux geometry. We leave the proof as an exercise.

Now we are in a position to prove the following uniqueness theorem.

Theorem 2.3. *Let $u, v \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ with $s \le u, v \le S$ be two solutions of* (0.1) *with initial data u_0, v_0, respectively. Assume that*

(i) *for a.e. $t, u(x+,t), v(x+,t), u(x-,t)$, and $v(x-,t)$ exist, and*

(ii) *u, v satisfy the interior entropy condition* (2.5).

Then for any $\overline{M} \ge M$, $a < 0, b > 0$, $b - a \ge 2\overline{M}t$, the function

$$
t \mapsto \int_{a+\overline{M}t}^{b-\overline{M}t} |u(x,t) - v(x,t)| dx
$$

is nonincreasing, and if $u_0 = v_0$ a.e., then it follows that $u = v$ a.e.

Proof. The proof of the above theorem follows exactly as in [1] and is also outlined in [3]. The proof is based on a doubling-of-variables argument (see [25]) coupled with a nonnegative sign of the interface entropy functional obtained in Lemma 2.2. ◻

So, in this section we have established the uniqueness part of the proof of Theorem 2.3 without any extra assumptions on the behavior of the solution at the interface. In the rest of the paper, we will show the existence part of Theorem 2.3.

3. Solutions to the Riemann problem. In this section, we give explicit solutions for the Riemann problem for (0.1) which satisfy both the interior entropy condition and the Rankine–Hugoniot condition (2.4) at the interface. For the sake of simplicity, we will deal only with fluxes satisfying (H_3) and $f, g \in C^2[s, S]$. For fluxes satisfying $(\overline{H_3},)$ the solutions to the Riemann problem can be similarly worked out. We remark that procedures for constructing the solutions of the Riemann problem have been described in [14] and [11]. We carry them out in this case. We are dealing with (0.1) with the following Riemann data:

$$u(x,0) = u_0(x) = \begin{cases} u_r & \text{if} \quad x > 0, \\ u_l & \text{if} \quad x < 0. \end{cases} \tag{3.1}$$

We start with the following case.

Case A_1. We have to distinguish the following subcases.

Case 1. $g(u_l) < g(S)$. From the shape of the fluxes, it is easy to see that there exists a unique $\theta \in (\theta_f, S]$ such that $f(\theta) = g(u_l)$. Let $s_0 = \frac{f(\theta)-f(u_r)}{\theta - u_r}$, $s_1 = \frac{g(u_l)-g(S)}{u_l - S}$, $s_2 = \frac{f(S)-f(u_r)}{S-u_r}$.

The solution in different subcases is given by the following:

Case 1.1. $\theta \leq u_r$.

$$u(x,t) = \begin{cases} u_l & \text{if} \quad x \leq 0, \\ \theta & \text{if} \quad 0 \leq x \leq tf'(\theta), \\ (f')^{-1}(\frac{x}{t}) & \text{if} \quad tf'(\theta) \leq x \leq tf'(u_r), \\ u_r & \text{if} \quad x \geq tf'(u_r). \end{cases}$$

Case 1.2. $\theta > u_r$.

$$\begin{cases} u_l & \text{if} \quad x \leq s_1 t, \\ S & \text{if} \quad s_1 t \leq x \leq s_2 t, \\ u_r & \text{if} \quad x \geq s_2 t. \end{cases}$$

Case 2. $g(u_l) \geq g(S)$. The solution in this case is given by the following:

$$u(x,t) = \begin{cases} u_l & \text{if} \quad x \leq s_1 t, \\ S & \text{if} \quad s_1 t \leq x \leq s_2 t, \\ u_r & \text{if} \quad x \geq s_2 t. \end{cases}$$

Case A_2. Again, we have to consider the following subcases.

Case 1. $f(u_r) > f(S)$. From the shape of the fluxes, it is easy to check that there exists a unique $\eta \in [\theta_g, S]$ such that $g(\eta) = f(u_r)$. Let $s_0 = \frac{g(\eta)-g(u_l)}{\eta - u_l}$, $s_1 = \frac{g(u_l)-g(S)}{u_l - S}$, $s_2 = \frac{f(S)-f(u_r)}{S-u_r}$.

The solution in different subcases is given by the following:

Case 1.1. $u_l > \eta$.

$$u(x,t) = \begin{cases} u_l & \text{if} \quad x \leq g'(u_l)t, \\ (g')^{-1}(\frac{x}{t}) & \text{if} \quad g'(u_l)t \leq x \leq g'(\eta)t, \\ \eta & \text{if} \quad g'(\eta)t \leq x \leq 0, \\ u_r & \text{if} \quad x \geq 0. \end{cases}$$

Case 1.2. $u_l \leq \eta$.

$$\begin{cases} u_l & \text{if} \quad x \leq s_0 t, \\ \eta & \text{if} \quad s_0 t \leq x \leq 0, \\ u_r & \text{if} \quad x \geq 0. \end{cases}$$

Case 2. $f(u_r) \leq f(S)$. The solution in this case is given by the following:

$$u(x,t) = \begin{cases} u_l & \text{if} \quad x \leq s_0 t, \\ \eta & \text{if} \quad s_0 t \leq x \leq 0, \\ u_r & \text{if} \quad x \geq 0. \end{cases}$$

Case B_1. Again, we have to consider the following cases.

Case 1. $g(u_l) \geq g(s)$. In this case, it is easy to see that there exists a unique $\theta \in [s, \theta_f]$ such that $f(\theta) = g(u_l)$. Let $s_0 = \frac{f(\theta)-f(u_r)}{\theta-u_r}$, $s_1 = \frac{g(s)-g(u_l)}{s-u_l}$, and $s_2 = \frac{f(s)-f(u_r)}{s-u_r}$.

The solution in different subcases is given by the following:

Case 1.1. $u_r < \theta$.

$$u(x,t) = \begin{cases} u_l & \text{if} \quad x \leq 0, \\ \theta & \text{if} \quad 0 \leq x \leq tf'(\theta), \\ (f')^{-1}(\frac{x}{t}) & \text{if} \quad tf'(\theta) \leq x \leq tf'(u_r), \\ u_r & \text{if} \quad x \geq tf'(u_r). \end{cases}$$

Case 1.2. $u_r \geq \theta$.

$$\begin{cases} u_l & \text{if} \quad x \leq 0, \\ \theta & \text{if} \quad 0 \leq x \leq s_0 t, \\ u_r & \text{if} \quad x \geq s_0 t. \end{cases}$$

Case 2. $g(u_l) < g(s)$. The solution is given by the following:

$$u(x,t) = \begin{cases} u_l & \text{if} \quad x \leq s_1 t, \\ s & \text{if} \quad s_1 t \leq x \leq s_2 t, \\ u_r & \text{if} \quad x \geq s_2 t. \end{cases}$$

Case B_2. We have to consider the following cases.

Case 1. $f(u_r) \leq f(s)$. In this case it is easy to see that there exists a unique $\eta \in [s, \theta_g]$ such that $f(u_r) = g(\eta)$. Let $s_0 = \frac{g(\eta)-g(u_l)}{\eta-u_l}$, $s_1 = \frac{g(s)-g(u_l)}{s-u_l}$, and $s_2 = \frac{f(s)-f(u_r)}{s-u_r}$.

The solution in different cases is given by the following:

Case 1.1. $u_l < \eta$.

$$u(x,t) = \begin{cases} u_l & \text{if} \quad x \leq g'(u_l)t, \\ (g')^{-1}(\frac{x}{t}) & \text{if} \quad g'(u_l)t \leq x \leq g'(\eta)t, \\ \eta & \text{if} \quad g'(\eta)t \leq x \leq 0, \\ u_r & \text{if} \quad x \geq 0. \end{cases}$$

Case 1.2. $u_l \geq \eta$.

$$\begin{cases} u_l & \text{if} \quad x \leq s_0 t, \\ \eta & \text{if} \quad s_0 t \leq x \leq 0, \\ u_r & \text{if} \quad x \geq 0. \end{cases}$$

Case 2. $f(u_r) > f(s)$.

$$u(x,t) = \begin{cases} u_l & \text{if} \quad x \leq s_1 t, \\ s & \text{if} \quad s_1 t \leq x \leq s_2 t, \\ u_r & \text{if} \quad x \geq s_2 t. \end{cases}$$

So we have given explicit solutions of the Riemann problem for (0.1) in all cases. The crucial fact about the solutions is that the solutions in some cases are the endpoints s and S, which are undercompressive in the sense that at s or S, we have $f' > 0, g' < 0$. A strange result of the flux geometry considered here is that we cannot avoid undercompressive waves at the interface, which is quite striking given that the interface entropy condition of [1], [3] essentially implies omitting undercompressive waves at the interface.

4. Description of the finite difference schemes. In this section, we seek to develop finite difference schemes for (0.1) which are of Godunov as well as Enquist–Osher type. As in [3], the key is to define the interface numerical fluxes; we start with the Godunov flux.

4.1. Godunov flux. Let h be a Lipschitz continuous function defined on $[s, S]$. We use the standard Godunov flux H defined by

$$\begin{aligned} H(a,b) &= \min_{\theta\in[a,b]} h(\theta) \quad \text{if } a \leq b \\ &= \max_{\theta\in[b,a]} h(\theta) \quad \text{if } a \geq b. \end{aligned}$$

We recall that H as defined above is Lipschitz in both its variables, nondecreasing in the first variable, and nonincreasing in the second variable. We use a similar method to define the interface Godunov flux. We use the explicit solutions of the Riemann problem in section 3 to define the interface Godunov flux. Let F and G be the standard Godunov flux corresponding to the fluxes f and g, respectively. We define the interface flux $\overline{F}$ in each case as follows.

Case A_1. $\overline{F}(a,b) = \min\{G(a,S), F(S,b)\}$.

Case A_2. $\overline{F}(a,b) = \max\{G(a,S), F(S,b)\}$.

Case B_1. $\overline{F}(a,b) = \max\{G(a,s), F(s,b)\}$.

Case B_2. $\overline{F}(a,b) = \min\{G(a,s), F(s,b)\}$.

Another way of writing down the above formulas is by defining the following function, which we define only in Case A_1 as follows: Let $\overline{f} : [s, S] \mapsto \mathbb{R}$, such that

$$\begin{aligned} \overline{f}(\theta) &= g(\theta) \quad \forall s \leq \theta \leq \overline{S} \\ &= g(S) \quad \text{otherwise.} \end{aligned} \tag{4.1}$$

We call the above-defined function $\overline{f}$ the interface function and claim that the interface Godunov flux $\overline{F}$ is the standard Godunov flux corresponding to $\overline{f}$. This fact is easy to check. The interface function corresponding to other cases can be easily constructed.

4.2. Enquist–Osher flux. One of the common approximate Riemann solvers that can be used in place of the Godunov flux is the Enquist–Osher flux developed in [13]. Let h be a lipschitz function; then the Enquist–Osher flux $\tilde{H}$ is given by

$$\tilde{H}(a,b) = \frac{1}{2}\left(h(a) + h(b) - \int_a^b |h'(\xi)|d\xi\right). \tag{4.2}$$

For a convex-type fluxes (fluxes with one minima and no maxima), we have a simple explicit formula given by

$$\tilde{H}(a,b) = h(\max(a,\theta)) + h(\min(\theta,b)) - h(\theta), \tag{4.3}$$

where θ is the unique minimum of the function h. The Enquist–Osher flux is similar to the Godunov flux except in the overcompressive case. Next, we have to define a suitable interface Enquist–Osher flux. As the Enquist–Osher fluxes are not based on exact solutions of the Riemann problem, we will employ the interface function $\overline{f}$ to construct the interface Enquist–Osher flux. Let $\tilde{F}$ and $\tilde{G}$ be the Enquist–Osher fluxes corresponding to f and g, respectively. We construct the interface Enquist–Osher flux only for Case A_1; other cases can be similarly treated. We define the interface Enquist–Osher flux ($\tilde{\overline{F}}$) in this case to be the standard Enquist–Osher flux corresponding to the interface function $\overline{f}$. We recall that the interface Godunov flux ($\overline{F}$) can be identified as the standard Godunov flux corresponding to the interface function ($\tilde{f}$), and we are following the same identification for defining the interface Enquist–Osher flux, but the concave-convex flux geometry in this case forces us to have the following proposition, which is easy to check.

PROPOSITION 4.1. *Let $\overline{F}$ and $\overline{\tilde{F}}$ be as defined above; then $\overline{F}(a,b) = \overline{\tilde{F}}(a,b)\ \forall a,b \in [s,S]$.*

Henceforth, we will also refer to the interface Enquist–Osher flux as $\overline{F}$. We collect some easy-to-verify facts regarding the interface flux in the following proposition.

PROPOSITION 4.2. *Let $\overline{F}$ be as given above and $a,b \in [s,S]$. Then the following holds:*

(a) *$\overline{F}$ is Lipschitz in each variable.*

(b) *$\overline{F}$ is nondecreasing in a and nonincreasing in b.*

(c) *$\overline{F}(s,s) = f(s) = g(s), \overline{F}(S,S) = f(s) = g(S)$.*

(d) *$\overline{F}$ is not consistent, i.e., there exists $a \in [s,S]$ such that $\overline{F}(a,a) \neq f(a) \neq g(a)$.*

Proof. (c) and (d) are easy to check from the definition of $\overline{F}$. (a) and (b) follow from the fact that $\overline{F}$ is the standard Godunov (Enquist–Osher) flux corresponding to the Lipschitz continuous function $\overline{f}$.

We remark that, in general, the interface Godunov and Enquist–Osher fluxes may not agree (for example, when both fluxes are convex), but in this case, on account of the flux geometry, they are exactly the same.

We are now in a position to describe the scheme. First, we describe the discretization in space and time as follows.

Let $h > 0$ and define the space grid points x_j as follows:

$$x_j = \left(\frac{2j-1}{2}\right)h \quad \text{for } j \geq 1, \quad x_j = \left(\frac{2j+1}{2}\right)h \quad \text{for } j \leq -1.$$

For time discretization, the time step $\Delta t > 0$, and let $t_n = n\Delta t$. We also introduce $\lambda = \frac{\Delta t}{h}$.

For a function $u_0 \in L^\infty(\mathbb{R})$ we define

$$u^0_{j+1} = \frac{1}{h}\int_{x_{j+1/2}}^{x_{j+3/2}} u_0(x)dx \quad \text{if } j \geq 0, \qquad u^0_{j-1} = \frac{1}{h}\int_{x_{j-3/2}}^{x_{j-1/2}} u_0(x)dx \quad \text{if } j \leq 0,$$

$$\begin{aligned} N_h(f,g,u_0) &= \sum_{i<-1} |G(u^0_i, u^0_{i+1}) - G(u^0_{i-1}, u^0_i)| + \sum_{i>1} |F(u^0_i, u^0_{i+1}) - F(u^0_{i-1}, u^0_i)| \\ &\quad + |\overline{F}(u^0_{-1}, u^0_1) - G(u^0_{-2}, u^0_{-1})| + |F(u^0_1, u^0_2) - \overline{F}(u^0_{-1}, u^0_1)|, \\ \tilde{N}_h(f,g,u_0) &= \sum_{i<-1} |\tilde{G}(u^0_i, u^0_{i+1}) - \tilde{G}(u^0_{i-1}, u^0_i)| + \sum_{i>1} |\tilde{F}(u^0_i, u^0_{i+1}) - \tilde{F}(u^0_{i-1}, u^0_i)| \\ &\quad + |\overline{F}(u^0_{-1}, u^0_1) - \tilde{G}(u^0_{-2}, u^0_{-1})| + |\tilde{F}(u^0_1, u^0_2) - \overline{F}(u^0_{-1}, u^0_1)|, \end{aligned}$$

$$N(f,g,u_0) = \sup_{h>0} \max\{N_h(f,g,u_0), \tilde{N}_h(f,g,u_0)\}.$$

It is easy to see that if $u_0 \in BV(\mathbb{R})$, then $N(f,g,u_0) \leq C||u_0||_{BV}$, where C is a constant depending only on the Lipschitz constants of f and g.

Now, we are in a position to describe our Godunov-type scheme. For every time level $(n+1)$, we calculate the discrete u_j^{n+1} as

$$\begin{aligned} u_j^{n+1} &= u_j^n - \lambda(F(u_j^n, u_{j+1}^n) - F(u_{j-1}^n, u_j^n)) \quad \forall j \geq 2, \\ u_1^{n+1} &= u_1^n - \lambda(F(u_1^n, u_2^n) - \overline{F}(u_{-1}^n, u_1^n)), \\ u_{-1}^{n+1} &= u_{-1}^n - \lambda(\overline{F}(u_{-1}^n, u_1^n) - G(u_{-2}^n, u_{-1}^n)), \\ u_j^{n+1} &= u_j^n - \lambda(G(u_i^n, u_{j+1}^n) - G(u_{j-1}^n, u_j^n)) \quad \forall j \leq -2. \end{aligned} \tag{4.4}$$

Similarly, we define the Enquist–Osher-type scheme for (0.1) as

$$\begin{aligned} u_j^{n+1} &= u_j^n - \lambda(\tilde{G}(u_j^n, u_{j+1}^n) - \tilde{G}(u_{j-1}^n, u_j^n)) \quad \forall j \leq -2, \\ u_{-1}^{n+1} &= u_{-1}^n - \lambda(\overline{F}(u_{-1}^n, u_1^n) - \tilde{G}(u_{-2}^n, u_{-1}^n)), \\ u_1^{n+1} &= u_1^n - \lambda(\tilde{F}(u_1^n, u_2^n) - \overline{F}(u_{-1}^n, u_1^n)), \\ u_j^{n+1} &= u_j^n - \lambda(\tilde{F}(u_j^n, u_{j+1}^n) - \tilde{F}(u_{j-1}^n, u_j^n)) \quad \forall j \geq 2. \end{aligned} \tag{4.5}$$

We now define the approximations in terms of the following piecewise constant functions:

$$u^h(x,t) = u_j^n \quad \text{if} \quad \tilde{u}^h(x,t) = \tilde{u}_j^n, \quad \text{if} \quad (x,t) \in [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}) \times [t^n, t^{n+1}). \tag{4.6}$$ □

In the next section, we shall show that these approximations u^h and $\tilde{u}^h$ are compact in the appropriate topology and converge to a weak solution of (0.1). In addition, the limit satisfies the interior entropy condition and is hence unique.

5. Convergence analysis. In this section, we carry out the convergence analysis for our schemes (4.4) and (4.5). We begin by showing that the solutions are bounded on account of the monotonicity of the scheme and the consistency of the fluxes at the endpoints. This gives compactness in the weak star topology which is not enough to pass to the limit in the nonlinear flux terms. In order to do so, we have to obtain derivative-type estimates, which for equations of type (1.1) is generally measured via the singular mapping introduced by Temple. This tool has been adopted in several papers on these equations. See [23], [28], [29]. To start with, we give some preliminary lemmas for convergence. We define the following functions to write our schemes in conservation form:

$$\begin{aligned} H_1(X,Y,Z) &= Y - \lambda(F(Y,Z) - F(X,Y)), \\ H_{-1}(X,Y,Z) &= Y - \lambda(G(Y,Z) - G(X,Y)), \\ H_2(X,Y,Z) &= Y - \lambda(F(Y,Z) - \overline{F}(X,Y)), \\ H_{-2}(X,Y,Z) &= Y - \lambda(\overline{F}(Y,Z) - G(X,Y)). \end{aligned}$$

Similarly we can define the corresponding $\overline{H}_i$'s for the Enquist–Osher scheme (4.5) by replacing the Godunov fluxes by the Enquist–Osher fluxes described in section 4. We have the following lemma.

LEMMA 5.1. *Let $2\lambda M \leq 1$ and $a \in [s,S]$; we have the following:*

1. $H_{\pm 1}(a,a,a) = \tilde{H}_{\pm 1}(a,a,a) = a, H_{\pm 2}(s,s,s) = \tilde{H}_{\pm 2}(s,s,s) = s$*,* $H_{\pm 2}(S,S,S) = \tilde{H}_{\pm 2}(S,S,S) = S$*.*
2. $H_i, \tilde{H}_i$ *are nondecreasing in each of their variables, and hence the schemes* (4.4), (4.5) *are monotone.*

Proof. Statement 1 follows directly from the definition of the fluxes and from Proposition 4.2. For statement 2, check that $H_{\pm 1}, \tilde{H}_{\pm 1}$ are monotone due to standard arguments. We check monotonicity only for (H_2); the other cases follow similarly. As (H_2) is Lipschitz, we differentiate it with respect to the variables and use Proposition 4.2 coupled with the CFL condition to get that

$$\begin{aligned}
\frac{\partial H_2}{\partial X} &= \lambda \frac{\partial(\overline{F})(X,Y)}{\partial a} \geq 0,\\
\frac{\partial H_2}{\partial Z} &= -\lambda \frac{\partial \overline{F}(Y,Z)}{\partial b} \geq 0,\\
\frac{\partial H_2}{\partial Y} &= 1 - \lambda \frac{\partial F(Y,Z)}{\partial a} + \frac{\partial \overline{F}(X,Y)}{\partial b} \geq 1 - 2\lambda M \geq 0;
\end{aligned}$$

hence we have shown that (H_2) is monotone in each variable and have proved the lemma. □

Now, we prove that the approximate solutions are uniformly bounded in L^∞ by the following invariant region principle.

LEMMA 5.2. *Let $u_0 \in L^\infty(\mathbb{R}, [s,S])$ be the initial data and let $u_j^n, \tilde{u}_j^n$ be the corresponding solutions calculated by the Godunov scheme* (4.4) *and the Enquist–Osher scheme* (4.5)*, respectively. Then*

$$s \leq u_j^n, \quad \tilde{u}_j^n \leq S \ \ \forall j, n. \tag{5.1}$$

Proof. We will prove for u_j^n only; the other case follows similarly. Since $s \leq u_0 \leq S$, and hence $\forall j,\ s \leq u_j^0 \leq S$. By induction, assume that (5.1) holds for n. Then from Lemma 5.1, we have

$$\begin{aligned}
&s = H_{-1}(s,s,s) \leq H_{-1}(u_{j-1}^n, u_j^n, u_{j+1}^n) = u_j^{n+1} \leq H_{-1}(S,S,S) = S \quad \text{if } \ j \leq -2,\\
&s = H_{1}(s,s,s) \leq H_{1}(u_{j-1}^n, u_j^n, u_{j+1}^n) = u_j^{n+1} \leq H_{1}(S,S,S) = S \quad \text{if } \ j \geq 2,\\
&s = H_{-2}(s,s,s) \leq H_{-2}(u_{-2}^n, u_{-1}^n, u_1^n) = u_{-1}^{n+1} \leq H_{-2}(S,S,S) = S,\\
&s = H_{2}(s,s,s) \leq H_{2}(u_{-1}^n, u_1^n, u_2^n) = u_1^{n+1} \leq H_{2}(S,S,S) = S.
\end{aligned}$$

This proves (5.1). □

Next we obtain the discrete L^1 contractivity estimate in the following.

LEMMA 5.3. *Let $u_0, v_0 \in L^\infty(\mathbb{R},\ [s,S])$ be the initial data and let $\{u_j^n\}$ and $\{v_j^n\}$ be the corresponding solutions calculated by the Godunov-type scheme* (4.4)*; let $\tilde{u_j^n}$ and $\tilde{v_j^n}$ be the corresponding solutions calculated by the Enquist–Osher scheme* (4.5)*. Then the following hold:*

$$\sum_{j\neq 0} |u_j^{n+1} - u_j^n| \leq \sum_{j\neq 0} |u_j^n - u_j^{n-1}|, \tag{5.2}$$

$$\sum_{j\neq 0} |\tilde{u}_j^{n+1} - \tilde{u}_j^n| \leq \sum_{j\neq 0} |\tilde{u}_j^n - \tilde{u}_j^{n-1}|. \tag{5.3}$$

Proof. The proof follows from monotonicity of the schemes and from the Crandall–Tartar lemma (see [16]).

Next we define the singular mappings. To start with, we use the standard notation $a \in \mathbb{R}$; then $a_+ = \max\{a,0\}, a_- = \min\{a,0\}, a = a_+ + a_-, |a| = a_+ - a_-$.

The singular mappings are given by

$$\psi_1(\theta) = \int_\alpha^\theta |g'(\xi)|d\xi, \quad \psi_2(\theta) = \int_\alpha^\theta |f'(\xi)|d\xi, \tag{5.4}$$

where $\alpha \in [s, S]$ is some number. Now we are in a position to define the transformed schemes for the discrete values of the solution as follows:

$$z_j^n = \begin{cases} \psi_1(u_j^n) & \text{if} \quad j \le -1, \\ \psi_1(u_{-1}^n) & \text{if} \quad j \ge -1, \end{cases} \qquad w_j^n = \begin{cases} \psi_2(u_1^n) & \text{if} \quad j \le 1, \\ \psi_2(u_j^n) & \text{if} \quad j \ge 1. \end{cases} \quad \square \tag{5.5}$$

We remark that unlike in [3], we define two sets of transformed variables in order to obtain the normalized variation bounds. This is a key difference in the analysis which enables us to show the convergence for finite difference approximations for the concave-convex case. This feature of using a set of singular mappings instead of the one considered so far in the literature appears to be novel and will be employed by us in the more general nonconvex case. We will show that the total variation of the approximate solutions under the above transformations is bounded in terms of the variation of the numerical fluxes. This is done in the following lemma, which gives a bound on the normalized variation in each cell.

LEMMA 5.4. *For all $j \in Z$, $\forall n \in \mathbb{N}$, we have the following inequalities:*

$$\forall j \le -3$$

$$(z_j^n - z_{j+1}^n)_+ \le |G(u_j^n, u_{j+1}^n) - G(u_{j-1}^n, u_j^n)| + |G(u_{j+1}^n, u_{j+2}^n) - G(u_j^n, u_{j+1}^n)|, \tag{5.6}$$

$$\forall j \ge 2$$

$$-(w_j^n - w_{j+1}^n)_- \le |F(u_j^n, u_{j+1}^n) - F(u_{j-1}^n, u_j^n)| + |F(u_{j+1}^n, u_{j+2}^n - F(u_j^n, u_{j+1}^n)|, \tag{5.7}$$

$$(z_{-2}^n - z_{-1}^n)_+ \le |G(u_{-2}^n, u_{-1}^n) - \overline{F}(u_{-1}^n, u_1^n)| + |G(u_{-3}^n, u_{-2}^n) - G(u_{-2}^n, u_{-1}^n)|, \tag{5.8}$$

$$-(w_1^n - w_2^n)_- \le |F(u_1^n, u_2^n) - \overline{F}(u_{-1}^n, u_1^n)| + |F(u_2^n, u_3^n) - F(u_1^n, u_2^n)|. \tag{5.9}$$

The above estimates also hold for the Enquist–Osher approximations by replacing u_j^n, z_j^n, w_j^n by $\tilde{u}_j^n, \tilde{z}_j^n, \tilde{w}_j^n$, respectively. We call these analogous estimates (5.6(a))*,* (5.7(a))*,* (5.8(a))*, and* (5.9(a))*, respectively.*

Proof. We provide proofs for (5.6), (5.8), and (5.7(a)). The other inequalities can be similarly proved.

Proof of (5.6). We fix t in the subsequent calculations and drop the superscript n in the notation. We have to consider three separate cases. First check that $(z_j - z_{j+1})_+ > 0$ iff $u_{j+1} < u_j$.

Case 1. $s \le u_{j+1} < u_j < \theta_g$. In this case, it is easy to check that we have the following:

$$(z_j - z_{j+1})_+ = g(u_j) - g(u_{j+1}), \quad G(u_j, u_{j+1}) = g(u_j). \tag{5.10}$$

For any $u_{j+2} \in [s, S]$, we have the following. If $u_{j+2} < u_{j+1}$, then we can check from the definition of G that $G(u_{j+1}, u_{j+2}) = g(u_{j+1})$, and if $u_{j+2} \ge u_{j+1}$, then by the fact that G is nonincreasing in the second variable and by its consistency, we get that $G(u_{j+1}, u_{j+2}) \le G(u_{j+1}, u_{j+1}) = g(u_{j+1})$. In either case, we have that $G(u_{j+1}, u_{j+2}) \le g(u_{j+1})$. By combining the above inequality with (5.10), we get the desired inequality:

$$(z_j^n - z_{j+1}^n)_+ \le |G(u_{j+1}^n, u_{j+2}^n) - G(u_j^n, u_{j+1}^n)|. \tag{5.11}$$

Case 2. $s \leq u_{j+1} < \theta_g \leq u_j$. In this case, we get directly from the definition that

$$(z_j - z_{j+1})_+ = g(\theta_g) - g(u_j) + g(\theta_g) - g(u_{j+1}), \quad G(u_j, u_{j+1}) = g(\theta_g). \tag{5.12}$$

For any $u_{j-1} \in [s, S]$, we have the following. If $u_{j-1} > u_j$, then we can check that $G(u_{j-1}, u_j) = g(u_j)$, and if $u_{j-1} \leq u_j$, then by the fact that G is nondecreasing in the first variable and by its consistency, we get that $G(u_{j-1}, u_j) \leq G(u_j, u_j) = g(u_j)$. In either case, we have that $G(u_{j-1}, u_j) \leq g(u_j)$. Now by combining the above estimate with (5.12) and the previous case, we get (5.6).

Case 3. $\theta_g \leq u_{j+1} < u_j \leq S$. In this case, from direct calculations we have that

$$(z_j - z_{j+1})_+ = g(u_{j+1}) - g(u_j), \quad G(u_j, u_{j+1}) = g(u_{j+1}). \tag{5.13}$$

Just by using (5.12) and (5.13), we get the desired inequality (5.8) and thus show it in all the three possible cases.

Proof of (5.8). As in the previous case, we have that $(z_{-2} - z_{-1})_+ > 0$ iff $u_{-2} > u_{-1}$, and we have to consider the following cases.

Case 1. $\theta_g \leq u_{-1} < u_{-2} \leq S$. The estimate (5.8) follows exactly as in Case 3 of the previous proof and we get

$$(z_{-2} - z_{-1})_+ \leq |G(u^n_{-2}, u^n_{-1}) - G(u^n_{-3}, u^n_{-2})|. \tag{5.14}$$

Case 2. $u_{-1} < u_{-2} \leq \theta_g$. In this case, from direct calculations we get that

$$(z_{-2} - z_{-1})_+ = g(u_{-2}) - g(u_{-1}), \quad G(u_{-2}, u_{-1}) = g(u_{-2}). \tag{5.15}$$

Now for any $u_1 \in [s, S]$, we have from the definition of the interface Godunov flux and its monotonicity properties that if $u_{-1} \leq \overline{S}$, $\overline{F}(u_{-1}, u_1) \leq \overline{F}(u_{-1}, s) = g(u_{-1})$; similarly, in the case $u_{-1} > \overline{S}$, we have $\overline{F}(u_{-1}, u_1) \leq g(S)$. In either case, we get that $\overline{F}(u_{-1}, u_1) \leq g(u_{-1})$. It is easy to check that (5.8) follows by combining the above inequality with (5.15).

Case 3. $u_{-1} \leq \theta_g < u_{-2}$. In this case, we verify that

$$(z_{-2} - z_{-1})_+ = g(\theta_g) - g(u_{-2}) + g(\theta_g) - g(u_{-1}), \quad G(u_{-2}, u_{-1}) = g(\theta_g). \tag{5.16}$$

By combining (5.16) with the estimates on the interface flux obtained in Case 2, we prove (5.8).

Proof of (5.7(a)). First check that $(\tilde{w}_j - \tilde{w}_{j+1})_- < 0$ iff $\tilde{u}_j < \tilde{u}_{j+1}$. We have to distinguish between three separate cases.

Case 1. $s \leq \tilde{u}_j < \tilde{u}_{j+1} \leq \theta_f$. In this case, we verify from direct calculations and the definition of the numerical flux $\tilde{F}$ that the following holds:

$$-(\tilde{w}_j - \tilde{w}_{j+1})_- = f(\tilde{u}_j) - f(\tilde{u}_{j+1}), \quad \tilde{F}(\tilde{u}_j, \tilde{u}_{j+1}) = f(\tilde{u}_{j+1}). \tag{5.17}$$

For any $\tilde{u}_{j-1} \in [s, S]$, we have the following three cases, i.e., if $\tilde{u}_{j-1} < \tilde{u}_j$, then from the definition of the Enquist–Osher flux it follows that $\tilde{F}(\tilde{u}_{j-1}, \tilde{u}_j) = f(\tilde{u}_j)$, and if $\tilde{u}_{j-1} > \tilde{u}_j$, then by the fact that the Enquist–Osher flux is nonincreasing in the first variable, it follows that $\tilde{F}(\tilde{u}_{j-1}, \tilde{u}_j) \geq \tilde{F}(\tilde{u}_j, \tilde{u}_j) = f(\tilde{u}_j)$. In either case we have $\tilde{F}(\tilde{u}_{j-1}, \tilde{u}_j) \geq f(\tilde{u}_j)$. So by combining this estimate with (5.17), we prove (5.7(a)) in this case.

Case 2. $\theta_f \leq \tilde{u}_j \leq \tilde{u}_{j+1} \leq S$. In this case, we verify from direct calculations and the definition of the numerical flux $\tilde{F}$ that the following holds:

$$-(\tilde{w}_j - \tilde{w}_{j+1})_- = f(\tilde{u}_{j+1}) - f(\tilde{u}_j), \quad \tilde{F}(\tilde{u}_j, \tilde{u}_{j+1}) = f(\tilde{u}_j). \tag{5.18}$$

For any $\tilde{u}_{j+2} \in [s, S]$, we have the following two cases, i.e., if $\tilde{u}_{j+1} < \tilde{u}_{j+2}$, then from the definition of the Enquist–Osher flux it follows that $\tilde{F}(\tilde{u}_{j+1}, \tilde{u}_{j+2}) = f(\tilde{u}_{j+1})$, and if $\tilde{u}_{j+2} \leq \tilde{u}_{j+1})$, then by the fact that the Enquist–Osher flux is nondecreasing in the first variable, it follows that $\tilde{F}(\tilde{u}_{j+1}, \tilde{u}_{j+2}) \geq \tilde{F}(\tilde{u}_{j+1}, \tilde{u}_{j+1}) = f(\tilde{u}_{j+1})$. In either case we have $\tilde{F}(\tilde{u}_{j+1}, \tilde{u}_{j+2}) \geq f(\tilde{u}_{j+1})$. So by combining the above with (5.18) we show (5.7(a)) in this case.

Case 3. $\tilde{u}_j \leq \theta_f \leq \tilde{u}_{j+1}$. In this case, we have that

$$-(\tilde{w}_j - \tilde{w}_{j+1})_- = f(\tilde{u}_j) - f(\theta_f) + f(\tilde{u}_{j+1}) - f\theta_f), \quad \tilde{F}(\tilde{u}_j, \tilde{u}_{j+1}) = f(\theta_f). \tag{5.19}$$

Now by combining (5.19) with the estimates obtained in the two previous cases, we prove (5.7(a)) in every case. Thus, we have shown the required estimates in all cases. Other estimates can be proved similarly, and we can complete the proof of Lemma 5.4. □

We use the above cell normalized variation inequalities in order to get the bounds on the total variation of the transformed schemes. More precisely, we have the following lemma.

LEMMA 5.5. *Let $z^n_j, w^n_j, \tilde{z}^n_j, \tilde{w}^n_j$ be as defined above. Then the following holds:*

$$\max\{TV(z^n_j), TV(\tilde{z}^n_j)\} \leq \frac{4}{\lambda} \max\left\{ \sum_{j\neq 0} |u^1_j - u^0_j|, \sum_{j\neq 0} |\tilde{u}^1_j - \tilde{u}^0_j| \right\}, \tag{5.20}$$

$$\max\{TV(\tilde{w}^n_j), TV(w^n_j)\} \leq \frac{4}{\lambda} \max\left\{ \sum_{j\neq 0} |u^1_j - u^0_j|, \sum_{j\neq 0} |\tilde{u}^1_j - \tilde{u}^0_j| \right\}.$$

Proof. We prove the above estimate for the sequence z^n_j. The other sequences are shown to satisfy the above inequality in the same way. First observe that $\forall j \geq -1$, we have $(z^n_j - z^n_{j+1})_+ \equiv 0$. So we have that $TV(z^n_j) = \sum |(z^n_j - z^n_{j+1}| = 2\sum(z^n_j - z^n_{j+1})_+$. Therefore by adding (5.6), (5.7), and the above inequality over all j, we get that

$$\begin{aligned}\sum(z^n_j - z^n_{j+1})_+ = 2\Bigg(&\sum_{j\leq -2} |G(u^n_j, u^n_{j+1}) - G(u^n_{j-1}, u^n_j)| + |G(u^n_{-2}, u^n_{-1}) - \overline{F}(u^n_{-1}, u^j_1)| \\ &+ |\overline{F}(u^n_{-1}, u^n_1) - F(u^n_1, u^n_2)| + \sum_{j\geq 2} |F(u^n_j, u^n_{j+1}) - F(u^n_{j-1}, u^n_j)|\Bigg) \\ &= \frac{2}{\lambda} \sum_{j\neq 0} |u^{n+1}_j - u^n_j|. \end{aligned} \tag{5.21}$$

From (5.2), we can get the required estimate on the right-hand side of (5.21), and the inequality follows. In a similar way, we can get the bounds for the other transformed sequences.

Now we define piecewise constant functions based on the transformed sequences and write some estimates on these functions, which follow in a straightforward way from Lemma 5.5. Define the piecewise constant functions $z^h, w^h, \tilde{z}^h, \tilde{w}^h$ by $z^h(x,t) = z^n_j, w^h(x,t) = w^n_j, \tilde{z}^h(x,t) = \tilde{z}^n_j, \tilde{w}^h(x,t) = \tilde{w}^n_j\ \forall (x,t) \in I^n_j$. Then the following estimates hold.

LEMMA 5.6. *With the functions defined as above and $\forall t \in \mathbb{R}_+$, we have*

$$\max\{TV(z^h), TV(w^h), TV(\tilde{z}^h), TV(\tilde{w}^h)\} \leq \frac{4}{\lambda} N(f, g, u_0). \tag{5.22}$$

Proof. The proof follows directly from Lemma 5.5.

We need time continuity estimates in L^1 for the approximations u^h, which is given in the following lemma.

LEMMA 5.7. *Let $u_0, v_0 \in L^\infty(\mathbb{R}, [s, S])$ such that $N(f, g, u_0) < \infty, N(f, g, v_0) < \infty$ are the initial data, let u_h and v_h be the corresponding solutions obtained by the Godunov scheme* (4.4)*, and let $\tilde{u}_h$ and $\tilde{v}_h$ be the corresponding solutions given by the Enquist–Osher-type scheme* (4.5)*. Then*

$$s \leq u_h(x,t), \tilde{u}_h(x,t) \leq S \qquad \forall (x,t) \in \mathbb{R} \times \mathbb{R}_+, \tag{5.23}$$

$$\max\left\{\int_{\mathbb{R}} |u_h(x,t) - u_h(x,\tau)|dx, \int_{\mathbb{R}} |\tilde{u}_h(x,t) - \tilde{u}^h(x,\tau)|dx\right\} \leq N(f,g,u_0)(2\Delta t + |t-\tau|).$$

Proof. The first inequality follows directly from (5.1). We prove the second inequality for the Godunov approximations; the Enquist–Osher approximations follow similarly.

Let $t_n \leq t < t_{n+1}$ and $t_m \leq \tau < t_{m+1}$, so

$$|n-m|\Delta t = |t_n - t_m| \leq |t_n - t| + |t - \tau| + |\tau - t_m| \leq 2\Delta t + |t - \tau|.$$

Hence from Lemma 5.3, we obtain

$$\begin{aligned}
\int_{\mathbb{R}} |u_h(x,t) - u_h(x,\tau)|dx &= h\sum_{j\neq 0} |u_j^n - u_j^m| \\
&\leq h\sum_{j\neq 0}\sum_{i=0}^{n-m+1} |u_j^{n-i} - u_j^{n-i-1}| \\
&\leq h|n-m|\sum_{j\neq 0} |u_j^1 - u_j^0| \\
&\leq \frac{\Delta t|n-m|}{\lambda}\sum_{j\neq 0} |u_j^1 - u_j^0| \\
&\leq (2\Delta t + |t-\tau|)N(f,g,u_0).
\end{aligned}$$

We complete the proof of the lemma. □

Now equipped with the above lemmas, we proceed to state and prove our main convergence theorem.

THEOREM 5.8. *Assume that λ, M satisfy the CFL condition $2\lambda M \leq 1$ and u_0 satisfy the hypotheses $(IN_1), (IN_2)$. Let $u_h, \tilde{u}_h$ be approximate solutions as defined above. Then there exists a subsequence (still denoted by h) such that u_h converge almost everywhere to a weak solution u of* (0.1)*. In fact, $u_h \to u$ in $L^\infty_{\text{loc}}(\mathbb{R}_+, L^1_{\text{loc}}(\mathbb{R}))$ as h goes to* 0*. Similarly, along a further subsequence still denoted by h, $\tilde{u_h}$ converge almost everywhere to a weak solution $\tilde{u}$ of* (0.1) *and, similarly, $\tilde{u_h} \to \tilde{u}$ in $L^\infty_{\text{loc}}(\mathbb{R}_+, L^1_{\text{loc}}(\mathbb{R}))$ as h goes to* 0*. Furthermore, both u and $\tilde{u}$ satisfy the Kruzkhov entropy condition and hence are identical to the entropy solution of* (0.1).

Proof. This is the main convergence theorem for our Godunov and Enquist–Osher-type schemes (4.4), (4.5). The proof follows easily from the BV bounds for the singular mappings and invertibility of ψ_1 in $x < 0$ and ψ_2 in $x > 0$. Entropy consistency follows from the Crandall–Majda entropy fluxes. We refer the reader to [28], [29] for details. □

We reiterate that the key step of the proof of convergence is Lemma 5.4, which used the set of singular mappings in order to get the normalized variation bounds.

Remark 5.1. The existence and uniqueness result presented in this paper and the Godunov and Enquist–Osher schemes can be extended to cover more general equations of the form (1.1) under the assumptions that

1. $k \in BV(\mathbb{R})$ with finitely many points of discontinuities and k being C^1 outside the set of discontinuities;
2. $f(k(x_1), s) = f(k(x_2), s), \quad f(k(x_1), S) = f(k(x_2), S) \quad \forall x_1, x_2 \in \mathbb{R}$;
3. $u \mapsto f(k, u)$ is Lipschitz and has at most one extrema in $[s, S]$ $\forall k \in \mathbb{R}$.

The proof of this general result is presented in a more general setting (by allowing finitely many extrema for the fluxes) in a forthcoming paper [6].

6. Numerical experience. We have tested the Godunov scheme (4.4) and the Enquist–Osher scheme (4.5) extensively and have compared the results with staggered mesh algorithms of Enquist–Osher type developed in [28], [29], and [19]. Due to the special nature of the fluxes considered in this paper, all entropy theories for (0.1) agree and the schemes give the same result.

In most of the test cases, we have observed that (4.4) and 4.5 resolve the interface discontinuity better than the staggered mesh algorithms. Regarding waves in the interior, both schemes of this paper as well as the staggered mesh algorithms seem to do equally well. Only when there is a rarefaction wave in the interior do the schemes (4.4) and (4.5) perform slightly better. This can be illustrated by the following example.

Consider the flux functions $f(u) = 2u^2 - u$ and $g(u) = 3u - 3u^2$ with the Riemann initial data of 0.25 when $x < 0$ and 1 when $x > 0$. From the exact solutions of the Riemann problem presented in section 3, the exact solution is given by a discontinuity at the interface and a rarefaction wave traveling to the right. The numerical results obtained by (4.4), which we have denoted the exact Riemann solver (ERS), and a staggered Godunov scheme developed in [28], which we denote as TS, are shown in Figure 6.1.

As seen in the figure, ERS resolves the solution very well even at low mesh sizes. TS also does well at approximating the solution, except that there is nonphysical nonmonotone traveling wave at the tip of the rarefaction. From the right side of Figure 6.1, we can observe that the wave decreases in amplitude as $h \to 0$, but its presence is puzzling. A way out of this nonphysical wave is suggested in [19] by moving

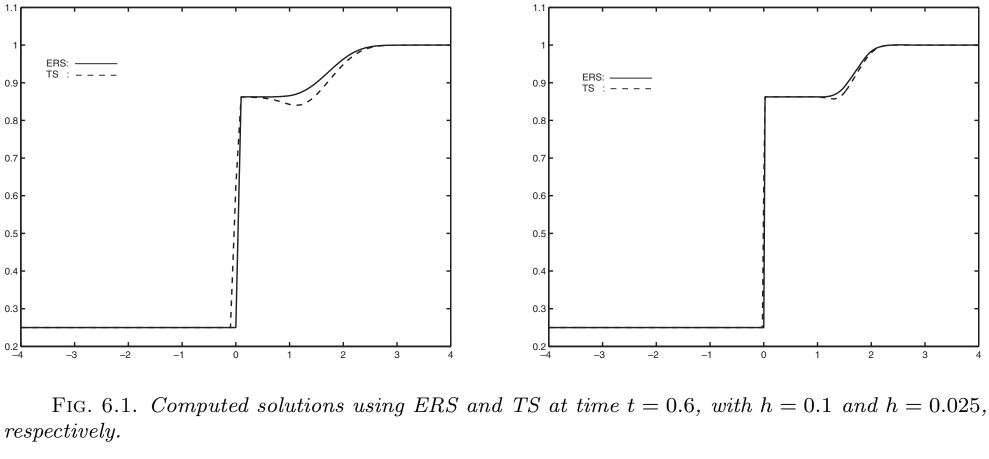


FIG. 6.1. *Computed solutions using ERS and TS at time* $t = 0.6$, *with* $h = 0.1$ *and* $h = 0.025$, *respectively.*

the interface discontinuity in the direction of the wave.

The main advantage of the staggered mesh algorithms as opposed to the 2 × 2 Riemann solvers is their simplicity. Although the scheme (4.4) is also based on exact solutions of the Riemann problem, it is as easy to implement as the staggered mesh schemes on account of the explicit formulae of the numerical fluxes. It also resolves the interface discontinuities better and avoids nonphysical waves and can be used as a substitute to the staggered mesh algorithms.

7. Conclusion.

Scalar conservation laws with discontinuous flux occur very frequently in physical and engineering applications, and hence we need to develop a well-posedness theory and stable numerical methods to approximate their solutions. An interesting case of such equations is when one of the fluxes is convex and the other concave, which can arise when the discontinuous coefficient is of indefinite sign. This case has received less attention in the literature on account of the difficulties in adapting the principal convergence tool of singular mapping.

In this paper, we have studied a scalar conservation law with discontinuous flux having the concave-convex flux geometry. The uniqueness of entropy solutions is shown by using a Kruzkhov-type argument. One strange feature of the flux geometry considered in this case is that we do not need to impose any additional entropy conditions at the interface. We give explicit solutions to the Riemann problem and use it to construct upwind finite difference schemes of Godunov and Enquist–Osher type. These schemes are shown to converge to the entropy solution by using a singular mapping technique. The novel feature of this paper is the use of a combination of singular mappings in order to show the convergence of our schemes to the entropy solutions. Numerical results are presented which show that our schemes resolve the solutions quite well. Comparison with the staggered mesh schemes indicate better performance of our schemes in some situations.

Acknowledgments. The author wishes to thank Professor Adimurthi and Professor G. D. Veerappa Gowda for their useful suggestions and discussions. The author also thanks the anonymous referees for their suggestions.

REFERENCES

[1] Adimurthi and G. D. V. Gowda, *Conservation laws with discontinuous flux*, J. Math. Kyoto Univ., 43 (2003), pp. 27–70.

[2] Adimurthi and G. D. V. Gowda, *Extensions of Godunov Scheme for Conservation Laws with Flux Function Discontinuous in the Space Variable*, preprint, 2003.

[3] Adimurthi, J. Jaffré, and G. D. V. Gowda, *Godunov-type methods for conservation laws with a flux function discontinuous in space*, SIAM J. Numer. Anal., 42 (2004), pp. 179–208.

[4] Adimurthi, S. Mishra, and G. D. V. Gowda, *Conservation laws with flux function discontinuous in the space variable.* I. *Optimal entropy solutions*, J. Hyperbolic Differ. Equ., submitted.

[5] Adimurthi, S. Mishra, and G. D. V. Gowda, *Conservation laws with flux function discontinuous in the space variable.* II. *Convex-concave type fluxes and generalised entropy solutions*, J. Comput. Appl. Math., submitted.

[6] Adimurthi, S. Mishra, and G. D. V. Gowda, *Convergence of Godunov Type Methods for Conservation Laws with a Spatially Varying Discontinuous Flux Function*, preprint.

[7] E. Audusse and B. Perthame, *Uniqueness for a Scalar Conservation Law with Discontinuous Flux via Adapted Entropies*, INRIA report-5261, 2004.

[8] P. Bagnerini and M. Rascle, *A multiclass homogenized hyperbolic model of traffic flow*, SIAM J. Math. Anal., 35 (2003), pp. 949–973.

[9] R. Burger, K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Well-posedness in* BV_t *and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units*, Numer. Math., 97 (2004), pp. 25–65.

[10] G. M. Coclite and N. H. Risebro, *Conservation Laws with Time Dependent Discontinuous Coefficients*, preprint.

[11] S. Diehl, *On scalar conservation laws with point source and discontinuous flux function*, SIAM J. Math. Anal., 26 (1995), pp. 1425–1451.

[12] S. Diehl, *A conservation law with point source and discontinuous flux function modelling continuous sedimentation*, SIAM J. Appl. Math., 56 (1996), pp. 388–419.

[13] B. Enquist and S. Osher, *One sided difference approximations for nonlinear conservation laws*, Math. Comp., 34 (1980), pp. 321–351.

[14] T. Gimse and N. H. Risebro, *Riemann problems with discontinuous flux function*, in Proceedings of the 3rd Annual International Conference on Hyperbolic Problems, Studentlitteratur, Uppsala, 1991, pp. 488–502.

[15] T. Gimse and N. H. Risebro, *Solution of the Cauchy problem for a conservation law with a discontinuous flux function*, SIAM J. Math. Anal., 23 (1992), pp. 635–648.

[16] E. Godlewski and P. A. Raviart, *Hyperbolic Systems of Conservation Laws*, Math. Appl. (Paris), Ellipses, Paris, 1991.

[17] K. H. Karlsen and J. D. Towers, *Convergence of the Lax–Friedrichs scheme and stability of conservation laws with a discontinuous space-time dependent flux*, Chinese Ann. Math. Ser B., 25 (2004), pp. 287–318.

[18] K. H. Karlsen, C. Klingenberg, and N. H. Risebro, *A relaxation scheme for conservation laws with a discontinuous coefficients*, Math. Comp., 73 (2004), pp. 1235–1259.

[19] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *Upwind difference approximations for degenerate parabolic convection-diffusion equations with a discontinuous coefficient*, IMA J. Numer. Anal., 22 (2002), pp. 623–664.

[20] K. H. Karlsen, N. H. Risebro, and J. D. Towers, *On a nonlinear degenerate parabolic transport-diffusion equation with a discontinuous coefficient*, Electron. J. Differential Equations, 2002 (93) (2002), 23 pp. (electronic).

[21] K. H. Karlsen, N. H. Risebro, and J. D. Towers, L^1 *stability for entropy solution of nonlinear degenerate parabolic convection-diffusion equations with discontinuous coefficients*, Skr. K. Nor. Vidensk. Selsk., no. 3, (2003), pp. 1–49.

[22] R. A. Klausen and N. H. Risebro, *Stability of conservation laws with discontinuous coefficients*, J. Differential Equations, 157 (1999), pp. 41–60.

[23] C. Klingenberg and N. H. Risebro, *Convex conservation laws with discontinuous coefficients, existence, uniqueness and asymptotic behaviour*, Comm. Partial Differential Equations, 20 (1995), pp. 1959–1990.

[24] C. Klingenberg and N. H. Risebro, *Stability of a resonant system of conservation laws modeling polymer flow with gravitation*, J. Differential Equations, 170 (2001), pp. 344–380.

[25] S. N. Kruzkhov, *First order quasilinear in several independent variables*, Math. USSR Sb., 10 (1970), pp. 217–243.

[26] S. Mishra, *Scalar Conservation Laws with Discontinuous Flux*, Master's thesis, Indian Institute of Science, Bangalore, India, 2003.

[27] B. Temple, *Global solution of the Cauchy problem for a class of* 2×2 *nonstrictly hyperbolic conservation laws*, Adv. in Appl. Math., 3 (1982), pp. 335–375.

[28] J. D. Towers, *Convergence of a difference scheme for conservation laws with a discontinuous flux*, SIAM J. Numer. Anal., 38 (2000), pp. 681–698.

[29] J. D. Towers, *A difference scheme for conservation laws with a discontinuous flux: The nonconvex case*, SIAM J. Numer. Anal., 39 (2001), pp. 1197–1218.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 578–603

ERROR ESTIMATE AND THE GEOMETRIC CORRECTOR FOR THE UPWIND FINITE VOLUME METHOD APPLIED TO THE LINEAR ADVECTION EQUATION[*]

DANIEL BOUCHE[†], JEAN-MICHEL GHIDAGLIA[‡], AND FRÉDÉRIC PASCAL[§]

Abstract. This paper deals with the upwind finite volume method applied to the linear advection equation on a bounded domain and with natural boundary conditions. We introduce what we call the geometric corrector, which is a sequence associated with every finite volume mesh in $\mathbf{R}^{nd}$ and every nonvanishing vector $\mathbf{a}$ of $\mathbf{R}^{nd}$. First we show that if the continuous solution is regular enough and if the norm of this corrector is bounded by the mesh size, then an order one error estimate for the finite volume scheme occurs. Afterwards we prove that this norm is indeed bounded by the mesh size in several cases, including the one where an arbitrary coarse conformal triangular mesh is uniformly refined in two dimensions. Computing numerically exactly this corrector allows us to state that this result might be extended under conditions to more general cases, such as the one with independent refined meshes.

Key words. finite volume method, consistency and accuracy, geometric corrector, unstructured meshes

AMS subject classifications. 65M06, 65M12, 65M15, 65M50

DOI. 10.1137/040605941

1. Introduction.

Finite volume methods (FVMs) were first used in the context of computational mechanics in situations in which solutions present discontinuities (see the monograph of Kröner [16], Godlewski and Raviart [13], and Eymard, Gallouët, and Herbin [10]). One reason for this is that they fundamentally rely on an integral version of the equations. This contrasts with finite difference methods (FDMs), for which smoothness of solutions is used in order to approximate derivatives by differential quotients. For finite element methods (FEMs), the situation is somewhat similar since trace theorems on hypersurfaces in H^1-type spaces exclude also the approximation of discontinuous solutions. It is the main reason why FVMs are widely used for the approximation of hyperbolic systems of conservation laws. Indeed, even when initial and boundary data are smooth, solutions of such equations produce in finite time discontinuous solutions. Let us also mention that other advantages of FVMs, like compact numerical stencil, effectiveness on unstructured grids, and simplicity in coding or data structures, are reasons for which these methods are used more and more often for elliptic and parabolic equations, albeit smoothness of the solution is guaranteed.

Concerning FVMs for hyperbolic equations, there are a lot of methods, and especially in computational fluid dynamics; historically these methods were natural

[*] Received by the editors March 30, 2004; accepted for publication (in revised form) October 13, 2004; published electronically June 30, 2005.
http://www.siam.org/journals/sinum/43-2/60594.html

[†] Commissariat à l'Énergie Atomique, DPTA, BP 12, F-91680 Bruyères-le-Châtel, France (daniel.bouche@cea.fr) and CMLA, ENS Cachan et CNRS UMR 8536, 61 avenue du Président Wilson, F-94235 Cachan Cedex, France.

[‡] Centre de Mathématiques et de Leurs Applications, ENS Cachan et CNRS UMR 8536, 61 avenue du Président Wilson, F-94235 Cachan Cedex, France (jmg@cmla.ens-cachan.fr).

[§] Laboratoire de Mathématiques, UMR 8628, CNRS et Université Paris-Sud, Bâtiment 425, Université Paris-Sud, F-91405 Orsay, France (Frederic.Pascal@math.u-psud.fr) and CMLA, ENS Cachan et CNRS UMR 8536, 61 avenue du Président Wilson, F-94235 Cachan Cedex, France.

extensions of the famous Godunov method for the Euler equations of gas dynamics to multidimensional problems. The Godunov method was known to be too diffusive, and this was attributed to the fact that it is a first order method. In the beginning of the eighties, Bram Van Leer proposed MUSCL methods, which can be seen as corrections that lead to second order schemes.

However, it is an open problem to determine the optimal rate of convergence of FVMs. This contrasts strongly with FDMs and FEMs, where such rates are known. Hence there is an apparent paradox between the fact that some methods are termed "first order" or even "second order" and the fact that corresponding error estimates are not proved. Indeed, there is some confusion in the literature: what is usually called a "first order" method, for example, corresponds to methods with a first order truncation error on *uniform Cartesian* grids. In practice—and this is the main interest of FVMs—these methods are used on unstructured meshes, and therefore it is not known whether "first order" methods lead to first order error estimates. In fact the situation is even more puzzling since "first order" methods on unstructured grids *do not lead in general to a truncation error that goes to zero* as the mesh size goes to zero (see section 3.3).

Even for the scalar linear advection equation, obtaining a priori optimal error estimates is still a challenging task. One of the main difficulties, as said already, lies in the fact that the nonuniformity of the mesh brings up an apparent loss of consistency, as has been observed by Hoffman [14], Turkel [27], and Pike [24]. In fact this loss of consistency is an artifact of standard convergence proof: for instance, the Lax theorem states that stability and consistency are sufficient conditions for a scheme to be convergent at the same rate that the truncation error converges to zero. Actually, consistency is not necessary; the scheme maintains the accuracy and the global error behaves better than the local error would indicate. This property of enhancement of the truncation error is called supraconvergence, and for second and higher order boundary value problems, this phenomenon, discovered by Tikhonov and Samarskij [26], was widely analyzed in various cases by Manteuffel and his coauthors in [19], [20], [15], [21] and in Garcia-Archilla and Sanz-Serna [11].

In these papers, the analysis relies on the fact that the truncation error (defined by applying the discrete operator to the exact solution) can be rewritten in the special form $L_h\tau_1 + \tau_2$, where τ_1 and τ_2 are of the optimal order $\mathcal{O}(h^p)$, L_h is the discrete operator, and $L_h\tau_1$ is only $\mathcal{O}(h^{p-1})$. Then an optimal discrete energy estimate for the global error can be derived. This idea is extended to finite difference schemes for hyperbolic systems with nonuniform one-dimensional grids and with irregular Cartesian multidimensional grids in Levermore, Manteuffel, and White [18] and in Monk and Süli [22].

This way to rewrite the truncation error can be seen as a correction of the error in order to cancel the leading part of the local error due to the nonuniformity of the mesh. Actually, in Wendroff and White [31], [30] and Wendroff [29], a grid function is introduced for suitably smooth solutions of hyperbolic systems in order to prove the optimal order of convergence of upwind finite difference and Lax–Wendroff schemes in one space dimension, and in two dimensions when an alternate direction method is used. Berger et al. [3], [2] successfully used the idea to get the optimal rate of convergence of the "h-box" scheme defined on a nonuniform Cartesian grid with embedded irregular small cells.

Recently Després [9], [7], [8], by using energy estimates, gave an elegant proof of an order $h^{1/2}$ error estimate with respect to the L^2 norm for the linear advection equation on regular finite element meshes in the particular case of periodic boundary conditions.

Finally, let us mention that when no regularity conditions are imposed on the exact solution, by suitable application of Kuznetsov [17] approximation theory, for instance, it is possible to establish rate of convergence of $h^{1/2}$ in the L^1 norm for nonregular Cartesian grids: see Eymard, Gallouët, and Herbin [10], Cockburn and Gremaud [4], [5], Cockburn, Gremaud, and Yang [6], Vila and Villedieu [28], and Teng and Zhang [25]. All these results deal with data (and solutions) which belong to the set of functions with bounded variations. As shown in the aforementioned references, the $h^{1/2}$ is optimal for these solutions. However, these papers raise the question of whether such a rate is due to the irregularity of the mesh. Our present work answers this open question by showing that this is not the case: on irregular meshes, provided the solution is smooth, the error estimate behaves like h. Hence the poor convergence behavior is due to the lack of smoothness of solutions.

In this paper we study the initial and boundary value problem for the linear advection equation posed on a polygonal domain of $\mathbf{R}^{nd}$ (nd is the space dimension of the problem under consideration). We construct first what we call "geometric correctors." They form a family of vectors in $\mathbf{R}^{nd}$, $\Gamma = (\Gamma_j)_{j=1,\dots,N}$, where N is the number of control volumes. This set of geometric correctors depends only on the mesh and on the advection vector but not on the solutions to the advection equation. Our first result in section 3 shows that if the solution is regular and if this family of vectors is uniformly bounded by the mesh size, i.e., $||\Gamma|| \le Ch$, then under a Courant–Friedrichs–Lewy (CFL) condition, the classical explicit first order upwind scheme for the advection equation is indeed first order: $||u_h - u|| \le C'h$, where u is the exact solution and u_h the FVM approximation. Since we are able to compute exactly Γ, numerical simulations allow us to study in which cases the hypothesis $||\Gamma|| \le Ch$ is satisfied. We prove in section 4 this estimation in several cases, including the one where an arbitrary coarse conformal triangular mesh is refined. Finally in section 5, we present numerical experiments that lead us to conjecture that this result holds true in the case of independent refined meshes if the advection vector is not parallel to a side of the polygonal domain. On the other hand, on the basis of numerical evidence, we conjecture that if the advection vector is parallel to a side of the polygonal domain, then the best estimate should be in $h^{1/2}$.

2. The continuous problem. We consider Ω a polygonal domain in $\mathbf{R}^{nd}$ with $nd \ge 1$, and we denote by $\mathbf{n}$ the unitary external normal vector on $\partial\Omega$. Let $\mathbf{a}$ be a nonzero vector and let us denote $\partial\Omega^- = \{x \in \partial\Omega, \mathbf{a}\cdot\mathbf{n}(x) < 0\}$. Given a function φ defined on Ω and a function ψ defined on $\partial\Omega^- \times [0,+\infty[$, the initial and boundary value problem for the advection equation on $\bar\Omega \times [0,+\infty[$ reads

$$\frac{\partial u}{\partial t} + (\mathbf{a}\cdot\nabla)u = 0\,, \quad (x,t) \in \Omega \times]0,+\infty[\,, \tag{2.1}$$

$$u(x,0) = \varphi(x)\,, \quad x \in \Omega\,, \tag{2.2}$$

$$u(x,t) = \psi(x,t)\,, \quad (x,t) \in \partial\Omega^- \times [0,+\infty[\,. \tag{2.3}$$

As is well known and understood, this problem has a unique smooth solution, provided the data φ and ψ are smooth and satisfy the so-called compatibility conditions (for example, the first compatibility condition is given below in (2.6)). In fact there are several methods for obtaining this solution. The first one, which uses functional analysis, is due to Bardos [1]. It consists of considering smooth solutions to the parabolic equation ($\nu > 0$)

$$\frac{\partial u^\nu}{\partial t} + (\mathbf{a} \cdot \nabla) u^\nu = \nu \Delta u^\nu\,, \qquad (x,t) \in \Omega \times \,]0, +\infty[\,, \tag{2.4}$$

with the complete Dirichlet boundary conditions

$$u^\nu(x,t) = \psi^\nu(x,t)\,, \qquad (x,t) \in \partial\Omega \times [0, +\infty[, \tag{2.5}$$

where ψ^ν is a smooth extension of ψ to $\Omega \times [0, +\infty[$. Then the solution to (2.1)–(2.3) is obtained at the limit $\nu = 0$. This produces a weak solution, whose regularity is obtained using the compatibility conditions. This method has the great advantage of applying to general (i.e., with variable nonnecessarily smooth coefficients) first order linear hyperbolic equations. However, it is not constructive in contrast with the so-called method of characteristics. In the case of the linear advection equation (2.1), this last method consists of considering the backward characteristics, a straight line here, defined by $(x - s\mathbf{a}, t - s)_{s \geq 0}$ and starting from an arbitrary point $(x,t) \in \Omega \times [0, +\infty[$, and then in looking for its intersection with the boundary of this cylinder. One sees easily that this point belongs either to the set $\Omega \times \{0\}$ or to the set $\partial\Omega^- \times [0, +\infty[$, and since (2.1) simply means that u is constant along these characteristics, one finds $u(x,t)$ by using either (2.2) or (2.3). Let us be more precise. Given $x \in \Omega$, we denote by $s(x)$ the first positive real number τ such that $x - \tau\mathbf{a}$ meets the boundary of Ω (one sees easily that necessarily $x - s(x)\mathbf{a} \in \partial\Omega^-$). Then there are three cases:

(i) if $0 < s(x) < t$, then $u(x,t) = \psi(x - s(x)\mathbf{a}, t - s(x))$;

(ii) if $s(x) > t$, then $u(x,t) = \varphi(x - t\mathbf{a})$;

(iii) if $s(x) = t$, then $u(x,t) = \psi(x - t\mathbf{a}, 0) = \varphi(x - t\mathbf{a})$.

The last case sheds light on the first compatibility condition between ψ and φ:

$$\psi(x, 0) = \varphi(x) \quad \forall x \in \partial\Omega^-\,. \tag{2.6}$$

3. A cell-centered finite volume discretization and associated error estimates.

3.1. Notation and geometric properties of meshes. Let $\mathcal{T} = \{K_j \,:\, j = 1, \ldots, N\}$ be a partition of the domain Ω in polyhedral volumes K_j (the control volumes) that forms a structured or unstructured triangulation of Ω and such that the hyperface between two adjacent volumes is included in a hyperplane. For a given j between 1 and N, there are two cases:

- In the first one, the volume K_j has no hyperface on the boundary $\partial\Omega$. Then we denote by $\mathcal{N}(j)$ the set of indices $k \neq j$ between 1 and N such that $K_k \cap K_j$ has $(nd-1)$ positive measure.
- In the second one, the boundary of the volume K_j meets the boundary $\partial\Omega$ in a set of $(nd-1)$ positive measure. Then we denote by $\mathcal{N}_0(j)$ the set of indices $k \neq j$ between 1 and N such that $K_k \cap K_j$ has $(nd-1)$ positive measure, and we complete this set into $\mathcal{N}(j)$ by negative integers numbering the hyperfaces of K_j which are on the boundary $\partial\Omega$. We denote the set of these negative integers by $\mathcal{N}_b(j)$.

In both cases we have $\mathcal{N}(j) = \mathcal{N}_0(j) \cup \mathcal{N}_b(j)$, since $\mathcal{N}_b(j)$ is empty when K_j has no hyperface on the boundary $\partial\Omega$.

Let $k \in \mathcal{N}(j)$. If $k \in \mathcal{N}_0(j)$, we denote by $\mathbf{n}_{j,k}$ the unit normal on $K_j \cap K_k$, which points out from K_j and by $\mathbf{N}_{j,k}$ the product $\mathbf{N}_{j,k} = |K_j \cap K_k| \mathbf{n}_{j,k}$, where $|K_j \cap K_k|$ denotes the $(nd-1)$ positive measure of $K_j \cap K_k$. If $k \in \mathcal{N}_b(j)$, we denote by K_k the symmetric of K_j with respect to the hyperface $K_j \cap \partial\Omega$ and keep the same notation as above. We shall use in what follows the partition of $\mathcal{N}(j)$:

$$\mathcal{N}(j) = \mathcal{N}^+(j) \cup \mathcal{N}^-(j) \cup \mathcal{N}^0(j)\,, \tag{3.1}$$

where $\mathcal{N}^+(j) = \{k \in \mathcal{N}(j), \mathbf{a}\cdot\mathbf{n}_{j,k} > 0\}$, $\mathcal{N}^-(j) = \{k \in \mathcal{N}(j), \mathbf{a}\cdot\mathbf{n}_{j,k} < 0\}$, and $\mathcal{N}^0(j) = \{k \in \mathcal{N}(j), \mathbf{a}\cdot\mathbf{n}_{j,k} = 0\}$. Similar definitions are extended to $\mathcal{N}_0^\varepsilon(j)$ and $\mathcal{N}_b^\varepsilon(j)$, $\varepsilon \in \{0,+,-\}$. Last, the centroid of K_j will be denoted by g_j, while the one of $K_j \cap K_k$ will be denoted by $g_{j,k}$.

Since we are interested in convergence results, we are going to consider families of triangulation $\mathcal{T}^h$ indexed by the real number $h = \max_{K_j\in\mathcal{T}^h} h_j$, where h_j is the diameter of the volume K_j. By definition of the parameter h, we have $|K_j| \le h^{nd}$ and $|K_j \cap K_k| \le h^{nd-1}$ for all $K_j, K_k \in \mathcal{T}^h$, and we assume that there exist $h_0 > 0$ and positive constants κ_1 and κ_2 such that for every $h < h_0$ we have

$$\frac{h_j^{nd}}{|K_j|} \le \kappa_1 \quad \text{and} \quad \sharp\mathcal{N}(j) \le \kappa_2 \quad \forall K_j \in \mathcal{T}^h\,. \tag{3.2}$$

Remark 1. The first assumption is equivalent to the shape-regularity assumption, and the second one means that the number of neighbors of each volume remains bounded when h tends to zero.

Let us recall the following properties, applications of the divergence theorem.

Proposition 3.1. *With the previous notation, we have the two vector identities*

$$\sum_{k\in\mathcal{N}(j)} \mathbf{N}_{j,k} = 0 \quad \text{and} \quad \sum_{k\in\mathcal{N}(j)} \mathbf{X}\cdot\mathbf{N}_{j,k}\; g_{j,k} = \mathbf{X}\;|K_j| \quad \forall \mathbf{X} \in \mathbf{R}^{nd}\,. \tag{3.3}$$

A straightforward application of the previous proposition leads to the following.

Proposition 3.2. *For every nonvanishing vector* $\mathbf{a} \neq 0$*, we have*

$$\mathcal{N}^+(j) \neq \emptyset \quad \text{and} \quad \sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k} \neq 0\,. \tag{3.4}$$

Concerning the time discretization of (2.1), we use a *finite difference* approach, and therefore we consider an increasing sequence $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} \longrightarrow \infty$ and set $\Delta t_n = t_{n+1} - t_n$.

We shall consider sequences $\xi = (\xi_j)_{j=1,\dots,N}$ of scalars and of vectors in $\mathbf{R}^{nd}$, and we will estimate the ℓ^p norm for $p \in [1,+\infty]$ with

$$||\xi||_\infty = \max_{1\le j\le N} |\xi_j| \quad \text{and} \quad ||\xi||_p = \left(\sum_{j=1}^{N} |K_j||\xi_j|^p\right)^{1/p} \quad \text{for } p \ge 1,$$

where $|\xi_j|$ is the ℓ^p norm in $\mathbf{R}^{nd}$ in the case of vectors.

3.2. The first order explicit upwind finite volume scheme. This scheme reads as follows:

$$\frac{u_j^{n+1} - u_j^n}{\Delta t_n} + \frac{1}{|K_j|}\left(\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k} u_j^n + \sum_{k\in\mathcal{N}^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k} u_k^n\right) = 0\,. \tag{3.5}$$

Indeed the underlying philosophy of the finite volume scheme is to approximate on each control volume K_j for $j \in 1, \ldots, N$ the mean value of the exact solution to the continuous equation (2.1)

$$U_j^n = \frac{1}{|K_j|} \int_{K_j} u(x, t_n) dx \tag{3.6}$$

by taking into account the direction from which the information comes. System (3.5) allows us to compute an approximation of U_j^n once the initial values, u_j^0 for $j \geq 1$, and the boundary ones, u_k^n for $k \in \cup_j \mathcal{N}_b^-(j)$ and for $n \geq 0$, have been provided.

The classical way to check if the goal is achieved is to evaluate the truncation error, which consists of replacing u_j^n by U_j^n in (3.5). Here we are going to use a truncation error based on the value $u(g_j, t_n)$ of the exact solution at the centroid of the control volumes. More precisely one computes

$$\begin{aligned} E_j^n = \frac{u(g_j, t_{n+1}) - u(g_j, t_n)}{\Delta t_n} + \frac{1}{|K_j|} \Bigg(& \sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} u(g_j, t_n) \\ & + \sum_{k \in \mathcal{N}_0^-(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} u(g_k, t_n) + \sum_{k \in \mathcal{N}_b^-(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} u_k^n \Bigg) . \end{aligned} \tag{3.7}$$

Then from (3.5) and (3.7), the error denoted by $\alpha_j^n = u_j^n - u(g_j, t_n)$ satisfies

$$\frac{\alpha_j^{n+1} - \alpha_j^n}{\Delta t_n} + \frac{1}{|K_j|} \left(\sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} \alpha_j^n + \sum_{k \in \mathcal{N}_0^-(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} \alpha_k^n \right) = -E_j^n , \tag{3.8}$$

which can also be written as

$$\alpha_j^{n+1} = (\mathcal{L}^n \alpha^n)_j - \Delta t_n E_j^n, \tag{3.9}$$

where we denote by $\mathcal{L}^n$ the following operator that acts on sequences $\xi = (\xi_j)_{j=1,\ldots,N}$:

$$(\mathcal{L}^n \xi)_j = \xi_j - \frac{\Delta t_n}{|K_j|} \left(\sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} \xi_j + \sum_{k \in \mathcal{N}_0^-(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} \xi_k \right) . \tag{3.10}$$

In classical finite difference theory, one transfers information on the smallness of the truncation error E_j^n to the error α_j^n via a stability property of the scheme, which amounts here to showing that the norm of the operator $\mathcal{L}^n$ is not greater than 1. Courant, Friedrichs, and Lewy, in their early study of the discretization of the one-dimensional advection equation by finite differences, introduced the CFL number as a limitation on the time step Δt_n in order to achieve stability. In the case of the scheme (3.5), their construction can be mimicked as follows. First we take on each volume the local time given by (observe that, thanks to Proposition 3.2, $\sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} \neq 0$)

$$\tau_j = \frac{|K_j|}{\sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k}} . \tag{3.11}$$

Then we set $\Delta t_* = \min_j \tau_j$ and assign to the CFL number the ratio

$$\lambda = \frac{\Delta t_n}{\Delta t_*} . \tag{3.12}$$

When $\lambda \leq 1$, the operator $\mathcal{L}^n$ has a norm not greater than 1. More precisely, we have the following result, whose proof is standard.

THEOREM 3.3. *Under the CFL condition $\lambda \leq 1$, for every $p \in [1, +\infty]$ the operator $\mathcal{L}^n$ satisfies*

$$||\mathcal{L}^n \xi||_p \leq ||\xi||_p \,. \tag{3.13}$$

This result, when combined with (3.9), has the following straightforward corollary.

COROLLARY 3.4. *Under the CFL condition $\lambda \leq 1$ and for every $p \in [1, +\infty]$ we have the estimate*

$$||\alpha^n||_p \leq ||\alpha^0||_p + \sum_{i=0}^{n-1} \Delta t_i ||E^i||_p \,. \tag{3.14}$$

This inequality shows that estimations on E^i transfer to ones on the error α^n.

3.3. On the truncation error. Let us consider the volume of control K_j. We see that the error E_j^n can be divided into three sums:

$$E_j^n = G_j^n + I_j^n + \frac{1}{|K_j|} \sum_{k \in \mathcal{N}_b^-(j)} \mathbf{a} \cdot \mathbf{N}_{j,k}(u_k^n - u(g_{j,k}, t_n)), \tag{3.15}$$

where

$$G_j^n = \frac{u(g_j, t_{n+1}) - u(g_j, t_n)}{\Delta t_n} + \frac{1}{|K_j|} \sum_{k \in \mathcal{N}(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} u(g_{j,k}, t_n) \tag{3.16}$$

and

$$\begin{aligned} |K_j| I_j^n = & \sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k}(u(g_j, t_n) - u(g_{j,k}, t_n)) \\ & + \sum_{k \in \mathcal{N}_0^-(j)} \mathbf{a} \cdot \mathbf{N}_{j,k}(u(g_k, t_n) - u(g_{j,k}, t_n)). \end{aligned} \tag{3.17}$$

Recalling the estimates (3.2) on the volumes, the third term in (3.15) that concerns the boundary conditions treatment behaves like $\mathcal{O}(h)$ if we assume, for instance, that

$$u_k^n = \frac{1}{\Delta t_n |K_j \cap K_k|} \int_{t_n}^{t_{n+1}} \int_{K_j \cap K_k} \psi(\sigma, t) d\sigma dt \,, \quad k \in \mathcal{N}_b^-(j) \,, \quad n \geq 0, \tag{3.18}$$

i.e., that the discrete numerical boundary treatment satisfies

$$|u_k^n - u(g_{j,k}, t_n)| = \mathcal{O}(h^2) \,, \quad k \in \mathcal{N}_b^-(j) \,, \quad n \geq 0 \,. \tag{3.19}$$

Using the relationship (3.3) on the normals, the assumptions (3.2) on the mesh, and, intensively, Taylor's expansions, we see that the second term in G_j^n can be written as

$$\begin{aligned} \sum_{k \in \mathcal{N}(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} u(g_{j,k}, t_n) &= \sum_{k \in \mathcal{N}(j)} \mathbf{a} \cdot \mathbf{N}_{j,k}(u(g_{j,k}, t_n) - u(g_j, t_n)) \,, \\ &= \sum_{k \in \mathcal{N}(j)} \mathbf{a} \cdot \mathbf{N}_{j,k}(g_{j,k} - g_j) \cdot \nabla u(g_j, t_n) + \mathcal{O}(h^{nd+1}) \,, \\ &= \left(\sum_{k \in \mathcal{N}(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} g_{j,k} \right) \cdot \nabla u(g_j, t_n) + \mathcal{O}(h^{nd+1}) \,, \\ &= |K_j| \mathbf{a} \cdot \nabla u(g_j, t_n) + \mathcal{O}(h^{nd+1}) \,. \end{aligned} \tag{3.20}$$

Finally, since

$$\frac{u(g_j,t_{n+1})-u(g_j,t_n)}{\Delta t_n}=\frac{\partial u}{\partial t}(g_j,t_n)+\mathcal{O}(\Delta t_n) \tag{3.21}$$

and since u is the solution of (2.1), we deduce that if $\Delta t_n \le ch$,

$$G_j^n=\mathcal{O}(h)\,. \tag{3.22}$$

But concerning I_j^n, a similar computation leads this time to $I_j^n=\mathcal{O}(1)$, and we find that $E_j^n=\mathcal{O}(1)$. Hence E_j^n does not even converge to zero as h goes to zero, so in a sense the scheme is inconsistent. One might think that the estimate above is not optimal, but in fact this is not the case since in the one-dimensional case the previous computations are much simpler and show that indeed E_j^n does not converge to zero as h goes to zero. This fact (the scheme is not pointwise-consistent) was already observed in two dimensions by Kröner [16, Lemma 3.2.8]. On the other hand, Lemma 4.1 of Chapter IV in Godlewski and Raviart [13] cannot be applied to the upwind scheme on a uniform triangulation by equilateral triangles, as they wrongly claim. In their proof, they use the identity $\frac{\partial\Phi}{\partial u}(u,u,\mathbf{n})+\frac{\partial\Phi}{\partial u}(u,u,-\mathbf{n})=0$, where $\Phi(u,v,\mathbf{n})$ denotes the numerical flux; that is, $\Phi(u,v,\mathbf{n})=(\mathbf{a}\cdot\mathbf{n})u$ for $\mathbf{a}\cdot\mathbf{n}>0$, $\Phi(u,v,\mathbf{n})=(\mathbf{a}\cdot\mathbf{n})v$ for $\mathbf{a}\cdot\mathbf{n}<0$ and $\Phi(u,v,\mathbf{n})=0$ for $\mathbf{a}\cdot\mathbf{n}=0$. And one can readily see that this identity is wrong. However, the result (that on such a triangulation the upwind scheme is first order accurate) is true, as it will follow from the proof of Theorem 4.2 in this paper.

3.4. A geometric corrector. Our goal is to construct a sequence of scalars $(\gamma_j^n)_{j=1,\dots,N}$ that satisfies the N equations

$$\sum_{k\in\mathcal{N}^+(j)}\mathbf{a}\cdot\mathbf{N}_{j,k}\gamma_j^n+\sum_{k\in\mathcal{N}_0^-(j)}\mathbf{a}\cdot\mathbf{N}_{j,k}\gamma_k^n=|K_j|I_j^n \tag{3.23}$$

in order to correct the errors $(\alpha_j^n)_{j=1,\dots,N}$ introduced in section 3.2 and to prove that under smoothness assumptions the finite volume scheme is first order accurate in spite of the fact that the truncation error is inconsistent. We first slightly modify the linear system (3.23) to be able to estimate in terms of h the norm of the sequence. Thus, we will consider the construction of a sequence of vectors in $\mathbf{R}^{nd}$, $(\Gamma_j)_{j=1,\dots,N}$ that satisfies the N following vector equations (we recall that g_j and $g_{j,k}$ belong to $\mathbf{R}^{nd}$):

$$\begin{aligned}&\sum_{k\in\mathcal{N}^+(j)}\mathbf{a}\cdot\mathbf{N}_{j,k}\Gamma_j+\sum_{k\in\mathcal{N}_0^-(j)}\mathbf{a}\cdot\mathbf{N}_{j,k}\Gamma_k\\&\qquad=\sum_{k\in\mathcal{N}^+(j)}\mathbf{a}\cdot\mathbf{N}_{j,k}(g_{j,k}-g_j)+\sum_{k\in\mathcal{N}_0^-(j)}\mathbf{a}\cdot\mathbf{N}_{j,k}(g_{j,k}-g_k)\,.\end{aligned} \tag{3.24}$$

We are going to prove, thanks to Theorem 3.8, that this system has one and only one solution, but let us first give the following definitions.

Definition 3.5. *Given a nonzero vector* $\mathbf{a}\in\mathbf{R}^{nd}$ *and a triangulation* $(K_j)_{j=1,\dots,N}$ *of a polygonal domain* Ω *in* $\mathbf{R}^{nd}$ *defined as in section* 3.1, *the geometric corrector for the advection equation* (2.1) *is the sequence of vectors in* $\mathbf{R}^{nd}$, $(\Gamma_j)_{j=1,\dots,N}$, *defined by the* $N\times N$ *system of equations* (3.24).

Definition 3.6. *For every control volume* $J\in\mathcal{T}$, *we denote by* $\mathcal{C}(J)$ *the cone of dependence of* J*:*

(3.25)
$$\mathcal{C}(J)=\{K\in\mathcal{T}\,/\,\exists\,J_1,\dots,J_p\in\mathcal{T}\,,\ \mathbf{a}\cdot\mathbf{N}_{J_1,K}<0,\ \mathbf{a}\cdot\mathbf{N}_{J_2,J_1}<0,\ \dots,\ \mathbf{a}\cdot\mathbf{N}_{J,J_p}<0\}\,.$$

Let us remark that

$$\forall K \in \mathcal{C}(J)\,,\ \big(j \in \mathcal{N}_0^-(K) \Rightarrow K_j \in \mathcal{C}(J)\big)\,. \tag{3.26}$$

PROPOSITION 3.7. *For every control volume $J \in \mathcal{T}$, there is at least one volume $K_k \in \mathcal{C}(J)$ that shares a face with $\partial\Omega^-$, i.e., such that*

$$\mathcal{N}_b^-(k) \neq \emptyset\,. \tag{3.27}$$

Proof. Indeed let us assume that there is no such volume in $\mathcal{C}(J)$. Since $\mathcal{C}(J)$ is composed of polyhedron, it is also a polyhedron, and there is at least one face (on the boundary of $\mathcal{C}(J)$) for which the external normal forms an obtuse angle with $\mathbf{a}$. Then for each volume $K \in \mathcal{C}(J)$ that meets this face, since from the assumption that K does not intersect $\partial\Omega^-$, there is an adjacent volume $L \in \mathcal{T} \setminus \mathcal{C}(J)$ such that $\mathbf{a} \cdot \mathbf{N}_{K,L} < 0$. From the definition of $\mathcal{C}(J)$, this last relation yields that L is in $\mathcal{C}(J)$, which is a contradiction. □

According to Proposition 3.2, $\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}$ does not vanish. We then define the $N \times N$ matrix B such that for an arbitrary sequence $X = (X_j)_{j=1,\dots,N}$ in $\mathbf{C}^N$ we have

$$(BX)_j = \frac{\sum_{k\in\mathcal{N}_0^-(j)}(-\mathbf{a}\cdot\mathbf{N}_{j,k})X_k}{\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}}\,. \tag{3.28}$$

We denote by $(\Delta_j)_{j=1,\dots,N}$ the sequence of vectors in $\mathbf{R}^{nd}$ defined by

$$\Delta_j = \frac{\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}(g_{j,k}-g_j) + \sum_{k\in\mathcal{N}_0^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}(g_{j,k}-g_k)}{\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}}\,. \tag{3.29}$$

It is clear that (3.24) can be written as

$$(Id - B)\Gamma = \Delta\,. \tag{3.30}$$

We are now able to prove that $Id - B$ is a nonsingular M-matrix.

THEOREM 3.8. *The spectrum $\sigma(B)$ of B satisfies*

$$\sigma(B) \subset \{z \in \mathbf{C}\,,\ |z| < 1\}\,, \tag{3.31}$$

i.e.,

$$(Id - B)^{-1} = \sum_{l=0}^{\infty} B^l\,. \tag{3.32}$$

Proof. (i) Let $X \neq 0$, $X \in \mathbf{C}^N$ such that $BX = \lambda X$. From definition (3.28), by observing that $-\mathbf{a}\cdot\mathbf{N}_{j,k} \geq 0$ for $k \in \mathcal{N}_0^-(j)$ and by using the relation $\sum_{k\in\mathcal{N}(j)} \mathbf{a}\cdot\mathbf{N}_{j,k} = 0$ rewritten as $\sum_{k\in\mathcal{N}_0^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k} + \sum_{k\in\mathcal{N}_b^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k} + \sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k} = 0$, we have

$$|(BX)_j| \leq \frac{\sum_{k\in\mathcal{N}_0^-(j)} -\mathbf{a}\cdot\mathbf{N}_{j,k}}{\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}} ||X||_\infty = \left(1 + \frac{\sum_{k\in\mathcal{N}_b^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}}{\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}}\right) ||X||_\infty. \tag{3.33}$$

Then $||BX||_\infty \leq ||X||_\infty$ and $|\lambda| \leq 1$.

(ii) Now let us assume that $\lambda \in \mathbf{C}$ with $|\lambda| = 1$ is an eigenvalue of B, i.e., that there is $X \neq 0$ such that $BX = \lambda X$ and $X \in \mathbf{C}^N$. Let us denote by $K_j \in \mathcal{T}$ a volume such that

$$|X_j| = \max_{1 \leq k \leq N} |X_k| \,. \tag{3.34}$$

From the jth component of $BX = \lambda X$, and since $\sum_{k \in \mathcal{N}(j)} \mathbf{N}_{j,k} = 0$, we get

$$\sum_{k \in \mathcal{N}_0^-(j)} (-\mathbf{a} \cdot \mathbf{N}_{j,k}) \left(|X_j| - \frac{X_k}{\lambda} \right) + \sum_{k \in \mathcal{N}_b^-(j)} (-\mathbf{a} \cdot \mathbf{N}_{j,k}) |X_j| = 0. \tag{3.35}$$

Thus

$$\sum_{k \in \mathcal{N}_0^-(j)} (-\mathbf{a} \cdot \mathbf{N}_{j,k}) \,\mathcal{R}e \left(|X_j| - \frac{X_k}{\lambda} \right) + \sum_{k \in \mathcal{N}_b^-(j)} (-\mathbf{a} \cdot \mathbf{N}_{j,k}) |X_j| = 0 \,, \tag{3.36}$$

$$\sum_{k \in \mathcal{N}_0^-(j)} (-\mathbf{a} \cdot \mathbf{N}_{j,k}) \,\mathcal{I}m \left(|X_j| - \frac{X_k}{\lambda} \right) = 0 \,. \tag{3.37}$$

Since $\mathcal{R}e(|X_j| - X_k/\lambda) \geq 0$ and $\mathcal{I}m(|X_j| - X_k/\lambda) \geq 0$, all the terms above are positive and thus equal to zero, so that $\mathcal{N}_b^-(j) = \emptyset$ and for all $k \in \mathcal{N}_0^-(j)$, $|X_j| = X_k/\lambda$, i.e., $|X_j| = |X_k|$. By induction, we find that for all k such that K_k belongs to $\mathcal{C}(K_j)$, $|X_k| = |X_j|$ and $\mathcal{N}_b^-(k) = \emptyset$, i.e., K_k does not meet $\partial\Omega^-$, which is in contradiction with Proposition 3.7. □

Remark 2. Let us denote

$$\delta_j = \sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} = \sum_{k \in \mathcal{N}^-(j)} |\mathbf{a} \cdot \mathbf{N}_{j,k}| > 0 \,, \tag{3.38}$$

so that (3.28) also reads as

$$(BX)_j = \frac{\sum_{k \in \mathcal{N}_0^-(j)} |\mathbf{a} \cdot \mathbf{N}_{j,k}| X_k}{\delta_j} \,. \tag{3.39}$$

We then observe that $(BX)_j$ is a barycenter of the X_k for $k \in \mathcal{N}^-(j)$ when K_j does not meet the boundary $\partial\Omega^-$, and of the X_k for $k \in \mathcal{N}_0^-(j)$ and 0 for $k \in \mathcal{N}_b^-(j)$ when K_j meets the boundary $\partial\Omega^-$. Now, given $m \geq 2$, we have

$$(B^m X)_j = \frac{\sum_{j_1 \in \mathcal{N}_0^-(j)} \cdots \sum_{j_m \in \mathcal{N}_0^-(j_{m-1})} |\mathbf{a} \cdot \mathbf{N}_{j,j_1}| \ldots |\mathbf{a} \cdot \mathbf{N}_{j_{m-1},j_m}| X_{j_m}}{\delta_j \delta_{j_1} \ldots \delta_{j_{m-1}}} \,. \tag{3.40}$$

Remark 3. The great advantage of the geometric corrector Γ_k defined by (3.24) on the corrector γ_j^n defined by (3.23) is that the former depends only on the advection vector $\mathbf{a}$ and on the geometry of the problem Ω and the elements K_j of the mesh and not on the initial and boundary data. On the other hand it might be that the introduction of this geometric corrector leads to a weaker estimate. We are going to show that this is not the case by considering the case where the data are affine functions. More precisely, we take for initial conditions (2.2) and boundary conditions (2.3)

$$\varphi(x) = \boldsymbol{\delta_1} \cdot x + \delta_2 \quad \text{and} \quad \psi(x,t) = \boldsymbol{\delta_1} \cdot x + \delta_2 - \boldsymbol{\delta_1} \cdot \mathbf{a} t \,, \quad t \geq 0, \tag{3.41}$$

where $\boldsymbol{\delta_1}$ is an arbitrary vector in $\mathbf{R}^{nd}$ and δ_2 is an arbitrary real number. Then the solution u to (2.1) is

$$u(x,t) = \boldsymbol{\delta_1} \cdot x + \delta_2 - \boldsymbol{\delta_1} \cdot \mathbf{a}t \tag{3.42}$$

with $\nabla u = \boldsymbol{\delta_1}$. In this case a straightforward computation yields that the exact solution of (3.23) has the form

$$\gamma_j^n = -\Gamma_j \cdot \boldsymbol{\delta_1} \equiv -\Gamma_j \cdot \nabla u \tag{3.43}$$

and indicates the optimality of the geometric corrector approach.

3.5. The error estimate. The previous remark suggests using the vector geometrical corrector $(\Gamma_j)_{j=1,\dots,N}$ defined by (3.24) to construct the scalar corrector γ_j^n by taking the formula

$$\gamma_j^n = -\Gamma_j \cdot \nabla u(g_j, t_n)\,. \tag{3.44}$$

Main result: we are now in a position to prove the following general result, where the required smoothness is due to the use of second order Taylor expansions.

THEOREM 3.9. *Let u be the smooth solution to* (2.1)–(2.3)*, where φ and ψ are arbitrary smooth functions satisfying the compatibility conditions (see Bardos* [1]*).*

Assume the local quasi uniformity of the family of meshes, i.e., that there is a positive constant κ_3 such that

$$\frac{1}{\kappa_3}|K_k| \le |K_j| \le \kappa_3 |K_k| \quad \forall h < h_0, \quad \forall K_j \in \mathcal{T}^h, \quad \forall k \in \mathcal{N}(j). \tag{3.45}$$

Let us make the following hypotheses on the discretization of the initial and boundary data: we assume that there exist two constants κ_4 and κ_5 such that for all $h < h_0$,

$$||(u_j^0 - \varphi(g_j))_{j=1,\dots,N}||_p \le \kappa_4 h\,, \tag{3.46}$$

$$|u_k^n - \psi(g_{j,k}, t_n)| \le \kappa_5 h^2 \quad \forall j\,, \quad \forall k \in \mathcal{N}_b^-(j) \text{ and } t_n \le T\,. \tag{3.47}$$

Under the CFL condition $\lambda \le 1$, for every $p \in [1, +\infty]$, if there exists C_p such that the geometric corrector Γ satisfies the estimate

$$||\Gamma||_p \le C_p h\,, \tag{3.48}$$

then the error for the explicit upwind finite volume scheme satisfies the first order estimate

$$||(u_j^n - U_j^n)_{j=1,\dots,N}||_p \le C_p' h, \quad t_n \le T\,. \tag{3.49}$$

Remark 4. If we have instead of (3.48) the weaker estimate

$$||\Gamma||_p \le C_p h^\alpha\,, \tag{3.50}$$

for some $\alpha \in]0,1]$, and provided we keep the same hypotheses, then the proof of Theorem 3.9 leads to the estimate

$$||(u_j^n - U_j^n)_{j=1,\dots,N}||_p \le C_p' h^\alpha, \quad t_n \le T\,. \tag{3.51}$$

Proof. Let us recall (3.9) about the error: $\alpha_j^{n+1} = (\mathcal{L}^n \alpha^n)_j - \Delta t_n E_j^n$. We propose to correct the error α with the corrector γ given by (3.44) as follows:

$$\underline{\alpha}_j^n = \alpha_j^n + \gamma_j^n \tag{3.52}$$

so that

$$\underline{\alpha}_j^{n+1} = (\mathcal{L}^n \underline{\alpha}^n)_j - \Delta t_n \underline{E}_j^n, \tag{3.53}$$

where the corrected truncation error has the form

$$\begin{aligned} \underline{E}_j^n &= E_j^n - \frac{1}{|K_j|}\left(\sum_{k \in \mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}\gamma_j^n + \sum_{k\in\mathcal{N}_0^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}\gamma_k^n \right) - \frac{\gamma_j^{n+1}-\gamma_j^n}{\Delta t_n} \\ &= G_j^n + \underline{I}_j^n + \frac{1}{|K_j|}\sum_{k\in\mathcal{N}_b^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}(u_k^n - u(g_{j,k},t_n)) - \frac{\gamma_j^{n+1}-\gamma_j^n}{\Delta t_n}\,. \end{aligned} \tag{3.54}$$

Now the previous bad behavior part in the error is changed to

$$\begin{aligned} \underline{I}_j^n = \frac{1}{|K_j|}\Bigg(&\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}\left(u(g_j,t_n) - u(g_{j,k},t_n) - \gamma_j^n\right) \\ &+ \sum_{k\in\mathcal{N}_0^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}\left(u(g_k,t_n) - u(g_{j,k},t_n) - \gamma_k^n\right)\Bigg)\,. \end{aligned} \tag{3.55}$$

Using the assumptions (3.2) and (3.45) on the mesh and the Taylor expansions

$$u(g_j,t_n) - u(g_{j,k},t_n) = (g_j - g_{j,k})\cdot\nabla u(g_j,t_n) + \mathcal{O}(h^2)\,, \tag{3.56}$$

$$u(g_k,t_n) - u(g_{j,k},t_n) = (g_k - g_{j,k})\cdot\nabla u(g_j,t_n) + \mathcal{O}(h^2)\,, \tag{3.57}$$

we have

$$\begin{aligned} \underline{I}_j^n = \frac{1}{|K_j|}\Bigg(&\sum_{k\in\mathcal{N}^+(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}\left((g_j - g_{j,k})\cdot\nabla u(g_j,t_n) - \gamma_j^n\right) \\ &+ \sum_{k\in\mathcal{N}_0^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}\left((g_k - g_{j,k})\cdot\nabla u(g_j,t_n) - \gamma_k^n\right)\Bigg) + \mathcal{O}(h)\,. \end{aligned} \tag{3.58}$$

The definitions (3.24) and (3.44) of the geometric corrector Γ and of γ_j^n yield that

$$\underline{I}_j^n = \frac{1}{|K_j|}\sum_{k\in\mathcal{N}_0^-(j)} \mathbf{a}\cdot\mathbf{N}_{j,k}\,\Gamma_k\cdot\left(\nabla u(g_k,t_n) - \nabla u(g_j,t_n)\right) + \mathcal{O}(h)\,. \tag{3.59}$$

Since Γ satisfies the estimate (3.48), we obtain

$$\underline{I}_j^n = \mathcal{O}(h). \tag{3.60}$$

On the other hand

$$\frac{\gamma_j^{n+1}-\gamma_j^n}{\Delta t_n} = \Gamma_j\cdot\left(\nabla\frac{\partial u}{\partial t}(g_j,t_n) + \mathcal{O}(\Delta t_n)\right)\,. \tag{3.61}$$

Then, again from the estimate (3.48) on Γ, we have

$$\frac{\gamma_j^{n+1}-\gamma_j^n}{\Delta t_n}=\mathcal{O}(h). \tag{3.62}$$

Thus using the assumption on the boundary conditions (3.47) and gathering results (3.22), (3.60), (3.62), we conclude that

$$\underline{E}_j^n=\mathcal{O}(h)\,. \tag{3.63}$$

According to the estimate (3.14) applied to (3.53) we deduce that

$$||(u_j^n-u(g_j,t_n)+\gamma_j^n)_{j=1,\dots,N}||_p\le||(u_j^0-u(g_j,0)+\gamma_j^0)_{j=1,\dots,N}||_p+\mathcal{O}(h)\,. \tag{3.64}$$

Now, the assumption on the initial values (3.46) shows that for $t_n\le T$,

$$||(u_j^n-u(g_j,t_n))_{j=1,\dots,N}||_p=\mathcal{O}(h)\,. \tag{3.65}$$

Finally the required results (3.49) follows from the Taylor expansion

$$U_j^n=u(g_j,t_n)+\mathcal{O}(h^2)\,. \qquad \square \tag{3.66}$$

Remark 5. All the results in this section extend to the implicit first order upwind scheme and to more general meshes where the intersection of two adjacent volumes are no longer included in a hyperplan but are composed of several $(nd-1)$-dimensional polygons. All these polygons have to be numbered and are distributed according to the angle they form with vector $\mathbf{a}$ as in (3.1). These extensions will be the subject of a forthcoming article.

Remark 6. In the definition (3.24) of the geometric corrector, the centroid g_j of the volume K_j can be replaced by any point $\tilde{g}_j$ such that $||g_j-\tilde{g}_j||\le C\,h$. Indeed, this will simply change Γ_j into $\Gamma_j+g_j-\tilde{g}_j$.

4. On the geometric corrector. We are now going to study from a theoretical point of view several cases where we have the following estimate on the geometric corrector: there exists C_p such that the sequence $\Gamma\equiv(\Gamma_j)_{j=1,\dots,N}$ satisfies

$$||\Gamma||_p\le C_p h\,. \tag{4.1}$$

4.1. The one-dimensional case. We consider the case $nd=1$. Here the advection vector is a scalar and we assume that $a>0$. The finite volume scheme reads here as

$$\frac{u_j^{n+1}-u_j^n}{\Delta t_n}+a\frac{u_j^n-u_{j-1}^n}{\Delta x_j}=0\,. \tag{4.2}$$

We have $\Omega=]A,B[$ and the "triangulation" of Ω is done by the control volumes $K_j=]x_{j-\frac12},x_{j+\frac12}[$ for $j=1,\dots,N$. The centroid of K_j is given by $g_j=x_j=(x_{j+\frac12}+x_{j-\frac12})/2$ and $|K_j|=\Delta x_j=x_{j+\frac12}-x_{j-\frac12}$. Equations (3.24) read in this case as

$$\left\{\begin{array}{rl} a\,\Gamma_1=a\,(x_{\frac32}-x_1)\,, & j=1\,,\\ a\,(\Gamma_j-\Gamma_{j-1})=a\,(x_{j+\frac12}-x_j)-a\,(x_{j-\frac12}-x_{j-1})\,, & j\ge2\,,\end{array}\right. \tag{4.3}$$

whose solution is simply $\Gamma_j = x_{j+\frac{1}{2}} - x_j = \frac{\Delta x_j}{2}$, $j \geq 1$. In this case, the estimation (4.1) is straightforward for all p and therefore Theorem 3.9 applies: the upwind FVM for the advection equation in one dimension leads to a first order approximation.

THEOREM 4.1. *We denote $h = \max_{j=1,\dots,N} \Delta x_j$ and assume that there is a constant κ_3 such that*

$$\frac{1}{\kappa_3} \leq \frac{\Delta x_j}{\Delta x_{j-1}} \leq \kappa_3 \quad \forall j \geq 2, \quad \forall h < h_0 . \tag{4.4}$$

If there exist two constants κ_4 and κ_5 such that the discretization of the initial and boundary data satisfies

$$|| \left(u_j^0 - \varphi(x_j)\right)_{j=1,\dots,N} ||_p \leq \kappa_4 h \quad \text{and} \quad |u_0^n - \psi(x_{\frac{1}{2}}, t_n)| \leq \kappa_5 h^2 , \quad t_n \leq T , \tag{4.5}$$

then under the CFL condition $\lambda \equiv a\Delta t_n / \min_{j=1,\dots,N} \Delta x_j \leq 1$, for every $p \in [1, +\infty]$, the error $u_j^n - U_j^n$ satisfies the first order estimate

$$||u^n - U^n||_p \leq C'_p h, \quad t_n \leq T . \tag{4.6}$$

Let us observe that the corrected error has the form

$$\underline{\alpha}_j^n = u_j^n - \Big(u(x_j, t_n) + (x_{j+\frac{1}{2}} - x_j) \cdot \nabla u(x_j, t_n)\Big) , \tag{4.7}$$

and this has a clear interpretation: u_j^n provided by the scheme should be compared to $u(x_{j+\frac{1}{2}}, t_n)$ rather than to $u(x_j, t_n)$. The upwinding introduces a bias which leads us to compare u_j^n with $u(x_{j+\frac{1}{2}}, t_n)$. In the case where $a < 0$ we would have found that u_j^n should be compared with $u(x_{j-\frac{1}{2}}, t_n)$. This has already been observed in Eymard, Gallouët, and Herbin [10] in their study of the one-dimensional linear advection equation.

Let us now make explicit the operator B in the one-dimensional case. For $a > 0$ one sees easily that for all $X \in \mathbf{C}^N$,

$$(BX)_1 = 0 \text{ and } (BX)_j = X_{j-1} \quad \text{for } j \geq 2 , \tag{4.8}$$

so that $B^N = 0$ while $B^{N-1} \neq 0$.

4.2. Back to the *nd*-dimensional case.

In this section, we show by straightforward computations that the geometric corrector is of order h for some two-dimensional structured meshes. We first consider the case of meshes composed of parallelograms.

4.2.1. Nonuniform nonorthogonal quadrilateral grid.

We consider a quadrilateral domain bounded from below by two half-lines, which we use as coordinate axes and that are uniquely defined such that the advection vector **a** is oriented from the exterior to the interior on these lines. Now we assume that the mesh is generated by half-lines parallel to these axes, as shown in Figure 4.1.

An element of the mesh is described by two indices (m, n), $m \geq 1$, along the x axis, $n \geq 1$ along the y axis, and $h(m)$ (respectively, $k(n)$) denotes the length of the element (m, n) along x (respectively, y). Here, for instance, sufficient conditions on the mesh to satisfy (3.2) and (3.45) read as

$$c_1 h \leq h(m) \leq c_2 h \quad \forall m \geq 1 , \quad c_3 h \leq k(n) \leq c_4 h \quad \forall n \geq 1 . \tag{4.9}$$

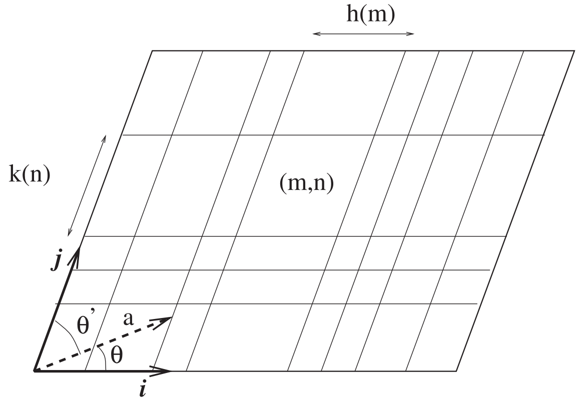


FIG. 4.1. *Nonuniform nonorthogonal quadrilateral grid.*

The corrector for the element (m,n) is therefore denoted by $\Gamma_{m,n}$, and the formula (3.24) for the geometric corrector suggests using

$$G_{m,n} = \Gamma_{m,n} - \frac{h(m)}{2}\vec{i} - \frac{k(n)}{2}\vec{j}, \tag{4.10}$$

where $\vec{i}$ (respectively, $\vec{j}$) denotes the unit vector along the x (respectively, y) axis. If we denote θ (respectively, θ') to be the angle of the advection vector $\mathbf{a}$ with the x axis (respectively, the y axis), it is straightforward to obtain that $G_{m,n}$ satisfy the following recursive formula:

$$G_{m,n} = \lambda_{m,n} G_{m,n-1} + (1-\lambda_{m,n}) G_{m-1,n} \quad \forall m \geq 1\,, \quad \forall n \geq 1, \tag{4.11}$$

where

$$\lambda_{m,n} = \frac{h(m)\sin(\theta)}{h(m) sin(\theta) + k(n)\sin(\theta')} \quad \forall m \geq 1, \quad \forall n \geq 1 \tag{4.12}$$

and where by convention we set

$$G_{m,0} = -\frac{h(m)}{2}\vec{i} \quad \text{and} \quad G_{0,n} = -\frac{h(n)}{2}\vec{j}\,. \tag{4.13}$$

Thus, $G_{m,n}$ is the barycenter of $G_{m,n-1}$ and $G_{m-1,n}$, with respective weights $\lambda_{m,n}$ and $1-\lambda_{m,n}$. Then a recursive computation leads $G_{m,n}$ to be a barycenter of $G_{p,0}$ with $1 \leq p \leq m$ and of $G_{0,q}$ with $1 \leq q \leq n$. From this property, we deduce that the absolute value of both components of $\Gamma_{m,n}$ is smaller than $h/2$ up to a multiplicative constant depending on c_i, $i = 1, \ldots, 4$, and that the estimate (4.1) is true for a nonuniform nonorthogonal quadrilateral mesh.

Remark 7. The above result can also be proved for quadrilateral meshes that satisfy only analogous conditions to (4.4) in each direction.

Remark 8. The above result is easily extended to a mesh composed of parallelepipeds in three dimensions.

4.2.2. Nonuniform nonorthogonal triangular grid. Let us divide each parallelogram of the previous mesh into two triangles according to Figure 4.2.

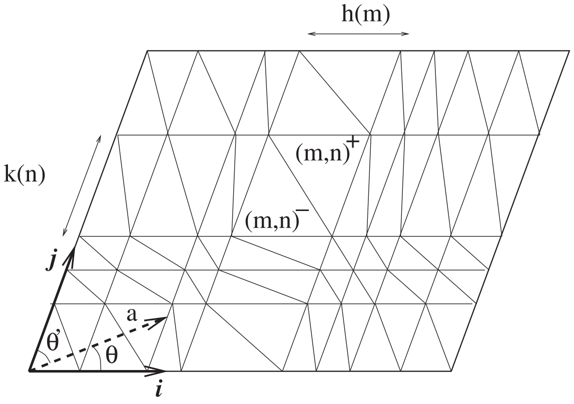


FIG. 4.2. *Nonuniform nonorthogonal triangular grid.*

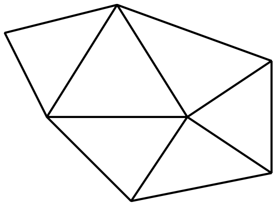


FIG. 4.3. $\mathcal{T}_0$*: a coarse triangular conformal mesh.*

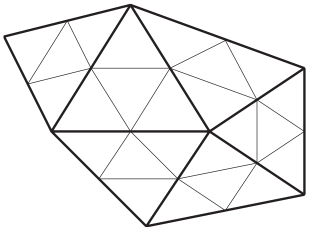


FIG. 4.4. $\mathcal{T}_1$*: 1-refinement of* $\mathcal{T}_0$*.*

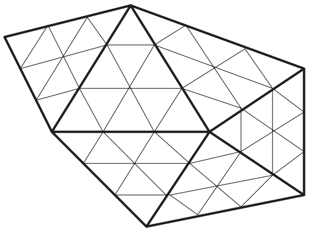


FIG. 4.5. $\mathcal{T}_2$*: 2-refinement of* $\mathcal{T}_0$*.*

The corrector for the downwind triangle of parallelogram (m, n) is denoted by $\Gamma^+_{m,n}$ and the corrector for the upwind one is denoted by $\Gamma^-_{m,n}$. Writing the equations that satisfy the geometric corrector in both triangles and eliminating $\Gamma^-_{m,n}$, we get an equation on $\Gamma^+_{m,n}$ of the same form as for the mesh composed of parallelogram. We now introduce

$$G^+_{m,n} = \Gamma^+_{m,n} - \frac{h(m)}{3}\vec{\imath} - \frac{k(n)}{3}\vec{\jmath} \tag{4.14}$$

and obtain that $G^+_{m,n}$ satisfy the recursive formula (4.11). As for the previous study, we conclude that $\Gamma^+_{m,n}$ is of order h. Using the expression of $\Gamma^-_{m,n}$, which is the sum of terms of order h, we prove estimate (4.1) for this type of meshes.

4.3. Asymptotically structured triangular meshes in two dimensions. The purpose of this section is to prove the estimate (4.1) in the case where a global refinement technique is applied to a given arbitrary and unstructured triangular conformal (in the finite element sense) mesh $\mathcal{T}_0$ as the one plotted in Figure 4.3. Given $\ell \in \mathbf{N}$ a positive integer, we denote by $\mathcal{T}_\ell$ the mesh obtained from $\mathcal{T}_0$ by dividing each triangle into $(\ell+1)^2$ congruent triangles, i.e., by introducing ℓ regularly spaced points on each edge, as can be seen in Figures 4.4 and 4.5.

THEOREM 4.2. *For every triangular conformal mesh* $\mathcal{T}_0$ *and for every nonvanishing vector* $\mathbf{a}$ *in* $\mathbf{R}^2$ *there exists a constant* $C(\mathbf{a}, \mathcal{T}_0)$ *such that for every* $\ell \in \mathbf{N}$ *we*

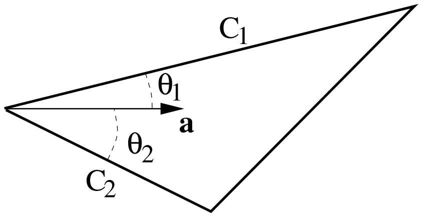


FIG. 4.6. *The case of* $\sharp\mathcal{N}^+(T) = 1$.

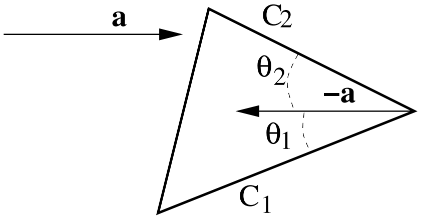


FIG. 4.7. *The case of* $\sharp\mathcal{N}^+(T) = 2$.

have the estimate

$$||\Gamma^{(\ell)}||_\infty \leq \frac{C(\mathbf{a}, \mathcal{T}_0)}{\ell + 1}\,, \tag{4.15}$$

where $\Gamma^{(\ell)}$ *denotes the geometric corrector sequence on* $\mathcal{T}_\ell$.

Denoting by $h^{(\ell)} = \max_{K_j \in \mathcal{T}_\ell} h_j > 0$, we have by construction $h^{(\ell)} = \frac{h^{(0)}}{\ell+1}$, and since $h^{(0)}$ depends only on $\mathcal{T}_0$, we deduce from (4.15) the proof of the estimate (4.1) in this case.

COROLLARY 4.3. *For every triangular conformal mesh* $\mathcal{T}_0$ *of a polygonal domain* Ω *and for every nonvanishing vector* $\mathbf{a}$ *in* $\mathbf{R}^2$*, there exists a constant* $C(\mathbf{a}, \mathcal{T}_0)$ *such that for every* $\ell \in \mathbf{N}$ *and for every real number* $p \in [1, \infty]$ *we have the estimate*

$$||\Gamma^{(\ell)}||_p \leq C(\mathbf{a}, \mathcal{T}_0)|\Omega|^{1/p} \max_{K_j \in \mathcal{T}_\ell} h_j\,. \tag{4.16}$$

Actually we are going to show a stronger result than Theorem 4.2. To state this result, we introduce the notation $\mathcal{N}^-(\Omega) = \{j$ such that K_j meets $\partial\Omega^-$ according a set of $nd - 1$ positive measure$\}$ and consider instead of (3.24)

$$\begin{aligned} &\sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} \Gamma_j + \sum_{k \in \mathcal{N}_0^-(j)} \mathbf{a} \cdot \mathbf{N}_{j,k} \Gamma_k \\ &\qquad = \sum_{k \in \mathcal{N}^+(j)} \mathbf{a} \cdot \mathbf{N}_{j,k}(g_{j,k} - g_j) + \sum_{k \in \mathcal{N}_0^-(j)} \mathbf{a} \cdot \mathbf{N}_{j,k}(g_{j,k} - g_k) + \epsilon_j\,, \end{aligned} \tag{4.17}$$

where ϵ_j vanishes for j not in $\mathcal{N}^-(\Omega)$.

THEOREM 4.4. *For every triangular conformal mesh* $\mathcal{T}_0$ *and for every nonvanishing vector* $\mathbf{a}$ *in* $\mathbf{R}^2$*, we consider on* $\mathcal{T}_\ell$ *a sequence* $\epsilon_j = \frac{\delta_j}{(\ell+1)^2}$*, where the sequence* δ_j *of vectors in* $\mathbf{R}^2$ *is uniformly bounded by a constant* $C_0(\mathbf{a}, \mathcal{T}_0)$ *and vanishes for* j *not in* $\mathcal{N}^-(\Omega)$*. Then there exists a constant* $C(a, \mathcal{T}_0)$ *such that for every* $\ell \in \mathbf{N}$ *we have the estimate* (4.15)*, where* $\Gamma^{(\ell)}$ *denotes the solution to* (4.17) *on* $\mathcal{T}_\ell$.

The proof of Theorem 4.4 proceeds by induction on M, the number of elements of $\mathcal{T}_0$.

4.3.1. The case where $\mathcal{T}_0$ has one element. In this case Ω is composed of a single triangle T and $\mathcal{T}_0 = T$. Besides the simpler case where the vector $\mathbf{a}$ is tangent to one of the edges of T, there are two cases that are presented in Figures 4.6 and 4.7, where $\theta_1 \neq 0$ and $\theta_2 \neq 0$: either $\sharp\mathcal{N}^+(T) = 1$ or $\sharp\mathcal{N}^+(T) = 2$. We index, as shown in

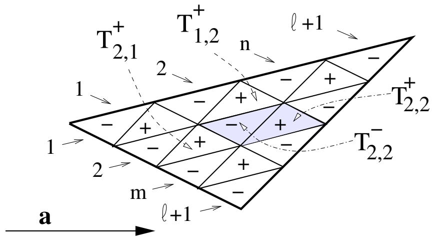


FIG. 4.8. $\sharp\mathcal{N}^+(T) = 1$*: an* ℓ*-refinement.*

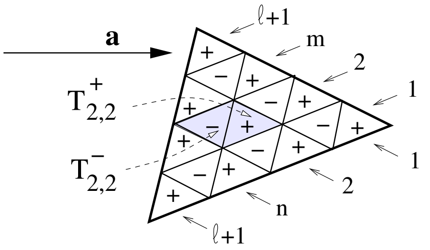


FIG. 4.9. $\sharp\mathcal{N}^+(T) = 2$*: an* ℓ*-refinement.*

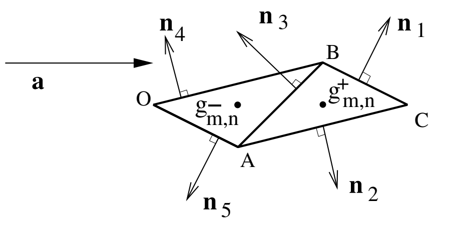


FIG. 4.10. $\sharp\mathcal{N}^+(T) = 1$*: the two subtriangles* $T^+_{m,n}$ *and* $T^-_{m,n}$ *(*$\mathbf{n}_i \equiv \mathbf{N}_i/||\mathbf{N}_i||$*).*

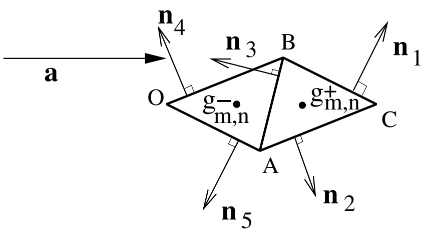


FIG. 4.11. $\sharp\mathcal{N}^+(T) = 2$*: the two subtriangles* $T^+_{m,n}$ *and* $T^-_{m,n}$ *(*$\mathbf{n}_i \equiv \mathbf{N}_i/||\mathbf{N}_i||$*).*

Figures 4.8 and 4.9, the $(\ell+1)^2$ triangles in $\mathcal{T}_\ell$, for $\ell \geq 1$, by a triplet (n, m, ϵ), where n and m are natural numbers satisfying $1 \leq n \leq \ell+1$, $1 \leq m \leq \ell+1$, $n+m \leq \ell+2$, and $\epsilon = \pm$. When $\epsilon = -$, we have $\sharp\mathcal{N}^+(T^\epsilon_{m,n}) = 1$, while when $\epsilon = +$, we have $\sharp\mathcal{N}^+(T^\epsilon_{m,n}) = 2$. We say that the triangle $T^\epsilon_{m,n}$ is interior when it has no edge on the boundary of T.

The case where $\sharp\mathcal{N}^+(T) = 1$*.*

- All the subtriangles $T^+_{m,n}$ (with $1 \leq m \leq \ell$, $1 \leq n \leq \ell$, and $m+n \leq \ell+1$) are interior ones. Hence if we adopt the notation explained in Figure 4.10, then we can write (4.17) in $T^+_{m,n}$ as follows:

$$\begin{aligned}(4.18)\qquad & \mathbf{a}\cdot\mathbf{N}_1\Gamma^+_{m,n} + \mathbf{a}\cdot\mathbf{N}_2\Gamma^+_{m,n} + \mathbf{a}\cdot\mathbf{N}_3\Gamma^-_{m,n} \\ &= \mathbf{a}\cdot\mathbf{N}_1\left(\frac{B+C}{2} - g^+_{m,n}\right) + \mathbf{a}\cdot\mathbf{N}_2\left(\frac{C+A}{2} - g^+_{m,n}\right) \\ &\quad + \mathbf{a}\cdot\mathbf{N}_3\left(\frac{A+B}{2} - g^-_{m,n}\right).\end{aligned}$$

If c_1, c_2, θ_1 and θ_2 are defined as in Figure 4.6, since $\mathbf{a}\cdot\mathbf{N}_1 + \mathbf{a}\cdot\mathbf{N}_2 = -\mathbf{a}\cdot\mathbf{N}_3 = \frac{c_1\sin\theta_1 + c_2\sin\theta_2}{l+1}$ and since the right-hand side of (4.18) is equal to $\frac{\boldsymbol{\zeta}_1}{(l+1)^2}$, where $\boldsymbol{\zeta}_1$ is a vector which depends only on T and $\mathbf{a}$, we deduce that

(4.19)
$$\Gamma^+_{m,n} - \Gamma^-_{m,n} = \frac{1}{\ell+1}\frac{\boldsymbol{\zeta}_1}{c_1\sin\theta_1 + c_2\sin\theta_2} \quad \text{for } 1 \leq m, n \leq \ell\,,\ m+n \leq \ell+1\,.$$

- Now let us consider the interior triangles $T^+_{m,n}$ for which we have $2 \leq m \leq \ell$, $2 \leq n \leq \ell$. Then the four triangles $T^+_{m,n}$, $T^-_{m,n}$, $T^+_{m-1,n}$, $T^+_{m,n-1}$ are also interior triangles and we can write (4.17) in the triangle $T^-_{m,n}$ as

$$\begin{aligned}(4.20)\qquad &\mathbf{a}\cdot\mathbf{N}_4\Gamma^+_{m-1,n} + \mathbf{a}\cdot\mathbf{N}_5\Gamma^+_{m,n-1} - \mathbf{a}\cdot\mathbf{N}_3\Gamma^-_{m,n}\\ &= \mathbf{a}\cdot\mathbf{N}_4\left(\frac{O+B}{2} - g^+_{m-1,n}\right) + \mathbf{a}\cdot\mathbf{N}_5\left(\frac{A+O}{2} - g^+_{m,n-1}\right)\\ &\quad - \mathbf{a}\cdot\mathbf{N}_3\left(\frac{A+B}{2} - g^-_{m,n}\right).\end{aligned}$$

By summing up (4.18) and (4.20), since their right-hand sides are opposite, we obtain

$$(4.21)\qquad \Gamma^+_{m,n} = \frac{c_1\sin\theta_1\Gamma^+_{m-1,n} + c_2\sin\theta_2\Gamma^+_{m,n-1}}{c_1\sin\theta_1 + c_2\sin\theta_2}.$$

- Let us now discuss the case where $T^-_{m,n}$ has at least one edge on the boundary of T. In this case the only change in (4.18) and (4.20) is that corresponding terms in the right- and left-hand sides are not present anymore when a triangle $T^\epsilon_{p,q}$ does not exist. Let us give these cases in full detail.
  - *The case* $T^-_{1,1}$. The analogue of (4.20) reads as

$$(4.22)\qquad -\mathbf{a}\cdot\mathbf{N}_3\Gamma^-_{1,1} = -\mathbf{a}\cdot\mathbf{N}_3\left(\frac{A+B}{2} - g^-_{1,1}\right) + \frac{\delta^-_{1,1}}{(\ell+1)^2}.$$

  Hence from (4.22) and (4.19)

$$(4.23)\qquad \Gamma^+_{1,1} = \frac{\boldsymbol{\zeta}_2}{\ell+1},$$

  where $\boldsymbol{\zeta}_2$ is a vector which depends only on T and $\mathbf{a}$.
  - *The case* $T^-_{m,1}$, $2 \leq m \leq \ell$. Equations (4.18), (4.20), and (4.21) are still valid in this case if we set $\Gamma^+_{m,0} \equiv \frac{B+C}{2} - g^+_{m,1} + \frac{\delta^-_{m,1}}{(\ell+1)c_2\sin\theta_2}$. Let us observe that this terms behaves like

$$(4.24)\qquad \Gamma^+_{m,0} = \frac{\boldsymbol{\zeta}_3}{\ell+1},$$

  where $\boldsymbol{\zeta}_3$ is a vector which depends only on T and $\mathbf{a}$.
  - *The case* $T^-_{1,n}$, $2 \leq n \leq \ell$. Here again (4.18), (4.20), and (4.21) are still valid in this case if we set $\Gamma^+_{0,n} \equiv \frac{A+C}{2} - g^+_{1,n} + \frac{\delta^-_{1,n}}{(\ell+1)c_1\sin\theta_1}$ and observe that

$$(4.25)\qquad \Gamma^+_{0,n} = \frac{\boldsymbol{\zeta}_4}{\ell+1},$$

  where $\boldsymbol{\zeta}_4$ is a vector which depends only on T and $\mathbf{a}$.
  - *The case* $T^-_{m,n}$ *with* $m+n = \ell+2$. Here we can only write (4.20), and we obtain

$$(4.26)\qquad \Gamma^-_{m,n} = \frac{c_1\sin\theta_1\Gamma^+_{m-1,n} + c_2\sin\theta_2\Gamma^+_{m,n-1}}{c_1\sin\theta_1 + c_2\sin\theta_2} + \frac{\boldsymbol{\zeta}_5}{\ell+1},$$

  where $\boldsymbol{\zeta}_5$ is a vector which depends only on T and $\mathbf{a}$ and where we set $\Gamma^+_{\ell+1,0} \equiv \frac{\delta^-_{\ell+1,1}}{(\ell+1)c_2\sin\theta_2}$ and $\Gamma^+_{0,\ell+1} \equiv \frac{\delta^-_{1,\ell+1}}{(\ell+1)c_1\sin\theta_1}$.

The barycenter structure of formula (4.21) is the key point and permits us to conclude. We find by induction on (m, n) the estimation on $\Gamma^+_{m,n}$ and then, by relation (4.19), the estimation on $\Gamma^-_{m,n}$ where the constant $C(\mathbf{a}, T)$ depends only on T and $\mathbf{a}$:

$$|\Gamma^\epsilon_{m,n}| \leq \frac{C(\mathbf{a}, T)}{\ell + 1}\,. \tag{4.27}$$

The case where $\sharp\mathcal{N}^+(T) = 2$. This case is slightly simpler because all the subtriangles $T^-_{m,n}$ are interior ones. Hence if we adopt the notation explained in Figure 4.11, for every (m, n) with $1 \leq n \leq \ell$, $1 \leq m \leq \ell$, $n + m \leq \ell + 1$, equation (4.17) in $T^-_{m,n}$ reads as

(4.28)

$$\begin{aligned}&\mathbf{a}\cdot\mathbf{N}_4\Gamma^+_{m+1,n} + \mathbf{a}\cdot\mathbf{N}_5\Gamma^+_{m,n+1} - \mathbf{a}\cdot\mathbf{N}_3\Gamma^-_{m,n}\\ &= \mathbf{a}\cdot\mathbf{N}_4\left(\frac{O+B}{2} - g^+_{m+1,n}\right) + \mathbf{a}\cdot\mathbf{N}_5\left(\frac{A+O}{2} - g^+_{m,n+1}\right) - \mathbf{a}\cdot\mathbf{N}_3\left(\frac{A+B}{2} - g^-_{m,n}\right).\end{aligned}$$

Concerning the subtriangles $T^+_{m,n}$, we have two cases.

- In the first one, $1 \leq n \leq \ell$, $1 \leq m \leq \ell$, $m + n \leq \ell + 1$. Then (4.17) in $T^+_{m,n}$ reads again as (4.18), and (4.19) is still valid. Combining (4.18) with (4.28), formula (4.21) is now replaced by

$$\Gamma^+_{m,n} = \frac{c_1 \sin\theta_1 \Gamma^+_{m+1,n} + c_2 \sin\theta_2 \Gamma^+_{m,n+1}}{c_1 \sin\theta_1 + c_2 \sin\theta_2}\,. \tag{4.29}$$

- In the second one, $m + n = \ell + 2$, and then (4.17) in $T^+_{m,n}$ reads as

  (4.30)

$$\begin{aligned}&\mathbf{a}\cdot\mathbf{N}_1\Gamma^+_{m,n} + \mathbf{a}\cdot\mathbf{N}_2\Gamma^+_{m,n}\\ &= \mathbf{a}\cdot\mathbf{N}_1\left(\frac{B+C}{2} - g^+_{m,n}\right) + \mathbf{a}\cdot\mathbf{N}_2\left(\frac{C+A}{2} - g^+_{m,n}\right) + \frac{\delta^+_{m,n}}{(\ell+1)^2}\,.\end{aligned}$$

  Hence we can say that

$$\Gamma^+_{m,n} = \frac{\boldsymbol{\zeta}}{\ell + 1} \text{ for } m + n = \ell + 2\,, \tag{4.31}$$

  where $\boldsymbol{\zeta}$ is a vector which depends only on T and $\mathbf{a}$.

In conclusion, we are able to show first by induction on $p = \ell + 3 - m - n$ that

$$|\Gamma^+_{m,n}| \leq \frac{|\boldsymbol{\zeta}|}{\ell + 1}\,. \tag{4.32}$$

Indeed, for $p = 1$ this assertion follows from (4.31). Assuming that (4.32) holds true for p, we take (m, n) with $m + n = \ell + 2 - p = \ell + 3 - (p + 1)$ and write (4.29). Since $m + n + 1 = \ell + 3 - p$ we can apply (4.32):

$$|\Gamma^+_{m+1,n}| \leq \frac{|\boldsymbol{\zeta}|}{\ell + 1}\,, |\Gamma^+_{m,n+1}| \leq \frac{|\boldsymbol{\zeta}|}{\ell + 1}\,. \tag{4.33}$$

So that thanks to (4.29), (4.32) holds true for $p + 1$. Now that (4.32) is shown, using (4.19) we conclude that (4.27) holds true in the case $\sharp\mathcal{N}^+(T) = 2$.

This ends the proof of Theorem 4.4 in the case where $\mathcal{T}_0$ has only one element.

4.3.2. The case where $\mathcal{T}_0$ has two elements. In this case, the domain Ω is the union of two triangles which share an edge. If $\mathbf{a}$ is tangent to this edge, then sequences of geometric correctors on one triangle are independent from sequences defined on the other one. Hence Theorem 4.4 follows in this case. Now when $\mathbf{a}$ is not tangent to this edge we denote these two triangles by T_1 and T_2 in order that $\mathbf{a}$ points from T_1 into T_2. We denote by $(\Gamma^{(\ell)})$ the solution to (4.17) on $\mathcal{T}_\ell$ the ℓ-refinement of $\mathcal{T}_0 = T_1 \cup T_2$. We also denote by $\mathcal{T}_{\ell,1}$ and $\mathcal{T}_{\ell,2}$ the ℓ-refinements of $\mathcal{T}_{0,1} = T_1$ and of $\mathcal{T}_{0,2} = T_2$. We are going to analyze subsequences of $(\Gamma^{(\ell)})$ associated to the triangle T_1 and the triangle T_2.

The first key observation is that since on $T_1 \cap T_2$ all the outer normals in subtriangles which are in T_1 form an acute angle with $\mathbf{a}$, the subsequence of $(\Gamma^{(\ell)})$ whose indices correspond to triangles which are in T_1 satisfies exactly (4.17) when it is written only for the refinement $\mathcal{T}_{\ell,1}$. But according to Theorem 3.8 we know that this solution is unique, and therefore this subsequence is equal to the sequence of geometric correctors defined on $\mathcal{T}_{\ell,1}$. We already proved that Theorem 4.4 applies to T_1 and deduce that

$$||\Gamma_j^{(\ell)}|| \le \frac{C_1(\mathbf{a}, \mathcal{T}_0)}{\ell+1} \quad \forall j \text{ corresponding to a subtriangle in } T_1 \,. \tag{4.34}$$

The next step is to estimate the remaining geometric correctors that correspond to subtriangles in T_2. By inspection of (4.17) when it is written for $\mathcal{T}_{\ell,2}$, one notices two facts:

1. If one considers a subtriangle which is in T_2 but which has no edge on $T_1 \cap T_2$, then the corrector of this subtriangle satisfies the same equation in both $\mathcal{T}_{\ell,2}$ and $\mathcal{T}_\ell$.
2. However, for a subtriangle j which is in T_2 and which has an edge on $T_1 \cap T_2$, we do not have the same equation since one term on each side of the equation is missing. But the difference is equal to

$$\mathbf{a} \cdot \mathbf{N}_{j,j_1} \left(\Gamma_{j_1}^{(\ell),1} + g_{j,j_1} - g_{j_1} \right) , \tag{4.35}$$

where we denote by j_1 the subtriangle in T_1 which shares an edge with the subtriangle j.

From (4.34), we observe that these differences slightly modify the right-hand side of equations associated to subtriangles of T_2 that share a boundary with ∂T_2^- by adding a term of the form $\frac{\delta(\mathbf{a},\mathcal{T}_0)}{(\ell+1)^2}$. Applying Theorem 4.4 to T_2, we deduce that

$$||\Gamma_j^{(\ell)}|| \le \frac{C_2(\mathbf{a}, \mathcal{T}_0)}{\ell+1} \quad \forall j \text{ corresponding to a subtriangle in } T_2 \,. \tag{4.36}$$

Then (4.34) and (4.36) allow us to conclude the proof of Theorem 4.4 in this case.

4.3.3. Proof of Theorem 4.4. Let $M \ge 1$ be given. We assume that this result is true for all triangulation $\mathcal{T}_0$ with a number of elements less than or equal to M. Let $\mathcal{T}_0$ be a triangulation with $M+1$ elements. First we are going to decompose this triangulation into two parts separated by a broken line made with consecutive edges having the same orientation with respect to $\mathbf{a}$ (see Figure 4.12). Then we will finish the proof by simply observing that it follows by extending the argument given in the case $M = 2$ in the previous section.

We prove the decomposition in two parts for a more general triangulation.

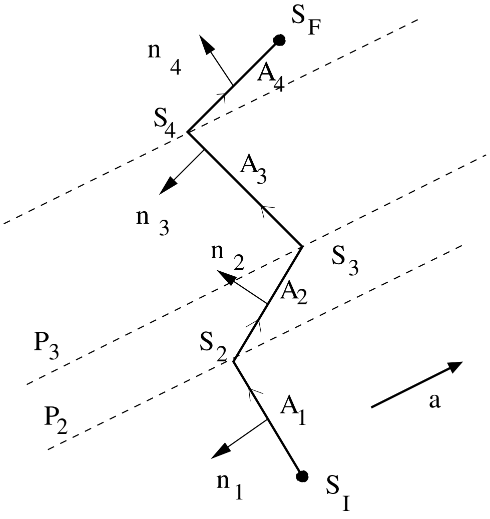


FIG. 4.12. *Broken line made with consecutive oriented edges* ($\mathbf{n}_i \equiv \mathbf{N}_i/||\mathbf{N}_i||$).

LEMMA 4.5. *Let $\mathcal{T}_0$ be a conformal mesh made with strictly convex polygons; i.e., the interior angle at vertices are strictly less than π. We assume that there is at least one interior edge not parallel to the vector* $\mathbf{a}$. *There exists a broken line made with consecutive oriented edges $A_1 = (S_I, S_2), \dots, A_i = (S_i, S_{i+1}), \dots, A_K = (S_K, S_F)$ which links two distinct vertices on the boundary S_I and S_F and such that $\mathbf{N}_i \cdot \mathbf{a} < 0$ for all $i = 1, \dots, K$, where the angle $(A_i, \mathbf{N}_i)$ is equal to $\frac{\pi}{2}$.*

Proof. First let us choose an arbitrary vertex on the boundary that is a vertex of an interior edge $A_1 = (S_I, S_2)$. We assume that the normal $\mathbf{N}_1$ to A_1 such that the angle $(A_1, \mathbf{N}_1)$ is equal to $\frac{\pi}{2}$ satisfies $\mathbf{N}_1 \cdot \mathbf{a} < 0$. If not, we take $\mathbf{a} = -\mathbf{a}$ and switch the role of S_I and S_F.

If S_2 is a vertex on the boundary, then we conclude with $S_F = S_2$. If not, let us denote by $\mathcal{P}_2$ the half-plane defined by the straight line $(S_2, \mathbf{a})$ (dashed line in Figure 4.12) and that does not contain the edge A_1. Since S_2 is an interior vertex and since interior angles of volumes sharing this vertex are strictly less than π, there is necessarily an edge $A_2 = (S_2, S_3)$ strictly in the half-plane $\mathcal{P}_2$. From the definition of $\mathcal{P}_2$, we have $\mathbf{N}_2 \cdot \mathbf{a} < 0$, where $\mathbf{N}_2$ is the normal to A_2 such that the angle $(A_2, \mathbf{N}_2)$ is equal to $\frac{\pi}{2}$.

Now, by induction we can determine a broken line of oriented edges $A_1, \dots, A_i$ such that $\mathbf{N}_1 \cdot \mathbf{a} < 0, \dots, \mathbf{N}_i \cdot \mathbf{a} < 0$ with the same convention for the orientation of $\mathbf{N}_i$. Since the distance of S_I to the line $(S_{i+1}, \mathbf{a})$ is strictly superior to the distance of S_I to the line $(S_i, \mathbf{a})$, there exists K such $S_{K+1} \equiv S_F$ is on the boundary. □

Thus we can realize Ω as the union of two adjacent polygonal sets Ω_1 and Ω_2 where $\mathbf{a}$ points from Ω_1 into Ω_2. Here again, the subsequence of geometric correctors on Ω whose indices correspond to triangles in Ω_1 is identical to the sequence of geometric correctors only defined in Ω_1. Since the coarse triangulation of Ω_1 has less than M elements, we can apply Theorem 4.4. For each subtriangle j of Ω_2 that shares an edge with the broken line defined in Lemma 4.5, the equation that satisfies the corrector in Ω and the equation that satisfies the corrector in Ω_2 differ from a combination of terms like (4.35), which are of the form $\frac{\delta_j}{(\ell+1)^2}$, and the δ_j are uniformly bounded by a constant which depends only on $\mathbf{a}$ and $\mathcal{T}_0$. Then, since Ω_2 has less than M elements, we can again apply Theorem 4.4 and therefore show this result for $M+1$. This concludes the proof of order h convergence of the upwind scheme for a uniform refinement of an arbitrary coarse conformal triangular mesh.

5. Numerical estimates for independent refinements. Since our analysis developed in section 3 is valid for arbitrary types of meshes, in order to validate from a numerical point of view the estimate (4.1), we perform some tests with a sequence of independent unstructured meshes where the mesh size decreases. In the present simulation, if we take two consecutive grids, one is not the refinement of the other one (by dividing, for instance, each triangle into four congruent subtriangles), but the mesh size is reduced. We consider a dodecagon with several meshes composed from 274 triangles to 286,514 triangles computed with the software Gmsh [12], developed by Jean-François Remacle and Christophe Geuzaine. In Figure 5.1, the pictures correspond to the three independent meshes of the sequence. We compute the corrector as solution of (3.24) for two advection vectors defined by the angle θ with the x axis: $\theta = 0$ and $\theta = \pi/4$.

In Figures 5.2, 5.3, and 5.4 the L^1 and L^∞ estimates of the geometric corrector and the L^1 norm of Δ defined by (3.29) are plotted versus the mesh size. The straight lines are the least-squares fits to the point and the slopes are 0.994 ($\theta = 0$) and 0.988 ($\theta = \pi/4$) for the L^1 norm of the corrector. We find 0.831 ($\theta = 0$) and 0.564 ($\theta = \pi/4$) for the L^∞ norm of the corrector and 0.969 ($\theta = 0$) and 0.954 ($\theta = \pi/4$) for $\|\Delta\|_1$. The L^1 norm of the geometric corrector and the norm of Δ behaves like $\mathcal{O}(h)$. Concerning the L^∞ norm of Γ, we have an unexpected behavior when the angle θ that defines the advection vector $\mathbf{a}$ is equal to $\pi/4$ (as shown in Figure 5.2), i.e., when the advection vector is parallel to two sides of the domain. Here, asymptotically we have $\|\Gamma\|_\infty = \mathcal{O}(h^{1/2})$, and although the right-hand side of (3.30) tends to zero with h in all cases, we observe that the estimation of the norm of the solution of this equation depends on the relative position of the advection vector with the boundary.

This behavior is similar to the loss of accuracy proved in Peterson [23]. More precisely, for the mesh and the subtle alignment with the direction of transport proposed by Peterson, we are able to prove that the L^∞ norm of the geometric corrector behaves like $\mathcal{O}(h^{1/2})$ and the L^1 norm is of order one. The technical proof is quite long and out of the scope of this paper and will be published elsewhere.

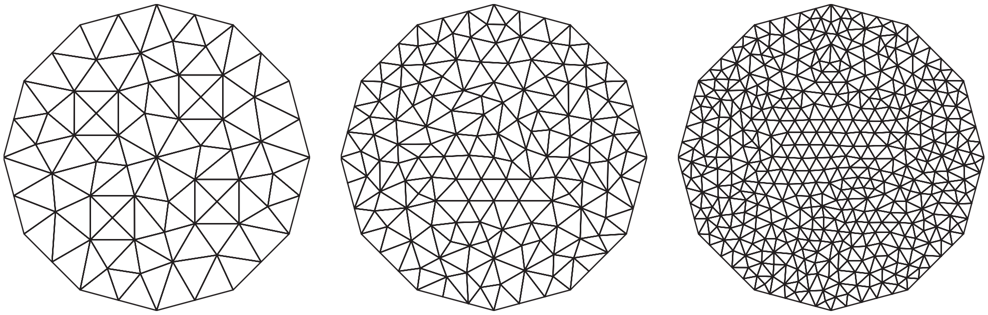


FIG. 5.1. *Some terms of the sequence of independent meshes of the dodecagon domain.*

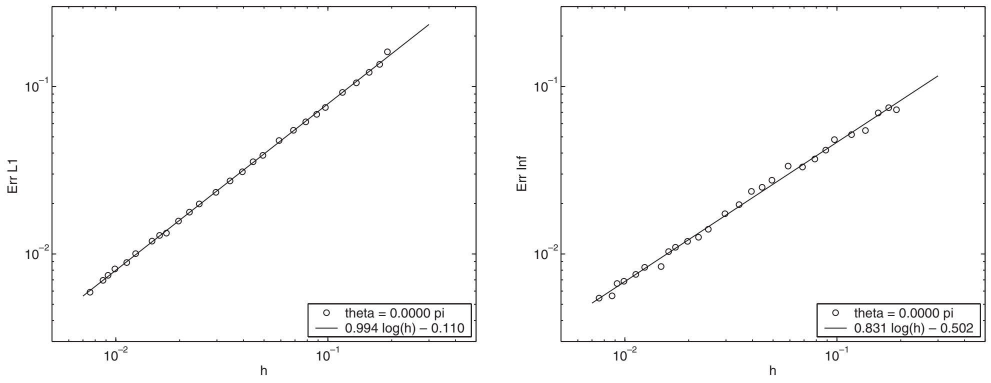


FIG. 5.2. *$\theta = 0$: $\|\Gamma\|_1$ and $\|\Gamma\|_\infty$ versus h for a sequence of independent meshes of the dodecagon domain.*

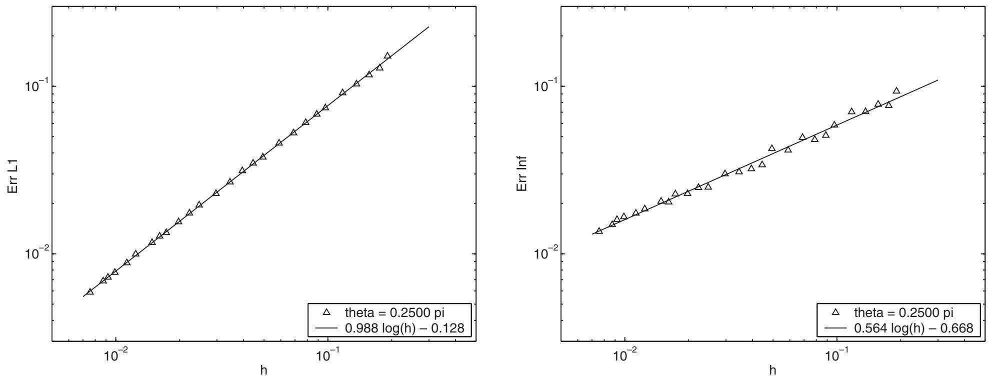


FIG. 5.3. *$\theta = \frac{\pi}{4}$: $\|\Gamma\|_1$ and $\|\Gamma\|_\infty$ versus h for a sequence of independent meshes of the dodecagon domain.*

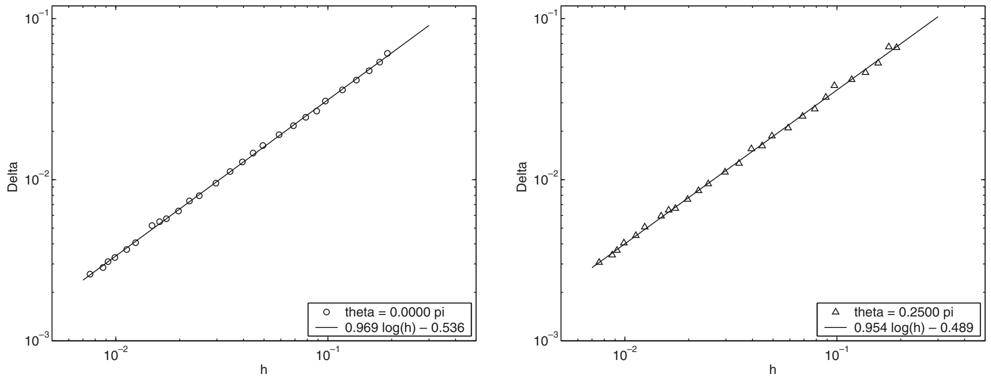


FIG. 5.4. *$\theta = 0$ and $\theta = \frac{\pi}{4}$: $\|\Delta\|_1$ versus h for a sequence of independent meshes of the dodecagon domain.*

6. Conclusion. For linear advection on bounded domain, we introduce the geometric corrector that depends only on the mesh and the velocity **a**. We first prove a general result that links the convergence of the finite volume scheme and the geometric corrector estimates. Then an analytical study of these correctors in the case of uniformly refined triangular meshes in two dimensions leads to the proof of the one order of convergence for the finite volume scheme applied to the linear advection with smooth solution and constant velocity. We plan to address the variable velocity case in a future work.

REFERENCES

[1] C. Bardos, *Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; application à l'équation de transport*, Ann. Sci. École Norm. Sup. (4), 3 (1970), pp. 185–233.

[2] M. J. Berger, C. Helzel, and R. J. LeVeque, *h-box methods for the approximation of hyperbolic conservation laws on irregular grids*, SIAM J. Numer. Anal., 41 (2003), pp. 893–918.

[3] M. J. Berger, R. J. LeVeque, and L. G. Stern, *Finite volume methods for irregular one-dimensional grids*, in Mathematics of Computation 1943–1993: A Half-century of Computational Mathematics (Vancouver, BC, 1993), Proc. Sympos. Appl. Math. 48, AMS, Providence, RI, 1994, pp. 255–259.

[4] B. Cockburn and P.-A. Gremaud, *A priori error estimates for numerical methods for scalar conservation laws.* I: *The general approach*, Math. Comput., 65 (1996), pp. 533–573.

[5] B. Cockburn and P.-A. Gremaud, *A priori error estimates for numerical methods for scalar conservation laws.* II: *Flux-splitting monotone schemes on irregular Cartesian grids*, Math. Comput., 66 (1997), pp. 547–572.

[6] B. Cockburn, P.-A. Gremaud, and J. X. Yang, *A priori error estimates for numerical methods for scalar conservation laws. Part* III: *Multidimensional flux-splitting monotone schemes on non-Cartesian grids*, SIAM J. Numer. Anal., 35 (1998), pp. 1775–1803.

[7] B. Després, *Théorème de lax et volumes finis*, in Proceedings du 32e CANUM, La Grande-Motte, Université de Montpellier et S.M.A.I., 2003.

[8] B. Després, *An explicit a priori estimate for a finite volume approximation of linear advection on non-Cartesian grids*, SIAM J. Numer. Anal., 42 (2004), pp. 484–504.

[9] B. Després, *Lax theorem and finite volume schemes*, Math. Comp., 73 (2004), pp. 1203–1234.

[10] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, P.-A. Ciarlet and J.-L. Lions, eds., Handb. Numer. Anal. 4, North-Holland, Amsterdam, 2000, pp. 713–1020.

[11] B. Garcia-Archilla and J. M. Sanz-Serna, *A finite difference formula for the discretization of d^3/dx^3 on nonuniform grids*, Math. Comp., 57 (1991), pp. 239–257.

[12] C. Geuzaine and J.-F. Remacle, *Gmsh: A three-dimensional finite element mesh generator with built-in pre- and post-processing facilities*; available online from http://www.geuz.org/gmsh/, 1997–2005.

[13] E. Godlewski and P.A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer-Verlag, New York, 1996.

[14] J. D. Hoffman, *Relationship between the truncation errors of centered finite-difference approximations on uniform and nonuniform meshes*, J. Comput. Phys., 46 (1982), pp. 469–474.

[15] H.-O. Kreiss, T.A. Manteuffel, B. Swartz, B. Wendroff, and A. B. White, Jr., *Supra-convergent schemes on irregular grids*, Math. Comp., 47 (1986), 537–554.

[16] D. Kröner, *Numerical Schemes for Conservation Laws*, Wiley-Teubner Ser. Adv. Numer. Math., Wiley, Chichester, UK, 1997.

[17] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation*, U.S.S.R. Comput. Math. Math. Phys., 16 (1976), pp. 105–119.

[18] C. D. Levermore, T. A. Manteuffel, and A. B. White, Jr., *Numerical solution of partial differential equations on irregular grids*, in Computational Techniques and Applications: CTAC-87 (Sydney, 1987), North-Holland, Amsterdam, 1988, pp. 417–426.

[19] T. A. Manteuffel and A. B. White, Jr., *The numerical solution of second order boundary value problems on nonuniform meshes*, Math. Comput., 47 (1986), pp. 511–535.

[20] T. A. Manteuffel and A. B. White, Jr., *On the efficient numerical solution of systems of second order boundary value problems*, SIAM J. Numer. Anal., 23 (1986), pp. 996–1006.

[21] T. A. Manteuffel and A. B. White, Jr., *A calculus of difference schemes for the solution of boundary-value problems on irregular meshes*, SIAM J. Numer. Anal., 29 (1992), pp. 1321–1346.

[22] P. Monk and E. Süli, *A convergence analysis of Yee's scheme on nonuniform grids*, SIAM J. Numer. Anal., 31 (1994), pp. 393–412.

[23] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[24] J. Pike, *Grid adaptive algorithms for the solution of the Euler equations on irregular grids*, J. Comput. Phys., 71 (1987), pp. 194–223.

[25] Z.-H. Teng and P. Zhang, *Optimal L^1-rate of convergence for the viscosity method and monotone scheme to piecewise constant solutions with shocks*, SIAM J. Numer. Anal., 34 (1997), pp. 959–978.

[26] A. N. Tikhonov and A. A. Samarskij, *Homogeneous difference schemes on non-uniform nets*, U.S.S.R. Comput. Math. Math. Phys., 3 (1964), pp. 927–953.

[27] E. Turkel, *Accuracy of schemes with nonuniform meshes for compressible fluid flows*, Appl. Numer. Math., 2 (1986), pp. 529–550.

[28] J.-P. Vila and P. Villedieu, *Convergence of an explicit finite volume scheme for first order symmetric systems*, Numer. Math., 94 (2003), pp. 573–602.

[29] B. Wendroff, *Supraconvergence in Two Dimensions*, Technical report LA-UR 95-3068, Los Alamos National Laboratory, Los Alamos, NM, 1995.

[30] B. Wendroff and A. B. White, Jr., *Some supraconvergent schemes for hyperbolic equations on irregular grids*, in Nonlinear Hyperbolic Equations—Theory, Computation Methods, and Applications (Aachen, 1988), Notes Numer. Fluid Mech. 24, Vieweg, Braunschweig, 1989, pp. 671–677.

[31] B. Wendroff and A. B. White, Jr., *A supraconvergent scheme for nonlinear hyperbolic systems*, Comput. Math. Appl., 18 (1989), pp. 761–767.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 604–622

INEXACT NEWTON REGULARIZATION USING CONJUGATE GRADIENTS AS INNER ITERATION*

ANDREAS RIEDER[†]

Abstract. In our papers [*Inverse Problems*, 15 (1999), pp. 309–327] and [*Numer. Math.*, 88 (2001), pp. 347–365] we proposed algorithm `REGINN`, an inexact Newton iteration for the stable solution of nonlinear ill-posed problems. `REGINN` consists of two components: the outer iteration, which is a Newton iteration stopped by the discrepancy principle, and an inner iteration, which computes the Newton correction by solving the linearized system. The convergence analysis presented in both papers covers virtually any *linear* regularization method as inner iteration, especially Landweber iteration, ν-methods, and Tikhonov–Phillips regularization. In the present paper we prove convergence rates for `REGINN` when the conjugate gradient method, which is *nonlinear*, serves as inner iteration. Thereby we add to a convergence analysis of Hanke, who had previously investigated `REGINN` furnished with the conjugate gradient method [*Numer. Funct. Anal. Optim.*, 18 (1997), pp. 971–993]. By numerical experiments we illustrate that the conjugate gradient method outperforms the ν-method as inner iteration.

Key words. nonlinear ill-posed problems, inexact Newton iteration, conjugate gradients, ν-methods, regularization

AMS subject classifications. 65J20, 65J22

DOI. 10.1137/040604029

1. Introduction.

Our goal is to find a stable approximate solution of the nonlinear ill-posed problem

$$F(x) = y^\delta, \tag{1.1}$$

where $F : \mathsf{D}(F) \subset X \to Y$ operates between the Hilbert spaces X and Y. Here, $\mathsf{D}(F)$ denotes the domain of definition of F, and y^δ is a noisy version of the exact but unknown data $y = F(x^+)$ satisfying

$$\|y - y^\delta\|_Y \leq \delta. \tag{1.2}$$

The nonnegative *noise level* δ is assumed to be known.

In [10, 11] we proposed algorithm `REGINN` for solving (1.1). As a Newton-type algorithm, `REGINN` updates the actual iterate x_n by adding a correction step s_n^δ obtained from solving a linearization of (1.1):

$$x_{n+1} = x_n + s_n^\delta, \quad n \in \mathbb{N}_0,$$

with an initial guess x_0. For obvious reasons we like to have s_n^δ as close as possible to the exact Newton step

$$s_n^{\mathrm{e}} = x^+ - x_n.$$

*Received by the editors August 2, 2004; accepted for publication (in revised form) September 27, 2004; published electronically June 30, 2005.
http://www.siam.org/journals/sinum/43-2/60402.html

[†]Institut für Praktische Mathematik und Institut für Wissenschaftliches Rechnen und Mathematische Modellbildung, Universität Karlsruhe, 76128 Karlsruhe, Germany (andreas.rieder@math.uni-karlsruhe.de, www.mathematik.uni-karlsruhe.de/∼rieder).

Assuming F to be Fréchet differentiable with derivative $F' : \mathsf{D}(F) \to \mathcal{L}(X, Y)$, the exact Newton step satisfies the linear equation

$$F'(x_n)s_n^{\mathrm{e}} = y - F(x_n) - E(x^+, x_n) =: b_n, \tag{1.3}$$

where $E(v, w) := F(v) - F(w) - F'(w)(v - w)$ is the linearization error.

Unfortunately, the above right-hand side b_n is not available; however, we know a perturbed version

$$b_n^\varepsilon := y^\delta - F(x_n) \quad \text{with } \|b_n - b_n^\varepsilon\|_Y \le \delta + \|E(x^+, x_n)\|_Y.$$

Therefore, we determine the correction step s_n^δ as a solution of

$$F'(x_n)s = b_n^\varepsilon. \tag{1.4}$$

Here, we have to take into account that the ill-posedness of (1.1) is passed on to (1.4). For instance, if F is completely continuous, then $F'(x_n)$ is a compact operator (see, e.g., Zeidler [13, Proposition 7.33]); hence, (1.4) is ill-posed.

Depending on how s_n^δ is stably obtained from (1.4), different methods arise, for instance, the nonlinear Landweber method (Hanke, Neubauer, and Scherzer [6]), the Gauß–Newton method (see, e.g., Bakushinskii [1] and Kaltenbacher [7]), and the Levenberg–Marquardt scheme (Hanke [4]).

In the next few lines we recall briefly how `REGINN` works. First, a regularization scheme is applied to the linear system (1.4), obtaining

$$s_{n,r} := g_r(A_n^* A_n)A_n^* b_n^\varepsilon,$$

where $A_n = F'(x_n)$ and $g_r : [0, \|A_n\|^2] \to \mathbb{R}$ is the piecewise continuous filter function of the chosen regularization method. The parameter $r \in \mathbb{N}$ is called the regularization parameter. For instance, the filter functions belonging to the Tikhonov–Phillips regularization, the Landweber iteration, and the ν-methods are explicitly known; see, e.g., [2, 12], where more examples can be found. The filter functions g_r of both latter examples are polynomials of degree $r - 1$. The conjugate gradients method (cg-method) can also be described by filter polynomials g_r of degree $r - 1$, which, however, do depend on the right-hand side b_n^ε: $g_r(\cdot) = g_r(\cdot, b_n^\varepsilon)$. Therefore, the cg-method is a *nonlinear* scheme in contrast to the other mentioned examples.

Now we have to select a regularization parameter r_n. In `REGINN` r_n is picked as the smallest number at which the relative (linear) residual is smaller than a given tolerance $\mu_n \in \,]0, 1]$, that is,

$$\|A_n s_{n,r_n} - b_n^\varepsilon\|_Y < \mu_n \|b_n^\varepsilon\|_Y \le \|A_n s_{n,i} - b_n^\varepsilon\|_Y, \quad i = 1, \ldots, r_n - 1. \tag{1.5}$$

The tolerances should not be too small to guarantee existence of r_n; see Lemma 2.1 below. A meaningful strategy to adapt the μ_n's dynamically was proposed in [10]. Setting $s_n^\delta := s_{n,r_n}$ we end up with the Newton iteration

$$x_{n+1} = x_n + g_{r_n}(A_n^* A_n)A_n^* b_n^\varepsilon, \quad n \in \mathbb{N}_0,$$

which has to be stopped in time to avoid noise amplification. A well-established stopping rule is the discrepancy principle: Choose $R > 0$ and accept iterate x_N as an approximate solution of (1.1) that fulfills

$$\|y^\delta - F(x_N)\|_Y \le R\,\delta < \|y^\delta - F(x_k)\|_Y, \quad k = 0, \ldots, N - 1. \tag{1.6}$$

```
REGINN($x, R, \{\mu_n\}$)
$n := 0;\ x_0 := x;$
while $\|F(x_n) - y^\delta\|_Y > R\,\delta$ do
{   $i := 0;$
    repeat
      $i := i + 1;$
      $s_{n,i} := g_i\big(F'(x_n)^*F'(x_n)\big)F'(x_n)^*\big(y^\delta - F(x_n)\big);$
    until $\|F'(x_n)\,s_{n,i} + F(x_n) - y^\delta\|_Y < \mu_n\,\|F(x_n) - y^\delta\|_Y$
    $x_{n+1} := x_n + s_{n,i};$
    $n := n + 1;$
}
$x := x_n;$
```

FIG. 1.1. *Algorithmic realization of* REGINN *(*REG*ularization based on* IN*exact* N*ewton iterations).*

For an algorithmic realization of REGINN, see Figure 1.1. The inner repeat-loop provides the Newton update s_{n,r_n} and the outer while-loop implements the Newton iteration stopped by the discrepancy principle.

In [11] we were able to verify (under reasonable assumptions) that REGINN with a *linear* regularization scheme $\{g_r\}_{r\in\mathbb{N}}$ is well defined and indeed terminates. Moreover, we proved the existence of a positive $\kappa_{\min} < 1$ such that the source condition[1]

$$x^+ - x_0 \in \mathsf{R}\big(\big|F'(x^+)\big|^\kappa\big) \qquad \text{for a } \kappa \in \,]\kappa_{\min}, 1] \tag{1.7}$$

implies the suboptimal convergence rate[2]

$$\|x^+ - x_{N(\delta)}\|_X = \mathrm{O}\big(\delta^{\,(\kappa-\kappa_{\min})/(1+\kappa)}\big) \qquad \text{as } \delta \to 0. \tag{1.8}$$

In the present paper we will improve upon the convergence results for REGINN: We will verify that (1.7) implies (1.8) even when the cg-method serves as inner iteration of REGINN. Thus we supplement a convergence analysis of Hanke [5], who had previously investigated REGINN with the cg-method as inner iteration: Under a slightly weaker version of our general assumption on the nonlinearity (see (2.1) below), Hanke proved convergence of $\{x_{N(\delta)}\}_{\delta>0}$ to a set of solutions of $F(x) = y$ as $\delta \to 0$.

This paper is structured as follows. In the next two sections we compile facts about REGINN and the cg-method which we will need later on in our analysis. In section 4 we show that REGINN is well defined under (1.7) and terminates with an approximation to x^+. Then the regularization property (1.8) will be verified (section 5). Finally, we present numerical experiments for a parameter identification model problem and end with concluding remarks in section 7. Some lengthy and technical proofs from sections 3 and 4 are shifted to Appendices A and B, respectively.

[1] By $\mathsf{R}(B)$ we denote the range of the operator B, and $|B|$ is the square root of B^*B.

[2] For linear inverse problems $Ax = y^\delta$ the regularization error cannot decrease faster than $\mathrm{O}(\delta^{\kappa/(1+\kappa)})$ as $\delta \to 0$ under the source condition $x^+ - x_0 \in \mathsf{R}(|A|^\kappa)$ in general; see, e.g., [2, section 3.2] or [12, Kapitel 3.2.3]. Regularization schemes attaining the maximal order are therefore called *order-optimal*.

2. General assumptions and termination of the `repeat`-loop. Throughout the paper we assume $F : \mathsf{D}(F) \subset X \to Y$ to be continuously Fréchet differentiable with derivative $F' : \mathsf{D}(F) \to \mathcal{L}(X, Y)$. Moreover, let $x^+ \in \mathsf{D}(F)$, $y = F(x^+)$, $y^\delta \in Y$ with $\|y - y^\delta\|_Y \le \delta$, $A = F'(x^+)$, and $A_n = F(x_n)$.

Our analysis relies heavily on the local factorization (2.1) of F': Let $Q : X \times X \to \mathcal{L}(X, Y)$ be a mapping such that

$$F'(v) = Q(v, w)\, F'(w) \quad \text{with } \|I - Q(v, w)\| \le C_Q \, \|v - w\|_X \tag{2.1}$$

for all v, $w \in B_\rho(x^+) \subset \mathsf{D}(F)$, the open ball of radius ρ about x^+. Here, C_Q is a positive constant. For a discussion of the nontrivial factorization (2.1) and for examples of meaningful operators satisfying (2.1), we refer to [6, 10, 11], [12, Kapitel 7.3], and the literature cited therein.

Let $C_Q \, \rho < 1$. Then (2.1) gives

$$\|F(v) \, - \, F(w)\|_Y \; \ge \; (1 \, - \, C_Q \, \rho) \; \|F'(w)\,(v - w)\|_Y \tag{2.2}$$

as well as

$$\|E(v, w)\|_Y \; \le \; \omega \; \|F(v) \, - \, F(w)\|_Y \quad \text{for all } \; v, w \in B_\rho(x^+), \tag{2.3}$$

where $\omega \; := \; C_Q \, \rho/(1 - C_Q \, \rho)$; see [10, section 3] or [12, Lemma 7.3.9]. Observe that $\omega < 1$ for $C_Q \, \rho < 1/2$.

In our subsequent analysis we will frequently use the following estimate: For $x, y \in B_\rho(x^+)$ and $C_Q\rho < 1/2$ we have

$$\| |F'(x)|^{-\kappa} \, |F'(y)|^{\kappa} \| \; \le \; (1 - 2C_Q\rho)^{-\kappa} =: C_{K,\kappa} \quad \text{for all } \; \kappa \in [0, 1], \tag{2.4}$$

which is due to Kaltenbacher [7, Lemma 2.2]; see also [12, Lemma 7.5.16].

Using (2.3) we will bound the data error $\|b_n^\varepsilon - b_n\|_Y$ in terms of δ, ω, and the nonlinear defect

$$d_n := \|y^\delta - F(x_n)\|_Y \; = \; \|b_n^\varepsilon\|_Y.$$

For $x_n \in B_\rho(x^+)$ we find

$$\|b_n^\varepsilon - b_n\|_Y \; \le \; (1 + \omega)\, \delta \; + \; \omega \, d_n \; := \; \varepsilon \; = \; \varepsilon(x_n, \delta).$$

We recall a result from [10] which gives conditions on μ_n to stop the `repeat`-loop.

LEMMA 2.1. *Let $\{g_r\}_{r \in \mathbb{N}}$ be the filter function of a linear or nonlinear regularization scheme for which the discrepancy principle returns a well-defined stopping index; that is, for $\tau > 1$ there exists a smallest index r_S with $\|A_n s_{n,r_S} - b_n^\varepsilon\|_Y \le \tau \, \varepsilon$. Further let (2.1) hold true with $C_Q \, \rho < 1/2$ and assume $x_n \in B_\rho(x^+)$, where $n < N$. If $R \ge (1 + \omega)/(1 - \omega)$, then the `repeat`-loop of algorithm `REGINN` terminates for any*

$$\mu_n \in \; \left]\omega + \frac{(1 + \omega)\, \delta}{d_n}, \, 1\right].$$

The lower bound on R in Lemma 2.1 guarantees that the interval for μ_n is nonempty. Since $b_n \in \mathsf{R}(A_n)$ (see (1.3)), all regularization methods mentioned in section 1 (Tikhonov–Phillips, Landweber, ν-method, cg-method) satisfy the requirement of Lemma 2.1; see, e.g., Engl, Hanke, and Neubauer [2, Chapter 4.3] or [12, Kapitel 3.4].

3. The method of conjugate gradients: Preliminaries. Here we recall some basic facts of the cg-method which we will need later in the paper. More details as well as all proofs can be found in, e.g., Engl, Hanke, and Neubauer [2, Chapter 7] or [12, Kapitel 5.3].

Let $T \in \mathcal{L}(X,Y)$ and $\boldsymbol{\eta} \in Y$. The cg-method is an iteration for solving the normal equation $T^*T\,\boldsymbol{\zeta} = T^*\boldsymbol{\eta}$. Starting with $\boldsymbol{\xi}_0 \in X$ the cg-method produces a sequence $\{\boldsymbol{\xi}_m\}_{m\in\mathbb{N}_0}$ with the minimization property

$$\|\boldsymbol{\eta} - T\boldsymbol{\xi}_m\|_Y = \min\bigl\{\|\boldsymbol{\eta} - T\boldsymbol{\xi}\|_Y \bigm| \boldsymbol{\xi} \in X,\ \boldsymbol{\xi} - \boldsymbol{\xi}_0 \in U_m\bigr\}, \quad m \ge 1,$$

where U_m is the mth Krylov space,

$$U_m := \operatorname{span}\bigl\{T^*r^0, (T^*T)T^*r^0, (T^*T)^2T^*r^0, \dots, (T^*T)^{m-1}T^*r^0\bigr\}$$

with $r^0 := \boldsymbol{\eta} - T\boldsymbol{\xi}_0$. Therefore, $\boldsymbol{\xi}_m$, $m \ge 1$, can be expressed by

$$\boldsymbol{\xi}_m = \boldsymbol{\xi}_0 + q_{m-1}(T^*T)T^*(\boldsymbol{\eta} - T\boldsymbol{\xi}_0)$$

with a polynomial q_{m-1} of degree $m-1$. Closely related to q_{m-1} is the residual polynomial $p_m(\lambda) = 1 - \lambda\, q_{m-1}(\lambda)$ of degree m satisfying

$$\boldsymbol{\eta} - T\boldsymbol{\xi}_m = p_m(TT^*)(\boldsymbol{\eta} - T\boldsymbol{\xi}_0).$$

Both polynomials depend on $\boldsymbol{\eta}$: $q_{m-1}(\cdot) = q_{m-1}(\cdot,\boldsymbol{\eta})$ and $p_m(\cdot) = p_m(\cdot,\boldsymbol{\eta})$. As soon as $T^*(\boldsymbol{\eta} - T\boldsymbol{\xi}_k) = 0$ holds true, the cg-sequence is finite, that is, $\boldsymbol{\xi}_m = \boldsymbol{\xi}_k$ for all $m \ge k$. Accordingly,

$$\boldsymbol{m}_{\mathrm{T}} := \sup\{m \in \mathbb{N} \,|\, T^*(\boldsymbol{\eta} - T\boldsymbol{\xi}_{m-1}) \ne 0\}$$

is called the ultimate termination index of the cg-method ($\boldsymbol{m}_{\mathrm{T}} = \infty$ is allowed and the supremum of the empty set is understood as zero).

The residual polynomials are orthogonal with respect to the inner product $\langle\varphi,\psi\rangle_\Pi := \langle\varphi(T^*T)T^*\boldsymbol{\eta}, \psi(T^*T)T^*\boldsymbol{\eta}\rangle_X$, which is defined on the space of all polynomials:

$$\langle p_i, p_j\rangle_\Pi = 0 \quad \text{for all } 1 \le i,j \le \boldsymbol{m}_{\mathrm{T}} \text{ with } i \ne j.$$

The orthogonality of $\{p_m\}_{1\le m\le \boldsymbol{m}_{\mathrm{T}}}$ has several consequences. The residual polynomials satisfy a three-term recursion which can be used to compute $\boldsymbol{\xi}_m$ iteratively from $\boldsymbol{\xi}_{m-1}$ in a rather cheap way; see Figure 3.1. Moreover, p_m has m simple roots $\lambda_{m,j} \in\,]0, \|T\|^2[$, $j = 1,\dots,m$, which we order by

$$0 < \lambda_{m,1} < \lambda_{m,2} < \cdots < \lambda_{m,m} < \|A\|^2.$$

Because of its normalization $p_m(0) = 1$, p_m decomposes into the following linear factors:

$$p_m(\lambda) \;=\; \prod_{j=1}^{m}(1 - \lambda/\lambda_{m,j}) \;=\; \prod_{j=1}^{m}\frac{\lambda_{m,j} - \lambda}{\lambda_{m,j}}. \tag{3.1}$$

Although we know neither q_{m-1} nor p_m explicitly, some useful information about both polynomials is available.

Lemma 3.1. *For $0 < \Lambda \le \lambda_{m,1}$ and $1 \le m \le \boldsymbol{m}_{\mathrm{T}}$, we have that*

$$\sup_{0\le\lambda\le\Lambda} |q_{m-1}(\lambda,\boldsymbol{\eta})| \;=\; q_{m-1}(0,\boldsymbol{\eta}) \;=\; -p_m'(0,\boldsymbol{\eta}) \;=\; \sum_{j=1}^{m}\lambda_{m,j}^{-1}.$$

```
cg-algorithm for $T \in \mathcal{L}(X,Y)$, $\boldsymbol{\eta} \in Y$ and starting guess $\boldsymbol{\xi}_0 \in X$.
$r^0 := \boldsymbol{\eta} - T\boldsymbol{\xi}_0$; $p^1 = a^0 := T^* r^0$;
$m := 1$;
while ($a^{m-1} \neq 0$)
    { $q^m := Tp^m$;
      $\alpha_m := \|a^{m-1}\|_X^2 / \|q^m\|_Y^2$;
      $\boldsymbol{\xi}_m := \boldsymbol{\xi}_{m-1} + \alpha_m \; p^m$;
      $r^m := r^{m-1} - \alpha_m \; q^m$;
      $a^m := T^* r^m$;
      $\beta_m := \|a^m\|_X^2 / \|a^{m-1}\|_X^2$;
      $p^{m+1} := a^m + \beta_m \; p^m$;
      $m := m+1$; }
```

FIG. 3.1. *Conjugate gradients algorithm.*

The next result is proved in Appendix A and will be used twice in our convergence analysis of REGINN with the cg-method as inner iteration.

LEMMA 3.2. *Let* $\{\boldsymbol{\xi}_m\}_{0\le m\le m_{\rm T}}$, $\boldsymbol{\xi}_0 = 0$, *be the cg-sequence with respect to* $T \in \mathcal{L}(X,Y)$ *and* $\boldsymbol{\eta} \in Y$. *Further, let* $\boldsymbol{\xi}$ *be in* $\mathsf{D}(|T|^{-\mu})$ *for a* $\mu \in [0,1]$. *Then, for any* $\nu \in [0,\mu]$, *we have that*

$$\begin{aligned} \||T|^{-\nu}(\boldsymbol{\xi}_m - \boldsymbol{\xi})\|_X \le q_{m-1}(0,\boldsymbol{\eta})^{(\nu+1)/2}\big(\|T(\boldsymbol{\xi}_m - \boldsymbol{\xi})\|_Y + \|\boldsymbol{\eta} - T\boldsymbol{\xi}\|_Y\big) \\ + \; q_{m-1}(0,\boldsymbol{\eta})^{(\nu-\mu)/2}\,\||T|^{-\mu}\boldsymbol{\xi}\|_X. \end{aligned} \tag{3.2}$$

4. Termination of REGINN with conjugate gradients.

The convergence of REGINN will be established by bounding the Newton corrections s_{n,r_n} sharply enough. Indeed, we will show that the Newton corrections decrease geometrically in n. Thus, the Newton iterates stay in a ball about x_0.

Recall the assumptions and notation from section 2 and let the cg-method be the inner iteration of REGINN exclusively throughout this section.

LEMMA 4.1. *Suppose* s_{n,r_n} *is well defined. Then*

$$\|s_{n,r_n}\|_X < 3\; q_{r_n-1}(0,b_n^\varepsilon)^{1/2}\; d_n. \tag{4.1}$$

Proof. We apply Lemma 3.2 with $T = A_n$, $\mu = \nu = 0$, $\boldsymbol{\xi} = 0$, $\boldsymbol{\eta} = b_n^\varepsilon$, $\boldsymbol{\xi}_m = s_{n,r_n} = q_{r_n-1}(A_n^*A, b_n^\varepsilon)A_n^* b_n^\varepsilon$, that is, $m = r_n$. Thus,

$$\|s_{n,r_n}\|_X \le q_{r_n-1}(0,b_n^\varepsilon)^{1/2}\,\big(\|A_n s_{n,r_n}\|_Y + \|b_n^\varepsilon\|_Y\big).$$

We are done by $\|A_n s_{n,r_n}\|_Y \le \|A_n s_{n,r_n} - b_n^\varepsilon\|_Y + \|b_n^\varepsilon\|_Y \le (\mu_n + 1)\,\|b_n^\varepsilon\|_Y$. □

In the following we bound each of the factors on the right-hand side of (4.1). From [10, Lemma 4.1] (see also [12, Lemma 7.5.9]) we already know that the nonlinear residuals d_n decrease linearly.

LEMMA 4.2. *Suppose that the nth iterate* x_n *of* REGINN *is well defined and lies in* $B_\rho(x^+)$. *Further, let* (2.3) *hold true with*

$$\omega < \eta/(2+\eta) \quad \textit{for one} \quad \eta < 1.^3$$

[3] This restriction is satisfied, for instance, if (2.1) holds true and ρ is small enough.

If, moreover,

$$R \geq \frac{1+\omega}{\eta - (2+\eta)\,\omega} \quad \text{and} \quad \mu_n \in \left]\omega + \frac{(1+\omega)\,\delta}{d_n},\ \eta - (1+\eta)\,\omega\right]$$

as well as $x_{n+1} \in B_\rho(x^+)$, *then*

$$\frac{d_{n+1}}{d_n} \;=\; \frac{\|y^\delta - F(x_{n+1})\|_Y}{\|y^\delta - F(x_n)\|_Y} \;<\; \frac{\mu_n + \omega}{1-\omega} \;\leq\; \eta.$$

4.1. Bounding $q_{r_n-1}(0, b_n^\varepsilon)$.

We assume the existence of $w \in X$ and $\kappa \in [0,1]$ such that

$$s_0^{\mathrm{e}} \;=\; x^+ - x_0 \;=\; |A|^\kappa w, \tag{4.2}$$

where $A = F'(x^+)$. To formulate the bound for $q_{r_k-1}(0, b_k^\varepsilon)$ we introduce the ratio

$$\tau_k := \mu_k\, d_k / \varepsilon(x_k, \delta), \tag{4.3}$$

which is greater than 1 under the hypotheses of Lemma 2.1.

LEMMA 4.3. *Let* (2.1) *hold true with* $C_Q\rho < 1/2$ *(thus,* $\omega < 1$ *in* (2.3)*) and assume that the first* $n < N$ *iterates* $\{x_1, \ldots, x_n\}$ *of* `REGINN` *exist and stay in* $B_\rho(x^+)$. *Further, let* $x_0 \in B_\rho(x^+)$ *satisfy the source condition* (4.2).

Then $s_k^{\mathrm{e}} = x^+ - x_k \in \mathsf{D}(|A_k|^{-\kappa})$, $0 \leq k \leq n$. *Moreover, if* $R \geq (1+\omega)/(1-\omega)$ *and if* $\mu_k \in]\omega + (1+\omega)\delta/d_k, 1]$, $0 \leq k \leq n$, *then for any* $\Theta \in]0,1[$ *such that* $\Theta \min\{\tau_0, \ldots, \tau_n\} > 1$ *we have*

$$q_{r_k-1}(0, b_k^\varepsilon)^{(\kappa+1)/2} \leq \frac{a_\Theta}{\Theta\,\tau_k - 1}\ \varepsilon(x_k, \delta)^{-1}\ \||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X, \quad 0 \leq k \leq n,$$

where a_Θ *is a positive constant depending* only *on* Θ *and* κ.

Proof. See Appendix B for the proof. □

Let us summarize what we found so far. Starting from (4.1) we are able to bound the Newton steps under the assumptions of Lemma 4.3 by

$$\|s_{k,r_k}\|_X \;<\; 3\left(\frac{a_\Theta}{\Theta\,\tau_k - 1}\right)^{1/(\kappa+1)} \varepsilon(x_k, \delta)^{-1/(\kappa+1)}\ \||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X^{1/(\kappa+1)}\ d_k. \tag{4.4}$$

4.2. Bounding $\||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X$.

Before we are able to establish termination of `REGINN` by (4.4), we have to know how $\||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X$ behaves in k. Since $s_k^{\mathrm{e}} = x^+ - x_k = x^+ - x_{k-1} - s_{k-1,r_{k-1}} = s_{k-1}^{\mathrm{e}} - s_{k-1,r_{k-1}}$ we conclude that

$$\begin{aligned} \||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X \;&=\; \||A_k|^{-\kappa}(s_{k-1}^{\mathrm{e}} - s_{k-1,r_{k-1}})\|_X \\ &\overset{(2.4)}{\leq}\; C_{K,\kappa}\, \||A_{k-1}|^{-\kappa}(s_{k-1}^{\mathrm{e}} - s_{k-1,r_{k-1}})\|_X. \end{aligned}$$

We estimate the norm on the right by applying Lemma 3.2 with $T = A_{k-1}$, $\mu = \nu = \kappa$, $\boldsymbol{\xi} = s^{\mathrm{e}}_{k-1}$ $\boldsymbol{\eta} = b^{\varepsilon}_{k-1}$, and $\boldsymbol{\xi}_m = s_{k-1,r_{k-1}}$. Hence,

$$\begin{aligned}
\||A_k|^{-\kappa} s^{\mathrm{e}}_k\|_X &\le C_{K,\kappa}\, q_{r_{k-1}-1}(0, b^{\varepsilon}_{k-1})^{(\kappa+1)/2} \big(\|A_{k-1} s_{k-1,r_{k-1}} - b_{k-1}\|_Y \\
&\qquad\qquad + \|b^{\varepsilon}_{k-1} - b_{k-1}\|_Y\big) + \||A_{k-1}|^{-\kappa} s^{\mathrm{e}}_{k-1}\|_X \\
&\le C_{K,\kappa}\, q_{r_{k-1}-1}(0, b^{\varepsilon}_{k-1})^{(\kappa+1)/2} \big(\underbrace{\mu_{k-1}\, d_{k-1}}_{\overset{(4.3)}{=}\ \tau_{k-1}\,\varepsilon(x_{k-1},\delta)} + 2\,\varepsilon(x_{k-1},\delta)\big) \\
&\qquad\qquad + \||A_{k-1}|^{-\kappa} s^{\mathrm{e}}_{k-1}\|_X \\
&\le 3\, C_{K,\kappa}\, q_{r_{k-1}-1}(0, b^{\varepsilon}_{k-1})^{(\kappa+1)/2}\ \tau_{k-1}\ \varepsilon(x_{k-1},\delta) \\
&\qquad\qquad + \||A_{k-1}|^{-\kappa} s^{\mathrm{e}}_{k-1}\|_X \\
&\le \Big(3\, C_{K,\kappa}\, a_\Theta\, \frac{\tau_{k-1}}{\Theta\,\tau_{k-1} - 1} + 1\Big)\ \||A_{k-1}|^{-\kappa} s^{\mathrm{e}}_{k-1}\|_X,
\end{aligned}$$

where we used Lemma 4.3 in the last step. Inductively, we end up with the following lemma.

LEMMA 4.4. *Let* (2.1) *hold true with* $C_Q\rho < 1/2$ *(thus,* $\omega < 1$ *in* (2.3)*) and assume that the first* $n < N$ *iterates* $\{x_1, \dots, x_n\}$ *of* `REGINN` *exist and stay in* $B_\rho(x^+)$. *Further, choose* $R \ge (1+\omega)/(1-\omega)$ *and let* $x_0 \in B_\rho(x^+)$ *satisfy the source condition* (4.2).

If $\mu_k \in\,]\omega + (1+\omega)\delta/d_k, 1]$, $0 \le k \le n$, *and if* $\Theta \in\,]0,1[$ *is such that* $\Theta\,\tau_k > 1$, $0 \le k \le n$, *then*

$$\||A_k|^{-\kappa} s^{\mathrm{e}}_k\|_X \le C_{K,\kappa}\, \Lambda_n^k\ \|w\|_X \quad \textit{for}\ \ k = 0, \dots, n \tag{4.5a}$$

with

$$\Lambda_n = 1 + 3\, C_{K,\kappa}\ a_\Theta\ \frac{t_n}{\Theta\, t_n - 1} \quad \textit{and} \quad t_n = \min\{\tau_0, \dots, \tau_n\}. \tag{4.5b}$$

4.3. Termination. We are now able to verify termination of `REGINN` with conjugate gradients as inner iteration: under reasonable technical assumptions all iterates remain in $B_\rho(x_0)$ and `REGINN` delivers an approximation $x_{N(\delta)}$ to x^+. The following theorem is the counterpart of Theorem 3.3 from [11] (see also [12, Satz 7.5.14]) and will be proved along the same lines.

THEOREM 4.5. *Let* (2.1) *hold true with* $C_Q\rho < 1/2$ *(thus,* $\omega < 1$ *in* (2.3)*). Let* $\tau > 1$ *and let* $\Theta \in\,]0,1[$ *be such that* $\Theta\,\tau > 1$. *Set*

$$\Lambda = 1 + 3\, C_{K,\kappa}\ a_\Theta\ \frac{\tau}{\Theta\,\tau - 1},$$

where $C_{K,\kappa} = (1 - C_Q\rho)^{-\kappa}$ *(see* (2.4)*), and* a_Θ *comes from the estimate in Lemma* 4.3. *Suppose that* (2.3) *is satisfied with*

$$\omega < \frac{\eta}{\eta + \tau + 1}, \quad \textit{where} \quad \eta\,\Lambda < 1.^{4}$$

Assume that the starting guess $x_0 \in B_{\rho/2}(x^+)$ *is chosen such that the source condition* (4.2) *applies for* κ *restricted to* $]\log_{1/\eta}\Lambda, 1]$ *and that the product* $\|w\|_X\, \|y -$

[4] This restriction is satisfied, for instance, if (2.1) holds true and ρ is small enough.

$F(x_0)\|_Y^\kappa$ is sufficiently small. If

$$R \geq \frac{\tau\,(1+\omega)}{\eta - \omega\,\big(\eta + \tau + 1\big)} \quad \textit{and} \quad \mu_k \in \left[\tau\,\left(\omega + \frac{(1+\omega)\,\delta}{d_k}\right),\ \eta - (1+\eta)\,\omega\right]$$

for $k \geq 0$, then there is an $N(\delta) \in \mathbb{N}$ and a $\overline{\delta} > 0$ such that all iterates $\{x_1, \dots, x_{N(\delta)}\}$ are well defined and stay in $B_\rho(x^+)$ for all noise levels $\delta \in\ ]0, \overline{\delta}]$. Moreover, the final iterate $x_{N(\delta)}$ satisfies the discrepancy principle (1.6) *and, for $d_0 > R\,\delta$,*

$$N(\delta)\ \leq\ \lfloor \log_\eta(R\,\delta/d_0) \rfloor + 1.[5] \tag{4.6}$$

Proof. We will prove Theorem 4.5 by induction. Therefore, assume that the first n iterates $\{x_0, \dots, x_n\}$ are well defined under the hypotheses of Theorem 4.5 and stay in $B_\rho(x^+)$.

If $d_n \leq R\,\delta$, the iteration will be stopped by (1.6) with $N(\delta) = n$. Otherwise, $d_n > R\,\delta$, and we show that the interval determining μ_n is not empty. The bound on ω implies that the denominator of the lower bound of R is positive. The lower bound on R guarantees that $\tau(\omega + (1+\omega)\delta/d_n) < \tau(\omega + (1+\omega)/R) < \eta - (1+\eta)\,\omega$.

According to Lemma 2.1, r_n and thus the Newton step s_{n,r_n} are well defined. By (4.4) and (4.5),

$$\|s_{n,r_n}\|_X \leq 3\,\left(\frac{C_{K,\kappa}\,a_\Theta\,\|w\|_X}{\Theta\,\tau_n - 1}\right)^{1/(\kappa+1)}\ \Lambda_n^{n/(\kappa+1)}\ \varepsilon(x_n,\delta)^{-1/(\kappa+1)}\ d_n.$$

The lower bound on the μ_k's yields $\tau_k \geq \tau > 1$, $k = 0, \dots, n$ (cf. (4.3)), that is, $\Lambda_n \leq \Lambda$. Moreover, $d_n/\varepsilon(x_n,\delta) \leq 1/\omega$. Taking Lemma 4.2 into account, we obtain

$$\|s_{n,i_n}\|_X \leq C_S\ \|w\|_X^{1/(\kappa+1)}\ d_0^{\kappa/(\kappa+1)}\ \sigma(\kappa)^n,$$

where $C_S = 3\big(C_{K,\kappa}\,a_\Theta/(\Theta\,\tau - 1)/\omega\big)^{1/(\kappa+1)}$ and

$$\sigma(\kappa) := \big(\Lambda\,\eta^\kappa\big)^{1/(\kappa+1)}\ < 1.[6] \tag{4.7}$$

We define the quantity

$$\alpha(\delta) := C_S\ \|w\|_X^{1/(\kappa+1)}\ \|F(x_0) - y^\delta\|_X^{\kappa/(\kappa+1)}\Big/\big(1 - \sigma(\kappa)\,\big). \tag{4.8}$$

In our formulation of Theorem 4.5 we assumed the product $\|w\|_X\|F(x_0) - y\|_X^\kappa$ to be sufficiently small. Now we can be more precise: assume that $\|w\|_X\|F(x_0) - y\|_X^\kappa$ is so small that $\alpha(0) < \rho/2$. Then there exists a $\overline{\delta} > 0$ yielding $\alpha(\delta) < \rho/2$ for all $\delta \in\ ]0, \overline{\delta}]$ and the new iterate $x_{n+1} = x_n + s_{n,r_n} = x_0 + \sum_{k=0}^n s_{k,r_k}$ is in $B_\rho(x^+)$:

$$\|x^+ - x_{n+1}\|_X \leq \|x^+ - x_0\|_X + \sum_{k=0}^n \|s_{k,r_k}\|_X \leq \rho/2 + \alpha(\delta) \leq \rho.$$

Further, $d_{n+1} \leq \eta^{n+1} d_0$ uniformly in $\delta \in\ ]0, \overline{\delta}]$ (Lemma 4.2). This completes the inductive step, thereby finishing the proof of Theorem 4.5. □

[5] Here, $\lfloor t \rfloor \in \mathbb{Z}$ for $t \in \mathbb{R}$ denotes the greatest integer: $\lfloor t \rfloor \leq t < \lfloor t \rfloor + 1$.

[6] Note that $\sigma(\kappa)$ is smaller than 1 since $\kappa > \log_{1/\eta}\,\Lambda$.

5. Convergence with rates.

5. Convergence with rates. Finally, we are able to verify the regularization property of `REGINN` with conjugate gradients as inner iteration, that is, we will show convergence of $x_{N(\delta)}$ to x^+ as the noise level δ decreases.

As an additional tool we will use the *interpolation inequality* (5.1): If $T \in \mathcal{L}(X,Y)$, then

$$\big\|(T^*T)^\alpha x\big\|_X \le \|(T^*T)^\beta x\|_X^{\alpha/\beta}\, \|x\|_X^{1-\alpha/\beta} \quad \text{for } \ 0 < \alpha \le \beta; \tag{5.1}$$

see, e.g., [2, 12].

Under the hypotheses of Theorem 4.5 we have to control the reconstruction error $s_k^{\mathrm{e}} = x^+ - x_k$ of the kth iterate, $0 \le k \le N(\delta)$:

$$\begin{aligned} \|s_k^{\mathrm{e}}\|_X^2 &= \langle |A_k|^\kappa s_k^{\mathrm{e}}, |A_k|^{-\kappa} s_k^{\mathrm{e}}\rangle_X \le \||A_k|^\kappa s_k^{\mathrm{e}}\|_X\, \||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X \\ &\overset{(5.1)}{\le} \||A_k| s_k^{\mathrm{e}}\|_X^\kappa\, \|s_k^{\mathrm{e}}\|_X^{1-\kappa}\, C_{K,\kappa}\, \Lambda^k\, \|w\|_X, \end{aligned}$$

where we also applied (4.5) with $\Lambda_k \le \Lambda$ to obtain the last inequality. Thus,

$$\begin{aligned} \|s_k^{\mathrm{e}}\|_X &\le C_{K,\kappa}^{1/(\kappa+1)}\, \Lambda^{k/(\kappa+1)}\, \|w\|_X^{1/(\kappa+1)}\, \|A_k s_k^{\mathrm{e}}\|_Y^{\kappa/(\kappa+1)} \\ &\overset{(2.2)}{\le} C_{K,\kappa}^{1/(\kappa+1)}\, \Lambda^{k/(\kappa+1)}\, \|w\|_X^{1/(\kappa+1)} \left(\frac{\|y - F(x_k)\|_Y}{1 - C_Q\,\rho}\right)^{\kappa/(\kappa+1)}. \end{aligned} \tag{5.2}$$

Relying on the above estimate, we are now able to copy the proof of Theorem 4.1 from [11] (see also [12, Satz 7.5.17]) to yield the announced convergence result.

THEOREM 5.1. *Adopt the assumptions of Theorem* 4.5*; especially, let the source condition* (4.2) *be satisfied with κ restricted to* $]\log_{1/\eta} \Lambda, 1]$*. Additionally, assume that* $\alpha(0) < \rho/2$ *(see* (4.8)*), as well as* $F(x_0) \ne y = F(x^+)$*. Then*

$$\|x^+ - x_{N(\delta)}\|_X \le C_\kappa\, \|w\|_X^{1/(\kappa+1)}\, \delta^{(\kappa - \log_{1/\eta}\Lambda)/(\kappa+1)} \quad \text{as } \ \delta \to 0, \tag{5.3}$$

where C_κ is a suitable constant.

In the noise-free situation, that is, $\delta = 0$, we have that

$$\|x^+ - x_k\|_X = \mathrm{O}\big(\sigma(\kappa)^k\big) \quad \text{as } \ k \to \infty$$

with $\sigma(\kappa)$ from (4.7)*.*

Proof. Plugging $k = N(\delta)$ into (5.2) and taking (1.6) into account give

$$\|s_{N(\delta)}^{\mathrm{e}}\|_X \le C_{K,\kappa}^{1/(\kappa+1)}\, \|w\|_X^{1/(\kappa+1)} \left(\frac{R+1}{1 - C_Q\,\rho}\right)^{\kappa/(\kappa+1)} \Lambda^{N(\delta)/(\kappa+1)}\, \delta^{\kappa/(\kappa+1)}.$$

Thus, (5.3) follows from (4.6). Convergence in the noise-free setting is obtained from (5.2) in combination with Lemma 4.2. □

6. Computational example.

6. Computational example. By computational experiments we will demonstrate the increase in numerical efficiency of `REGINN` when replacing the ν-method by the conjugate gradient iteration as inner iteration. We distinguish the two variants by ν-`REGINN` and cg-`REGINN`. Throughout this section let $\nu = 1$.[7]

[7] Any $\nu \ge 1$ is admissible [11, Example 2.1]. However, larger ν slow down ν-`REGINN` in the numerical computations presented here.

For our numerical experiments we select a model problem which satisfies our main assumption (2.1). We like to identify the bivariate parameter $c \geq 0$ in the two-dimensional elliptic PDE

$$\begin{aligned} -\Delta u + c\,u &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega \end{aligned} \tag{6.1}$$

from the knowledge of u in the box $\Omega =]0,1[^2$. In (6.1), $-\Delta$ is the Laplacian. Further, let f and g be continuous functions. If u has no zeros in Ω, then c can be recovered explicitly by $c = (f + \Delta u)/u$. Thus, c is uniquely determined by u but does not depend continuously on it. In the case of noise-corrupted data the inversion formula is useless. Further details about our model problem can be found in Hanke, Neubauer, and Scherzer [6, Example 4.2]. Since we already used our model problem for numerical experiments in [10, 11, 12] we will be brief in what follows.

We discretize (6.1) by finite differences with respect to the grid points $(x_i, y_j) = (i\,h, j\,h)$, $1 \leq i,j \leq m$, where $m \in \mathbb{N}$ and $h = 1/(m+1)$ is the discretization step size. Ordering the grid points lexicographically yields the $m^2 \times m^2$ linear system

$$\big(\mathbf{A} + \operatorname{diag}(\mathbf{c})\big)\,\mathbf{u} = \mathbf{f},$$

where $\mathbf{A}$ comes from the difference star of the Laplacian $-\Delta$ and where the components of $\mathbf{c} = (\mathbf{c}_1, \ldots, \mathbf{c}_{m^2})^t$ are given by $\mathbf{c}_{\ell(i,j)} = c(x_i, y_j)$ with $\ell : \{1, \ldots, m\}^2 \to \{1, \ldots, m^2\}$ denoting lexicographical ordering. The boundary values g are incorporated into the right-hand side $\mathbf{f}$. From the convergence theory for finite differences (see, e.g., Hackbusch [3]), we know that the solution $\mathbf{u}$ of the above linear system satisfies $\mathbf{u}_{\ell(i,j)} = u(x_i, y_j) + \mathrm{O}(h^2)$ as $h \to 0$ whenever u is sufficiently smooth.

In this discrete setting we like to reconstruct $\mathbf{c}$ from $\mathbf{u}$. Thus, we have to solve the nonlinear equation

$$F(\mathbf{c}) = \mathbf{u} \tag{6.2}$$

with $F : \mathbb{R}^{m^2}_{\geq 0} \to \mathbb{R}^{m^2}$, $F(\mathbf{c}) = (\mathbf{A} + \operatorname{diag}(\mathbf{c}))^{-1}\,\mathbf{f}$. The function F is differentiable with Jacobi-matrix $F'(\mathbf{c})w = -\big(\mathbf{A} + \operatorname{diag}(\mathbf{c})\big)^{-1}\big(F(\mathbf{c}) \odot w\big)$, where $\odot$ denotes componentwise multiplication of vectors. Moreover, F' can be factorized according to (2.1) in a neighborhood of any $\mathbf{c} > 0$ (componentwise), where also $F(\mathbf{c}) > 0$.[8] In our numerical experiments the parameter to be identified is

$$c^+(x,y) = 1.5\,\sin(2\pi\,x)\,\sin(3\pi\,y) + 3\,\big((x-0.5)^2 + (y-0.5)^2\big) + 2.$$

We have chosen f and g such that $u(x,y) = 16\,x\,(x-1)\,y\,(1-y) + 1$ solves (6.1) with respect to c^+. As perturbed right-hand side $\mathbf{u}^\delta$ for (6.2) we worked with $\mathbf{u}^\delta = \widetilde{\mathbf{u}} + \delta\,\mathbf{v}$, where $\widetilde{\mathbf{u}}_{\ell(i,j)} = u(x_i, y_j)$ and $\mathbf{v} = \mathbf{z}/\|\mathbf{z}\|_h$. The entries of the random vector $\mathbf{z}$ are uniformly distributed in $[-1,1]$. Therefore, $\|\mathbf{u} - \mathbf{u}^\delta\|_h \leq \delta + \mathrm{O}(h^2)$ measured in the weighted Euclidean norm $\|\mathbf{z}\|_h = h\,\big(\sum_{i=0}^{m^2} \mathbf{z}_i^2\big)^{1/2}$, which approximates the L^2-norm over Ω.

In all computations below we started `REGINN` with initial guess $\mathbf{c}_0$, where $(\mathbf{c}_0)_{\ell(i,j)} = c_0(x_i, x_j)$ and

$$c_0(x,y) = 3\,\big((x-0.5)^2 + (y-0.5)^2\big) + 2 + 128\,x\,(x-1)\,y\,(1-y).$$

[8] The implication $\mathbf{c} > 0 \Rightarrow F(\mathbf{c}) > 0$ holds true, for instance, if $\mathbf{f} > 0$ since $\mathbf{A} + \operatorname{diag}(\mathbf{c})$ is an M-matrix whose inverse has only nonnegative entries.

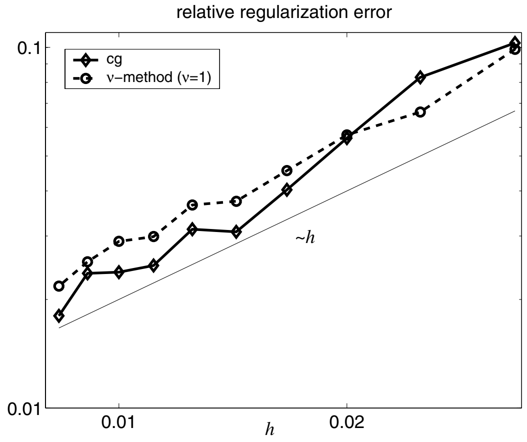


FIG. 6.1. *Relative reconstruction error* (6.4) *of* REGINN *as a function of* h *(solid line with* $\diamond$*: cg-method as inner iteration; dashed line with* $\circ$*:* ν*-method as inner iteration). The thin solid line indicates decay* $\mathrm{O}(h)$ *as* $h \to 0$.

Further, we always used $R = 1.5$, and we adapted the tolerances $\{\mu_k\}$ in (1.5) dynamically according to scheme (6.3) below, which was proposed in [10, section 6] (see also [12, Kapitel 7.5.3.4]): Initialize $\mu_{\text{start}} = 0.1$, $\gamma = 0.9$, $\mu_{\max} = 0.999$, and $\widetilde{\mu}_0 = \widetilde{\mu}_1 := \mu_{\text{start}}$. For $k = 0, \dots, N(\delta) - 1$ set

$$\mu_k := \mu_{\max} \max\big\{R\,\delta/\|F(\mathbf{c}_k) - \mathbf{u}^\delta\|_h,\, \widetilde{\mu}_k\big\}, \tag{6.3}$$

where $\mathbf{c}_k$ is the kth iterate and

$$\widetilde{\mu}_k := \begin{cases} 1 - \frac{r_{k-2}}{r_{k-1}}\,(1 - \mu_{k-1}) & : \quad r_{k-1} \geq r_{k-2}, \\ \gamma\,\mu_{k-1} & : \quad \text{else.} \end{cases}$$

Figure 6.1 shows relative reconstruction errors by ν-REGINN and cg-REGINN. More precisely, we plotted

$$\operatorname{err}(h) := \|\mathbf{c}_{N(\delta(h))} - \mathbf{c}^+\|_h / \|\mathbf{c}^+\|_h \quad \text{with } \delta(h) = 10\,h^2 \tag{6.4}$$

as a function of $h \in \{(10\,k)^{-1} \mid k = 3, \dots, 12\}$, where $\mathbf{c}^+_{\ell(i,j)} = c^+(x_i, y_j)$ and where $\mathbf{c}_{N(\delta(h))}$ is the output of either ν-REGINN or cg-REGINN. The auxiliary line in Figure 6.1 represents exact decay $\mathrm{O}(h)$ as $h \to 0$. Thus, our computations indicate that $\operatorname{err}(h) = \mathrm{O}(h)$ as $h \to 0$. Since $\|\mathbf{u} - \mathbf{u}^{\delta(h)}\|_h \leq \delta(h) + \mathrm{O}(h^2) := \boldsymbol{\delta}(h) = \mathrm{O}(h^2)$ the regularization error achieves maximal order of convergence according to Theorem 5.1, namely, $\operatorname{err}(\boldsymbol{\delta}) = \mathrm{O}(\boldsymbol{\delta}^{1/2})$ as $\boldsymbol{\delta} \to 0$, that is, $\kappa = 1$ and $\Lambda = 1$.

Both variants of REGINN deliver errors in comparable magnitude. In this respect there is not much difference between cg-REGINN and ν-REGINN. However, looking at the numerical efficiency we observe a significant difference. In Figure 6.2 we plotted the ratio

$$q(h) := \frac{\text{cpu-time of } \nu\text{-REGINN to compute } \mathbf{c}_{N(\delta(h))}}{\text{cpu-time of cg-REGINN to compute } \mathbf{c}_{N(\delta(h))}}, \tag{6.5}$$

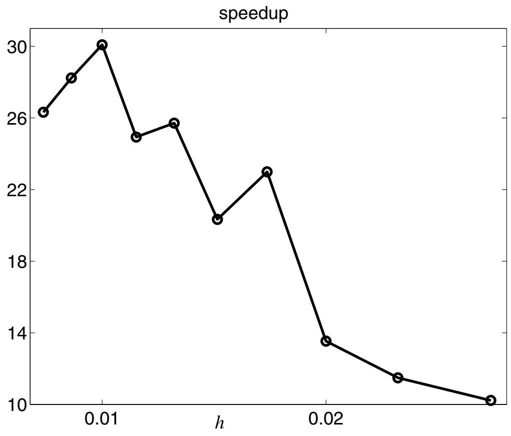


FIG. 6.2. *Speedup q (6.5) of cg-*`REGINN` *over ν-*`REGINN`.

TABLE 6.1
Convergence history of ν-`REGINN` *for $h = 0.01$, $\delta(h) = 10\,h^2$, with respect to the tolerance selection* (6.3)*, where $\mu_{\text{start}} = 0.1$, $\mu_{\max} = 0.999$, and $\gamma = 0.9$. Overall cpu-time:* 240.72 *seconds.*

k	μ_{k-1}	r_{k-1}	d_k/d_{k-1}	$d_k/(R\,\delta)$	e_k
1	0.0999	40	0.2464	43.68	0.4280
2	0.0999	75	0.1128	4.925	0.2204
3	0.5194	99	0.5153	2.538	0.1311
4	0.6353	105	0.6324	1.605	0.0836
5	0.6555	176	0.6550	1.051	0.0443
6	0.9504	242	0.9503	0.999	0.0290

TABLE 6.2
Convergence history of cg-`REGINN` *for $h = 0.01$, $\delta(h) = 10\,h^2$, with respect to the tolerance selection* (6.3)*, where $\mu_{\text{start}} = 0.1$, $\mu_{\max} = 0.999$, and $\gamma = 0.9$. Overall cpu-time:* 8.00 *seconds.*

k	μ_{k-1}	r_{k-1}	d_k/d_{k-1}	$d_k/(R\,\delta)$	e_k
1	0.0999	1	0.3997	70.86	0.5598
2	0.0999	1	0.1188	8.418	0.3090
3	0.1187	6	0.1223	1.030	0.0239
4	0.9704	1	0.9629	0.991	0.0238

where we did not count cpu-time for preprocessing steps performed by both variants.[9] Figure 6.2 reveals that cg-`REGINN` is 10 to 30 times faster than ν-`REGINN` in our example.

Tables 6.1 and 6.2 record the convergence history of ν-`REGINN` and cg-`REGINN` in full detail for the discretization step size $h = 0.01$. In both tables

$$d_k := \|F(\mathbf{c}_k) - \mathbf{u}^{\delta(h)}\|_h \quad \text{and} \quad e_k := \|\mathbf{c}_k - \mathbf{c}^+\|_h / \|\mathbf{c}^+\|_h$$

denote the nonlinear defect and the relative L^2-error of the kth iterate, respectively.

Among all Krylov-subspace methods the conjugate gradient iteration is the most efficient one when the discrepancy principle is the used stopping rule; see, e.g., [2,

[9] The experiments were carried out under MATLAB 6.5 on a 2.6GHz Intel Pentium 4 processor.

Chapter 7.1] or [12, Kapitel 5.3.6]. As expected, cg-`REGINN` outperforms ν-`REGINN` since it takes much fewer inner iterations to yield the correction step which we can observe clearly in the tables (one iteration step of the cg-methods is only slightly more expensive than one iteration step of the ν-method).

7. Concluding remarks.

In this paper we proved local convergence with rates for a regularization scheme of inexact Newton type with the cg-method as inner iteration. Theoretical aspects are emphasized; ideas and techniques have been presented to cope with the nonlinearity of the conjugate gradient iteration.

As far as the author knows, the restrictive factorization assumption (2.1) has not been verified for real applications such as, e.g., impedance tomography, ultrasound tomography, and SPECT (single photon emission computed tomography). Therefore the most pressing improvement of the presented analysis is to weaken or to get rid of (2.1).

Nevertheless the practitioner may benefit from our theoretical results in at least two ways: (1) The adaptive tolerance selection scheme (6.3) has a sound justification for cg-`REGINN` and can be expected to perform well also for more general nonlinearities. (2) A potential convergence analysis of cg-`REGINN` for a specific application, which does not fall into the general category considered, can be based upon techniques developed here.

Appendix A. Proof of Lemma 3.2.

For the sake of simplicity we only prove Lemma 3.2 for a compact operator T (the general result will follow by integration over the spectral family of T^*T). Most of our arguments have been used before by Plato [9, Lemma 5.4] (see also [12, Lemma 5.3.11]) to prove another error estimate for the cg-method.

Let $\{(\sigma_j; v_j, u_j) | j \in \mathbb{N}\} \subset \,]0,\infty[\, \times X \times Y$ be the singular system of T, that is, $Tx = \sum_{j=1}^{\infty} \sigma_j \langle x, v_j \rangle_X u_j$ with $\lim_{j\to\infty} \sigma_j = 0$ monotonically, and $\{v_j\}$ and $\{u_j\}$ are orthonormal bases in $\mathsf{N}(T)^\perp$ and $\overline{\mathsf{R}(T)}$, respectively.[10]

We introduce the spectral family $\{\mathcal{E}_\Lambda\}_{\Lambda>0} \subset \mathcal{L}(X)$ of T^*T by[11]

$$\mathcal{E}_\Lambda x := \sum_{j\in\mathcal{J}(\Lambda)} \langle x, v_j \rangle_X v_j + P_{\mathsf{N}(T)} x, \qquad \mathcal{J}(\Lambda) := \{j \in \mathbb{N} \,|\, \sigma_j^2 \le \Lambda\}, \tag{A.1}$$

and start with

$$\||T|^{-\nu}(\boldsymbol{\xi}_m - \boldsymbol{\xi})\|_X \le \||T|^{-\nu}(I - \mathcal{E}_\Lambda)(\boldsymbol{\xi}_m - \boldsymbol{\xi})\|_X + \||T|^{-\nu}\mathcal{E}_\Lambda(\boldsymbol{\xi}_m - \boldsymbol{\xi})\|_X.$$

We proceed by

$$\begin{aligned}\||T|^{-\nu}(I - \mathcal{E}_\Lambda)(\boldsymbol{\xi}_m - \boldsymbol{\xi})\|_X^2 &= \sum_{k=1}^{\infty} \sigma_k^{-2(\nu+1)} \left|\left\langle |T|(I - \mathcal{E}_\Lambda)(\boldsymbol{\xi}_m - \boldsymbol{\xi}), v_k \right\rangle_X\right|^2 \\ &= \sum_{k\notin\mathcal{J}(\Lambda)} \sigma_k^{-2(\nu+1)} \left|\left\langle |T|(\boldsymbol{\xi}_m - \boldsymbol{\xi}), v_k \right\rangle_X\right|^2 \\ &\le \Lambda^{-(\nu+1)} \|T(\boldsymbol{\xi}_m - \boldsymbol{\xi})\|_Y^2\end{aligned}$$

[10] $\mathsf{N}(B)$ and $\mathsf{R}(B)$ denote the null space and the range of a linear operator, respectively.

[11] $P_M \in \mathcal{L}(X)$ denotes the orthogonal projection onto the closed subspace M of X.

and

$$
\begin{aligned}
&\||T|^{-\nu}\mathcal{E}_\Lambda(\boldsymbol{\xi}_m-\boldsymbol{\xi})\|_X\\
&\quad\le \||T|^{-\nu}\mathcal{E}_\Lambda p_m(T^*T,\boldsymbol{\eta})\boldsymbol{\xi}\|_X+\||T|^{-\nu}\mathcal{E}_\Lambda q_{m-1}(T^*T,\boldsymbol{\eta})T^*(\boldsymbol{\eta}-T\boldsymbol{\xi})\|_X\\
&\quad\le \|\mathcal{E}_\Lambda|T|^{\mu-\nu}p_m(T^*T,\boldsymbol{\eta})\|_X\,\||T|^{-\mu}\boldsymbol{\xi}\|_X\\
&\qquad\qquad\qquad+\||T|^{-\nu}\mathcal{E}_\Lambda q_{m-1}(T^*T,\boldsymbol{\eta})T^*\|_X\,\|\boldsymbol{\eta}-T\boldsymbol{\xi}\|_X\\
&\quad\le \||T|^{-\mu}\boldsymbol{\xi}\|_X\sup_{0\le\lambda\le\Lambda}\lambda^{(\mu-\nu)/2}\,|p_m(\lambda,\boldsymbol{\eta})|\\
&\qquad\qquad\qquad+\|\mathcal{E}_\Lambda q_{m-1}(T^*T,\boldsymbol{\eta})|T|^{-\nu}T^*\|_X\,\|\boldsymbol{\eta}-T\boldsymbol{\xi}\|_X.
\end{aligned}
$$

Further,

$$
\begin{aligned}
&\|\mathcal{E}_\Lambda q_{m-1}(T^*T,\boldsymbol{\eta})|T|^{-\nu}T^*\|^2\\
&\qquad=\|\mathcal{E}_\Lambda\,q_{m-1}(T^*T,\boldsymbol{\eta})|T|^{2(1-\nu)}q_{m-1}(T^*T,\boldsymbol{\eta})\mathcal{E}_\Lambda\|\\
&\qquad\le\|\mathcal{E}_\Lambda\,q_{m-1}(T^*T,\boldsymbol{\eta})\|\;\||T|^{2(1-\nu)}q_{m-1}(T^*T,\boldsymbol{\eta})\mathcal{E}_\Lambda\|\\
&\qquad\le\sup_{0\le\lambda\le\Lambda}|q_{m-1}(\lambda,\boldsymbol{\eta})|\sup_{0\le\lambda\le\Lambda}\lambda^{1-\nu}\,|q_{m-1}(\lambda,\boldsymbol{\eta})|.
\end{aligned}
$$

By Lemma 3.1 we have

$$
\sup_{0\le\lambda\le\Lambda}|q_{m-1}(\lambda,\boldsymbol{\eta})|\le q_{m-1}(0,\boldsymbol{\eta})\quad\text{whenever }0<\Lambda\le\lambda_{m,1}.
$$

The representation (3.1) of p_m shows that $0\le p_m(\lambda)\le 1$ for $\lambda\in[0,\lambda_{m,1}]$. Since $p_m(\lambda)=1-\lambda\,q_{m-1}(\lambda)$ we derive that

$$
\begin{aligned}
\sup_{0\le\lambda\le\Lambda}\lambda^{1-\nu}\,|q_{m-1}(\lambda,\boldsymbol{\eta})|&\le\big(\sup_{0\le\lambda\le\Lambda}\lambda\,|q_{m-1}(\lambda,\boldsymbol{\eta})|\big)^{1-\nu}\big(\sup_{0\le\lambda\le\Lambda}|q_{m-1}(\lambda,\boldsymbol{\eta})|\big)^{\nu}\\
&\le q_{m-1}(0,\boldsymbol{\eta})^{\nu}
\end{aligned}
$$

whenever $0<\Lambda\le\lambda_{m,1}$. Finally, for $0<\Lambda\le\lambda_{m,1}$, we obtain that

$$
\begin{aligned}
\||T|^{-\nu}(\boldsymbol{\xi}_m-\boldsymbol{\xi})\|_X\le\Lambda^{-(\nu+1)/2}\,\|T(\boldsymbol{\xi}_m-\boldsymbol{\xi})\|_Y+\||T|^{-\mu}\boldsymbol{\xi}\|_X\,\Lambda^{(\mu-\nu)/2}&\\
+\,q_{m-1}(0,\boldsymbol{\eta})^{(\nu+1)/2}\,\|\boldsymbol{\eta}-T\boldsymbol{\xi}\|_X&,
\end{aligned}
$$

which yields the stated inequality (3.2) by setting $\Lambda=1/q_{m-1}(0,\boldsymbol{\eta})$. This choice for Λ is admissible since $1/q_{m-1}(0,\boldsymbol{\eta})<\lambda_{m,1}$; see Lemma 3.1.

Appendix B. Proof of Lemma 4.3. Before we can prove Lemma 4.3, we need some auxiliary results (Lemmas B.1 and B.2 and Corollary B.3 below). Here we rely on arguments laid out by Plato [9, Kapitel 5] and Nemirovskii [8] (see also [12, Kapitel 5.3]).

Suppose that the first n iterates $\{x_1,\ldots,x_n\}$ of `REGINN` exist and stay in $B_\rho(x^+)$. Point of departure is the inequality

$$
\text{(B.1)}\qquad\|b_k^\varepsilon-A_k s_{k,m}\|_Y\le\|\mathcal{F}_{\lambda_{m,1}}\varphi_m(A_kA_k^*,b_k^\varepsilon)b_k\|_Y+\varepsilon,\quad 1\le m\le\boldsymbol{m}_{\mathrm{T}},
$$

where $0\le k\le n$, $\boldsymbol{m}_{\mathrm{T}}=\boldsymbol{m}_{\mathrm{T}}(k)$, and $\mathcal{F}_\Lambda\in\mathcal{L}(Y)$, $\Lambda>0$, is the orthogonal projection

$$
\text{(B.2)}\qquad\mathcal{F}_\Lambda y:=\sum_{i\in\mathcal{J}(\Lambda)}\langle y,u_i\rangle_Y\,u_i+P_{\mathsf{N}(A_k^*)}y
$$

with index set $\mathscr{J}$ as in (A.1). In defining $\mathcal{F}_\Lambda$ we used the singular system of A_k.[12] Further, the function $\varphi_m(\cdot, b_k^\varepsilon) \in \mathcal{C}(\mathbb{R})$ in (B.1) is

$$\varphi_m(\lambda, b_k^\varepsilon) := p_m(\lambda, b_k^\varepsilon)\,\Big(\frac{\lambda_{m,1}}{\lambda_{m,1}-\lambda}\Big)^{1/2},$$

where $\lambda_{m,1}$ is the smallest zero of the mth residual polynomial $p_m(\cdot, b_k^\varepsilon)$ of the cg-method with respect to A_k and b_k^ε. A proof of (B.1) is presented, e.g., by Engl, Hanke, and Neubauer [2, Proof of Theorem 7.10].

As $s_k^{\mathrm{e}} = x^+ - x_k = s_0^{\mathrm{e}} - \sum_{j=0}^{k-1} s_{j,r_j}$ we obtain

$$s_k^{\mathrm{e}} \overset{(4.2)}{=} |A|^\kappa w - \sum_{j=0}^{k-1} A_j^*\, q_{r_j-1}(A_jA_j^*, b_j^\varepsilon)\, b_j^\varepsilon.$$

Note that $s_k^{\mathrm{e}} \in \mathsf{D}(|A_i|^{-\kappa})$, $i = 0, \dots, n$. Indeed, in using $A_j = Q_{j,i}A_i$ with $Q_{j,i} = Q(x_j, x_i)$ (see (2.1)), we obtain that

$$\begin{aligned} \||A_i|^{-\kappa} s_k^{\mathrm{e}}\|_X \;&\leq\; \||A_i|^{-\kappa}|A|^\kappa w\|_X + \sum_{j=0}^{k-1} \||A_i|^{-\kappa} A_j^*\, q_{r_j-1}(A_jA_j^*, b_j^\varepsilon)\, b_j^\varepsilon\|_X \\ &\overset{(2.4)}{\leq}\; C_{K,\kappa}\,\|w\|_X + \underbrace{\||A_i|^{-\kappa}A_i^*\|}_{=\|A_i\|^{1-\kappa}} \sum_{j=0}^{k-1} \|Q_{j,i}^* q_{r_j-1}(A_jA_j^*, b_j^\varepsilon)\, b_j^\varepsilon\|_Y. \end{aligned}$$

Thus, by $A_k s_k^{\mathrm{e}} = b_k$,

$$\begin{aligned} &\|\mathcal{F}_{\lambda_{m,1}}\varphi_m(A_kA_k^*, b_k^\varepsilon) b_k\|_Y \\ &\qquad = \|\mathcal{F}_{\lambda_{m,1}}\varphi_m(A_kA_k^*, b_k^\varepsilon) A_k|A_k|^\kappa|A_k|^{-\kappa} s_k^{\mathrm{e}}\|_Y \\ &\qquad \leq \sup_{0\leq\lambda\leq\lambda_{m,1}} \lambda^{(\kappa+1)/2}\,\varphi_m(\lambda, b_k^\varepsilon)\; \||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X. \end{aligned} \tag{B.3}$$

Techniques from elementary calculus, together with Lemma 3.1, yield

$$\sup_{0\leq\lambda\leq\lambda_{m,1}} \lambda^{(\kappa+1)/2}\,\varphi_m(\lambda, b_k^\varepsilon) \leq 2\, q_{m-1}(0, b_k^\varepsilon)^{-(\kappa+1)/2}$$

for $\kappa \in [0,1]$; see, e.g., [2, (7.8)] or [12, (5.65)]. Hence,

$$\|\mathcal{F}_{\lambda_{m,1}}\varphi_m(A_kA_k^*, b_k^\varepsilon) b_k\|_Y \leq 2\, q_{m-1}(0, b_k^\varepsilon)^{-(\kappa+1)/2}\, \||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X. \tag{B.4}$$

Finally, (B.1) and (B.4) yield the following lemma.

LEMMA B.1. *Let* (2.1) *hold true and assume that the first n iterates $\{x_1, \dots, x_n\}$ of* `REGINN` *exist and stay in $B_\rho(x^+)$. If $x_0 \in B_\rho(x^+)$ satisfies* (4.2), *then, for $0 \leq k \leq n$ and $1 \leq m \leq \boldsymbol{m}_{\mathrm{T}}(k)$,*

$$\|b_k^\varepsilon - A_k s_{k,m}\|_Y \leq \varepsilon + 2\, q_{m-1}(0, b_k^\varepsilon)^{-(\kappa+1)/2}\; \||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X.$$

We need a second auxiliary lemma.

[12] More precisely, $\{\mathcal{F}_\Lambda\}_{\Lambda>0}$ is the spectral family of $A_kA_k^*$.

LEMMA B.2. *Let* (2.1) *hold true and assume that the first* n *iterates* $\{x_1, \ldots, x_n\}$ *of* `REGINN` *exist and stay in* $B_\rho(x^+)$. *Further, let* $x_0 \in B_\rho(x^+)$ *satisfy* (4.2). *Choose* $\vartheta > 2$ *and* $2 < r \leq 2(\vartheta - 1)$. *Let* $0 \leq k \leq n$ *and* $1 \leq m \leq \boldsymbol{m}_{\mathrm{T}}(k)$.

If $\vartheta\, q_{m-2}(0, b_k^\varepsilon) \leq q_{m-1}(0, b_k^\varepsilon)$, *then*

$$\frac{r-2}{r-1}\ \|b_k^\varepsilon - A_k s_{k,m-1}\|_Y \leq \varepsilon + \alpha^{(\kappa+1)/2}\, q_{m-1}(0, b_k^\varepsilon)^{-(\kappa+1)/2}\ \||A_k|^{-\kappa} s_k^{\mathrm{e}}\|_X,$$

where $\alpha = r/(1 - \vartheta^{-1})$.

Proof. Under our assumptions Plato [9, (5.16)] (see also [12, (5.67)]) established the bound

$$\frac{r-2}{r-1}\ \|b_k^\varepsilon - A_k s_{k,m-1}\|_Y \leq \|\mathcal{F}_{r\lambda_{m,1}} p(A_k A_k^*) b_k^\varepsilon\|_Y$$

with the polynomial $p(\lambda) = p_m(\lambda, b_k^\varepsilon)/(1 - \lambda/\lambda_{m,1})$ of degree $m - 1$. The triangle inequality leads to

$$\|\mathcal{F}_{r\lambda_{m,1}} p(A_k A_k^*) b_k^\varepsilon\|_Y \leq \varepsilon \sup_{0 \leq \lambda \leq r\lambda_{m,1}} |p(\lambda)| + \|\mathcal{F}_{r\lambda_{m,1}} p(A_k A_k^*) b_k\|_Y.$$

To bound $\|\mathcal{F}_{r\lambda_{m,1}} p(A_k A_k^*) b_k\|_Y$ we are able to apply exactly the same arguments as were used in estimating $\|\mathcal{F}_{\lambda_{m,1}} \varphi_m(A_k A_k^*) b_k\|_Y$; cf. (B.3). Accordingly, if

$$\sup_{0 \leq \lambda \leq r\lambda_{m,1}} |p(\lambda)| \leq 1 \tag{B.5a}$$

as well as

$$\sup_{0 \leq \lambda \leq r\lambda_{m,1}} \lambda^{(\kappa+1)/2}\, |p(\lambda)| \leq \alpha^{(\kappa+1)/2}\, q_{m-1}(0, b_k^\varepsilon)^{-(\kappa+1)/2}, \tag{B.5b}$$

then Lemma B.2 is true. Assume, for the moment, that

$$r\lambda_{m,1} < 2\,\lambda_{m,2} \quad \text{and} \quad \lambda_{m,1} \leq q_{m-1}(0, b_k^\varepsilon)^{-1}/(1 - \vartheta^{-1}) \tag{B.6}$$

hold true. The left inequality yields $r\lambda_{m,1}/\lambda_{m,j} < 2$, $j = 2, \ldots, m$, whence

$$|p(\lambda)| = \prod_{j=2}^{m} |1 - \lambda/\lambda_{m,j}| < 1 \quad \text{for } 0 < \lambda \leq r\lambda_{m,1}.$$

Therefore, (B.6) implies (B.5) and we are left with verifying (B.6).

First we look into the estimate on the right in (B.6). Since the residual polynomials $\{p_m(\cdot, b_k^\varepsilon)\}_{1 \leq m \leq \boldsymbol{m}_{\mathrm{T}}}$ are orthogonal, their zeros interlace, that is, $\lambda_{m-1,j} < \lambda_{m,j+1}$, $j = 1, \ldots, m-1$. By Lemma 3.1 we therefore have

$$\begin{aligned} q_{m-1}(0, b_k^\varepsilon) &= \lambda_{m,1}^{-1} + \sum_{j=1}^{m-1} \lambda_{m,j+1}^{-1} \\ &< \lambda_{m,1}^{-1} + \sum_{j=1}^{m-1} \lambda_{m-1,j}^{-1} = \lambda_{m,1}^{-1} + q_{m-2}(0, b_k^\varepsilon). \end{aligned} \tag{B.7}$$

The hypothesis $q_{m-2}(0, b_k^\varepsilon) \le \vartheta^{-1}\, q_{m-1}(0, b_k^\varepsilon)$ implies $q_{m-1}(0, b_k^\varepsilon) \le \lambda_{m,1}^{-1} + \vartheta^{-1} q_{m-1}(0, b_k^\varepsilon)$ and thus the right inequality in (B.6). Since

$$\begin{aligned}(\vartheta - 1)\ \lambda_{m-1,1}^{-1} \quad &< \quad (\vartheta - 1) \sum_{j=1}^{m-1} \lambda_{m-1,j}^{-1} \ = \ \vartheta\ q_{m-2}(0, b_k^\varepsilon) - q_{m-2}(0, b_k^\varepsilon) \\ &\overset{\substack{\text{by} \\ \text{assumpt.}}}{\le} \quad q_{m-1}(0, b_k^\varepsilon) - q_{m-2}(0, b_k^\varepsilon) \ \overset{\text{(B.7)}}{<} \ \lambda_{m,1}^{-1},\end{aligned}$$

we obtain $\lambda_{m,1} < (\vartheta - 1)^{-1}\, \lambda_{m-1,1} < (\vartheta - 1)^{-1}\, \lambda_{m,2}$ (interlacing property). In view of $r/(\vartheta - 1) \le 2$ we conclude that the left inequality in (B.6) holds true as well, thereby finishing the proof of Lemma B.2. $\square$

Both latter lemmas merge in the next corollary.

COROLLARY B.3. *Let* (2.1) *hold true and assume the first* n *iterates* $\{x_1, \ldots, x_n\}$ *of* REGINN *exist and stay in* $B_\rho(x^+)$. *Further, let* $x_0 \in B_\rho(x^+)$ *satisfy* (4.2). *Then, to any* $\Theta \in\,]0, 1[$, *there exists a number* a_Θ *such that, for* $0 \le k \le n$,

$$\Theta\, \| b_k^\varepsilon - A_k s_{k,m-1} \|_Y \le \varepsilon + a_\Theta\, q_{m-1}(0, b_k^\varepsilon)^{-(\kappa+1)/2}\ \| |A_k|^{-\kappa} s_k^{\mathrm{e}} \|_X ,$$

where $1 \le m \le \boldsymbol{m}_{\mathrm{T}}(k)$. *The number* a_Θ only *depends on* Θ *and* κ.

Proof. There is exactly one $r = r(\Theta) > 2$ such that $\Theta = \frac{r-2}{r-1}$. Let this r be fixed and define $\vartheta = r/2 + 1 > 2$, that is, $r = 2(\vartheta - 1)$. Exactly one of the following two cases holds true.

1. In the case of $\vartheta\, q_{m-2}(0, b_k^\varepsilon) \le q_{m-1}(0, b_k^\varepsilon)$ the assertion follows immediately from Lemma B.2 when setting

$$a_{\Theta,1} := \Big(\frac{r}{1 - \vartheta^{-1}} \Big)^{(\kappa+1)/2} = \Big(\frac{r}{1 - (r/2 + 1)^{-1}} \Big)^{(\kappa+1)/2} .$$

2. In the case of $q_{m-1}(0, b_k^\varepsilon) < \vartheta\, q_{m-2}(0, b_k^\varepsilon)$ we argue with Lemma B.1. Since $\Theta < 1$ and $q_{m-2}(0, b_k^\varepsilon)^{-1} < \vartheta\, q_{m-1}(0, b_k^\varepsilon)^{-1}$ we have

$$\begin{aligned}\Theta\, \| b_k^\varepsilon - A s_{k,m-1} \|_Y &\le \varepsilon + 2\, q_{m-2}(0, b_k^\varepsilon)^{-(\kappa+1)/2}\ \| |A_k|^{-\kappa} s_k^{\mathrm{e}} \|_X \\ &< \varepsilon + a_{\Theta,2}\, q_{m-1}(0, b_k^\varepsilon)^{-(\kappa+1)/2}\ \| |A_k|^{-\kappa} s_k^{\mathrm{e}} \|_X\end{aligned}$$

with

$$a_{\Theta,2} := 2\ \vartheta^{(\kappa+1)/2} = 2\, (r/2 + 1)^{(\kappa+1)/2} .$$

So, the assertion of Corollary B.3 is verified with $a_\Theta = \max\{a_{\Theta,1}, a_{\Theta,2}\}$. $\square$

Finally we are in a position to verify Lemma 4.3: The μ_k's and R satisfy the requirements of Lemma 2.1. Hence, $\tau_k > 1$ (see (4.3)), and $\tau_k\, \varepsilon \le \| b_k^\varepsilon - A_k s_{k, r_k - 1} \|_Y$ (see (1.5)). Since $1 \le r_k \le \boldsymbol{m}_{\mathrm{T}}(k)$ we obtain

$$\Theta\, \tau_k\, \varepsilon \le \Theta\, \| b_k^\varepsilon - A_k s_{k, r_k - 1} \|_Y \le \varepsilon + a_\Theta\, q_{r_k - 1}(0, b_k^\varepsilon)^{-(\kappa+1)/2}\ \| |A_k|^{-\kappa} s_k^{\mathrm{e}} \|_X$$

by Corollary B.3. A simple rearrangement of terms yields the assertion of Lemma 4.3.

REFERENCES

[1] A. B. BAKUSHINSKII, *The problem of the convergence of the iteratively regularized Gauss-Newton method*, Comput. Math. Phys., 32 (1992), pp. 1353–1359.

[2] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Math. Appl. 375, Kluwer Academic, Dordrecht, The Netherlands, 1996.

[3] W. Hackbusch, *Elliptic Differential Equations: Theory and Numerical Treatment*, Springer Ser. Comput. Math. 18, Springer-Verlag, Heidelberg, 1992.

[4] M. Hanke, *A regularizing Levenberg-Marquardt scheme, with applications to inverse groundwater filtration problems*, Inverse Problems, 13 (1997), pp. 79–95.

[5] M. Hanke, *Regularizing properties of a truncated Newton-CG algorithm for non-linear inverse problems*, Numer. Funct. Anal. Optim., 18 (1997), pp. 971–993.

[6] M. Hanke, A. Neubauer, and O. Scherzer, *A convergence analysis of the Landweber iteration for nonlinear ill-posed problems*, Numer. Math., 72 (1995), pp. 21–37.

[7] B. Kaltenbacher, *A posteriori parameter choice strategies for some Newton type methods for the regularization of nonlinear ill-posed problems*, Numer. Math., 79 (1998), pp. 501–528.

[8] A. S. Nemirovskii, *The regularizing properties of the adjoint gradient method in ill-posed problems*, USSR Comp. Math. Math. Phys., 26 (1986), pp. 7–16.

[9] R. Plato, *Über die Diskretisierung und Regularisierung schlecht gestellter Probleme*, Ph.D. thesis, Fachbereich Mathematik der Technischen Universität Berlin, Berlin, 1990. Available online from www.math.tu-berlin.de/∼plato/promo.ps.

[10] A. Rieder, *On the regularization of nonlinear ill-posed problems via inexact Newton iterations*, Inverse Problems, 15 (1999), pp. 309–327.

[11] A. Rieder, *On convergence rates of inexact Newton regularizations*, Numer. Math., 88 (2001), pp. 347–365.

[12] A. Rieder, *Keine Probleme mit Inversen Problemen*, Vieweg, Wiesbaden, Germany, 2003.

[13] E. Zeidler, *Nonlinear Functional Analysis and Its Applications* I: *Fixed-Point Theorems*, Springer-Verlag, New York, 1993.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 623–644

SPECTRAL APPROXIMATION OF THE HELMHOLTZ EQUATION WITH HIGH WAVE NUMBERS*

JIE SHEN[†] AND LI-LIAN WANG[†]

Abstract. A complete error analysis is performed for the spectral-Galerkin approximation of a model Helmholtz equation with high wave numbers. The analysis presented in this paper does not rely on the explicit knowledge of continuous/discrete Green's functions and does not require any mesh condition to be satisfied. Furthermore, new error estimates are also established for multidimensional radial and spherical symmetric domains. Illustrative numerical results in agreement with the theoretical analysis are presented.

Key words. Helmholtz equation, high wave numbers, spectral-Galerkin approximation, error analysis

AMS subject classifications. 65N35, 65N22, 35J05, 65F05

DOI. 10.1137/040607332

1. Introduction.

Time harmonic wave propagations appear in many applications, e.g., wave scattering and transmission, noise reduction, fluid-solid interaction, and sea and earthquake wave propagation. In many situations, time harmonic wave propagations are governed by the following Helmholtz equation in an exterior domain with the so-called Sommerfeld radiation boundary condition:

$$\begin{aligned} &-\Delta u - k^2 u = f \qquad \text{in } R^n \backslash D, \\ &u|_{\partial D} = 0, \;\; \partial_r u - \mathrm{i}ku = o\left(\|x\|^{\frac{1-n}{2}}\right) \qquad \text{as } \|x\| \to \infty, \end{aligned} \tag{1.1}$$

where D is a bounded domain in R^n $(n = 1, 2, 3)$, ∂_r is the radial derivative, and k is the nondimensional wave number: $k = \frac{\omega L}{c}$, where ω is a given frequency, L is the measure of the domain, and c is the sound speed in the acoustic medium.

Problem (1.1) presents a great challenge to numerical analysts and computational scientists because (i) the domain is unbounded, and (ii) the solution is highly oscillatory (when k is large) and decays slowly. There is abundant literature on different numerical techniques that have been developed for this problem, such as boundary element methods [5], infinite element methods [11], methods using nonreflecting boundary conditions [14], perfectly matched layers (PML) [2], among others. In many of these approaches, an essential step is to solve the following problem:

$$\begin{aligned} &-\Delta u - k^2 u = f \qquad \text{in } \Omega := B \backslash D, \\ &u|_{\partial D} = 0, \;\; (\partial_r u - \mathrm{i}ku)|_{\partial B} = g, \end{aligned} \tag{1.2}$$

where ∂_r is the outward normal derivative, f, g are given data, and B is a sufficiently large ball containing D.

The analysis and implementation of numerical schemes for (1.2) are challenging when the wave number k is large. The Galerkin finite element method (FEM) for (1.2)

*Received by the editors April 26, 2004; accepted for publication (in revised form) October 11, 2004; published electronically June 30, 2005.

http://www.siam.org/journals/sinum/43-2/60733.html

[†]Department of Mathematics, Purdue University, West Lafayette, IN 47907 (shen@math.purdue.edu, lwang@math.purdue.edu). The work of the first author was partially supported by NSF grant DMS-0311915.

in the one-dimensional case was first carried out in [8], where the well-posedness and error estimates of the Galerkin FEM were established under the condition $k^2h \lesssim 1$ using the Green's function and an argument due to Schatz [21]. A refined analysis for (1.2) in the one-dimensional case was performed in [18] (resp., [19]) for the h version (resp., hp version) of FEM, where the well-posedness and error estimates were established under the condition $kh \lesssim 1$ using the discrete Green's functions. The proofs in these works rely heavily on the use of explicit forms of continuous and/or discrete Green's functions. Hence, it is extremely complicated, if not impossible, to extend to more general cases and higher space dimensions.

On the other hand, the error estimates in the aforementioned papers concluded that the mesh condition $k^2h \lesssim 1$ has to be verified for the error estimates to be independent of k. This so-called pollution effect associated with high wave numbers was discussed in detail in [1]. It is well known [13] that spectral methods are suitable for problems with highly oscillatory solutions since they require fewer grid points per wavelength compared with finite difference methods and FEMs. Furthermore, since the convergence rate of spectral methods increases with the smoothness of the solution, the effect of pollution on the convergence rate of spectral methods is much less significant for smooth (but highly oscillatory) solutions. Hence, it is advantageous to use a spectral method for the Helmholtz equation (1.2) with high wave numbers.

In a recent work [7], Cummings and Feng obtained sharp regularity results for (1.2) in general two- or three-dimensional domains by using Rellich identities instead of using representations in terms of double-layer potentials (cf. [10]). Their analysis not only leads to sharper regularity results but also greatly simplifies the usual process for obtaining a priori estimates and is applicable to general and multidimensional star-shaped domains. Unfortunately, the technique used in [7] cannot be directly applied to Galerkin FEMs because the finite element subspaces do not contain the special test functions used in [7]. However, the situation is different in a spectral-Galerkin method, for which the procedure in [7] can be applied.

We consider in this paper the spectral-Galerkin method for the Helmholtz equation with high wave numbers. In the next section, we set up a prototypical one-dimensional Helmholtz equation which is derived from a multidimensional Helmholtz equation, and we establish its well-posedness; then we derive a priori estimates which are essential for the error analysis. In section 3, we introduce the spectral-Galerkin method and use the same arguments for the space continuous problem to establish the well-posedness and a priori estimates for the discrete problem; then we employ some new optimal Jacobi approximation results to carry out a complete error analysis. In section 4, we consider an alternative formulation which leads to an efficient numerical algorithm and present some illustrative numerical results. We extend our analysis to multidimensional domains in section 5.

We now introduce some notation. Let $\omega(x)$ be a given real weight function in $I = (a, b)$, which is not necessary in $L^1(I)$. We denote by $L^2_\omega(I)$ a Hilbert space of real or complex functions with inner product and norm

$$(u, v)_\omega = \int_I u(r)\overline{v(r)}\omega(r)dr, \qquad \|u\|_\omega = (u, u)_\omega^{\frac{1}{2}},$$

where $\bar{v}$ is the complex conjugate of v. Then the weighted Sobolev spaces $H^s_\omega(I)$ $(s = 0, 1, 2, \dots)$ can be defined as usual with inner products, norms, and seminorms denoted by $(\cdot, \cdot)_{s,\omega}$, $\|\cdot\|_{s,\omega}$, and $|\cdot|_{s,\omega}$, respectively. For real $s > 0$, $H^s_\omega(I)$ is defined by space interpolation. The subscript ω will be omitted from the notation in the case $\omega \equiv 1$. For simplicity, we denote $\partial^l_r v = \frac{d^l v}{dr^l}$, $l \geq 1$.

2. Model equation and a priori estimates. Since a global spectral method is most efficient on regular domains, we shall restrict our attention to the following special cases ($b > a \geq 0$):

- One-dimensional case (1-D): $D = (0, a)$ and $B = (0, b)$.
- Two-dimensional case (2-D): $D = \{(x, y) : x^2 + y^2 < a^2\}$ and $B = \{(x, y) : x^2 + y^2 < b^2\}$.
- Three-dimensional case (3-D): $D = \{(x, y, z) : x^2 + y^2 + z^2 < a^2\}$ and $B = \{(x, y, z) : x^2 + y^2 + z^2 < b^2\}$.

In the 2-D (resp., 3-D) case, we expand functions in polar (resp., spherical) coordinates, i.e., $u = \sum u_m(r)e^{im\theta}$ (resp., $u = \sum u_{lm}(r)Y_{l,m}(\theta, \phi)$, where $\{Y_{l,m}(\theta, \phi)\}$ are the usual spherical harmonic functions). Hence, the problem (1.2) reduces, after a polar (when $n = 2$) or spherical (when $n = 3$) transform, to a sequence (for each m in 2-D and (l, m) in 3-D) of 1-D equations (for brevity, we use u to denote u_m/u_{lm}, and likewise for f and g, below):

$$-\frac{1}{r^{n-1}}\partial_r(r^{n-1}\partial_r u) + d_m\frac{u}{r^2} - k^2 u = f, \quad r \in (a, b),\ n = 1, 2, 3,\ m \geq 0 \tag{2.1}$$

($d_m = 0, m^2, m(m+1)$ for $n = 1, 2, 3$, respectively), with suitable boundary conditions to be specified below.

If $a > 0$, the coefficients r^{n-1} and r^{-2} in (2.1) are uniformly bounded, so (2.1) with $a > 0$ is easier to handle than the case $a = 0$. Hence, for brevity of presentation, we shall be concerned mainly with the case $a = 0$, while some results for $a > 0$ will be stated without proof in section 5. On the other hand, the character of (2.1) does not change with the change of variable: $r \to rb$. Consequently, it suffices to consider the problem (2.1) in $I := (0, 1)$. The appropriate boundary conditions for (2.1) are the pole conditions at $r = 0$,

$$u(0) = 0 \quad \text{if } n = 1 \text{ and if } n = 2 \text{ with } m > 0, \tag{2.2}$$

and the Robin boundary condition (derived from the Sommerfeld radiation boundary condition) at $r = 1$,

$$u'(1) - \mathrm{i}ku(1) = g. \tag{2.3}$$

We note that error estimates for finite element approximations to the Helmholtz equation (2.1) with high wave numbers were derived in [8, 18, 19] for the 1-D case, and in [9, 6] for 2-D cases and in [12] for the 1-D Bessel equation reduced from a 3-D spherical domain, respectively.

Let $\mathbb{N}$ be the set of all nonnegative integers and let P_N be the space of all polynomials of degree at most N. We shall use c to denote a generic positive constant independent of any function, the wave frequency k, the radial/spherical frequency m, and the number of modes N. We use the expression $A \lesssim B$ to mean that there exists a generic positive constant c such that $A \leq cB$.

2.1. Variational formulation and weak solution. Let us denote $\omega^\alpha(r) = r^\alpha$ and $\omega(r) = r$. We define a Hilbert space,

$$X := X(m, n) := \{u \in H^1_{\omega^{n-1}}(I) :\ u \in L^2_{\omega^{n-3}}(I) \text{ for } n = 2, 3;\ u \text{ satisfies (2.2)}\},$$

and a sesquilinear form on $X \times X$,

$$\begin{aligned} \mathcal{B}(u, v) := \mathcal{B}_{mn}(u, v) := (\partial_r u, \partial_r v)_{\omega^{n-1}} + d_m(u, v)_{\omega^{n-3}} - k^2(u, v)_{\omega^{n-1}} \\ - \mathrm{i}ku(1)\overline{v(1)}. \end{aligned} \tag{2.4}$$

Note that to lighten the presentation, we will often omit m and n from the notation.

Then the weak formulation of (2.1)–(2.2) is to find $u \in X$ such that

$$\mathcal{B}(u,v) = (f,v)_{\omega^{n-1}} + g\overline{v(1)} \;\; \forall v \in X, \;\; n = 1,2,3. \tag{2.5}$$

THEOREM 2.1. *Given $f \in X'$, the problem* (2.5) *admits a unique weak solution.*

Proof. This result with $n = 1$ was established in [8, 18]. Hence, we shall prove only the cases with $n = 2$ and 3.

We first consider the uniqueness. It suffices to show that $u = 0$ is the only solution of the problem (2.5) with $f \equiv 0$ and $g = 0$.

Taking $v = u$ in (2.5) with $f \equiv 0$ and $g = 0$, we find from (2.4) that

$$\mathcal{B}(u,u) = \|\partial_r u\|^2_{\omega^{n-1}} + d_m\|u\|^2_{\omega^{n-3}} - k^2\|u\|^2_{\omega^{n-1}} - \mathrm{i}k|u(1)|^2 = 0, \tag{2.6}$$

which implies immediately $u(1) = 0$.

Next, let $J_\mu(r)$ be the Bessel function of the first kind of order μ. We recall that

$$\phi_m(r;h,n) := \begin{cases} J_m(hr), & n = 2, \; r,h > 0, \\ \frac{1}{\sqrt{r}} J_{m+\frac{1}{2}}(hr), & n = 3, \; r,h > 0, \end{cases} \tag{2.7}$$

is the solution of the modified Bessel equation (cf. [25]):

$$-\frac{1}{r^{n-1}}\partial_r(r^{n-1}\partial_r\phi_m) - \Big(h^2 - \frac{d_m}{r^2}\Big)\phi_m = 0, \;\; n = 2,3, \; m \geq 0. \tag{2.8}$$

Let $\{\xi_j\}_{j=1}^\infty$ be the set of all positive real zeros of the Bessel function $J_{m+\frac{n}{2}-1}(r)$. Then $\{\phi_m(r;\xi_j,n)\}_{j=1}^\infty$ forms a complete orthogonal system in $L^2_{\omega^{n-1}}(I)$ (cf. [26]), namely,

$$\begin{aligned} &\int_0^1 \phi_m(r;\xi_j,n)\phi_m(r;\xi_l,n) r^{n-1} dr \\ &= \int_0^1 J_{m+\frac{n}{2}-1}(r\xi_j) J_{m+\frac{n}{2}-1}(r\xi_l) r dr = \frac{1}{2} J^2_{m+\frac{n}{2}}(\xi_j)\delta_{j,l}. \end{aligned} \tag{2.9}$$

Since $u \in L^2_{\omega^{n-1}}(I)$, we can write

$$u(r) = \sum_{j=1}^\infty \tilde{u}_m^{(j)} \phi_m(r;\xi_j,n), \tag{2.10}$$

with

$$\tilde{u}_m^{(j)} = \frac{1}{\gamma_m^{(j)}} \int_0^1 u(r)\phi_m(r;\xi_j,n) r^{n-1} dr, \quad \gamma_m^{(j)} = \frac{1}{2} J^2_{m+\frac{n}{2}}(\xi_j). \tag{2.11}$$

Thanks to $u(1) = 0$, we derive from (2.8) with $h = \xi_j$, (2.11), and integration by parts that

$$\begin{aligned} 0 = \mathcal{B}(u,\phi_m) &= \int_0^1 u(r)\Big\{ -\frac{1}{r^{n-1}}\partial_r(r^{n-1}\partial_r\phi_m) - \Big(k^2 - \frac{d_m}{r^2}\Big)\phi_m\Big\} r^{n-1} dr \\ &= (\xi_j^2 - k^2)\int_0^1 u(r)\phi_m(r;\xi_j,n) r^{n-1} dr = (\xi_j^2 - k^2)\gamma_m^{(j)}\tilde{u}_m^{(j)}. \end{aligned} \tag{2.12}$$

If $J_{m+\frac{n}{2}-1}(k) \neq 0$ (i.e., $k \neq \xi_j$ for all $j \geq 1$), then (2.12) implies $\tilde{u}_m^{(j)} = 0$ for all j. Accordingly, we have $u \equiv 0$ (cf. (2.10)).

On the other hand, if $J_{m+\frac{n}{2}-1}(k) = 0$, then $k = \xi_{j_0}$ for some $j_0 \geq 1$. We then derive from (2.12) that $\tilde{u}_m^{(j)} = 0$ for all $j \neq j_0$. Thus, by (2.10),

$$u(r) = \tilde{u}_m^{(j_0)} \phi_m(r; \xi_{j_0}, n), \tag{2.13}$$

and it remains to verify $\tilde{u}_m^{(j_0)} = 0$. Due to $u(1) = 0$, integration by parts yields

$$\int_0^1 \partial_r u(r) \partial_r (r^m) r^{n-1} dr = -d_m \int_0^1 u(r) r^{m+n-3} dr, \quad n = 2, 3,\ m \geq 0. \tag{2.14}$$

Taking $v = r^m (\in X)$ in (2.5), we obtain from (2.13) that

$$\begin{aligned} 0 &= \mathcal{B}(u, r^m) = \tilde{u}_m^{(j_0)} \mathcal{B}(\phi_m(\cdot; \xi_{j_0}, n), r^m) \\ &= -k^2 \tilde{u}_m^{(j_0)} \int_0^1 \phi_m(r; \xi_{j_0}, n) r^{m+n-1} dr = -k^2 \tilde{u}_m^{(j_0)} \int_0^1 J_{m+\frac{n}{2}-1}(r\xi_{j_0}) r^{m+\frac{n}{2}} dr. \end{aligned} \tag{2.15}$$

We recall that $r^\mu,\ \mu \geq 0$, can be expanded as (see [26, p. 581])

$$r^\mu = \sum_{j=1}^{\infty} \frac{2 J_\mu(r\xi_j)}{\xi_j J_{\mu+1}(\xi_j)}, \quad 0 \leq r < 1. \tag{2.16}$$

Inserting (2.16) with $\mu = m + \frac{n}{2} - 1$ into (2.15) and using the orthogonality (2.9) lead to

$$0 = -k^2 \tilde{u}_m^{(j_0)} \int_0^1 J_{m+\frac{n}{2}-1}(r\xi_{j_0}) r^{m+\frac{n}{2}} dr = -k^2 \tilde{u}_m^{(j_0)} \frac{J_{m+\frac{n}{2}}(\xi_{j_0})}{\xi_{j_0}}.$$

This implies $\tilde{u}_m^{(j_0)} = 0$. Hence, we have $u \equiv 0$, which implies the uniqueness.

To prove the existence, we note from (2.6) that the following Gårding-type inequality holds:

$$\mathrm{Re}(\mathcal{B}(u, u)) \geq \|\partial_r u\|_{\omega^{n-1}}^2 + d_m \|u\|_{\omega^{n-3}}^2 - k^2 \|u\|_{\omega^{n-1}}^2. \tag{2.17}$$

Since all the arguments above apply also to the dual problem of (2.5), by the classical Fredholm alternative argument (see, for instance, [20, p. 194]); problem (2.5) either has a nontrivial solution with $f \equiv 0$ and $g = 0$ or it has at least one solution for every $f \in X'$. Since the uniqueness is proved, existence follows from the above argument. □

2.2. A priori estimates.

THEOREM 2.2. *If $f \in L^2_{\omega^{n-1}}(I)$, we have*

$$\|\partial_r u\|_{\omega^{n-1}} + \sqrt{d_m} \|u\|_{\omega^{n-3}} + k \|u\|_{\omega^{n-1}} \lesssim |g| + \|f\|_{\omega^{n-1}}, \quad n = 1, 2, 3. \tag{2.18}$$

Proof. The proof consists of taking two different test functions in (2.5). The first test function is the usual one. As in [7], the key step is to choose a suitable second test function which enables us to obtain a priori estimates without using the Green's functions as in [8, 18, 19]. In the following proof, $\varepsilon_j > 0,\ 1 \leq j \leq 5$, are adjustable real numbers.

Step 1. We take $v = u$ in (2.5) whose imaginary and real parts are as follows:

$$\begin{aligned} &-k|u(1)|^2 = \mathrm{Im}(g\overline{u(1)}) + \mathrm{Im}(f,u)_{\omega^{n-1}},\\ &\|\partial_r u\|^2_{\omega^{n-1}} + d_m\|u\|^2_{\omega^{n-3}} - k^2\|u\|^2_{\omega^{n-1}} = \mathrm{Re}(g\overline{u(1)}) + \mathrm{Re}(f,u)_{\omega^{n-1}}. \end{aligned} \tag{2.19}$$

Applying the Cauchy–Schwarz inequality to the imaginary part, we obtain

$$\begin{aligned} k|u(1)|^2 &\le |\mathrm{Im}(g\overline{u(1)})| + |\mathrm{Im}(f,u)_{\omega^{n-1}}|,\\ &\le \frac{k}{2}|u(1)|^2 + \frac{1}{2k}|g|^2 + \frac{\varepsilon_1 k}{2}\|u\|^2_{\omega^{n-1}} + \frac{1}{2\varepsilon_1 k}\|f\|^2_{\omega^{n-1}}; \end{aligned} \tag{2.20}$$

likewise, we obtain from the real part that

$$\begin{aligned} &\|\partial_r u\|^2_{\omega^{n-1}} + d_m\|u\|^2_{\omega^{n-3}} \le k^2\|u\|^2_{\omega^{n-1}} + |\mathrm{Re}(g\overline{u(1)})| + |\mathrm{Re}(f,u)_{\omega^{n-1}}|\\ &\quad\le k^2\|u\|^2_{\omega^{n-1}} + \varepsilon_2 k^2|u(1)|^2 + \frac{1}{4\varepsilon_2 k^2}|g|^2 + \frac{\varepsilon_3 k^2}{2}\|u\|^2_{\omega^{n-1}} + \frac{1}{2\varepsilon_3 k^2}\|f\|^2_{\omega^{n-1}}. \end{aligned} \tag{2.21}$$

Therefore, by (2.20),

$$|u(1)|^2 \le \varepsilon_1\|u\|^2_{\omega^{n-1}} + \frac{1}{k^2}|g|^2 + \frac{1}{\varepsilon_1 k^2}\|f\|^2_{\omega^{n-1}}, \tag{2.22}$$

and by (2.21)–(2.22) with $\varepsilon_2 = \frac{\varepsilon_3}{2\varepsilon_1}$,

$$\begin{aligned} \|\partial_r u\|^2_{\omega^{n-1}} + d_m\|u\|^2_{\omega^{n-3}} &\le (1+\varepsilon_3)k^2\|u\|^2_{\omega^{n-1}}\\ &+ \Big(\frac{\varepsilon_3}{2\varepsilon_1} + \frac{\varepsilon_1}{2\varepsilon_3 k^2}\Big)|g|^2 + \Big(\frac{\varepsilon_3}{2\varepsilon_1^2} + \frac{1}{2\varepsilon_3 k^2}\Big)\|f\|^2_{\omega^{n-1}}. \end{aligned} \tag{2.23}$$

It remains to bound $k^2\|u\|^2_{\omega^{n-1}}$.

Step 2. Using a usual regularity argument, one can easily derive that, for $f \in L^2_{\omega^{n-1}}(I)$, the weak solution of (2.5) satisfies $r\partial_r u \in X$, and we now consider the real part of (2.5) with $v = 2r\partial_r u$. After integrating by parts, the first three terms become

$$2\mathrm{Re}(\partial_r u, \partial_r(r\partial_r u))_{\omega^{n-1}} = |\partial_r u(1)|^2 + (2-n)\|\partial_r u\|^2_{\omega^{n-1}}; \tag{2.24a}$$

$$2\mathrm{Re}(u, r\partial_r u)_{\omega^{n-3}} = |u(1)|^2 + (2-n)\|u\|^2_{\omega^{n-3}}; \tag{2.24b}$$

$$-2k^2\mathrm{Re}(u, r\partial_r u)_{\omega^{n-1}} = -k^2|u(1)|^2 + nk^2\|u\|^2_{\omega^{n-1}}. \tag{2.24c}$$

Consequently, the real part of (2.5) with $v = 2r\partial_r u$ is

$$\begin{aligned} &(2-n)\Big(\|\partial_r u\|^2_{\omega^{n-1}} + d_m\|u\|^2_{\omega^{n-3}}\Big) + nk^2\|u\|^2_{\omega^{n-1}} + |\partial_r u(1)|^2 + d_m|u(1)|^2\\ &\qquad = k^2|u(1)|^2 + 2\mathrm{Re}\Big((\mathrm{i}ku(1) + g)\overline{\partial_r u(1)}\Big) + 2\mathrm{Re}(f, r\partial_r u)_{\omega^{n-1}}. \end{aligned} \tag{2.25}$$

We now proceed separately for the three different cases.

Case (i): $n = 1$. Thanks to $d_m = 0$, we derive from (2.25) and the Cauchy–Schwarz inequality that

$$\begin{aligned} \|\partial_r u\|^2 + k^2\|u\|^2 + |\partial_r u(1)|^2 &\le k^2|u(1)|^2 + \frac{1}{2}|\partial_r u(1)|^2\\ &+ 2k^2|u(1)|^2 + 2|g|^2 + \frac{1}{2}\|\partial_r u\|^2 + 2\|f\|^2. \end{aligned} \tag{2.26}$$

Hence, we obtain from (2.22) that

$$\begin{aligned}\frac{1}{2}\|\partial_r u\|^2 + k^2\|u\|^2 + \frac{1}{2}|\partial_r u(1)|^2 &\le 3\varepsilon_1 k^2\|u\|^2 + c\Big(|g|^2 + (\varepsilon_1^{-1}+2)\|f\|^2\Big)\\ &\le \frac{k^2}{2}\|u\|^2 + c(|g|^2+\|f\|^2),\end{aligned} \tag{2.27}$$

where we have taken $\varepsilon_1 = \frac{1}{6}$ to derive the last inequality. This implies (2.18) with $n = 1$.

Case (ii): $n = 2$. Similarly, we have from (2.22), (2.23), and (2.25) that

$$\begin{aligned}2k^2\|u\|_\omega^2 + |\partial_r u(1)|^2 + d_m|u(1)|^2 &\le \frac{1}{2}|\partial_r u(1)|^2 + 3k^2|u(1)|^2\\ &\quad + 2|g|^2 + \varepsilon_4\|\partial_r u\|_\omega^2 + \varepsilon_4^{-1}\|f\|_\omega^2\\ &\le \frac{1}{2}|\partial_r u(1)|^2 + \Big(3\varepsilon_1 + \varepsilon_4(1+\varepsilon_3)\Big)k^2\|u\|_\omega^2 + C_1|g|^2 + C_2\|f\|_\omega^2,\end{aligned}$$

where C_1 and C_2 are two positive constants in terms of $\varepsilon_1, \varepsilon_3$, and ε_4. We take $\varepsilon_1 = 1/6$, $\varepsilon_3 = 1$, $\varepsilon_4 = 1/4$ and obtain that

$$k^2\|u\|_\omega^2 + d_m|u(1)|^2 + \frac{1}{2}|\partial_r u(1)|^2 \lesssim |g|^2 + \|f\|_\omega^2. \tag{2.28}$$

A combination of (2.23) and (2.28) leads to (2.18) with $n = 2$.

Case (iii): $n = 3$. As in the derivation of Case (ii), using (2.22), (2.23), and (2.25) yields

$$\begin{aligned}3k^2\|u\|_{\omega^2}^2 + |\partial_r u(1)|^2 + d_m|u(1)|^2 &\le \|\partial_r u\|_{\omega^2}^2 + d_m\|u\|^2 + \frac{1}{2}|\partial_r u(1)|^2 + 3k^2|u(1)|^2\\ &\quad + 2|g|^2 + \varepsilon_5\|\partial_r u\|_{\omega^2}^2 + \varepsilon_5^{-1}\|f\|_{\omega^2}^2\\ &\le \frac{1}{2}|\partial_r u(1)|^2 + \Big(3\varepsilon_1 + (1+\varepsilon_5)(1+\varepsilon_3)\Big)k^2\|u\|_{\omega^2}^2 + C_3|g|^2 + C_4\|f\|_{\omega^2}^2,\end{aligned}$$

where C_3 and C_4 are two positive constants depending only on $\varepsilon_1, \varepsilon_3$, and ε_5. Taking $\varepsilon_1 = 2/27$ and $\varepsilon_3 = \varepsilon_5 = 1/3$ such that $3\varepsilon_1 + (1+\varepsilon_5)(1+\varepsilon_3) = 2$ gives

$$k^2\|u\|_{\omega^2}^2 + d_m|u(1)|^2 + \frac{1}{2}|\partial_r u(1)|^2 \lesssim |g|^2 + \|f\|_{\omega^2}^2. \tag{2.29}$$

This completes the proof. □

Remark 2.1. We have also proved that

$$|\partial_r u(1)| + \sqrt{d_m}|u(1)| + k|u(1)| \lesssim |g| + \|f\|_{\omega^{n-1}}, \quad n = 1, 2, 3. \tag{2.30}$$

Remark 2.2. A combination of (2.1) and (2.18) leads to

$$|u|_2 \lesssim k|g| + (1+k)\|f\| \quad \text{if } n = 1, \tag{2.31a}$$

$$\|D^2 u\| \lesssim k|g| + (1+k)\|f\|_{\omega^{n-1}} \quad \text{if } n = 2, 3, \tag{2.31b}$$

where $D^2 u = -\partial_r(r^{n-1}\partial_r u) + d_m r^{n-3}u$.

3. Spectral-Galerkin approximation. In this section, we shall present the spectral-Galerkin scheme and analyze its errors in suitably weighted Sobolev spaces.

3.1. Spectral-Galerkin solution. Let us denote $X_N := X \cap P_N$, where P_N is the space of all polynomials of degree at most N. The spectral-Galerkin approximation of (2.5) is to find $u_N \in X_N$ such that

$$\mathcal{B}(u_N, v_N) = (f, v_N)_{\omega^{n-1}} + g\overline{v_N(1)} \quad \forall v_N \in X_N. \tag{3.1}$$

We observe that the sesquilinear form $\mathcal{B}(\cdot,\cdot)$ is not coercive in $X_N \times X_N$. To prove the well-posedness of (3.1) with $n = 1$, Douglas et al. [8] used an argument due to Schatz [21] for the (finite element) discrete system under the condition $k^2h \lesssim 1$, while Ihlenburg and Babuška [18] used an inf-sup argument due to Babuška and Brezzi under the condition $kh \lesssim 1$. However, the spectral-Galerkin approximation space X_N, unlike in the Galerkin FEM, has the following property: *For $u_N \in X_N$, we have $r\partial_r u_N \in X_N$. Hence, the proof of Theorem* 2.2 *is also valid for the discrete system* (3.1); i.e., we have the following.

THEOREM 3.1. *Let u_N be a solution of* (3.1). *Then Theorem* 2.2 *holds with u_N in place of u.*

An immediate consequence is the following.

COROLLARY 3.1. *The problem* (3.1) *admits a unique solution.*

Proof. Since (3.1) is a finite-dimensional linear system, it suffices to prove the uniqueness. Now, let u_N be a solution of (3.1) with $f \equiv 0$ and $g = 0$. We derive from Theorem 3.1 that $u_N \equiv 0$, which implies the uniqueness. □

Remark 3.1. It is interesting to note that while the existence of a solution for finite element approximations to the Helmholtz equation is guaranteed only under a mesh condition $kh \lesssim 1$ (see, for instance, [8, 18]), the spectral-Galerkin approximation (3.1) always admits a unique solution, just as (2.1) itself.

3.2. Error estimates. Thanks to Theorems 2.2 and 3.1, we can analyze the errors of the proposed scheme by comparing the numerical solution with some orthogonal projection of the exact solution as usual. For this purpose, let $\Pi^{1,m}_{N,n} : X \to X_N$ be an orthogonal projection, defined by

$$(\partial_r(u - \Pi^{1,m}_{N,n}u),\ \partial_r v_N)_{\omega^{n-1}} = 0 \quad \forall v_N \in X_N,\ \ n = 1, 3\ \ \forall m \text{ and } n = 2 \text{ with } m = 0. \tag{3.2}$$

In order to estimate the errors between u and u_N, we have to analyze the approximation properties of the projector $\Pi^{1,m}_{N,n}$ for functions in the following suitably weighted Sobolev spaces:

$$\widetilde{H}^s_{\omega^{n-1}}(I) := \{u\ :\ u \in L^2_{\omega^{n-1}}(I),\ (r - r^2)^{\frac{k-1}{2}}\partial^k_r u \in L^2_{\omega^{n-1}}(I),\ 1 \le k \le s\},$$

with the norm and seminorm

$$\|u\|_{\widetilde{H}^s_{\omega^{n-1}}} = \Big(\|u\|^2_{\omega^{n-1}} + \sum_{k=1}^{s} \|(r - r^2)^{\frac{k-1}{2}}\partial^k_r u\|^2_{\omega^{n-1}}\Big)^{\frac{1}{2}},$$
$$|u|_{\widetilde{H}^s_{\omega^{n-1}}} = \|(r - r^2)^{\frac{s-1}{2}}\partial^s_r u\|_{\omega^{n-1}}, \quad s \ge 1,\ s \in \mathbb{N}.$$

LEMMA 3.1. *For any $u \in X \cap \widetilde{H}^s_{\omega^{n-1}}(I)$, with $s \ge 1$ and $s \in \mathbb{N}$,*

$$\|\Pi^{1,m}_{N,n}u - u\|_{\mu,\omega^{n-1}} \lesssim N^{\mu-s}\|(r - r^2)^{\frac{s-1}{2}}\partial^s_r u\|_{\omega^{n-1}}, \tag{3.3}$$
$$\mu = 0, 1, \quad n = 1, 3\ \ \forall m \text{ and } n = 2 \text{ with } m = 0.$$

Proof. This result for $n = 1$ can be derived from [4] with an improvement of the norm in terms of the weights $(r - r^2)^{\frac{s-1}{2}}$ given in [16]. For $n = 2$ with $m = 0$ and $n = 3$, one can refer to [15, 16] for the proofs. □

Next, we shall estimate $e_N = u_N - \Pi^{1,m}_{N,n} u$. We denote $\tilde{e}_N = u - \Pi^{1,m}_{N,n} u$.

LEMMA 3.2. *Let u and u_N be, respectively, the solutions of* (2.5) *and* (3.1). *Then we have, for $n = 1, 3$ for all m and $n = 2$ with $m = 0$,*

$$\begin{aligned}(3.4)\qquad &\|\partial_r e_N\|_{\omega^{n-1}} + \sqrt{d_m}\|e_N\|_{\omega^{n-3}} + k\|e_N\|_{\omega^{n-1}} \\ &\quad\lesssim \sqrt{d_m}\Big(\|\partial_r \tilde{e}_N\|_{\omega^{n-1}} + \|\tilde{e}_N\|_{\omega^{n-3}}\Big) + k^2\|\tilde{e}_N\|_{\omega^{n-1}} + k(1 + d_m k^{-2})|\tilde{e}_N(1)|.\end{aligned}$$

Proof. By (2.5) and (3.1), we have $\mathcal{B}(u - u_N, v_N) = 0$ for all $v_N \in X_N$. Hence, we derive from (2.5) and (3.2) that, for any $v_N \in X_N$,

$$\begin{aligned}(3.5)\qquad \mathcal{B}(e_N, v_N) &= \mathcal{B}(u - \Pi^{1,m}_{N,n} u, v_N) \\ &= d_m(\tilde{e}_N, v_N)_{\omega^{n-3}} - k^2(\tilde{e}_N, v_N)_{\omega^{n-1}} - \mathrm{i}k\tilde{e}_N(1)\overline{v_N(1)}.\end{aligned}$$

We can view (3.5) in the form of (2.5) with $u = e_N$, $g = -\mathrm{i}k\tilde{e}_N(1)$, $f = -k^2\tilde{e}_N$ plus an extra term $d_m(\tilde{e}_N, v_N)_{\omega^{n-3}}$. Hence, as in the proof of Theorem 2.2, we take two different test functions $v_N = e_N, r\partial_r e_N \in X_N$ and estimate the extra term by

$$\begin{aligned} d_m|(\tilde{e}_N, e_N)_{\omega^{n-3}}| &\le \varepsilon_6 d_m\|e_N\|^2_{\omega^{n-3}} + \frac{d_m}{4\varepsilon_6}\|\tilde{e}_N\|^2_{\omega^{n-3}}, \\ d_m|(\tilde{e}_N, r\partial_r e_N)_{\omega^{n-3}}| &= d_m|\tilde{e}_N(1)\overline{e_N(1)} - (\partial_r\tilde{e}_N, e_N)_{\omega^{n-2}} - (n-2)(\tilde{e}_N, e_N)_{\omega^{n-3}}| \\ &\le \varepsilon_7 k^2|e_N(1)|^2 + \frac{d_m^2}{4k^2\varepsilon_7}|\tilde{e}_N(1)|^2 + \varepsilon_8 d_m\|e_N\|^2_{\omega^{n-3}} \\ &\quad + \frac{c d_m}{4\varepsilon_8}\Big(\|\partial_r\tilde{e}_N\|^2_{\omega^{n-1}} + \|\tilde{e}_N\|^2_{\omega^{n-3}}\Big).\end{aligned}$$

Thus, choosing suitable constants $\{\varepsilon_j\}_{j=6}^{8}$, and following a procedure similar to the proof of Theorem 2.2, we can derive

$$\begin{aligned}(3.6)\qquad &\|\partial_r e_N\|^2_{\omega^{n-1}} + d_m\|e_N\|^2_{\omega^{n-3}} + k^2\|e_N\|^2_{\omega^{n-1}} \\ &\quad\lesssim d_m(\|\partial_r\tilde{e}_N\|^2_{\omega^{n-1}} + \|\tilde{e}_N\|^2_{\omega^{n-3}}) + k^4\|\tilde{e}_N\|^2_{\omega^{n-1}} + k^2(1 + d_m^2 k^{-4})|\tilde{e}_N(1)|^2,\end{aligned}$$

which leads to the desired result. □

We now recall the following inequalities.

LEMMA 3.3.

$$(3.7a)\qquad |u(1)| \lesssim \|u\|^{\frac{1}{2}}_{\omega^{n-1}}\|u\|^{\frac{1}{2}}_{1,\omega^{n-1}} \quad \forall u \in H^1_{\omega^{n-1}}(I), \ \ n = 1, 2, 3,$$

$$(3.7b)\qquad \|u\| \lesssim \|u\|_{1,\omega^2} \quad \forall u \in H^1_{\omega^2}(I).$$

Proof. By the Sobolev inequality (see the appendix in [4]),

$$|u(1)| \lesssim \|u\|^{\frac{1}{2}}_{L^2(1/2,1)}\|u\|^{\frac{1}{2}}_{H^1(1/2,1)} \lesssim \|u\|^{\frac{1}{2}}_{L^2_{\omega^{n-1}}(1/2,1)}\|u\|^{\frac{1}{2}}_{H^1_{\omega^{n-1}}(1/2,1)} \lesssim \|u\|^{\frac{1}{2}}_{\omega^{n-1}}\|u\|^{\frac{1}{2}}_{1,\omega^{n-1}}.$$

Here, we used the fact that the weight function r^{n-1} is uniformly bounded on $[1/2, 1]$. The inequality (3.7b) follows directly from formula (13.5) of [3]. □

As a consequence of (3.7b) and Lemma 3.1, we derive that for $n=3$,

$$\|\Pi_{N,n}^{1,m}u-u\|_{\omega^{n-3}} \lesssim \|\Pi_{N,n}^{1,m}u-u\|_{1,\omega^{n-1}} \lesssim N^{1-s}\|(r-r^2)^{\frac{s-1}{2}}\partial_r^s u\|_{\omega^{n-1}}. \tag{3.8}$$

With the above preparations, we can now prove our main results.

THEOREM 3.2. *Let u and u_N be, respectively, the solutions of* (2.5) *and* (3.1) *such that $u \in X \cap \widetilde{H}^s_{\omega^{n-1}}(I)$ with $s \geq 1, s \in \mathbb{N}$.*

(i) *For $n=1$ or $n=2,3, m=0$,*

$$\|\partial_r(u-u_N)\|_{\omega^{n-1}} + k\|u-u_N\|_{\omega^{n-1}} \lesssim (1+k^2N^{-1})N^{1-s}\|(r-r^2)^{\frac{s-1}{2}}\partial_r^s u\|_{\omega^{n-1}}. \tag{3.9}$$

(ii) *For $n=3$ and $m>0$,*

$$\begin{aligned}&\|\partial_r(u-u_N)\|_{\omega^2} + \sqrt{d_m}\|u-u_N\| + k\|u-u_N\|_{\omega^2}\\ &\quad\lesssim \Big(\sqrt{d_m}+d_m^2k^{-4}+k^2N^{-1}\Big)N^{1-s}\|(r-r^2)^{\frac{s-1}{2}}\partial_r^s u\|_{\omega^2},\end{aligned} \tag{3.10}$$

where $d_m = m(m+1)$.

Proof. We first prove (3.9). Since

$$\begin{aligned}\|\partial_r(u-u_N)\|_{\omega^{n-1}} + k\|u-u_N\|_{\omega^{n-1}} &\lesssim \|\partial_r(\Pi_{N,n}^{1,m}u-u)\|_{\omega^{n-1}}\\ &\quad+ k\|\Pi_{N,n}^{1,m}u-u\|_{\omega^{n-1}} + \|\partial_r e_N\|_{\omega^{n-1}} + k\|e_N\|_{\omega^{n-1}},\end{aligned}$$

formula (3.9) follows from Lemmas 3.1 and 3.2 and (3.7a).

Similarly, for $n=3$ and $m>0$, we derive from (3.7a) and Lemmas 3.1 and 3.2 that

$$\begin{aligned}&\|\partial_r(u-u_N)\|_{\omega^{n-1}} + \sqrt{d_m}\|u-u_N\|_{\omega^{n-3}} + k\|u-u_N\|_{\omega^{n-1}}\\ &\quad\lesssim \sqrt{d_m}\Big(\|\partial_r(\Pi_{N,n}^{1,m}u-u)\|_{\omega^{n-1}} + \|\Pi_{N,n}^{1,m}u-u\|_{\omega^{n-3}}\Big)\\ &\qquad+ k^2\|\Pi_{N,n}^{1,m}u-u\|_{\omega^{n-1}} + k(1+d_mk^{-2})|(\Pi_{N,n}^{1,m}u-u)(1)|\\ &\quad\lesssim \sqrt{d_m}\Big(\|\partial_r(\Pi_{N,n}^{1,m}u-u)\|_{\omega^{n-1}} + \|\Pi_{N,n}^{1,m}u-u\|_{\omega^{n-3}}\Big)\\ &\qquad+ 2k^2\|\Pi_{N,n}^{1,m}u-u\|_{\omega^{n-1}} + (1+d_mk^{-2})^2\|\Pi_{N,n}^{1,m}u-u\|_{1,\omega^{n-1}}\\ &\quad\lesssim \Big(\sqrt{d_m}+(1+d_mk^{-2})^2+k^2N^{-1}\Big)N^{1-s}\|(r-r^2)^{\frac{s-1}{2}}\partial_r^s u\|_{\omega^{n-1}}\\ &\qquad+ \sqrt{d_m}\|\Pi_{N,n}^{1,m}u-u\|_{\omega^{n-3}}.\end{aligned} \tag{3.11}$$

Hence, we can obtain (3.10) by using (3.8) to estimate the last term in (3.11). □

Remark 3.2. For $n=1$, an error estimate of the same order as in (3.9) was derived in [19] for the *hp* FEM under the condition $kh \lesssim 1$. Our estimate is valid without any restriction on k and N and is bounded by a weaker weighted seminorm.

Although we believe that the estimate (3.10), modulo perhaps a logarithmic term, is also valid for the case $n=2$ with $m>0$, the above proof cannot be directly extended to this case due to a breakdown in the Hardy inequality (cf. [17]) as $\varepsilon \to 0$,

$$\int_0^1 \frac{u^2}{r^2}r^{1-\varepsilon}dr \leq \frac{4}{\varepsilon}\int_0^1 (\partial_r u)^2 r^{1-\varepsilon}dr, \tag{3.12}$$

which indicates that $\|\Pi^{1,m}_{N,n}u - u\|_{\omega^{-1}}$ in the last term of (3.11) cannot be bounded by $\|\partial_r(\Pi^{1,m}_{N,n}u - u)\|_\omega$.

Next, we perform the error estimate for the case $n = 2$ with $m > 0$ by using a different approach.

Let $a_m(u, v) := (\partial_r u, \partial_r v)_\omega + d_m(u, v)_{\omega^{-1}}$ and define the orthogonal projection $\pi^{1,m}_N : X \to X_N$ by

$$a_m(\pi^{1,m}_N u - u, v_N) = 0 \quad \forall v_N \in X_N. \tag{3.13}$$

To analyze the approximation properties of the above projector, we first consider an auxiliary projection. Let $\hat\omega = r(1-r)$, let $P^0_N := \{u \in P_N : u(0) = u(1) = 0\}$, and let π_N be the $L^2_{\hat\omega^{-1}}$-orthogonal projection onto P^0_N defined by

$$(\pi_N u - u, v_N)_{\hat\omega^{-1}} = 0 \quad \forall v_N \in P^0_N.$$

The following result can be derived directly from the generalized Jacobi approximation with parameters $\alpha = \beta = -1$ (cf. Theorem 3.1 of [24]).

Lemma 3.4. *For any $u \in L^2_{\hat\omega^{-1}}(I) \cap \widetilde{H}^s(I)$ with $s \geq 1, s \in \mathbb{N}$,*

$$\|\partial_r(\pi_N u - u)\| + N\|(\pi_N u - u)\|_{\hat\omega^{-1}} \lesssim N^{1-s}\|(r - r^2)^{\frac{s-1}{2}}\partial^s_r u\|. \tag{3.14}$$

Corollary 3.2. *There exists an operator $\pi^1_N : H^1(I) \to P_N$ such that $(\pi^1_N u)(r) = u(r)$ for $r = 0, 1$ and for any $u \in \widetilde{H}^s(I)$, with $s \geq 1, s \in \mathbb{N}$,*

$$\|\partial_r(\pi^1_N u - u)\| + N\|\pi^1_N u - u\|_{\hat\omega^{-1}} \lesssim N^{1-s}\|(r - r^2)^{\frac{s-1}{2}}\partial^s_r u\|. \tag{3.15}$$

Proof. Let $u_*(r) = (1-r)u(0) + ru(1) \in P_1$ for all $u \in H^1(I)$. By construction, we have $(u - u_*)(r) = 0$ for $r = 0, 1$. Next, we derive from the Hardy inequality (cf. [17]) that

$$\begin{aligned} \Big(\int_0^1 (u-u_*)^2(r-r^2)^{-1}dr\Big)^{\frac12} &\lesssim \Big(\int_0^1 (\partial_r(u-u_*))^2 dr\Big)^{\frac12} \\ &\lesssim \|\partial_r u\| + |u(1) - u(0)| \lesssim \|\partial_r u\| + \int_0^1 |\partial_r u| dr \lesssim \|\partial_r u\|. \end{aligned} \tag{3.16}$$

Hence, $u - u_* \in L^2_{\hat\omega^{-1}}(I)$ and we can define

$$\pi^1_N u = \pi_N(u - u_*) + u_* \in P_N \quad \forall u \in H^1(I).$$

Clearly, $(\pi^1_N u)(r) = u(r)$ for $r = 0, 1$, and by Lemma 3.4,

$$\|\partial_r(\pi^1_N u - u)\| + N\|(\pi^1_N u - u)\|_{\hat\omega^{-1}} \lesssim N^{1-s}\|(r - r^2)^{\frac{s-1}{2}}\partial^s_r(u - u_*)\|. \tag{3.17}$$

Since $\partial^s_r u_* \equiv 0$ for $s \geq 2$, and $\partial_r u_* = u(1) - u(0)$, which implies that

$$\|\partial_r u_*\| = |u(1) - u(0)| \lesssim \|\partial_r u\|,$$

the desired result follows from (3.17). □

Using the above corollary leads to the following lemma.

LEMMA 3.5. *For any $u \in X \cap \widetilde{H}^s(I)$ with $s \geq 1, s \in \mathbb{N}$,*

$$\begin{aligned}\|\partial_r(\pi_N^{1,m}u - u)\|_\omega + \sqrt{d_m}\|\pi_N^{1,m}u - u\|_{\omega^{-1}}&\\ \lesssim (1 + \sqrt{d_m}N^{-1})N^{1-s}\|(r-r^2)^{\frac{s-1}{2}}\partial_r^s u\|;&\end{aligned} \tag{3.18a}$$

$$\|\pi_N^{1,m}u - u\|_\omega \lesssim (d_m^{-\frac{1}{2}} + N^{-1})N^{1-s}\|(r-r^2)^{\frac{s-1}{2}}\partial_r^s u\|. \tag{3.18b}$$

Proof. The definition (3.13) implies that for any $\phi \in X_N$,

$$a_m(\pi_N^{1,m}u - u, \pi_N^{1,m}u - u) \leq a_m(\phi - u, \phi - u). \tag{3.19}$$

Taking $\phi = \pi_N^1 u \in X_N$ in (3.19), we obtain from Corollary 3.2 that

$$\begin{aligned}\|\partial_r(\pi_N^{1,m}u - u)\|_\omega + \sqrt{d_m}\|\pi_N^{1,m}u - u\|_{\omega^{-1}} &\lesssim \|\partial_r(\pi_N^1 u - u)\| + \sqrt{d_m}\|\pi_N^1 u - u\|_{\hat{\omega}^{-1}}\\ &\lesssim (1 + \sqrt{d_m}N^{-1})N^{1-s}\|(r-r^2)^{\frac{s-1}{2}}\partial_r^s u\|.\end{aligned}$$

Since $\|\pi_N^{1,m}u - u\|_\omega \leq \|\pi_N^{1,m}u - u\|_{\omega^{-1}}$, (3.18b) follows from (3.18a). □

We can now derive an error estimate for the case $n = 2$ with $m > 0$.

THEOREM 3.3. *If $u \in X \cap \widetilde{H}^s(I)$, with $s \geq 1$ and $s \in \mathbb{N}$, we have*

$$\begin{aligned}&\|\partial_r(u - u_N)\|_\omega + \sqrt{d_m}\|u - u_N\|_{\omega^{-1}} + k\|u - u_N\|_\omega\\ &\lesssim \Big((1 + \sqrt{d_m}N^{-1} + d_m^2 k^{-4}) + k^2(d_m^{-\frac{1}{2}} + N^{-1})\Big)N^{1-s}\|(r-r^2)^{\frac{s-1}{2}}\partial_r^s u\|.\end{aligned} \tag{3.20}$$

Proof. Let us still denote $e_N = u_N - \pi_N^{1,m}u$ and $\tilde{e}_N = u - \pi_N^{1,m}u$. Due to (3.13), the error equation (3.5) becomes

$$\mathcal{B}(e_N, v_N) = -k^2(\tilde{e}_N, v_N)_\omega - \mathrm{i}k\tilde{e}_N(1)\overline{v_N(1)}.$$

Consequently, (3.6) is changed to

$$\|\partial_r e_N\|_\omega^2 + d_m\|e_N\|_{\omega^{-1}}^2 + k^2\|e_N\|_\omega^2 \lesssim k^4\|\tilde{e}_N\|_\omega^2 + k^2(1 + d_m^2 k^{-4})|\tilde{e}_N(1)|^2.$$

Thus, following a procedure similar to that in the proof of Theorem 3.2, and thanks to Lemma 3.5, we can obtain (3.20). □

4. An alternate formulation and its numerical implementation.

In this section, we shall give an alternate formulation for problem (2.1)–(2.3), which is more suitable for implementation and also leads to a convergence rate similar to that of Theorem 3.2.

4.1. The formulation.

We make the transform

$$u(r) = v(r)e^{\mathrm{i}kr}, \quad f(r) = h(r)e^{\mathrm{i}kr}, \quad r \in I, \tag{4.1}$$

and we convert the problem (2.1)–(2.3) to

$$\begin{aligned}&-\frac{1}{r^{n-1}}\partial_r(r^{n-1}\partial_r v) + d_m\frac{v}{r^2} - \mathrm{i}k\Big(2\partial_r v + (n-1)\frac{v}{r}\Big) = h,\\ &r \in I := (0,1),\ n = 1,2,3,\ m \geq 0,\end{aligned} \tag{4.2}$$

where v satisfies the Dirichlet boundary condition (2.2) and the Neumann boundary condition:

$$v'(1) = \tilde{g} := ge^{-\mathrm{i}k}. \tag{4.3}$$

Let the spaces X and X_N be the same as before. The weak formulation of (4.2) with (2.2) and (4.3) is to find $v \in X$ such that

$$\begin{aligned}\widetilde{\mathcal{B}}(v,w) := (\partial_r v, \partial_r w)_{\omega^{n-1}} + d_m(v,w)_{\omega^{n-3}} - 2\mathrm{i}k(\partial_r v, w)_{\omega^{n-1}} \\ - (n-1)\mathrm{i}k(v,w)_{\omega^{n-2}} = (h,w)_{\omega^{n-1}} + \tilde{g}\overline{w(1)} \quad \forall w \in X.\end{aligned} \tag{4.4}$$

The well-posedness of this formulation is guaranteed by (4.1) and Theorem 2.1.

The spectral-Galerkin approximation to (4.4) is to seek $v_N \in X_N$ such that

$$\widetilde{\mathcal{B}}_N(v_N, w_N) = (h, w_N)_{\omega^{n-1}} + \tilde{g}\overline{w_N(1)} \quad \forall w_N \in X_N. \tag{4.5}$$

Using a procedure similar to the one used before, we can derive corresponding a priori estimates and error estimates. For simplicity, we consider the case $g = 0$.

THEOREM 4.1. *Let v and v_N be the solutions of* (4.4) *and* (4.5) *with $\tilde{g} = 0$ and $h \in L^2_{\omega^{n-1}}(I)$. Then*

$$\|\partial_r v\|_{\omega^{n-1}} + \sqrt{d_m}\|v\|_{\omega^{n-3}} \lesssim \|h\|_{\omega^{n-1}}, \tag{4.6}$$

$$\|\partial_r v_N\|_{\omega^{n-1}} + \sqrt{d_m}\|v_N\|_{\omega^{n-3}} \lesssim \|h\|_{\omega^{n-1}}. \tag{4.7}$$

Proof. As in the proof of Theorem 2.2, we take two different test functions in (4.4). We first take $w = v$ in (4.4), whose real part is

$$\|\partial_r v\|^2_{\omega^{n-1}} + d_m\|v\|^2_{\omega^{n-3}} + 2k\mathrm{Im}(\partial_r v, v)_{\omega^{n-1}} = \mathrm{Re}(h,v)_{\omega^{n-1}}, \tag{4.8}$$

and using integration by parts, its imaginary part becomes

$$-2k\mathrm{Re}(\partial_r v, v)_{\omega^{n-1}} - (n-1)k\|v\|^2_{\omega^{n-2}} = -k|v(1)|^2 = \mathrm{Im}(h,v)_{\omega^{n-1}}. \tag{4.9}$$

Here, in the derivation of (4.8) (likewise for (4.10) below), we have used the fact $\mathrm{Re}(\mathrm{i}(u,v)) = -\mathrm{Im}(u,v)$.

Next, we take $w = 2r\partial_r v$ ($\in X$) in (4.4), and thanks to (2.24a)–(2.24b), its real part becomes

$$\begin{aligned}(2-n)(\|\partial_r v\|^2_{\omega^{n-1}} + d_m\|v\|^2_{\omega^{n-3}}) + d_m|v(1)|^2 \\ + 2(n-1)k\mathrm{Im}(v, \partial_r v)_{\omega^{n-1}} = 2\mathrm{Re}(h, r\partial_r v)_{\omega^{n-1}}.\end{aligned} \tag{4.10}$$

As a consequence of (4.10), we have that for $n = 1$ (we recall that $d_m = 0$ in this case),

$$\|\partial_r v\|^2 \le 2\|h\|_{\omega^2}\|\partial_r v\| \le 2\|h\|\|\partial_r v\|, \tag{4.11}$$

which implies (4.6) with $n = 1$.

It remains to prove (4.6) with $n = 2, 3$. Since $\partial_r v(1) = 0$, it is easy to verify

$$\mathrm{Im}(\partial_r v, v)_{\omega^{n-1}} = -\mathrm{Im}(v, \partial_r v)_{\omega^{n-1}}. \tag{4.12}$$

Therefore, multiplying (4.8) by $n-1$ and adding the resulting equation to (4.10), we derive from the Cauchy–Schwarz inequality that

$$\begin{aligned}\|\partial_r v\|^2_{\omega^{n-1}} + d_m\|v\|^2_{\omega^{n-3}} + d_m|v(1)|^2 = (n-1)\mathrm{Re}(h,v)_{\omega^{n-1}} \\ + 2\mathrm{Re}(h, r\partial_r v)_{\omega^{n-1}} \le 2\|h\|_{\omega^{n-1}}\|v\|_{\omega^{n-1}} + \frac{1}{4}\|\partial_r v\|^2_{\omega^{n-1}} + 4\|h\|^2_{\omega^{n-1}}.\end{aligned} \tag{4.13}$$

Clearly, we have

$$|v(1)|^2 = \int_0^1 \partial_r(|v(r)|^2 r^n)dr = n\int_0^1 |v(r)|^2 r^{n-1}dr + 2\int_0^1 \partial_r v(r)\overline{v(r)}r^n dr,$$

and by the Cauchy–Schwarz inequality,

$$n\|v\|^2_{\omega^{n-1}} \le |v(1)|^2 + 2\|v\|_{\omega^{n-1}}\|\partial_r v\|_{\omega^{n+1}} \le |v(1)|^2 + \frac{n}{2}\|v\|^2_{\omega^{n-1}} + \frac{2}{n}\|\partial_r v\|^2_{\omega^{n-1}},$$

which together with (4.9) leads to

$$\begin{aligned}\|v\|^2_{\omega^{n-1}} &\le \frac{2}{n}|v(1)|^2 + \frac{4}{n^2}\|\partial_r v\|^2_{\omega^{n-1}} \le \frac{2}{nk}|\mathrm{Im}(h,v)_{\omega^{n-1}}| + \frac{4}{n^2}\|\partial_r v\|^2_{\omega^{n-1}}\\ &\le \frac{1}{2}\|v\|^2_{\omega^{n-1}} + \frac{2}{n^2k^2}\|h\|^2_{\omega^{n-1}} + \frac{4}{n^2}\|\partial_r v\|^2_{\omega^{n-1}}.\end{aligned} \tag{4.14}$$

As a result of (4.13) and (4.14), we obtain

$$\begin{aligned}\|\partial_r v\|^2_{\omega^{n-1}} + d_m\|v\|^2_{\omega^{n-3}} &\le 2\|h\|_{\omega^{n-1}}\Big(\frac{2}{nk}\|h\|_{\omega^{n-1}} + \frac{2\sqrt{2}}{n}\|\partial_r v\|_{\omega^{n-1}}\Big)\\ &+ \frac{1}{4}\|\partial_r v\|^2_{\omega^{n-1}} + 4\|h\|^2_{\omega^{n-1}} \le \frac{1}{2}\|\partial_r v\|^2_{\omega^{n-1}} + \Big(\frac{4}{nk} + \frac{32}{n^2} + 4\Big)\|h\|^2_{\omega^{n-1}}.\end{aligned}$$

This completes the proof of (4.6).

Since $r\partial_r v_N \in X_N$, we have the same results for the numerical solution v_N. □

Thanks to the above theorem, we can derive the following convergence result by using an argument similar to the proof of Theorem 3.2.

THEOREM 4.2. *Let* v *and* v_N *be, respectively, the solutions of* (4.4) *and* (4.5) *with* $\tilde{g} = 0$, *and we have*

(i) *for* $n = 1, 3$ *or* $n = 2, m = 0$, *and* $v \in X \cap \widetilde{H}^s_{\omega^{n-1}}(I)$ *with* $s \ge 1$ *and* $s \in \mathbb{N}$,

$$\|\partial_r(v - v_N)\|_{\omega^{n-1}} + \sqrt{d_m}\|v - v_N\|_{\omega^{n-3}} \lesssim (k + \sqrt{d_m})N^{1-s}\|(r - r^2)^{\frac{s-1}{2}}\partial_r^s v\|_{\omega^{n-1}}; \tag{4.15}$$

(ii) *for* $n = 2,\ m > 0$, *and* $v \in X \cap \widetilde{H}^s(I)$, *with* $s \ge 1$ *and* $s \in \mathbb{N}$,

$$\begin{aligned}&\|\partial_r(v - v_N)\|_{\omega} + \sqrt{d_m}\|v - v_N\|_{\omega^{-1}}\\ &\quad\lesssim \Big((1 + \sqrt{d_m}N^{-1}) + k^2(d_m^{-\frac{1}{2}} + N^{-1})\Big)N^{1-s}\|(r - r^2)^{\frac{s-1}{2}}\partial_r^s v\|.\end{aligned} \tag{4.16}$$

Proof. Let $\Pi^{1,m}_{N,n}$ be the orthogonal projection defined in (3.2), and denote $e_N = v_N - \Pi^{1,m}_{N,n}v$ and $\hat{e}_N = v - \Pi^{1,m}_{N,n}v$. Like (3.5), the error equation is

$$\widetilde{\mathcal{B}}(e_N, w_N) = d_m(\hat{e}_N, w_N)_{\omega^{n-3}} - 2\mathrm{i}k(\partial_r\hat{e}_N, w_N)_{\omega^{n-1}} - (n-1)\mathrm{i}k(\hat{e}_N, w_N)_{\omega^{n-2}}.$$

Therefore, taking the test function $w_N = e_N, r\partial_r e_N$, setting $h = -2\mathrm{i}k\partial_r\hat{e}_N - (n-1)\mathrm{i}kr^{-1}\hat{e}_N$ in (4.5), and dealing with the term $d_m(\hat{e}_N, w_N)_{\omega^{n-3}}$ the same as that in the proof of Lemma 3.2, we obtain

$$\|\partial_r e_N\|^2_{\omega^{n-1}} + d_m\|e_N\|^2_{\omega^{n-3}} \lesssim (k^2 + d_m)(\|\partial_r\hat{e}_N\|^2_{\omega^{n-1}} + \|\hat{e}_N\|^2_{\omega^{n-3}}).$$

The rest of the proof of (4.15) is similar to that of Theorem 3.2.

The estimate (4.16) can be proved in the same fashion by using the results in Lemma 3.5. □

4.2. Numerical implementations.

4.2.1. Choice of basis functions. Without loss of generality, we still assume that $\tilde{g} = 0$ in (4.2). For computational convenience, we transform $I = (0, 1)$ to the reference interval $\hat{I} = (-1, 1)$ with $x = 2r - 1$, $r = \frac{1}{2}(1 + x)$, $r \in I$, $x \in \hat{I}$. As demonstrated in [22, 23], it is advantageous to construct basis function satisfying the underlying homogeneous boundary conditions by using compact combinations of orthogonal polynomials. Hence, we define

$$W_N = W_N^{(m,n)} := \{w \in P_N : w'(1) = 0;\ \ w(-1) = 0 \text{ if } n = 1 \text{ and if } n = 2 \text{ with } m > 0\},$$

and we let $L_l(x)$ denote the Legendre polynomial of degree l. Define

$$\begin{aligned} &\phi_j(x) := (L_j(x) + L_{j+1}(x)) - \Big(\frac{j+1}{j+2}\Big)^2 (L_{j+1}(x) + L_{j+2}(x)); \\ &\psi_j(x) := L_j(x) - \frac{j}{j+2} L_{j+1}(x). \end{aligned} \tag{4.17}$$

Since $L_l(-1) = (-1)^l$ and $L_l'(1) = \frac{1}{2} l(l + 1)$, one can verify easily that

$$\phi_j(-1) = \phi_j'(1) = \psi_j'(1) = 0. \tag{4.18}$$

Hence, for $n = 1$ or $n = 2$ with $m > 0$, $W_N^{(m,n)} = \text{span}\{\phi_j : j = 0, 1, \ldots, N - 2\}$; and for $n = 3$ or $n = 2$ with $m = 0$, $W_N^{(m,n)} = \text{span}\{\psi_j : j = 0, 1, \ldots, N - 1\}$.

Now, let us write

$$v_N(r) := w_N^R(x) + \mathrm{i} w_N^I(x), \quad 2^{n-3} r^{n-1} h(r) := q^R(x) + \mathrm{i} q^I(x), \tag{4.19}$$

where w_N^R, w_N^I, q^R, and q^I are real functions in $\hat{I}$. Our spectral-Galerkin algorithm is to seek $w_N^R, w_N^I \in W_N$ such that for any real polynomials $\phi, \psi \in W_N$,

$$\begin{aligned} &((1 + x)^{n-1} \partial_x w_N^R, \partial_x \phi) + d_m((1 + x)^{n-3} w_N^R, \phi) + k((1 + x)^{n-1} \partial_x w_N^I, \phi) \\ &\qquad + \frac{n-1}{2} k((1 + x)^{n-2} w_N^I, \phi) = (q^R, \phi); \\ &((1 + x)^{n-1} \partial_x w_N^I, \partial_x \psi) + d_m((1 + x)^{n-3} w_N^I, \psi) - k((1 + x)^{n-1} \partial_x w_N^R, \psi) \\ &\qquad - \frac{n-1}{2} k((1 + x)^{n-2} w_N^R, \phi) = (q^I, \psi). \end{aligned} \tag{4.20}$$

Thanks to the nice properties of the Legendre polynomials, one can find that the coefficient matrix of the above system is sparse, and its nonzero entries can be determined exactly.

4.2.2. Numerical results. We present some numerical results for the problem (2.1)–(2.3) by using the schemes proposed above.

Example 1. We consider (2.1)–(2.3) with $n = 2$, $d_m = 100$, and $g = 0$ and set the exact solution to be

$$u(r) = v(r) e^{\mathrm{i}kr}, \quad r \in I, \tag{4.21}$$

where $v(r) = (\cos 2k - \cos(2k(1-r))) + \mathrm{i}(\frac{1}{k}(\sin 2k - \sin(2k(1-r))) - 2r \cos(2k(1-r)))$ is the exact solution of the transformed problem (4.2).

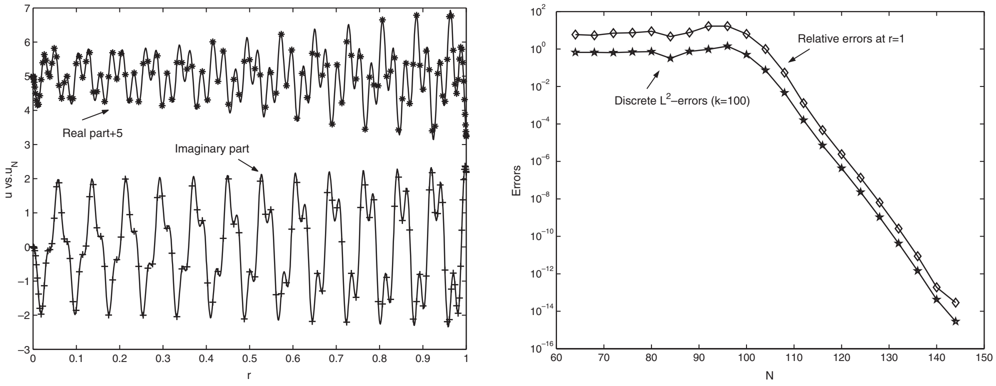


FIG. 4.1. *Left: exact solution vs. numerical solution. Right: errors vs. N $(k = 100)$.*

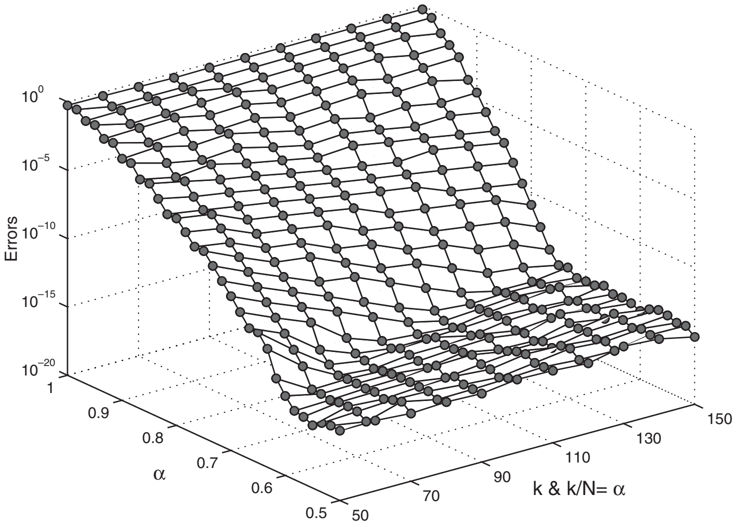


FIG. 4.2. *Errors vs. $\alpha \in [0.5, 1]$ and $k \in [50, 150]$ with $\frac{k}{N} = \alpha$.*

In Figure 4.1 (left), we plot the numerical solution at Legendre–Gauss–Lobatto points with $k = 80$ and $N = 96$ (asterisk-markers for the real part (raised by 5 unit) and plus-markers for the imaginary part) vs. the exact solution (solid line).

We now examine the convergence rate. According to Theorem 3.3, the predicted order of convergence for the exact solution (4.21) is

$$\|u - u_N\|_\omega \sim k^{1+s} N^{1-s}, \quad N \gg 1, \;\; k > 0, \;\; s \geq 1. \tag{4.22}$$

In Figure 4.1 (right), we fix the wave number $k = 100$ and plot the discrete L^2-errors and relative errors at $r = 1$ vs. different modes N. As expected, an exponential convergence rate is observed once N is large enough to resolve the oscillation.

Next, we fix $\alpha = \frac{k}{N}$ and examine the error behavior with respect to α. In Figure 4.2, we plot the discrete L^2-errors with $0.5 \leq \alpha \leq 1, \; 50 \leq k \leq 150$, and $N = \frac{k}{\alpha}$. The results indicate that the proposed scheme can provide very accurate approximations to highly oscillatory solutions under the condition $\frac{k}{N} = \alpha < 1$, which is necessary for

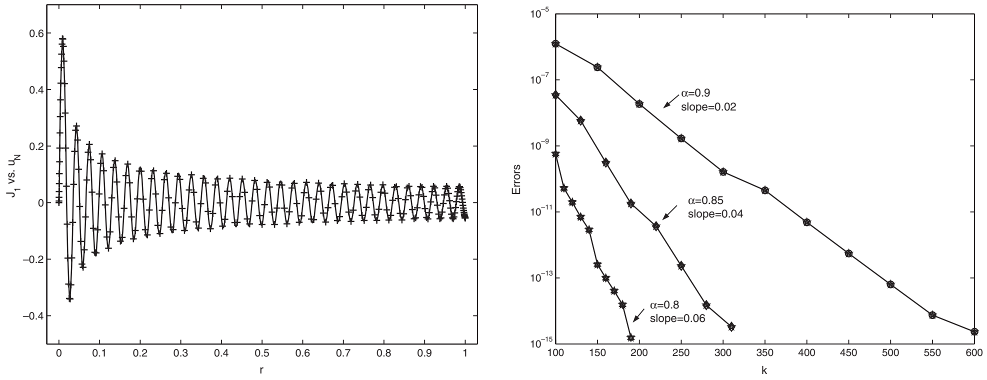


FIG. 4.3. *Left: exact solution vs. numerical solution. Right: errors vs. wave number k with* $\alpha = \frac{k}{N}$ *fixed.*

convergence (cf. [13]).

Example 2. We consider the problem (2.1)–(2.3) with $n = 2$ and $d_m = 1$. An exact solution is

$$u(r) = J_1(kr), \quad \text{with} \quad f \equiv 0 \;\text{ and }\; g = k(J_1'(k) - \mathrm{i}J_1(k)), \tag{4.23}$$

where $J_1(\cdot)$ is the first degree Bessel function of the first kind. As pointed out in [26], we have the following asymptotic property:

$$u(r) = J_1(kr) = \sqrt{\frac{2}{\pi kr}} \cos\left(kr - \frac{3}{4}\pi\right) + O((kr)^{-\frac{3}{2}}) \quad \text{if } kr \gg 1. \tag{4.24}$$

Hence, the solution is highly oscillating when the wave number k is large (see Figure 4.3 (left)). We derive from (4.24) that the expected convergence rate is $k^{\frac{1}{2}+s}N^{1-s}$. In Figure 4.3 (left), we plot the exact solution vs. the numerical solution with $k = 200$ and $N = 256$. In this case, the discrete L^2-error is 2.45×10^{-15} and relative error at $r = 1$ is 3.84×10^{-13}. The error behaviors with several fixed $\alpha = \frac{k}{N}$ are plotted in Figure 4.3 (right), which demonstrates that the spectral-Galerkin method is capable of providing very accurate results even for α close to 1.

5. Extensions to multidimensional cases. The results we derived for the prototypical 1-D problem (2.1)–(2.3) (with $n = 2, 3$) can be used to derive error estimates for the spectral-Galerkin approximation to the multidimensional problem (1.2). As an example, we consider the case $n = 3$:

$$\begin{aligned} -\Delta U - k^2 U = F \quad & \text{in } \hat{\Omega} := \{(x,y,z) \;:\; a^2 < x^2+y^2+z^2 < b^2\}, \\ \partial_r U - \mathrm{i}kU = G \quad & \text{on } S_b := \{(x,y,z) \;:\; x^2+y^2+z^2 = b^2\}, \\ U = 0 \quad & \text{on } \;S_a := \{(x,y,z) \;:\; x^2+y^2+z^2 = a^2\} \text{ if } a > 0. \end{aligned} \tag{5.1}$$

Applying the spherical transformation

$$x = r\cos\theta\sin\phi, \quad y = r\sin\theta\sin\phi, \quad z = r\cos\phi \tag{5.2}$$

to (5.1) and setting $u(r,\theta,\phi)=U(x,y,z),\ f(r,\theta,\phi)=F(x,y,z),\ g(\theta,\phi)=G(x,y,z)$, and $S:=[0,2\pi)\times[0,\pi)$, we obtain

$$\begin{aligned} -\Big(\frac{\partial^2}{\partial r^2}+\frac{2}{r}\frac{\partial}{\partial r}+\frac{1}{r^2}\Delta_S\Big)u-k^2u&=f \quad \text{in } \Omega:=(a,b)\times S,\\ \partial_r u-\mathrm{i}ku&=g \quad \text{on } S_b,\\ u&=0 \quad \text{on } S_a \ \text{ if } a>0, \end{aligned} \tag{5.3}$$

where Δ_S is the Laplace–Beltrami operator (the Laplacian on the unit sphere S):

$$\Delta_S=\frac{1}{\sin^2\phi}\frac{\partial^2}{\partial\theta^2}+\frac{\cos\phi}{\sin\phi}\frac{\partial}{\partial\phi}+\frac{\partial^2}{\partial^2\phi}. \tag{5.4}$$

We recall that the spherical harmonic functions $\{Y_{l,m}\}$ are the eigenfunctions of the Laplace–Beltrami operator (see [25])

$$-\Delta_S Y_{l,m}(\theta,\phi)=m(m+1)Y_{l,m}(\theta,\phi) \tag{5.5}$$

and are defined by

$$Y_{l,m}(\theta,\phi)=\sqrt{\frac{(2m+1)(m-l)!}{4\pi(m+l)!}}e^{il\theta}P_m^l(\cos\phi), \quad m\ge|l|\ge0,$$

where $P_m^l(x)$ is the associated Legendre functions given by

$$P_m^l(x)=\frac{(-1)^l}{2^m m!}(1-x^2)^{\frac{l}{2}}\frac{d^{m+l}}{dx^{m+1}}\{(x^2-1)^m\}.$$

The set of harmonic functions forms a complete orthonormal system in $L^2(S)$, i.e.,

$$\int_0^{2\pi}\int_0^{\pi}Y_{l,m}(\theta,\phi)\overline{Y_{l',m'}}(\theta,\phi)\sin\phi\, d\phi d\theta=\delta_{l,l'}\delta_{m,m'}. \tag{5.6}$$

Hence, for any function $U(x,y,z)\in L^2(\hat\Omega)$, the function $u(r,\theta,\phi)=U(x,y,z)$ can be expanded as

$$u=\sum_{|l|=0}^{\infty}\sum_{m\ge|l|}^{\infty}u_{lm}(r)Y_{l,m}(\theta,\phi), \ \text{ with } \ u_{lm}(r)=\int_S u(r,\theta,\phi)\overline{Y}_{l,m}(\theta,\phi)dS, \tag{5.7}$$

and we have

$$\|u\|^2_{L^2_{\omega^2}(\Omega)}=\sum_{|l|=0}^{\infty}\sum_{m\ge|l|}^{\infty}\|u_{lm}\|^2_{\omega^2}=\|U\|^2_{L^2(\hat\Omega)} \quad (\omega^2=r^2). \tag{5.8}$$

For a scalar function v on S, the gradient operator $\vec\nabla_S$ on the unit sphere is defined by $\vec\nabla_S v=\big(\frac{1}{\sin\phi}\partial_\theta v,\partial_\phi v\big)$. One can verify readily that

$$-(\Delta_S u,v)_S=(\vec\nabla_S u,\vec\nabla_S v)_S \quad \forall u,v\in\mathcal{D}(\Delta_S), \tag{5.9}$$

where $\mathcal{D}(\Delta_S)$ is the domain of the Laplace–Beltrami operator Δ_S. In particular, as a consequence of (5.5)–(5.9), we have

$$(\vec\nabla_S Y_{l,m},\vec\nabla_S Y_{l,m})_S=m(m+1), \quad m\ge|l|\ge0. \tag{5.10}$$

Accordingly, we can define the Sobolev space on S:

$$H^1(S) := \{u : u \text{ is measurable on } S \text{ and } \|u\|^2_{H^1(S)} < \infty\},$$

where $\|u\|_{H^1(S)} = \big(\|u\|^2_{L^2(S)} + \|\vec{\nabla}_S u\|^2_{L^2(S)}\big)^{\frac{1}{2}}$.

The variational formulation of (5.3) is to find $u \in V := H^1_{\omega^2}(I; L^2(S)) \cap L^2(I; H^1(S))$ such that $(\omega^2 = r^2)$

$$\begin{aligned} a(u, v) &:= (\partial_r u, \partial_r v)_{\omega^2,\Omega} + (\vec{\nabla}_S u, \vec{\nabla}_S v)_\Omega - k^2(u, v)_{\omega^2,\Omega} \\ &\quad - \mathrm{i}kb^2(u(b,\cdot), v(b,\cdot))_S = (f, v)_{\omega^2,\Omega} + b^2(g, v(b,\cdot))_S \quad \forall v \in V. \end{aligned} \tag{5.11}$$

The spectral-Galerkin approximation of (5.11) is to find $u_{MN} \in V_{MN}$ such that

$$a(u_{MN}, v) = (f, v)_{\omega^2,\Omega} + b^2(g, v(b,\cdot))_S \quad \forall v \in V_{MN}, \tag{5.12}$$

where $V_{MN} := W_M \times X_N$, and

$$W_M := \text{span}\{Y_{l,m} :\ 0 \le |l| \le m \le M\},\ X_N := \{u \in P_N\ :\ u(a) = 0 \ \text{ if }\ a > 0\}.$$

Hence, we can write

$$(u(r,\theta,\phi), f(r,\theta,\phi), g(\theta,\phi)) = \sum_{|l|=0}^{\infty} \sum_{m\ge|l|}^{\infty} (u_{lm}(r), f_{lm}(r), g_{lm}) Y_{l,m}(\theta,\phi); \tag{5.13a}$$

$$u_{MN}(r,\theta,\phi) = \sum_{|l|=0}^{M} \sum_{m\ge|l|}^{M} u^N_{lm}(r) Y_{l,m}(\theta,\phi). \tag{5.13b}$$

In order to describe the error bounds, we define a nonisotropic space $\widetilde{H}^s_{\omega^2}(I; H^t(S))$ as follows:

$$\widetilde{H}^s_{\omega^2}(I; H^t(S)) = \left\{ u \in L^2_{\omega^2}(\Omega) : \sum_{|l|=0}^{\infty} \sum_{m\ge|l|}^{\infty} m^t(m+1)^t \|u_{lm}\|^2_{\widetilde{H}^s_{\omega^2}(I)} < +\infty \right\}, \tag{5.14}$$

where $\{u_{lm}\}$ are the expansion coefficients of u in terms of $Y_{l,m}$ as in (5.7). Thanks to (5.10), we can define the norm on $\widetilde{H}^s_{\omega^2}(I; H^t(S))$ by

$$\|u\|_{\widetilde{H}^s_{\omega^2}(I;H^t(S))} = \left(\sum_{|l|=0}^{\infty} \sum_{m\ge|l|}^{\infty} m^t(m+1)^t \|u_{lm}\|^2_{\widetilde{H}^s_{\omega^2}(I)} \right)^{\frac{1}{2}} \tag{5.15}$$

and its seminorm by replacing $\|u_{lm}\|_{\widetilde{H}^s_{\omega^2}(I)}$ with $|u_{lm}|_{\widetilde{H}^s_{\omega^2}(I)}$. In particular, $L^2_{\omega^2}(I; H^t(S)) = \widetilde{H}^0_{\omega^2}(I; H^t(S))$ and $\widetilde{H}^s_{\omega^2}(I; L^2(S)) = \widetilde{H}^s_{\omega^2}(I; H^0(S))$.

5.1. In a sphere ($a = 0$). Without loss of generality, we assume that $b = 1$. In this case, we can show that $\{u_{lm}\}$ (resp., $\{u^N_{lm}\}$) satisfy the 1-D problem (2.5) (resp., (3.1)) with $n = 3$ and f, g being replaced by f_{lm} and g_{lm}, respectively.

THEOREM 5.1. *Let u and u_{MN} be, respectively, the solutions of* (5.11) *and* (5.12), *and denote $e = u - u_{MN}$. Then if*

$$u \in L^2(I; H^t(S)) \cap H^1_{\omega^2}(I; H^{t-1}(S)) \cap \widetilde{H}^s_{\omega^2}(I; L^2(S)), \quad s, t \ge 1,\ s, t \in \mathbb{N}, \tag{5.16}$$

we have

$$\begin{aligned} &\|\partial_r e\|_{L^2_{\omega^2}(\Omega)} + \|\vec{\nabla}_S e\|_{L^2(\Omega)} + k\|e\|_{L^2_{\omega^2}(\Omega)} \\ &\qquad \lesssim C_*\Big((M + M^4k^{-4} + k^2N^{-1})N^{1-s} + M^{1-t}(1 + kM^{-1})\Big), \end{aligned} \tag{5.17}$$

where C_* is a positive constant depending only on the seminorms of u in the spaces mentioned in (5.16).

Proof. Let $e_{lm}(r) = u_{lm}(r) - u_{lm}^N(r)$. We deduce from Theorem 3.2 that

$$\begin{aligned} &\|\partial_r e_{lm}\|_{L^2_{\omega^2}(I)} + \sqrt{d_m}\|e_{lm}\|_{L^2(I)} + k\|e_{lm}\|_{L^2_{\omega^2}(I)} \\ &\qquad \lesssim \Big(1 + \sqrt{d_m} + d_m^2k^{-4} + k^2N^{-1}\Big)N^{1-s}|e_{lm}|_{\widetilde{H}^s_{\omega^2}(I)}, \end{aligned} \tag{5.18}$$

where $d_m = m(m+1)$. Therefore, by (5.6)–(5.10) and (5.13b)–(5.14),

$$\begin{aligned} &\|\partial_r e\|^2_{L^2_{\omega^2}(\Omega)} + \|\vec{\nabla}_S e\|^2_{L^2(\Omega)} + k^2\|e\|^2_{L^2_{\omega^2}(\Omega)} \\ &\quad = \sum_{|l|=0}^{M}\sum_{m\ge|l|}^{M}\Big(\|\partial_r e_{lm}\|^2_{L^2_{\omega^2}(I)} + d_m\|e_{lm}\|^2_{L^2(I)} + k^2\|e_{lm}\|^2_{L^2_{\omega^2}(I)}\Big) \\ &\qquad + \left(\sum_{|l|=0}^{\infty}\sum_{m>M}^{\infty} + \sum_{|l|>M}^{\infty}\sum_{m\ge|l|}^{\infty}\right)\Big(\|\partial_r u_{lm}\|^2_{L^2_{\omega^2}(I)} + d_m\|u_{lm}\|^2_{L^2(I)} + k^2\|u_{lm}\|^2_{L^2_{\omega^2}(I)}\Big) \\ &\quad \lesssim \Big(1 + \sqrt{d_M} + d_M^2k^{-4} + k^2N^{-1}\Big)^2 N^{2-2s}\sum_{|l|=0}^{M}\sum_{m\ge|l|}^{M}|u_{lm}|^2_{\widetilde{H}^s_{\omega^2}(I)} \\ &\qquad + d_M^{1-t}\sum_{|l|=0}^{\infty}\sum_{m\ge|l|}^{\infty}\Big(d_m^{t-1}(\|\partial_r u_{lm}\|^2_{L^2_{\omega^2}(I)} + d_m\|u_{lm}\|^2_{L^2(I)} + k^2\|u_{lm}\|^2_{L^2_{\omega^2}(I)})\Big) \\ &\quad \lesssim (M + M^4k^{-4} + k^2N^{-1})^2N^{2-2s}|u|^2_{\widetilde{H}^s_{\omega^2}(I;L^2(S))} \\ &\qquad + d_M^{1-t}\Big(|u|^2_{H^1_{\omega^2}(I;H^{t-1}(S))} + |u|^2_{L^2(I;H^t(S))} + k^2d_m^{-1}|u|^2_{L^2_{\omega^2}(I;H^t(S))}\Big), \end{aligned}$$

which implies the desired result. □

5.2. In a spherical shell ($a > 0$). In this case, $\{u_{lm}\}$ are the solutions of

$$\widehat{\mathcal{B}}_{lm}(u_{lm}, v) = (f_{lm}, v)_{\omega^2} + b^2 g_{lm}\overline{v(b)} \quad \forall v \in X, \;\; 0 \le |l| \le m, \tag{5.19}$$

where $X := \{u \in H^1(I) \;:\; u(a) = 0\}$, and

$$\widehat{\mathcal{B}}_{lm}(u, v) := (\partial_r u, \partial_r v)_{\omega^2} + d_m(u, v) - k^2(u, v)_{\omega^2} - \mathrm{i}kb^2u(b)\overline{v(b)}, \tag{5.20}$$

with $\omega^2 = r^2$, $d_m = m(m+1)$. The numerical approximations u_{lm}^N $(0 \le |l| \le m,\ m = 0, 1, \ldots, M)$ are defined by

$$\widehat{\mathcal{B}}_{lm}(u_{lm}^N, v_N) = (f_{lm}, v_N)_{\omega^2} + b^2 g_{lm}\overline{v_N(b)} \quad \forall v_N \in X_N := X \cap P_N. \tag{5.21}$$

Since u_{lm}, $(r-a)u_{lm} \in X$ (resp., u_{lm}^N, $(r-a)u_{lm}^N \in X_N$), we can use them as test functions in (5.19) (resp., (5.21)), and derive the following results using an argument analogous to that in the proof of Theorem 2.2.

LEMMA 5.1. *Let* $\{u_{lm}\}$ *and* $\{u_{lm}^N\}$ *be, respectively, the solution of* (5.19) *and* (5.21)*. Then there exists* $\xi \in (a,b)$ *such that for* $C_\xi := (2-\frac{2a}{\xi})^{-1}$*, we have*

$$\|\partial_r u_{lm}\|^2_{L^2_{\omega^2}(I)} + d_m\|u_{lm}\|^2_{L^2(I)} + k^2\|u_{lm}\|^2_{L^2_{\omega^2}(I)} \lesssim C_\xi b^3(|g_{lm}|^2 + b^2\|f_{lm}\|^2_{L^2_{\omega^2}(I)}),$$
$$\|\partial_r u^N_{lm}\|^2_{L^2_{\omega^2}(I)} + d_m\|u^N_{lm}\|^2_{L^2(I)} + k^2\|u^N_{lm}\|^2_{L^2_{\omega^2}(I)} \lesssim C_\xi b^3(|g_{lm}|^2 + b^2\|f_{lm}\|^2_{L^2_{\omega^2}(I)}).$$

The above a priori estimates allow us to perform the error analysis for the spherical shell case. Similar to the case $a = 0$, we can prove the following.

THEOREM 5.2. *Let* u *and* u_{MN} *be, respectively, the solutions of* (5.11) *and* (5.12)*, and denote* $e = u - u_{MN}$*. Then if*

$$u \in L^2((a,b);H^t(S)) \cap H^1_{\omega^2}((a,b);H^{t-1}(S)) \cap \widetilde{H}^s_{\omega^2}((a,b);L^2(S)), \quad s,t \geq 1,\ s,t \in \mathbb{N},$$

there exists $\xi \in (a,b)$ *such that for* $C_\xi := (2-\frac{2a}{\xi})^{-1}$*, we have*

$$\|\partial_r e\|_{L^2_{\omega^2}(\Omega)} + \|\vec{\nabla}_S e\|_{L^2(\Omega)} + k\|e\|_{L^2_{\omega^2}(\Omega)}$$
$$\lesssim C_* b^2(1+\sqrt{C_\xi})\Big((M + M^4k^{-4} + k^2N^{-1})N^{1-s} + M^{1-t}(1+kM^{-1})\Big),$$

where C_* *is a positive constant depending only on the seminorms of* u *in the spaces mentioned in* (5.16)*.*

Remark 5.1. A similar procedure can be performed for the Helmholtz equation (1.2) in a 2-D axisymmetric domain ($n = 2$) by using a Fourier expansion in the θ-direction.

6. Concluding remarks. We presented in this paper a complete error analysis and an efficient numerical algorithm for the spectral-Galerkin approximation of the Helmholtz equation with high wave numbers in a 1-D domain as well as in multidimensional radial and spherical symmetric domains.

Our analysis is made possible by using two new arguments: (i) we employed a new procedure advocated in [7] which allowed us to obtain sharp (in terms of k) a priori estimates for both the continuous and discrete problems; (ii) we used new Jacobi and generalized Jacobi approximation results developed recently in [16, 24] which enabled us to derive optimal estimates for the cases $n = 2, 3$ which involve degenerate/singular coefficients.

Unlike in most of the previous studies on the approximation of the Helmholtz equation with high wave numbers, our analysis does not rely on explicit knowledge of continuous/discrete Green's functions and is valid without any restriction on the wave number k and the discretization parameter N. Hence, it is possible to extend our results to more complex problems such as Helmholtz equations in an inhomogeneous medium and to more complex domains through a suitable mapping or a domain perturbation technique.

Acknowledgment. The authors would like to thank Dr. Xiaobing Feng for many stimulating discussions and helpful suggestions.

REFERENCES

[1] I. M. BABUŠKA AND S. A. SAUTER, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM J. Numer. Anal., 34 (1997), pp. 2392–2423.

[2] J. P Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[3] C. Bernardi and Y. Maday, *Spectral methods*, in Handbook of Numerical Analysis, Vol. 5 (Part 2), P. G. Ciarlet and L. L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 209–485.

[4] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.

[5] R. D. Ciskowski and C. A. Brebbia, *Boundary Element Methods in Acoustics*, Elsevier, London, 1991.

[6] P. Cummings, *Analysis of Finite Element Based Numerical Methods for Acoustic Waves, Elastic Waves, and Fluid-Solid Iterations in the Frequency Domain*, Ph.D. thesis, University of Tennessee, Knoxville, 2001.

[7] P. Cummings and X. B. Feng, *Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations*, Math. Models Methods Appl. Sci., to appear.

[8] J. Douglas, J. E. Santos, D. Sheen, and L. S. Bennethum, *Frequency domain treatment of one-dimensional scalar waves*, Math. Models Methods Appl. Sci., 3 (1993), pp. 171–194.

[9] J. Douglas, Jr., D. Sheen, and J. E. Santos, *Approximation of scalar waves in the space-frequency domain*, Math. Models Methods Appl. Sci., 4 (1994), pp. 509–531.

[10] X. Feng and D. Sheen, *An elliptic regularity coefficient estimate for a problem arising from a frequency domain treatment of waves*, Trans. Amer. Math. Soc., 346 (1994), pp. 475–487.

[11] K. Gerdes and L. Demkowicz, *Solution of* 3*D-Laplace and Helmholtz equations in exterior domains using hp-infinite elements*, Comput. Methods Appl. Mech. Engrg., 137 (1996), pp. 239–273.

[12] K. Gerdes and F. Ihlenburg, *On the pollution effect in FE solutions of the* 3*D-Helmholtz equation*, Comput. Methods Appl. Mech. Engrg., 170 (1999), pp. 155–172.

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

[14] M. J. Grote and J. B. Keller, *On non-reflecting boundary conditions*, J. Comput. Phys., 122 (1995), pp. 231–243.

[15] B. Y. Guo and L. L. Wang, *Jacobi interpolation approximations and their applications to singular differential equations*, Adv. Comput. Math., 14 (2001), pp. 227–276.

[16] B. Y. Guo and L. L. Wang, *Jacobi approximations in non-uniformly Jacobi-weighted Sobolev spaces*, J. Approx. Theory, 128 (2004), pp. 1–41.

[17] G. H. Hardy, J. E. Littlewood, and G. Pólya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1952.

[18] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number, Part* I*: The h-version of FEM*, Comput. Math. Appl., 30 (1995), pp. 9–37.

[19] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number. Part* II: *The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.

[20] F. John, *Partial Differential Equations*, 4th ed., Springer-Verlag, New York, 1982.

[21] A. H. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.

[22] J. Shen, *Efficient spectral-Galerkin method.* I*. Direct solvers for second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.

[23] J. Shen, *Efficient spectral-Galerkin methods* III: *Polar and cylindrical geometries*, SIAM J. Sci. Comput., 18 (1997), pp. 1583–1604.

[24] J. Shen and L. L. Wang, *Error analysis for mapped Jacobi spectral methods*, J. Sci. Comput., to appear.

[25] I. N. Sneddon, *Special Functions of Mathematical Physics and Chemistry*, 3rd ed., Longman, New York, 1980.

[26] G. N. Watson, *A Treatise of the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, UK, 1966.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 645–671

AN ADAPTIVE PERFECTLY MATCHED LAYER TECHNIQUE FOR TIME-HARMONIC SCATTERING PROBLEMS*

ZHIMING CHEN† AND XUEZHE LIU‡

Abstract. We develop an adaptive perfectly matched layer (PML) technique for solving the time-harmonic scattering problems. The PML parameters such as the thickness of the layer and the fictitious medium property are determined through sharp a posteriori error estimates. The derived finite element a posteriori estimate for adapting meshes has the nice feature that it decays exponentially away from the boundary of the fixed domain where the PML layer is placed. This property makes the total computational costs insensitive to the thickness of the PML absorbing layers. Numerical experiments are included to illustrate the competitive behavior of the proposed adaptive method.

Key words. adaptivity, perfectly matched layer, a posteriori error analysis, scattering problems

AMS subject classifications. 65N30, 78A45, 35Q60

DOI. 10.1137/040610337

1. Introduction.

We propose and study an adaptive perfectly matched layer (PML) technique for solving Helmholtz-type scattering problems with perfectly conducting boundary:

$$\Delta u + k^2 u = 0 \quad \text{in } \mathbb{R}^2 \backslash \bar{D}, \tag{1.1a}$$

$$\frac{\partial u}{\partial \mathbf{n}} = -g \quad \text{on } \Gamma_D, \tag{1.1b}$$

$$\sqrt{r}\left(\frac{\partial u}{\partial r} - \mathbf{i}ku\right) \to 0 \quad \text{as } r = |x| \to \infty. \tag{1.1c}$$

Here $D \subset \mathbb{R}^2$ is a bounded domain with Lipschitz boundary Γ_D, $g \in H^{-1/2}(\Gamma_D)$ is determined by the incoming wave, and $\mathbf{n}$ is the unit outer normal to Γ_D. We assume the wave number $k \in \mathbb{R}$ is a constant. We remark that the results in this paper can be easily extended to solve the scattering problems with other boundary conditions such as Dirichlet or the impedance boundary condition on Γ_D, or to solve the acoustic wave propagation through inhomogeneous media with a variable wave number $k^2(x)$ inside some bounded domain.

Since the work of Berenger [3] which proposed a PML technique for solving with the time-dependent Maxwell equations, various constructions of PML absorbing layers have been proposed and studied in the literature (cf., e.g., Turkel and Yefet [20], Teixeira and Chew [19] for the reviews). Under the assumption that the exterior solution is composed of outgoing waves only, the basic idea of the PML technique is to surround the computational domain by a layer of finite thickness with specially designed

*Received by the editors June 22, 2004; accepted for publication (in revised form) February 17, 2005; published electronically June 30, 2005.
http://www.siam.org/journals/sinum/43-2/61033.html

†LSEC, Institute of Computational Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (zmchen@lsec.cc.ac.cn). The research of this author was supported in part by China NSF grant 10025102 and by China MOST grant G1999032802.

‡Institute of Computational Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, People's Republic of China (liuxz@lsec.cc.ac.cn).

model medium that would either slow down or attenuate all the waves that propagate from inside the computational domain. The PML equation for the time-harmonic scattering problem (1.1a) is derived in Collino and Monk [10] by a complex extension of the solution u in the exterior domain. It is proved in Lassas and Somersalo [13] and Hohage, Schmidt, and Zschiedrich [12] that the resultant PML solution converges exponentially to the solution of the original scattering problem as the thickness of the PML layer tends to infinity. We remark that in practical applications involving PML techniques, one cannot afford to use a very thick PML layer if uniform finite element meshes are used because it requires excessive grid points and hence more computer time and more storage. On the other hand, a thin PML layer requires a rapid variation of the artificial material property which deteriorates the accuracy if too coarse mesh is used in the PML layer.

A posteriori error estimates are computable quantities in terms of the discrete solution and data that measure the actual discrete errors without the knowledge of exact solutions. They are essential in designing algorithms for mesh modification which equidistribute the computational effort and optimize the computation. Ever since the pioneering work of Babuška and Rheinboldt [2], the adaptive finite element methods based on a posteriori error estimates have become a central theme in scientific and engineering computations. The ability of error control and the asymptotically optimal approximation property (see, e.g., Morin, Nochetto, and Siebert [17] and Chen and Dai [5]) make the adaptive finite element method attractive for complicated physical and industrial processes (cf., e.g., Chen and Dai [4] and Chen, Nochetto, and Schmidt [7]). For the efforts to solve scattering problems using adaptive methods based on a posterior error estimate, we refer to the recent work of Monk [15] and Monk and Süli [16].

It is proposed in Chen and Wu [8] for scattering problem by periodic structures (the grating problem) that one can use the a posteriori error estimate to determine the PML parameters. Moreover, the derived a posteriori error estimate in [8] has the nice feature of exponential decay in terms of the distance to the boundary of the fixed domain where the PML layer is placed. This property leads to coarse mesh size away from the fixed domain and thus makes the total computational costs insensitive to the thickness of the PML absorbing layer.

In this paper we extend the idea of using a posteriori error estimates to determine the PML parameters and propose an adaptive PML technique for solving the scattering problem (1.1a)–(1.1c). The main difficulty of the analysis is that in contrast to the grating problems in which there are only finite number of outgoing modes [8], now there are infinite number of outgoing modes expressed in terms of Hankel functions. We overcome this difficulty by exploiting the following uniform estimate for the Hankel functions H^1_ν, $\nu \in \mathbb{R}$:

$$|H_\nu^{(1)}(z)| \le e^{-\mathrm{Im}\,(z)\left(1-\frac{\Theta^2}{|z|^2}\right)^{1/2}} |H_\nu^{(1)}(\Theta)| \tag{1.2}$$

for any $z \in \mathbb{C}_{++}, \Theta \in \mathbb{R}$ such that $0 < \Theta \le |z|$, where $\mathbb{C}_{++} = \{z \in \mathbb{C} : \mathrm{Im}\,(z) \ge 0, \mathrm{Re}\,(z) \ge 0\}$. To our knowledge this sharp estimate is new and allows us to prove the exponentially decaying property of the PML solution without resorting to the integral equation technique in [13] or the representation formula in [12]. We remark that in [13], [12], it is required that the fictitious absorbing coefficient must be linear after certain distance away from the boundary where the PML layer is placed. The estimate (1.2) is proved in Lemma 2.2 which depends on the Macdonald formula for the modified Bessel functions. We also remark that since (1.2) is valid for all real order ν, the results of this paper can be extended directly to study three-dimensional Helmholtz-type scattering problems. We will report progress in this direction as well as the study of the electromagnetic scattering problems elsewhere in future.

Let $\Omega^{\rm PML} = B_\rho \backslash \bar{B}_R$, where $0 < R < \rho$ and B_a denotes the circle of radius $a > 0$. Let $\alpha(r) = 1 + \mathbf{i}\sigma(r)$ be the fictitious medium property. In practical applications, σ is usually taken as power functions:

$$\sigma = \sigma(r) = \sigma_0 \left(\frac{r-R}{\rho - R} \right)^m \quad \text{for some constant } \sigma_0 > 0 \text{ and integer } m \geq 1.$$

Under the assumption that the Dirichlet problem of the PML equation in the PML layer is uniquely solvable, we prove the following key estimate between the Dirichlet-to-Neumann mapping for the original scattering problem $T : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$ and the PML problem $\hat{T}$ (cf. Lemma 2.5), where $\Gamma_R = \partial B_R$,

$$\| T - \hat{T} \|_{L(H^{1/2}(\Gamma_R), H^{-1/2}(\Gamma_R))} \leq C(1+kR)^2 |\alpha_0|^2 e^{-k{\rm Im}\,(\tilde{\rho})\left(1 - \frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}},$$

where $\alpha_0 = 1 + \mathbf{i}\sigma_0$ and $\tilde{\rho} = \int_0^\rho \alpha(t)dt$ is the complex radius corresponding to ρ. We remark that the assumption of the unique solvability of the PML Dirichlet problem in the PML layer is rather mild in practical applications because standard Fredholm alternative theory implies that the PML Dirichlet problem in the PML layer is uniquely solvable for all but a discrete number of real k. Moreover, in the appendix of this paper, we show that for any given ρ, R, the Dirichlet PML problem in the PML layer is uniquely solvable for sufficiently large $\sigma_0 > 0$.

The layout of the paper is as follows. In section 2 we recall the PML formulation for (1.1a)–(1.1c), derive the key estimates for Hankel functions, and study the properties of the PML equation in the PML layer. Existence, uniqueness, and convergence of the PML formulation are considered. In section 3 we introduce the finite element discretization. In section 4 we derive the sharp a posteriori error estimate which lays down the basis of the combined adaptive PML and finite element methods. In section 5 we discuss the implementation of the adaptive method and present several numerical examples to illustrate the competitive behavior of the method. Finally in the appendix we show the unique solvability of the Dirichlet PML problem in the PML layer for sufficiently large σ_0.

2. The PML formulation. Let D be contained in the interior of the circle $B_R = \{x \in \mathbb{R}^2 : |x| < R\}$. We start by introducing an equivalent variational formulation of (1.1a)–(1.1c) in the bounded domain $\Omega_R = B_R \backslash \bar{D}$. In the domain $\mathbb{R}^2 \backslash \bar{B}_R$, the solution u of (1.1a)–(1.1c) can be written under the polar coordinates as follows:

$$u(r,\theta) = \sum_{n \in \mathbb{Z}} \frac{H_n^{(1)}(kr)}{H_n^{(1)}(kR)} \hat{u}_n e^{\mathbf{i}n\theta}, \quad \hat{u}_n = \frac{1}{2\pi} \int_0^{2\pi} u(R,\theta) e^{-\mathbf{i}n\theta} d\theta, \tag{2.1}$$

where $H_n^{(1)}$ is the Hankel function of the first kind and order n. The series in (2.1) converges uniformly for $r > R$ (cf., e.g., Colten and Kress [11]). Let $T : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$, where $\Gamma_R = \partial B_R$, be the Dirichlet-to-Neumann operator defined as follows: for any $f \in H^{1/2}(\Gamma_R)$,

$$Tf = \sum_{n \in \mathbb{Z}} k \frac{H_n^{(1)\prime}(kR)}{H_n^{(1)}(kR)} \hat{f}_n e^{\mathbf{i}n\theta}, \quad \hat{f}_n = \frac{1}{2\pi} \int_0^{2\pi} f e^{-\mathbf{i}n\theta} d\theta. \tag{2.2}$$

It is known that T is well-defined and the solution u written as in (2.1) satisfies

$$\frac{\partial u}{\partial \mathbf{n}}\Big|_{\Gamma_R} = Tu.$$

Let $a : H^1(\Omega_R) \times H^1(\Omega_R) \to \mathbb{C}$ be the sesquilinear form:

$$a(\varphi, \psi) = \int_{\Omega_R} \left(\nabla\varphi \cdot \nabla\bar{\psi} - k^2\varphi\bar{\psi}\right) dx - \langle T\varphi, \psi\rangle_{\Gamma_R}, \tag{2.3}$$

where $\langle\cdot,\cdot\rangle_{\Gamma_R}$ stands for the inner product on $L^2(\Gamma_R)$ or the duality pairing between $H^{-1/2}(\Gamma_R)$ and $H^{1/2}(\Gamma_R)$. Similar notation applies for $\langle\cdot,\cdot\rangle_{\Gamma_D}$, $\langle\cdot,\cdot\rangle_{\Gamma_\rho}$. The scattering problem (1.1a)–(1.1c) is equivalent to the following weak formulation (cf., e.g., [11]): given $g \in H^{-1/2}(\Gamma_D)$, find $u \in H^1(\Omega_R)$ such that

$$a(u, \psi) = \langle g, \psi\rangle_{\Gamma_D} \quad \forall \psi \in H^1(\Omega_R). \tag{2.4}$$

The existence of a unique solution of the variational problem (2.4) is known (cf., e.g., Colton and Kress [11] and McLean [14]). Then the general theory in Babuška and Aziz [1, Chapter 5] implies that there exists a constant $\mu > 0$ such that the following inf-sup condition holds:

$$\sup_{0\neq\psi\in H^1(\Omega_R)} \frac{|a(\varphi,\psi)|}{\|\,\psi\,\|_{H^1(\Omega_R)}} \geq \ \mu\|\,\varphi\,\|_{H^1(\Omega_R)} \quad \forall \varphi \in H^1(\Omega_R). \tag{2.5}$$

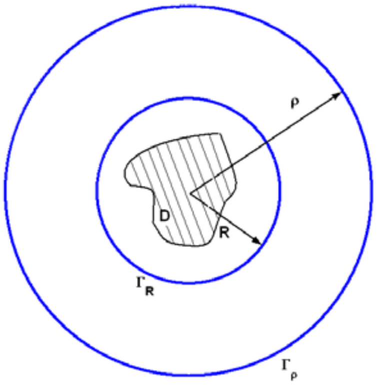


FIG. 2.1. *Setting of the scattering problem with the PML layer.*

Now we turn to the introduction of the absorbing PML layer. We surround the domain Ω_R with a PML layer $\Omega^{\mathrm{PML}} = \{x \in \mathbb{R}^2 : R < |x| < \rho\}$. The specially designed model medium in the PML layer should basically be so chosen that either the wave never reaches its external boundary or the amplitude of the reflected wave is so small that it does not essentially contaminate the solution in Ω_R. Throughout the paper we assume $\rho \leq CR$ for some generic fixed constant $C > 0$.

Let $\alpha(r) = 1 + \mathbf{i}\sigma(r)$ be the model medium property which satisfies

$$\sigma \in C(\mathbb{R}), \quad \sigma \geq 0, \quad \sigma = 0 \quad \text{for } r \leq R.$$

Denote by $\tilde{r}$ the complex radius defined by

$$\tilde{r} = \tilde{r}(r) = \begin{cases} r & \text{if } r \leq R, \\ \displaystyle\int_0^r \alpha(t)dt = r\beta(r) & \text{if } r \geq R. \end{cases} \tag{2.6}$$

Following [10], we introduce the PML equation

$$\nabla \cdot (A\nabla w) + \alpha\beta k^2 w = 0 \quad \text{in } \Omega^{\text{PML}}, \tag{2.7}$$

where $A = A(x)$ is a matrix which satisfies, in polar coordinates,

$$\nabla \cdot (A\nabla) = \frac{1}{r}\frac{\partial}{\partial r}\left(\frac{\beta r}{\alpha}\frac{\partial}{\partial r}\right) + \frac{\alpha}{\beta r^2}\frac{\partial^2}{\partial \theta^2}. \tag{2.8}$$

The PML solution $\hat{u}$ in $\Omega_\rho = B_\rho \backslash \bar{D}$ is defined as the solution of the following system:

$$\nabla \cdot (A\nabla \hat{u}) + \alpha\beta k^2 \hat{u} = 0 \quad \text{in } \Omega_\rho, \tag{2.9a}$$

$$\frac{\partial \hat{u}}{\partial \mathbf{n}} = -g \quad \text{on } \Gamma_D, \qquad \hat{u} = 0 \quad \text{on } \Gamma_\rho. \tag{2.9b}$$

This problem can be reformulated in the bounded domain Ω_R by imposing the boundary condition

$$\frac{\partial \hat{u}}{\partial \mathbf{n}}\Big|_{\Gamma_R} = \hat{T}\hat{u},$$

where the operator $\hat{T} : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$ is defined as follows: given $f \in H^{1/2}(\Gamma_R)$,

$$\hat{T}f = \frac{\partial \zeta}{\partial \mathbf{n}}\Big|_{\Gamma_R},$$

where $\zeta \in H^1(\Omega^{\text{PML}})$ satisfies

$$\nabla \cdot (A\nabla \zeta) + \alpha\beta k^2 \zeta = 0 \quad \text{in } \Omega^{\text{PML}}, \tag{2.10a}$$

$$\zeta = f \quad \text{on } \Gamma_R, \qquad \zeta = 0 \quad \text{on } \Gamma_\rho. \tag{2.10b}$$

The existence and uniqueness of the solutions of the PML problems (2.10a)–(2.10b) will be studied in the subsection 2.2 below.

On the basis of operator $\hat{T}$, we introduce the sesquilinear form $\hat{a} : H^1(\Omega_R) \times H^1(\Omega_R) \to \mathbb{C}$ by

$$\hat{a}(\varphi, \psi) = \int_{\Omega_R} \left(A\nabla\varphi \cdot \nabla\bar{\psi} - k^2\alpha\beta\varphi\bar{\psi}\right) dx - \langle \hat{T}\varphi, \psi \rangle_{\Gamma_R}. \tag{2.11}$$

Then the weak formulation for (2.9a)–(2.9b) is, given $g \in H^{-1/2}(\Gamma_D)$, find $\hat{u} \in H^1(\Omega_R)$ such that

$$\hat{a}(\hat{u}, \psi) = \langle g, \psi \rangle_{\Gamma_D} \quad \forall \psi \in H^1(\Omega_R). \tag{2.12}$$

The well-posedness of the PML problem (2.12) and the convergence of its solution to the solution of the original scattering problem (2.4) will be studied in the subsection 2.3. In the following we first derive some basic estimates for the Hankel function $H_n^{(1)}$ which play a key role in the analysis in this paper.

2.1. Hankel functions. For $\nu \in \mathbb{C}$, the two Hankel functions $H_\nu^{(1)}(z), H_\nu^{(2)}(z)$, where $z \in \mathbb{C}$, are two fundamental solutions of the Bessel equation for functions of order ν:

$$z^2\frac{d^2y}{dz^2} + z\frac{dy}{dz} + (z^2 - \nu^2)y = 0, \tag{2.13}$$

which satisfy the following asymptotic behaviors as $|z| \to \infty$:

$$H_\nu^{(1)}(z) \sim \left(\frac{2}{\pi z}\right)^{\frac{1}{2}} e^{\mathbf{i}\left(z-\frac{1}{2}\nu\pi-\frac{1}{4}\pi\right)}, \quad H_\nu^{(2)}(z) \sim \left(\frac{2}{\pi z}\right)^{\frac{1}{2}} e^{-\mathbf{i}\left(z-\frac{1}{2}\nu\pi-\frac{1}{4}\pi\right)}. \tag{2.14}$$

We also need the Bessel functions of purely imaginary argument $K_\nu(z)$, also called the modified Bessel functions, which is the solution of the differential equation

$$z^2\frac{d^2y}{dz^2} + z\frac{dy}{dz} - (z^2 + \nu^2)y = 0. \tag{2.15}$$

It is connected with $H_\nu^{(1)}(z)$ through the relation

$$K_\nu(z) = \frac{1}{2}\pi\mathbf{i}e^{\frac{1}{2}\nu\pi\mathbf{i}}H_\nu^{(1)}(\mathbf{i}z). \tag{2.16}$$

The importance of the function $K_\nu(z)$ in mathematical physics lies in the fact that it is a solution of (2.15) which tends to zero exponentially as $z \to \infty$ through positive values. We refer to the treatise by Watson [21] for extensive studies on the functions $H_\nu^{(1)}(z), H_\nu^{(2)}(z)$, and $K_\nu(z)$.

The following lemma is proved in [21, p. 439].

LEMMA 2.1 (Macdonald formula). *For any $\nu \in \mathbb{C}$ and $z_1, z_2 \in \mathbb{C}$ satisfying*

$$|\arg z_1| < \pi, \quad |\arg z_2| < \pi, \quad |\arg(z_1 + z_2)| < \frac{1}{4}\pi,$$

we have

$$K_\nu(z_1)K_\nu(z_2) = \frac{1}{2}\int_0^\infty e^{-\frac{v}{2}-\frac{z_1^2+z_2^2}{2v}} K_\nu\left(\frac{z_1z_2}{v}\right)\frac{dv}{v}.$$

An important consequence of this lemma is that for real ν, $K_\nu(z)$ has no zeros if $|\arg z| \le \frac{1}{2}\pi$ [21, p. 511], which, by (2.16), implies that $H_\nu^{(1)}(z)$ has no zeros when $\mathrm{Im}\,(z) \le 0$. In particular, we have $H_n^{(1)}(kR) \neq 0$ for any $n \in \mathbb{Z}, R > 0$. This justifies the writing of $H_n^{(1)}(kR)$ in the denominator in (2.1), (2.2).

LEMMA 2.2. *For any $\nu \in \mathbb{R}, z \in \mathbb{C}_{++} = \{z \in \mathbb{C} : \mathrm{Im}\,(z) \ge 0, \mathrm{Re}\,(z) \ge 0\}$, and $\Theta \in \mathbb{R}$ such that $0 < \Theta \le |z|$, we have*

$$|H_\nu^{(1)}(z)| \le e^{-\mathrm{Im}\,(z)\left(1-\frac{\Theta^2}{|z|^2}\right)^{1/2}}|H_\nu^{(1)}(\Theta)|. \tag{2.17}$$

This estimate, which to our knowledge is new, will play an important role in the analysis of this paper. The importance of the estimate (2.17) lies in the fact that it is uniform with respect to ν. We remark that the large argument asymptotic expansions such as (2.14) in the literature usually depend on ν and thus are insufficient for our purpose.

Proof. By (2.16) we know that

$$|H_\nu^{(1)}(z)|^2 = H_\nu^{(1)}(z)\overline{H_\nu^{(1)}(z)} = \frac{4}{\pi^2} K_\nu(-\mathbf{i}z)\overline{K_\nu(-\mathbf{i}z)} = \frac{4}{\pi^2} K_\nu(-\mathbf{i}z)K_\nu(\mathbf{i}\bar{z}),$$

where we have used the formula $K_\nu(\bar{z}) = \overline{K_\nu(z)}$ for real ν. Since $z \in \mathbb{C}_{++}$, we know that $|\arg(-\mathbf{i}z)| < \pi$, $|\arg(\mathbf{i}\bar{z})| < \pi$, and $|\arg(-\mathbf{i}z + \mathbf{i}\bar{z})| = 0 < \frac{\pi}{4}$. Thus by Lemma 2.1 we obtain

$$|H_\nu^{(1)}(z)|^2 = \frac{2}{\pi^2} \int_0^\infty e^{-\frac{v}{2} - \frac{-z^2 - \bar{z}^2}{2v}} K_\nu\left(\frac{|z|^2}{v}\right) \frac{dv}{v}.$$

After the change of variable $w = |z|^2/v$, we get

$$|H_\nu^{(1)}(z)|^2 = \frac{2}{\pi^2} \int_0^\infty e^{-\frac{|z|^2}{2w} + \frac{z^2 + \bar{z}^2}{2|z|^2} w} K_\nu(w) \frac{dw}{w},$$

which, for any $\Theta > 0$, we rewrite as

$$|H_\nu^{(1)}(z)|^2 = \frac{2}{\pi^2} \int_0^\infty e^{-\frac{|z|^2 - \Theta^2}{2w} - \frac{2|z|^2 - z^2 - \bar{z}^2}{2|z|^2} w} \cdot e^{-\frac{\Theta^2}{2w} + w} K_\nu(w) \frac{dw}{w}.$$

Now for $0 < \Theta \le |z|$, by Cauchy–Schwarz inequality, we deduce that

$$e^{-\frac{|z|^2 - \Theta^2}{2w} - \frac{2|z|^2 - z^2 - \bar{z}^2}{2|z|^2} w} = e^{-\frac{|z|^2 - \Theta^2}{2w} - \frac{2\mathrm{Im}\,(z)^2}{|z|^2} w} \le e^{-2\mathrm{Im}\,(z)\left(1 - \frac{\Theta^2}{|z|^2}\right)^{1/2}}.$$

Therefore,

$$\begin{aligned} |H_\nu^{(1)}(z)|^2 &\le e^{-2\mathrm{Im}\,(z)\left(1 - \frac{\Theta^2}{|z|^2}\right)^{1/2}} \frac{2}{\pi^2} \int_0^\infty e^{-\frac{\Theta^2}{2w} + w} K_\nu(w) \frac{dw}{w} \\ &= e^{-2\mathrm{Im}\,(z)\left(1 - \frac{\Theta^2}{|z|^2}\right)^{1/2}} |H_\nu^{(1)}(\Theta)|^2. \end{aligned}$$

This completes the proof. □

To proceed further, we recall the following Nicholson integral [21, p. 441]:

$$J_\nu^2(z) + Y_\nu^2(z) = \frac{8}{\pi^2} \int_0^\infty K_0(2z \sinh t) \cosh(2\nu t) dt \quad \text{for } z \in \mathbb{C}, \mathrm{Re}\,(z) > 0.$$

Here $K_0(z)$ is the modified Bessel function of order zero in (2.16). Since $\cosh(t) = (e^t + e^{-t})/2$ is an increasing function in $\mathbb{R}^+$, we have, for $\Theta > 0$, $n \ge 1$, that

$$\begin{aligned} J_{n-1}^2(\Theta) + Y_{n-1}^2(\Theta) &= \frac{8}{\pi^2} \int_0^\infty K_0(2\Theta \sinh t) \cosh(2(n-1)t) dt \\ &\le \frac{8}{\pi^2} \int_0^\infty K_0(2\Theta \sinh t) \cosh(2nt) dt \\ &= J_n^2(\Theta) + Y_n^2(\Theta). \end{aligned}$$

Thus,

$$|H_{n-1}^{(1)}(\Theta)| \le |H_n^{(1)}(\Theta)| \quad \text{for any } \Theta > 0, n \ge 1. \tag{2.18}$$

LEMMA 2.3. *For any $z \in \mathbb{C}_{++}$ and $\Theta \in \mathbb{R}$ such that $0 < \Theta \le |z|$, we have*

$$|H_n^{(1)\prime}(z)| \le e^{-\mathrm{Im}\,(z)\left(1-\frac{\Theta^2}{|z|^2}\right)^{1/2}}\left(1+\frac{|n|}{|z|}\right)|H_n^{(1)}(\Theta)| \quad \text{for } n \in \mathbb{Z}, |n| \ge 1, \tag{2.19}$$

$$|H_0^{(1)\prime}(z)| \le e^{-\mathrm{Im}\,(z)\left(1-\frac{\Theta^2}{|z|^2}\right)^{1/2}}|H_0^{(1)\prime}(\Theta)|. \tag{2.20}$$

Proof. Since $H_{-n}^{(1)} = e^{\mathbf{i}n\pi}H_n^{(1)}(z)$, we only need to prove (2.19) for $n \in \mathbb{Z}$, $n \ge 1$. By the formula

$$z\frac{dH_n^{(1)}(z)}{dz} + nH_n^{(1)}(z) = zH_{n-1}^{(1)}(z),$$

Lemma 2.2, and (2.18), we know that

$$\begin{aligned}|H_n^{(1)\prime}(z)| &\le |H_{n-1}^{(1)}(z)| + \frac{n}{|z|}|H_n^{(1)}(z)|\\ &\le e^{-\mathrm{Im}\,(z)\left(1-\frac{\Theta^2}{|z|^2}\right)^{1/2}}\left(|H_{n-1}^{(1)}(\Theta)| + \frac{n}{|z|}|H_n^{(1)}(\Theta)|\right)\\ &\le e^{-\mathrm{Im}\,(z)\left(1-\frac{\Theta^2}{|z|^2}\right)^{1/2}}\left(1+\frac{n}{|z|}\right)|H_n^{(1)}(\Theta)|.\end{aligned}$$

This proves (2.19). The estimate (2.20) can be proved similarly by using the formula $dH_0^{(1)}(z)/dz = -H_1^{(1)}(z)$. This completes the proof. □

2.2. The PML equation in the layer. In this subsection we consider the Dirichlet problem of the PML equation in the layer Ω^{PML}:

$$\nabla \cdot (A\nabla w) + \alpha\beta k^2 w = 0 \quad \text{in } \Omega^{\mathrm{PML}}, \tag{2.21a}$$

$$w = 0 \quad \text{on } \Gamma_R, \qquad w = q \quad \text{on } \Gamma_\rho, \tag{2.21b}$$

where $q \in H^{1/2}(\Gamma_\rho)$. Let $\hat{b} : H^1(\Omega^{\mathrm{PML}}) \times H^1(\Omega^{\mathrm{PML}}) \to \mathbb{C}$ be the sesquilinear form:

$$\hat{b}(\varphi,\psi) = \int_R^\rho\int_0^{2\pi}\left(\frac{\beta r}{\alpha}\frac{\partial\varphi}{\partial r}\frac{\partial\bar{\psi}}{\partial r} + \frac{\alpha}{\beta r}\frac{\partial\varphi}{\partial\theta}\frac{\partial\bar{\psi}}{\partial\theta} - \alpha\beta k^2 r\varphi\bar{\psi}\right)dr\,d\theta. \tag{2.22}$$

Then from (2.8) we know that the weak formulation for (2.21a)–(2.21b) is the following: Given $q \in H^{1/2}(\Gamma_\rho)$, find $w \in H^1(\Omega^{\mathrm{PML}})$ such that $w = 0$ on Γ_R, $w = q$ on Γ_ρ, and

$$\hat{b}(w,\varphi) = 0 \quad \forall\varphi \in H_0^1(\Omega^{\mathrm{PML}}). \tag{2.23}$$

We make the following assumption on the fictitious medium property σ, which is rather mild in the practical application of the PML techniques:

(H1) $\sigma = \sigma_0\left(\frac{r-R}{\rho-R}\right)^m$ for some constant $\sigma_0 > 0$ and some integer $m \ge 1$.

From (H1) we know that $\beta(r) = 1 + \mathbf{i}\hat{\sigma}(r)$, where

$$\hat{\sigma}(r) = \frac{1}{r}\int_R^r \sigma(t)dt = \frac{\sigma_0}{m+1}\frac{r-R}{r}\left(\frac{r-R}{\rho-R}\right)^m.$$

Thus $\hat{\sigma} \leq \sigma$ for all $r \geq R$. Notice that for $\alpha = 1 + \mathbf{i}\sigma$, $\beta = 1 + \mathbf{i}\hat{\sigma}$, we have

$$\operatorname{Re}\left(\frac{\beta}{\alpha}\right) = \frac{1+\sigma\hat{\sigma}}{1+\sigma^2}, \qquad \operatorname{Re}\left(\frac{\alpha}{\beta}\right) = \frac{1+\sigma\hat{\sigma}}{1+\hat{\sigma}^2}, \qquad \operatorname{Re}(\alpha\beta) = 1 - \sigma\hat{\sigma}$$

and, consequently,

$$\operatorname{Re}[\hat{b}(v,v)] = \int_R^\rho \int_0^{2\pi} \left[\frac{1+\sigma\hat{\sigma}}{1+\sigma^2} r \left|\frac{\partial v}{\partial r}\right|^2 + \frac{1+\sigma\hat{\sigma}}{1+\hat{\sigma}^2} \frac{1}{r} \left|\frac{\partial v}{\partial \theta}\right|^2 + (\sigma\hat{\sigma} - 1)k^2 r |v|^2 \right] dr\, d\theta.$$

Since

$$\frac{1+\sigma\hat{\sigma}}{1+\sigma^2} \geq \frac{1}{1+\sigma^2} \geq |\alpha_0|^{-2}, \qquad \frac{1+\sigma\hat{\sigma}}{1+\hat{\sigma}^2} \geq 1 \geq |\alpha_0|^{-2}, \tag{2.24}$$

where $\alpha_0 = 1 + \mathbf{i}\sigma_0$, by using the analytic Fredholm alternative theorem we know that the PML problem in the layer (2.23) exists a unique solution for every real k except possibly for a discrete set of values of k (cf., e.g., the argument in [10, Theorem 2]). In this paper we will not elaborate on this issue and simply make the following assumption:

(H2) There exists a unique solution to the Dirichlet PML problem (2.23) in the layer.

For any $\varphi \in H^1(\Omega^{\mathrm{PML}})$, define

$$\|\varphi\|_{*,\Omega^{\mathrm{PML}}} = \left[\int_R^\rho \int_0^{2\pi} \left(\frac{1+\sigma\hat{\sigma}}{1+\sigma^2} r \left|\frac{\partial \varphi}{\partial r}\right|^2 + \frac{1+\sigma\hat{\sigma}}{1+\hat{\sigma}^2} \frac{1}{r} \left|\frac{\partial \varphi}{\partial \theta}\right|^2 + (1+\sigma\hat{\sigma})k^2 r |\varphi|^2 \right) \right]^{1/2}.$$

It is easy to see that $\|\cdot\|_{*,\Omega^{\mathrm{PML}}}$ is an equivalent norm on $H^1(\Omega^{\mathrm{PML}})$. By using the general theory in [1, Chapter 5], (H2) implies that there exists a constant $\hat{C} > 0$ such that

$$\sup_{0 \neq \psi \in H_0^1(\Omega^{\mathrm{PML}})} \frac{|\hat{b}(\varphi,\psi)|}{\|\psi\|_{*,\Omega^{\mathrm{PML}}}} \geq \hat{C} \|\varphi\|_{*,\Omega^{\mathrm{PML}}} \quad \forall \varphi \in H_0^1(\Omega^{\mathrm{PML}}). \tag{2.25}$$

The constant $\hat{C}$ depends in general on the domain Ω^{PML} and the wave number k. In the appendix of the paper, however, we will show that for sufficiently large σ_0, (H2) can be proved and $\hat{C}$ can be chosen as independent of Ω^{PML} and k. Without loss of generality we assume $\hat{C} \leq 1$.

To proceed, we introduce the following notation. For any function ξ defined on a circle $\Gamma_a = \{x \in \mathbb{R}^2 : |x| = a\}$ having the Fourier expansion:

$$\xi = \sum_{n \in \mathbb{Z}} \hat{\xi}_n e^{\mathbf{i}n\theta}, \qquad \hat{\xi}_n = \frac{1}{2\pi} \int_0^{2\pi} \xi e^{-\mathbf{i}n\theta} d\theta,$$

we define

$$\|\xi\|^2_{H^{1/2}(\Gamma_a)} = 2\pi \sum_{n \in \mathbb{Z}} (1+n^2)^{1/2} |\hat{\xi}_n|^2, \qquad \|\xi\|^2_{H^{-1/2}(\Gamma_a)} = 2\pi \sum_{n \in \mathbb{Z}} (1+n^2)^{-1/2} |\hat{\xi}_n|^2.$$

The following theorem is the main objective of this subsection.

THEOREM 2.4. *Let* (H1)–(H2) *be satisfied. There exists a constant* $C > 0$ *independent of* k, R, ρ, *and* σ_0 *such that the following estimates are satisfied:*

$$\| \, |\alpha|^{-1} \nabla w \, \|_{L^2(\Omega^{\rm PML})} \le C\hat{C}^{-1}(1+kR)|\alpha_0| \| \, q \, \|_{H^{1/2}(\Gamma_\rho)}, \tag{2.26}$$

$$\left\| \frac{\partial w}{\partial \mathbf{n}} \right\|_{H^{-1/2}(\Gamma_R)} \le C\hat{C}^{-1}(1+kR)^2|\alpha_0|^2 \| \, q \, \|_{H^{1/2}(\Gamma_\rho)}, \tag{2.27}$$

where $\alpha_0 = 1 + \mathbf{i}\sigma_0$.

Proof. We first show that there exists a constant C independent of k, ρ, R, and σ_0 such that

$$|\hat{b}(\varphi,\psi)| \le C(1+kR)|\alpha_0| \| \, \psi \, \|_{*,\Omega^{\rm PML}} |||\varphi|||_{H^1(\Omega^{\rm PML})}, \tag{2.28}$$

where $|||\varphi|||_{H^1(\Omega^{\rm PML})} = (\| \, \nabla\varphi \, \|^2_{L^2(\Omega^{\rm PML})} + R^{-2} \| \, \varphi \, \|^2_{L^2(\Omega^{\rm PML})})^{1/2}$ is the weighted H^1-norm. In fact, since $\hat{\sigma} \le \sigma \le \sigma_0$, we have

$$\begin{aligned}
&\left| \int_R^\rho \int_0^{2\pi} \left(\frac{\beta}{\alpha} r \frac{\partial\varphi}{\partial r}\frac{\partial\bar{\psi}}{\partial r} + \frac{\alpha}{\beta r}\frac{\partial\varphi}{\partial\theta}\frac{\partial\bar{\psi}}{\partial\theta} - \alpha\beta k^2 r\varphi\bar{\psi} \right) dr\, d\theta \right| \\
&\le \left(\int_R^\rho \int_0^{2\pi} \frac{1+\sigma\hat{\sigma}}{1+\sigma^2} r \left| \frac{\partial\psi}{\partial r} \right|^2 \right)^{1/2} \left(\int_R^\rho \int_0^{2\pi} \frac{1+\hat{\sigma}^2}{1+\sigma\hat{\sigma}} r \left| \frac{\partial\varphi}{\partial r} \right|^2 \right)^{1/2} \\
&\quad + \left(\int_R^\rho \int_0^{2\pi} \frac{1+\sigma\hat{\sigma}}{1+\hat{\sigma}^2} \frac{1}{r} \left| \frac{\partial\psi}{\partial\theta} \right|^2 \right)^{1/2} \left(\int_R^\rho \int_0^{2\pi} \frac{1+\sigma^2}{1+\sigma\hat{\sigma}} \frac{1}{r} \left| \frac{\partial\varphi}{\partial\theta} \right|^2 \right)^{1/2} \\
&\quad + \left(\int_R^\rho \int_0^{2\pi} k^2(1+\sigma\hat{\sigma}) r|\psi|^2 \right)^{1/2} \left(\int_R^\rho \int_0^{2\pi} k^2 r \frac{|\alpha\beta|^2}{1+\sigma\hat{\sigma}} |\varphi|^2 \right)^{1/2} \\
&\le C(1+kR)|\alpha_0| \| \, \psi \, \|_{*,\Omega^{\rm PML}} |||\varphi|||_{H^1(\Omega^{\rm PML})}.
\end{aligned}$$

This implies the estimate (2.28).

Now we turn to the proof the estimate (2.26). Let $\psi \in H^1(\Omega^{\rm PML})$ such that $\psi = 0$ on Γ_R and $\psi = q$ on Γ_ρ. By taking $\varphi = w - \psi \in H^1_0(\Omega^{\rm PML})$ in (2.23), we know from (2.28) that

$$|\hat{b}(\varphi,\varphi)| = |\hat{b}(w-\psi,\varphi)| = |\hat{b}(\psi,\varphi)| \le C(1+kR)|\alpha_0| \| \, \varphi \, \|_{*,\Omega^{\rm PML}} |||\psi|||_{H^1(\Omega^{\rm PML})},$$

which implies by (2.25) that

$$\| \, \varphi \, \|_{*,\Omega^{\rm PML}} \le C\hat{C}^{-1}(1+kR)|\alpha_0| |||\psi|||_{H^1(\Omega^{\rm PML})}.$$

Notice that

$$\| \, \psi \, \|_{*,\Omega^{\rm PML}} \le C(1+kR)|\alpha_0| |||\psi|||_{H^1(\Omega^{\rm PML})},$$

we get

$$\| \, w \, \|_{*,\Omega^{\rm PML}} = \| \, \varphi + \psi \, \|_{*,\Omega^{\rm PML}} \le C\hat{C}^{-1}(1+kR)|\alpha_0| |||\psi|||_{H^1(\Omega^{\rm PML})}.$$

Since the above estimate is valid for any $\psi \in H^1(\Omega^{\rm PML})$ such that $\psi = 0$ on Γ_R, $\psi = q$ on Γ_ρ, we deduce by standard scaling argument using the assumption $\rho \le CR$ that

$$\| \, w \, \|_{*,\Omega^{\rm PML}} \le C\hat{C}^{-1}(1+kR)|\alpha_0| \| \, q \, \|_{H^{1/2}(\Gamma_\rho)}. \tag{2.29}$$

This shows the estimate (2.26) upon using (2.24).

To show (2.27) we multiply the (2.21a) by any function $\varphi \in H^1(\Omega^{\rm PML})$ such that $\varphi = 0$ on Γ_ρ and integrate over $\Omega^{\rm PML}$ to obtain

$$-\int_{\Omega^{\rm PML}} A\nabla w \cdot \nabla\varphi dx - \int_{\Gamma_R} \frac{\partial w}{\partial r}\varphi ds + \int_{\Omega^{\rm PML}} \alpha\beta k^2 w\varphi dx = 0.$$

Thus

$$\left|\int_{\Gamma_R} \frac{\partial w}{\partial r}\varphi ds\right| = |\hat{b}(w,\bar{\varphi})| \le C(1+kR)|\alpha_0| \|\, w \,\|_{*,\Omega^{\rm PML}} |||\varphi|||_{H^1(\Omega^{\rm PML})}$$

for any $\varphi \in H^1(\Omega^{\rm PML})$ such that $\varphi = 0$ on Γ_ρ. This implies by (2.29) that

$$\left|\int_{\Gamma_R} \frac{\partial w}{\partial r}\varphi ds\right| \le C\hat{C}^{-1}(1+kR)^2|\alpha_0|^2 \|\, q \,\|_{H^{1/2}(\Gamma_\rho)} \|\varphi\|_{H^{1/2}(\Gamma_R)} \qquad \forall \varphi \in H^{1/2}(\Gamma_R).$$

This completes the proof of the theorem. □

2.3. Convergence of the PML problem. In this subsection we consider the convergence of the PML problem (2.12) to the original scattering problem (2.4). Following an idea in [13], for any function $f \in H^{1/2}(\Gamma_R)$, we introduce the propagation operator $P : H^{1/2}(\Gamma_R) \to H^{1/2}(\Gamma_\rho)$:

$$P(f) = \sum_{n\in\mathbb{Z}} \frac{H_n^{(1)}(k\tilde{\rho})}{H_n^{(1)}(kR)} \hat{f}_n e^{\mathbf{i}n\theta}, \qquad \hat{f}_n = \frac{1}{2\pi}\int_0^{2\pi} f e^{-\mathbf{i}n\theta} d\theta. \tag{2.30}$$

By Lemma 2.2, it is easy to see that $P : H^{1/2}(\Gamma_R) \to H^{1/2}(\Gamma_\rho)$ is well-defined, and

$$\|\, P(f) \,\|_{H^{1/2}(\Gamma_\rho)} \le e^{-k{\rm Im}\,(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \|\, f \,\|_{H^{1/2}(\Gamma_R)} \qquad \forall r \ge R. \tag{2.31}$$

Moreover, by Theorem 2.4, under the assumptions (H1)–(H2), the operator $\hat{T} : H^{1/2}(\Gamma_R) \to H^{-1/2}(\Gamma_R)$, which is defined through the Dirichlet problem of the PML equation in the layer, is also well-defined. Furthermore, we have the following estimate.

Lemma 2.5. *Let* (H1)–(H2) *be satisfied. We have*

$$\|\, Tf - \hat{T}f \,\|_{H^{-1/2}(\Gamma_R)} \le C\hat{C}^{-1}(1+kR)^2|\alpha_0|^2 e^{-k{\rm Im}\,(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \|\, f \,\|_{H^{1/2}(\Gamma_R)}.$$

Proof. For any $f \in H^{1/2}(\Gamma_R)$, we know that

$$Tf - \hat{T}f = \frac{\partial w}{\partial \mathbf{n}}\Big|_{\Gamma_R},$$

where $w \in H^1(\Omega^{\rm PML})$ satisfies

$$\nabla\cdot(A\nabla w) + \alpha\beta k^2 w = 0 \quad \text{in } \Omega^{\rm PML},$$
$$w = 0 \ \text{ on } \Gamma_R, \qquad w = P(f) \quad \text{on } \Gamma_\rho.$$

By (2.27) and (2.31) we then have

$$\begin{aligned}\left\|\frac{\partial w}{\partial \mathbf{n}}\right\|_{H^{-1/2}(\Gamma_R)} &\le C\hat{C}^{-1}(1+kR)^2|\alpha_0|^2 \|\, P(f) \,\|_{H^{1/2}(\Gamma_\rho)} \\ &\le C\hat{C}^{-1}(1+kR)^2|\alpha_0|^2 e^{-k{\rm Im}\,(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \|\, f \,\|_{H^{1/2}(\Gamma_R)}.\end{aligned}$$

This completes the proof. □

The following theorem is the main result of this section.

THEOREM 2.6. *Let* (H1)–(H2) *be satisfied. Then for sufficiently large $\sigma_0 > 0$, the PML problem* (2.12) *has a unique solution $\hat{u} \in H^1(\Omega_\rho)$. Moreover, we have the following estimate:*

$$\| u - \hat{u} \|_{H^1(\Omega_R)} \le C\hat{C}^{-1}(1+kR)^2|\alpha_0|^2 e^{-k\mathrm{Im}\,(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \| \hat{u} \|_{H^{1/2}(\Gamma_R)}. \tag{2.32}$$

Proof. The existence of a unique solution for (2.12) follows from Lemma 2.5 by using the same argument as in [8, Theorem 2.4]. Next, by (2.4) and (2.12), we have

$$a(u-\hat{u},\varphi) = \hat{a}(\hat{u},\varphi) - a(\hat{u},\varphi) = \langle T\hat{u} - \hat{T}\hat{u}, \varphi\rangle_{\Gamma_R} \quad \forall \varphi \in H^1(\Omega^{\mathrm{PML}}).$$

This implies the desired estimate (2.32) upon using Lemma 2.5 and (2.5). □

3. Finite element approximations. In this section we introduce the finite element approximations of the PML problems (2.9a)–(2.9b). From now on we assume $g \in L^2(\Gamma_D)$. Let $b : H^1(\Omega_\rho) \times H^1(\Omega_\rho) \to \mathbb{C}$ be the sesquilinear form given by

$$b(\varphi,\psi) = \int_{\Omega_\rho} \left(A\nabla\varphi \cdot \nabla\bar{\psi} - \alpha\beta k^2 \varphi\bar{\psi}\right) dx. \tag{3.1}$$

Denote by $H^1_{(0)}(\Omega_\rho) = \{v \in H^1(\Omega_\rho) : v = 0 \text{ on } \Gamma_\rho\}$. Then the weak formulation of (2.9a)–(2.9b) is, given $g \in L^2(\Gamma_D)$, find $\hat{u} \in H^1_{(0)}(\Omega_\rho)$ such that

$$b(\hat{u},\psi) = \int_{\Gamma_D} g\bar{\psi} ds \qquad \forall \psi \in H^1_{(0)}(\Omega_\rho). \tag{3.2}$$

Let Γ^h_ρ, which consists of piecewise segments whose vertices lie on Γ_ρ, be an approximation of Γ_ρ. Let Ω^h_ρ be the subdomain of Ω_ρ bounded by Γ_D and Γ^h_ρ. Let $\mathcal{M}_h$ be a regular triangulation of the domain Ω^h_ρ. We assume the elements $K \in \mathcal{M}_h$ may have one curved edge align with Γ_D so that $\Omega^h_\rho = \cup_{K\in\mathcal{M}_h} K$.

Let $V_h \subset H^1(\Omega^h_\rho)$ be the conforming linear finite element space over Ω^h_ρ, and $\overset{\circ}{V}_h = \{v_h \in V_h : v_h = 0 \text{ on } \Gamma^h_\rho\}$. In the following we will always assume that the functions in $\overset{\circ}{V}_h$ are extended to the domain Ω_ρ by zero so that any function $v_h \in \overset{\circ}{V}_h$ is also a function in $H^1_{(0)}(\Omega_\rho)$. The finite element approximation to the PML problems (2.9a)–(2.9b) reads as follows: Find $u_h \in \overset{\circ}{V}_h$ such that

$$b(u_h,\psi_h) = \int_{\Gamma_D} g\bar{\psi}_h ds \qquad \forall\ \psi_h \in \overset{\circ}{V}_h. \tag{3.3}$$

Following the general theory in [1, Chapter 5], the existence of unique solution of the discrete problem (3.3) and the finite element convergence analysis depend on the following discrete inf-sup condition:

$$\sup_{0\neq\psi_h\in\overset{\circ}{V}_h} \frac{|b(\varphi_h,\psi_h)|}{\| \psi_h \|_{H^1(\Omega_\rho)}} \ge \hat{\mu}\ \| \varphi_h \|_{H^1(\Omega_\rho)} \quad \forall\ \varphi_h \in \overset{\circ}{V}_h, \tag{3.4}$$

where the constant $\hat{\mu} > 0$ is independent of the finite element mesh size. Since the continuous problem (3.2) has a unique solution by Theorem 2.6, the sesquilinear form

$b : H^1_{(0)}(\Omega_\rho) \times H^1_{(0)}(\Omega_\rho) \to \mathbb{C}$ satisfies the continuous inf-sup condition. Then a general argument of Schatz [18] implies (3.4) is valid for sufficiently small mesh size $h < h^*$. On the basis of (3.4), appropriate a priori error estimate can also be derived which depends on the regularity of the PML solution $\hat{u}$. In this paper, we are interested in a posterior error estimates and the associated adaptive algorithm. Thus in the following we simply assume the discrete problem (3.3) has a unique solution $u_h \in \overset{\circ}{V}_h$.

For any $K \in \mathcal{M}_h$, we denote by h_K its diameter. Let $\mathcal{B}_h$ denote the set of all sides that do not lie on Γ_D and Γ^h_ρ. For any $e \in \mathcal{B}_h$, h_e stands for its length. For any $K \in \mathcal{M}_h$, we introduce the residual:

$$R_h := \nabla \cdot (A \nabla u_h|_K) + \alpha\beta k^2 u_h|_K. \tag{3.5}$$

For any interior side $e \in \mathcal{B}_h$ which is the common side of K_1 and $K_2 \in \mathcal{M}_h$, we define the jump residual across e:

$$J_e := (A\nabla u_h|_{K_1} - A\nabla u_h|_{K_2}) \cdot \nu_e, \tag{3.6}$$

using the convention that the unit normal vector ν_e to e points from K_2 to K_1. If $e = \Gamma_D \cap \partial K$ for some element $K \in \mathcal{M}_h$, then we define the jump residual:

$$J_e := 2(\nabla u_h|_K \cdot \mathbf{n} + g). \tag{3.7}$$

For any $K \in \mathcal{M}_h$, denote by η_K the local error estimator which is defined by

$$\eta_K = \max_{x \in \tilde{K}} \omega(x) \cdot \left(\|h_K R_h\|^2_{L^2(K)} + \frac{1}{2} \sum_{e \subset \partial K} h_e \| \, J_e \, \|^2_{L^2(e)} \right)^{1/2}, \tag{3.8}$$

where $\tilde{K}$ is the union of all elements having nonempty intersection with K, and

$$\omega(x) = \begin{cases} 1 & \text{if } x \in \overline{\Omega_R}, \\ |\alpha_0 \alpha| e^{-k \mathrm{Im}\,(\tilde{r}) \left(1 - \frac{r^2}{|\tilde{r}|^2}\right)^{1/2}} & \text{if } x \in \Omega^{\mathrm{PML}}. \end{cases}$$

The following theorem is the main result of this paper.

THEOREM 3.1. *There exists a constant C depending only on the minimum angle of the mesh $\mathcal{M}_h$ such that the following a posterior error estimate is valid:*

$$\begin{aligned} \| \, u - u_h \, \|_{H^1(\Omega_R)} \le\; & C\hat{C}^{-1} \Lambda(kR)^{1/2}(1 + kR) \left(\sum_{K \in \mathcal{M}_h} \eta_K^2 \right)^{1/2} \\ & + C\hat{C}^{-1}(1 + kR)^2 |\alpha_0|^2 e^{-k \mathrm{Im}\,(\tilde{\rho}) \left(1 - \frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \| \, u_h \, \|_{H^{1/2}(\Gamma_R)}. \end{aligned} \tag{3.9}$$

Here $\Lambda(kR)$ is defined in Lemma 4.3 *below.*

The proof of this theorem will be given in section 4. The important exponentially decaying factor $e^{-k\mathrm{Im}\,(\tilde{r})\left(1 - \frac{r^2}{|\tilde{r}|^2}\right)^{1/2}}$ in the PML region Ω^{PML} allows us to take thicker PML layers without introducing unnecessary fine meshes away from the fixed domain Ω_R. Recall that thicker PML layers allow smaller PML medium property, which enhances numerical stability.

4. A posteriori error estimates.

In this section we prove the a posteriori error estimates in Theorem 3.1.

4.1. Error representation formula. For any $\varphi \in H^1(\Omega_R)$, let $\tilde{\varphi}$ be its extension in $\Omega^{\rm PML}$ such that

$$\nabla\cdot(\bar{A}\nabla\tilde{\varphi}) + \overline{\alpha\beta}k^2\tilde{\varphi} = 0 \quad \text{in } \Omega^{\rm PML}, \tag{4.1a}$$

$$\tilde{\varphi} = \varphi \quad \text{on } \Gamma_R, \qquad \tilde{\varphi} = 0 \quad \text{on } \Gamma_\rho. \tag{4.1b}$$

Lemma 4.1. *Let* (H2) *be satisfied. For any $\varphi, \psi \in H^1(\Omega^{\rm PML})$, we have*

$$\langle \hat{T}\varphi, \psi\rangle_{\Gamma_R} = \langle \hat{T}\bar{\psi}, \bar{\varphi}\rangle_{\Gamma_R}.$$

Proof. By definition, $\hat{T}\varphi = \partial w/\partial\mathbf{n}$ on Γ_R, where w satisfies

$$\nabla\cdot(A\nabla w) + \alpha\beta k^2 w = 0 \quad \text{in } \Omega^{\rm PML},$$
$$w = \varphi \quad \text{on } \Gamma_R, \qquad w = 0 \quad \text{on } \Gamma_\rho.$$

Thus

$$w(x) = \sum_{n\in\mathbb{Z}} \left(a_n H_n^{(1)}(k\tilde{r}) + b_n H_n^{(2)}(k\tilde{r})\right) e^{\mathbf{i}n\theta}$$

with the coefficients a_n, b_n being determined by the boundary conditions in (4.1b)

$$a_n H_n^{(1)}(kR) + b_n H_n^{(2)}(kR) = \hat{\varphi}_n, \qquad a_n H_n^{(1)}(k\tilde{\rho}) + b_n H_n^{(2)}(k\tilde{\rho}) = 0,$$

where $\hat{\varphi}_n = \frac{1}{2\pi}\int_0^{2\pi}\varphi(R,\theta)e^{-\mathbf{i}n\theta}d\theta$ is the nth Fourier coefficient of $\varphi|_{\Gamma_R}$. Denote by

$$H_n(k\tilde{r}) = H_n^{(1)}(k\tilde{r})H_n^{(2)}(k\tilde{\rho}) - H_n^{(2)}(k\tilde{r})H_n^{(1)}(k\tilde{\rho}).$$

Then since by (H2) the Dirichlet PML problem in the layer has a unique solution, we get $H_n(kR) \neq 0$, and

$$a_n = \frac{H_n^{(2)}(k\tilde{\rho})}{H_n(kR)}\hat{\varphi}_n, \qquad b_n = -\,\frac{H_n^{(1)}(k\tilde{\rho})}{H_n(kR)}\hat{\varphi}_n.$$

Thus

$$w = w(r,\theta) = \sum_{n\in\mathbb{Z}} \frac{H_n(k\tilde{r})}{H_n(kR)}\hat{\varphi}_n e^{\mathbf{i}n\theta},$$

which, since $\tilde{r}'(R) = \alpha(R) = 1$ and $\tilde{R} = R$, implies

$$\hat{T}\varphi|_{\Gamma_R} = \sum_{n\in\mathbb{Z}} k\frac{H_n'(kR)}{H_n(kR)}\hat{\varphi}_n e^{\mathbf{i}n\theta}.$$

Therefore,

$$\langle \hat{T}\varphi, \psi\rangle_{\Gamma_R} = \sum_{n\in\mathbb{Z}} k\frac{H_n'(kR)}{H_n(kR)}\hat{\varphi}_n\bar{\hat{\psi}}_n \qquad \forall \varphi, \psi \in H^1(\Omega^{\rm PML}).$$

This completes the proof. □

Whenever no confusion of the notation incurred, we shall write in the following $\tilde{\varphi}$ as φ in $\Omega^{\rm PML}$.

LEMMA 4.2 (error representational formula). *For any $\varphi \in H^1(\Omega_R)$, which is extended to be a function in $H^1(\Omega_\rho)$ according to* (4.1a)–(4.1b), *and $\varphi_h \in \overset{\circ}{V}_h$, we have*

$$a(u-u_h,\varphi) = \int_{\Gamma_D} g(\overline{\varphi-\varphi_h}) - b(u_h,\varphi-\varphi_h) + \langle Tu_h - \hat{T}u_h, \varphi\rangle_{\Gamma_R}. \tag{4.2}$$

Proof. By (2.4) and the definitions (2.3) and (3.1),

$$\begin{aligned} &a(u-u_h,\varphi)\\ &\quad = \int_{\Gamma_D} g\bar{\varphi} - \int_{\Omega_R} (A\nabla u_h\cdot\nabla\bar{\varphi} - \alpha\beta k^2 u_h\bar{\varphi}) + \langle Tu_h,\varphi\rangle_{\Gamma_R}\\ &\quad = \int_{\Gamma_D} g\bar{\varphi} - b(u_h,\varphi) + \int_{\Omega^{\rm PML}} (A\nabla u_h\cdot\nabla\bar{\tilde{\varphi}} - \alpha\beta k^2 u_h\bar{\tilde{\varphi}}) + \langle Tu_h,\varphi\rangle_{\Gamma_R}. \end{aligned} \tag{4.3}$$

On the other hand, by multiplying (4.1a) by $\bar{u}_h$, integrating by parts, and recalling that $\mathbf{n}$ is the unit outer normal to Γ_R which points outside Ω_R, we deduce that

$$-\int_{\Omega^{\rm PML}} (\bar{A}\nabla\tilde{\varphi}\cdot\nabla\bar{u}_h - \overline{\alpha\beta}k^2\tilde{\varphi}\bar{u}_h) - \left\langle \frac{\partial\tilde{\varphi}}{\partial\mathbf{n}}, u_h\right\rangle_{\Gamma_R} = 0,$$

which is equivalent to

$$\int_{\Omega^{\rm PML}} (A\nabla u_h\cdot\nabla\bar{\tilde{\varphi}} - \alpha\beta k^2 u_h\bar{\tilde{\varphi}}) = -\left\langle \frac{\partial\bar{\tilde{\varphi}}}{\partial\mathbf{n}}, \bar{u}_h\right\rangle_{\Gamma_R}. \tag{4.4}$$

Since by the definition of $\hat{T}: H^{1/2}(\Gamma_R)\to H^{-1/2}(\Gamma_R)$,

$$\frac{\partial\bar{\tilde{\varphi}}}{\partial\mathbf{n}}\Big|_{\Gamma_R} = \hat{T}\bar{\varphi},$$

we obtain by substituting (4.4) into (4.3) that

$$a(u-u_h,\varphi) = \int_{\Gamma_D} g\bar{\varphi} - b(u_h,\varphi) + \langle Tu_h,\varphi\rangle - \langle \hat{T}\bar{\varphi},\bar{u}_h\rangle.$$

This completes the proof upon using Lemma 4.1 and (3.3). □

4.2. Estimates for the extension. For any $\varphi\in H^1(\Omega_R)$, we define, for $r\ge R$,

$$\phi = \phi(r,\theta) = \sum_{n\in\mathbb{Z}} \frac{H_n^{(1)}(k\tilde{r})}{H_n^{(1)}(kR)}\bar{\hat{\varphi}}_n e^{\mathbf{i}n\theta}, \qquad \hat{\varphi}_n = \frac{1}{2\pi}\int_0^{2\pi} \varphi(R,\theta)e^{-\mathbf{i}n\theta}d\theta. \tag{4.5}$$

The function ϕ satisfies

$$\nabla\cdot(A\nabla\phi) + \alpha\beta k^2\phi = 0 \quad \text{in } \mathbb{R}^2\backslash\bar{B}_R, \tag{4.6a}$$

$$\phi = \bar{\varphi} \quad \text{on } \Gamma_R, \tag{4.6b}$$

$$|\phi| \text{ is uniformly bounded as } r = |x| \to \infty. \tag{4.6c}$$

By Lemma 2.2, it is easy to see that

$$\| \phi \|_{H^{1/2}(\Gamma_\rho)} \le e^{-k\mathrm{Im}\,(\tilde\rho)\left(1-\frac{R^2}{|\tilde\rho|^2}\right)^{1/2}} \| \varphi \|_{H^{1/2}(\Gamma_R)}. \tag{4.7}$$

Set

$$\gamma(r) = e^{k\mathrm{Im}\,(\tilde r)\left(1-\frac{r^2}{|\tilde r|^2}\right)^{1/2}}.$$

Since $\tilde r = r(1+\mathbf{i}\hat\sigma)$, we obatin by simple calculation that

$$\gamma'(r) = \gamma(r)\cdot k\left(\frac{\sigma\hat\sigma}{(1+\hat\sigma^2)^{1/2}} + \frac{r\hat\sigma\hat\sigma'}{(1+\hat\sigma^2)^{3/2}}\right),$$

which, together with $r\hat\sigma' = \sigma - \hat\sigma \le \sigma$, implies

$$0 \le \gamma'(r) \le 2\sigma k\gamma(r) \quad \forall r \ge R. \tag{4.8}$$

Lemma 4.3. *Let $\Lambda(kR) = \max(1, \frac{|H_0^{(1)\prime}(kR)|}{|H_0^{(1)}(kR)|})$. Then there exists a constant $C > 0$ independent of k, R, ρ, and σ_0 such that*

$$\| |\alpha|^{-1}\gamma\nabla\phi \|_{L^2(\Omega^{\mathrm{PML}})} \le C\Lambda(kR)^{1/2}(1+kR)|\alpha_0|\| \varphi \|_{H^{1/2}(\Gamma_R)}.$$

Proof. We multiply (4.6a) by $\gamma^2\bar\phi$ and integrate over Ω^{PML} to obtain

$$\begin{aligned}
&\int_R^\rho\int_0^{2\pi}\gamma^2\left(\frac{\beta r}{\alpha}\left|\frac{\partial\phi}{\partial r}\right|^2 + \frac{\alpha}{\beta r}\left|\frac{\partial\phi}{\partial\theta}\right|^2\right)dr\,d\theta\\
&\quad = -\int_R^\rho\int_0^{2\pi}\left(\frac{\beta r}{\alpha}\frac{\partial\phi}{\partial r}(\gamma^2)'\bar\phi - \alpha\beta k^2 r\gamma^2|\phi|^2\right)dr\,d\theta\\
&\qquad + \int_0^{2\pi}\left[\frac{\beta r}{\alpha}\gamma^2\frac{\partial\phi}{\partial r}\bar\phi\right](\rho)d\theta - \int_0^{2\pi}\left[\frac{\beta r}{\alpha}\gamma^2\frac{\partial\phi}{\partial r}\bar\phi\right](R)d\theta.
\end{aligned}$$

Taking the real part of the equation we get

$$\begin{aligned}
&\int_R^\rho\int_0^{2\pi}\gamma^2\left(\frac{1+\sigma\hat\sigma}{1+\sigma^2}r\left|\frac{\partial\phi}{\partial r}\right|^2 + \frac{1+\sigma\hat\sigma}{1+\hat\sigma^2}\frac{1}{r}\left|\frac{\partial\phi}{\partial\theta}\right|^2\right)dr\,d\theta\\
&\quad \le \int_R^\rho\int_0^{2\pi}\left|\frac{\beta r}{\alpha}\frac{\partial\phi}{\partial r}2\gamma\gamma'\bar\phi\right|dr\,d\theta + \int_R^\rho\int_0^{2\pi}|\alpha\beta|k^2r\gamma^2|\phi|^2dr\,d\theta\\
&\qquad + \int_0^{2\pi}\left|\left[\frac{\beta r}{\alpha}\gamma^2\frac{\partial\phi}{\partial r}\bar\phi\right](\rho)\right|d\theta + \int_0^{2\pi}\left|\left[\frac{\beta r}{\alpha}\gamma^2\frac{\partial\phi}{\partial r}\bar\phi\right](R)\right|d\theta\\
&\quad := \mathrm{I}_1 + \cdots + \mathrm{I}_4.
\end{aligned} \tag{4.9}$$

Since $\gamma' \le 2k\sigma\gamma$ by (4.8), we obtain by Cauchy–Schwarz inequality and the fact $\hat\sigma \le \sigma$ that

$$\begin{aligned}
\mathrm{I}_1 &\le \left(\int_R^\rho\int_0^{2\pi}\gamma^2\frac{1+\sigma\hat\sigma}{1+\sigma^2}r\left|\frac{\partial\phi}{\partial r}\right|^2\right)^{1/2}\left(\int_R^\rho\int_0^{2\pi}16k^2\sigma^2\gamma^2\left|\frac{\beta}{\alpha}\right|^2\frac{1+\sigma^2}{1+\sigma\hat\sigma}\ r|\phi|^2\right)^{1/2}\\
&\le 4\left(\int_R^\rho\int_0^{2\pi}\gamma^2\frac{1+\sigma\hat\sigma}{1+\sigma^2}r\left|\frac{\partial\phi}{\partial r}\right|^2dr\,d\theta\right)^{1/2}\left(\int_R^\rho\int_0^{2\pi}k^2\sigma^2\gamma^2r|\phi|^2dr\,d\theta\right)^{1/2}.
\end{aligned}$$

On the other hand, by (4.5) and Lemma 2.2, we know that

$$\begin{aligned}\int_R^\rho \int_0^{2\pi} k^2\sigma^2\gamma^2 r|\phi|^2 dr\,d\theta &= 2\pi\sum_{n\in\mathbb{Z}}\int_R^\rho k^2\sigma^2\gamma^2 r\left|\frac{H_n^{(1)}(k\tilde r)}{H_n^{(1)}(kR)}\right|^2 dr\cdot|\hat\phi_n|^2\\ &\le 2\pi\sum_{n\in\mathbb{Z}}\int_R^\rho k^2\sigma^2 r dr\cdot|\hat\phi_n|^2\\ &\le C(1+kR)^2|\alpha_0|^2\|\,\phi\,\|^2_{L^2(\Gamma_R)}.\end{aligned}\tag{4.10}$$

Hence

$$\mathrm{I}_1 \le \frac{1}{2}\int_R^\rho\int_0^{2\pi}\gamma^2\frac{1+\sigma\hat\sigma}{1+\sigma^2}r\left|\frac{\partial\phi}{\partial r}\right|^2 dr\,d\theta + C(1+kR)^2|\alpha_0|^2\|\,\phi\,\|^2_{L^2(\Gamma_R)}.$$

By (4.10) we also have

$$\mathrm{I}_2 \le C(1+kR)^2|\alpha_0|^2\|\,\phi\,\|^2_{L^2(\Gamma_R)}.$$

Next, since $\tilde r'(r)=\alpha(r)$, by (4.5) and Lemma 2.3, we have

$$\begin{aligned}\mathrm{I}_3 &\le 2\pi\sum_{n\in\mathbb{Z}}\left|k\rho\beta(\rho)\gamma(\rho)^2\frac{H_n^{(1)\prime}(k\tilde\rho)}{H_n^{(1)}(kR)}\frac{H_n^{(1)}(k\tilde\rho)}{H_n^{(1)}(kR)}\right|\cdot|\hat\phi_n|^2\\ &\le 2\pi|\alpha_0|\sum_{n\ne0}k\rho\left(1+\frac{|n|}{|k\tilde\rho|}\right)|\hat\phi_n|^2 + 2\pi|\alpha_0|k\rho\left|\frac{H_0^{(1)\prime}(kR)}{H_0^{(1)}(kR)}\right|\cdot|\hat\phi_0|^2\\ &\le 2\pi|\alpha_0|\sum_{n\ne0}(k\rho+|n|)|\hat\phi_n|^2 + 2\pi|\alpha_0|k\rho\Lambda(kR)|\hat\phi_0|^2,\end{aligned}$$

where in the last inequality we have used the relation $\rho\le|\tilde\rho|$. Since $k\rho+|n|\le(1+k\rho)(1+n^2)^{1/2}$, we deduce finally

$$\mathrm{I}_3 \le C\Lambda(kR)(1+k\rho)|\alpha_0|\|\,\phi\,\|^2_{H^{1/2}(\Gamma_R)} \le C\Lambda(kR)(1+kR)|\alpha_0|\|\,\phi\,\|^2_{H^{1/2}(\Gamma_R)}.$$

Similarly, we can prove

$$\mathrm{I}_4 \le C\Lambda(kR)(1+kR)\|\,\phi\,\|^2_{H^{1/2}(\Gamma_R)}.$$

Substituting the estimates for $\mathrm{I}_1,\dots,\mathrm{I}_4$ into (4.9), we conclude that

$$\begin{aligned}&\int_R^\rho\int_0^{2\pi}\gamma^2\left(\frac{1+\sigma\hat\sigma}{1+\sigma^2}r\left|\frac{\partial\phi}{\partial r}\right|^2+\frac{1+\sigma\hat\sigma}{1+\hat\sigma^2}\frac{1}{r}\left|\frac{\partial\phi}{\partial\theta}\right|^2\right)dr\,d\theta\\ &\quad\le C\Lambda(kR)(1+kR)^2|\alpha_0|^2\|\,\phi\,\|^2_{H^{1/2}(\Gamma_R)}.\end{aligned}$$

This completes the proof. □

The following lemma is the main objective of this subsection.

LEMMA 4.4. *For any $\varphi\in H^1(\Omega_R)$, which is extended to be a function $\tilde\varphi\in H^1(\Omega^{\mathrm{PML}})$ according to* (4.1a)–(4.1b), *we have the following estimate:*

$$\|\,|\alpha|^{-1}\gamma\nabla\tilde\varphi\,\|_{L^2(\Omega^{\mathrm{PML}})} \le C\hat C^{-1}\Lambda(kR)^{1/2}(1+kR)|\alpha_0|\|\,\varphi\,\|_{H^{1/2}(\Gamma_R)}.$$

Proof. Let $w = \tilde{\varphi} - \bar{\phi}$; then from (4.1a)–(4.1b) and (4.6a)–(4.6b) we know that w satisfies

$$\nabla \cdot (A\nabla w) + \alpha\beta k^2 w = 0 \quad \text{in } \Omega^{\mathrm{PML}},$$
$$w = 0 \ \text{ on } \Gamma_R, \qquad w = -\bar{\phi} \quad \text{on } \Gamma_\rho.$$

By Theorem 2.4 and (4.7) we have

$$\begin{aligned}\| \,|\alpha|^{-1}\nabla w \,\|_{L^2(\Omega^{\mathrm{PML}})} &\le C\hat{C}^{-1}(1+kR)|\alpha_0|\| \,w\, \|_{H^{1/2}(\Gamma_\rho)}\\ &= C\hat{C}^{-1}(1+kR)|\alpha_0|\| \,\phi\, \|_{H^{1/2}(\Gamma_\rho)}\\ &\le C\hat{C}^{-1}(1+kR)|\alpha_0|e^{-k\mathrm{Im}\,(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}}\| \,\varphi\, \|_{H^{1/2}(\Gamma_R)}.\end{aligned}$$

By (4.8), γ is a increasing function, and we know that, for $r \le \rho$,

$$\gamma(r)e^{-k\mathrm{Im}\,(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \le \gamma(\rho)e^{-k\mathrm{Im}\,(\tilde{\rho})\left(1-\frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \le 1.$$

Hence

$$\| \,|\alpha|^{-1}\gamma\nabla w \,\|_{L^2(\Omega^{\mathrm{PML}})} \le C\hat{C}^{-1}(1+kR)|\alpha_0|\| \,\varphi\, \|_{H^{1/2}(\Gamma_R)}.$$

This completes the proof upon using Lemma 4.3. ◻

To conclude this subsection we remark that a direct consequence of this lemma is that

$$\| \,\omega^{-1}\nabla\varphi \,\|_{L^2(\Omega^{\mathrm{PML}})} \le C\hat{C}^{-1}\Lambda(kR)^{1/2}(1+kR)\| \,\varphi\, \|_{H^{1/2}(\Gamma_R)}. \tag{4.11}$$

4.3. Proof of Theorem 3.1. Since we are going to interpolate nonsmooth functions, we resort to an interpolation operator $\Pi_h : H^1_{(0)}(\Omega_\rho) \to \overset{\circ}{V}_h$ of Clement-type [9], where $H^1_{(0)}(\Omega_\rho) = \{v \in H^1(\Omega_\rho) : v = 0 \ \text{ on } \Gamma_\rho\}$. Let $\mathcal{N}_h = \{a_i\}_{i=1}^N$ be the set of the nodes of $\mathcal{M}_h$ which is interior to Ω^h_ρ or on the boundary Γ_D, and let $\{\phi_i\}_{i=1}^N$ be the corresponding nodal basis of V_h. Define $\Delta_i = \operatorname{supp}\phi_i \cap \Omega_\rho$. Then the interpolation operator $\Pi_h : H^1_{(0)}(\Omega_\rho) \to V_h$ is defined by

$$\Pi_h v(x) = \sum_{i=1}^N \left(\frac{1}{|\Delta_i|}\int_{\Delta_i} v(x)dx\right)\phi_i(x).$$

Since the nodes on Γ^h_ρ are not included in the definition of Π_h, we know that $\Pi_h v \in \overset{\circ}{V}_h$. Moreover, by slightly modifying the argument in [6, Lemmas 3.1 and 3.2], one can show that the operator Π_h enjoys the following interpolation estimates, for any $v \in H^1_{(0)}(\Omega_\rho)$,

$$\| \,v - \Pi_h v\, \|_{L^2(K)} \le Ch_K\| \,\nabla v\, \|_{L^2(\tilde{K})}, \quad \| \,v - \Pi_h v\, \|_{L^2(e)} \le Ch_e^{1/2}\| \,\nabla v\, \|_{L^2(\tilde{e})}, \tag{4.12}$$

where $\tilde{K}$ and $\tilde{e}$ are the union of all elements in $\mathcal{M}_h$ having nonempty intersection with $K \in \mathcal{M}_h$ and the side e, respectively.

Now we take $\varphi_h = \Pi_h\varphi \in \overset{\circ}{V}_h$ in the error representation formula (4.2) to get

$$\begin{aligned}a(u-u_h,\varphi) &= \int_{\Gamma_D} g(\overline{\varphi - \Pi_h\varphi}) - b(u_h, \varphi - \Pi_h\varphi) + \langle Tu_h - \hat{T}u_h, \varphi\rangle_{\Gamma_R}\\ &:= \mathrm{II}_1 + \mathrm{II}_2 + \mathrm{II}_3.\end{aligned} \tag{4.13}$$

We observe that, by integration by parts and using (3.5)–(3.7),

$$\mathrm{II}_1 + \mathrm{II}_2 = \sum_{K \in \mathcal{M}_h} \left(\int_K R_h (\overline{\varphi - \Pi_h \varphi}) dx + \sum_{e \subset \partial K} \frac{1}{2} \int_e J_e (\overline{\varphi - \Pi_h \varphi}) ds \right).$$

Standard argument in the a posteriori error analysis using (4.12) and (4.11) implies

$$\begin{aligned} |\mathrm{II}_1 + \mathrm{II}_2| &\leq C \sum_{K \in \mathcal{M}_h} \left(\| h_K R_h \|^2_{L^2(K)} + \frac{1}{2} \sum_{e \subset \partial K} \| h_e^{1/2} J_e \|^2_{L^2(e)} \right)^{1/2} \| \nabla \varphi \|_{L^2(\tilde{K})} \\ &\leq C \sum_{K \in \mathcal{M}_h} \eta_K \| \omega^{-1} \nabla \varphi \|_{L^2(\tilde{K})} \\ &\leq C \hat{C}^{-1} \Lambda (kR)^{1/2} (1 + kR) \left(\sum_{K \in \mathcal{M}_h} \eta_K^2 \right)^{1/2} \| \varphi \|_{H^{1/2}(\Gamma_R)}. \end{aligned}$$

By Lemma 2.5, we obtain

$$|\mathrm{II}_3| \leq C \hat{C}^{-1} (1 + kR)^2 |\alpha_0|^2 e^{-k \mathrm{Im}\,(\tilde{\rho}) \left(1 - \frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \| u_h \|_{H^{1/2}(\Gamma_R)} \| \varphi \|_{H^{1/2}(\Gamma_R)}.$$

Therefore, by the inf-sup condition (2.5), we finally get

$$\begin{aligned} \| u - u_h \|_{H^1(\Omega_R)} &\leq C \sup_{0 \neq \varphi \in H^1(\Omega_R)} \frac{|a(u - u_h, \varphi)|}{\| \varphi \|_{H^1(\Omega_R)}} \\ &\leq C \hat{C}^{-1} \Lambda (kR)^{1/2} (1 + kR) \left(\sum_{K \in \mathcal{M}_h} \eta_K^2 \right)^{1/2} \\ &\quad + C \hat{C}^{-1} (1 + kR)^2 |\alpha_0|^2 e^{-k \mathrm{Im}\,(\tilde{\rho}) \left(1 - \frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \| u_h \|_{H^{1/2}(\Gamma_R)}. \end{aligned}$$

This completes the proof. ◻

5. Implementation and numerical examples. The implementation of the adaptive algorithm in this section is based on the PDE toolbox of MATLAB. We use the a posteriori error estimate in Theorem 3.1 to determine the PML parameters. According to the discussion in section 2, we choose the PML medium property as the power function and thus we need only to specify the thickness $\rho - R$ of the layer and the medium parameter σ_0. Recall from Theorem 3.1 that the a posteriori error estimate consists of two parts: the PML error and the finite element discretization error. In our implementation we first choose ρ and σ_0 such that the exponentially decaying factor:

$$\hat{\omega} = e^{-k \mathrm{Im}\,(\tilde{\rho}) \left(1 - \frac{R^2}{|\tilde{\rho}|^2}\right)^{1/2}} \leq 10^{-8}, \tag{5.1}$$

which makes the PML error negligible compared with the finite element discretization errors. Once the PML region and the medium property are fixed, we use the standard finite element adaptive strategy to modify the mesh according to the a posteriori error estimate. Now we describe the adaptive algorithm we used in the paper.

ALGORITHM 5.1. Given tolerance TOL > 0. Let $m = 2$.

- Choose ρ and σ_0 such that the exponentially decaying factor $\hat{\omega} \leq 10^{-8}$.

- Set the computational domain $\Omega_\rho = B_\rho \backslash \bar{\Gamma}_D$ and generate an initial mesh $\mathcal{M}_h$ over Ω_ρ.
- While $\mathcal{E}_{FEM} = \left(\sum_{K\in\mathcal{M}_h} \eta_K^2\right)^{1/2} > \text{TOL}$ do

  – refine the mesh $\mathcal{M}_h$ according to the strategy:

  $$\text{if } \eta_K > \tfrac{1}{2}\max_{K\in\mathcal{M}_h}\eta_K, \text{ refine the element } K \in \mathcal{M}_h$$

  – solve the discrete problem (3.3) on $\mathcal{M}_h$
  – compute error estimators on $\mathcal{M}_h$
  end while

In the following we report two numerical examples to demonstrate the competitive behavior of the proposed algorithm. In the computations we first prescribe ρ and then determine σ_0 according to (5.1). We scale the error estimator for determining finite element meshes by a factor 0.15 as in the PDE toolbox of MATLAB.

Example 1. Let the scatter D be unit circle. We consider the scattering problem whose exact solution is known: $u = H_0^{(1)}(kr)$, where $r = |x|$. We take $R = 2$ and $k = 1$. Table 5.1 shows the different choices of the PML parameters ρ and σ_0 determined by the relation (5.1).

TABLE 5.1
The PML parameters for Examples 1 *and* 2.

Example 1		Example 2	
ρ	σ_0	ρ	σ_0
2R	30	2R	4
3R	15	3R	2
4R	10	4R	1
8R	5		

Figure 5.1 shows the $\log N_k$-$\log \| \nabla(u-u_k) \|_{L^2(\Omega_R)}$ curves, where N_k is the number of nodes of the mesh $\mathcal{M}_k$ and u_k is the finite element solution of (3.3) over the mesh $\mathcal{M}_k$. It indicates that the meshes and the associated numerical complexity are quasi-optimal: $\| \nabla(u - u_k) \|_{L^2(\Omega_R)} = CN_k^{-1/2}$ is valid asymptotically.

One of the important quantities in the scattering problems is the far field pattern:

$$u_\infty(\hat{x}) = \frac{e^{i\frac{\pi}{4}}}{\sqrt{8\pi k}} \int_{\partial D} \left(u(y) \frac{\partial e^{-ik\hat{x}\cdot y}}{\partial v(y)} - \frac{\partial u(y)}{\partial v(y)} e^{-ik\hat{x}\cdot y} \right) ds(y), \qquad \hat{x} = \frac{x}{|x|}.$$

We compute the far field $u_\infty(\hat{x})$, $\hat{x} = (\cos(\theta), \sin(\theta))^T$ in the observation direction $\theta = \pi/4$. Figure 5.2 shows the far fields for different choices of PML parameters ρ and σ_0. We observe that our adaptive algorithm is robust with respect to the choice of the thickness of PML layer: the far fields of the scattering solutions are insensitive to the choices of the PML parameters.

Example 2. This example is taken from [10] which concerns the scattering of the plane wave $u_I = e^{\mathbf{i}kx_1}$ from a perfectly conducting metal. The scatter D is contained in the box $\{x \in \mathbb{R} :\ -2 < x_1 < 2.2, -0.7 < x_2 < 0.7\}$ as plotted in Figure 5.3. We take $R = 3$ and $k = 2\pi$. The different choices of PML parameters ρ and σ_0 determined by the relation (5.1) are shown in Table 5.1.

Figure 5.4 shows the $\log N_k$–$\log \mathcal{E}_k$ curves, where N_k is the number of nodes of the mesh $\mathcal{M}_k$ and the $\mathcal{E}_k = (\sum_{K\in\mathcal{M}_k} \eta_K^2)^{1/2}$ is the associated a posteriori error estimate. It indicates that the meshes and the associated numerical complexity are quasi-optimal: $\mathcal{E}_k = CN_k^{-1/2}$ is valid asymptotically.

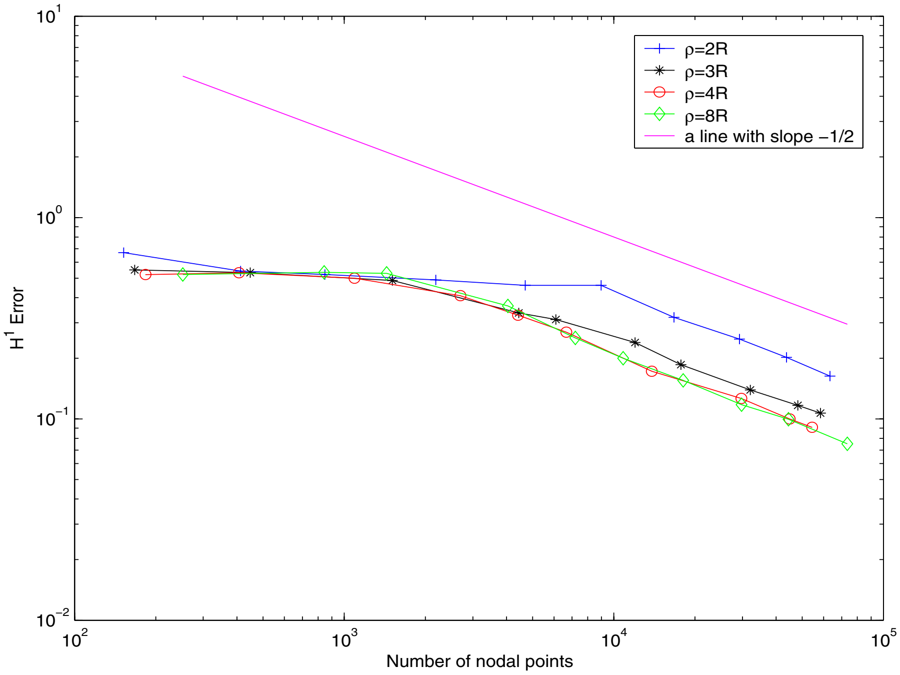


FIG. 5.1. *Quasi-optimality of the adaptive mesh refinements of the error* $\|\nabla(u - u_h)\|_{L^2(\Omega_R)}$ *for Example* 1.

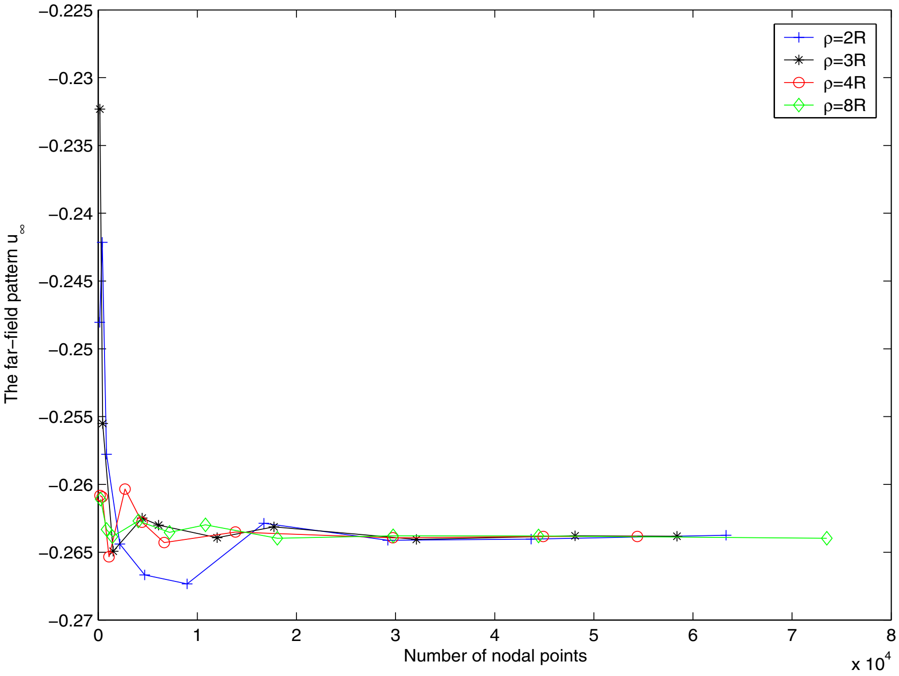


FIG. 5.2. *The real part of the far fields when the observing angle* $\theta = \pi/4$ *for Example* 1.

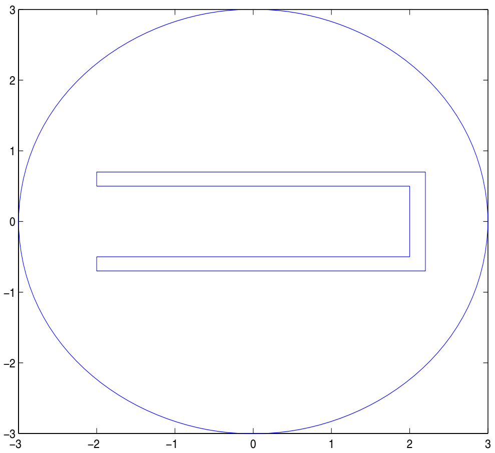


FIG. 5.3. *The geometry of the scatter for Example* 2.

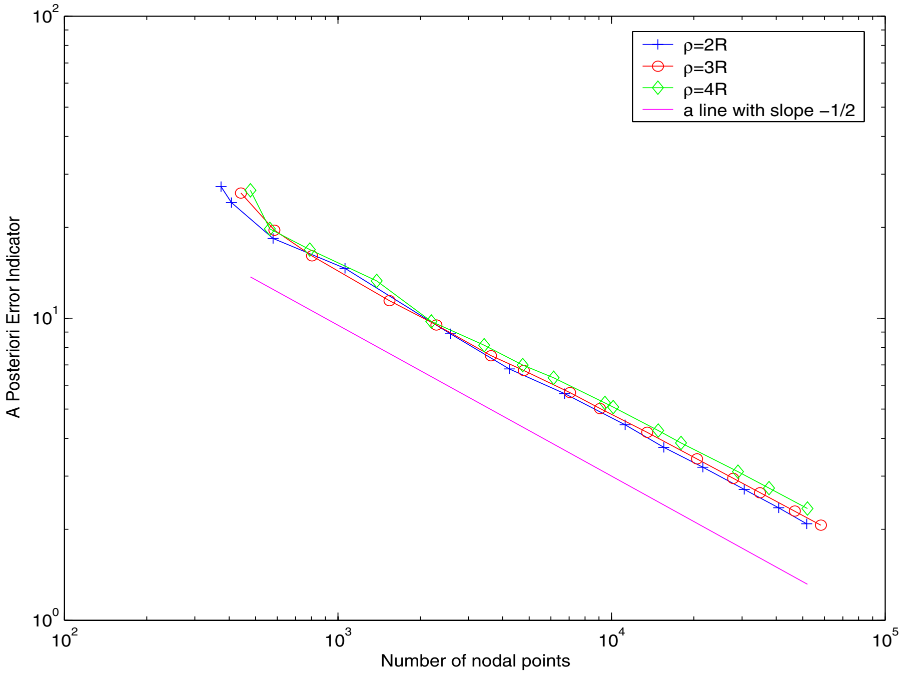


FIG. 5.4. *Quasi-optimality of the adaptive mesh refinements of the a posteriori error estimator for Example* 2.

Figures 5.5 and 5.6 show the far fields in the incident direction $\theta = 0$ and the reflective direction $\theta = \pi$. Again we observe that the far fields are insensitive to the choices of PML parameters.

In Figure 5.7 we show the mesh after 13 adaptive iterations when $\rho = 3R$. We observe that the mesh near the boundary Γ_ρ is rather coarse, as a consequence of the exponentially decaying factor in our finite element a posteriori error estimator.

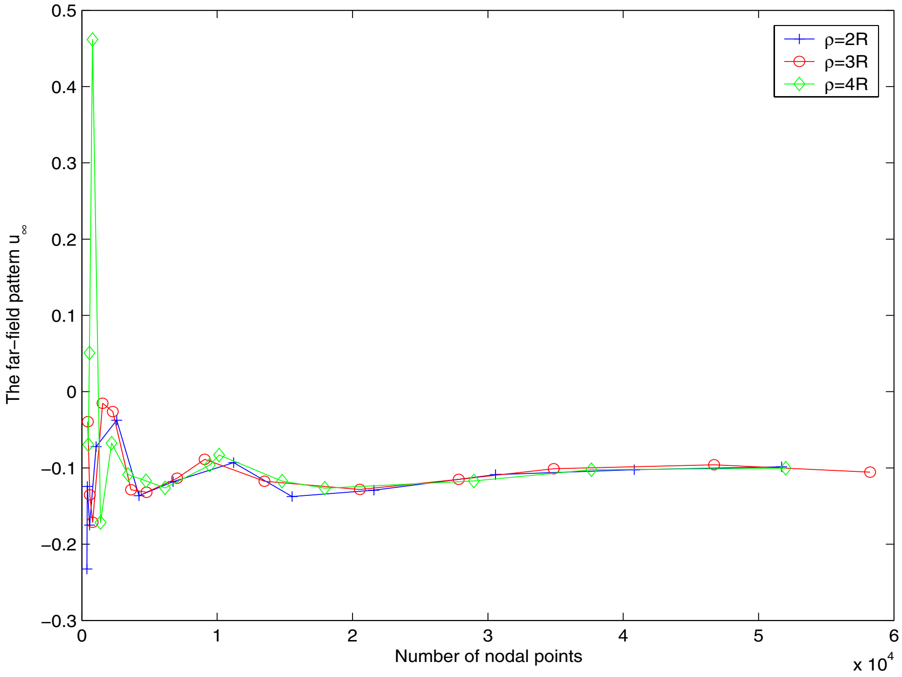


FIG. 5.5. *The real part of the far fields in the incident direction for Example* 2.

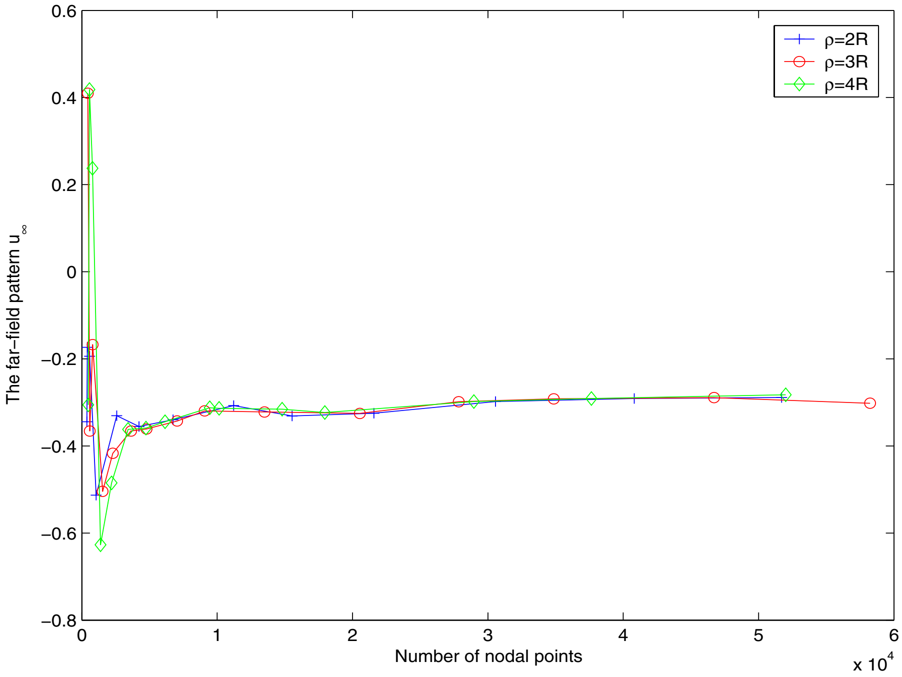


FIG. 5.6. *The real part of the far fields in the reflective direction for Example* 2.

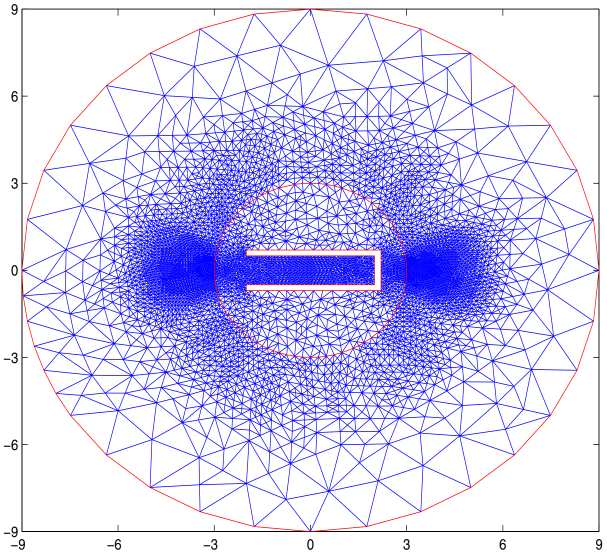


FIG. 5.7. *The mesh of* 7048 *nodes after* 13 *adaptive iterations when* $\rho = 3R$ *for Example* 2.

Appendix. The PML equation in the layer for large σ_0. The purpose of the appendix is to show that for sufficiently large σ_0, the PML problem in the layer (2.23) has a unique solution w. Moreover, the constant $\hat{C}$ in (2.25) can be chosen independent of Ω^{PML} and k.

From (H1) we know that $\beta(r) = 1 + \mathbf{i}\hat{\sigma}(r)$, where

$$\hat{\sigma}(r) = \frac{1}{r}\int_R^r \sigma(t)dt = \frac{\sigma_0}{m+1}\frac{r-R}{r}\left(\frac{r-R}{\rho-R}\right)^m.$$

Define

$$\zeta(r) := \frac{2\sigma_0^2}{\sigma\hat{\sigma}(r)} = \frac{2(m+1)r(\rho-R)^{2m}}{(r-R)^{2m+1}} \qquad \forall r > R.$$

It is clear that $\zeta : (R,\infty) \to \mathbb{R}$ is strictly monotone decreasing and $\zeta(r) \to \infty$ as $r \to R$, $\zeta(r) \to 0$ as $r \to \infty$. Thus, for any $\sigma_0 > 0$, there exists a unique $\hat{R} = \hat{R}(\sigma_0) > R$ such that

$$\sigma_0^2 = \zeta(\hat{R}) = \frac{2(m+1)\hat{R}(\rho-R)^{2m}}{(\hat{R}-R)^{2m+1}}. \tag{5.1}$$

Hence, since $\sigma\hat{\sigma} : (R,\infty) \to \mathbb{R}$ is increasing, we have

$$\sigma\hat{\sigma}(r) \ge \sigma\hat{\sigma}(\hat{R}) = \frac{2\sigma_0^2}{\zeta(\hat{R})} = 2 \qquad \text{for } r \ge \hat{R}. \tag{5.2}$$

In this appendix we make the following assumption on the choice of σ_0:

(H3) $\sigma_0^2 \ge \zeta(\hat{R}_{\max})$, where $\hat{R}_{\max} := \max\{r \in (R,\rho) : \theta(r) \le 1\}$ with

$$\theta(r) = k^2R^2\left[\left(\frac{r^2}{R^2} - 1\right)\ln\frac{r}{R} + \frac{2r^2(r-R)}{(2m+1)R^3}\right] \qquad \forall r \ge R.$$

Since the function $\theta : (R, \rho) \to \mathbb{R}$ is strictly monotone increasing and $\theta(R) = 0$, $\hat{R}_{\max}$ is well-defined.

LEMMA 5.1. *Let* (H1) *and* (H3) *be satisfied. Then*

$$\int_R^{\hat{R}} k^2 r \left(\int_R^r \frac{1 + \sigma^2(t)}{t} dt \right) dr \leq \frac{1}{2}.$$

Proof. First we have

$$\int_R^{\hat{R}} r \left(\int_R^r \frac{1}{t} dt \right) dr = \int_R^{\hat{R}} r \ln \frac{r}{R} dr \leq \frac{1}{2}(\hat{R}^2 - R^2) \ln \frac{\hat{R}}{R}.$$

Next by (H1) and (5.1) we know that

$$\begin{aligned}
\int_R^{\hat{R}} r \left(\int_R^r \frac{\sigma^2}{t} dt \right) dr &\leq \frac{\hat{R}}{R} \int_R^{\hat{R}} \left(\int_R^r \sigma^2(t) dt \right) dr \\
&= \frac{\hat{R}}{R} \int_R^{\hat{R}} \frac{\sigma_0^2}{2m+1} \frac{(r - R)^{2m+1}}{(\rho - R)^{2m}} dr \\
&= \frac{\hat{R}}{R} \frac{\sigma_0^2}{(2m+1)(2m+2)} \frac{(\hat{R} - R)^{2m+2}}{(\rho - R)^{2m}} \\
&= \frac{\hat{R}^2(\hat{R} - R)}{(2m+1)R}.
\end{aligned}$$

Thus

$$\begin{aligned}
\int_R^{\hat{R}} k^2 r \left(\int_R^r \frac{1 + \sigma^2(t)}{t} dt \right) dr &\leq \frac{1}{2} k^2 R^2 \left[\left(\frac{\hat{R}^2}{R^2} - 1 \right) \ln \frac{\hat{R}}{R} + \frac{2\hat{R}^2(\hat{R} - R)}{(2m+1)R^3} \right] \\
&= \frac{1}{2} \theta(\hat{R}).
\end{aligned}$$

Now if $\sigma_0^2 \geq \zeta(\hat{R}_{\max})$, we know from the monotonicity of ζ that $\hat{R} = \hat{R}(\sigma_0) \leq \hat{R}_{\max}$. Thus $\theta(\hat{R}) \leq \theta(\hat{R}_{\max}) \leq 1$ by (H3). This completes the proof. □

Now we are ready to prove the main result of this appendix.

THEOREM 5.2. *Under the assumptions* (H1) *and* (H3) *there exists a constant* $C > 0$ *independent of* k, R, ρ, *and* σ_0 *such that*

$$\operatorname{Re}[\hat{b}(v, v)] \geq C \| v \|^2_{*, \Omega^{\mathrm{PML}}} \qquad \forall v \in H_0^1(\Omega^{\mathrm{PML}}).$$

Proof. For any $v \in H_0^1(\Omega^{\mathrm{PML}})$, we have

$$\operatorname{Re}[\hat{b}(v, v)] = \int_R^{\rho} \int_0^{2\pi} \left[\frac{1 + \sigma\hat{\sigma}}{1 + \sigma^2} r \left| \frac{\partial v}{\partial r} \right|^2 + \frac{1 + \sigma\hat{\sigma}}{1 + \hat{\sigma}^2} \frac{1}{r} \left| \frac{\partial v}{\partial \theta} \right|^2 + (\sigma\hat{\sigma} - 1) k^2 r |v|^2 \right].$$

By (5.2) we know that

$$\begin{aligned}
&\int_R^{\rho} \int_0^{2\pi} (\sigma\hat{\sigma} - 1) k^2 r |v|^2 dr \, d\theta \\
&\quad = \int_R^{\hat{R}} \int_0^{2\pi} (\sigma\hat{\sigma} - 1) k^2 r |v|^2 dr \, d\theta + \int_{\hat{R}}^{\rho} \int_0^{2\pi} (\sigma\hat{\sigma} - 1) k^2 r |v|^2 dr \, d\theta \\
&\quad \geq -\frac{3}{2} \int_R^{\hat{R}} \int_0^{2\pi} k^2 r |v|^2 dr \, d\theta + \frac{1}{4} \int_R^{\rho} \int_0^{2\pi} (1 + \sigma\hat{\sigma}) k^2 r |v|^2 dr \, d\theta.
\end{aligned}$$

Notice that since $v = 0$ on Γ_R,

$$|v(r)| = \left| \int_R^r \frac{\partial v}{\partial r} dr \right| \leq \left(\int_R^r \frac{1+\sigma\hat{\sigma}}{1+\sigma^2} t \left| \frac{\partial v}{\partial r} \right|^2 dt \right)^{1/2} \left(\int_R^r \frac{1}{t} \frac{1+\sigma^2}{1+\sigma\hat{\sigma}} dt \right)^{1/2},$$

which, by Lemma 5.1, yields

$$\begin{aligned} \int_R^{\hat{R}} \int_0^{2\pi} k^2 r |v|^2 dr &\leq \left(\int_R^{\hat{R}} \int_0^{2\pi} \frac{1+\sigma\hat{\sigma}}{1+\sigma^2} r \left| \frac{\partial v}{\partial r} \right|^2 \right) \cdot \int_R^{\hat{R}} k^2 r \left(\int_R^r \frac{1}{t} \frac{1+\sigma^2}{1+\sigma\hat{\sigma}} dt \right) \\ &\leq \frac{1}{2} \int_R^{\hat{R}} \int_0^{2\pi} \frac{1+\sigma\hat{\sigma}}{1+\sigma^2} r \left| \frac{\partial v}{\partial r} \right|^2. \end{aligned}$$

Thus

$$\operatorname{Re}[\hat{b}(v,v)] \geq \frac{1}{4} \int_R^{\rho} \int_0^{2\pi} \left[\frac{1+\sigma\hat{\sigma}}{1+\sigma^2} r \left| \frac{\partial v}{\partial r} \right|^2 + \frac{1+\sigma\hat{\sigma}}{1+\hat{\sigma}^2} \frac{1}{r} \left| \frac{\partial v}{\partial \theta} \right|^2 + (1+\sigma\hat{\sigma}) k^2 r |v|^2 \right].$$

This completes the proof. □

Acknowledgments. The authors wish to thank Peter Monk for several inspiring discussions and the referees for constructive comments.

REFERENCES

[1] I. Babuška and A. Aziz, *Survey lectures on mathematical foundations of the finite element method,* in The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations, A. Aziz, ed., Academic Press, New York, 1973, pp. 5–359.

[2] I. Babuška and C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[3] J.-P. Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[4] Z. Chen and S. Dai, *Adaptive Galerkin methods with error control for a dynamical Ginzburg–Landau model in superconductivity,* SIAM J. Numer. Anal., 38 (2001), pp. 1961–1985.

[5] Z. Chen and S. Dai, *On the efficiency of adaptive finite element methods for elliptic problems with discontinuous coefficients*, SIAM J. Sci. Comput., 24 (2002), pp. 443–462.

[6] Z. Chen and R.H. Nochetto, *Residual type a posteriori error estimates for elliptic obstacle problems*, Numer. Math., 84 (2000), pp. 527–548.

[7] Z. Chen, R.H. Nochetto, and A. Schmidt, *A characteristic Galerkin method with adaptive error control for continuous casting problem*, Comput. Methods Appl. Mech. Engrg., 189 (2000), pp. 249–276.

[8] Z. Chen and H. Wu, *An adaptive finite element method with perfectly matched absorbing layers for the wave scattering by periodic structures*, SIAM J. Numer. Anal., 41 (2003), pp. 799–826.

[9] Ph. Clement, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numer., 9 (1975), pp. 77–84.

[10] F. Collino and P.B. Monk, *The perfectly matched layer in curvilinear coordinates*, SIAM J. Sci. Comput., 19 (1998), pp. 2061–2090.

[11] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley and Sons, New York, 1983.

[12] T. Hohage, F. Schmidt, and L. Zschiedrich, *Solving time-harmonic scattering problems based on the pole condition.* II: *Convergence of the PML method*, SIAM J. Math. Anal., 35 (2003), 547–560.

[13] M. Lassas and E. Somersalo, *On the existence and convergence of the solution of PML equations*, Computing, 60 (1998), pp. 229–241.

[14] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, 2000.

[15] P. Monk, *A posteriori error indicators for Maxwell's equations*, J. Comput. Appl. Math., 100 (1998), 173–190.

[16] P. Monk and E. Süli, *The adaptive computation of far-field patterns by a posteriori error estimation of linear functionals*, SIAM J. Numer. Anal., 36 (1998), pp. 251–274.

[17] P. Morin, R.H. Nochetto, and K.G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[18] A.H. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.

[19] F.L. Teixeira and W.C. Chew, *Advances in the theory of perfectly matched layers*, in Fast and Efficient Algorithms in Computational Electromagnetics, W.C. Chew et al., eds., Artech House, Boston, 2001, pp. 283–346.

[20] E. Turkel and A. Yefet, *Absorbing PML boundary layers for wave-like equations*, Appl. Numer. Math., 27 (1998), pp. 533–557.

[21] G.N. Watson, *A Treatise on The Theory of Bessel Functions,* Cambridge University Press, Cambridge, 1922.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 672–685

ANALYSIS OF FIRST ORDER ERRORS IN SHOCK CALCULATIONS IN TWO SPACE DIMENSIONS*

MALIN SIKLOSI† AND GUNILLA EFRAIMSSON‡

Abstract. Numerical computations show that solutions of hyperbolic conservation laws obtained by second or higher order shock capturing methods in many cases are only first order accurate downstream of shocks (see, e.g., [M. H. Carpenter and J. H. Casper, *AIAA J.*, 37 (1999), pp. 1072–1079]). We use matched asymptotic expansions to analyze the degeneration in order of accuracy for stationary solutions of hyperbolic conservation laws in two space dimensions.

Key words. hyperbolic conservation laws, shock wave, shock capturing, artificial viscosity, asymptotic analysis

AMS subject classifications. 35L65, 35L67, 65M06, 65M02

DOI. 10.1137/040603462

1. Introduction.

Numerical results presented in this and other papers (see, e.g., [2], [3], [4], [5], [7], and [16]) show that solutions of hyperbolic conservation laws in one and two space dimensions obtained by formally second or higher order accurate shock capturing schemes degenerate to first order downstream of shock layers. A numerical study by Carpenter and Casper [3] shows that the error depends on flow conditions and generally increases with increasing shock strength. They further conclude that the first order error terms seem to be nearly independent of the design order of accuracy of the method.

The first order term for reasonable mesh-sizes seems to be small in many cases. However, in applications where the small scale behavior is of significance, such as aeroacoustics, the degeneration to first order accuracy can be troublesome. It is also important to understand the phenomenon more deeply in order, hopefully, to be able to design new methods which do not suffer from this deficiency.

This paper is the fourth in a series of papers. In [5], an explanation of the degeneracy in accuracy was given for the case of steady-state solutions of systems with source terms in one space dimension, and, in [16], it was shown how to raise the accuracy and remove the first order error with a specific choice of a matrix valued artificial viscosity. In [23], the same method of analysis was extended to time-dependent solutions of hyperbolic systems in one space dimension. Also in that case, it was possible to avoid the downstream error by using a specific matrix valued artificial viscosity. In this paper we consider steady-state solutions of hyperbolic conservation laws in two space dimensions. We present an analysis which reveals how the downstream error is related to the dissipative terms in the numerical scheme.

Shock capturing methods for hyperbolic conservation laws consist of a discretization of the conservation law, augmented by some mechanism for adding dissipation in the vicinity of discontinuities in the solution without explicitly keeping track of

*Received by the editors January 23, 2004; accepted for publication (in revised form) April 19, 2005; published electronically August 31, 2005.
http://www.siam.org/journals/sinum/43-2/60346.html

†Department of Numerical Analysis and Computer Science, Royal Institute of Technology, SE-100 44 Stockholm, Sweden (malins@nada.kth.se).

‡Department of Aeronautical and Vehicle Engineering, Royal Institute of Technology, SE-100 44 Stockholm, Sweden (gef@kth.se).

the position of the discontinuities. The dissipative terms are either intrinsic in the scheme or explicitly added (artificial viscosity). In order to avoid spurious oscillations around shocks, the dissipative terms must be $\mathcal{O}(h)$ in the shock layer. Here, h is the grid size. In smooth regions, the dissipative terms are of order $\mathcal{O}(h^p)$ or smaller, where p is the order of accuracy of the discretization of the conservation law. Many shock capturing schemes for problems in several space dimensions are obtained as a straightforward generalization of the corresponding one-dimensional scheme. This means that the dissipative terms are turned on separately in the x- and y-directions. Hence, at shocks aligned to one of the coordinate axes, dissipative terms only act normally to the shock. For shocks that are oblique relative to the grid, dissipative terms are switched on in both space directions.

A widely used technique to model the behavior of numerical schemes is to study the solutions of the so-called modified equation; see [19]. The modified equation is obtained by Taylor series expansions. The goal is to obtain a PDE which approximates the discrete numerical solution better than the original, continuous PDE. The technique has mostly been used to investigate the dispersive and dissipative properties of PDE schemes. It has, in particular, been widely used to investigate the behavior of numerical schemes in the vicinity of shocks. In the shock region, the modified equation can be proved to be valid only for weak shocks; see [9]. Due to the strong gradients in the shock region, it is not obvious what terms in the Taylor expansion will be dominant; see, e.g., [14]. Considering the wide use, there is surprisingly little literature considering the validity and scope of the method of modified equations; see [10], [24], and references therein.

Many shock capturing schemes that are in wide use, such as the scheme due to Jameson, Schmidt, and Turkel [12], are based on a relatively simple finite difference/finite volume discretization of the conservation law. Mechanisms for the extra dissipation that is needed at discontinuities are then explicitly added. The dissipation mechanisms usually consist of some kind of a switch function multiplying a finite difference/finite volume discretization of a second derivative term. It is straightforward to obtain a modified equation, modeling the finite difference/finite volume discretizations of the conservation law and the second derivative term. The construction of the switch functions is often rather complicated and it is not as easy to obtain a good model. However, e.g., for the switch introduced by Jameson, Schmidt, and Turkel in [12], it is reasonable to model the switch function by a smooth function, which is one in the shock region and vanishes a fixed number of grid points away from the shock; see computations in [6].

For more advanced schemes, where the dissipative mechanisms are more intrinsic in the scheme, e.g., higher order WENO schemes, it is difficult to find an analytic formulation of the modified equation and, hence, to construct a model for the numerical scheme. Theoretical studies of "continuous discrete shock profiles" (see [13], [20], [22], and references therein) and numerical investigations, such as the work by Arora and Roe [1], indicate, however, that for many of these more advanced methods, the numerical solution, also in the vicinity of shock waves, can be modeled by a continuous PDE.

In section 2, we suggest a model for numerical solutions obtained by shock capturing schemes. The model consists of the modified equation with boundary conditions. We then analyze this modified problem by a matched asymptotic analysis. It is assumed that an inner solution is valid in the shock layer and an outer solution is valid elsewhere. The two solutions are matched in a region of overlap. From the analysis, we see that generally the outer solution contains a term of $\mathcal{O}(h)$ downstream of the shock.

The phenomenon has also been studied by other methods in [7] and [4]. In [7], analytic examples in one dimension are constructed where the numerical solution is only first order accurate downstream of a shock, although the numerical scheme is formally second order. It is also shown that a converging numerical method will yield solutions having the formal order of accuracy in domains where no characteristics have passed through a shock. In [4], the first order downstream error is numerically detected in solutions of a one-dimensional (1D) shock–sound interaction problem solved by a fourth order ENO method. A scalar, linear equation is used to model the problem. It can be seen that the solution of the model problem computed with the fourth order ENO method behaves qualitatively differently depending on if the discontinuity is located on a cell interface or in the interior of a cell. In the first case, the solution is fourth order in all of the domain, but in the second case the solution is only first order downstream of the discontinuity. On the basis of this observation, the numerical method is modified such that the shock position will always be on a cell interface, and the fourth order accuracy of the solution of the shock–sound interaction problem is obtained both upstream and downstream.

2. Analysis.

In this section we will consider a hyperbolic system in two space dimensions with a stationary solution containing a shock. We will model the numerical solution and present an analysis that shows how the $\mathcal{O}(1)$ error in the shock region causes an $\mathcal{O}(h)$ error downstream of the shock. From the analysis, we see that the first order downstream error can only appear if there is a variation of the solution tangentially to the shock, i.e., if the solution is truly two-dimensional (2D).

Many numerical schemes for problems in several space dimensions are obtained as a straightforward generalization of the corresponding 1D scheme. This means that the dissipative terms are turned on separately in the x- and y-directions. Hence, at shocks aligned to one of the coordinate axes, dissipative terms only act normally to the shock. At a shock which is oblique relative to the grid, however, there are dissipative terms in both coordinate directions.

First, we consider a straight shock which is aligned with the y-axis. Then, we show how the analysis can be extended to a more general shock. The essential conclusion is the same in both cases; downstream the solution is only first order accurate.

2.1. The inviscid problem.

We consider a system of conservation laws

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x + \mathbf{g}(\mathbf{u})_y = 0, \quad (x, y) \in \Omega, \quad t > 0, \tag{2.1}$$

where $\mathbf{u}(x, y, t) \in \mathbf{R}^m$ and $\mathbf{f}, \mathbf{g} : \mathbf{R}^m \to \mathbf{R}^m$. We assume that the linearization of (2.1) is strongly hyperbolic, i.e., there exists a smooth matrix function $S(\mathbf{u}, \bar{\omega})$ such that $S\breve{\mathbf{f}}'S^{-1}$ is diagonal with real entries. The Jacobian $\breve{\mathbf{f}}'$ is defined by $\breve{\mathbf{f}}'(\mathbf{u}, \bar{\omega}) = \omega^x \mathbf{f}'(\mathbf{u}) + \omega^y \mathbf{g}'(\mathbf{u})$, and $\bar{\omega} = (\omega^x, \omega^y)$ is an arbitrary unit vector. The eigenvalues of $\breve{\mathbf{f}}'$ are denoted by $\breve{\lambda}_k(\mathbf{u}, \bar{\omega})$, $k = 1, 2, \ldots, m$, and are in the increasing order.

Across discontinuities, $\mathbf{u}$ satisfies the Rankine–Hugoniot condition; see, e.g., [11]. If $\mathbf{u}$ has a discontinuity across a curve Σ in the (x, y)-plane, the Rankine–Hugoniot condition can be formulated as

$$n^x[\mathbf{f}] + n^y[\mathbf{g}] + s[\mathbf{u}] = 0, \tag{2.2}$$

where $\bar{n} = (n^x, n^y)$ is the unit normal of Σ, and $s = -\bar{v} \cdot \bar{n}$ is the speed of the discontinuity in the negative normal direction; see Figure 2.1. By the notation $[\mathbf{u}]$,

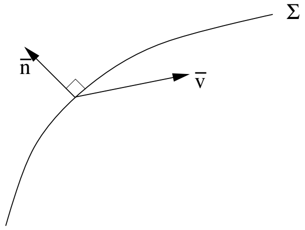


FIG. 2.1. *A discontinuity surface Σ moving with velocity $\bar{v}$. Here, $\bar{n}$ denotes the unit normal of Σ.*

we denote the jump in $\mathbf{u}$ across the discontinuity. We will define this notation in a precise way later.

We call the discontinuity a k-shock if it satisfies the Lax entropy condition, i.e., there is an integer k, $1 \leq k \leq m$, such that

$$\breve{\lambda}_k(\mathbf{u}^-, \bar{n}) > s > \breve{\lambda}_k(\mathbf{u}^+, \bar{n})$$

so that the k-characteristic impinges on the discontinuity from both sides, while the other characteristics cross the discontinuity:

$$\begin{aligned} &\breve{\lambda}_j(\mathbf{u}^-, \bar{n}) < s \quad \text{and} \quad \breve{\lambda}_j(\mathbf{u}^+, \bar{n}) < s \quad \text{for } j = 1, 2, \dots, k-1, \\ &\breve{\lambda}_j(\mathbf{u}^-, \bar{n}) > s \quad \text{and} \quad \breve{\lambda}_j(\mathbf{u}^+, \bar{n}) > s \quad \text{for } j = k+1, k+2, \dots, m. \end{aligned}$$

Here, $\mathbf{u}^\pm$ denotes the value of $\mathbf{u}$ on each side of the shock, and this notation will be made precise later. The Lax entropy condition is a necessary condition for linearized stability of the shock front; see, e.g., [21]. It is also necessary for linearized stability of the shock front that the diagonalization matrix S at the shock front satisfies a certain determinant condition. For the general formulation of this determinant condition; see, e.g., [25] or [21]. In Assumption 1, we will formulate the determinant condition for the special case we consider.

We let the domain Ω be a strip along the positive y-axis, which has width $2a$ and is infinitely long, i.e.,

$$\Omega = \{(x, y) : |x| \leq a, y > 0\}.$$

Assumption 1. We assume that there exists a steady-state solution of (2.1) which satisfies the Rankine–Hugoniot condition (2.2) at discontinuities. We denote the steady-state solution $\mathbf{u} = \mathbf{u}(x, y)$. We further assume that the initial and boundary conditions are chosen such that $\mathbf{u}$ contains a single 1-shock, which is straight and aligned with the y-axis, where the y-dependence is introduced via the boundary conditions. Also,

$$\mathbf{u}(x, 0) = \begin{cases} \mathcal{U}^+(x) & \text{for } x > 0, \\ \mathcal{U}^-(x) & \text{for } x < 0, \end{cases}$$

where $\mathcal{U}^\pm(x) \in \mathbf{R}^m$ are smooth functions.

We will use the following notation:

$$[\mathbf{u}(\cdot, y)] = \mathbf{u}^+(y) - \mathbf{u}^-(y), \quad \text{where } \mathbf{u}^\pm(y) = \lim_{\xi \to 0^+} \mathbf{u}(\pm\xi, y).$$

Corresponding notation for other quantities will be used frequently.

Let the $m \times m$ matrix D be defined by

$$D = (S_{II}^{+} \quad [\mathbf{g}]). \tag{2.3}$$

Here the columns of S_{II}^{+} are the eigenvectors of $\mathbf{f}'(\mathbf{u}^{+}(y))$ corresponding to the eigenvalues $\lambda_j(\mathbf{u}^{+}),\ j = 2, 3, \dots, m$. We assume

$$\det D \neq 0.$$

We consider a boundary condition at $x = \pm a$ of the form

$$\mathbf{u} = \mathbf{h}(y) \quad \text{at } x = -a, \tag{2.4}$$

$$R(\mathbf{u}) = r(y) \quad \text{at } x = a, \tag{2.5}$$

where $\mathbf{h}(y) \in \mathbf{R}^n$ and $r(y) \in \mathbf{R}$ are given. Here, $R(\mathbf{u}) : \mathbf{R}^n \to \mathbf{R}$ is a nonlinear function. Hence, one boundary condition is given for each ingoing characteristic. On both sides of the y-axis, $\mathbf{u}(x, y)$ is assumed to be smooth without discontinuities. We assume that the eigenvalues of $\mathbf{f}'(\mathbf{u})$ have constant sign on each side of the shock interface.

Under the conditions in Assumption 1, the Rankine–Hugoniot conditions simplify to

$$[\mathbf{f}(\cdot, y)] = 0, \tag{2.6}$$

and the shock is a 1-shock if

$$\begin{gathered} \lambda_1(\mathbf{u}^-) > 0 > \lambda_1(\mathbf{u}^+), \\ \lambda_j(\mathbf{u}^-) > 0, \quad \text{and} \quad \lambda_j(\mathbf{u}^+) > 0 \quad \text{for } j = 2, 3, \dots, m, \end{gathered}$$

where $\lambda_j(\mathbf{u}),\ j = 1, 2, \dots, m$, are the eigenvalues of the Jacobian $\mathbf{f}'(\mathbf{u})$.

Remark. For 1-shocks and n-shocks there is just one downstream side. Hence, the first order error appears only on one side of the shock. For other Lax shocks, both sides of the shock are downstream sides, and the first order error appears on both sides. The phenomenon can be analyzed by the same method in both cases, but the analysis becomes less involved when only one side must be considered. For clarity and without loss of generality, we analyze a 1-shock.

Remark. Instead of boundary conditions of the form (2.4) and (2.5), one could use more general boundary conditions. This would not change the result of the analysis.

2.2. The modified problem. On the basis of the considerations in the introduction, we use the equations given below to model a generic second order shock capturing scheme for solving (2.1) in the case specified in Assumption 1.

Assumption 2. Let $\mathbf{u}^{\varepsilon}$ be the solution of

$$\mathbf{f}(\mathbf{u}^{\varepsilon})_x + \mathbf{g}(\mathbf{u}^{\varepsilon})_y = \varepsilon \left(\phi \left(\frac{x - x^{\varepsilon}(y)}{\varepsilon} \right) \mathbf{u}_x^{\varepsilon} \right)_x + c_2 \varepsilon^2 (\mathbf{u}_{xx}^{\varepsilon} + \mathbf{u}_{yy}^{\varepsilon}), \quad (x, y) \in \Omega, \tag{2.7}$$

where $\varepsilon = c_1 h$ and c_1 and c_2 are scalar constants. The function $\phi(z)$ is a smooth one-variable function which models the switch mechanism, i.e.,

$$\phi(z) = \begin{cases} 1 & \text{for } |z| \leq K_0, \\ 0 & \text{for } |z| \geq K_1, \end{cases}$$

where $K_0 < K_1$ are positive constants with K_0 sufficiently large.

The boundary conditions for $\mathbf{u}^\varepsilon$ are

$$\mathbf{u}^\varepsilon = \mathbf{h}(y) \quad \text{at } x = -a, \tag{2.8}$$

$$R(\mathbf{u}^\varepsilon) = r(y) \quad \text{at } x = a \tag{2.9}$$

together with boundary conditions that model the numerical boundary conditions, such as extrapolation of outgoing characteristic variables. We assume that the numerical boundary conditions are chosen such that the possible boundary layer effects are $\mathcal{O}(h^2)$ or smaller. Also,

$$\mathbf{u}^\varepsilon(x,0) = \begin{cases} \mathcal{U}^+(x) & \text{for } \frac{x}{\varepsilon} > K_1, \\ \hat{\mathbf{U}}(x/\varepsilon, 0) & \text{for } |\frac{x}{\varepsilon}| \leq K_1, \\ \mathcal{U}^-(x) & \text{for } \frac{x}{\varepsilon} < K_1, \end{cases}$$

where $\hat{\mathbf{U}}(\tilde{x}, y)$ is the solution of (2.27). We define the position of the viscous shock layer as the smallest x-value such that $\mathbf{u}^{\varepsilon(1)}(x,y) = (\mathbf{u}^{-(1)}(y)+\mathbf{u}^{+(1)}(y))/2$ and denote this x-value by $x^\varepsilon(y)$. That is, the viscous shock position is described by $(x^\varepsilon(y), y)$. Hence we have

$$x^\varepsilon(0) = 0. \tag{2.10}$$

Remark. We consider the same boundary condition for $\mathbf{u}^\varepsilon$ as for $\mathbf{u}$. When (2.1) is solved numerically, the boundary conditions must be augmented by $m-1$ numerical boundary conditions at $x = a$. Correspondingly, additional boundary conditions that model the numerical boundary conditions are needed for the parabolic PDE (2.7). Numerical boundary conditions can introduce boundary layers in the solution. We consider numerical boundary conditions where such effects are $\mathcal{O}(h^2)$ or smaller. Extrapolation of outgoing characteristic variables is a commonly used numerical boundary condition, but other numerical boundary conditions are also possible, and would not change the result of the analysis.

2.3. Exploring the modified problem using asymptotic expansions. To explore the behavior of the model stated in Assumption 2, we will use the technique of matched asymptotic expansions. In [8], an introduction on how to use asymptotics and matching for internal layers is given. For a comprehensive description of asymptotic techniques, see, e.g., [15] or [18]. In this subsection, we will follow the methods outlined in [8] and use asymptotic expansions and matching for constructing an alternative formulation of the model stated in Assumption 2.

Assumption 3. We assume that $\mathbf{u}^\varepsilon$ can be approximated by truncations of a formal power series

$$\mathbf{u}^\varepsilon(x,y) \sim \sum_{i=0}^{\infty} \varepsilon^i \mathbf{u}_i(x,y) \tag{2.11}$$

in regions away from the inviscid shock interface at $x = 0$. The "outer" functions $\mathbf{u}_i$ may be discontinuous at $x = 0$ but are uniformly smooth up to $x = 0$. We assume that $\mathbf{u}_0 = \mathbf{u}$, i.e., the leading order term of the outer expansion is equal to the solution of the corresponding inviscid problem stated in Assumption 1.

Near $x = 0$, $\mathbf{u}^\varepsilon$ can be approximated by truncations of another formal power series:

$$\mathbf{u}^\varepsilon(x, y) \sim \sum_{i=0}^{\infty} \varepsilon^i \mathbf{U}_i(\tilde{x}, y). \tag{2.12}$$

The "inner" representation of $\mathbf{u}^\varepsilon$ is expressed using the variables $(\tilde{x}, y)$, where

$$\tilde{x} = \frac{x}{\varepsilon}.$$

Also, the position of the viscid shock interface, $x^\varepsilon(y)$, can be expanded in ε:

$$x^\varepsilon(y) \sim 0 + \varepsilon \tilde{x}_1 + \varepsilon^2 \tilde{x}_2 + \cdots .$$

The two expansions (2.11) and (2.12) can be matched in a region of overlap. The more the terms included in the truncated series, the better the approximation. Also, the region of overlap depends on how many terms are included in the truncated series; if more terms are included, the region of overlap decreases. Let $\Sigma_\delta = \{(x, y) : |x| \leq \delta\}$. Let $\mathcal{D}_\delta$ be the complement of Σ_δ in Ω. We denote truncations of series (2.11) including terms up to order ε^N by $\mathbf{u}^{\varepsilon(N)}_{\text{outer}}$, i.e.,

$$\mathbf{u}^{\varepsilon(N)}_{\text{outer}} = \sum_{i=0}^{N} \varepsilon^i \mathbf{u}_i(x, y),$$

and use the corresponding notation for truncations of series (2.12). We assume that there exists a function $\delta(N, \varepsilon)$, with $\delta \to 0$ as $\varepsilon \to 0$, such that

$$|\mathbf{u}^\varepsilon - \mathbf{u}^{\varepsilon(N)}_{\text{outer}}| = \mathcal{O}(\varepsilon^{N+1}) \quad \text{as } \varepsilon \to 0 \tag{2.13}$$

uniformly for $(x, y) \in \mathcal{D}_{\delta(N,\varepsilon)}$. This is to be true for all N up to some integer N_0 that depends on the context. We also assume that there is a function $\tilde{K}(\varepsilon, N)$ such that $\tilde{K} \to \infty$ as $\varepsilon \to 0$ so that

$$|\mathbf{u}^\varepsilon - \mathbf{u}^{\varepsilon(N)}_{\text{inner}}| = \mathcal{O}(\varepsilon^{N+1}) \quad \text{as } \varepsilon \to 0$$

uniformly for $|\tilde{x}| < \tilde{K}(\varepsilon, N)$. Again, this is to be true for all N up to N_0.

The asymptotic expansions above can be viewed as an alternative formulation of the model of the numerical solution, as described in Assumption 2. Hence, instead of analyzing the differential equation (2.7), we can analyze the different terms in the truncated expansions. We model a numerical method which gives a formally second order accurate approximation of the solution of (2.1) away from the shock region. We claim that the solution will be second order accurate upstream of the shock, but only first order accurate downstream, i.e., we must show that $\mathbf{u}_1 = 0$ is upstream and $\mathbf{u}_1 \neq 0$ is downstream. Hence, for our purposes, it is sufficient to truncate the asymptotic expansions after the first two terms and then analyze the truncated expansions with the purpose to show $\mathbf{u}_1 \neq 0$ downstream. Therefore, we choose $N_0 = 1$.

To obtain equations for the terms in the outer and inner expansions, we substitute the expansions into (2.7), Taylor expand and collect terms multiplying the same power of ε. The equation for $\mathbf{U}_0$ is

$$(\phi(\tilde{x} - \tilde{x}_0)\mathbf{U}_{0\tilde{x}})_{\tilde{x}} - \mathbf{f}(\mathbf{U}_0)_{\tilde{x}} = 0, \tag{2.14}$$

where we have used that

$$\frac{\partial}{\partial x} = \frac{1}{\varepsilon}\frac{\partial}{\partial \tilde{x}}.$$

We have also used the fact that in the $(\tilde{x}, y)$-coordinate system, all derivatives of $\mathbf{U}_i$ are $\mathcal{O}(1)$. If $\tilde{x}_0$ denotes the position of the shock layer in $\mathbf{U}_0$, i.e.,

$$\mathbf{U}_0^{(1)}(\tilde{x}_0, y) = (\mathbf{u}^{-(1)}(y) + \mathbf{u}^{+(1)}(y))/2,$$

then Assumption 3 gives $\tilde{x}_0 = \tilde{x}_1 + \mathcal{O}(\varepsilon)$. On each side of $x = 0$ the equation for $\mathbf{u}_1$ is

$$(\mathbf{f}'(\mathbf{u})\mathbf{u}_1)_x + (\mathbf{g}'(\mathbf{u})\mathbf{u}_1)_y = 0. \tag{2.15}$$

We also need boundary conditions for $\mathbf{u}_1$ on the two domains $[-a, 0]$ and $[0, a]$. Using (2.4) and (2.8) in (2.13) we obtain

$$\mathbf{u}_1(-a, y) = 0.$$

No further boundary conditions are needed (or allowed) for $\mathbf{u}_1$ on the upstream side, since all characteristics of (2.15) are going into the domain at $x = -a$ and going out of the domain at $x = 0$. Hence, $\mathbf{u}_1 \equiv 0$ in the upstream region. By using (2.5) and (2.9) in (2.13) we obtain the boundary condition for $\mathbf{u}_1$ at $x = a$,

$$R'(\mathbf{u})\mathbf{u}(a, y) = 0.$$

The boundary condition for $\mathbf{u}_1^+(y)$ remains to be determined. We will do that in the next subsection.

We will now construct the region of overlap on the downstream side. The region of overlap is obtained analogously on the upstream side. Expressed in the stretched variable $\tilde{x}$, the region of overlap, which we will denote by $\mathcal{J}$, is defined by

$$\mathcal{J}: \quad \varepsilon^{-1}\delta(\varepsilon, 1) \le \tilde{x} \le \tilde{K}(\varepsilon, 1).$$

Both the inner and outer expansions are valid in $\mathcal{J}$; hence they must be equal in $\mathcal{J}$, i.e.,

$$\mathbf{U}_0(\tilde{x}, y) + \varepsilon\mathbf{U}_1(\tilde{x}, y) = \mathbf{u}(\varepsilon\tilde{x}, y) + \varepsilon\mathbf{u}_1(\varepsilon\tilde{x}, y) + \mathcal{O}(\varepsilon^2) \quad \text{in } \mathcal{J}.$$

Both $\mathbf{u}$ and $\mathbf{u}_1$ are assumed to be smooth up to $x = 0$. Using the Taylor series expansion around $\tilde{x} = 0$, we arrive at

$$\left(\mathbf{U}_0(\tilde{x}, y) - \mathbf{u}^+(y)\right) + \varepsilon\left(\mathbf{U}_1(y) - \mathbf{u}_x^+(y)\tilde{x} - \mathbf{u}_1^+(y)\right) = \mathcal{O}(\varepsilon^2\tilde{x}^2) \quad \text{in } \mathcal{J}. \tag{2.16}$$

We will need $\varepsilon^{-1}\delta(\varepsilon, 1) \to \infty$ as $\varepsilon \to 0$. Hence, we make the ansatz

$$\delta = \varepsilon^{\alpha}, \quad K = \varepsilon^{-\beta}, \quad 0 < \alpha < 1, \quad 1 - \alpha < \beta,$$

i.e.,

$$\mathcal{J}: \quad \varepsilon^{\alpha-1} < \tilde{x} < \varepsilon^{-\beta}.$$

Both the terms $(\mathbf{U}_0(\tilde{x}, y) - \mathbf{u}^+(y))$ and $\varepsilon\left(\mathbf{U}_1(\tilde{x}, y) - \mathbf{u}_x^+(y)\tilde{x} - \mathbf{u}_1^+(y)\right)$ in (2.16) are individually $\mathcal{O}(\varepsilon^2\tilde{x}^2)$ in $\mathcal{J}$. To see this, we now consider ε as a function of $\tilde{x}$. In

$\mathcal{J}$ we have $\varepsilon = \tilde{x}^{-\gamma}$, where $\gamma \in (1/\beta, 1/(1-\alpha))$. Substituting this on the left-hand side of (2.16), we arrive at

$$\big(\mathbf{U}_0(\tilde{x}, y) - \mathbf{u}^+(\cdot, y)\big) + \tilde{x}^{-\gamma}\left(\mathbf{U}_1(\tilde{x}, y) - \mathbf{u}_x^+(y)\tilde{x} - \mathbf{u}_1^+(y)\right) = \mathcal{O}(\varepsilon^2\tilde{x}^2) \quad \text{in } \mathcal{J}. \tag{2.17}$$

We may then write (2.17) in the form

$$\frac{V_0(\tilde{x}, y) + V_1(\tilde{x}, y)}{\varepsilon^2\tilde{x}^2}, \quad \text{which is bounded in } \mathcal{J}. \tag{2.18}$$

If γ is chosen such that $\tilde{x}$ is strictly within $\mathcal{J}$ for $\varepsilon = \tilde{x}^{-\gamma}$, $\tilde{x}$ will remain inside $\mathcal{J}$, also for $\varepsilon = \tilde{x}^{-\gamma'}$, if $\gamma' = \gamma + \eta$ for $\eta > 0$ sufficiently small, i.e.,

$$\frac{V_0(\tilde{x}, y) + \tilde{x}^{-\eta}V_1(\tilde{x}, y)}{\varepsilon^2\tilde{x}^2} \quad \text{is bounded in } \mathcal{J}. \tag{2.19}$$

Subtracting (2.19) from (2.18) and resubstituting $\tilde{x}^{-\gamma} = \varepsilon$, we obtain

$$\mathbf{U}_1(\tilde{x}, y) - \mathbf{u}_x^+(y)\tilde{x} - \mathbf{u}_1^+(y) = \mathcal{O}(\varepsilon\tilde{x}^2).$$

Then it follows that

$$\mathbf{U}_0(\tilde{x}, y) - \mathbf{u}^+(y) = \mathcal{O}(\varepsilon^2\tilde{x}^2).$$

Hence, in $\mathcal{J}$, we must have $\varepsilon\tilde{x}^2 = \mathbf{o}(1)$. This is true, e.g., if $\beta = 1/3$. We can then choose, e.g., $\alpha = 1/4$. The construction of $\mathcal{J}$ is then complete. The matching conditions are

$$\begin{aligned} &\mathbf{U}_0(\pm\infty, y) = \mathbf{u}^\pm(y), \\ &\mathbf{U}_1(\tilde{x}, y) - \mathbf{u}_x^\pm(y)\tilde{x} - \mathbf{u}_1^\pm(y) = \mathbf{o}(1) \quad \text{as } \tilde{x} \to \pm\infty. \end{aligned} \tag{2.20}$$

Since $\phi(z)$ vanishes for $|z| \geq K_1$, we conclude from (2.14) that $\mathbf{U}_0$ must have reached $\mathbf{u}^\pm(y)$ at $\tilde{x}_0 \pm K_1$. Hence, the matching condition (2.20) can be reformulated as

$$\mathbf{U}_0(\tilde{x}_0 - K_1, y) = \mathbf{u}^-(y), \quad \mathbf{U}_0(\tilde{x}_0 + K_1, y) = \mathbf{u}^+(y).$$

2.4. Downstream boundary condition for the first order term. In this subsection, we will derive the necessary boundary conditions for $\mathbf{u}_1$ at $x = 0^+$.

Let x_m^- denote one point in the upstream matching region and x_m^+ denote one point in the downstream matching region. Integration of the viscous (2.7) from x_m^- to x_m^+ gives

$$[\mathbf{f}(\mathbf{u}^\varepsilon)]_{x_m^-}^{x_m^+} + \int_{x_m^-}^{x_m^+} \mathbf{g}(\mathbf{u}^\varepsilon)_y - c_2\varepsilon^2\mathbf{u}_{yy}^\varepsilon \, dx = \mathcal{O}(\varepsilon^2), \tag{2.21}$$

where we have used that $\phi(x)$ vanishes in the matching regions. Using the outer expansion of $\mathbf{u}^\varepsilon$ we obtain

$$[\mathbf{f}(\mathbf{u}^\varepsilon)]_{x_m^-}^{x_m^+} = [\mathbf{f}(\mathbf{u})]_{x_m^-}^{x_m^+} + \varepsilon[\mathbf{f}'(\mathbf{u})\mathbf{u}_1]_{x_m^-}^{x_m^+} + \mathcal{O}(\varepsilon^2). \tag{2.22}$$

By integrating the inviscid (2.1) over the same interval we find

$$[\mathbf{f}(\mathbf{u})]_{x_m^-}^{x_m^+} = -\int_{x_m^-}^{0^-} \mathbf{g}(\mathbf{u})_y \, dx - \int_{0^+}^{x_m^+} \mathbf{g}(\mathbf{u})_y \, dx. \tag{2.23}$$

Note that $\mathbf{u}$ is discontinuous at $x=0$ and that the Rankine–Hugoniot condition (2.6) applies across the discontinuity. Taking into account that $\mathbf{u}_1 \equiv 0$ upstream of the shock layer and introducing (2.22) and (2.23) into (2.21), and Taylor expanding $\mathbf{u}$ and $\mathbf{u}_1$ around $\tilde{x} = 0$, we arrive at

$$\varepsilon \mathbf{f}'(\mathbf{u}^+(y))\mathbf{u}_1^+(y) + I_1(y) - I_2(y) = \mathcal{O}(\varepsilon^2), \tag{2.24}$$

where I_1 and I_2 are defined by

$$\begin{aligned} I_1 &= \int_{x_m^-}^{0^-} (\mathbf{g}(\mathbf{u}^\varepsilon) - \mathbf{g}(\mathbf{u}))_y \, dx + \int_{0^+}^{x_m^+} (\mathbf{g}(\mathbf{u}^\varepsilon) - \mathbf{g}(\mathbf{u}))_y \, dx, \\ I_2 &= \int_{x_m^-}^{x_m^+} c_2 \varepsilon^2 \mathbf{u}_{yy}^\varepsilon \, dx. \end{aligned} \tag{2.25}$$

First, we consider I_2. Expressed using the inner expansion of $\mathbf{u}^\varepsilon$ and the $(\tilde{x}, y)$-coordinate system, we have

$$I_2 = \varepsilon \int_{\tilde{x}_m^-}^{\tilde{x}_m^+} c_2 \varepsilon^2 (\mathbf{U}_0 + \varepsilon \mathbf{U}_1 + \mathcal{O}(\varepsilon^2))_y \, d\tilde{x}.$$

In the $(\tilde{x}, y)$-coordinate system, all derivatives of $\mathbf{U}_i$ are $\mathcal{O}(1)$. Hence,

$$I_2 = \mathbf{o}(\varepsilon^2), \tag{2.26}$$

where we have used that $\varepsilon \tilde{x} = \mathbf{o}(1)$ in the matching region.

Next, we consider I_1. We use the inner expansion of $\mathbf{u}^\varepsilon$, the Taylor expansion of $\mathbf{u}$ around $x = 0^\pm$ and change coordinates. Then we obtain

$$\begin{aligned} I_1 &= \varepsilon \tilde{I}_1, \\ \tilde{I}_1 &= \int_{\tilde{x}_m^-}^{0} \big(\mathbf{g}(\mathbf{U}_0) - \mathbf{g}(\mathbf{u}^-)\big)_y \, d\tilde{x} + \int_{0}^{\tilde{x}_m^+} \big(\mathbf{g}(\mathbf{U}_0) - \mathbf{g}(\mathbf{u}^+)\big)_y \, d\tilde{x} + \mathbf{o}(1), \end{aligned}$$

where we have used that $\varepsilon \tilde{x}^2 = \mathbf{o}(1)$ in the matching region.

For the further analysis, it is convenient to introduce $\hat{\mathbf{U}}(\tilde{x}, y)$ defined by

$$\begin{aligned} \big(\phi(\tilde{x})\hat{\mathbf{U}}_{\tilde{x}}\big)_{\tilde{x}} - \mathbf{f}(\hat{\mathbf{U}})_{\tilde{x}} &= 0, \\ \hat{\mathbf{U}}(-K_1, y) &= \mathbf{u}^-, \\ \hat{\mathbf{U}}(K_1, y) &= \mathbf{u}^+, \end{aligned} \tag{2.27}$$

i.e., $\mathbf{U}_0(\tilde{x}, y) = \hat{\mathbf{U}}(\tilde{x} - \tilde{x}_0, y)$. Hence, using $\hat{\mathbf{U}}$ we obtain

$$\begin{aligned} \tilde{I}_1 = &\int_{\tilde{x}_m^-}^{0} (\mathbf{g}(\hat{\mathbf{U}}(\tilde{x} - \tilde{x}_0(y)) - \mathbf{g}(\mathbf{u}^-))_y \, d\tilde{x} \\ &+ \int_{0}^{\tilde{x}_m^+} (\mathbf{g}(\hat{\mathbf{U}}(\tilde{x} - \tilde{x}_0(y)) - \mathbf{g}(\mathbf{u}^+))_y \, d\tilde{x} + \mathbf{o}(1). \end{aligned}$$

We now again change the coordinate system, introducing $\hat{x} = \tilde{x} - \tilde{x}_0(y), \hat{y} = y$. The relations between derivatives in the $(\tilde{x}, y)$-coordinate system and the $(\hat{x}, \hat{y})$-coordinate system are

$$\begin{aligned}\frac{\partial}{\partial \tilde{x}} &= \frac{\partial}{\partial \hat{x}},\\ \frac{\partial}{\partial y} &= -\tilde{x}_0'(\hat{y})\frac{\partial}{\partial \hat{x}} + \frac{\partial}{\partial \hat{y}}.\end{aligned}$$

After applying the change of coordinates, we arrive at

$$\tilde{I}_1 = (-\tilde{x}_0(\hat{y})[\mathbf{g}])_{\hat{y}} + \hat{I}_{1\hat{y}} + \mathbf{o}(1),$$

where

$$\hat{I}_1 = \int_{-K_1}^{0} \mathbf{g}(\hat{\mathbf{U}}) - \mathbf{g}(\mathbf{u}^-)\,d\hat{x} + \int_0^{K_1} \mathbf{g}(\hat{\mathbf{U}}) - \mathbf{g}(\mathbf{u}^+)\,d\hat{x}.$$

The leading terms of (2.24) together with condition (2.10) give that the equations for $\mathbf{u}_1^+(y)$ and $\tilde{x}_1(y)$ are

$$\mathbf{f}'(\mathbf{u}^+)\mathbf{u}_1^+ - (\tilde{x}_1[\mathbf{g}])_y = -\hat{I}_{1y}, \tag{2.28}$$

$$\tilde{x}_1(0) = 0. \tag{2.29}$$

We have used that $\tilde{x}_0 = \tilde{x}_1 + \mathcal{O}(\varepsilon)$ and that $\mathbf{u}$ and $\mathbf{u}_1$ are independent of ε. Note that the terms $\tilde{x}_1[\mathbf{g}(\cdot, y)]$ and $\hat{I}_1$ are one-variable functions; hence there is no difference between $\hat{y}$ and y.

(2.28) and (2.29) constitute the boundary condition for $\mathbf{u}_1$ at $x = 0^+$. To make (2.28) and (2.29) easier to understand, we rewrite them using the characteristic variables of $\mathbf{u}_1$. Let w_I be the characteristics of $\mathbf{u}_1$ going into the shock, and let $\mathbf{w}_{II}$ be the characteristic variables going out of the shock. We then have

$$\mathbf{u}^+ = (S_I^+ S_{II}^+)\begin{pmatrix}\mathrm{w}_I\\ \mathbf{w}_{II}\end{pmatrix}.$$

Expressed in characteristic variables, (2.28) can be written as

$$\begin{pmatrix}\mathbf{w}_{II}\\ \tilde{x}_1'\end{pmatrix} = D^{-1}\begin{pmatrix}\Lambda_{II}^+ & 0\\ 0 & -1\end{pmatrix}^{-1}(\tilde{x}_1[\mathbf{g}] - S_I^+\mathrm{w}_I - \hat{I}_{1y}), \tag{2.30}$$

where D is defined in (2.3) and $\Lambda_{II}^+ = \mathrm{diag}(\lambda_2^+, \lambda_3^+, \ldots, \lambda_m^+)$. By solving (2.30) and (2.29) for $\tilde{x}_1$, we can express $\mathbf{w}_{II}$ in w_I and known functions of time. The Laplace transform method (see, e.g., [17]) shows that the equation and boundary conditions for $\mathbf{w}$ constitute a well-posed problem. Well-posedness implies that for any $\hat{I}_{1y}$ there exists a unique solution. The boundary condition for $\mathbf{w}$ at $x = 0^+$ is homogeneous if $\hat{I}_{1y} \equiv 0$, and nonhomogeneous otherwise. Since $\mathbf{w}$ is a transformation of $\mathbf{u}_1$, the same applies for $\mathbf{u}_1$.

2.5. Main result.

We summarize the conclusions from the analysis in the following theorem.

THEOREM 2.1. *If Assumptions* 1–3 *are satisfied, then* $\mathbf{u}_1 \equiv 0$ *on the upstream side and* $\mathbf{u}_1$ *together with* $\tilde{x}_1$ *on the downstream side satisfy the well-posed problem consisting of* (2.15) *with the boundary conditions* (2.28) *and* (2.29) *on the domain*

$$0 < x < a, \quad y > 0.$$

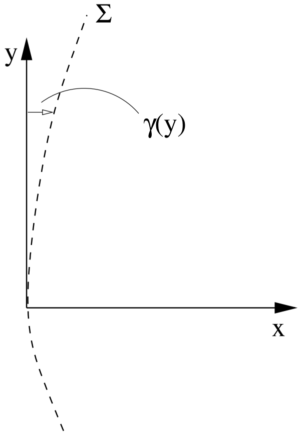


FIG. 2.2. *For a curved shock, we define the shock curve* Σ *by* $\Sigma = \{(x, y) : x = \gamma(y)\}$.

If $\hat{I}_{1y} \equiv 0$*, then* $\mathbf{u}_1 \equiv 0$*, and* $\mathbf{u}^\varepsilon$ *is a second order accurate approximation of* $\mathbf{u}$*. In the general case* $\mathbf{u}_1 \neq 0$*, and* $\mathbf{u}^\varepsilon$ *will be a first order accurate approximation of* $\mathbf{u}$*.*

Remark. In [16] and [23], similar analysis was presented for steady-state solutions of systems with source terms in one space dimension and time-dependent solutions of systems in one space dimension, respectively. In both cases, integral conditions similar to $\hat{I}_{1y} = 0$ are necessary for second order accuracy of the numerical schemes. In the 1D cases, we were able to design a matrix valued viscosity coefficient such that the integral condition is satisfied. Numerical computations verified that the numerical solutions obtained are indeed second order accurate both upstream and downstream of the shock. In the 1D cases, however, the integral condition involved $\mathbf{u}$ and $\hat{U}$, while the integral condition in two dimensions involved $\mathbf{g}(\mathbf{u})$ and $\mathbf{g}(\hat{U})$. In general, $\mathbf{g}$ is a complicated nonlinear function. We have not been able to design a matrix valued viscosity coefficient such that the integral condition $\hat{I}_{1y} = 0$ is satisfied.

2.6. Curved shocks. In this subsection we sketch how the analysis in the previous sections is altered if we consider curved shocks. The shock curve Σ is defined by $\Sigma = \{(x, y) : x = \gamma(y)\}$; see Figure 2.2.

The normal of Σ is $(-1, \gamma')/\sqrt{1 + |\gamma'|^2}$. As long as $|\gamma'|$ is bounded we have from (2.2)

$$-[\mathbf{f}] + \gamma'[\mathbf{g}] = 0.$$

After the coordinate transformation

$$\hat{x} = x - \gamma(y), \quad \hat{y} = y, \tag{2.31}$$

the inviscid steady-state solution

$$\mathbf{f}(\mathbf{u})_x - \mathbf{g}(\mathbf{u})_y = 0$$

becomes

$$(\mathbf{f}(\mathbf{u}) - \gamma'(\hat{y})\mathbf{g}(\mathbf{u}))_{\hat{x}} + \mathbf{g}(\mathbf{u})_{\hat{y}} = 0.$$

In numerical computations the dissipation is usually treated separately in each grid direction. Therefore, instead of (2.7), we now consider the model equation

$$\mathbf{f}(\mathbf{u}^\varepsilon)_x + \mathbf{g}(\mathbf{u}^\varepsilon)_y = \varepsilon\big(\phi_1 \mathbf{u}^\varepsilon_x\big)_x + \varepsilon\big(\phi_2 \mathbf{u}^\varepsilon_y\big)_y + c_2\varepsilon^2\big(\mathbf{u}^\varepsilon_{xx} + \mathbf{u}^\varepsilon_{yy}\big) \tag{2.32}$$

with the appropriate changes of boundary conditions, etc. The functions ϕ_1 and ϕ_2 are smooth functions of one variable modeling the switch mechanism. The model equation (2.32) expressed in the $(\hat{x}, \hat{y})$-variables is

$$\begin{aligned}(2.33)\quad &(\mathbf{f}(\mathbf{u}^\varepsilon) - \gamma'(\hat{y})\mathbf{g}(\mathbf{u}^\varepsilon))_{\hat{x}} + \mathbf{g}(\mathbf{u}^\varepsilon)_{\hat{y}} \\ &\quad = \varepsilon\big(\big(\phi_1 \mathbf{u}^\varepsilon_{\hat{x}}\big)_{\hat{x}} - \gamma'\big(\phi_2\big(-\gamma'\mathbf{u}^\varepsilon_{\hat{x}} + \mathbf{u}^\varepsilon_{\hat{y}}\big)\big)_{\hat{x}} + \big(\phi_2\big(-\gamma'\mathbf{u}^\varepsilon_{\hat{x}} + \mathbf{u}^\varepsilon_{\hat{y}}\big)\big)_{\hat{y}}\big) + \mathcal{O}(\varepsilon^2).\end{aligned}$$

One of the most important assumptions in the analysis in the previous subsections was that the derivatives in the x-direction were $\mathcal{O}(1/\varepsilon)$, whereas the y-derivatives were $\mathcal{O}(1)$. As long as $|\gamma'| \ll 1$, this also applies in this more general setting. After introducing

$$\hat{\mathbf{f}}(\mathbf{u}) = \mathbf{f}(\mathbf{u}) - \gamma'(\hat{y})\mathbf{g}(\mathbf{u})$$

the same analysis as before can be performed. Consequently, the downstream boundary condition for the first order term is of the same form as (2.28) and (2.29).

In the case with a straight shock, the first order downstream error was driven only by the variation along the shock of the states in the inviscid solution. From (2.27), however, we see that in the case of a curved shock, the first order downstream error is also driven by the curvature of the shock.

For a general shock, the analysis presented in this section can be applied after a rotation. There is no essential difference in the analysis, but the calculations become more tedious. The conclusion is still that the solution will be only first order accurate downstream of a shock interface.

3. Summary.

In this paper we present an analysis that yields a possible explanation to the reduction in order of accuracy, when formally second and higher order shock capturing schemes are used for solutions containing shocks. A detailed analysis of the so-called modified equation for a generic shock capturing scheme is presented for the case of a steady-state solution containing a shock that is straight and aligned with one of the grid directions. We then show how the analysis can be extended to curved shocks. The analysis yields that a reduction to first order accuracy in 2D steady-state solutions in general is due to the variation of the solution in the direction tangential to the shock layer in the vicinity of the shock. Note that in the 1D case the reduction to first order accuracy in a steady-state solution is only possible with a lower order term present in the conservation law. Such a term is not necessary in the 2D case.

Acknowledgment. The authors would like to thank Prof. Gunilla Kreiss for fruitful discussions and helpful suggestions throughout the work on this paper.

REFERENCES

[1] M. Arora and P. L. Roe, *On postshock oscillations due to capturing schemes in unsteady flows*, J. Comput. Phys., 130 (1997), pp. 25–40.

[2] M. H. Carpenter and J. H. Casper, *Computational considerations for the simulation of discontinuous flows*, in Barriers and Challenges in Computational Fluid Dynamics, V. Venkatakrishnan, ed., Kluwer Academic Publishers, Dordrecht, the Netherlands, 1997.

[3] M. H. Carpenter and J. H. Casper, *Accuracy of shock capturing in two spatial dimensions*, AIAA J., 37 (1999), pp. 1072–1079.

[4] J. H. Casper and M. H. Carpenter, *Computational considerations for the simulation of shock-induced sound*, SIAM J. Sci. Comput., 19 (1998), pp. 813–828.

[5] G. Efraimsson and G. Kreiss, *A remark on numerical errors downstream of slightly viscous shocks*, SIAM J. Numer. Anal., 36 (1999), pp. 853–863.

[6] P. Eliasson, *Dissipation Mechanisms and Multigrid Solutions in a Multiblock Solver for Compressible Flow*, TRITA-NA-R9314, Royal Institute of Technology, Stockholm, Sweden, 1993.

[7] B. Engquist and B. Sjögreen, *The convergence rate of finite difference schemes in the presence of shocks*, SIAM J. Numer. Anal., 35 (1998), pp. 2464–2485.

[8] P. Fife, *Dynamics of Internal Layers and Diffusive Interfaces*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 53, SIAM, Philadelphia, 1988.

[9] J. Goodman and A. Majda, *The validity of the modified equation for nonlinear shock waves*, J. Comput. Phys., 58 (1985), pp. 336–348.

[10] D. F. Griffiths and J. M. Sanz-Serna, *On the scope of the method of modified equations*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 994–1008.

[11] C. Hirsch, *Numerical Computation of Internal and External Flows*, Vol. I, Wiley, New York, 1988.

[12] A. Jameson, W. Schmidt, and E. Turkel, *Numerical Solutions of the Euler Equations by Finite Volume Methods Using Runge–Kutta Time-Stepping Schemes*, AIAA Paper 81–1259, 1981.

[13] G.-S. Jiang and S.-H. Yu, *Discrete shocks for finite difference approximations to scalar conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 749–772.

[14] S. Karni and S. Čanić, *Computations of slowly moving shocks*, J Comput. Phys., 136 (1997), pp. 132–139.

[15] J. Kevorkian and J. D. Cole, *Perturbation Methods in Applied Mathematics*, Springer-Verlag, Berlin, 1981.

[16] G. Kreiss, G. Efraimsson, and J. Nordström, *Elimination of first order errors in shock calculations*, SIAM J. Numer. Anal., 38 (2001), pp. 1986–1998.

[17] H.-O. Kreiss and J. Lorenz, *Initial-Boundary Value Problems and the Navier–Stokes Equations*, Academic Press, New York, 1989.

[18] P. A. Lagerstrom, *Matched Asymptotic Expansions*, Springer-Verlag, Berlin, 1988.

[19] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.

[20] T.-P. Liu and S.-H. Yu, *Continuum shock profiles for discrete conservation laws* I*: Construction*, Comm. Pure Appl. Math., 52 (1999), pp. 85–127.

[21] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer-Verlag, Berlin, 1984.

[22] A. Majda and J. Ralston, *Discrete shock profiles for systems of conservation laws*, Comm. Pure Appl. Math., 32 (1979), pp. 445–482.

[23] M. Siklosi and G. Kreiss, *Elimination of first order errors in time dependent shock calculations*, SIAM J. Numer. Anal., 41 (2003), pp. 2131–2148.

[24] F. R. Villatoro and J. I. Ramos, *On the method of modified equations.* I*: Asymptotic analysis of the Euler forward difference method*, Appl. Math. Comput., 103 (1999), pp. 111–139.

[25] K. Zumbrun and D. Serre, *Viscous and inviscid stability of multidimensional planar shock fronts*, Indiana Univ. Math. J., 44 (1999), pp. 937–992.

SIAM J. Numer. Anal.
Vol. 43, No. 2, pp. 686–706

MULTILEVEL PRECONDITIONERS FOR NON–SELF-ADJOINT OR INDEFINITE ORTHOGONAL SPLINE COLLOCATION PROBLEMS*

RAKHIM AITBAYEV†

Abstract. Efficient numerical algorithms are developed and analyzed that implement symmetric multilevel preconditioners for the solution of an orthogonal spline collocation (OSC) discretization of a Dirichlet boundary value problem with a non–self-adjoint or an indefinite operator. The OSC solution is sought in the Hermite space of piecewise bicubic polynomials. It is proved that the proposed additive and multiplicative preconditioners are uniformly spectrally equivalent to the operator of the normal OSC equation. The preconditioners are used with the preconditioned conjugate gradient method, and numerical results are presented that demonstrate their efficiency.

Key words. orthogonal spline collocation, multilevel methods, preconditioner, non–self-adjoint or indefinite operator, elliptic boundary value problem

AMS subject classifications. 65N35, 65N55, 65F10

DOI. 10.1137/040609884

1. Introduction.

Let Ω be a unit square $(0,1)\times(0,1)$ with boundary $\partial\Omega$, and let $x=(x_1,x_2)$. We consider a Dirichlet boundary value problem (BVP)

$$Lu = f \text{ in } \Omega \quad \text{and} \quad u=0 \text{ on } \partial\Omega, \tag{1.1}$$

where we let

$$Lu\,(x) = \sum_{i,j=1}^{2} a_{ij}(x)u_{x_ix_j}(x) + \sum_{i=1}^{2} b_i(x)u_{x_i}(x) + c(x)u(x). \tag{1.2}$$

With respect to BVP (1.1), we make the following assumptions. The functions $\{a_{ij}\}_{i,j=1}^2$, $\{b_i\}_{i=1}^2$, c, and f are sufficiently smooth, and $a_{12}=a_{21}$. The differential operator L satisfies the uniform ellipticity condition; that is, there is $\nu>0$ such that

$$\sum_{i,j=1}^{2} a_{ij}(x)\,\eta_i\,\eta_j \ge \nu(\eta_1^2+\eta_2^2), \quad x\in\Omega,\ (\eta_1,\eta_2)\in R^2. \tag{1.3}$$

For any $f\in L^2(\Omega)$, BVP (1.1) has a unique solution $u(x)$ in

$$\tilde{H}_0^2(\Omega) = \{v\in H^2(\Omega):\ v=0 \text{ on } \partial\Omega\}, \tag{1.4}$$

where $L^2(\Omega)$ and $H^2(\Omega)$ are the Sobolev spaces.

We approximate BVP (1.1) by an orthogonal spline collocation (OSC) scheme, in which the discrete solution is sought in the Hermite space of piecewise bicubic polynomials, and it satisfies the differential equation exactly at the special set of collocation points. The primary advantages of the OSC method are as follows: it has low computational cost of forming a linear system of algebraic equations; it has relatively easy

[*] Received by the editors June 11, 2004; accepted for publication (in revised form) December 13, 2004; published electronically August 31, 2005.

http://www.siam.org/journals/sinum/43-2/60988.html

[†] Department of Mathematics, New Mexico Institute of Mining and Technology, Socorro, NM 87801 (aitbayev@nmt.edu).

application of higher order finite elements; the OSC solution possesses optimal order error estimates [7]; and the solution exhibits the so-called superconvergence property [8, 15]. The matrix of a linear system resulting from the OSC discretization is sparse and can be expressed as a sum of tensor products of so-called almost block diagonal matrices [2].

The solution of the OSC equations by a banded Gaussian elimination requires $\mathrm{O}(N^2)$ arithmetic operations, where N is the number of unknowns [22, 23, 34]. If the differential operator is separable and the partition is uniform in one direction, then the OSC problem can be solved by a fast direct algorithm with the cost $\mathrm{O}(N \log N)$ [10].

Classical iterative methods, such as Jacobi, Gauss–Seidel, and SOR, for the OSC solution of Poisson's equation on a uniform partition were studied in [21, 26, 37]. ADI methods for solving OSC problems with separable operators were investigated in [5, 14, 18].

Modern techniques are underdeveloped for the solution of OSC equations in comparison with finite element Galerkin or finite difference methods. Only a few optimal cost algorithms have been proposed to solve the OSC discretization of self-adjoint positive definite BVPs. In [19], the multigrid method was applied to the OSC problem and compared with a multigrid finite difference method. The author concluded that the proposed multigrid OSC method is less efficient than a multigrid finite difference method. A domain decomposition–based fast solver for the OSC discretization of the Dirichlet problem for Poisson's equation was developed in [6] requiring $\mathrm{O}(N \log\log N)$ arithmetic operations. In [13], multigrid methods were developed and analyzed for quadratic spline collocation equations. Numerical results were presented indicating that a multigrid iteration is an efficient solver for the quadratic spline collocation equations.

It is well known that the primary issue in the efficient application of an iterative algorithm for solving a BVP is the construction of a good preconditioner. In [24], the authors studied preconditioning of a non–self-adjoint or an indefinite OSC operator by a finite element operator and investigated H^1 condition numbers and the distribution of singular values of the preconditioned matrices. Additive and multiplicative multilevel preconditioners were proposed in [9] for the iterative solution of the OSC discretization of a self-adjoint positive definite Dirichlet BVP. It was proved that the preconditioners are uniformly spectrally equivalent to the OSC operator corresponding to a BVP with the Laplacian and require $\mathrm{O}(N)$ arithmetic operations. An efficient two-level domain decomposition–type "edge" preconditioner was proposed in [29] that requires $\mathrm{O}(N)$ arithmetic operations. The preconditioner is applied with the GMRES method, and the number of iterations is independent of the partition stepsize h.

Numerical techniques developed for self-adjoint positive definite BVPs are usually inefficient or even fail when applied to non–self-adjoint or indefinite BVPs, and hence, special, more sophisticated methods are required to obtain the solution [4, 11, 12, 27, 28]. A fast direct preconditioning algorithm for the solution of the normal OSC equation approximating non–self-adjoint or indefinite BVPs was proposed in [1].

In this work, we develop additive and multiplicative multilevel preconditioners for the computation of the solution of the normal OSC equation. Results and algorithms presented in this paper are closely related to those in [1, 7, 9, 31, 32, 39, 40]. To prove uniform spectral equivalence of our preconditioners, we use the approach described in [31] and [32] based on the equivalence of a norm of a certain Besov space with the Sobolev H^2-norm. We note that the approach used in [39] and [40] requires higher regularity of the solution of BVP (1.1). Our main conclusion in this work is that the general framework of multilevel methods can be applied to the OSC discretization of BVPs to construct efficient preconditioners. Rather general non–self-adjoint or indefinite OSC problems can be preconditioned quite well by the proposed multilevel OSC preconditioners.

The outline of this article is as follows. We introduce our notation and define the OSC problem in section 2. In section 3, we present auxiliary facts that are used to prove main results of this work. In section 4, we define additive and multiplicative OSC preconditioners and prove that they are uniformly spectrally equivalent to the operator of the normal OSC equation. In section 5, we introduce the matrix-vector form of the OSC problem in the standard Hermite finite element basis and obtain recurrence relations for the computation of the OSC approximations and other quantities required by the multilevel algorithms. In section 6, we describe implementations of the additive and the multiplicative preconditioners, and in section 7, we present numerical results of the application of the preconditioners with the preconditioned conjugate gradient (PCG) method to solve test problems.

2. OSC problem.

In this section, we introduce our notation and define the OSC problem. Throughout this paper, C, C_1, and $C_2 \geq C_1$ denote generic positive constants independent of the partition stepsize, the number of partition levels, and other variables in the expressions where the constants appear. By $\|\cdot\|_{L^2(\Omega)}$ and $\|\cdot\|_{H^2(\Omega)}$ we denote the standard Sobolev norms.

Construction of nested spaces. We set $\pi_0 = \Omega$ and, for integer $K > 0$, we construct a sequence of partitions $\{\pi_k\}_{k=0}^{K}$ by subdividing each rectangular element of partition π_{k-1} into four congruent rectangular elements of partition π_k. Let $h_k = 2^{-k}$ denote the stepsize of partition π_k. In what follows, if not stated otherwise, the index variable k takes all values in $\{0, 1, \ldots, K\}$. We note that there are $K + 1$ partition levels, and integer K is an important parameter in our analysis.

Let V_k be the vector space of Hermite piecewise bicubic polynomials that vanish on $\partial\Omega$, which has the dimension $N_k = 4^{k+1}$ (see Chapter 3 in [35]). The sequence of vector spaces $\{V_k\}$ is nested as follows:

$$V_0 \subset V_1 \subset \cdots \subset V_K \subset \tilde{H}_0^2(\Omega). \tag{2.1}$$

We denote $h = h_K$, $N = N_K$, $\pi_h = \pi_K$, and $V_h = V_K$.

Original OSC equation. Let $\mathcal{G}_h$ be the set of nodes of the two-dimensional composite Gaussian quadrature on partition π_h with 4 nodes in each element of π_h. In the OSC discretization of BVP (1.1), we seek $u_h \in V_h$ that satisfies the OSC equations

$$Lu_h(\xi) = f(\xi), \quad \xi \in \mathcal{G}_h, \tag{2.2}$$

where the differential operator L is defined in (1.2). Existence and uniqueness of a solution and a convergence analysis of problem (2.2) are given in [7].

The OSC problem (2.2) can be written as the operator equation

$$L_h u_h = f_h, \tag{2.3}$$

where the OSC operator L_h from V_h into V_h and the vector $f_h \in V_h$ are defined by

$$\begin{aligned} (L_h v)(\xi) &= Lv(\xi), \quad \text{for any } \xi \in \mathcal{G}_h \text{ and for any } v \in V_h, \\ f_h(\xi) &= f(\xi), \quad \text{for any } \xi \in \mathcal{G}_h. \end{aligned} \tag{2.4}$$

Both L_h and f_h are well defined since a function in V_h is uniquely determined by its values at $\mathcal{G}_h$ (see Lemma 5.1 in [33]). We call (2.3) the original OSC equation.

Normal OSC equation. The vector space V_h is a Hilbert space with the inner product

$$(v,w)_h = \frac{h^2}{4} \sum_{\xi \in \mathcal{G}_h} v(\xi) w(\xi), \tag{2.5}$$

which corresponds to the composite Gaussian quadrature on π_h. Let L_h^* be the adjoint to L_h with respect to the inner product $(\cdot,\cdot)_h$. Applying L_h^* on (2.3), we obtain the normal OSC equation

$$L_h^* L_h u_h = L_h^* f_h. \tag{2.6}$$

We introduce a bilinear form

$$a_h(w,v) = (L_h^* L_h w, v)_h, \quad w, v \in V_h, \tag{2.7}$$

and consider the following variational form of problem (2.6): find $u_h \in V_h$ that satisfies

$$a_h(u_h, v) = (L_h^* f_h, v)_h \ \text{ for all } \ v \in V_h. \tag{2.8}$$

The problems (2.8) and (2.2) are equivalent; hence, problem (2.8) has a unique solution. In this work, we develop and analyze multilevel preconditioners for the iterative solution of (2.8).

Space decompositions. In what follows, we denote $\sum_k$ and $\sum_{k,i}$ for $\sum_{k=0}^K$ and $\sum_{k=0}^K \sum_{i=1}^{N_k}$, respectively, where N_k is the dimension of V_k.

Let $\{\psi_i^k\}_{i=1}^{N_k}$ be a finite element basis of V_k that satisfies

$$C_1 h_k^{1-|\alpha|} \le \|\partial^\alpha \psi_i^k\|_{L^2(\Omega)} \le C_2 h_k^{1-|\alpha|}, \quad |\alpha| \le 2, \tag{2.9}$$

where $\alpha = (\alpha_1, \alpha_2)$ is a multi-index (see Theorem 5.7 in [3]). Let

$$V_{ki} = \operatorname{span}(\psi_i^k), \quad 1 \le i \le N_k, \tag{2.10}$$

be one-dimensional subspaces of V_k. Based on (2.1), we consider the following two space decompositions:

$$\sum_k V_k = V_h \quad \text{and} \quad \sum_{k,i} V_{ki} = V_h.$$

For $v \in V_h$, let

$$\mathcal{V}_1(v) = \left\{ \{v_k\} : \sum_k v_k = v,\ v_k \in V_k,\ 0 \le k \le K \right\},$$

$$\mathcal{V}_2(v) = \left\{ \{v_{ki}\} : \sum_{ki} v_{ki} = v,\ v_{ki} \in V_{ki},\ 1 \le i \le N_k,\ 0 \le k \le K \right\}.$$

We call an element in $\mathcal{V}_1(v)$ and in $\mathcal{V}_2(v)$ a representation of v. The sets $\mathcal{V}_1(v)$ and $\mathcal{V}_2(v)$ will be used to define auxiliary equivalent norms on V_h.

3. Auxiliary results.

In this section we present auxiliary facts that are used to prove main results of this work. The following inequalities are proved in Theorem 3.1 in [1].

Lemma 3.1. *For h sufficiently small,*

$$C_1 \|v\|^2_{H^2(\Omega)} \le a_h(v,v) \le C_2 \|v\|^2_{H^2(\Omega)}, \quad v \in V_h. \tag{3.1}$$

Remark. In what follows, by sufficiently small h, we mean values of h for which the statement of Lemma 3.1 holds.

It follows from (3.1) and (2.7) that, for h sufficiently small, the bilinear form $a_h(\cdot,\cdot)$ is an inner product on V_h. The following is the estimate (8.24) in Chapter 3 of [25].

LEMMA 3.2.

$$\|v\|_{H^2(\Omega)} \leq C\|\Delta v\|_{L^2(\Omega)}, \quad v \in \tilde{H}_0^2(\Omega). \tag{3.2}$$

Inequalities (3.1), (3.2), and

$$\|\Delta v\|^2_{L^2(\Omega)} \leq C \int_\Omega (v^2_{x_1x_1} + v^2_{x_2x_2})dx, \quad v \in \tilde{H}_0^2(\Omega), \tag{3.3}$$

imply that, for h sufficiently small,

$$C_1\|\Delta v\|^2_{L^2(\Omega)} \leq a_h(v,v) \leq C_2\|\Delta v\|^2_{L^2(\Omega)}, \quad v \in V_h. \tag{3.4}$$

LEMMA 3.3. *Let $v \in V_k$, and let*

$$v = \sum_{i=1}^{N_k} c_j\psi_i^k \ \ \textit{and} \ \ \vec{c}_v = (c_1,\ldots,c_{N_k})^t. \tag{3.5}$$

Then,

$$C_1h_k|\vec{c}_v| \leq \|v\|_{L^2(\Omega)} \leq C_2h_k|\vec{c}_v|, \quad v \in V_k, \tag{3.6}$$

where $|\cdot|$ is the 2-norm on R^{N_k}.

Proof. Using (3.5), we obtain

$$\|v\|^2_{L^2(\Omega)} = \int_\Omega \sum_{i=1}^{N_k} c_i\psi_i^k \sum_{j=1}^{N_k} c_j\psi_j^k\,dx = \sum_{i,j=1}^{N_k} c_ic_j \int_\Omega \psi_i^k\psi_j^k\,dx.$$

The inequalities in (3.6) follow from the last identity and the fact that the eigenvalues of the mass matrix corresponding to the finite element basis $\{\psi_i^k\}_{i=1}^{N_k}$ belong to the interval $[C_1h_k^2, C_2h_k^2]$ (see (5.103) in [3]). □

Let

$$\|v\|_{*,h} = \left(\inf_{\mathcal{V}_1(v)} \sum_{k=0}^{K} h_k^{-4}\|v_k\|^2_{L^2(\Omega)}\right)^{1/2}, \quad v \in V_h, \tag{3.7}$$

where $\inf_{\mathcal{V}_1(v)}$ denotes the infimum with respect to all representations $\{v_k\}$ in $\mathcal{V}_1(v)$. The following important statement is similar to Corollary 2.1 in [32] and Lemma 2 in [31].

LEMMA 3.4. *The norms $\|\cdot\|_{H^2(\Omega)}$ and $\|\cdot\|_{*,h}$ are uniformly equivalent on V_h; that is,*

$$C_1\|v\|_{H^2(\Omega)} \leq \|v\|_{*,h} \leq C_2\|v\|_{H^2(\Omega)}, \quad v \in V_h. \tag{3.8}$$

Proof. First, we prove the second inequality in (3.8). It is known that the Besov space $B^2_{2,2}(\Omega)$ coincides, up to equivalent norms, with the Sobolev space $H^2(\Omega)$ (see part (b) of Theorem 4.6.1 and (3) of section 4.2.1 in [38]). Since $\tilde{H}_0^2(\Omega)$ is a closed subspace of $H^2(\Omega)$, it is a closed subspace of $B^2_{2,2}(\Omega)$. We note that functions in $\tilde{H}_0^2(\Omega)$ are continuous on $\overline{\Omega}$, and therefore, for any $v \in \tilde{H}_0^2(\Omega)$, $v(x) = 0$ for all $x \in \partial\Omega$; in particular, the trace of $v \in \tilde{H}_0^2(\Omega)$ is continuous. In a natural way, we extend the definition of $\{V_k\}$ for $k > K$. It follows from Theorem 5.1 in [16] that

$$\|v\|_{A^2_{2,2}} = \left(\|v\|^2_{L^2(\Omega)} + \sum_{k=0}^{\infty} 2^{4k} \left(\inf_{z \in V_k} \|v - z\|_{L^2(\Omega)} \right)^2 \right)^{1/2}$$

is a norm on $\tilde{H}_0^2(\Omega)$ equivalent to the standard norm in $B^2_{2,2}(\Omega)$ (see [30] for a definition of the approximation space $A^2_{2,2}$). Therefore, using the equivalence of norms $\|\cdot\|_{H^2(\Omega)}$ and $\|\cdot\|_{A^2_{2,2}}$ on $\tilde{H}_0^2(\Omega)$ and the fact $V_h \subset \tilde{H}_0^2(\Omega)$, we get

$$\|v\|_{A^2_{2,2}} \le C\|v\|_{H^2(\Omega)}, \quad v \in V_h. \tag{3.9}$$

Let us prove

$$\|v\|_{*,h} \le C\|v\|_{A^2_{2,2}}, \quad v \in V_h. \tag{3.10}$$

We note that

$$\|v\|_{A^2_{2,2}} = \left(\|v\|^2_{L^2(\Omega)} + \sum_{k=0}^{K-1} 2^{4k} \left(\inf_{z \in V_k} \|v - z\|_{L^2(\Omega)} \right)^2 \right)^{1/2}, \quad v \in V_h, \tag{3.11}$$

since

$$\inf_{z \in V_k} \|v - z\|_{L^2(\Omega)} = 0, \;\; k \ge K, \;\; v \in V_h.$$

Take any $v \in V_h$ and set $v_0 = z_0$ and $v_k = z_k - z_{k-1}$ for $1 \le k \le K$, where z_k is the orthogonal L^2-projection of v into V_k. Note that $v = z_K = \sum_{k=0}^K v_k$. Using $\|v_0\|_{L^2(\Omega)} \le \|v\|_{L^2(\Omega)}$, the triangle inequality, the definition of $\{z_k\}$, and (3.11), we obtain

$$\begin{aligned} \sum_{k=0}^{K} h_k^{-4} \|v_k\|^2_{L^2(\Omega)} &= \|v_0\|^2_{L^2(\Omega)} + \sum_{k=1}^{K} h_k^{-4} \|z_k - v + v - z_{k-1}\|^2_{L^2(\Omega)} \\ &\le \|v\|^2_{L^2(\Omega)} + 4 \sum_{k=0}^{K-1} h_k^{-4} \|v - z_k\|^2_{L^2(\Omega)} \\ &= \|v\|^2_{L^2(\Omega)} + 4 \sum_{k=0}^{K-1} 2^{4k} \left(\inf_{z \in V_k} \|v - z\|_{L^2(\Omega)} \right)^2 \le 4\|v\|^2_{A^2_{2,2}}. \end{aligned}$$

Taking the infimum over $\mathcal{V}_1(v)$, we get (3.10). Inequalities (3.10) and (3.9) imply the second inequality in (3.8).

Let us prove the first inequality in (3.8). Take any $v \in V_h$ and let $\{v_k\} \in \mathcal{V}_1(v)$. Using the strengthened Cauchy–Schwarz inequality,

$$\left| \int_\Omega \Delta v_k \, \Delta v_l dx \right| \le C 2^{-|k-l|/2} (h_k h_l)^{-2} \|v_k\|_{L^2(\Omega)} \|v_l\|_{L^2(\Omega)}, \quad v_k \in V_k, \;\; v_l \in V_l,$$

(see Lemma 5.1 in [40]) and the fact that the spectral radius of matrix $B = (b_{kl})$ with the entries

$$b_{kl} = 2^{-|k-l|/2}, \quad 0 \le k, l \le K,$$

is bounded by the maximum norm $\|B\|_\infty \le 3 + 2\sqrt{2}$, we get

$$\begin{aligned}\|\Delta v\|^2_{L^2(\Omega)} &= \int_\Omega (\Delta v)^2 dx = \int_\Omega \left(\sum_{k=0}^K \Delta v_k\right)\left(\sum_{l=0}^K \Delta v_l\right) dx = \sum_{k,\, l=0}^K \int_\Omega \Delta v_k \, \Delta v_l dx \\ &\le C \sum_{k,\, l=0}^K 2^{-|k-l|/2}(h_k h_l)^{-2} \|v_k\|_{L^2(\Omega)} \|v_l\|_{L^2(\Omega)} \\ &\le C(3 + 2\sqrt{2}) \sum_k h_k^{-4} \|v_k\|^2_{L^2(\Omega)}.\end{aligned}$$

From the last estimate, using (3.2) and taking the infimum over $\mathcal{V}_1(v)$, we obtain the first inequality in (3.8).

To finish the proof of the lemma, we establish that $\|\cdot\|_{*,h}$ is a norm on V_h. It follows from the inequalities in (3.8) that $\|v\|_{*,h} \ge 0$ for any $v \in V_h$, and $\|v\|_{*,h} = 0$ if and only if $v = 0$. Let us show that $\|cv\|_{*,h} = |c| \, \|v\|_{*,h}$ for any $v \in V_h$ and $c \in R$. Since the case $c = 0$ is trivial, assume $c \ne 0$. Using the fact that the infimum over $\mathcal{V}_1(cv)$ equals the infimum over $\mathcal{V}_1(v)$, we obtain

$$\|cv\|^2_{*,h} = \inf_{\{w_k\} \in \mathcal{V}_1(cv)} \sum_{k=0}^K h_k^{-4} \|w_k\|^2_{L^2(\Omega)} = \inf_{\{v_k\} \in \mathcal{V}_1(v)} \sum_{k=0}^K h_k^{-4} \|cv_k\|^2_{L^2(\Omega)} = c^2 \|v\|^2_{*,h},$$

which implies the required identity.

To prove the triangle inequality for $\|\cdot\|_{h,*}$, using the triangle inequality for the L^2-norm and the Minkowski inequality, we obtain, for any v and w in V_h,

$$\begin{aligned}\|v + w\|^2_{*,h} &= \inf_{\{z_k\} \in \mathcal{V}_1(v+w)} \sum_{k=0}^K h_k^{-4} \|z_k\|^2_{L^2(\Omega)} \le \inf_{\mathcal{V}_1(v)} \inf_{\mathcal{V}_1(w)} \sum_{k=0}^K h_k^{-4} \|v_k + w_k\|^2_{L^2(\Omega)} \\ &\le \inf_{\mathcal{V}_1(v)} \inf_{\mathcal{V}_1(w)} \sum_{k=0}^K h_k^{-4} (\|v_k\|_{L^2(\Omega)} + \|w_k\|_{L^2(\Omega)})^2 \le (\|v\|_{*,h} + \|w\|_{*,h})^2.\end{aligned}$$

Thus, $\|\cdot\|_{*,h}$ is a norm on V_h which is equivalent to the H^2-norm. □

Remark. The result of Lemma 3.4 is analogous to those formulated in [32, Corollary 2.1] and [31, Lemma 2], where relations similar to (3.8) were proved first for Sobolev spaces with equivalent norms involving infinite number series, and then the versions for finite-dimensional subspaces were obtained. Our proof of Lemma 3.4 is somewhat different since it is based on the representation (3.11).

Let

$$\|v\|_{\Sigma,\Delta} = \left(\inf_{\mathcal{V}_2(v)} \sum_{k,i} \|\Delta v_{ki}\|^2_{L^2(\Omega)} \right)^{1/2}, \quad v \in V_h. \tag{3.12}$$

Lemma 3.5.

$$C_1 \|v\|_{H^2(\Omega)} \le \|v\|_{\Sigma,\Delta} \le C_2 \|v\|_{H^2(\Omega)}, \quad v \in V_h. \tag{3.13}$$

Proof. We call nonnegative quantities $A(h)$ and $B(h)$ uniformly equivalent with respect to h and write $A(h) \approx B(h)$ if

$$C_1 B(h) \leq A(h) \leq C_2 B(h).$$

Our proof consists of establishing the following sequence of equivalence relations:

$$\|v\|^2_{\Sigma,\Delta} \approx \inf_{\mathcal{V}_2(v)} \sum_{k,i} h_k^{-4} \|v_{ki}\|^2_{L^2(\Omega)} \approx \|v\|^2_{*,h} \approx \|v\|^2_{H^2(\Omega)}. \tag{3.14}$$

The last equivalence relation in (3.14) is stated and proved in Lemma 3.4.

Let us prove the first equivalence relation in (3.14). Take any $v \in V_h$ and consider a representation $\{v_{ki}\} \in \mathcal{V}_2(v)$. Using (2.10), (3.3), (3.2), and (2.9) with $|\alpha| = 2$ and $\alpha = (0,0)$, we obtain

$$C_1 \|\Delta v_{ki}\|^2_{L^2(\Omega)} \leq h_k^{-4} \|v_{ki}\|^2_{L^2(\Omega)} \leq C_2 \|\Delta v_{ki}\|^2_{L^2(\Omega)}.$$

Summing these inequalities with respect to k and i and taking the infimum over $\mathcal{V}_2(v)$, we obtain the first equivalence relation in (3.14).

We now prove the second equivalence relation in (3.14):

$$C_1 \inf_{\mathcal{V}_2(v)} \sum_{k,i} h_k^{-4} \|v_{ki}\|^2_{L^2(\Omega)} \leq \|v\|^2_{*,h} \leq C_2 \inf_{\mathcal{V}_2(v)} \sum_{k,i} h_k^{-4} \|v_{ki}\|^2_{L^2(\Omega)}, \quad v \in V_h. \tag{3.15}$$

Take any $v \in V_h$. Using uniqueness of the representation

$$v_k = \sum_{i=1}^{N_k} v_{ki}, \quad v_{ki} \in V_{ki}, \quad 1 \leq i \leq N_k,$$

for $v_k \in V_k$, we define injection mappings $\mathcal{V}_1(v) \to \mathcal{V}_2(v)$ and $\mathcal{V}_2(v) \to \mathcal{V}_1(v)$ by

$$\sum_k v_k = \sum_{ki} v_{ki}.$$

The Schroeder–Bernstein theorem implies that there is a bijection $\mathcal{V}_1(v) \to \mathcal{V}_2(v)$.

Consider any representation $\{v_k\} \in \mathcal{V}_1(v)$ and the representation $\{v_{ki}\} \in \mathcal{V}_2(v)$ given by the bijection $\mathcal{V}_1(v) \to \mathcal{V}_2(v)$. Let c_{ki} be such that $v_{ki} = c_{ki}\psi_i^k$. Using (2.9) with $\alpha = (0,0)$, we have

$$C_1\, c_{ki}^2 h_k^2 \leq \|v_{ki}\|^2_{L^2(\Omega)} \leq C_2\, c_{ki}^2 h_k^2.$$

Summing the last inequalities with respect to i and using (3.6), we obtain

$$C_1 \sum_{i=1}^{N_k} \|v_{ki}\|^2_{L^2(\Omega)} \leq \|v_k\|^2_{L^2(\Omega)} \leq C_2 \sum_{i=1}^{N_k} \|v_{ki}\|^2_{L^2(\Omega)}. \tag{3.16}$$

Multiplying (3.16) by h_k^{-4}, summing for $k = 0, 1, \ldots, K$, taking the infimum over $\mathcal{V}_1(v)$, and using (3.7), we obtain

$$C_1 \inf_{\mathcal{V}_1(v)} \sum_{k,i} h_k^{-4} \|v_{ki}\|^2_{L^2(\Omega)} \leq \|v\|^2_{*,h} \leq C_2 \inf_{\mathcal{V}_1(v)} \sum_{k,i} h_k^{-4} \|v_{ki}\|^2_{L^2(\Omega)}. \tag{3.17}$$

Since there is a bijection $\mathcal{V}_1(v) \to \mathcal{V}_2(v)$, the infimum over $\mathcal{V}_1(v)$ is equal to the infimum over $\mathcal{V}_2(v)$; hence, (3.17) implies (3.15). □

4. Uniform spectral equivalence. In this section, we define additive and multiplicative OSC preconditioners and prove that they are uniformly spectrally equivalent to the OSC operator $L_h^* L_h$.

Additive preconditioner. For $0 \leq k \leq K$ and $1 \leq i \leq N_k$, let T_i^k be a linear operator from V_h into V_{ki} defined as follows: for any $w \in V_h$, $T_i^k w$ satisfies

$$a_h(T_i^k w, v) = a_h(w, v) \quad \text{for all} \quad v \in V_{ki}. \tag{4.1}$$

The following is our main result for the additive preconditioner.

Theorem 4.1. *Assume that h is sufficiently small. Linear operator*

$$T_A = \sum_{k,i} T_i^k \tag{4.2}$$

is self-adjoint positive definite on V_h in the inner product $a_h(\cdot,\cdot)$. Linear operator

$$B_A = L_h^* L_h T_A^{-1} \tag{4.3}$$

is self-adjoint positive definite on V_h in the inner product $(\cdot,\cdot)_h$, and

$$C_1 (B_A v, v)_h \leq (L_h^* L_h v, v)_h \leq C_2 (B_A v, v)_h, \quad v \in V_h. \tag{4.4}$$

Proof. Let

$$\|v\|^2_{\Sigma,a_h} = \inf_{\mathcal{V}_2(v)} \sum_{k,i} a_h(v_{ki}, v_{ki}), \quad v \in V_h. \tag{4.5}$$

First, let us prove the equivalence relation

$$C_1 a_h(v,v) \leq \|v\|^2_{\Sigma,a_h} \leq C_2 a_h(v,v), \quad v \in V_h. \tag{4.6}$$

Using (3.4), (3.12), and (4.5), we obtain inequalities

$$C_1 \|v\|_{\Sigma,\Delta} \leq \|v\|_{\Sigma,a_h} \leq C_2 \|v\|_{\Sigma,\Delta}, \quad v \in V_h,$$

which, by Lemma 3.5, imply

$$C_1 \|v\|_{H^2(\Omega)} \leq \|v\|_{\Sigma,a_h} \leq C_2 \|v\|_{H^2(\Omega)}, \quad v \in V_h.$$

The last inequalities and Lemma 3.1 imply (4.6). We note that the second inequality in (4.6) is one of the key assumptions in the abstract theory of Schwarz methods (see Assumption 1 in section 5.2 of [36]).

Using (4.2), (4.1), the second inequality in (4.6), and (3.1), we obtain

$$a_h(T_A v, v) \geq C a_h(v,v) \geq C \|v\|^2_{H^2(\Omega)}, \quad v \in V_h$$

(see Theorem 1 in [17]). Therefore, operator T_A is positive definite. Operators T_i^k are self-adjoint since $a_h(\cdot,\cdot)$ is a symmetric bilinear form; hence, T_A is self-adjoint (see Lemma 2 in section 5.2 of [36]). Thus, operator T_A^{-1} is self-adjoint positive definite. It follows from (4.3), (2.7), and Lemma 1 in section 5.2 of [36] that

$$(B_A v, v)_h = a_h(T_A^{-1} v, v) = \|v\|^2_{\Sigma,a_h}, \quad v \in V_h.$$

The last relation, along with (4.6) and (2.7), gives (4.4). $\square$

Multiplicative preconditioner. Let

$$T_M = I_h - \left[\prod_{k=K}^{0} \prod_{i=1}^{N_k} \left(I_h - T_i^k\right) \right] \left[\prod_{k=0}^{K} \prod_{i=N_k}^{1} \left(I_h - T_i^k\right) \right], \tag{4.7}$$

where I_h is the identity operator on V_h.

THEOREM 4.2. *Assume that h is sufficiently small. Linear operator T_M is self-adjoint positive definite on V_h in the inner product $a_h(\cdot,\cdot)$. Linear operator*

$$B_M = L_h^* L_h T_M^{-1} \tag{4.8}$$

is self-adjoint positive definite on V_h in the inner product $(\cdot,\cdot)_h$, and

$$C_1(B_M v, v)_h \le (L_h^* L_h v, v)_h \le C_2(B_M v, v)_h, \quad v \in V_h. \tag{4.9}$$

Proof. Since operators $\{T_i^k\}$ are self-adjoint on V_h with respect to the inner product $a_h(\cdot,\cdot)$, it is easy to see that T_M is also a self-adjoint operator. Hence, B_M is self-adjoint in the inner product $(\cdot,\cdot)_h$. Inequalities in (4.9) follow from Lemma 4 in section 5.2 of [36] with $\omega = 1$, the second inequality in (4.6), and Lemma 6.1 in [40] formulated for $\{V_k\}$. □

Remark. Using the multigrid terminology, we note that the multiplicative preconditioner B_M corresponds to the V-cycle multigrid algorithm with the Gauss–Seidel smoother.

Iterative method. Since the operator of (2.6) is self-adjoint positive definite, we can use our multilevel preconditioners with the PCG algorithm to compute the OSC solution (see Algorithm 9.4.14 in [20]).

Let $\lambda_{\min,h}$ and $\lambda_{\max,h}$ be, respectively, the smallest and the largest eigenvalues of the preconditioned operator

$$\tilde{A}_h = M_h^{-1/2} L_h^* L_h M_h^{-1/2},$$

where M_h is a preconditioner. It is well known that the convergence rate of the PCG is bounded from above by

$$(\sqrt{\kappa_h} - 1)/(\sqrt{\kappa_h} + 1),$$

where $\kappa_h = \lambda_{\max,h}/\lambda_{\min,h}$ is the spectral condition number of $\tilde{A}_h$ (see Theorem 9.4.14 in [20]). It follows from Theorems 4.1 and 4.2 that

$$\kappa_h \le C_2/C_1 < \infty \ \text{ as } \ h \to \infty. \tag{4.10}$$

The estimate (4.10) implies that it takes $O(|\ln \epsilon|)$ iterations of the PCG algorithm with the multilevel OSC preconditioners to approximate the solution of (2.6) with tolerance ϵ; that is, the number of iterations is bounded by a constant independent of h and K.

5. OSC matrix-vector representation.

In this section, we introduce the matrix-vector representation of the OSC problem in the standard Hermite finite element basis and obtain recurrence relations for the computation of the OSC approximations and other required quantities.

Representation of Hermite piecewise cubic polynomials. Let V_k^1 denote a vector space of Hermite piecewise cubic polynomials vanishing at $t = 0$ and $t = 1$ and corresponding to the partition $\{t_i^k = i/2^k\}_{i=0}^{2^k}$ of the interval $[0,1]$. The dimension of V_k^1 is $M_k = 2^{k+1}$. Let

$$\Phi_k = \{s_0^k, v_1^k, s_1^k, \dots, v_{2^k-1}^k, s_{2^k-1}^k, s_{2^k}^k\} \equiv \{\phi_i\}_{i=1}^{M_k} \tag{5.1}$$

be the standard basis of V_k^1 consisting of nodal value and nodal slope basis functions $v_i(t)$ and $s_i(t)$, respectively, defined for $0 \le i \le 2^k$ by

$$\begin{aligned} &v_i^k(t_j^k) = \delta_{ij}, \quad (v_i^k)'(t_j^k) = 0, \quad 0 \le j \le 2^k, \\ &s_i^k(t_j^k) = 0, \quad (s_i^k)'(t_j^k) = h_k^{-1}\delta_{ij}, \quad 0 \le j \le 2^k, \end{aligned} \tag{5.2}$$

where δ_{ij} is the Kronecker delta. We note that the basis functions v_0^k and $v_{2^k}^k$ corresponding to the boundary points $t = 0$ and $t = 1$ are not included in Φ_k.

The value and the slope basis functions in Φ_k corresponding to an interior partition node are uniquely expressed as a linear combination of five basis functions in Φ_{k+1} as follows:

$$\begin{aligned} v_i^k &= \tfrac{1}{2}v_{2i-1}^{k+1} + \tfrac{3}{4}s_{2i-1}^{k+1} + v_{2i}^{k+1} + \tfrac{1}{2}v_{2i+1}^{k+1} - \tfrac{3}{4}s_{2i+1}^{k+1}, \\ s_i^k &= -\tfrac{1}{8}v_{2i-1}^{k+1} - \tfrac{1}{8}s_{2i-1}^{k+1} + \tfrac{1}{2}s_{2i}^{l+1} + \tfrac{1}{8}v_{2i+1}^{k+1} - \tfrac{1}{8}s_{2i+1}^{k+1}. \end{aligned} \tag{5.3}$$

Let

$$P = \begin{pmatrix} 4 & -1 \\ 6 & -1 \end{pmatrix}, \; Q = \begin{pmatrix} 8 & 0 \\ 0 & 4 \end{pmatrix}, \; \text{and } R = \begin{pmatrix} 4 & 1 \\ -6 & -1 \end{pmatrix} \tag{5.4}$$

be matrices corresponding to the representation (5.3). Let P_k^1 be a $2M_k \times M_k$ matrix obtained from

$$\frac{1}{8}\begin{pmatrix} Q & O & O & \dots & O \\ R & P & O & & \\ O & Q & O & & \\ O & R & P & & \\ \vdots & & & \ddots & \vdots \\ O & & & \dots & Q \end{pmatrix} \in R^{(2M_k+2)\times(M_k+2)}, \tag{5.5}$$

where O is the 2×2 zero matrix, by removing the first and next-to-last rows and columns. For $v \in V_k^1$, let $[v]_{\mathcal{H},k}$ denote the vector representation of v in the basis Φ_k. It follows from (5.3), (5.4), and (5.5) that

$$[v]_{\mathcal{H},k+1} = P_k^1[v]_{\mathcal{H},k}, \quad v \in V_k^1, \quad k \in \{0,1,2,\dots,K-1\}. \tag{5.6}$$

Representation of Hermite piecewise bicubic polynomials. We note that $V_k = V_k^1 \otimes V_k^1$, where the symbol $\otimes$ denotes a vector space tensor product. Set

$$\psi_{M_k(i-1)+j}^k(x) = \phi_i(x_1)\phi_j(x_2) \quad \text{for } 1 \le i,j \le M_k, \tag{5.7}$$

where basis functions ϕ_i, for $1 \le i \le M_k$, are defined by (5.1) and (5.2). The set $\Psi_k = \{\psi_j^k\}_{j=1}^{N_k}$, where $N_k = M_k^2 = 4^{k+1}$, is the standard basis for V_k in the standard ordering.

It follows from (5.3) and (5.7) that a basis function in Ψ_k corresponding to an interior partition node is uniquely expressed as a linear combination of 25 basis functions in Ψ_{k+1}. Let $[v]_{\mathcal{H},k}$ denote the vector representation of $v \in V_k$ in the basis Ψ_k, and let $[v]_{\mathcal{H}} = [v]_{\mathcal{H},K}$ for $v \in V_h$. It is obvious that

$$[\psi_j^k]_{\mathcal{H},k} = \vec{e}_j^{\,k}, \quad 1 \le j \le N_k, \tag{5.8}$$

where $\vec{e}_j^{\,k}$ is the jth standard basis vector in R^{N_k}.

We set

$$P_k = P_k^1 \otimes P_k^1 \in R^{2N_k \times N_k}, \tag{5.9}$$

where P_k^1 is the one-dimensional interpolation matrix in (5.6) and the symbol $\otimes$ now denotes the matrix tensor product. Matrix P_k corresponds to the piecewise bicubic Hermite interpolation in V_{k+1}, and we call P_k the interpolation matrix from level k to level $k+1$. It follows from (5.6), (5.7), and (5.9) that

$$[v]_{\mathcal{H},k+1} = P_k [v]_{\mathcal{H},k} \quad \text{for } \ 0 \le k \le K-1 \ \text{ and } \ v \in V_k. \tag{5.10}$$

Applying formula (5.10) recurrently, we obtain

$$[v]_{\mathcal{H}} = P_{K-1} \cdots P_k [v]_{\mathcal{H},k}, \quad v \in V_k. \tag{5.11}$$

In particular, replacing v in (5.11) by ψ_j^k and using (5.8), we get

$$[\psi_j^k]_{\mathcal{H}} = P_{K-1} \cdots P_k \vec{e}_j^{\,k}, \quad 1 \le j \le N_k. \tag{5.12}$$

Matrix-vector form of the OSC problem. Assume that the set of Gauss points $\mathcal{G}_h = \{\xi_i\}_{i=1}^N$ is ordered, and let

$$[v]_{\mathcal{G}} = [v(\xi_1), \dots, v(\xi_N)]^t$$

for any function v defined on $\mathcal{G}_h$. From (2.5), we have

$$(v, w)_h = (h^2/4)[w]_{\mathcal{G}}^t [v]_{\mathcal{G}}, \quad v, w \in V_h. \tag{5.13}$$

Let $[L_h]$ be the OSC matrix of size $N \times N$, corresponding to the differential operator L in (1.2), with entries $L\psi_j^K(\xi_i)$, where i is the row index. For a continuous function $g(x)$, let

$$D(g) = \operatorname{diag}\left(g(\xi_1), \dots, g(\xi_N)\right)$$

be a diagonal matrix. We note that

$$\begin{aligned}[L_h] = {} & D(a_{11})(\hat{A} \otimes \hat{B}) + 2D(a_{12})(\hat{C} \otimes \hat{C}) + D(a_{22})(\hat{B} \otimes \hat{A}) \\ & + D(b_1)(\hat{C} \otimes \hat{B}) + D(b_2)(\hat{B} \otimes \hat{C}) + D(c)(\hat{B} \otimes \hat{B}),\end{aligned}$$

and the matrices $\hat{A}$, $\hat{C}$, and $\hat{B}$ have the following almost block diagonal structure:

$$\begin{pmatrix} \tilde{W} & Z & O & O & \cdots & O & \tilde{O} \\ \tilde{O} & W & Z & O & & & \\ \tilde{O} & O & W & Z & & & \\ \vdots & \vdots & & & \ddots & \vdots & \vdots \\ \tilde{O} & O & & & \dots & W & \tilde{Z} \end{pmatrix} \in R^{M_K \times M_K},$$

where W, Z, O and $\tilde{W}$, $\tilde{Z}$, $\tilde{O}$ are, respectively, 2×2 and 2×1 blocks; O and $\tilde{O}$ are zero matrices. The (i,j) entries of matrices $\hat{A}$, $\hat{C}$, and $\hat{B}$ are $\phi_i''(\eta_j)$, $\phi_i'(\eta_j)$, and $\phi_i(\eta_j)$, respectively, where $\{\eta_j\}_{j=1}^{M_K}$ is the set of Gauss points in interval $[0,1]$ corresponding to partition π_h.

It is easy to see that

$$[L_h v]_{\mathcal{G}} = [L_h][v]_{\mathcal{H}}, \quad v \in V_h. \tag{5.14}$$

Using (2.7), (5.13), and (5.14), we obtain, for any v and w in V_h,

$$a_h(v,w) = (L_h v, L_h w)_h = (h^2/4)[L_h w]_{\mathcal{G}}^t [L_h v]_{\mathcal{G}} = (h^2/4)[w]_{\mathcal{H}}^t [L_h]^t [L_h][v]_{\mathcal{H}}. \tag{5.15}$$

Let $A_k = (a_{ij}^k)$ be an $N_k \times N_k$ matrix with entries

$$a_{ij}^k = (4/h^2) a_h(\psi_j^k, \psi_i^k), \tag{5.16}$$

and let $A = A_K$. From (5.16), using (5.15), we get

$$a_{ij}^k = [\psi_j^k]_{\mathcal{H}}^t [L_h]^t [L_h][\psi_i^k]_{\mathcal{H}}. \tag{5.17}$$

From (5.17) for $k = K$, using (5.8), we obtain

$$A = [L_h]^t [L_h]. \tag{5.18}$$

Similarly, using (5.13), (5.14), the relation $[f]_{\mathcal{G}} = [f_h]_{\mathcal{G}}$, and (5.8), we obtain

$$\begin{aligned}(4/h^2)(L_h^* f_h, \psi_i^K)_h &= (4/h^2)(f_h, L_h \psi_i^K)_h = [f_h]_{\mathcal{G}}^t [L_h \psi_i^K]_{\mathcal{G}} \\ &= [f]_{\mathcal{G}}^t [L_h][\psi_i^K]_{\mathcal{H}} = (\vec{e}_i^{\,K})^t [L_h]^t [f]_{\mathcal{G}}.\end{aligned}$$

Thus, the variational OSC equation (2.8) has the matrix-vector form

$$A\,[u_h]_{\mathcal{H}} = [L_h]^t [f]_{\mathcal{G}}. \tag{5.19}$$

6. Implementation.

In this section, we describe implementations of both the additive and the multiplicative OSC preconditioners.

Additive preconditioner. Let us describe the computation of $w = B_A^{-1} v$ for $v \in V_h$, where B_A is defined by (4.3). Let

$$w_k = \sum_{i=1}^{N_k} w_{ki}, \quad \text{where } \; w_{ki} = T_i^k (L_h^* L_h)^{-1} v. \tag{6.1}$$

Using (4.3), (4.2), and (6.1), we get

$$w = T_A (L_h^* L_h)^{-1} v = \left(\sum_{k,i} T_i^k \right) (L_h^* L_h)^{-1} v = \sum_{k,i} w_{ki} = \sum_k w_k. \tag{6.2}$$

Thus, to compute w, we need to compute and sum w_k for $k = 0, 1, \ldots, K$.

Using (4.1) and (2.7), we obtain, for $1 \le i \le N_k$,

$$a_h(w_{ki}, \psi_i^k) = a_h(T_i^k (L_h^* L_h)^{-1} v, \psi_i^k) = a_h((L_h^* L_h)^{-1} v, \psi_i^k) = (v, \psi_i^k)_h. \tag{6.3}$$

Let

$$w_{ki} = c_{ki} \psi_i^k, \quad c_{ki} \in R, \tag{6.4}$$

$$[v]_k = (4/h^2)[(v, \psi_1^k)_h, \ldots, (v, \psi_{N_k}^k)_h]^t \in R^{N_k}. \tag{6.5}$$

Substituting (6.4) into (6.3) and using (6.5) and (5.16), we rewrite (6.3) in the form

$$\operatorname{diag}(A_k)\vec{w}_k = [v]_k, \tag{6.6}$$

where $\vec{w}_k = (c_{k1}, \ldots, c_{kN_k})^t = [w_k]^t_{\mathcal{H},k}$ and $\operatorname{diag}(A_k) = \operatorname{diag}(a^k_{11}, \ldots, a^k_{N_kN_k})$.

Let $[I_h]$ be an $N \times N$ matrix with entries $\psi^K_j(\xi_i)$, where i is the row index. Matrix $[I_h]$ is nonsingular and maps $[v]_{\mathcal{H}}$ to $[v]_{\mathcal{G}}$, that is,

$$[v]_{\mathcal{G}} = [I_h]\,[v]_{\mathcal{H}}, \quad v \in V_h. \tag{6.7}$$

For $[v]_k$ defined by (6.5), let us show

$$[v]_K = [I_h]^t[v]_{\mathcal{G}}, \tag{6.8}$$

$$[v]_k = P^t_k[v]_{k+1} \quad \text{for} \quad k = K-1, K-2, \ldots, 0, \tag{6.9}$$

where the interpolation matrix P_k is defined by (5.9) and (5.5). Using (6.5), (5.13), and (6.7), we obtain

$$(\vec{e}^{\,k}_j)^t[v]_k = (4/h^2)(v, \psi^k_j)_h = [\psi^k_j]^t_{\mathcal{G}}[v]_{\mathcal{G}} = [\psi^k_j]^t_{\mathcal{H}}[I_h]^t[v]_{\mathcal{G}}, \quad 1 \le j \le N_k. \tag{6.10}$$

Relation (6.8) follows from (6.10) with $k = K$ and (5.8). Using (6.10), (5.12), and (6.8), we obtain

$$[v]_k = P^t_k \cdots P^t_{K-1}[v]_K,$$

which implies (6.9).

The multiplication by P^t_k is carried out using the representation

$$P^t_k = ((P^1_k)^t \otimes I_k)(I_k \otimes (P^1_k)^t),$$

where I_k is the identity matrix of size $M_k \times M_k$. Matrices $\{A_k\}$ can be precomputed using the recurrence formula

$$A_k = P^t_k A_{k+1} P_k \quad \text{for} \quad k = K-1, K-2, \ldots, 0, \tag{6.11}$$

which follows from (5.17), (5.12), and (5.18). Finally, to compute $w = \sum_k w_k$, that is, $[w]_{\mathcal{H}}$, we implement

$$\vec{w}_{k+1} \leftarrow \vec{w}_{k+1} + P_k\vec{w}_k, \quad k = 0, 1, \ldots, K-1.$$

The additive preconditioning algorithm is presented in Figure 6.1. It is easy to see that the computational cost of the additive algorithm is $O(h^{-2}) = O(4^K)$.

Multiplicative preconditioner. We now consider the computation of $w = B^{-1}_M v$ for $v \in V_h$, where B_M is defined in (4.8). Let

$$u = (L^*_h L_h)^{-1}v. \tag{6.12}$$

Using (4.8) and (6.12), we get

$$w = B^{-1}_M v = T_M(L^*_h L_h)^{-1}v = T_M u,$$

which, by (4.7), implies

$$u - w = \left[\prod_{k=K}^{0}\prod_{i=1}^{N_k}(I_h - T^k_i)\right]\left[\prod_{k=0}^{K}\prod_{i=N_k}^{1}(I_h - T^k_i)\right]u. \tag{6.13}$$

```
input: $K$, $[v]_K$, $\{\mathrm{diag}(A_k)\}_{k=0}^K$
output: $[w]_{\mathcal{H}}$
$\vec{v}_K \leftarrow [v]_K$
for $k = K, K-1, \ldots, 0$
    if $(k < K)$ $\vec{v}_k = P_k^t \vec{v}_{k+1}$ end
    solve $\mathrm{diag}(A_k)\vec{w}_k = \vec{v}_k$
end
for $k = 0, 1, \ldots, K-1$
    $\vec{w}_{k+1} \leftarrow \vec{w}_{k+1} + P_k \vec{w}_k$
end
$[w]_{\mathcal{H}} \leftarrow \vec{w}_K$
```

FIG. 6.1. *Additive preconditioning algorithm.*

Let S be an ordered set of pairs (k, i) with the ordering corresponding to that of factors in (6.13) from right to left. Setting $y = u - w$, we see that $u - w$ can be computed by

$$y \leftarrow u; \;\; y \leftarrow (I_h - T_i^k)y \;\text{ for }\; (k, i) \in S,$$

which is equivalent to

$$w \leftarrow 0; \;\; w \leftarrow w + T_i^k(u - w) \;\text{ for }\; (k, i) \in S. \tag{6.14}$$

Let us develop an efficient implementation of the algorithm in (6.14). Using (4.1), (2.7), and (6.12), we obtain

$$a_h(T_i^k(u - w), \psi_i^k) = a_h(u - w, \psi_i^k) = (v, \psi_i^k)_h - a_h(w, \psi_i^k), \quad \psi_i^k \in V_{ki}. \tag{6.15}$$

Substituting $T_i^k(u - w) = c_{ki}\psi_i^k$ into (6.15) and (6.14), we get

$$c_{ki} = g_i^k(w)/a_{ii}^k, \tag{6.16}$$

$$w \leftarrow w + c_{ki}\psi_i^k, \tag{6.17}$$

where a_{ii}^k is as defined in (5.16), and

$$g_{ki}(w) = (4/h^2)\left[(v, \psi_i^k)_h - a_h(w, \psi_i^k)\right], \quad 1 \le i \le N_k. \tag{6.18}$$

Let $\vec{g}_k(w) = (g_{k1}(w), \ldots, g_{kN_k}(w))^t$. We note that, each time the value of w is changed by (6.17), all entries of vector $\vec{g}^k(w)$ should be updated by (6.18), and such computation requires a matrix-vector product. It is more efficient to use a recurrent assignment

$$\vec{g}_k(w) \leftarrow \vec{g}_k(w) - c_{ki}(a_{1i}^k, \ldots, a_{N_k,i}^k)^t, \tag{6.19}$$

which is obtained by multiplying (6.17) by ψ_j^k in the $a_h(\cdot,\cdot)$ inner product and subtracting $(v, \psi_j^k)_h$ from both sides of the resulting assignment.

For $k = K-1, K-2, \ldots, 0$, let w_k be the value of w after implementing

$$w \leftarrow w + T_i^l(u - w), \quad i = 1, \ldots, N_l, \;\; l = K, K-1, \ldots, k+1,$$

```
input: $K$, $[v]_K$, $\{A_k\}_{k=0}^K$
output: $[w]_{\mathcal{H}}$
$\vec{g}_K \leftarrow [v]_K$
for $k = K, \dots, 0$
    solve $L_k \vec{w}_k = \vec{g}_k$
    if $(k > 0)$ $\vec{g}_{k-1} \leftarrow P_{k-1}^t(\vec{g}_k - A_k \vec{w}_k)$ end
end
for $k = 0, \dots, K$
    if $(k > 0)$ $\vec{w}_k \leftarrow \vec{w}_k + P_{k-1} \vec{w}_{k-1}$ end
    solve $L_k^t \vec{w} = \vec{g}_k - A_k \vec{w}_k$
    $\vec{w}_k \leftarrow \vec{w}_k + \vec{w}$
end
$[w]_{\mathcal{H}} \leftarrow \vec{w}_K$
```

FIG. 6.2. *Multiplicative preconditioning algorithm.*

and let

$$\vec{g}_k = \vec{g}_k(w_k). \tag{6.20}$$

We note that (6.16) followed by (6.19) for $i = 1, \dots, N_k$ is the column-oriented algorithm of solving a lower triangular linear system $L_k \vec{w}_k = \vec{g}_k$, where matrix L_k contains the lower triangular part of A_k and $\vec{w}_k = (c_{k1}, \dots, c_{k,N_k})^t$.

Let us show that vectors $\{\vec{g}_k\}$ can be computed by the recurrence formula

$$\vec{g}_{k-1} = P_{k-1}^t\,(\vec{g}_k - A_k \vec{w}_k). \tag{6.21}$$

Using (6.19), the definition of w_{k-1}, and (6.20), we obtain

$$\vec{g}_k(w_{k-1}) = \vec{g}_k - A_k \vec{w}_k. \tag{6.22}$$

Applying (6.18), (6.5), (5.15), (5.12), we get

$$\vec{g}_k(w) = [v]_k - P_k^t \cdots P_{K-1}^t [L_h]^t [L_h][w]_{\mathcal{H}}. \tag{6.23}$$

From (6.23) with k replaced by $k-1$ and (6.9), we have

$$\vec{g}_{k-1}(w) = P_{k-1}^t \left([v]_k - P_k^t \cdots P_{K-1}^t [L_h]^t [L_h][w]_{\mathcal{H}}\right) = P_{k-1}^t \vec{g}_k(w),$$

which, by (6.22), implies (6.21).

In the ascend phase, for $k = 0, \dots, K$, an upper triangular linear system with the matrix L_k^t is solved. The algorithm implementing the multiplicative preconditioning is given in Figure 6.2. It is easy to see that the cost of the multiplicative algorithm is $O(h^{-2}) = O(4^K)$.

7. Numerical results. In this section, we present numerical results for solving test problems by the PCG method with the multilevel OSC preconditioners developed in this work. For a chosen exact solution $u(x)$ of BVP (1.1), we set $f = Lu$. PCG iterations are stopped when the relative residual norm, that is, the ratio of the 2-norm of the residual of (5.19) to the 2-norm of the right-hand side, becomes less than tolerance ϵ.

Table 7.1

Comparison of multilevel preconditioners to solve the normal and the original OSC equations by PCG ($L = \Delta$, $\epsilon = 10^{-12}$).

	Normal OSC equation				Original OSC equation			
	Additive		Multiplicative		Additive		Multiplicative	
h	κ_h	Iter.	κ_h	Iter.	κ_h	Iter.	κ_h	Iter.
1/16	4.490	26	1.434	12	7.551	30	1.169	9
1/32	5.016	29	1.476	13	7.794	30	1.167	9
1/64	5.488	32	1.501	13	7.908	30	1.165	9
1/128	5.845	34	1.516	14	7.962	29	1.164	9
1/256	6.162	35	1.525	14	7.991	28	1.162	9

The spectral condition number κ_h of the preconditioned OSC operator satisfies (4.10), that is, $\kappa_h < C_2/C_1$ as $h \to 0$. To demonstrate this fact numerically, we compute quantities that approximate κ_h on a sequence of nested partitions using the PCG iteration parameters and present these approximations in the following tables under κ_h. Under "Iter.", we report the numbers of iterations to reduce the relative residual norm within the specified tolerance.

In the first set of experiments, we compare our preconditioners with those proposed in [9], where both additive and multiplicative multilevel preconditioners were developed to solve the original OSC equation (2.3) for a self-adjoint L. As in [9], we take $L = \Delta$, the Laplacian, and

$$u(x) = 10x_1^2(1 - x_1)x_2^2(1 - x_2).$$

Since $u(x)$ is a bicubic polynomial, it is also the solution of the OSC problem; hence, the discretization error is zero. The numerical results are presented in Table 7.1, and they indicate that, for $L = \Delta$, PCG with multilevel preconditioners for the original OSC equation is slightly more efficient than that for the normal OSC equation. For smaller values of h, the approximations to κ_h and the numbers of iterations are relatively small and change insignificantly.

In the next set of experiments, we solve the same problems as in [1], where the PCG algorithm was tested with a direct solver preconditioner. The operator L in (1.2) is taken with the coefficients

$$\begin{aligned} &a_{11}(x) = e^{x_1x_2}, && b_1(x) = x_2e^{x_1x_2} + \beta_1 \cos[\pi(x_1 + x_2)], \\ &a_{12}(x) = \alpha/(1 + x_1 + x_2), && b_2(x) = -x_1e^{-x_1x_2} + \beta_2 \sin(2\pi x_1x_2), \\ &a_{22}(x) = e^{-x_1x_2}, && c(x) = \gamma[1 + 1/(1 + x_1 + x_2)], \end{aligned} \tag{7.1}$$

where α, β_1, β_2, and γ are parameters, and the exact solution of BVP (1.1) is set to

$$u(x) = e^{x_1+x_2}x_1x_2(1 - x_1)(1 - x_2).$$

Using the PCG with the multiplicative preconditioner, we solve the following four test problems corresponding to the differential operator L, which is defined in each problem:

P1 – self-adjoint negative definite, $\alpha = \beta_1 = \beta_2 = \gamma = 0$.

P2 – self-adjoint indefinite, $\alpha = \beta_1 = \beta_2 = 0$ and $\gamma = 100$.

P3 – non–self-adjoint indefinite, $\beta_2 = 100$ and $\alpha = \beta_1 = \gamma = 0$.

P4 – non–self-adjoint indefinite, $\alpha = 0.5$, $\beta_1 = 10$, $\beta_2 = \gamma = 50$.

Table 7.2

Approximations to the spectral condition number κ_h and PCG iteration numbers for the variable coefficient L ($\epsilon = 10^{-10}$). Top part: multilevel preconditioner. Bottom part: additive preconditioner.

	P1		P2		P3		P4	
h	κ_h	Iter.	κ_h	Iter.	κ_h	Iter.	κ_h	Iter.
1/16	2.570	15 (26)	363.0	68 (46)	624.5	43 (34)	516.1	51 (68)
1/32	3.646	17 (30)	363.2	70 (51)	499.8	42 (38)	457.8	61 (75)
1/64	4.922	20 (33)	361.6	71 (54)	459.3	46 (40)	402.9	56 (81)
1/128	6.189	23 (34)	361.5	72 (55)	398.6	52 (42)	382.0	58 (84)
1/256	7.253	25	360.3	73	376.0	54	377.8	59
1/16	30.3	43	5835.0	103	9764.6	71	6961.2	67
1/32	50.9	58	5802.3	117	7396.7	79	5999.5	83
1/64	75.6	70	5801.5	128	6562.9	98	5131.5	104
1/128	99.9	81	5803.4	140	5414.9	122	5008.0	121
1/256	117.6	89	5805.2	147	5025.1	139	5213.8	137

In the top part of Table 7.2, we report results with the multiplicative preconditioner, and, in parentheses, we reproduce results reported in [1].

For all four problems, the numbers of iterations increase slowly with decreasing h. It takes the least number of iterations to solve the self-adjoint negative definite problem P1, and the most number of iterations to solve P2, the "most indefinite" problem of P1–P4. For P1, approximations to the spectral ratio κ_h are much smaller than those for P2–P4, where the approximations to κ_h are about the same for smaller values of h. It is interesting to note that κ_h monotonically increases for P1 and decreases for P2–P4 as h decreases. We observe that the approximations to κ_h for P2–P4 are significantly larger than those for P1. The numbers of iterations for P1 and P4 favor the multilevel multiplicative preconditioner rather than the direct solver preconditioner developed in [1], although the preconditioner in [1] produces smaller numbers of iterations for P2 and P3.

Results for the additive preconditioner are presented in the bottom part of Table 7.2, and they suggest that, for the tested problems, the additive preconditioner is less efficient, in terms of numbers of iterations, than the multiplicative preconditioner. The approximations of κ_h computed with the additive preconditioner are approximately 15 times larger than those computed with the multiplicative preconditioner, and the indicated difference is reflected by a larger number of iterations for the additive preconditioner. This result is intuitively expected based on known properties of Jacobi and Gauss–Seidel smoothers for finite difference operators.

In Figure 7.1, we display plots of residual curves for $h = 1/256$. We observe monotone convergence only for the self-adjoint negative definite problem P1; for P2–P4, the residual curves are plotted relatively close to each other.

In the last set of experiments, we solve the Helmholtz equation $\Delta u + k^2 u = f$ for several values of k^2. Numerical results were obtained using the multiplicative preconditioner, and they are presented in Table 7.3. We see that the approximations to κ_h change insignificantly when h is decreased for all taken values of k^2. For $k^2 = 1000$, the approximations to k_h are very large; however, for $k^2 = 1400$ they are the smallest. The approximations to k_h are large because of small values of the smallest eigenvalue $\lambda_{\min,h}$ of the preconditioned operator. In Figure 7.2, we plot the eigenvalues of the OSC matrix $[L_h]$ with $h = 1/32$ ($N = 4,096$) for the Helmholtz equation with $k^2 = 1400$. We note that the eigenvalues are widely spread over the complex plane.

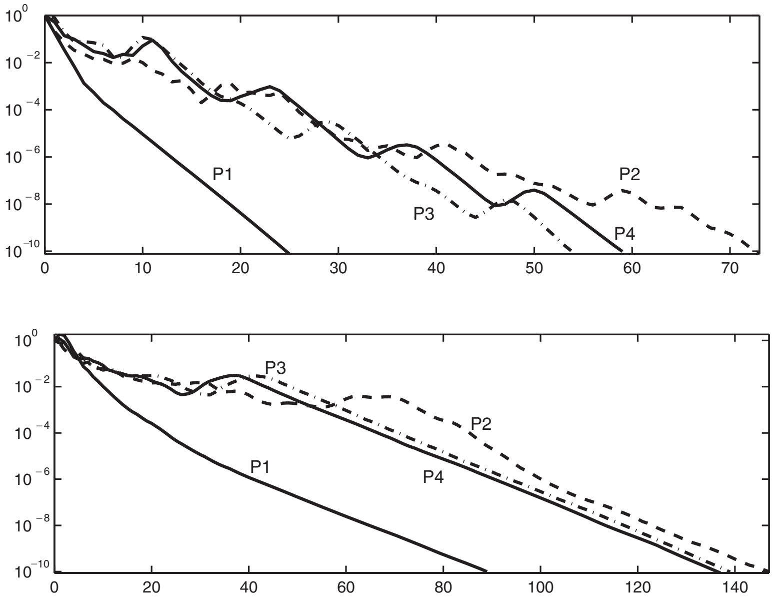


Fig. 7.1. *Logarithmic plots of the relative residual norm versus iteration number* ($h = 1/256$). *Top figure: multiplicative preconditioner. Bottom figure: additive preconditioner.*

Table 7.3

Approximations to the spectral condition number κ_h *and PCG iteration numbers for the Helmholtz equation* ($\epsilon = 10^{-10}$).

	$k^2 = 100$		$k^2 = 500$		$k^2 = 1000$		$k^2 = 1400$	
h	κ_h	Iter.	κ_h	Iter.	κ_h	Iter.	κ_h	Iter.
1/16	1468.6	51	1808.6	108	133089.6	200	301.4	98
1/32	1446.7	51	1465.0	117	34202.1	175	514.3	133
1/64	1442.1	51	1437.6	125	26442.1	182	545.2	141
1/128	1441.2	51	1435.0	128	25981.5	183	546.1	151
1/256	1441.0	51	1434.8	144	25947.7	199	545.9	154

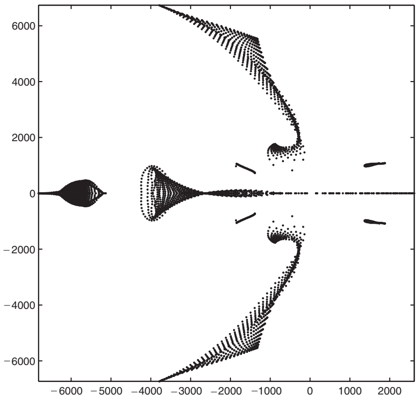


Fig. 7.2. *Eigenvalues of the OSC matrix* $[L_h]$ *for the Helmholtz equation on the complex plane* ($k^2 = 1400,\ \ h = 1/32$).

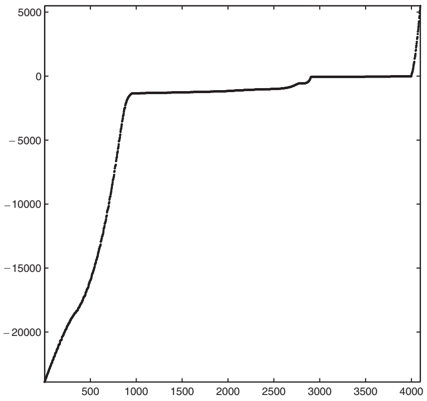


FIG. 7.3. *Eigenvalues of the OSC operator L_h for the Helmholtz equation plotted against their ordering numbers* ($k^2 = 1400, \;\; h = 1/32$).

In Figure 7.3, we plot the eigenvalues of the symmetric matrix $[I_h]^t[L_h]$, which are the eigenvalues of the OSC operator L_h. We note that L_h has large numbers of both positive and negative eigenvalues.

Acknowledgments. The author wishes to thank Prof. Bernard Bialecki for his assistance during the preparation of this paper and the referees for their valuable comments.

REFERENCES

[1] R. AITBAYEV AND B. BIALECKI, *A preconditioned conjugate gradient method for nonselfadjoint or indefinite orthogonal spline collocation problems*, SIAM J. Numer. Anal., 41 (2003), pp. 589–604.

[2] U. ASCHER, S. PRUESS, AND R. D. RUSSELL, *On spline basis selection for solving differential equations*, SIAM J. Numer. Anal., 20 (1983), pp. 121–142.

[3] O. AXELSSON AND V. A. BARKER, *Finite Element Solution of Boundary Value Problems: Theory and Computation*, SIAM, Philadelphia, 2001.

[4] R. E. BANK, *A comparison of two multilevel iterative methods for nonsymmetric and indefinite elliptic finite element equations*, SIAM J. Numer. Anal., 18 (1981), pp. 724–743.

[5] B. BIALECKI, *An alternating direction implicit method for orthogonal spline collocation linear systems*, Numer. Math., 59 (1991), pp. 413–429.

[6] B. BIALECKI, *A fast domain decomposition Poisson solver on a rectangle for Hermite bicubic orthogonal spline collocation*, SIAM J. Numer. Anal., 30 (1993), pp. 425–434.

[7] B. BIALECKI, *Convergence analysis of orthogonal spline collocation for elliptic boundary value problems*, SIAM J. Numer. Anal., 35 (1998), pp. 617–631.

[8] B. BIALECKI, *Superconvergence of the orthogonal spline collocation solution of Poisson's equation*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 285–303.

[9] B. BIALECKI AND M. DRYJA, *Multilevel additive and multiplicative methods for orthogonal spline collocation problems*, Numer. Math., 77 (1997), pp. 35–58.

[10] B. BIALECKI, G. FAIRWEATHER, AND K. R. BENNETT, *Fast direct solvers for piecewise Hermite bicubic orthogonal spline collocation equations*, SIAM J. Numer. Anal., 29 (1992), pp. 156–173.

[11] J. H. BRAMBLE, D. Y. KWAK, AND J. E. PASCIAK, *Uniform convergence of multigrid V-cycle iterations for indefinite and nonsymmetric problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1746–1763.

[12] A. BRANDT AND I. LIVSHITS, *Wave-ray multigrid method for standing wave equations*, Electron. Trans. Numer. Anal., 6 (1997), pp. 162–181.

[13] C. CHRISTARA AND B. F. SMITH, *Multigrid and multilevel methods for quadratic spline collocation*, BIT, 37 (1997), pp. 781–803.

[14] K. D. Cooper and P. M. Prenter, *Alternating direction collocation for separable elliptic partial differential equations*, SIAM J. Numer. Anal., 28 (1991), pp. 711–727.

[15] C. de Boor and B. Swartz, *Collocation at Gaussian points*, SIAM J. Numer. Anal., 10 (1973), pp. 582–606.

[16] R. DeVore and V. Popov, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 397–414.

[17] M. Dryja and O. B. Widlund, *Schwarz methods of Neumann-Neumann type for* 3*-dimensional elliptic finite-element problems*, Comm. Pure Appl. Math., 48 (1995), pp. 121–155.

[18] W. R. Dyksen, *Tensor product generalized ADI methods for separable elliptic problems*, SIAM J. Numer. Anal., 24 (1987), pp. 59–76.

[19] J. Gary, *The multigrid iteration applied to the collocation method*, SIAM J. Numer. Anal., 18 (1981), pp. 211–224.

[20] W. Hackbusch, *Iterative Solution of Large Sparse Systems of Equations*, Springer-Verlag, New York, 1994.

[21] A. Hadjidimos, E. N. Houstis, J. R. Rice, and E. Vavalis, *Modified successive overrelaxation (MSOR) and equivalent* 2*-step iterative methods for collocation matrices*, J. Comput. Appl. Math., 42 (1992), pp. 375–393.

[22] E. N. Houstis, W. F. Mitchell, and J. R. Rice, *Algorithms INTCOL and HERMCOL: Collocation on rectangular domains with bicubic Hermite polynomials*, ACM Trans. Math. Software, 11 (1985), pp. 416–418.

[23] E. N. Houstis, W. F. Mitchell, and J. R. Rice, *Collocation software for second-order partial differential equations*, ACM Trans. Math. Software, 11 (1985), pp. 379–412.

[24] S. D. Kim and S. V. Parter, *Preconditioning cubic spline collocation discretizations of elliptic problems*, Numer. Math., 72 (1995), pp. 39–72.

[25] O. A. Ladyzhenskaja and N. N. Ural'ceva, *Linear and Quasilinear Elliptic Equations*, Academic Press, New York, 1968.

[26] Y.-L. Lai, A. Hadjidimos, E. N. Houstis, and J. R. Rice, *On the iterative solution of Hermite collocation equations*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 254–277.

[27] J. Mandel, *Multigrid convergence for nonsymmetric, indefinite variational problems and one smoothing step*, Appl. Math. Comput., 19 (1986), pp. 201–216.

[28] T. A. Manteuffel and S. V. Parter, *Preconditioning and boundary conditions*, SIAM J. Numer. Anal., 27 (1990), pp. 656–694.

[29] G. Mateescu, C. J. Ribbens, and L. T. Watson, *A domain decomposition preconditioner for Hermite collocation problems*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 135–151.

[30] P. Oswald, *On function spaces related to finite element approximation theory*, Z. Anal. Anwendungen, 9 (1990), pp. 43–64.

[31] P. Oswald, *On discrete norm estimates related to multilevel preconditioners in the finite element method*, in Proceedings of the International Conference on Constructive Theory of Functions (Varna 1991), K. G. Ivanov, P. Petrushev, and B. Sendov, eds., Publ. House Bulgarian Academy of Sciences, Sofia, 1992, pp. 203–214.

[32] P. Oswald, *Multilevel preconditioners for discretizations of the biharmonic equation by rectangular finite elements*, Numer. Linear Algebra Appl., 2 (1995), pp. 487–505.

[33] P. Percell and M. F. Wheeler, *A C^1 finite element collocation method for elliptic equations*, SIAM J. Numer. Anal., 17 (1980), pp. 605–622.

[34] J. R. Rice and R. F. Boisvert, *Solving Elliptic Problems Using ELLPACK*, Springer-Verlag, New York, 1985.

[35] M. H. Schultz, *Spline Analysis*, Prentice–Hall, Inc., Englewood Cliffs, NJ, 1973.

[36] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[37] W. Sun, *Block iterative algorithms for solving Hermite bicubic collocation equations*, SIAM J. Numer. Anal., 33 (1996), pp. 589–601.

[38] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North–Holland, Amsterdam, New York, 1978.

[39] X. Zhang, *Multilevel Schwarz methods*, Numer. Math., 63 (1992), pp. 521–539.

[40] X. Zhang, *Multilevel Schwarz methods for the biharmonic Dirichlet problem*, SIAM J. Sci. Comput., 15 (1994), pp. 621–644.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 707–732

ADAPTIVE MULTIVARIATE APPROXIMATION USING BINARY SPACE PARTITIONS AND GEOMETRIC WAVELETS[*]

S. DEKEL[†] AND D. LEVIATAN[‡]

Abstract. The binary space partition (BSP) technique is a simple and efficient method to adaptively partition an initial given domain to match the geometry of a given input function. As such, the BSP technique has been widely used by practitioners, but up until now no rigorous mathematical justification for it has been offered. Here we attempt to put the technique on sound mathematical foundations, and we offer an enhancement of the BSP algorithm in the spirit of what we are going to call *geometric wavelets*. This new approach to sparse geometric representation is based on recent developments in the theory of multivariate nonlinear piecewise polynomial approximation. We provide numerical examples of n-term geometric wavelet approximations of known test images and compare them with dyadic wavelet approximation. We also discuss applications to image denoising and compression.

Key words. binary space partitions, geometric wavelets, piecewise polynomial approximation, nonlinear approximation, adaptive multivariate approximation

AMS subject classifications. 41A15, 41A25, 41A17, 41A63, 65T60, 68U10

DOI. 10.1137/040604649

1. Introduction.

The binary space partition (BSP) technique is widely used in image processing and computer graphics [15, 17, 19], and can be described as follows. Given an initial convex domain in $\mathbb{R}^d$, such as $[0,1]^d$, and a function $f \in L_p([0,1]^d)$, $0 < p < \infty$, one subdivides the initial domain into two subdomains by intersecting it with a hyperplane. The subdivision is performed so that a given cost function is minimized. This subdivision process then proceeds recursively on the subdomains until some exit criterion is met. To be specific, we describe the algorithm of [17], which is a BSP algorithm, for the purpose of finding a compact geometric description of the target function, in this case a digital image ($d = 2$).

In [17], at each stage of the BSP process, for a given convex polytope Ω, the algorithm finds two subdomains Ω', Ω'' and two bivariate (linear) polynomials $Q_{\Omega'}$, $Q_{\Omega''}$ that minimize the quantity

$$\|f - Q_{\Omega'}\|^p_{L_p(\Omega')} + \|f - Q_{\Omega''}\|^p_{L_p(\Omega'')}$$

over all pairs Ω', Ω'' of polyhedral domains that are the result of a binary space partition of Ω. The polynomials $Q_{\Omega'}$, $Q_{\Omega''}$ are found using the least-squares technique with $p = 2$. The goal in [17] is to encode a *cut* of the BSP tree, i.e., a sparse piecewise polynomial approximation of the original digital image based on a union of disjoint polytopes from the BSP tree. Also, to meet a given bit target, rate-distortion optimization strategies are used (see also [21]).

Inspired by recent progress in multivariate piecewise polynomial approximation, made by Karaivanov, Petrushev, and collaborators [13, 14], we propose a modification to the above method which can be described as a *geometric wavelets* approach. Let

[*] Received by the editors March 2, 2004; accepted for publication (in revised form) November 1, 2004; published electronically August 31, 2005.

http://www.siam.org/journals/sinum/43-2/60464.html

[†] RealTimeImage, 6 Hamasger St., Or-Yehuda 60408, Israel (shai.dekel@turboimage.com).

[‡] School of Mathematical Sciences, Sackler Faculty of Exact Sciences, Tel-Aviv University, Tel-Aviv 69978, Israel (leviatan@math.tau.ac.il).

Ω' be a *child* of Ω in a BSP tree; i.e., $\Omega' \subset \Omega$ and Ω' has been created by a BSP partition of Ω. We use the polynomial approximations Q_Ω, $Q_{\Omega'}$ that were found for these domains by the local optimization algorithm above and define

$$\psi_{\Omega'} := \psi_{\Omega'}(f) := \mathbf{1}_{\Omega'}(Q_{\Omega'} - Q_\Omega) \tag{1.1}$$

as the geometric wavelet associated with the subdomain Ω' and the function f. A reader familiar with wavelets (see, e.g., [3, 7]) will notice that $\psi_{\Omega'}$ is a "local difference" component that belongs to the detail space between two levels in the BSP tree, a "low resolution" level associated with Ω and a "high resolution" level associated with Ω'. Also, these wavelets have what may be regarded as the "zero moments" property; i.e., if f is locally a polynomial over Ω, then we get $Q_{\Omega'} = Q_\Omega = f$ and $\psi_{\Omega'} = 0$. However, the BSP method is highly nonlinear; both the partition and the geometric wavelets are so dependent on the function f that one cannot expect some of the familiar properties of wavelets like a two-scale relation, a partition of unity, or spanning of some a priori given spaces.

Our modified BSP algorithm proceeds as follows. We apply the BSP algorithm and create a "full" BSP tree $\mathcal{P}$. Obviously, in applications, the subdivision process is terminated when the leaves of the tree are subdomains of sufficiently small volume, or equivalently, in image processing, when the subdomains contain only a few pixels. We shall see that under certain mild conditions on the partition $\mathcal{P}$ and the function f we have

$$f = \sum_{\Omega \in \mathcal{P}} \psi_\Omega(f) \quad \text{a.e. in } [0,1]^d,$$

where

$$\psi_{[0,1]^d} := \psi_{[0,1]^d}(f) := \mathbf{1}_{[0,1]^d} Q_{[0,1]^d}.$$

We then compute all the geometric wavelets (1.1) and sort them according to their L_p norms, i.e.,

$$\|\psi_{\Omega_{k_1}}\|_p \geq \|\psi_{\Omega_{k_2}}\|_p \geq \|\psi_{\Omega_{k_3}}\|_p \cdots . \tag{1.2}$$

Given an integer $n \in \mathbb{N}$, we approximate f by the n-term geometric wavelet sum

$$\sum_{j=1}^{n} \psi_{\Omega_{k_j}}. \tag{1.3}$$

The sum (1.3) is, in some sense, a generalization of the classical n-term wavelet approximation (see [7] and references therein), where the wavelets are constructed over dyadic cubes.

A key observation is that the BSP algorithm described above is a *geometric greedy* algorithm. At each stage of the algorithm we try to find a locally optimal partition of a given subdomain. Indeed, the problem of finding an optimal triangulation or partition is associated with an NP-hard problem (see the discussion in [6, section 4] and references therein).

It is known in classical wavelet theory (see, e.g., [7]) that the energy of the wavelet basis coefficients in some l_τ-norm, $0 < \tau < p$, is a valid gauge for the "sparseness" of the wavelet representation of the given function. We follow this idea, extending it to our geometric wavelet setup. Thus we take as a reasonable benchmark by which to measure the efficiency of the greedy algorithm, a BSP partition that "almost" minimizes, over all possible partitions, the sum of energies of the geometric wavelets of a given function, namely,

$$\left(\sum_{\Omega\in\mathcal{P}}\|\psi_\Omega\|_p^\tau\right)^{1/\tau}, \tag{1.4}$$

for some $0<\tau<p$.

We note the following geometric suboptimality of the BSP algorithm (see [12, 25] and references therein). We say that a BSP for n disjoint objects in a given convex domain is a recursive dissection of the domain into convex regions such that each object (or part of an object) is in a distinct region. Ideally, every object should be in one convex region, but sometimes it is inevitable that some of the objects are dissected. The size of the BSP is defined as the number of leaves in the resulting BSP tree.

It can be shown that for a collection of n disjoint line segments in the plane, there exists a BSP of complexity $O(n\log n)$. Recently, Tóth [24] showed a lower bound of $\Omega(n\log n/\log\log n)$, meaning that for $d=2$, in the worst case, the BSP algorithm might need slightly more elements to "capture" arbitrary linear geometry. In higher dimensions, the performance of the BSP in the worst case decreases. For example, the known lower bound for the BSP of a collection of n disjoint rectangles in $\mathbb{R}^3$ is $\Omega(n^2)$.

The paper is organized as follows. In section 2, we outline the algorithmic aspects of the geometric wavelet approach so that the reader who is less interested in the rigorous mathematics may skip section 3 and proceed directly to section 4. In section 3, we review the more theoretical aspects of our approach, and we provide some details on the approximation spaces that are associated with the method. It is interesting to note that, while the approximation spaces corresponding to nonlinear n-term wavelet approximation are linear Besov spaces (see [7] for details), the adaptive nature of the geometric wavelets implies that the corresponding approximation spaces are nonlinear. Nevertheless, it turns out that the problem at hand is "tamed" enough so as to enable the application of the classical machinery of the Jackson and Bernstein inequalities (see, e.g., [7]). Specifically, the analysis can be carried out because we are adaptively selecting one nested fixed partition for a given function, from which we select n-term geometric wavelets for any n. (In contrast, general adaptive piecewise polynomial n-term approximation [6] allows for each n, the selection of any n pieces, with no assumptions that they are taken from a fixed partition.) We conclude the paper with section 4, where we provide some numerical examples of n-term geometric wavelet approximation of digital images and discussion of possible applications in image denoising and compression.

2. Adaptive BSP partitions and the geometric wavelet approximation algorithm. Let $\Pi_{r-1}:=\Pi_{r-1}(\mathbb{R}^d)$ denote the multivariate polynomials of total degree $r-1$ (order r) in d variables. Given a bounded domain $\Omega\subset\mathbb{R}^d$, we denote the *degree (error) of polynomial approximation* of a function $f\in L_p(\Omega)$, $0<p\le\infty$, by

$$E_{r-1}(f,\Omega)_p:=\inf_{P\in\Pi_{r-1}}\|f-P\|_{L_p(\Omega)}.$$

Recall that the greedy BSP algorithm consists of finding, at each step, an optimal dissection of some domain Ω, and computing the polynomials $Q_{\Omega'}$ and $Q_{\Omega''}$ that best approximate the target function f in the p-norm over the children $\Omega', \Omega'' \subset \Omega$. In practice, we will have a suboptimal dissection, and *near-best* approximation. Thus, we are going to assume that for each $\Omega \in \mathcal{P}$, Q_Ω is a near-best approximation, i.e.,

$$\|f - Q_\Omega\|_{L_p(\Omega)} \le CE_{r-1}(f, \Omega)_p, \tag{2.1}$$

where C is independent of f and Ω but may depend on parameters like d, r, and possibly p. We shall see in section 3 that for the purpose of analysis when $p \le 1$, we need the stronger assumption that Q_Ω is a (possibly not unique) best approximation.

Let $\mathcal{P}$ be a partition of $[0,1]^d$, and let Ω' be a child of $\Omega \in \mathcal{P}$. For $f \in L_p([0,1]^d)$, $0 < p < \infty$, we set $\psi_{\Omega'}$ as in (1.1). As noted in the introduction, the function $\psi_{\Omega'}$ in (1.1) may be regarded as a local wavelet component of the function f that corresponds to the partition $\mathcal{P}$. For $0 < \tau \le p$ we denote the τ-energy of the sequence of geometric wavelets by the l_τ-norm of its L_p-norms,

$$\mathcal{N}_\tau(f, \mathcal{P}) := \left(\sum_{\Omega \in \mathcal{P}} \|\psi_\Omega\|_p^\tau \right)^{1/\tau}. \tag{2.2}$$

We will show that, under some mild conditions, the geometric wavelet expansion converges to the function. Namely, we introduce a weak constraint on the BSP partitions, which allows the analysis below to be carried out (see, for example, the proof of Theorem 3.5 below). We say that $\mathcal{P}$ is in $BSP(\rho)$, $3/4 < \rho < 1$, if for any child Ω' of Ω we have

$$|\Omega'| \le \rho |\Omega|, \tag{2.3}$$

where $|V|$ denotes the volume of a bounded set $V \subset \mathbb{R}^d$.

THEOREM 2.1. *Assume that* $\mathcal{N}_\tau(f, \mathcal{P}) < \infty$*, for some* $f \in L_p([0,1]^d)$, $0 < p < \infty$, $0 < \tau < p$*, and* $\mathcal{P} \in \mathrm{BSP}(\rho)$*. Then*

1. $f = \sum_\Omega \psi_\Omega$*, absolutely, a.e. in* $[0,1]^d$,
2. $\|f\|_p \le C(d, r, p, \tau, \rho) \mathcal{N}_\tau(f, \mathcal{P})$.

Proof. The proof is almost identical to the proof of [13, Theorem 2.17], except that here we take $\eta = p$, and we replace [13, Lemma 2.7] by Lemma 2.4 below. □

Thus, it is expedient to look for partitions (and τ) that yield finite energy or, better still, that minimize the energy. Obviously, this is not always possible or it may be too costly, and we are willing to settle for somewhat less. To this end, we define the following.

DEFINITION 2.2. *For* $f \in L_p([0,1]^d)$ *and* $0 < \tau < p < \infty$*, we say that* $\mathcal{P}_\tau(f) \in \mathrm{BSP}(\rho)$ *is a* near-best partition *if*

$$\mathcal{N}_\tau(f, \mathcal{P}_\tau(f)) \le C \inf_{\mathcal{P} \in \mathrm{BSP}(\rho)} \mathcal{N}_\tau(f, \mathcal{P}). \tag{2.4}$$

Let $\mathcal{P}_D$ be the BSP partition that gives the classical subdivision of $[0,1]^d$ into dyadic cubes. This can be done, for example, in the case $d = 2$ by partitioning $[0,1]^2$ along the line $x_1 = 1/2$ and then partitioning the two resulting rectangles along the line $x_2 = 1/2$. We get four dyadic cubes, and we proceed on each one recursively in the same manner. In section 3 we show the following relationship between $\mathcal{N}_\tau(f, \mathcal{P}_\tau(f))$ and the Besov seminorm of f (compare with the classical dyadic wavelet-type characterization of Besov spaces [10] and, in particular, the quantities $N_3(f)$ and $N_4(f)$ therein).

We will show that for $f \in L_p([0,1]^d)$, $0 < p < \infty$, $\alpha > 0$, and $1/\tau = \alpha + 1/p$, we have

$$\mathcal{N}_\tau(f, \mathcal{P}_\tau(f)) \le C\mathcal{N}_\tau(f, \mathcal{P}_D) \approx |f|_{B^{d\alpha,r}_\tau}, \tag{2.5}$$

where $B^{\gamma,r}_\tau$, $\gamma > 0$, is the classical Besov space (see Definition 3.1 below). The proof follows from the discussion beyond (3.6), and especially from (3.16).

We note that (2.2) was already defined in [16] for the special case of partitions over dyadic boxes. Also in [16], the author gives an algorithm to find the best dyadic box partition (see also [11]), thereby providing a complete solution to a restricted version of (2.4).

For $1 < p < \infty$, a more subtle but sharper definition of $\mathcal{P}_\tau(f)$ would be to define it as an "almost" minimizer of the weak ℓ_τ-norm of its corresponding geometric wavelets instead of the ℓ_τ-norm (2.2). Recall that the weak ℓ_τ-norm of a sequence $\{a_k\}$ is defined by

$$\|\{a_k\}\|_{w\ell_\tau} := \inf\{M : \#\{k : |a_k| > M\varepsilon^{1/\tau}\} \le \varepsilon^{-1}\ \forall \varepsilon > 0\}$$

and satisfies $\|\{a_k\}\|_{w\ell_\tau} \le \|\{a_k\}\|_{\ell_\tau}$. This corresponds to a well-known fact that n-term wavelet approximation can be estimated using the weaker p-norm when $1 < p < \infty$ (see [13, Theorem 3.3] for details, and see [7, Theorem 7.2.5] for the case of classic dyadic wavelets).

As we shall see, $\mathcal{N}_\tau(f, \mathcal{P})$ may serve as a "quality gauge" for partitions, when τ takes certain values strictly smaller than p. The following example demonstrates the role of τ.

Example 2.3. Let $\widetilde{\Omega} \subset [0,1]^d$ be a convex polytope, and define $f(x) := \mathbf{1}_{\widetilde{\Omega}}(x)$. Assume $\mathcal{P}$ is a partition such that for each $\Omega \in \mathcal{P}$ either $\widetilde{\Omega} \subseteq \Omega$, $\Omega \subseteq \widetilde{\Omega}$, or $\operatorname{int}(\Omega \cap \widetilde{\Omega}) = \emptyset$, where $\operatorname{int}(E)$ denotes the interior of $E \subset \mathbb{R}^d$. Then for $p = 2$ and $r = 1$ it is easy to see that

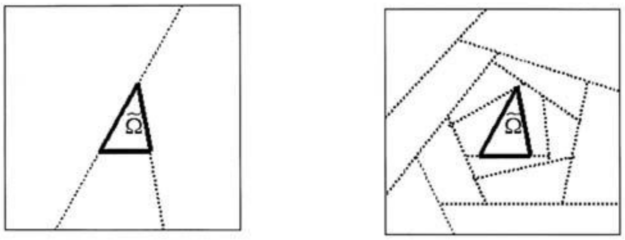
$$Q_\Omega = \begin{cases} \dfrac{|\widetilde{\Omega}|}{|\Omega|}, & \widetilde{\Omega} \subseteq \Omega, \\ 0, & \operatorname{int}(\Omega \cap \widetilde{\Omega}) = \emptyset. \end{cases}$$

Therefore we have $\psi_{[0,1]^d} = |\widetilde{\Omega}|\mathbf{1}_{[0,1]^d}$ and, for $\Omega, \Omega' \in \mathcal{P}$ with Ω' a child of Ω,

$$\|\psi_{\Omega'}\|_2^\tau = \|Q_{\Omega'} - Q_\Omega\|_{L_2(\Omega')}^\tau = \begin{cases} |\widetilde{\Omega}|^\tau \left(\dfrac{1}{|\Omega'|} - \dfrac{1}{|\Omega|}\right)^\tau |\Omega'|^{\tau/2}, & \widetilde{\Omega} \subseteq \Omega', \\ |\widetilde{\Omega}|^\tau \dfrac{1}{|\Omega|^\tau} |\Omega'|^{\tau/2}, & \operatorname{int}(\widetilde{\Omega}) \subset \Omega \setminus \Omega', \\ 0, & \operatorname{int}(\Omega \cap \widetilde{\Omega}) = \emptyset \text{ or } \Omega \subseteq \widetilde{\Omega}. \end{cases}$$

Thus, the energy of the geometric wavelets is given by the formal sum

$$\begin{aligned} \mathcal{N}_\tau^\tau(f, \mathcal{P}) &= \sum_{\Omega \in \mathcal{P}} \|\psi_\Omega\|_2^\tau \\ &= |\widetilde{\Omega}|^\tau \left(1 + \sum_{\substack{\widetilde{\Omega} \subseteq \Omega' \\ \Omega' \text{ child of } \Omega}} \left(\frac{1}{|\Omega'|} - \frac{1}{|\Omega|}\right)^\tau |\Omega'|^{\tau/2} + \frac{1}{|\Omega|^\tau}(|\Omega| - |\Omega'|)^{\tau/2}\right). \end{aligned} \tag{2.6}$$



$\mathcal{P}^{(1)}$: "Good" BSP

$\mathcal{P}^{(2)}$: "Bad" BSP

FIG. 1. *Two BSPs with* $\mathcal{N}_2(f,\mathcal{P}^{(1)}) = \mathcal{N}_2(f,\mathcal{P}^{(2)}) = \|f\|_2$.

The above sum converges, for example, if $\mathcal{P}$ is in BSP(ρ), for some $\rho < 1$. In the special case $\tau = 2$ we get

$$\begin{aligned}\mathcal{N}_2^2(f,\mathcal{P}) &= |\widetilde{\Omega}|^2\left(1 + \sum_{\substack{\widetilde{\Omega}\subseteq\Omega' \\ \Omega' \text{ child of } \Omega}} \left(\frac{1}{|\Omega'|} - \frac{1}{|\Omega|}\right)^2 |\Omega'| + \frac{1}{|\Omega|^2}(|\Omega| - |\Omega'|)\right)\\ &= |\widetilde{\Omega}|^2\left(1 + \sum_{\substack{\widetilde{\Omega}\subseteq\Omega' \\ \Omega' \text{ child of } \Omega}} \left(\frac{1}{|\Omega'|} - \frac{1}{|\Omega|}\right)\right)\\ &= |\widetilde{\Omega}|,\end{aligned}$$

which implies that $\mathcal{N}_2(f,\mathcal{P}) = \|f\|_2$. Since this equality holds for any partition that satisfies the above conditions, it follows that $\mathcal{N}_2(f,\mathcal{P})$ is not a good sparsity gauge for adaptive partitions when $p = 2$.

Referring to Figure 1, we see that the partition $\mathcal{P}^{(1)}$ is optimal since its BSP lines coincide with the hyperplanes that describe $\partial\widetilde{\Omega}$, while $\mathcal{P}^{(2)}$ contains "unnecessary" subdomains. Nevertheless, the equality $\mathcal{N}_2(f,\mathcal{P}^{(1)}) = \mathcal{N}_2(f,\mathcal{P}^{(2)}) = \|f\|_2$ holds. However, things change dramatically when we choose a sufficiently small τ. In this case, the ℓ_τ-norm serves almost as a counting measure, and since the sum (2.6) contains significantly fewer nonzero elements in the case of $\mathcal{P}^{(1)}$, we obtain that $\mathcal{N}_\tau(f,\mathcal{P}^{(1)})$ is much smaller than $\mathcal{N}_\tau(f,\mathcal{P}^{(2)})$.

Thus, we wish to address the issue of the expected range of the parameter τ for digital images and $p = 2$. If the image contains a curve singularity that is not a straight line, then the theory of section 3 below suggests that we should take $\tau \geq 2/5$. Since, in a way, dyadic wavelets are a special case of geometric wavelets, we can obtain an upper bound estimate on τ using the ideas of [8]. One needs to compute the discrete dyadic wavelet transform of the image and then compute the rate of convergence of the n-term wavelet approximation, by fitting the error function with the exponent $e(f,n) := C(f)n^{-\alpha(f)}$. Since we expect geometric wavelets to perform at least at the rate of dyadic wavelets, we should take $\tau \leq 2/(2\alpha(f)+1)$.

Going back to the greedy BSP step described in the introduction, let $(\Omega',\Omega'') \in \mathrm{BSP}(\Omega)$, and let Q_Ω, $Q_{\Omega'}$, $Q_{\Omega''}$ be the near-best polynomial approximations for their corresponding subdomains. Then we have, by (1.1),

$$\begin{aligned}&\|\psi_{\Omega'}\|_p^\tau + \|\psi_{\Omega''}\|_p^\tau\\ &\quad\leq C(\|f - Q_\Omega\|_{L_p(\Omega)}^p + \|f - Q_{\Omega'}\|_{L_p(\Omega')}^p + \|f - Q_{\Omega''}\|_{L_p(\Omega'')}^p).\end{aligned} \tag{2.7}$$

Observing that Q_Ω has been already been determined at a previous (greedy) step, we have that the local greedy optimization step of [17] will capture the geometry in which the local geometric wavelet components of f are relatively small. If we denote the levels of a BSP partition $\mathcal{P}$ of $[0,1]^d$ by $\{\mathcal{P}_m\}_{m\in\mathbb{N}}$, we say that $\Omega' \in \mathcal{P}_{m+1}$ is a *child* of $\Omega \in \mathcal{P}_m$ if $\Omega' \subset \Omega$. Then we note that our analysis also suggests that a significant improvement may be obtained if the local optimization step is carried out for several levels at once. Namely, given $\Omega \in \mathcal{P}_m$, try to minimize, for some (small) $J \geq 2$,

$$\sum_{j=1}^{J} \sum_{\substack{\tilde{\Omega}\subset\Omega \\ \tilde{\Omega}\in\mathcal{P}_{m+j}}} \|f - Q_{\tilde{\Omega}}\|^p_{L_p(\tilde{\Omega})}. \tag{2.8}$$

Finally, we return to the proof of Theorem 2.1. Condition (2.3) implies that

$$(1-\rho)|\Omega| \leq |\Omega'| \leq \rho|\Omega|. \tag{2.9}$$

This condition for BSPs corresponds to the *weak local regularity (WLR)* condition that is assumed for triangulations in [13]. Observe that a BSP still allows the polytopes of the partition to be adaptive to the geometry of the function to be approximated; i.e., the polytopes may become as thin as one may wish, so long as the "thinning" process occurs over a sequence of levels of the partition. Also, note that we have not limited the complexity of the polytopes. Indeed, polytopes at the mth level may be of complexity m.

We need the following results on norms of polynomials over convex domains.

LEMMA 2.4. *Let $P \in \Pi_{r-1}(\mathbb{R}^d)$, and let $0 < \rho < 1$ and $0 < p, q \leq \infty$.*

(a) *Assume that $\Omega', \Omega \subset \mathbb{R}^d$ are bounded convex domains such that $\Omega' \subseteq \Omega$ and $(1-\rho)|\Omega| \leq |\Omega'|$. Then*

$$\|P\|_{L_p(\Omega)} \leq C(d,r,p,\rho)\|P\|_{L_p(\Omega')}.$$

(b) *For any bounded convex domain $\Omega \subset \mathbb{R}^d$,*

$$\|P\|_{L_q(\Omega)} \approx |\Omega|^{1/q-1/p}\|P\|_{L_p(\Omega)},$$

with constants of equivalency depending only on d, r, p, and q.

(c) *If Ω' is a child of Ω in a* BSP *partition $\mathcal{P} \in \mathrm{BSP}(\rho)$, then*

$$\|P\|_{L_q(\Omega)} \approx \|P\|_{L_q(\Omega')} \approx |\Omega|^{1/q-1/p}\|P\|_{L_p(\Omega')},$$

with constants of equivalency depending only on d, r, p, q, and ρ.

Proof. The proof of (a) and (b) can be found in [5, Lemma 3.1] and the first part of the proof of [5, Lemma 3.2], respectively. Assertion (c) follows from (a) and (b), since, by the properties of $\mathcal{P}$, we have that all the domains concerned are convex, and the following equivalence of volumes holds:

$$(1-\rho)|\Omega| \leq |\Omega'| \leq (1-\rho)^{-1}|\Omega \setminus \Omega'|. \qquad \square$$

We conclude this section by outlining the steps of the adaptive geometric wavelet approximation algorithm:

1. Given $f \in L_p([0,1]^d)$, find a BSP using local steps of optimal partitions and polynomial approximations (see discussion above (2.8)).

2. For each subdomain of the partition, $\Omega \in \mathcal{P}$, compute the p-norm of the corresponding geometric wavelet ψ_Ω.

3. Sort the geometric wavelets according their energy as in (1.2). As in the case of classical dyadic wavelets, this step can be simplified by using thresholding (see [7, section 7.8]).

4. For any $n \geq 1$, construct the n-term geometric wavelet sum (1.3).

3. Theoretical aspects of the geometric wavelet approach. One of the greatest challenges in approximation theory is the characterization of adaptive multivariate piecewise polynomial approximation (see the discussion in [7, section 6.5] and [6]). Given $f \in L_p([0,1]^d)$, we wish to understand the behavior of the degree of nonlinear approximation

$$\inf_{S \in \Sigma_n^r} \|f - S\|_{L_p([0,1]^d)}, \tag{3.1}$$

where Σ_n^r is the collection $\sum_{k=1}^n \mathbf{1}_{\Omega_k} P_k$; $\{\Omega_k\}$ are convex polytopes with disjoint interiors such that $\bigcup_{k=1}^n \Omega_k = [0,1]^d$; and $P_k \in \Pi_{r-1}$, $1 \leq k \leq n$. Usually $\{\Omega_k\}$ are assumed to be simplices (triangles in the bivariate case), so as to keep their complexity bounded. However, when using the BSP approach, the polytopes $\{\Omega_k\}$ can be of arbitrary complexity, and descendant polytopes are contained in their ancestors.

In the univariate case there is a certain equivalence between the two n-term approximation methods, wavelets and piecewise polynomials. Namely, the approximation spaces associated with the two methods are characterized by the same Besov spaces [7]. The advantage of wavelet approximation over piecewise polynomial approximation is the simplicity and efficiency with which one can implement it. When $d \geq 2$, these two methods are no longer equivalent. Wavelet approximation is still characterized by the (linear) Besov spaces, while the approximation spaces associated with piecewise polynomials are known to be nonlinear spaces [6], and their characterization remains an open problem.

While the geometric wavelet algorithm of section 2 is highly adaptive and geometrically flexible, it is nothing but a "tamed" version of the piecewise polynomial method (see also discussion in [13]). To explain this, for a given BSP partition $\mathcal{P}$, denote by $\Sigma_n^r(\mathcal{P})$ the collection

$$\sum_{k=1}^n \mathbf{1}_{\Omega_k} P_k, \quad \Omega_k \in \mathcal{P}, \quad P_k \in \Pi_{r-1}, \quad 1 \leq k \leq n. \tag{3.2}$$

Observe that the n-term geometric wavelet sum (1.3) is in $\Sigma_n^r(\mathcal{P})$, for the given partition $\mathcal{P}$. Let $\mathcal{P}_\tau(f) \in \mathrm{BSP}(\rho)$ be the near-best partition of Definition 2.2 for $f \in L_p([0,1]^d)$, $0 < \tau < p$. Then, the degree of nonlinear approximation from the near-best partition is given by

$$\sigma_{n,r,\tau}(f)_p := \inf_{S \in \Sigma_n^r(\mathcal{P}_\tau(f))} \|f - S\|_p. \tag{3.3}$$

We see that the main difference between (3.1) and (3.3) is that in the latter the n-term approximations are taken from a fixed partition. This is a major advantage, as one of the main difficulties one encounters when trying to analyze the degree of approximation of n-term piecewise polynomial approximation (where the supports have disjoint interiors) is that for $S_1, S_2 \in \Sigma_n^r$ we may have, in the worst case, that $S_1 + S_2$ is of complexity $O(n^d)$, that is, supported on n^d domains with disjoint interiors.

On the other hand, if we have a fixed partition $\mathcal{P}$, and two piecewise polynomials $S_1, S_2 \in \Sigma_n^r(\mathcal{P})$, then $S_1 + S_2 \in \Sigma_{2n}^r(\mathcal{P})$. Still, even for a fixed partition, it is hard to find a solution to (3.3). As we demonstrate below, a good method for computing an n-term piecewise polynomial approximation is to take the n-term geometric wavelet sum (1.3) (see the proof of Theorem 3.6).

The goal of this section is to provide some characterization of the adaptive geometric wavelet approximation, where the n-terms are taken from a near-best adaptive partition $\mathcal{P}_\tau(f)$, which we consider as a benchmark to any of the greedy algorithms discussed above. To this end we denote by $A_{q,\tau}^{\gamma,r}(L_p)$, $\gamma > 0$, $0 < q \le \infty$, $0 < \tau < p$, the *approximation space* corresponding to nonlinear approximation from $\mathcal{P}_\tau(f)$. This is the collection of all functions $f \in L_p([0,1]^d)$ for which the error (3.3) roughly "decays" at the rate $n^{-\gamma}$, i.e., $f \in L_p([0,1]^d)$ for which

$$(f)_{A_{q,\tau}^{\gamma,r}(L_p)} := \begin{cases} \left(\sum_{m=0}^{\infty} (2^{m\gamma} \sigma_{2^m,r,\tau}(f)_p)^q \right)^{1/q}, & 0 < q < \infty, \\ \sup_{m \ge 0} (2^{m\gamma} \sigma_{2^m,r,\tau}(f)_p), & q = \infty, \end{cases}$$

is finite.

Recall that for $f \in L_\tau(\Omega)$, $0 < \tau \le \infty$, $h \in \mathbb{R}^d$, and $r \in \mathrm{N}$, we denote the rth order difference operator by

$$\Delta_h^r(f, x) := \Delta_h^r(f, \Omega, x) := \begin{cases} \sum_{k=0}^{r} (-1)^{r+k} \binom{r}{k} f(x + kh), & [x, x + rh] \subset \Omega, \\ 0, & \text{otherwise}, \end{cases}$$

where $[x, y]$ denotes the line segment connecting the points $x, y \in \mathbb{R}^d$. The *modulus of smoothness of order* r of $f \in L_\tau(\Omega)$ (see, e.g., [7, 9]) is defined by

$$\omega_r(f, t)_{L_\tau(\Omega)} := \sup_{|h| \le t} \|\Delta_h^r(f, \Omega, \cdot)\|_{L_\tau(\Omega)}, \quad t > 0,$$

where for $h \in \mathbb{R}^d$, $|h|$ denotes the length of h. We also define

$$\omega_r(f, \Omega)_\tau := \omega_r(f, \operatorname{diam}(\Omega))_{L_\tau(\Omega)}. \tag{3.4}$$

Definition 3.1. *For $\gamma > 0$, $\tau > 0$, and $r \in N$, the* Besov space $B_\tau^{\gamma,r}$ *is the collection of functions $f \in L_\tau([0,1]^d)$ for which*

$$|f|_{B_\tau^{\gamma,r}} := \left(\sum_{m=0}^{\infty} \left(2^{\gamma m} \omega_r \left(f, 2^{-m} \right)_{L_\tau([0,1]^d)} \right)^\tau \right)^{1/\tau} < \infty.$$

Definition 3.2. *For $0 < p < \infty$, $\alpha > 0$, $\rho > 0$, and $1/\tau := \alpha + 1/p$, we define the* geometric B-space $\mathcal{GB}_\tau^{\alpha,r}$, *$r \in N$, as the set of functions $f \in L_p([0,1]^d)$ for which*

$$(f)_{\mathcal{GB}_\tau^{\alpha,r}} := \left(\inf_{\mathcal{P} \in \mathrm{BSP}(\rho)} \sum_{\Omega \in \mathcal{P}} (|\Omega|^{-\alpha} \omega_r(f, \Omega)_\tau)^\tau \right)^{1/\tau} < \infty. \tag{3.5}$$

Note that the smoothness measure $(\cdot)_{\mathcal{GB}_\tau^{\alpha,r}}$ is not a (quasi-)seminorm, since the triangle inequality, in general, is not satisfied. However, it is easy to show that for $\alpha_1 \le \alpha_2$ and $1/\tau_k = \alpha_k + 1/p$, $k = 1, 2$, we have $\mathcal{GB}_{\tau_2}^{\alpha_2,r} \subseteq \mathcal{GB}_{\tau_1}^{\alpha_1,r}$, so just as in the case of Besov spaces, a larger α implies a smaller class of functions with "more smoothness." Also, the smoothness measure $(\cdot)_{\mathcal{GB}_\tau^{\alpha,r}}$ of a function is bounded by the Besov (quasi-)seminorm of the function in $B_\tau^{d\alpha,r}$. Indeed, let $\mathcal{P}_D$ denote the BSP partition that gives the classical dyadic partition. If we denote the collection of dyadic cubes of side length 2^{-m} by $\mathcal{D}_m$, then

$$\begin{aligned}(f)_{\mathcal{GB}_\tau^{\alpha,r}} &\le \left(\sum_{\Omega\in\mathcal{P}_D} (|\Omega|^{-\alpha}\omega_r(f,\Omega)_\tau)^\tau\right)^{1/\tau} \\ &\le C\left(\sum_{m=0}^{\infty}\sum_{I\in\mathcal{D}_m} (2^{d\alpha m}\omega_r(f,I)_\tau)^\tau\right)^{1/\tau} \\ &\le C|f|_{B_\tau^{d\alpha,r}}.\end{aligned} \tag{3.6}$$

For a geometric B-space $\mathcal{GB}$ we introduce the (nonlinear) *K-functional* corresponding to the pair L_p and $\mathcal{GB}$

$$K(f,t) := K(f,t,L_p,\mathcal{GB}) := \inf_{g\in\mathcal{GB}}\{\|f-g\|_p + t\cdot(g)_{\mathcal{GB}}\}, \quad t>0. \tag{3.7}$$

The (nonlinear) *interpolation space* $(L_p,\mathcal{GB})_{\lambda,q}$, $\lambda>0$, $0<q\le\infty$, is defined as the set of all $f\in L_p([0,1]^d)$ such that

$$(f)_{(L_p,\mathcal{GB})_{\lambda,q}} := \begin{cases}\left(\sum_{m=0}^{\infty}(2^{m\lambda}K(f,2^{-m}))^q\right)^{1/q}, & 0<q<\infty,\\ \sup_{m\ge 0} 2^{m\lambda}K(f,2^{-m}), & q=\infty,\end{cases}$$

is finite. Although the interpolation spaces $(L_p,\mathcal{GB})_{\lambda,q}$ are nonlinear, we can still apply the Jackson and Bernstein machinery that one usually applies in the case of linear spaces defined over fixed geometry, such as dyadic partitions [7] or fixed triangulations [13, 5]. We obtain the following characterization.

THEOREM 3.3. *Let* $0<\gamma<\alpha$, $0<q\le\infty$, *and* $0<p<\infty$; *then*

$$A_{q,\tau}^{\gamma,r}(L_p) = (L_p,\mathcal{GB}_\tau^{\alpha,r})_{\frac{\gamma}{\alpha},q}, \tag{3.8}$$

where $1/\tau := \alpha + 1/p$.

The remainder of this section is devoted to the proof of Theorem 3.3.

In [5] we proved that for all bounded convex domains $\Omega\subset\mathbb{R}^d$ and functions $f\in L_\tau(\Omega)$, $0<\tau\le\infty$, we have the equivalence

$$E_{r-1}(f,\Omega)_\tau \approx \omega_r(f,\Omega)_\tau, \tag{3.9}$$

where the constants of equivalency depend only on d, r, and τ.

To proceed with our analysis, we have to show that the polynomial approximations Q_Ω in (2.1), which are near-best approximations in the p-norm, are also near-best approximations for some $0<\eta<p$. Indeed we show the following.

LEMMA 3.4. *Let $\Omega \subset \mathbb{R}^d$ be a bounded convex domain and let $f \in L_p(\Omega)$, $0 < p < \infty$. Then for any $r \in \mathrm{N}$ there exists a polynomial $Q \in \Pi_{r-1}$ such that for all $0 < \eta \le p$ if $0 < p \le 1$, and for all $1 \le \eta \le p$ if $1 < p < \infty$, we have*

$$\|f - Q\|_{L_\eta(\Omega)} \le C E_{r-1}(f, \Omega)_\eta, \tag{3.10}$$

where for $1 < p < \infty$, $C = C(r, d)$, and for $0 < p \le 1$, $C = C(r, d, \eta) \le C(r, d, \eta_0)$, $\eta_0 \le \eta \le p$.

Proof. We begin with the case $1 < p < \infty$. Given a convex domain $\Omega \subset \mathbb{R}^d$, in [4] we have constructed for any $g \in C^r(\Omega)$ a near-best polynomial $\widetilde{Q} \in \Pi_{r-1}$ such that

$$\|g - \widetilde{Q}\|_{L_\eta(\Omega)} \le C(r, d) E_{r-1}(g, \Omega)_\eta, \quad 1 \le \eta < \infty. \tag{3.11}$$

Let $f \in L_p(\Omega)$, and let $\{g_n\}$ be a sequence in $C^r(\Omega)$ such that $\|f - g_n\|_p \to 0$ as $n \to \infty$. By Hölder's inequality, it follows that for all $1 \le \eta \le p$, $\|f - g_n\|_\eta \to 0$ as $n \to \infty$. Now let Q_n be the near-best approximation to g_n guaranteed by (3.11). Then $\|g_n - Q_n\|_p \le C(r, d)\|g\|_p$, and since we may assume that $\|f - g_n\|_p \le \|f\|_p$, we obtain

$$\|Q_n\|_\infty \le C(r, d)|\Omega|^{-1/p}\|Q_n\|_p \le C(r, d)|\Omega|^{-1/p}\|f\|_p.$$

Hence, the set of polynomials Q_n is compact in $C(\Omega)$, and we may assume that $\{Q_n\}$ converges in the uniform norm to a polynomial Q. Now

$$\|f - Q\|_\eta \le \|f - g_n\|_\eta + \|g_n - Q_n\|_\eta + \|Q_n - Q\|_\eta, \quad 1 \le \eta \le p,$$

whence

$$\|f - Q\|_\eta \le \lim_{n \to \infty} C(r, d) E_{r-1}(g_n, \Omega)_\eta = C(r, d) E_{r-1}(f, \Omega)_\eta, \quad 1 \le \eta \le p.$$

This proves (3.10) for $1 < p < \infty$.

For the case $0 < p \le 1$, we first make the following observation. Let A be a nonsingular affine mapping on $\mathbb{R}^d$, given by $A(x) := Mx + b$, where M is a nonsingular $d \times d$ matrix, and let $f \in L_p(\Omega)$. Define $\tilde{f} := f(A\cdot)$, $\tilde{Q} := Q(A\cdot)$, and $\widetilde{\Omega} := A^{-1}\Omega$. Then $\tilde{f} \in L_p(\widetilde{\Omega})$, and

$$\|f - Q\|_{L_\eta(\Omega)} = |\det M|^{1/\eta} \|\tilde{f} - \tilde{Q}\|_{L_\eta(\widetilde{\Omega})}, \quad 0 < \eta \le p. \tag{3.12}$$

Therefore

$$E_{r-1}(f, \Omega)_\eta = |\det M|^{1/\eta} E_{r-1}(\tilde{f}, \widetilde{\Omega})_\eta, \quad 0 < \eta \le p. \tag{3.13}$$

By John's theorem (see [4, 5] and references therein), for any bounded convex domain $\Omega \subset \mathbb{R}^d$ there exists a nonsingular affine mapping A such that

$$B(0, 1) \subseteq \widetilde{\Omega} \subseteq B(0, d), \tag{3.14}$$

where $B(x_0, R)$ denotes the ball of radius R with center at x_0. Then we follow [1] (see also [9, Theorem 3.10.4]), and for $\tilde{f} \in L_p(\widetilde{\Omega})$ obtain $\tilde{Q} \in \Pi_{r-1}$, a so-called polynomial of best approximation in $L_1(\widetilde{\Omega})$, which satisfies

$$\|\tilde{f} - \tilde{Q}\|_{L_\eta(\widetilde{\Omega})} \le C(r, d, \eta) E_{r-1}(\tilde{f}, \widetilde{\Omega})_\eta, \quad \eta \le 1, \tag{3.15}$$

where $C(r, d, \eta) \le C(r, d, \eta_0)$, $\eta_0 < \eta \le p$. Now, (3.10) for $0 < p \le 1$ follows by virtue of (3.12) and (3.13). $\square$

THEOREM 3.5. *For $0 < p < \infty$, $\alpha > 0$, $1/\tau = \alpha + 1/p$, and $f \in L_p([0,1]^d)$, we have the equivalence*

$$(f)_{\mathcal{GB}_\tau^{\alpha,r}} \approx \mathcal{N}_\tau(f, \mathcal{P}_\tau(f)), \tag{3.16}$$

with constants of equivalency depending only on α, d, r, p, and ρ.

Proof. Let $\mathcal{P} \in \mathrm{BSP}(\rho)$ be a given partition. For $0 < \mu \le p$ and $\Omega \in \mathcal{P}$, denote by $Q_{\Omega,\mu}$ a near-best polynomial approximation of $f \in L_\mu(\Omega)$. Note that with this notation, the near-best polynomials used in (1.1) are $Q_\Omega = Q_{\Omega,p}$. We define

$$\mathcal{N}_{\tau,\mu}(f,\mathcal{P}) := \left(\sum_{\Omega\in\mathcal{P}} \|\psi_{\Omega,\mu}\|_p^\tau\right)^{1/\tau},$$

where $\psi_{\Omega,\mu}$ are defined in (1.1) with the near-best polynomials $Q_{\Omega,\mu}$, and

$$\widetilde{\mathcal{N}}_{\tau,\mu}(f,\mathcal{P}) := \left(\sum_{\Omega\in\mathcal{P}} (|\Omega|^{1/p-1/\mu}\omega_r(f,\Omega)_\mu)^\tau\right)^{1/\tau}.$$

By Lemma 3.4 we know that there is a $\tau < \eta < p$ such that for any $\Omega \in \mathcal{P}$ we may take $\psi_{\Omega,\eta} = \psi_{\Omega,p} = \psi_\Omega$. Therefore, in order to prove (3.16), it suffices to prove that for any $\mathcal{P} \in \mathrm{BSP}(\rho)$

$$\mathcal{N}_{\tau,\eta}(f,\mathcal{P}) \approx \widetilde{\mathcal{N}}_{\tau,\tau}(f,\mathcal{P}) \tag{3.17}$$

holds with constants of equivalency that depend only on d, r, p, τ, η, and ρ.

To this end, take $\tau \le \mu \le \eta$, and recall that if Ω' is a child of Ω, then

$$\begin{aligned} \|\psi_{\Omega',\mu}\|_\mu &\le C(\|f - Q_{\Omega,\mu}\|_{L_\mu(\Omega')} + \|f - Q_{\Omega'},\mu\|_{L_\mu(\Omega')}) \\ &\le C(E_{r-1}(f,\Omega)_\mu + E_{r-1}(f,\Omega')_\mu), \end{aligned} \tag{3.18}$$

where $C = C(r,d,\mu)$. Hence

$$\begin{aligned} \mathcal{N}_{\tau,\mu}(f,\mathcal{P}) &= \left(\sum_{\Omega\in\mathcal{P}} \|\psi_{\Omega,\mu}\|_p^\tau\right)^{1/\tau} \\ &\le C\left(\sum_{\Omega\in\mathcal{P}} (|\Omega|^{1/p-1/\mu}\|\psi_{\Omega,\mu}\|_\mu)^\tau\right)^{1/\tau} \\ &\le C\left(\sum_{\Omega\in\mathcal{P}} (|\Omega|^{1/p-1/\mu}E_{r-1}(f,\Omega)_\mu)^\tau\right)^{1/\tau} \\ &\le C\left(\sum_{\Omega\in\mathcal{P}} (|\Omega|^{1/p-1/\mu}\omega_r(f,\Omega)_\mu)^\tau\right)^{1/\tau} \\ &= C\widetilde{\mathcal{N}}_{\tau,\mu}(f,\mathcal{P}), \end{aligned} \tag{3.19}$$

where for the first inequality we applied Lemma 2.4, for the second we applied (3.18) and (2.9), and finally for the third inequality we applied (3.9).

Next we show that for $\tau \le \mu \le \eta$

$$\widetilde{\mathcal{N}}_{\tau,\eta}(f,\mathcal{P}) \le \mathcal{N}_{\tau,\mu}(f,\mathcal{P}). \tag{3.20}$$

We may assume that $\mathcal{N}_{\tau,\mu}(f,\mathcal{P}) < \infty$, because otherwise there is nothing to prove. Since $\mu < p$, we have that $f \in L_\mu([0,1]^d)$, and Theorem 2.1 implies

$$f = \sum_{\Omega\in\mathcal{P}} \psi_{\Omega,\mu} \quad \text{a.e.}$$

Therefore,

$$\begin{aligned}
\omega_r(f,\Omega)_\eta^\tau &= \omega_r\left(f - \sum_{\widetilde{\Omega}\in\mathcal{P},\,\widetilde{\Omega}\supseteq\Omega} \psi_{\widetilde{\Omega},\mu}, \Omega\right)_\eta^\tau \\
&\le C\left\|\sum_{\widetilde{\Omega}\in\mathcal{P},\,\widetilde{\Omega}\subset\Omega} |\psi_{\widetilde{\Omega},\mu}|\right\|_\eta^\tau \\
&\le C\sum_{\widetilde{\Omega}\in\mathcal{P},\,\widetilde{\Omega}\subset\Omega} \|\psi_{\widetilde{\Omega},\mu}\|_\eta^\tau \\
&\le C\sum_{\widetilde{\Omega}\in\mathcal{P},\,\widetilde{\Omega}\subset\Omega} |\widetilde{\Omega}|^{\tau(1/\eta-1/\tau)}\|\psi_{\widetilde{\Omega},\mu}\|_\tau^\tau,
\end{aligned}$$

where for the equality we used the fact that for $\Omega \subseteq \widetilde{\Omega}$ the geometric wavelet $\psi_{\widetilde{\Omega},\mu}$ is a polynomial of total degree $\le r-1$, for the second inequality we applied [13, Theorem 3.3], and for the third inequality we applied Lemma 2.4. Therefore,

$$\begin{aligned}
\widetilde{\mathcal{N}}_{\tau,\eta}(f,\mathcal{P})^\tau &\le C\sum_{\Omega\in\mathcal{P}} |\Omega|^{\tau(1/p-1/\eta)} \sum_{\widetilde{\Omega}\subset\Omega} |\widetilde{\Omega}|^{\tau(1/\eta-1/\tau)}\|\psi_{\widetilde{\Omega},\mu}\|_\tau^\tau \\
&= C\sum_{\Omega\in\mathcal{P}}\sum_{\widetilde{\Omega}\subset\Omega}\left(\frac{|\widetilde{\Omega}|}{|\Omega|}\right)^{\tau(1/\eta-1/p)} (|\widetilde{\Omega}|^{1/p-1/\tau}\|\psi_{\widetilde{\Omega},\mu}\|_\tau)^\tau \\
&= C\sum_{\widetilde{\Omega}\in\mathcal{P}} (|\widetilde{\Omega}|^{1/p-1/\tau}\|\psi_{\widetilde{\Omega},\mu}\|_\tau)^\tau \sum_{\substack{\Omega\in\mathcal{P}\\ \Omega\supset\widetilde{\Omega}}}\left(\frac{|\widetilde{\Omega}|}{|\Omega|}\right)^{\tau(1/\eta-1/p)}.
\end{aligned}$$

Now, if $\widetilde{\Omega} \in \mathcal{P}_m$ and $\Omega \in \mathcal{P}_{m-k}$, $k > 0$, is one of its ancestors, then by (2.9),

$$|\widetilde{\Omega}| \le |\Omega|\rho^k.$$

Hence

$$\sum_{\Omega\in\mathcal{P},\,\Omega\supset\widetilde{\Omega}}\left(\frac{|\widetilde{\Omega}|}{|\Omega|}\right)^{\tau(1/\eta-1/p)} \le C\sum_{k=1}^{\infty}\rho^{k\tau(1/\eta-1/p)} \le C(p,\eta,\tau,\rho).$$

We conclude that

$$\begin{aligned}
\widetilde{\mathcal{N}}_{\tau,\eta}(f,\mathcal{P})^\tau &\le C\sum_{\widetilde{\Omega}\in\mathcal{P}} (|\widetilde{\Omega}|^{1/p-1/\tau}\|\psi_{\widetilde{\Omega},\mu}\|_\tau)^\tau \\
&\le C\sum_{\widetilde{\Omega}\in\mathcal{P}} \|\psi_{\widetilde{\Omega},\mu}\|_p^\tau = C\mathcal{N}_{\tau,\mu}(f,\mathcal{P})^\tau,
\end{aligned}$$

where for the last inequality we again applied Lemma 2.4. This proves (3.20).

Now combining (3.19) with $\mu = \eta$, (3.20) with $\mu = \tau$, and then (3.19) with $\mu = \tau$, we obtain

$$\mathcal{N}_{\tau,\eta}(f,\mathcal{P}) \le C\widetilde{\mathcal{N}}_{\tau,\eta}(f,\mathcal{P}) \le C\mathcal{N}_{\tau,\tau}(f,\mathcal{P}) \le C\widetilde{\mathcal{N}}_{\tau,\tau}(f,\mathcal{P}),$$

which proves one direction in (3.17). In order to prove the opposite direction, we observe that it follows from Hölder's inequality that

$$\widetilde{\mathcal{N}}_{\tau,\tau}(f,\mathcal{P}) \le \widetilde{\mathcal{N}}_{\tau,\eta}(f,\mathcal{P}).$$

Using (3.20) with $\mu = \eta$ yields

$$\widetilde{\mathcal{N}}_{\tau,\eta}(f,\mathcal{P}) \le C\mathcal{N}_{\tau,\eta}(f,\mathcal{P}),$$

which gives

$$\widetilde{\mathcal{N}}_{\tau,\tau}(f,\mathcal{P}) \le \widetilde{\mathcal{N}}_{\tau,\eta}(f,\mathcal{P}) \le C\mathcal{N}_{\tau,\eta}(f,\mathcal{P}).$$

This completes the proof of the opposite direction in (3.17) and concludes our proof. □

In view of the above, one may draw the following conclusion. There are cases of functions that are not in the Besov space of scale $d\alpha$ and therefore cannot be approximated by n-term wavelet approximation at the "rate" $n^{-\alpha}$ (see [7]). Yet, there might exist an adaptive partition which captures the geometry (if it exists!) of the function's singularities and leads to a finite smoothness measure (3.5) for the scale α. In fact we show that such a partition can also provide n-term geometric wavelet approximation at the rate $n^{-\alpha}$.

Theorem 3.6 (Jackson estimate). *Let $0 < p < \infty$, $\alpha > 0$, and $r \in \mathrm{N}$. If $f \in \mathcal{GB}_\tau^{\alpha,r}$, $1/\tau = \alpha + 1/p$, then*

$$\sigma_{n,r,\tau}(f)_p \le Cn^{-\alpha}(f)_{\mathcal{GB}_\tau^{\alpha,r}}, \tag{3.21}$$

where $C := C(\alpha,d,r,p,\rho)$.

Proof. Given f, p, and τ, we select the near-best adaptive partition $\mathcal{P}_\tau(f)$. Applying [13, Theorem 3.4] with the collection $\{\Phi_m\} := \{\psi_\Omega\}_{\Omega\in\mathcal{P}_\tau(f)}$ and then (3.16), we obtain

$$\begin{aligned}\sigma_{n,r,\tau}(f)_p &\le Cn^{-\alpha}\mathcal{N}_\tau(f,\mathcal{P}_\tau(f))\\ &\le Cn^{-\alpha}(f)_{\mathcal{GB}_\tau^{\alpha,r}}. \qquad \square\end{aligned}$$

Let $\phi \in L_p([0,1]^d)$ and let $\mathcal{P} \in \mathrm{BSP}(\rho)$ be a *fixed* partition. Then, the smoothness of ϕ with respect to the fixed partition $\mathcal{P}$ is

$$|\phi|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{P})} := \left(\sum_{\Omega\in\mathcal{P}} (|\Omega|^{-\alpha}\omega_r(\phi,\Omega)_\tau)^\tau\right)^{1/\tau}.$$

For a fixed partition $\mathcal{P}$, the smoothness quantity $|\cdot|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{P})}$ is a quasi seminorm. Therefore we obtain the Bernstein estimate for BSPs in much the same way that it was proved for triangulations in the bivariate case in [13], and in arbitrary dimension $d \ge 2$ in [5]. Namely, we have the following.

THEOREM 3.7 (Bernstein estimate). *Let $\mathcal{P} \in \mathrm{BSP}(\rho)$, and let $\phi \in \Sigma_n^r(\mathcal{P})$. Then for all $0 < p < \infty$, $\alpha > 0$, and $1/\tau = \alpha + 1/p$,*

$$|\phi|_{\mathcal{B}_\tau^{\alpha,r}(\mathcal{P})} \le C n^\alpha \|\phi\|_p, \tag{3.22}$$

where $C := C(\alpha, d, r, p, \rho)$.

We are now ready to prove Theorem 3.3.

Proof of Theorem 3.3. The proof is similar to the proof of [9, Theorem 7.9.1]. The proof that the right-hand side of (3.8) is contained in the left-hand side readily follows by the Jackson inequality. Indeed, it is a standard technique to show that (3.21) implies that for every $f \in L_p$

$$\sigma_{n,r,\tau}(f)_p \le CK(f, n^{-\alpha}, L_p, \mathcal{GB}_\tau^{\alpha,r}).$$

Hence by the first part of the proof of [9, Theorem 7.9.1]

$$(f)_{A_{q,\tau}^{\gamma,r}} \le C\Big(\|f\|_p + (f)_{(L_p, \mathcal{GB}_\tau^{\alpha,r})_{\frac{\gamma}{\alpha},q}}\Big).$$

In order to prove that the left-hand side of (3.8) is contained in the right-hand side, we have to estimate the appropriate K-functional. Namely, we replace the proof of [9, Theorem 7.5.1(ii)] with the estimate

$$K(f, 2^{-m\alpha}, L_p, \mathcal{GB}_\tau^{\alpha,r}) \le C 2^{-m\alpha} \left(\sum_{j=1}^m (2^{j\alpha} \sigma_{2^{j-1}}(f)_p)^\mu + \|f\|_p^\mu \right)^{1/\mu}, \tag{3.23}$$

where $K(f, \cdot, L_p, \mathcal{GB}_\tau^{\alpha,r})$ is defined by (3.7), $\sigma_{2^j}(f)_p := \sigma_{2^j,r,\tau}(f)_p$, $m \ge 1$, and $\mu := \min(\tau, 1)$. Note that, in proving this, special attention is needed to circumvent the fact that $(\cdot)_{\mathcal{GB}_\tau^{\alpha,r}}$ is not a (quasi-)seminorm. Indeed, for each $j \ge 0$ we take a geometric wavelet sum $S_j \in \Sigma_{2^j}^r(\mathcal{P}_\tau(f))$ such that

$$\|f - S_j\|_{L_p([0,1]^d)} \le 2\sigma_{2^j}(f)_p.$$

Since $\mathcal{P}_\tau(f)$ is a fixed nested partition, we have that $\phi_j := S_j - S_{j-1} \in \Sigma_{2^{j+1}}^r(\mathcal{P}_\tau(f))$, $j \ge 1$, and

$$\|\phi_j\|_p \le \|f - S_j\|_p + \|f - S_{j-1}\|_p \le 2\sigma_{2^{j-1}}(f)_p, \quad j \ge 1.$$

We also set $\phi_0 := S_0$. Since S_0 is a single geometric wavelet component, we conclude that (3.9) implies that $\|\phi_0\|_p \le C\|f\|_p$. Now, we substitute $g := S_m = \sum_{j=0}^m \phi_j$ in (3.7) and apply the Bernstein inequality (3.22) on the fixed partition $\mathcal{P}_\tau(f)$ to obtain

$$\begin{aligned}
K(f, 2^{-m\alpha}, L_p, \mathcal{GB}_\tau^{\alpha,r}) &\le \|f - S_m\|_p + 2^{-m\alpha}(S_m)_{\mathcal{GB}_\tau^{\alpha,r}} \\
&\le C(\sigma_{2^m}(f)_p + 2^{-m\alpha}|S_m|_{B_\tau^{\alpha,r}(P_\tau(f))}) \\
&\le C\left(\sigma_{2^m,r}(f)_p + 2^{-m\alpha}\left(\sum_{j=0}^m |\phi_j|^\mu_{B_\tau^{\alpha,r}(\mathcal{P}_\tau(f))}\right)^{1/\mu}\right) \\
&\le C\left(\sigma_{2^m}(f)_p + 2^{-m\alpha}\left(\sum_{j=0}^m (2^{(j+1)\alpha}\|\phi_j\|_p)^\mu\right)^{1/\mu}\right) \\
&\le C 2^{-m\alpha}\left(\sum_{j=1}^m (2^{j\alpha}\sigma_{2^{j-1}}(f)_p)^\mu + \|f\|_p^\mu\right)^{1/\mu}.
\end{aligned}$$

We leave the rest of the proof to the reader. □

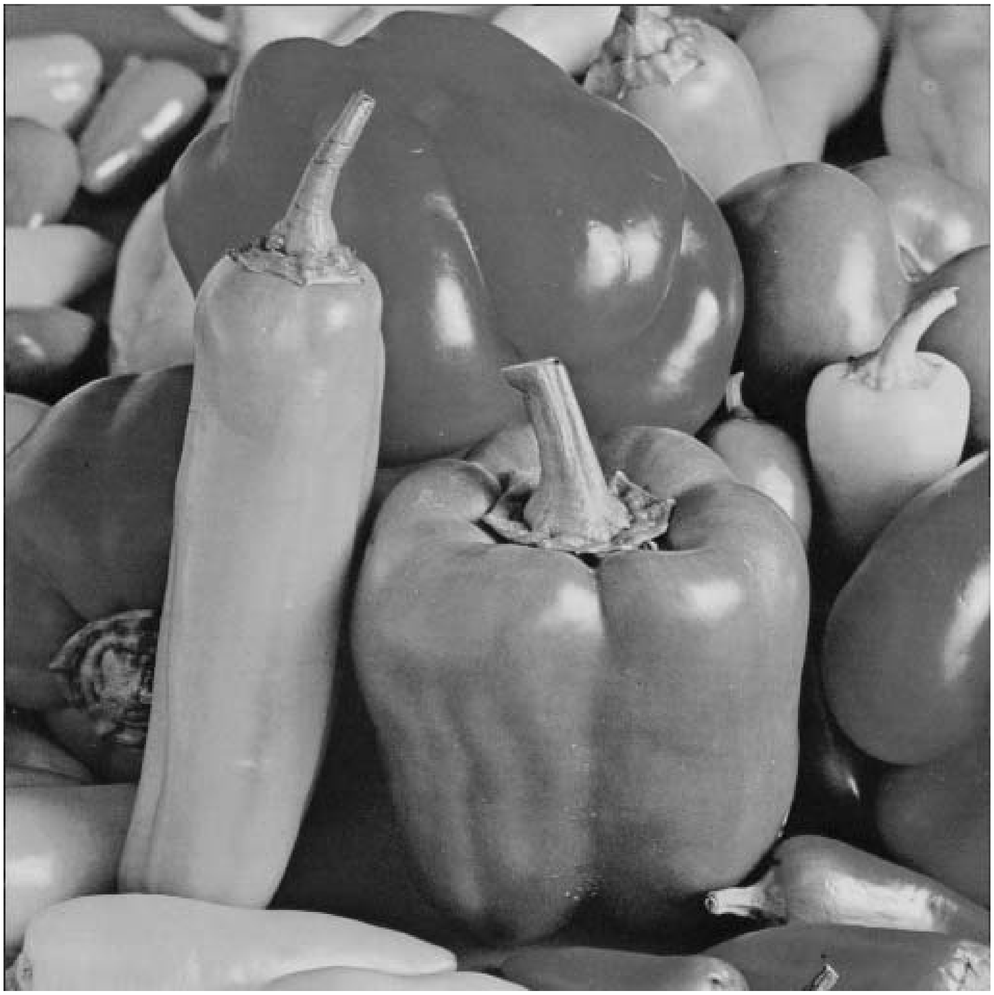


Fig. 2. *The "peppers" image* 512×512.

4. Simulation results and discussion. We implemented the geometric wavelet algorithm for the purpose of finding sparse representations of digital images with $r = 2$ (linear polynomials) and $p = 2$. We point out that, in our current implementation, condition (2.3) does not come into play.

To reduce the time complexity of the implementation, the images were subdivided into tiles of size 64×64, and a BSP tree was constructed over each of the tiles separately. Although JPEG-like artifacts, resulting from the tiles' boundaries, are visible in the examples below, this approach ensures that the time complexity of the algorithm is almost linear with respect to the image size. Once all the BSP trees were constructed over the 64×64 tiles, and the geometric wavelets were computed, we extracted a global n-term approximation (1.3) from the joint list of all the geometric wavelets over all the tiles. Our experiments show that in most cases increasing the tile size does not have a significant impact on the results.

To further improve the time complexity of the algorithm, we performed coarse

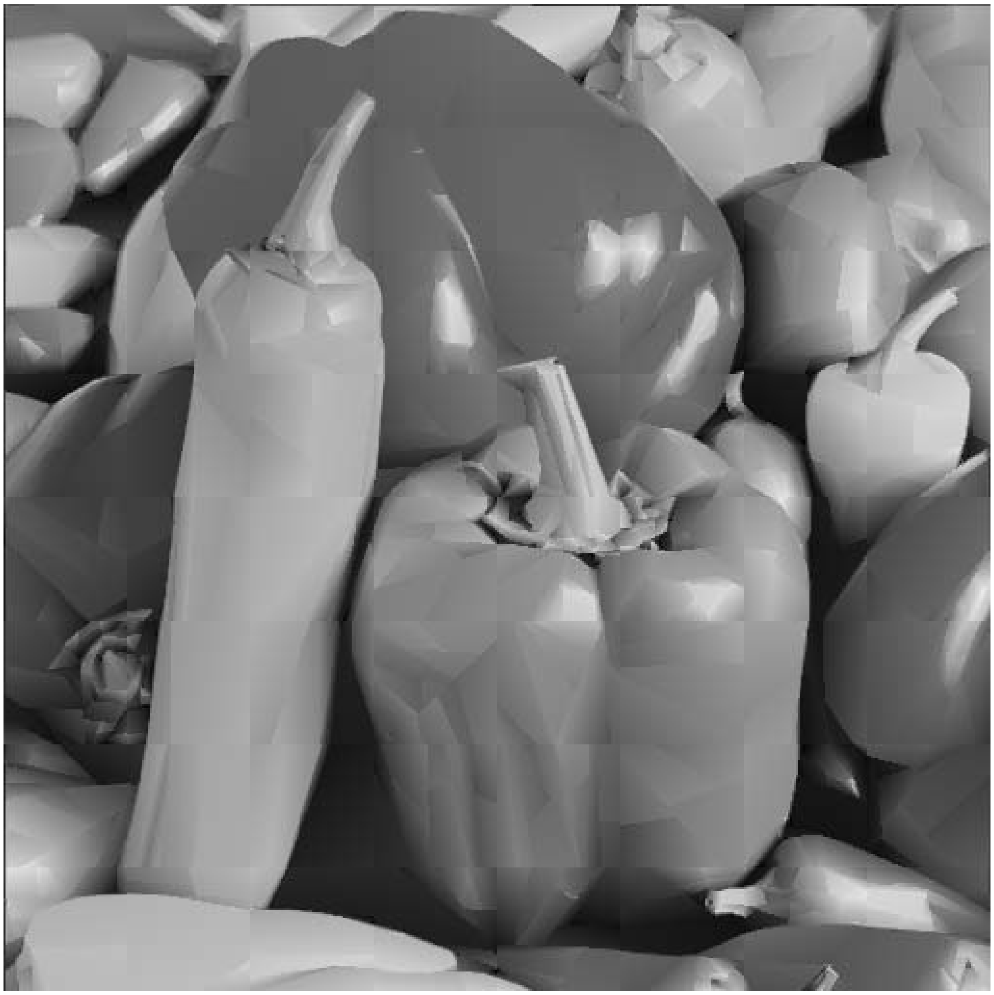


FIG. 3. *Geometric wavelet approximation of the peppers image with* $n = 2048$, PSNR $= 31.32$.

partition searches at lower levels of the BSP tree and fine searches at the higher levels. The search for the optimal partition was done by advancing two points on a domain's boundary, computing the two subdomains created by the line that goes through these points, and then computing the two least-squares linear polynomials over each of these subdomains. In lower levels of the BSP tree, this march was done in larger steps and in finer levels, and the step size was set to 1, the pixel resolution. In some sense, the idea of finer partitions at higher resolutions is related to the way curvelets [2] have "more directions" at higher resolutions.

In Figure 3 we see an n-term geometric wavelet approximation of the known test image peppers (cf. original in Figure 2) of size 512×512 with 2048 elements and PSNR (peak signal-to-noise ratio) 31.32. In Figure 4 we see an n-term dyadic wavelet approximation with twice as many elements, 4096, and still somewhat worse PSNR, 29.22. In all the examples below, we used a ratio of 1:2 (peppers, Figures 3–4; Lena, Figures 6–7), 1:3 (Barbara, Figures 13–14) or 1:4 (cameraman, Figures 9–11)

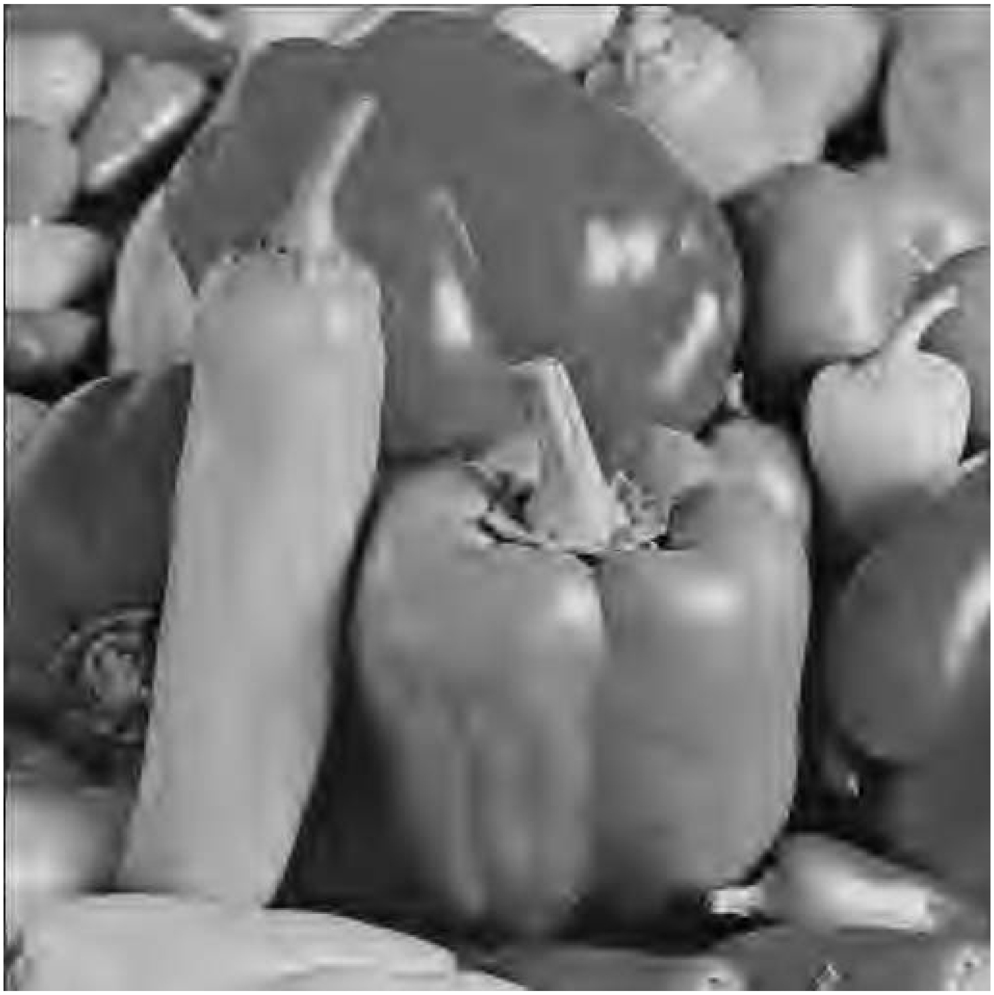


FIG. 4. *Dyadic biorthogonal wavelet approximation of the peppers image with* $n = 4096$*,* PSNR = 29.22*.*

between the number of geometric wavelets and dyadic wavelets, so as to make the comparison more relevant. Observe that on the more "geometric" images, peppers and cameraman, i.e., images that are roughly composed of smooth regions and strong distinct edges, the geometric wavelets seem to perform relatively better. For example, for the cameraman image the 512-term geometric wavelet approximation gives the same PSNR as the 2048-term dyadic wavelet approximation.

For the dyadic wavelets approximation we used the MATLAB wavelet toolbox, where we selected the well-known biorthogonal wavelet basis $(4, 4)$ (see [3]), also known as the "nine-seven" in the engineering community. This biorthogonal wavelet has four zero moments, corresponding to $r = 4$. We note that we actually allowed the dyadic wavelet approximation to use even slightly more elements than claimed in the figures, so as to compensate for MATLAB handling of the image boundaries by a somewhat overredundant wavelet decomposition. The results are summarized in Table 1.

In Figure 15 we see an example of image denoising using geometric wavelets. To

Fig. 5. *The "Lena" image* 512 × 512.

Table 1
Comparison of n-term dyadic and geometric wavelets.

Image	N-term dyadic	N-term geometric	Ratio	PSNR dyadic	PSNR geometric
peppers	4096	2048	2:1	29.22	31.32
Lena	4096	2048	2:1	30.18	31.26
cameraman	2048	512	4:1	26.72	26.71
		1024			28.93
Barbara	12288	4096	3:1	27.54	27.10

compare with results in [22], we added Gaussian white noise to the Lena test image with standard deviation of 20, which gives a noisy image with PSNR = 22.14. Following the usual "sparse representation" methodology [22], we applied the geometric wavelet algorithm to the noisy image and extracted an n-term approximation (1.3) to the original image. We see that geometric features are recovered quite well in the pro-

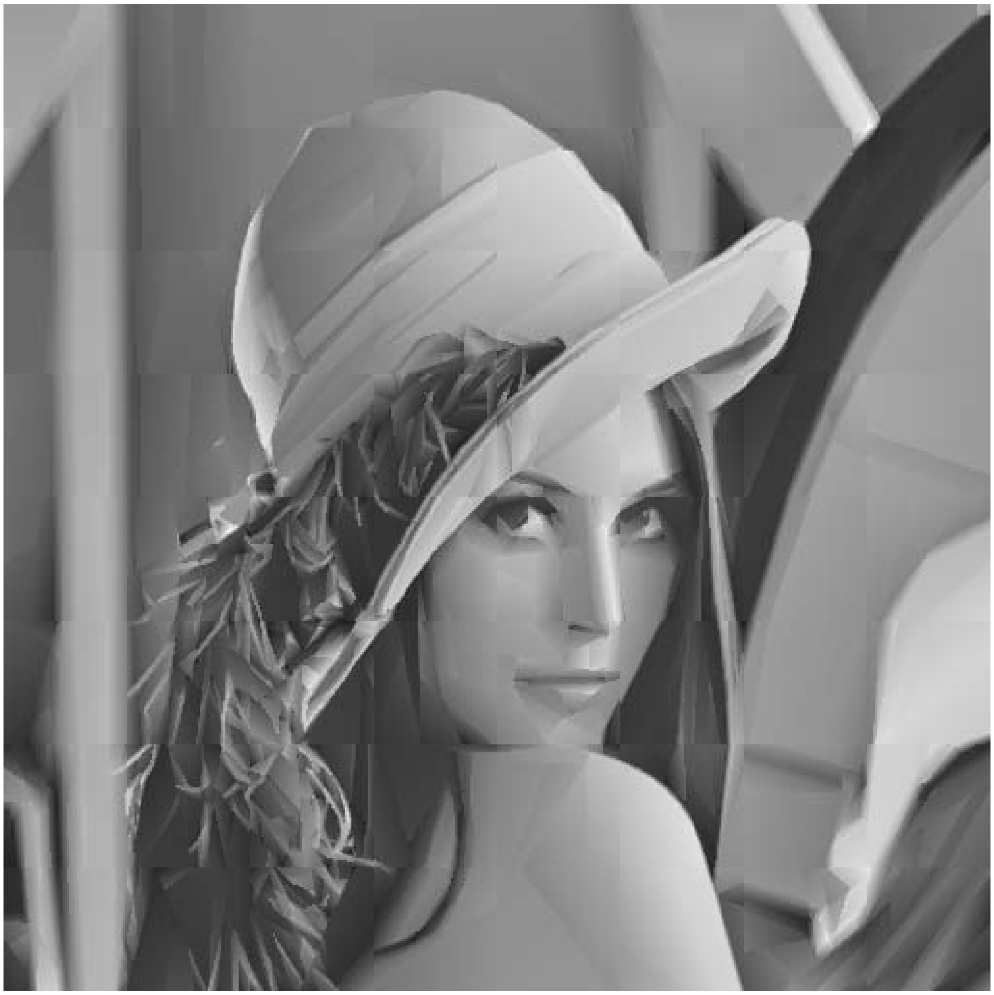


FIG. 6. *Geometric wavelet approximation of the Lena image with* $n = 2048$, PSNR $= 31.26$.

cess, in a manner which is very competitive with curvelets. The algorithm produced a restored image with PSNR $= 29.76$.

As with classical wavelets, the n-term strategy can be used for progressive coding and rate-distortion control, where more geometric wavelets are added according to their order of appearance in (1.2). It is important to note that when trying to encode the approximation (1.3) it should be remembered that for a geometric wavelet located in a "deep" level of the BSP tree, one needs to encode the sequence of binary partitions that created it. Thus, if the wavelet ψ_Ω is located at the mth level of the BSP partition, $O(m)$ bits are required to encode its location. Therefore, encoding geometric wavelets at higher levels is more expensive when considering bit allocation. However, this is no different from dyadic wavelet compression, where encoding the index of a dyadic wavelet located at the resolution m also requires $O(m)$ bits. Recall that at lower levels of the BSP tree we perform coarse partitions and at higher levels, fine partitions. As pointed out in [17], this also improves the coding performance, since it facilitates the quantization and encoding of the partitions.

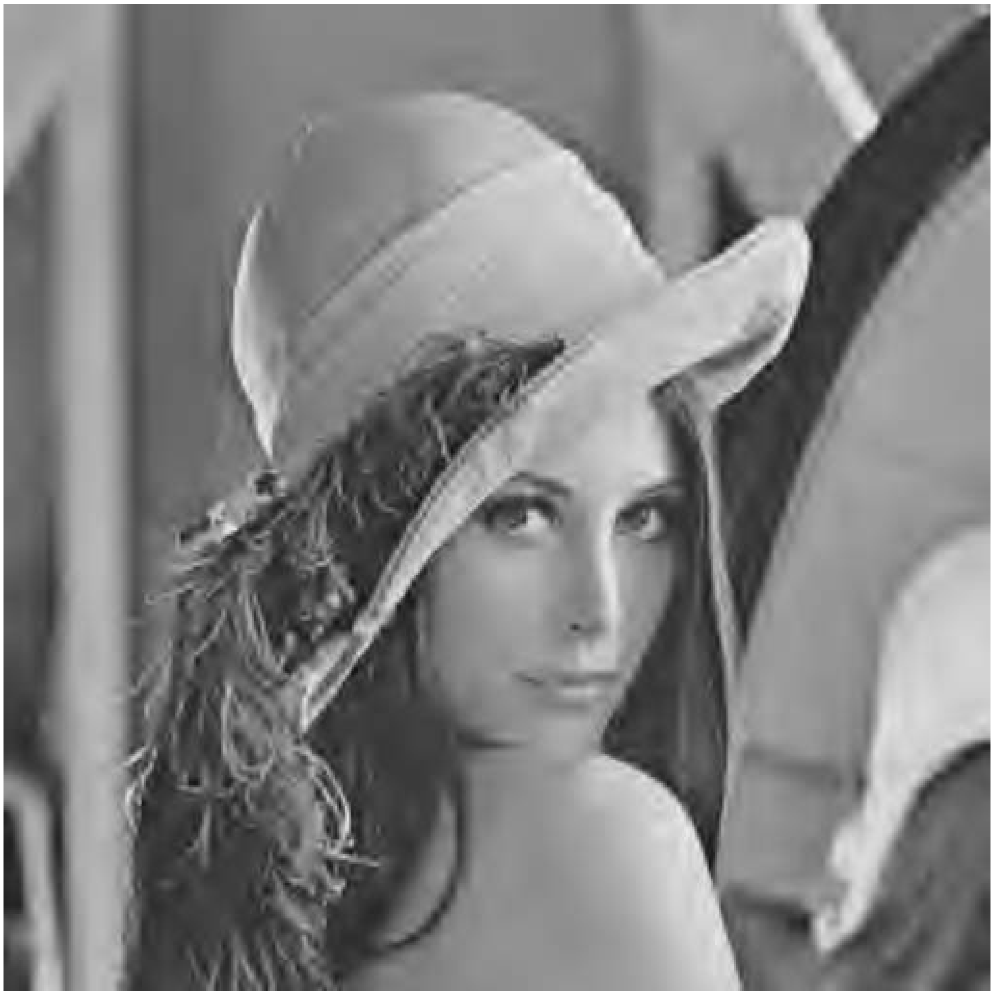


FIG. 7. *Dyadic biorthogonal wavelet approximation of the Lena image with* $n = 4096$*,* PSNR $=$ 30.18*.*

Although image coding using geometric wavelets is ongoing work, we anticipate that the problem of encoding geometric side-information can be solved by using zerotree-type encoding [18, 20] and rate-distortion optimization techniques [21, 23]. Furthermore, we plan to incorporate a geometric rate-distortion optimization technique borrowed from the wavelet coding algorithm WedgePrints [26]. Namely, at each node of the BSP tree, one may allocate a flag (bit) to signal to the decoder a decision about whether all further partitions of this domain are uniform (nonadaptive) or geometrically adaptive. Encoding geometric wavelets whose supports lie in a "uniform" ancestor domain is similar to dyadic wavelet encoding, where only an index of the geometric wavelet in a uniform partition needs to be encoded and the support of the geometric wavelet is known from the uniform partition of the ancestor. Thus, using rate-distortion optimization techniques, one would choose at each node of the BSP whether to use an adaptive partition whose geometry needs to be encoded, or a uniform nonadaptive partition.

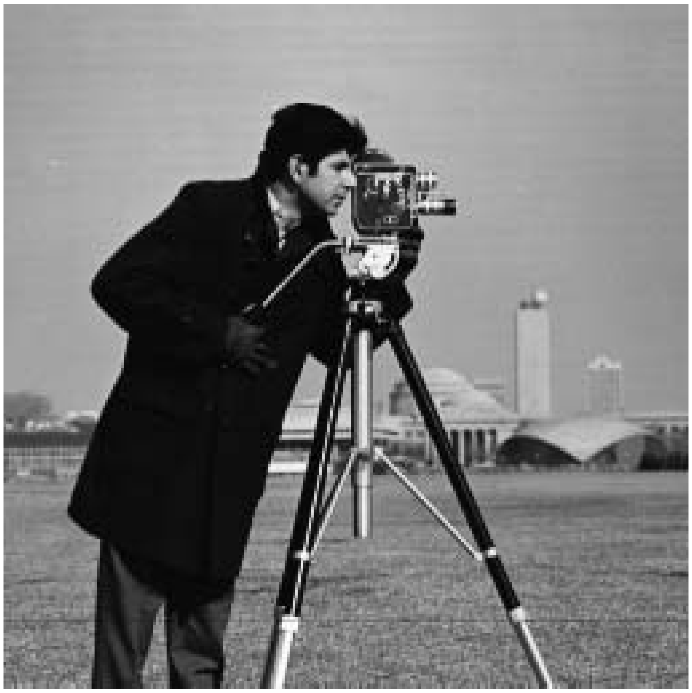


FIG. 8. *The "cameraman" image* 256×256.

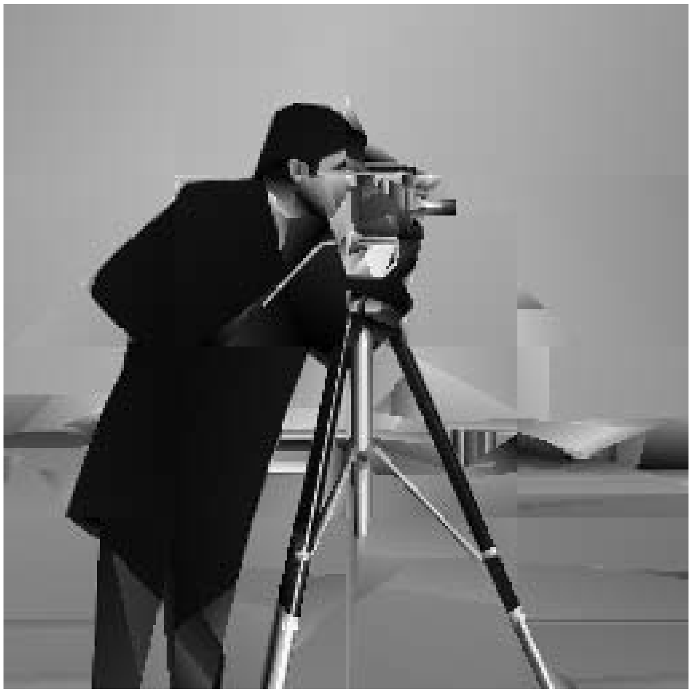


FIG. 9. *Geometric wavelet approximation of the cameraman image with* $n = 512$, $\mathrm{PSNR} = 26.71$.

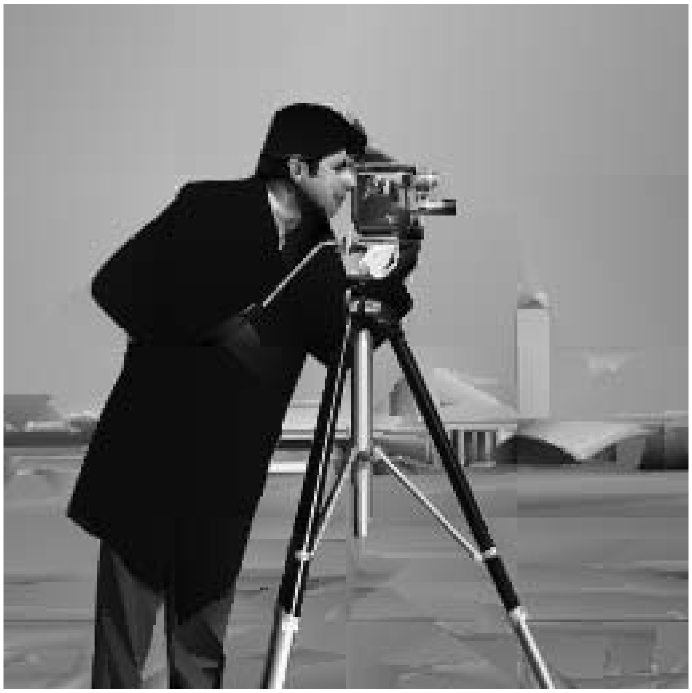


FIG. 10. *Geometric wavelet approximation of the cameraman image with* $n = 1024$, PSNR = 28.93.

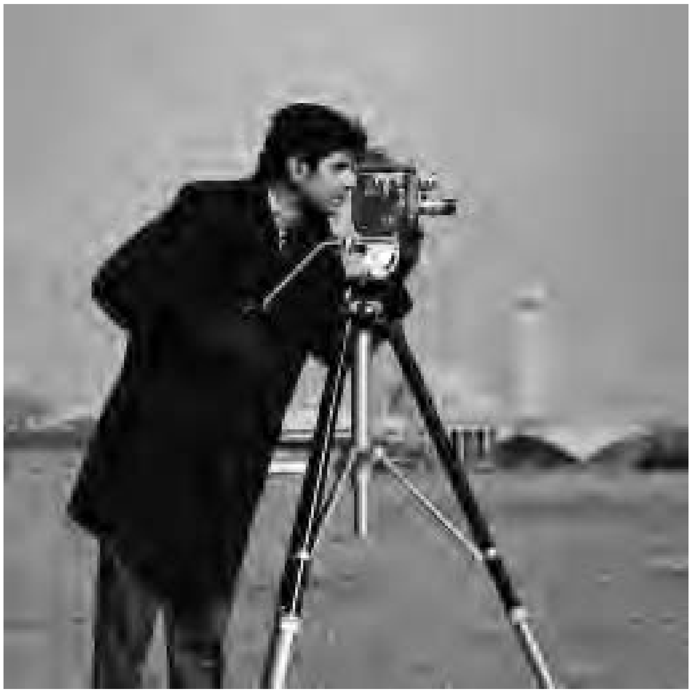


FIG. 11. *Dyadic biorthogonal wavelet approximation of the cameraman image with* $n = 2048$, PSNR = 26.72.

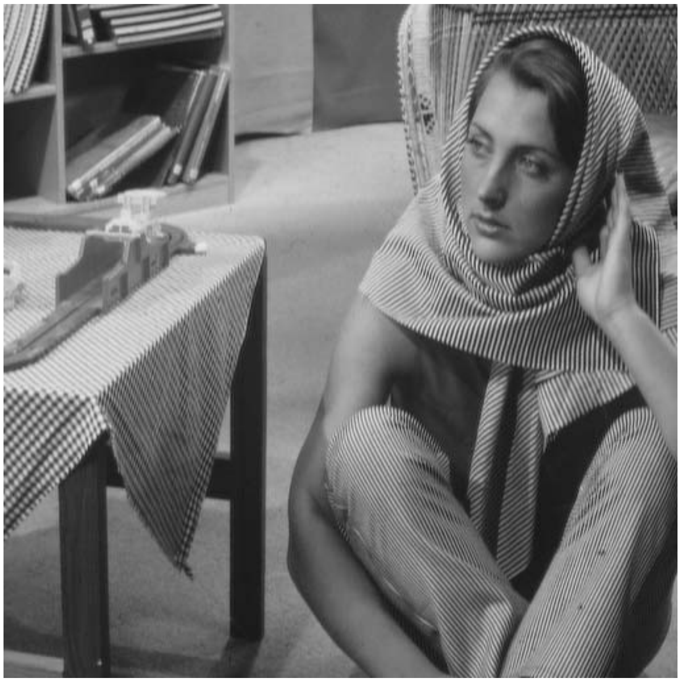


Fig. 12. *The "Barbara" image* 512×512.

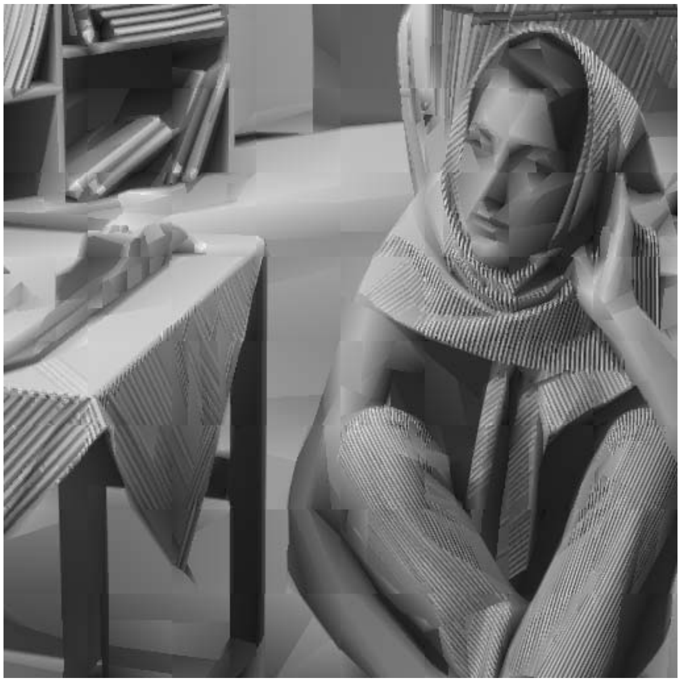


Fig. 13. *Geometric wavelet approximation of the Barbara image with* $n = 4096$, PSNR $= 27.10$.

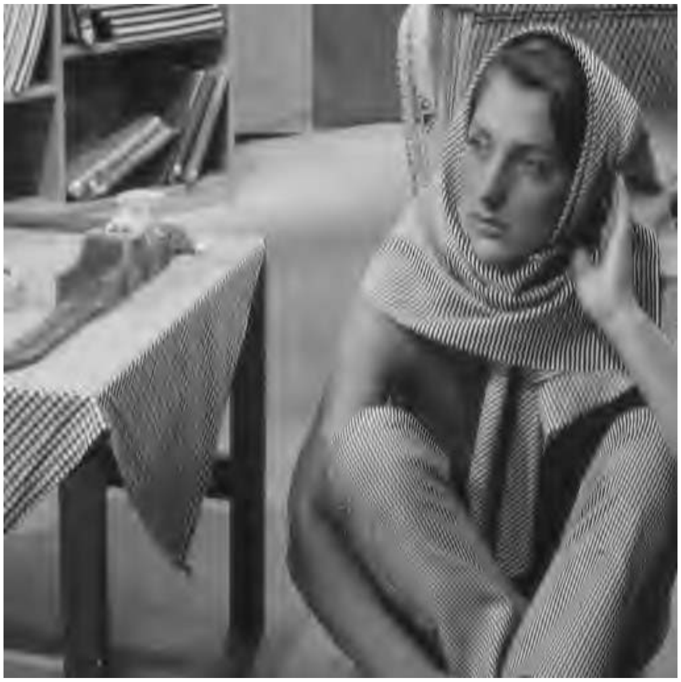


Fig. 14. *Dyadic biorthogonal wavelet approximation of the Barbara image with* $n = 12288$, PSNR $= 27.54$.

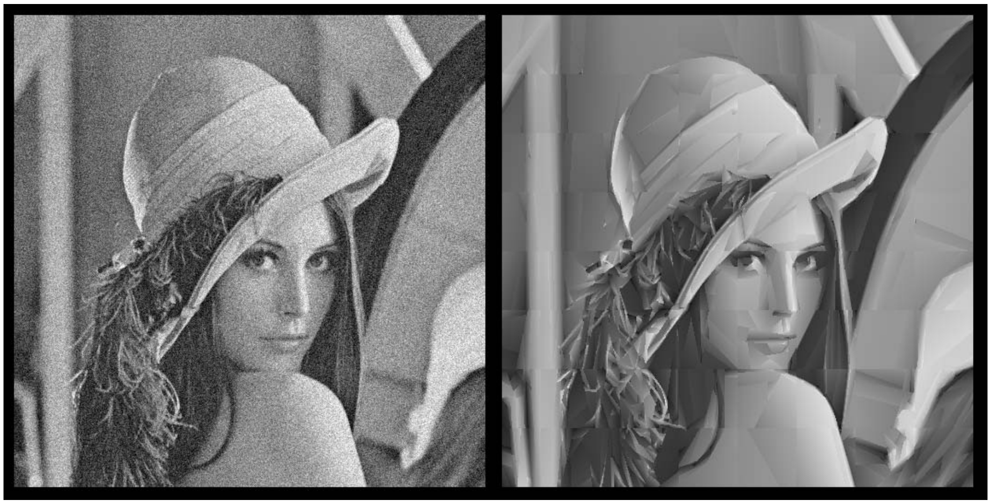


Fig. 15. *Geometric wavelet denoising. Noisy image* PSNR $= 22.14$*; restored image* PSNR $= 29.76$.

REFERENCES

[1] L. Brown and B. Lucier, *Best approximations in L^1 are near best in L^p, $p < 1$*, Proc. Amer. Math. Soc., 120 (1994), pp. 97–100.

[2] E. Candès and D. Donoho, *New Tight Frames of Curvelets and Optimal Representations of Objects with Smooth Singularities*, Technical report, Stanford, CA, 2002.

[3] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Reg. Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[4] S. Dekel and D. Leviatan, *The Bramble–Hilbert lemma for convex domains*, SIAM J. Math. Anal., 35 (2004), pp. 1203–1212.

[5] S. Dekel and D. Leviatan, *Whitney estimates for convex domains with applications to multivariate piecewise polynomial approximation*, Found. Comput. Math., 4 (2004), pp. 345–368.

[6] S. Dekel, D. Leviatan, and M. Sharir, *On bivariate smoothness spaces associated with nonlinear approximation*, Constr. Approx., 20 (2004), pp. 625–646.

[7] R. DeVore, *Nonlinear approximation*, Acta Numer., 7 (1998), pp. 51–150.

[8] R. DeVore, B. Jawerth, and B. Lucier, *Image compression through wavelet transform coding*, IEEE Trans. Inform. Theory, 38 (1992), pp. 719–746.

[9] R. DeVore and G. Lorentz, *Constructive Approximation*, Springer-Verlag, New York, 1991.

[10] R. DeVore and V. Popov, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 397–414.

[11] D. Donoho, *CART and best-ortho-basis: A connection*, Ann. Statist., 25 (1997), pp. 1870–1911.

[12] J. Hershberger and S. Suri, *Binary space partitions for* $3D$ *subdivisions*, in Proceedings of the 14th Annual ACM-SIAM Joint Symposium on Discrete Algorithms, Baltimore, MD, 2003, SIAM, Philadelphia, 2003, pp. 100–108.

[13] B. Karaivanov and P. Petrushev, *Nonlinear piecewise polynomial approximation beyond Besov spaces*, Appl. Comput. Harmon. Anal., 15 (2003), pp. 177–223.

[14] B. Karaivanov, P. Petrushev, and R. Sharpley, *Algorithms for nonlinear piecewise polynomial approximation: Theoretical aspects*, Trans. Amer. Math. Soc., 355 (2003), pp. 2585–2631.

[15] M. S. Paterson and F. F. Yao, *Efficient binary space partitions for hidden-surface removal and solid modeling*, Discrete Comput. Geom., 5 (1990), pp. 485–503.

[16] P. Petrushev, *Multivariate* n*-term rational and piecewise polynomial approximation*, J. Approx. Theory, 121 (2003), pp. 158–197.

[17] H. Radha, M. Vetterli, and R. Leonardi, *Image compression using binary space partitioning trees*, IEEE Trans. Image Process., 5 (1996), pp. 1610–1624.

[18] A. Said and W. Pearlman, *A new fast and efficient image codec based on set partitioning in hierarchical trees*, IEEE Trans. Circuits Systems Video Technol., 6 (1996), pp. 243–250.

[19] P. Salembier and L. Garrido, *Binary partition tree as an efficient representation for image processing, segmentation, and information retrieval*, IEEE Trans. Image Process., 9 (2000), pp. 561–576.

[20] M. Shapiro, *An embedded hierarchical image coder using zerotrees of wavelet coefficients*, IEEE Trans. Signal Process., 41 (1993), pp. 3445–3462.

[21] R. Shukla, P. L. Dragotti, M. N. Do, and M. Vetterli, *Rate-distortion optimized tree structured compression algorithms for piecewise polynomial images*, IEEE Trans. Image Process., 14 (2005), pp. 343–359.

[22] J. L. Starck, E. Candès, and D. L. Donoho, *The curvelet transform for image denoising*, IEEE Trans. Image Process., 11 (2000), pp. 670–684.

[23] D. Taubman, *High performance scalable image compression with EBCOT*, IEEE Trans. Image Process., 9 (2000), pp. 1151–1170.

[24] C. Tóth, *A note on binary plane partitions*, in Proceedings of the 17th ACM Symposium on Computational Geometry, ACM, New York, 2001, pp. 151–156.

[25] C. Tóth, *Binary space partitions for line segments with a limited number of directions*, SIAM J. Comput., 32 (2003), pp. 307–325.

[26] M. Wakin, J. Romberg, H. Choi, and R. Baraniuk, *Geometric methods for wavelet-based image compression*, in Wavelets: Applications in Signal and Image Processing X, M. Unser, A. Aldroubi, and A. Laine, eds., SPIE, Bellingham, WA, 2003, pp. 507–520.

SIAM J. Numer. Anal.
Vol. 43, No. 2, pp. 733–749

OPTIMAL ERROR ESTIMATES FOR LINEAR PARABOLIC PROBLEMS WITH DISCONTINUOUS COEFFICIENTS*

RAJEN KUMAR SINHA† AND BHUPEN DEKA†

Abstract. A finite element discretization is proposed and analyzed for a linear parabolic problems with discontinuous coefficients. Due to low global regularity of the solution, it seems difficult to achieve optimal order of convergence with classical finite element methods [*Numer. Math.*, 79 (1998), pp. 175–202]. In this paper, we have used a finite element discretization, where interface triangles are assumed to be curved triangles instead of straight triangles as in classical finite element methods. Optimal order error estimates in L^2 and H^1 norms are shown to hold even if the regularity of the solution is low on the whole domain. While the continuous time Galerkin method is discussed for the spatially discrete scheme, the discontinuous Galerkin method is analyzed for the fully discrete scheme. The interfaces and boundaries of the domains are assumed to be smooth for our purpose.

Key words. parabolic, interface, discontinuous coefficients, finite element approximation, semidiscrete and fully discrete schemes, optimal error estimates, discontinuous Galerkin method

AMS subject classifications. 65N30, 65F10

DOI. 10.1137/040605357

1. Introduction.

In this paper, we consider a linear parabolic interface problem of the form

$$u_t + \mathcal{L}u = f(x,t) \quad \text{in } \Omega \times (0,T] \tag{1.1}$$

with initial and boundary conditions

$$u(x,0) = u_0 \ \text{ in } \Omega; \ \ u(x,t) = 0 \quad \text{on } \partial\Omega \times (0,T] \tag{1.2}$$

and interface conditions

$$[u] = 0, \ \ \left[\mathcal{A}\frac{\partial u}{\partial \mathbf{n}}\right] = g(x,t) \quad \text{along } \Gamma, \tag{1.3}$$

where Ω is a bounded domain in $\mathbb{R}^2$ with smooth boundary $\partial\Omega$, $\Omega_1 \subset \Omega$ is an open domain with C^2 boundary Γ, and $\Omega_2 = \Omega\backslash\Omega_1$ (see Figure 1). The operator $\mathcal{L}$ is a second order elliptic partial differential operator of the form

$$\mathcal{L}v = -\nabla \cdot (\mathcal{A}\nabla v).$$

The symbol $[v]$ is a jump of a quantity v across the interface Γ and $\mathbf{n}$ denotes the unit outward normal to the boundary $\partial\Omega_1$. We assume that the coefficient matrix $\mathcal{A} = (a_{ij}(x))_{i,j=1}^2$ is symmetric and uniformly positive definite in Ω. Moreover, the matrix $\mathcal{A}$ is assumed to be discontinuous along Γ but piecewise smooth in each subdomain Ω_1 and Ω_2, i.e.,

$$\mathcal{A} = \mathcal{A}_l = (a_{ij}^l(x))_{i,j=1}^2 \quad \text{for } x \in \Omega_l, \ l = 1, 2.$$

Here for each l, $\mathcal{A}_l$ is a uniformly positive definite matrix.

[*] Received by the editors March 17, 2004; accepted for publication (in revised form) November 23, 2004; published electronically August 31, 2005.

http://www.siam.org/journals/sinum/43-2/60535.html

[†] Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati, Assam, 781039, India (rajen@iitg.ernet.in, bdeka@iitg.ernet.in). The research of the first author was supported by the Department of Science and Technology, Government of India.

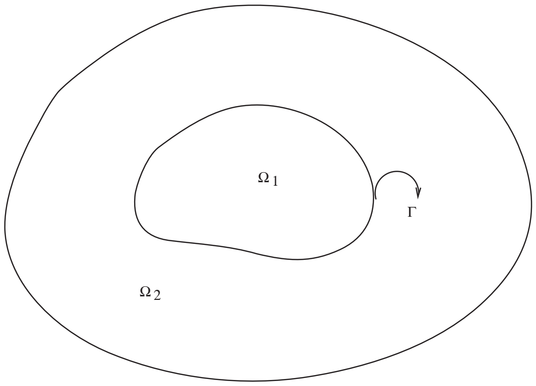


FIG. 1. *Domain Ω and its subdomains Ω_1, Ω_2 with interface Γ.*

Parabolic equations (1.1) with discontinuous coefficients occur in many applications, such as in material sciences and fluid dynamics, where domains consist of two or more heterogeneous media, i.e., when two distinct materials or fluids with different conductivities, densities, or diffusions are involved. Because of the discontinuity of the coefficients along the interface, the solution of such a problem has low regularity on the whole physical domain (cf. [10] and [5]).

In recent years, numerical methods for solving parabolic problems, by means of finite element methods under minimal regularity assumption on the true solutions, are the subject of much interest; see [5] and [6]. Due to low global regularity of the true solution, achieving higher order accuracy by the classical finite element method seems very difficult (cf. [2] and [5]). In [5], the authors studied the convergence of the finite element method for the elliptic and parabolic problems by approximating the smooth interface by a polygon and the interface function by its interpolant. They obtained suboptimal order error estimates in both energy and L^2 norms. It is also mentioned [5, p. 177] that classical analysis is difficult to apply in the convergence analysis for the interface problem. Subsequently, the author of [13] analyzed the error in the finite element method applied to self-adjoint elliptic interface problems and obtained optimal order error estimate in the H^1 norm. More recently, in [16] the authors studied elliptic interface type problems by means of the finite element method and proved optimal rates of convergence when the global regularity of the solution is low.

In the present paper, we propose a finite element discretization for the parabolic interface problem (1.1)–(1.3) by allowing interface triangles to be curved triangles instead of straight triangles. The analysis presented shows that the finite element solutions approximate the true solutions with an optimal order even if the global regularity is very low. More precisely, for the spatially discrete scheme, optimal order error estimates are derived in L^2 and H^1 norms. The key to the present analysis is the introduction of some auxiliary projections and duality arguments. Further, the discontinuous Galerkin method is analyzed for the time discretization, and related error estimates are obtained. To the best of our knowledge, optimal error estimates using conforming finite element methods for the parabolic interface problem have not been established earlier.

The previous work on finite element analysis of elliptic and parabolic problems without interface can be found in [7], [8], [9], [12], [17], and references therein. For literature on the discontinuous Galerkin methods, we refer to [3], [11], [14], [15], and [17].

A brief outline of this paper is as follows. In section 2, we introduce some notation, recall some basic results from the literature, and obtain the a priori estimate for the solution. In section 3, we describe a finite element discretization for the problem (1.1)–(1.3) and prove some approximation properties related to the auxiliary projection used in our analysis. Section 4 is devoted to the error estimates for the spatially discrete scheme. Finally, the discontinuous Galerkin method is analyzed for the fully discrete scheme, and related error estimates are derived in section 5.

Throughout this paper, C denotes a generic positive constant that does not depend on the spatial and time discretization parameters h and k, respectively.

2. Preliminaries.

2.1. Basic notation. We shall use standard notation for Sobolev spaces and norms. For $m \geq 0$ and real p with $1 \leq p \leq \infty$, we use $W^{m,p}(\Omega)$ to denote a Sobolev space of order m with norm $\|.\|_m$ and, in particular, for $p = 2$, we write $W^{m,2}(\Omega) = H^m(\Omega) = H^m$. $H_0^m(\Omega)$ is a closed subspace of $H^m(\Omega)$, which is also a closure of $C_0^\infty(\Omega)$ (the set of all C^∞ functions with compact support) with respect to the norm of $H^m(\Omega)$. For a fractional number s, Sobolev space H^s is defined in [1]. For a given Banach space $\mathcal{B}$, we define for $m = 0, 1$,

$$H^m(0,T;\mathcal{B}) = \left(u(t) \in \mathcal{B} \text{ for a.e. } t \in (0,T) \text{ and } \sum_{j=0}^{m} \int_0^T \left\| \frac{\partial^j u(t)}{\partial t^j} \right\|_{\mathcal{B}}^2 dt < \infty \right)$$

equipped with the norm

$$\|u\|_{H^m(0,T;\mathcal{B})} = \left(\sum_{j=0}^{m} \int_0^T \left\| \frac{\partial^j u(t)}{\partial t^j} \right\|_{\mathcal{B}}^2 dt \right)^{\frac{1}{2}}.$$

We write $\|u\|_{H^1(\Omega)}^2 \equiv \|u\|_{H^1(\Omega_1)}^2 + \|u\|_{H^1(\Omega_2)}^2$ and $L^2(0,T;\mathcal{B}) = H^0(0,T;\mathcal{B})$.

In addition, we shall also work on the following spaces:

$$X = H^1(\Omega) \cap H^2(\Omega_1) \cap H^2(\Omega_2) \quad \text{and} \quad Y = L^2(\Omega) \cap H^1(\Omega_1) \cap H^1(\Omega_2)$$

equipped with the norms

$$\|v\|_X = \|v\|_{H^1(\Omega)} + \|v\|_{H^2(\Omega_1)} + \|v\|_{H^2(\Omega_2)}$$

and

$$\|v\|_Y = \|v\|_{L^2(\Omega)} + \|v\|_{H^1(\Omega_1)} + \|v\|_{H^1(\Omega_2)},$$

respectively.

In order to introduce the weak formulation of the problem, we now define the bilinear form $A(\cdot,\cdot) : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ by

$$A(u,v) = \int_\Omega \mathcal{A}\nabla u \cdot \nabla v dx \quad \forall u, v \in H^1(\Omega).$$

Then the weak formulation of the interface problem (1.1)–(1.3) is stated as follows. Find $u \in H_0^1(\Omega)$ such that

$$(u_t, v) + A(u,v) = (f,v) + \langle g, v\rangle_\Gamma \quad \forall v \in H_0^1(\Omega), \;\; t \in (0,T], \tag{2.1}$$

with $u(0) = u_0$. Here, $(\cdot,\cdot)$ and $\langle\cdot,\cdot\rangle$ are used to denote the inner products of the $L^2(\Omega)$ and $L^2(\Gamma)$ spaces, respectively.

2.2. A priori estimate. Due to the presence of discontinuous coefficients, the solution u, in general, does not belong to $H^2(\Omega)$. But one can expect higher local regularity of the solution when the coefficients are locally smoother (cf. [10]). In the theorem below, we prove the a priori estimate for the solution u of the interface problem (1.1)–(1.3) under appropriate regularity conditions on f and g.

THEOREM 2.1. *Let $f \in H^1(0,T;L^2(\Omega))$, $g \in H^1(0,T;H^{\frac{1}{2}}(\Gamma))$, and $u_0 \in H_0^1(\Omega)$. Then the problem (1.1)–(1.3) has a unique solution $u \in L^2(0,T;X) \cap H^1(0,T;Y)$. Further, u satisfies the a priori estimate*

$$\|u\|_{L^2(0,T;X)} \leq C\Big\{\|f\|_{L^2(0,T;L^2(\Omega))} + \|u_0\|_{H^1(\Omega)} + \|g(0)\|_{H^{\frac{1}{2}}(\Gamma)} + \|g(T)\|_{H^{\frac{1}{2}}(\Gamma)} + \|g\|_{H^1(0,T;H^{\frac{1}{2}}(\Gamma))}\Big\}. \tag{2.2}$$

Proof. The proof of the existence of a unique solution is in [10]. Next, to obtain the a priori estimate (2.2), we first transform the problem (1.1)–(1.3) into the following equivalent problem.

For a.e. $t \in (0,T]$, find $u = u(x,t) \in H_0^1(\Omega) \cap X$ satisfying

$$\begin{aligned} \mathcal{L}u &= f(x,t) - u_t \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \\ [u] = 0, \quad \left[\mathcal{A}\frac{\partial u}{\partial \mathbf{n}}\right] &= g(x,t) \quad \text{along } \Gamma. \end{aligned} \tag{2.3}$$

From the elliptic regularity estimate for the elliptic interface problem (cf. [5]), it follows that

$$\|u\|_X \leq C\left(\|f - u_t\|_{L^2(\Omega)} + \|g\|_{H^{\frac{1}{2}}(\Gamma)}\right). \tag{2.4}$$

Multiply both sides of (2.3) by u_t and then integrate over Ω to obtain

$$\|u_t\|^2_{L^2(\Omega)} + (\mathcal{L}u, u_t) = (f, u_t). \tag{2.5}$$

Note that $u \in H^1(0,T;X)$ and $[u] = 0$ on Γ imply $[u_t] = 0$ on Γ. Hence, an integration by parts leads to

$$\begin{aligned} (\mathcal{L}u, u_t) &= \int_{\Omega_1} \mathcal{A}_1 \nabla u \cdot \nabla u_t dx + \int_{\Omega_2} \mathcal{A}_2 \nabla u \cdot \nabla u_t dx + \int_\Gamma \left[\mathcal{A}\frac{\partial u}{\partial \mathbf{n}}\right] u_t dS \\ &= A^1(u, u_t) + A^2(u, u_t) + \langle g, u_t\rangle_\Gamma, \end{aligned} \tag{2.6}$$

where $A^l(.,.) : H^1(\Omega_l) \times H^1(\Omega_l) \to \mathbb{R}$ is given by

$$A^l(w, v) = \int_{\Omega_l} \mathcal{A}_l \nabla w \cdot \nabla v dx, \quad l = 1, 2.$$

Equation (2.5), together with (2.6), yields

$$\|u_t\|^2_{L^2(\Omega)} + \frac{1}{2}\frac{d}{dt}\left(\sum_{i=1}^2 A^i(u,u)\right) = (f, u_t) - \frac{d}{dt}\langle g, u\rangle_\Gamma + \langle g_t, u\rangle_\Gamma.$$

Integrate the above equation from 0 to T. Then apply the Cauchy–Schwarz inequality and the trace theorem (cf. [1]) to obtain

$$\begin{aligned}
&\int_0^T \|u_t\|^2_{L^2(\Omega)}ds + \|u(T)\|^2_{H^1(\Omega_1)} + \|u(T)\|^2_{H^1(\Omega_2)}\\
&\le C\left(\int_0^T \|f\|_{L^2(\Omega)}\|u_t\|_{L^2(\Omega)}ds\right.\\
&\qquad + \|g(T)\|_{L^2(\Gamma)}\|u(T)\|_{L^2(\Gamma)} + \|g(x,0)\|_{L^2(\Gamma)}\|u_0\|_{L^2(\Gamma)}\\
&\qquad \left. + \int_0^T \|g_t\|_{L^2(\Gamma)}\|u\|_{L^2(\Gamma)}ds + \|u_0\|^2_{H^1(\Omega_1)} + \|u_0\|^2_{H^1(\Omega_2)}\right)\\
&\le C\left(\int_0^T \|f\|_{L^2(\Omega)}\|u_t\|_{L^2(\Omega)}ds + \|g(T)\|_{H^{\frac{1}{2}}(\Gamma)}\|u(T)\|_{H^1(\Omega)}\right.\\
&\qquad \left. + \|g(x,0)\|_{H^{\frac{1}{2}}(\Gamma)}\|u_0\|_{H^1(\Omega)} + \int_0^T \|g_t\|_{H^{\frac{1}{2}}(\Gamma)}\|u\|_{H^1(\Omega)}ds + \|u_0\|^2_{H^1(\Omega)}\right).
\end{aligned}$$

Use a standard kickback argument to obtain

$$\begin{aligned}
&\|u_t\|^2_{L^2(0,T;L^2(\Omega))} + \|u(T)\|^2_{H^1(\Omega)}\\
&\le C\left(\int_0^T \|f\|^2_{L^2(\Omega)}ds + \|g(T)\|^2_{H^{\frac{1}{2}}(\Gamma)}\right.\\
&\qquad \left. + \|g(0)\|^2_{H^{\frac{1}{2}}(\Gamma)} + \int_0^T \|g_t\|^2_{H^{\frac{1}{2}}(\Gamma)}ds + \|u_0\|^2_{H^1(\Omega)}\right)\\
&\quad + C\int_0^T \|u(s)\|^2_{H^1(\Omega)}ds.
\end{aligned}$$

Finally, an application of Gronwall's lemma completes the proof. □

3. Finite element discretization and some auxiliary results. In this section, we shall describe a finite element discretization, introduce some auxiliary projections, and prove their approximation properties.

For the purpose of finite element approximation we now describe the triangulation of Ω as follows: Let $\mathcal{T}_h$ be a triangulation of Ω with mesh parameter h, $0 < h < 1$. We first approximate the domain Ω_1 by a domain Ω_1^h with the polygonal boundary Γ_h whose vertices all lie on the interface Γ. Let Ω_2^h be the approximation for the domain Ω_2 with polygonal exterior $\partial\Omega_2^h$ and interior boundary Γ_h. Further, let $\{P_j\}_{j=1}^{m_h}$ be the set of all nodes of the triangulation $\mathcal{T}_h$ lying on the interface Γ, and let $\{e_j\}(j = 1, \ldots, m_h)$ be the edge connecting the two neighboring points P_j and P_{j+1} such that $P_{m_h+1} = P_1$. Let $\mathcal{T}_h^*$ be a triangulation obtained with a modification of $\mathcal{T}_h$. $\mathcal{T}_h^*$ is obtained by changing those triangles of $\mathcal{T}_h$ having one e_j edge (for some $1 \le j \le m_h$) into curved triangles having two original edges unchanged but having their third edge e_j replaced with the curved segment (cf. Figure 2). The element $K \in \mathcal{T}_h^*$ with one curved edge along the interface Γ is called the interface curved triangle.

Triangulation $\mathcal{T}_h^*$ of the domain Ω satisfies the following conditions:

($\mathcal{A}1$) $\overline{\Omega} = \cup_{K\in\mathcal{T}_h^*}K$.

($\mathcal{A}2$) If K_1, $K_2 \in \mathcal{T}_h^*$, and $K_1 \ne K_2$, then either $K_1 \cap K_2 = \emptyset$ or $K_1 \cap K_2$ is a common vertex, or edge, or one curved edge of both triangles.

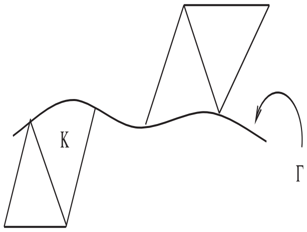


FIG. 2. *Interface curved triangle K.*

($\mathcal{A}3$) Each interface triangle K intersects Γ (interface) in at most two vertices and has at most one curved edge.

($\mathcal{A}4$) For each triangle $K \in \mathcal{T}_h^*$, let r_K, $\overline{r}_K$ be the radii of its inscribed and circumscribed circles, respectively. Let $h = \max\{\overline{r}_K : K \in \mathcal{T}_h^*\}$. We assume that, for some fixed $h_0 > 0$, there exists two positive constants C_0 and C_1 independent of h such that

$$C_0 h \leq diam(K) \leq C_1 h \quad \forall K \in \mathcal{T}_h^*, \quad \forall h \in (0, h_0).$$

Assumption ($\mathcal{A}4$) allows us to relate L^2 and H^1 norms of the polynomials in each element of $\mathcal{T}_h^*$ by

$$\|v\|_{H^1(K)} \leq Ch^{-1}\|v\|_{L^2(K)} \quad \forall K \in \mathcal{T}_h^* \tag{3.1}$$

for any polynomial $v \in P_1(K)$ (cf. [4, Lem. 4.5.3]).

Let V_h be a family of finite element subspaces of $H_0^1(\Omega)$ defined on $\mathcal{T}_h^*$ consisting of piecewise linear polynomials vanishing on the boundary $\partial\Omega$. Note that the construction of such finite element spaces is not straightforward. We refer to [13] for the construction and examples of various types of finite element spaces V_h.

Further, we assume that V_h satisfy the inverse estimate

$$\|\phi\|_{H^1(\Omega)} \leq Ch^{-1}\|\phi\|_{L^2(\Omega)} \quad \forall\, \phi \in V_h, \tag{3.2}$$

and this follows immediately from the estimate (3.1).

For $v \in X$, let

$$f^* = -\nabla \cdot (\mathcal{A}_l \nabla v) \quad \text{in } \Omega_l,\ l = 1, 2.$$

We now define an operator $R_h : X \cap H_0^1(\Omega) \to V_h$ by

$$A(R_h v, \phi) = (f^*, \phi) \quad \forall \phi \in V_h,\ v \in X \cap H_0^1(\Omega). \tag{3.3}$$

It now follows from the definition of f^* that

$$(f^*, \phi) = A(v, \phi) \quad \forall \phi \in V_h,\ v \in X \cap H_0^1(\Omega) \tag{3.4}$$

which, together with (3.3), yields

$$A(R_h v, \phi) = (f^*, \phi) = A(v, \phi) \quad \forall \phi \in V_h,\ v \in X \cap H_0^1(\Omega). \tag{3.5}$$

Below, we present a proof that shows optimal error bounds for the projection operator R_h. This lemma is very crucial for our later analysis.

LEMMA 3.1. *Let R_h be defined by* (3.5). *Then there is a positive constant C independent of h such that*

$$\|v - R_h v\| + h\|v - R_h v\|_1 \le Ch^2 \|v\|_X \quad \forall v \in X \cap H_0^1(\Omega).$$

Proof. By the definition of f^*, it is easy to see that

$$A(v, \phi) = (f^*, \phi) \quad \forall \phi \in H_0^1(\Omega). \tag{3.6}$$

For $k = 1, 2$, define $f_k : \Omega \to \mathbb{R}$ by

$$f_k = \begin{cases} f^*|_{\Omega_k} & \text{in } \Omega_k, \\ 0 & \text{otherwise.} \end{cases}$$

Clearly, $f_k \in L^2(\Omega)$ and $f^* = f_1 + f_2$ a.e. in Ω. We now consider the following interface problems: Let $w_k \in H_0^1(\Omega)$ be the solution of the interface problem

$$A(w_k, \phi) = (f_k, \phi) \quad \forall \phi \in H_0^1(\Omega). \tag{3.7}$$

Then by the coercivity of the operator A it follows immediately that $v = w_1 + w_2$.

Let $w_h^k \in V_h$ be the finite element approximation to w_k defined by

$$A(w_h^k, \phi) = (f_k, \phi) \quad \forall \phi \in V_h. \tag{3.8}$$

Again by the coercivity of the operator A and definition (3.5) of the R_h operator, it follows that $R_h v = w_h^1 + w_h^2$.

Since $f_k|_{\Omega_s} = 0$, $s = 1(2)$ if $k = 2(1)$, we have for the elliptic interface problem (3.7)–(3.8) (see [13, Thm. 6.1])

$$\begin{aligned} \|w_k - w_h^k\|_{H^1(\Omega)} &\le Ch(\|w_k\|_{H^2(\Omega_1)} + \|w_k\|_{H^2(\Omega_2)}) \\ &\le Ch\|w_k\|_X. \end{aligned} \tag{3.9}$$

Then by the elliptic regularity (cf. [5]) we have

$$\begin{aligned} \|w_k\|_X &\le C\|f_k\|_{L^2(\Omega)} \\ &\le C\|f^*\|_{L^2(\Omega_k)} \le C\|v\|_{H^2(\Omega_k)}. \end{aligned}$$

This, together with (3.9), yields

$$\|w_k - w_h^k\|_{H^1(\Omega)} \le Ch\|v\|_{H^2(\Omega_k)}. \tag{3.10}$$

Then

$$\begin{aligned} \|v - R_h v\|_{H^1(\Omega)} &\le \|w_1 - w_h^1\|_{H^1(\Omega)} + \|w_2 - w_h^2\|_{H^1(\Omega)} \\ &\le C\,(h\|w_1\|_X + h\|w_2\|_X) \\ &\le Ch\left(\|v\|_{H^2(\Omega_1)} + \|v\|_{H^2(\Omega_2)}\right) \\ &\le Ch\|v\|_X \quad \forall v \in X \cap H_0^1(\Omega). \end{aligned} \tag{3.11}$$

For the L^2 norm estimate, let us consider the following problem: Find $w \in H_0^1(\Omega)$ such that

$$A(w, \phi) = (v - R_h v, \phi) \quad \forall \phi \in H_0^1(\Omega). \tag{3.12}$$

By setting $\phi = v - R_h v \in H_0^1(\Omega)$ in (3.12) and using definition (3.5) of the R_h operator, we have

$$\begin{aligned}\|v - R_h v\|^2_{L^2(\Omega)} &= A(w, v - R_h v)\\ &= A(w - R_h w, v - R_h v) + A(R_h w, v - R_h v)\\ &= A(w - R_h w, v - R_h v)\\ &\le C\|w - R_h w\|_{H^1(\Omega)}\|v - R_h v\|_{H^1(\Omega)}\\ &\le Ch^2\|v\|_X\|w\|_X;\end{aligned}$$

in the last inequality, we used (3.11). Then applying the elliptic regularity estimate for the interface problem (3.12), we get

$$\begin{aligned}\|v - R_h v\|^2_{L^2(\Omega)} &\le Ch^2\|v\|_X\|w\|_X\\ &\le Ch^2\|v - R_h v\|_{L^2(\Omega)}\|v\|_X.\end{aligned}$$

This completes the proof of Lemma 3.1. □

Let $L_h : L^2(\Omega) \to V_h$ be the standard L^2 projection defined by

$$(L_h v, \phi) = (v, \phi) \quad \forall v \in L^2(\Omega), \ \ \phi \in V_h, \tag{3.13}$$

satisfying the stability estimate

$$\|L_h v\|_{H^1(\Omega)} \le C\|v\|_{H^1(\Omega)} \quad \forall v \in H_0^1(\Omega). \tag{3.14}$$

It is well known that $L_h v \in V_h$ is the best approximation in the L^2 norm to $v \in L^2(\Omega)$. The following lemma shows that $L_h v$ is a quasi-best approximation to $v \in H_0^1(\Omega) \cap X$ in the H^1 norm.

Lemma 3.2. *Let R_h and L_h be defined by* (3.5) *and* (3.13)*, respectively. Then we have*

$$\|L_h v - v\|_{H^1(\Omega)} \le C\|R_h v - v\|_{H^1(\Omega)} \quad \forall v \in H_0^1(\Omega) \cap X.$$

Proof. For any $v \in H_0^1(\Omega) \cap X$, we know that there exists a unique solution $w \in H_0^1(\Omega)$ for the elliptic interface problem

$$A(w, \phi) = (R_h v - v, \phi) \quad \forall \phi \in H_0^1(\Omega). \tag{3.15}$$

Equation (3.15), together with (3.5) and Lemma 3.1, leads to

$$\begin{aligned}\|R_h v - v\|^2_{L^2(\Omega)} &= A(w - R_h w, R_h v - v) + A(R_h w, R_h v - v)\\ &\le C\|w - R_h w\|_{H^1(\Omega)}\|v - R_h v\|_{H^1(\Omega)}\\ &\le Ch\|w\|_X\|v - R_h v\|_{H^1(\Omega)}\\ &\le Ch\|v - R_h v\|_{L^2(\Omega)}\|v - R_h v\|_{H^1(\Omega)}.\end{aligned}\tag{3.16}$$

Here we used the fact that $\|w\|_X \le C\|v - R_h v\|_{L^2(\Omega)}$. Use of the triangle inequality and (3.2) yields

$$\begin{aligned}\|L_h v - v\|_{H^1(\Omega)} &\le \|R_h v - v\|_{H^1(\Omega)} + \|L_h v - R_h v\|_{H^1(\Omega)}\\ &\le \|R_h v - v\|_{H^1(\Omega)} + Ch^{-1}\|L_h v - R_h v\|_{L^2(\Omega)}\\ &\le \|R_h v - v\|_{H^1(\Omega)} + Ch^{-1}\{\|v - R_h v\|_{L^2(\Omega)} + \|L_h v - v\|_{L^2(\Omega)}\}.\end{aligned}\tag{3.17}$$

We know $L_h v$ is the best approximation of $v \in L^2(\Omega)$ with respect to the L^2 norm. Since $v \in X \cap H_0^1(\Omega)$, $R_h v \in V_h$. Thus, $\|v - L_h v\|_{L^2(\Omega)} \le C\|v - R_h v\|_{L^2(\Omega)}$. Therefore, (3.17) implies

$$\|L_h v - v\|_{H^1(\Omega)} \le \|R_h v - v\|_{H^1(\Omega)} + Ch^{-1}\|v - R_h v\|_{L^2(\Omega)}. \tag{3.18}$$

The desired estimate now follows from (3.16) and (3.18). □

4. The continuous time Galerkin approximation. This section deals with the error analysis for the spatially discrete scheme. Optimal order of convergence in both L^2 and H^1 norms is established when the global regularity of the solution is low on the entire domain.

The continuous time Galerkin finite element approximation to (2.1) is stated as follows: Find $u_h(t) \in V_h$ such that

$$(u_{ht}, v_h) + A(u_h, v_h) = (f, v_h) + \langle g, v_h\rangle_\Gamma \quad \forall v_h \in V_h, \quad t \in (0, T], \tag{4.1}$$

with $u_h(0) = L_h u_0$. Subtracting (4.1) from (2.1) we have

$$(u_t - u_{ht}, v_h) + A(u - u_h, v_h) = 0 \quad \forall v_h \in V_h. \tag{4.2}$$

Define the error $e(t)$ as $e(t) = u(t) - u_h(t)$. Setting $v_h = L_h u$ in (4.2) and using (3.13), we obtain

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\|e\|^2_{L^2(\Omega)} &+ A(e, e)\\
&= A(u - u_h, u - L_h u) + (u_t - u_{ht}, u - L_h u)\\
&= A(u - u_h, u - L_h u) + (u_t - L_h u_t, u - L_h u)\\
&\quad + (L_h u_t - u_{ht}, u - L_h u)\\
&= A(u - u_h, u - L_h u) + (u_t - (L_h u)_t, u - L_h u)\\
&\quad + (L_h u_t - u_{ht}, u - L_h u)\\
&= A(u - u_h, u - L_h u) + \frac{1}{2}\frac{d}{dt}(u - L_h u, u - L_h u).
\end{aligned} \tag{4.3}$$

In the last equality we used the fact that $L_h u_t - u_{ht} \in V_h$ and the definition (3.13) of the L_h operator. Integrate the above equation from 0 to t. Then apply the Cauchy–Schwarz inequality and Young's inequality to obtain

$$\begin{aligned}
\|e\|^2_{L^2(\Omega)} &+ \int_0^t \|e\|^2_{H^1(\Omega)} ds\\
&\le C\left(\int_0^t \|u - L_h u\|^2_{H^1(\Omega)} ds + \|u - L_h u\|^2_{L^2(\Omega)} + \|u_0 - L_h u_0\|^2_{L^2(\Omega)}\right).
\end{aligned}$$

An application of Lemma 3.2 leads to

$$\begin{aligned}
\|e\|^2_{L^2(\Omega)} &+ \int_0^t \|e\|^2_{H^1(\Omega)} ds\\
&\le C\left(\int_0^t \|u - R_h u\|^2_{H^1(\Omega)} ds + \|u - L_h u\|^2_{L^2(\Omega)} + \|u_0 - L_h u_0\|^2_{L^2(\Omega)}\right),
\end{aligned}$$

which, together with Lemma 3.1 and the fact $\|L_h v - v\|_{L^2(\Omega)} \le Ch\|v\|_{H^1(\Omega)} \forall v \in H^1_0(\Omega)$ yields

$$\int_0^t \|e\|^2_{H^1(\Omega)} ds \le h^2 \left\{\|u_0\|^2_{H^1(\Omega)} + \|u\|^2_X + \|u\|^2_{L^2(0,T;X)}\right\}.$$

Thus we have proved the following optimal H^1 norm estimate.

THEOREM 4.1. *Let u and u_h be the solutions of the problem* (1.1)–(1.3) *and* (4.1), *respectively. Then, for $u_0 \in H^1_0(\Omega)$, $f \in H^1(0, T; L^2(\Omega))$, and $g \in H^1(0, T; H^{\frac{1}{2}}(\Gamma))$, we have*

$$\|u - u_h\|_{L^2(0,T;H^1(\Omega))} \le Ch\left\{\|u_0\|_{H^1(\Omega)} + \|u\|_X + \|u\|_{L^2(0,T;X)}\right\}.$$

Next, for the L^2 norm error estimate, we shall use the duality argument. For this purpose, we now consider the following auxiliary problem: Find $z \in H_0^1(\Omega)$ such that

$$A(z, v) = (u - u_h, v) \quad \forall v \in H_0^1(\Omega), \quad t \in (0, T], \tag{4.4}$$

with $[\mathcal{A}\frac{\partial z}{\partial \mathbf{n}}] = 0$ across the interface Γ. Then its finite element approximation is defined to be a function $z_h \in V_h$ satisfying

$$A(z_h, v_h) = (u - u_h, v_h) \quad \forall v_h \in V_h, \quad t \in (0, T]. \tag{4.5}$$

Setting $v = u - u_h$ in (4.4) and using (4.2), we obtain

$$\begin{aligned} \|u - u_h\|^2_{L^2(\Omega)} &= A(z, u - u_h) \\ &= A(z - z_h, u - u_h) + A(z_h, u - u_h) \\ &= A(z - z_h, u - u_h) - (u_t - u_{ht}, z_h). \end{aligned} \tag{4.6}$$

Differentiating (4.5) with respect to t, we obtain

$$A(z_{ht}, v_h) = (u_t - u_{ht}, v_h).$$

Thus, we have

$$\frac{1}{2}\frac{d}{dt}A(z_h, z_h) = A(z_{ht}, z_h) = (u_t - u_{ht}, z_h),$$

and hence, integrating (4.6) from 0 to T we obtain

$$\begin{aligned} &\|u - u_h\|^2_{L^2(0,T;L^2(\Omega))} + \frac{1}{2}A(z_h, z_h) \\ &\leq C\int_0^T \|z - z_h\|_{H^1(\Omega)}\|u - u_h\|_{H^1(\Omega)}ds + \frac{1}{2}A(z_h(0), z_h(0)). \end{aligned}$$

Further, using the regularity estimate for the elliptic interface problem (4.4)–(4.5) and Lemma 3.1, we obtain

$$\begin{aligned} \|z - z_h\|_{H^1(\Omega)} &\leq C\|z - R_h z\|_{H^1(\Omega)} \\ &\leq Ch\|z\|_X \\ &\leq Ch\|u - u_h\|_{L^2(\Omega)}, \end{aligned}$$

and hence

$$\begin{aligned} &\|u - u_h\|^2_{L^2(0,T;L^2(\Omega))} \\ &\leq C\int_0^T h\|u - u_h\|_{L^2(\Omega)}\|u - u_h\|_{H^1(\Omega)}ds + \frac{1}{2}A(z_h(0), z_h(0)). \end{aligned} \tag{4.7}$$

Taking $t \to 0$, it now follows from (4.5) that

$$A(z_h(0), z_h(0)) = (u_0 - L_h u_0, z_h(0)) = 0.$$

This, together with (4.7) and Theorem 4.1, leads to

$$\begin{aligned} &\|u - u_h\|^2_{L^2(0,T;L^2(\Omega))} \\ &\leq Ch\left(\int_0^T \|u - u_h\|^2_{L^2(\Omega)}ds\right)^{\frac{1}{2}}\left(\int_0^T \|u - u_h\|^2_{H^1(\Omega)}ds\right)^{\frac{1}{2}} \\ &\leq Ch^2\|u - u_h\|_{L^2(0,T;L^2(\Omega))}\left\{\|u_0\|_{H^1(\Omega)} + \|u\|_X + \|u\|_{L^2(0,T;X)}\right\}. \end{aligned}$$

Thus we have proved the following optimal L^2 norm estimate.

THEOREM 4.2. *Let u and u_h be the solutions of the problems* (1.1)–(1.3) *and* (4.1), *respectively. Then, for $u_0 \in H_0^1(\Omega)$, $f \in H^1(0,T;L^2(\Omega))$, $g \in H^1(0,T;H^{\frac{1}{2}}(\Gamma))$, we have*

$$\|u-u_h\|_{L^2(0,T;L^2(\Omega))} \leq Ch^2 \left\{ \|u_0\|_{H^1(\Omega)} + \|u\|_X + \|u\|_{L^2(0,T;X)} \right\}.$$

Remark 4.1. Note that Theorems 4.1 and 4.2 yield an optimal order of convergence, assuming that the interface triangles are curved instead of straight. In contrast to [5] we do not require the interface function $g \in C(\Gamma)$, but $g \in H^1(0,T;H^{1/2}(\Gamma))$ suffices for the present analysis.

5. Discrete time discontinuous Galerkin method. In this section, we apply the discontinuous Galerkin method in the direction of the time variable. In this method an approximation to the solution is sought as a piecewise constant polynomial function in t, which is not necessarily continuous at the nodes of the defining partition.

We first divide the interval $[0,T]$ into M equally spaced subintervals by the points

$$0 = t_0 < t_1 < \cdots < t_M = T$$

with $t_n = nk$, $k = T/M$ being the time step. Let $I_n = (t_{n-1}, t_n]$ be the nth subinterval. In order to discretize (2.1) in time, we shall use the finite dimensional space

$$V_{hk} = \{\phi : [0,T] \to V_h \ : \phi|_{I_n} \in V_h \text{ is constant in time}\}.$$

For $\phi \in V_{hk}$, we denote ϕ^n and ϕ_+^n to be the value of ϕ and its limit from the above at t_n, respectively. Further, we write V_{hk}^n for the restriction to I_n of the functions in V_{hk}.

Remark 5.1. The functions belonging to V_{hk} need not be continuous at the nodes but are taken to be continuous to the left there.

Now we introduce the backward difference quotient

$$\Delta_k \phi^n = \frac{\phi^n - \phi^{n-1}}{k}$$

for a given sequence $\{\phi^n\}_{n=0}^M \subset L^2(\Omega)$. For a given Banach space $\mathcal{B}$ and some function $\xi \in L^2(0,T;\mathcal{B})$, we write

$$\overline{\xi}^{\,n} = k^{-1} \int_{I_n} \xi(x,t)dt. \tag{5.1}$$

The fully discrete finite element approximation to the problem (2.1) is defined as follows: Find $U^n \in V_{hk}$, for $n = 1, 2, \ldots, M$, such that

$$(\Delta_k U^n, v_h) + A(U^n, v_h) = (\overline{f}^{\,n}, v_h) + \langle \overline{g}^{\,n}, v_h \rangle_\Gamma \quad \forall v_h \in V_{hk}^n \tag{5.2}$$

with $U^0 = R_h u_0$ and $\overline{f}^{\,n}$ defined as in (5.1).

Below we prove the stability result for the solution U^n satisfying (5.2).

LEMMA 5.1. *Let U^n be satisfy* (5.2). *Then we have*

$$\|U^M\|_{L^2(\Omega)}^2 + \sum_{n=1}^M k\|U^n\|_{H^1(\Omega)}^2 \leq C\left(\|f\|_{L^2(0,T;L^2(\Omega))}^2 + \|u_0\|_{H^1(\Omega)}^2 + \|g\|_{L^2(0,T;H^{\frac{1}{2}}(\Gamma))}^2 \right).$$

Proof. Setting $v_h = U^n$ in (5.2) and then using the Cauchy–Schwarz inequality we obtain

$$\begin{aligned}\|U^n\|^2_{L^2(\Omega)} &+ k\|U^n\|^2_{H^1(\Omega)}\\ &\le k(\overline{f}^{\,n}, U^n) + k\langle\overline{g}^{\,n}, U^n\rangle_\Gamma + (U^{n-1}, U^n)\\ &\le k\|\overline{f}^{\,n}\|_{L^2(\Omega)}\|U^n\|_{L^2(\Omega)} + k\|\overline{g}^{\,n}\|_{L^2(\Gamma)}\|U^n\|_{L^2(\Gamma)}\\ &\qquad + \|U^{n-1}\|_{L^2(\Omega)}\|U^n\|_{L^2(\Omega)}.\end{aligned}$$

Applying Young's inequality, summing over n from $n = 1$ to $n = M$, and noting that $\|R_h u_0\|_{H^1(\Omega)} \le C\|u_0\|_{H^1(\Omega)}$, we obtain

$$\begin{aligned}\|U^M\|^2_{L^2(\Omega)} &+ \sum_{n=1}^{M} k\|U^n\|^2_{H^1(\Omega)}\\ &\le C\left(\|u_0\|^2_{H^1(\Omega)} + \sum_{n=1}^{M} k\|\overline{f}^{\,n}\|^2_{L^2(\Omega)} + \sum_{n=1}^{M} k\|\overline{g}^{\,n}\|^2_{L^2(\Gamma)}\right).\end{aligned}$$

It follows from simple calculation that

$$\sum_{n=1}^{M} k\|\overline{f}^{\,n}\|^2_{L^2(\Omega)} \le C\|f\|^2_{L^2(0,T;L^2(\Omega))} \quad \text{and} \quad \sum_{n=1}^{M} k\|\overline{g}^{\,n}\|^2_{L^2(\Gamma)} \le C\|g\|^2_{L^2(0,T;H^{\frac{1}{2}}(\Gamma))}.$$

Altogether these estimates lead to the desired result and complete the proof. □

Now we introduce the interpolant $P_k \in V_{hk}$ of u defined by

$$\begin{aligned}\int_{I_n} A(P_k - u, \phi)\, ds &= 0 \quad \forall \phi \in V_{hk},\\ \text{i.e. } P_k|_{I_n} &= k^{-1}\int_{I_n} R_h u ds\\ &= \overline{P}_k^{\,n}.\end{aligned} \tag{5.3}$$

It is easy to notice from Lemma 3.1 that

$$\left(\sum_{n=1}^{M} k\|\overline{u}^{\,n} - \overline{P}_k^{\,n}\|^2_{H^m(\Omega)}\right)^{\frac{1}{2}} \le Ch^{2-m}\|u\|_{L^2(0,T;X)}, \quad m = 0, 1. \tag{5.4}$$

Now we state the main results of this section in the following theorems.

THEOREM 5.2. *Let u and U be the solutions of* (1.1)–(1.3) *and* (5.2), *respectively. Then, for $u_0 \in X \cap H^1_0(\Omega)$, $f \in H^1(0,T;L^2(\Omega))$, and $g \in H^1(0,T;H^{\frac{1}{2}}(\Gamma))$, there exists a constant C independent of h and k such that*

$$\|u - U\|_{L^2(0,T;L^2(\Omega))} \le C(k + h^2)\left\{\|u_0\|_X + \|u\|_{L^2(0,T;X)} + \|u_t\|_{L^2(0,T;L^2(\Omega))}\right\}.$$

THEOREM 5.3. *Let u and U be the solutions of* (1.1)–(1.3) *and* (5.2), *respectively. Then, for $u_0 \in X \cap H^1_0(\Omega)$, $f \in H^1(0,T;L^2(\Omega))$, and $g \in H^1(0,T;H^{\frac{1}{2}}(\Gamma))$, there exists a positive constant C independent of h and k such that*

$$\|u - U\|_{L^2(0,T;H^1(\Omega))} \le C(k + h)\left\{\|u_0\|_X + \|u\|_{L^2(0,T;X)} + \|u_t\|_{L^2(0,T;Y)}\right\}.$$

The proofs of the above theorems require some preparation. We now appeal to the parabolic duality arguments. Consider the following auxiliary problem: Find $z^n \in V_{hk}$ such that

$$(5.5)\qquad (-\nabla_k z^n, v_h) + A(z_+^{n-1}, v_h) = (\overline{u}^{\,n} - U^n, v_h) \quad \forall v_h \in V_{hk}^n,\ 1 \le n \le M,$$

with $z_+^M = 0$, $\nabla_k z^n = \frac{z_+^n - z_+^{n-1}}{k}$. The following stability result of the solution z^n of (5.5) is very crucial for the convergence analysis.

LEMMA 5.4. *Let z^n be the solution of* (5.5). *Then we have*

$$\sum_{n=1}^{M} k\|\nabla_k z^n\|_{L^2(\Omega)}^2 + \|z_+^0\|_{H^1(\Omega)}^2 \le C \sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|_{L^2(\Omega)}^2.$$

Proof. Taking $v_h = -k\nabla_k z^n$ in (5.5) and applying the Cauchy–Schwarz inequality, we obtain

$$(5.6)\qquad k\|\nabla_k z^n\|_{L^2(\Omega)}^2 + A(z_+^{n-1}, z_+^{n-1} - z_+^n) \le Ck\|\overline{u}^{\,n} - U^n\|_{L^2(\Omega)}^2.$$

It is easy to notice that

$$A(z_+^{n-1}, z_+^{n-1} - z_+^n) = \frac{k^2}{2} A(\nabla_k z^n, \nabla_k z^n) - \frac{1}{2} A(z_+^n, z_+^n) + \frac{1}{2} A(z_+^{n-1}, z_+^{n-1}).$$

This, combined with (5.6), yields

$$k\|\nabla_k z^n\|_{L^2(\Omega)}^2 + \frac{k^2}{2} A(\nabla_k z^n, \nabla_k z^n) - \frac{1}{2} A(z_+^n, z_+^n) + \frac{1}{2} A(z_+^{n-1}, z_+^{n-1}) \le Ck\|\overline{u}^{\,n} - U^n\|_{L^2(\Omega)}^2.$$

Summing over n from $n = 1$ to $n = M$, we obtain

$$\sum_{n=1}^{M} k\|\nabla_k z^n\|_{L^2(\Omega)}^2 + A(z_+^0, z_+^0) \le \sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|_{L^2(\Omega)}^2.$$

This completes the proof. □

Proof of Theorem 5.2. Choose $v_h = k(\overline{P}_k^{\,n} - U^n) \in V_{hk}^n$ in (5.5). Observing that $A(z_+^{n-1}, \overline{P}_k^{\,n}) = A(z_+^{n-1}, \overline{u}^{\,n})$, we have

$$\begin{aligned}
&k\|\overline{u}^{\,n} - U^n\|_{L^2(\Omega)}^2 \\
&= k(\overline{u}^{\,n} - U^n, \overline{u}^{\,n} - \overline{P}_k^{\,n}) + k(-\nabla_k z^n, \overline{P}_k^{\,n} - U^n) + kA(z_+^{n-1}, \overline{P}_k^{\,n} - U^n) \\
&= k(\overline{u}^{\,n} - U^n, \overline{u}^{\,n} - \overline{P}_k^{\,n}) + k(-\nabla_k z^n, \overline{P}_k^{\,n} - U^n) + kA(z_+^{n-1}, \overline{u}^{\,n} - U^n) \\
&= k(\overline{u}^{\,n} - \overline{P}_k^{\,n}, \overline{u}^{\,n} - U^n) + k(-\nabla_k z^n, \overline{P}_k^{\,n} - u^n) \\
(5.7)\quad &+ k(-\nabla_k z^n, u^n - U^n) + kA(z_+^{n-1}, \overline{u}^{\,n} - U^n).
\end{aligned}$$

Now summing over n from $n=1$ to $n=M$, we obtain

$$\begin{aligned}\sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|^2_{L^2(\Omega)} &= \sum_{n=1}^{M}\Big\{k(\overline{u}^{\,n} - \overline{P}_k^{\,n}, \overline{u}^{\,n} - U^n)\Big\}\\ &\quad + \sum_{n=1}^{M}\Big\{k(-\nabla_k z^n, \overline{P}_k^{\,n} - u^n)\Big\}\\ &\quad + \sum_{n=1}^{M} k\left\{(-\nabla_k z^n, u^n - U^n) + A(z_+^{n-1}, \overline{u}^{\,n} - U^n)\right\}\\ &=: I_1 + I_2 + I_3. \end{aligned}\tag{5.8}$$

Before estimating the three terms in (5.8), we first rewrite the term I_3. Note that for all $v \in H_0^1(\Omega)$, we have

$$(\Delta_k u^n, v) + A(\overline{u}^{\,n}, v) = (\overline{f}^{\,n}, v) + \langle \overline{g}^{\,n}, v\rangle_\Gamma, \quad 1 \le n \le M. \tag{5.9}$$

Now, taking $v_h = v = z_+^{n-1}$ in both (5.2) and (5.9), subtracting one from the other, and summing the resulting equation over n, we obtain

$$\sum_{n=1}^{M} k\left\{(\Delta_k(u^n - U^n), z_+^{n-1}) + A(\overline{u}^{\,n} - U^n, z_+^{n-1})\right\} = 0 \tag{5.10}$$

which, together with (5.8), yields

$$\begin{aligned}&\sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|^2_{L^2(\Omega)}\\ &= I_1 + I_2 + \sum_{n=1}^{M} k\bigg((-\nabla_k z^n, u^n - U^n) + (-\Delta_k(u^n - U^n), z_+^{n-1})\bigg)\\ &=: I_1 + I_2 + I_4.\end{aligned}\tag{5.11}$$

Using the fact that $z_+^M = 0$, and applying the identity

$$\sum_{n=1}^{M}(a_n - a_{n-1})b_n = a_M b_M - a_0 b_0 - \sum_{n=1}^{M} a_{n-1}(b_n - b_{n-1})$$

to I_4 with $a_n = z_+^n$ and $b_n = u^n - U^n$, we obtain

$$\begin{aligned} I_4 &= \sum_{n=1}^{M} k\left\{(-\nabla_k z^n, u^n - U^n) + (-\Delta_k(u^n - U^n), z_+^{n-1})\right\}\\ &= (z_+^0, u_0 - R_h u_0).\end{aligned}\tag{5.12}$$

Using (5.4) and the Cauchy–Schwarz inequality, it now follows that

$$\begin{aligned}|I_1| &\le \left(\sum_{n=1}^{M} k\|\overline{u}^{\,n} - \overline{P}_k^{\,n}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \left(\sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}\\ &\le Ch^2\|u\|_{L^2(0,T;X)} \left(\sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}.\end{aligned}\tag{5.13}$$

Similarly, for I_2, use of (5.4) and Lemma 5.4 leads to

$$\begin{aligned}|I_2| &\le C\left(\sum_{n=1}^{M} k\|\nabla_k z^n\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \\ &\quad\times\left[\left(\sum_{n=1}^{M} k\|\overline{P}_k{}^n - \overline{u}^{\,n}\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} + \left(\sum_{n=1}^{M} k\|\overline{u}^{\,n} - u^n\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}\right] \\ &\le C(k+h^2)\left(\|u\|^2_{L^2(0,T;X)} + \|u_t\|^2_{L^2(0,T;L^2(\Omega))}\right)^{\frac{1}{2}}\left(\sum_{n=1}^{M} k\|\nabla_k z^n\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}} \\ &\le C(k+h^2)\left(\|u\|^2_{L^2(0,T;X)} + \|u_t\|^2_{L^2(0,T;L^2(\Omega))}\right)^{\frac{1}{2}} \\ &\quad\times\left(\sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}. \qquad (5.14)\end{aligned}$$

Finally, using Lemmas 3.1 and 5.4, the term I_4 is estimated as

$$\begin{aligned}|I_4| &\le \|z_+^0\|_{H^1(\Omega)}\|u_0 - R_h u_0\|_{L^2(\Omega)} \\ &\le h^2\|u_0\|_X\left(\sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}. \qquad (5.15)\end{aligned}$$

By simple calculation, it follows that

$$\begin{aligned}\|u-U\|_{L^2(0,T;L^2(\Omega))} &\le Ck\|u_t\|_{L^2(0,T;L^2(\Omega))} \\ &\quad + C\left(\sum_{n=1}^{M} k\|\overline{u}^{\,n} - U^n\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}, \qquad (5.16)\end{aligned}$$

and hence, the desired result now follows from (5.11) and the estimates (5.13)–(5.16). This completes the proof. □

For the H^1 norm estimate, as for the L^2 norm, we analyze the following auxiliary discrete problem: For $1 \le n \le M$, find $w^n \in V_{hk}$ such that

$$(-\nabla_k w^n, v_h) + A(w_+^{n-1}, v_h) = (\nabla(\overline{u}^{\,n} - U^n), \nabla v_h) \quad \forall v_h \in V_{hk}^n, \qquad (5.17)$$

with $w_+^M = 0$. Applying the standard arguments (cf. [5]) and (3.14), we have the following stability result of the solution w satisfying (5.17). This stability result is crucial for the H^1 norm estimate.

LEMMA 5.5. *Let w^n satisfy* (5.17). *Then the following stability results hold:*

$$\max_{1\le n\le M}\|w_+^{n-1}\|^2_{L^2(\Omega)} + \sum_{n=1}^{M} k\|w^{n-1}\|^2_{H^1(\Omega)} \le \sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - U^n)\|^2_{L^2(\Omega)}$$

and

$$\sum_{n=1}^{M} k\|\nabla_k w^n\|^2_{H^{-1}(\Omega)} \le \sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - U^n)\|^2_{L^2(\Omega)}.$$

Proof of Theorem 5.3. Now choose $v_h = k(\overline{P}_k^{\,n} - U^n)$ in (5.17) and repeat the same analysis as for deriving I_4 in (5.12) to obtain

$$\begin{aligned}
&\sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - U^n)\|^2_{L^2(\Omega)} \\
&= \sum_{n=1}^{M} k(\nabla(\overline{u}^{\,n} - \overline{P}_k^{\,n}), \nabla(\overline{u}^{\,n} - U^n)) + \sum_{n=1}^{M} k(-\nabla_k w^n, \overline{P}_k^{\,n} - u^n) \\
&\quad + \sum_{n=1}^{M} \left(k(-\nabla_k w^n, u^n - U^n) + kA(w_+^{n-1}, \overline{u}^{\,n} - U^n)\right) \\
&= \sum_{n=1}^{M} k(\nabla(\overline{u}^{\,n} - \overline{P}_k^{\,n}), \nabla(\overline{u}^{\,n} - U^n)) + \sum_{n=1}^{M} k(-\nabla_k w^n, \overline{P}_k^{\,n} - u^n) \\
&\quad + (w_+^0, u_0 - R_h u_0) \\
&=: II_1 + II_2 + II_3. \qquad (5.18)
\end{aligned}$$

For the term II_1, use the Cauchy–Schwarz inequality and (5.4) to obtain

$$\begin{aligned}
|II_1| &\le \sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - \overline{P}_k^{\,n})\|_{L^2(\Omega)} \|\nabla(\overline{u}^{\,n} - U^n)\|_{L^2(\Omega)} \\
&\le \left\{\sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - \overline{P}_k^{\,n})\|^2_{L^2(\Omega)}\right\}^{\frac{1}{2}} \left\{\sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - U^n)\|^2_{L^2(\Omega)}\right\}^{\frac{1}{2}} \\
&\le Ch\|u\|_{L^2(0,T;X)} \left\{\sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - U^n)\|^2_{L^2(\Omega)}\right\}^{\frac{1}{2}}. \qquad (5.19)
\end{aligned}$$

The term II_2 is estimated in a manner similar to I_2 as in (5.14). Thus, using (5.4) and Lemma 5.5, we obtain

$$\begin{aligned}
|II_2| &\le \left(\sum_{n=1}^{M} k\|\nabla_k w^n\|^2_{H^{-1}(\Omega)}\right)^{\frac{1}{2}} \\
&\quad \cdot \left[\left\{\sum_{n=1}^{M} k\|\overline{u}^{\,n} - \overline{P}_k^{\,n}\|^2_{H^1(\Omega)}\right\}^{\frac{1}{2}} + \left\{\sum_{n=1}^{M} k\|\overline{u}^{\,n} - u^n\|^2_{H^1(\Omega)}\right\}^{\frac{1}{2}}\right] \\
&\le C(k+h)\left(\|u\|^2_{L^2(0,T;X)} + \|u_t\|^2_{L^2(0,T;Y)}\right)^{\frac{1}{2}} \\
&\quad \cdot \left(\sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - U^n)\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}. \qquad (5.20)
\end{aligned}$$

An application of Lemmas 3.1 and 5.5 yields

$$|II_3| \le Ch\|u_0\|_X \left(\sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n} - U^n)\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}}. \qquad (5.21)$$

Again, by an easy calculation (cf. [5]), it follows that

$$(5.22)\ \|u-U\|_{L^2(0,T;H^1(\Omega))} \leq k\|u_t\|_{L^2(0,T;Y)} + \left(\sum_{n=1}^{M} k\|\nabla(\overline{u}^{\,n}-U^n)\|^2_{L^2(\Omega)}\right)^{\frac{1}{2}},$$

and hence, the desired estimate now follows from (5.18)–(5.22). This completes the proof. □

Acknowledgments. The authors wish to thank the anonymous referees for their careful reading of the manuscript and for their valuable comments and suggestions that improved some results as well as the presentation.

REFERENCES

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

[2] I. Babuška, *The finite element method for elliptic equations with discontinuous coefficients*, Computing, 5 (1970), pp. 207–213.

[3] Ch. G. Makridakis and I. Babuška, *On the stability of the discontinuous Galerkin method for the heat equation*, SIAM J. Numer. Anal., 34 (1997), pp. 389–401.

[4] S. C. Brenner and L. R. Scott, *The mathematical theory of finite element methods*, Springer-Verlag, New York, 1994.

[5] Z. Chen and J. Zou, *Finite element methods and their convergence for elliptic and parabolic interface problems*, Numer. Math., 79 (1998), pp. 175–202.

[6] K. Chrysafinos and L. S. Hou, *Error estimates for semidiscrete finite element approximations of linear and semilinear parabolic equations under minimal regularity assumptions*, SIAM J. Numer. Anal., 40 (2002), pp. 282–306.

[7] P. G. Ciarlet, *The finite element method for elliptic problems*, North–Holland, Amsterdam, 1975.

[8] J. Douglas, Jr., and T. Dupont, *Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., 7 (1970), pp. 575–626.

[9] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.

[10] O. A. Ladyzhenskaya, V. Ya. Rivkind, and N. N. Ural'tseva, *The classical solvability of diffraction problems*, Proc. Steklov Inst. Math., 92 (1966), pp. 63–104.

[11] S. Larsson, V. Thomée, and L. B. Wahlbion, *Numerical solution of parabolic integro-differential equations by the discontinuous Galerkin method*, Math. Comp., 67 (1998), pp. 45–71.

[12] M. Luskin and R. Rannacher, *On the smoothing property of the Galerkin method for parabolic equations*, SIAM J. Numer. Anal., 19 (1982), pp. 93–113.

[13] B. F. Nielsen, *Finite element discretizations of elliptic problems in the presence of arbitrarily small ellipticity: An error analysis*, SIAM J. Numer. Anal., 36 (1999), pp. 368–392.

[14] J. T. Oden and L. C. Wellford, *A theory of discontinuous finite element Galerkin approximations of shock waves in nonlinear elastic solids part* 1: *Variational theory*, Comput. Methods Appl. Mech. Engrg., 8 (1976), pp. 1–16.

[15] J. T. Oden and L. C. Wellford, *A theory of discontinuous finite element Galerkin approximations of shock waves in nonlinear elastic solids part* 2: *Accuracy and convergence*, Comput. Methods Appl. Mech. Engrg., 8 (1976), pp. 17–36.

[16] R. K. Sinha and B. Deka, *On the convergence of finite element method for second order elliptic interface problems*, Numer. Funct. Anal. Optim., to appear.

[17] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, New York, 1997.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 750–766

POLYNOMIALS AND POTENTIAL THEORY FOR GAUSSIAN RADIAL BASIS FUNCTION INTERPOLATION*

RODRIGO B. PLATTE† AND TOBIN A. DRISCOLL†

Abstract. We explore a connection between Gaussian radial basis functions and polynomials. Using standard tools of potential theory, we find that these radial functions are susceptible to the Runge phenomenon, not only in the limit of increasingly flat functions, but also in the finite shape parameter case. We show that there exist interpolation node distributions that prevent such phenomena and allow stable approximations. Using polynomials also provides an explicit interpolation formula that avoids the difficulties of inverting interpolation matrices, while not imposing restrictions on the shape parameter or number of points.

Key words. Gaussian radial basis functions, RBF, potential theory, Runge phenomenon, convergence, stability

AMS subject classifications. 65D05, 41A30

DOI. 10.1137/040610143

1. Introduction.

Radial basis functions (RBFs) have been popular for some time in high-dimensional approximation [3] and are increasingly being used in the numerical solution of partial differential equations [7, 14, 17, 20]. Given a set of *centers* $\xi_0, \dots, \xi_N$ in R^d, an RBF approximation takes the form

$$F(x) = \sum_{k=0}^{N} \lambda_k \, \phi\big(\|x - \xi_k\|\big), \tag{1.1}$$

where $\|\cdot\|$ denotes the Euclidean distance between two points and $\phi(r)$ is a function defined for $r \geq 0$. The coefficients $\lambda_0, \dots, \lambda_N$ may be chosen by interpolation or other conditions at a set of *nodes* that typically coincide with the centers. In this article, however, we give special attention to the case in which the locations of centers and nodes differ. Moreover, we shall consider equally spaced centers in most parts of this exposition.

Common choices for ϕ fall into two main categories:

- infinitely smooth and containing a free parameter, such as multiquadrics ($\phi(r) = \sqrt{r^2 + c^2}$) and Gaussians ($\phi(r) = e^{-(r/c)^2}$);
- piecewise smooth and parameter-free, such as cubics ($\phi(r) = r^3$) and thin plate splines ($\phi(r) = r^2 \ln r$).

Convergence analysis of RBF interpolation has been carried out by several researchers—see, e.g., [18, 19, 25]. For smooth ϕ, spectral convergence has been proved for functions belonging to a certain reproducing kernel Hilbert space $\mathcal{F}_\phi$ [19]. This space, however, is rather small since the Fourier transform of functions in $\mathcal{F}_\phi$ must decay very quickly or have compact support [25]. More recently, in [26] Yoon obtained spectral orders on Sobolev spaces, and in [11] error analysis was performed by considering the simplified case of equispaced periodic data. In this article, we use standard

*Received by the editors June 17, 2004; accepted for publication (in revised form) November 18, 2004; published electronically August 31, 2005. This research was supported by National Science Foundation grant DMS-0104229.

http://www.siam.org/journals/sinum/43-2/61014.html

†Department of Mathematical Sciences, University of Delaware, Newark, DE 19716 (platte@math.udel.edu, driscoll@math.udel.edu).

tools of polynomial interpolation and potential theory to study several properties of Gaussian RBF (GRBF) interpolation in one dimension, including convergence and stability.

As is well known in polynomial interpolation, a proper choice of interpolation nodes is essential for good approximations. It is also known that, for fixed N in the limit $c \to \infty$, RBF interpolation is equivalent to polynomial interpolation on the same nodes [6]; hence, the classical Runge phenomenon, and its remedy through node spacing, applies. For practical implementations it is well appreciated that node clustering near the boundaries is helpful [10, 20], but to our knowledge there has been no clear statement about the Runge phenomenon or asymptotically stable interpolation nodes for finite-parameter RBFs. The question has perhaps been obscured somewhat by the fact that the straightforward approach to computing the λ_k is itself numerically ill-conditioned when the underlying approximations are accurate [22].

In this paper we explore the fact that GRBFs with equally spaced centers are related to polynomials through a simple change of variable. Using this connection, in section 2 we demonstrate a Runge phenomenon using GRBFs on equispaced and classical Chebyshev nodes, and we compute asymptotically optimal node densities using potential theory. Numerical calculations suggest that these node densities give Lebesgue constants that grow at logarithmic rates, allowing stable approximations. In section 3 we explore the algorithmic implications of the connections we have made and derive a barycentric interpolation formula that circumvents the difficulty of inverting a poorly conditioned matrix, so approximations can be carried out to machine precision without restrictions on the values of the shape parameter c and number of centers N. Finally, section 4 contains observations on multiquadrics and other possible extensions of the methods presented.

2. Gaussian RBFs as polynomials.

In (1.1) we now choose $d = 1$, Gaussian shape functions, and centers $\xi_k = -1 + 2k/N = -1 + kh$, $k = 0, \ldots, N$. Hence the GRBF approximation is

$$F(x) = \sum_{k=0}^{N} \lambda_k e^{-(x+1-kh)^2/c^2} = e^{-(x+1)^2/c^2} \sum_{k=0}^{N} \lambda_k e^{(2kh-k^2h^2)/c^2} e^{2kxh/c^2}. \tag{2.1}$$

Making the definition $\beta = 2h/c^2 = 4/(Nc^2)$ and using the transformation

$$s = e^{\beta x}, \quad s \in [e^{-\beta}, e^{\beta}],$$

we find that

$$G(s) = F\left(\frac{\log(s)}{\beta}\right) = e^{-\frac{N}{4\beta}(\log s + \beta)^2} \sum_{k=0}^{N} \tilde{\lambda}_k s^k = \psi_\beta^N(s) \sum_{k=0}^{N} \tilde{\lambda}_k s^k, \tag{2.2}$$

where the $\tilde{\lambda}_k$ are independent of s. In this section we regard β as a fixed parameter of the GRBF method. In the literature this is sometimes called the stationary case [2].

From (2.2) it is clear that G/ψ_β^N is a polynomial of degree no greater than N. If F is chosen by interpolation to a given f at $N+1$ nodes, then we can apply standard potential theory to find necessary convergence conditions on the singularities of f in the complex plane $z = x + iy$.

LEMMA 2.1. *Suppose that f is analytic in a closed simply connected region R that lies inside the strip $-\pi/(2\beta) < \mathrm{Im}(z) < \pi/(2\beta)$ and that C is a simple, closed,* *rectifiable curve that lies in R and contains the interpolation points $x_0, x_1, \dots, x_N$. Then the remainder of the GRBF interpolation for f at x can be represented as the contour integral*

$$f(x) - F(x) = \frac{\beta \eta_N(x)}{2\pi i} \int_C \frac{f(z) e^{\beta z}}{\eta_N(z)(e^{\beta z} - e^{\beta x})} dz,$$

where $\eta_N(x) = e^{-\frac{N\beta}{4}(x+1)^2} \prod_{k=0}^{N} (e^{\beta x} - e^{\beta x_k})$.

Proof. Consider the conformal map $w = e^{\beta z}$, and let $g(s) = f(\log(s)/\beta)$. Under this transformation, the region R is mapped to a closed simply connected region that lies in the half-plane $\mathrm{Re}(w) > 0$. Thus g/ψ_β^N is analytic in this region in the w-plane, and we can use the Hermite formula for the error in polynomial interpolation [5],

$$\begin{aligned} g(s) - G(s) &= \psi_\beta^N(s) \left(\frac{g(s)}{\psi_\beta^N(s)} - \sum_{k=0}^{N} \tilde{\lambda}_k s^k \right) \\ &= \frac{\psi_\beta^N(s) \prod_{k=0}^{N}(s - s_k)}{2\pi i} \int_{\mathcal{C}} \frac{g(w)}{(w-s)\psi_\beta^N(w) \prod_{k=0}^{N}(w - s_k)} dw, \end{aligned}$$

where $s_k = e^{\beta x_k}$ and $\mathcal{C}$ is the image of C in the w-plane. A change of variables completes the proof. □

We now turn our attention to necessary conditions for uniform convergence of the interpolation process. To this end, we need the concept of limiting node density functions. These functions describe how the density of node distributions varies over $[-1, 1]$ as $N \to \infty$ [10, 16]. Given a node density function μ, it follows that the node locations x_j satisfy [10]

$$\frac{j}{N} = \int_{-1}^{x_j} \mu(x) dx, \qquad j = 0, \dots, N.$$

Since our analysis parallels the convergence proof for polynomial interpolation (see, e.g., [5, 16, 24]), define

$$u_\beta(z) = \frac{\beta}{4} \mathrm{Re}\left[(z+1)^2\right] - \int_{-1}^{1} \log(|e^{\beta z} - e^{\beta t}|) \mu(t) dt. \tag{2.3}$$

We shall refer to this function as the *logarithmic potential* and to its level curves as *equipotentials.*

In the theorem below we shall assume that μ is such that there exist a and b, $a < b$, with the property that if $K \in [a, b]$, then there exists a simple, closed, rectifiable curve that satisfies $u_\beta(z) = K$ and contains the interval $[-1, 1]$ in its interior. We denote this curve by C_K and by R_K the part of the plane which lies inside it. We also require that if $K_1 > K_2$, then $R_{K_1} \subset R_{K_2}$. To illustrate this feature, consider the logarithmic potential for uniformly distributed nodes on $[-1, 1]$ and $\beta = 1$. In this case we have that $\mu(t) = 1/2$. The level curves of u_1 are presented in Figure 2.1. In this instance one could choose $a = 0.5$ and $b = 0.7$.

THEOREM 2.2. *Suppose μ satisfies the properties above, and let B be the closure of R_b. If f is an analytic function in an open region R which lies inside the strip $-\pi/(2\beta) < \mathrm{Im}(z) < \pi/(2\beta)$ and contains B in its interior, then the GRBF interpolation described above converges uniformly with respect to $z \in B$.*

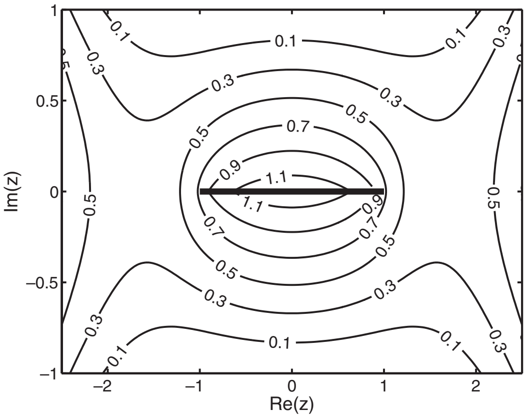


FIG. 2.1. *Level curves of the logarithmic potential for* $\beta = 1$ *and* $\mu(t) = 1/2$. *The straight line represents the interval* $[-1, 1]$.

Proof. Since R is open and B is closed, there exist K_1 and K_2 such that $K_1 < K_2 < b$ and $R_{K_1} \cup C_{K_1}$ lies inside R. Using Lemma 2.1, we have that for any x on C_{K_2},

$$|f(x) - F(x)| \leq \frac{\beta M}{2\pi\delta} \int_{C_{K_1}} \frac{|\eta_N(x)|}{|\eta_N(z)|} dz, \tag{2.4}$$

where M is the largest value of $|f(z)e^{\beta z}|$ on C_{K_1} and δ is the smallest value of $|e^{\beta z} - e^{\beta x}|$ for $z \in C_{K_1}$ and $x \in C_{K_2}$.

We also have that

$$\frac{|\eta_N(x)|}{|\eta_N(z)|} = \exp\left\{-N\left(\log|\eta_N(z)|^{\frac{1}{N}} - \log|\eta_N(x)|^{\frac{1}{N}}\right)\right\}. \tag{2.5}$$

A bound on this exponential can be obtained using the limiting logarithmic potential. Notice that

$$\lim_{N\to\infty} \log|\eta_N(z)|^{\frac{1}{N}} = -u_\beta(z) = -K_1 \quad \text{for } z \in C_{K_1}$$

and

$$\lim_{N\to\infty} \log|\eta_N(x)|^{\frac{1}{N}} = -u_\beta(x) = -K_2 \quad \text{for } x \in C_{K_2}.$$

Hence, for any given ϵ, $0 < \epsilon < (K_2 - K_1)/2$, there exists N_ϵ such that for $N > N_\epsilon$

$$\begin{aligned} &-K_1 - \epsilon < \log|\eta_N(z)|^{\frac{1}{N}} < -K_1 + \epsilon, \\ &-K_2 - \epsilon < \log|\eta_N(x)|^{\frac{1}{N}} < -K_2 + \epsilon, \end{aligned}$$

which implies that

$$\log|\eta_N(z)|^{\frac{1}{N}} - \log|\eta_N(x)|^{\frac{1}{N}} < m_\epsilon, \tag{2.6}$$

where $m_\epsilon = K_2 - K_1 + 2\epsilon > 0$.

Combining (2.4), (2.5), and (2.6) gives

$$|f(x) - F(x)| \leq \frac{\beta M \kappa}{2\pi\delta} e^{-Nm_\epsilon}, \quad N > N_\epsilon,\ x \in C_{K_2}, \tag{2.7}$$

where κ is the length of C_{K_1}.

This last inequality implies that $|f - F| \to 0$ uniformly as $N \to \infty$ on C_{K_2}. Since $f - F$ is analytic in R_{K_2}, by the maximum modulus principle we have that F converges uniformly to f in R_{K_2}. □

We point out that, as happens in polynomial interpolation, the convergence in (2.7) is exponential with a rate that is governed by the equipotentials induced by the nodes.

2.1. The Runge phenomenon. The Runge phenomenon is well understood in polynomial interpolation in one dimension [5, 9] . Even if a function is smooth on the interpolation interval $[-1, 1]$, polynomial interpolants will not converge to it uniformly as $N \to \infty$ unless the function is analytic in a larger complex region whose shape depends on the interpolation nodes. Clustering nodes more densely near the ends of the interval avoids this difficulty. Specifically, points distributed with density $\pi^{-1}(1 - x^2)^{-1/2}$, such as Chebyshev extreme points $x_j = -\cos(j\pi/N)$ and zeros of Chebyshev and Legendre polynomials, are common choices of interpolation nodes on $[-1, 1]$. Uniform convergence of polynomial interpolants is guaranteed for these nodes as long as the function being interpolated is analytic inside an ellipse with foci ± 1 and semiminor larger than δ, for some $\delta > 0$ [9].

In this section we show that for GRBFs uniform convergence may be lost, not only in the polynomial limit $c \to \infty$ but also for constant β (which implies $c \to 0$ as $N \to \infty$), if the distribution of interpolation nodes is not chosen appropriately. Theorem 2.2 can be used to state the regularity requirements of the function being interpolated using a given node distribution, enabling us to determine whether the interpolation process is convergent.

We point out that, for $\beta \ll 1$,

$$u_\beta(z) = -\log(\beta) - \int_{-1}^{1} \log|z - t| \mu(t) dt + O(\beta). \tag{2.8}$$

In this case, the level curves of u_β are similar to equipotentials of polynomial interpolation, and the convergence of the GRBF interpolation process can be predicted from the well-known behavior of polynomial interpolation.

Equipotentials for $\beta = 0.1, 0.8, 2, 5$ are presented in Figure 2.2. On the left of this figure, we present contour maps obtained with a uniform node distribution, and on the right, contour maps obtained with the Chebyshev extreme points. Equipotentials for $\beta = 0.1$ are similar to equipotentials for polynomial interpolation [9], as expected. By Theorem 2.2, convergence is guaranteed if the function is analytic inside the contour line that surrounds the smallest equipotential domain that includes $[-1, 1]$, whereas any singularity inside this region leads to spurious oscillations that usually grow exponentially. Therefore, it is desirable to have the region where the function is required to be analytic be as small as possible. In this sense, we note that for $\beta = 0.1$ the Chebyshev distribution is close to optimal, and for $\beta = 5$ a uniform distribution seems to be more appropriate. We also note that, for large β, Chebyshev density overclusters the nodes near the ends of the interval. In fact, if this clustering is used with $\beta = 5$, even the interpolation of $f \equiv 1$ is unstable; in this case there is no equipotential region that encloses $[-1, 1]$.

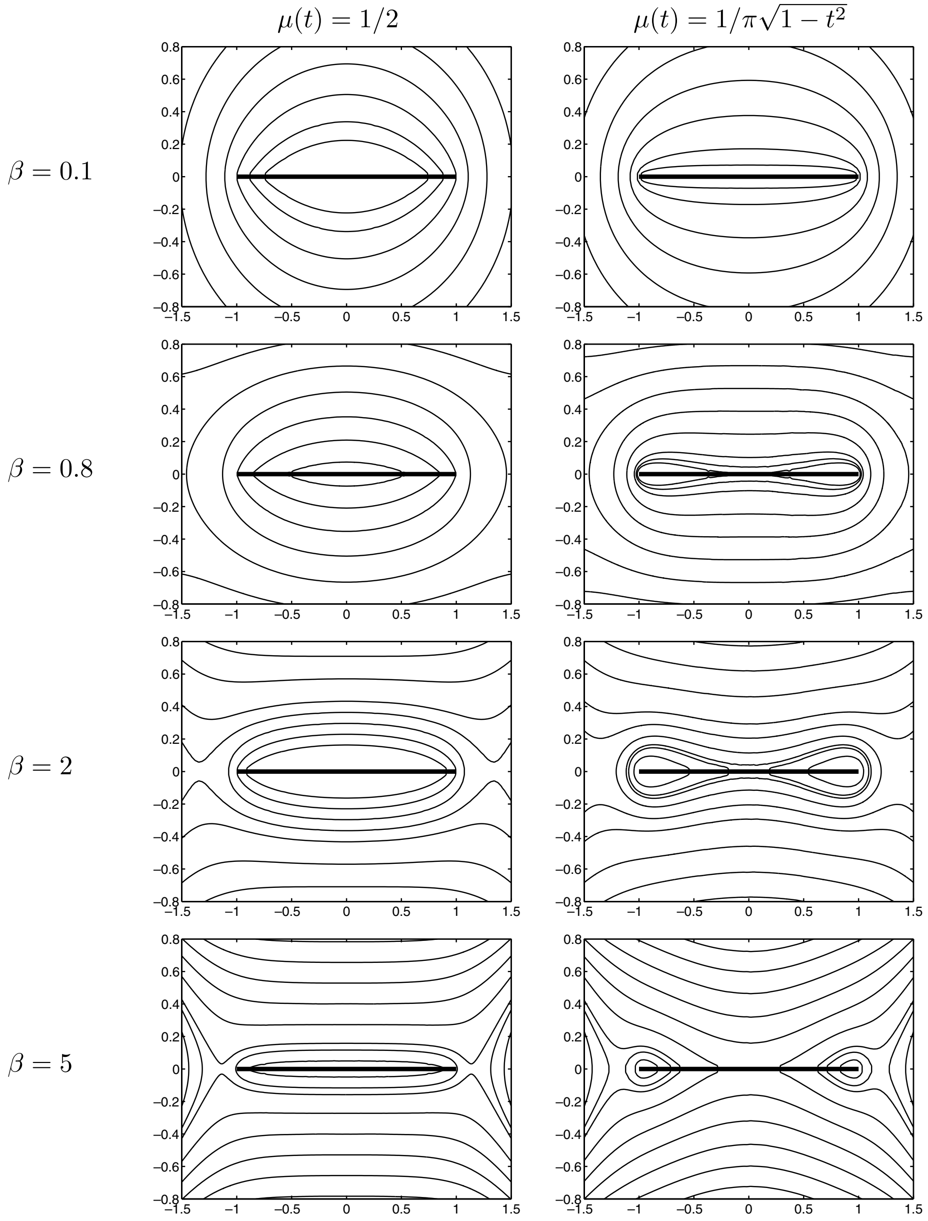
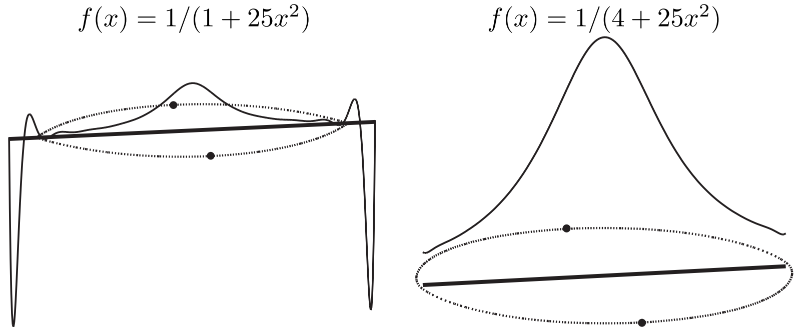


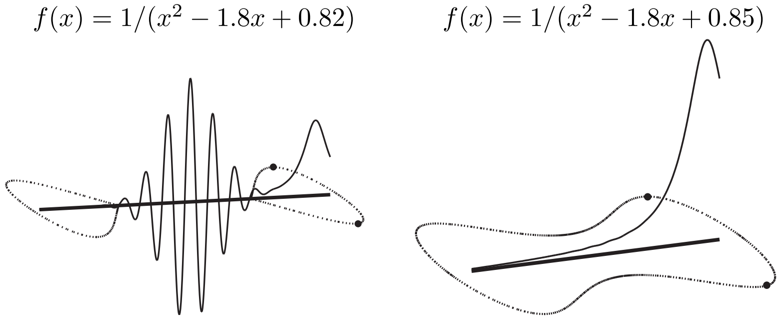
Fig. 2.2. *Contour maps of the logarithmic potential. Plots on the left were obtained with uniform node distribution. Plots on the right were obtained with Chebyshev distribution.*

To demonstrate how the equipotentials and singularities of the interpolated function restrict the convergence of GRBF interpolation, in Figures 2.3 and 2.4 we show two pairs of interpolants. Each pair consists of one function that leads to the Runge phenomenon and one that leads to a stable interpolation process. In Figure 2.3, equispaced nodes were used. The interpolation of $f(x) = 1/(4 + 25x^2)$ is convergent, while the interpolation of $f(x) = 1/(1 + 25x^2)$ is not. Notice from Figure 2.2 that the former function is singular at points inside the smallest equipotential domain,



$f(x) = 1/(1+25x^2)$ $f(x) = 1/(4+25x^2)$

FIG. 2.3. *Interpolation of f with 25 equispaced nodes and $\beta = 0.8$. Closed curves are level curves of the logarithmic potential, dots mark the singularities of f, and straight lines represent the interval* $[-1, 1]$.



$f(x) = 1/(x^2 - 1.8x + 0.82)$ $f(x) = 1/(x^2 - 1.8x + 0.85)$

FIG. 2.4. *Interpolation of f with 41 Chebyshev nodes and $\beta = 2$. Closed curves are level curves of the logarithmic potential, dots mark the singularities of f, and straight lines represent the interval* $[-1, 1]$.

and the singularities of the latter function lie outside this region. For Chebyshev nodes and $\beta = 2$, interpolation of $f(x) = 1/(x^2 - 1.8x + 0.82)$ generates spurious oscillation in the center of the interval. Interpolation of a slightly different function, $f(x) = 1/(x^2 - 1.8x + 0.85)$, gives a well-behaved interpolant.

2.2. Lebesgue constants. Although Theorem 2.2 guarantees convergence for sufficiently smooth functions and properly chosen interpolation points, approximations may not converge in the presence of rounding errors due to the rapid growth of the Lebesgue constant. For GRBF interpolation, we define the Lebesgue constant by

$$\Lambda_N^{GRBF} = \max_{x\in[-1,1]} \sum_{k=0}^{N} |L_k(x)|, \tag{2.9}$$

where

$$L_k(x) = e^{-\frac{N\beta}{4}\left((x+1)^2-(x_k+1)^2\right)} \prod_{\substack{j=0\\ j\neq k}}^{N} \frac{(e^{\beta x} - e^{\beta x_j})}{(e^{\beta x_k} - e^{\beta x_j})} \tag{2.10}$$

is the GRBF cardinal function. Notice that $L_k(x_k) = 1$, $L_k(x_j) = 0$ $(j \neq k)$, and by (2.2), $L_k(x) \in \operatorname{Span}\{e^{-(x-\xi_k)^2/c^2}\}$. Thus, the unique GRBF interpolant can be

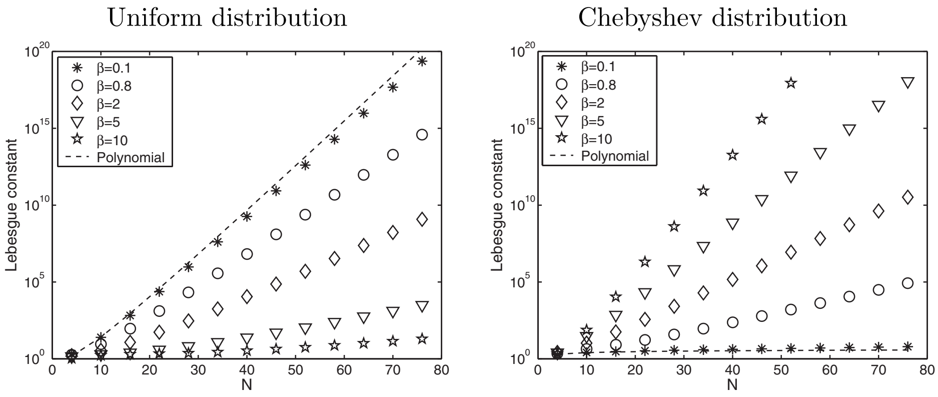


FIG. 2.5. *Lebesgue constant for different values of β. Dashed lines mark the Lebesgue constant values for polynomial interpolation.*

written as

$$F(x) = \sum_{k=0}^{N} L_k(x) f(x_k), \tag{2.11}$$

and it follows that

$$\|F - f\|_\infty \le (1 + \Lambda_N^{GRBF}) \|F^{opt} - f\|_\infty, \tag{2.12}$$

where F^{opt} is the best approximation to f in the GRBF subspace with respect to the infinity norm.

Figure 2.5 illustrates how the GRBF Lebesgue constant grows with N for equispaced nodes (left) and Chebyshev nodes (right). As expected, for small β the GRBF Lebesgue constants approximate the polynomial Lebesgue constants, which behave asymptotically as $O(2^N/(N \log N))$ for equispaced nodes and $O(\log N)$ for Chebyshev nodes [9, 23]. This figure shows that the Lebesgue constants grow exponentially for both node distributions, except for large values of β for uniform nodes and small values of β for Chebyshev nodes.

In the presence of rounding errors, (2.12) indicates that if computations are carried out with precision ε, then the solution will generally be contaminated by errors of size $\varepsilon \Lambda_N^{GRBF}$ [23]. For instance, if $f(x) = 1/(x^2 - 1.8x + 0.85)$ and $\beta = 2$, the convergence of the interpolation process on Chebyshev nodes in double precision stops at $N = 80$, with a minimum residue of $O(10^{-7})$ due to rounding error. Similar results have been observed on equispaced nodes if β is small.

2.3. Stable interpolation nodes. Our goal now is to find node distributions that lead to a convergent interpolation process whenever the function is analytic on $[-1, 1]$. This happens only if $[-1, 1]$ is itself an equipotential, as is the case for Chebyshev density in polynomial interpolation. Therefore, we seek a density function μ that satisfies

$$\frac{\beta}{4}(x+1)^2 = \int_{-1}^{1} \log(|e^{\beta x} - e^{\beta t}|)\mu(t)dt \ + \ \text{constant}, \quad x \in [-1, 1]. \tag{2.13}$$

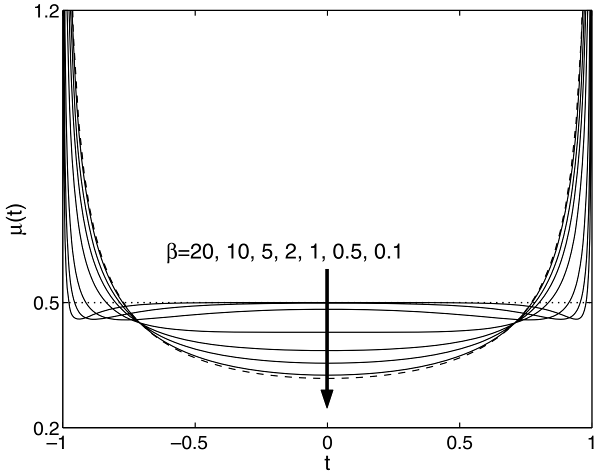


FIG. 2.6. *Numerical approximations of the optimal density functions for several values of β. The dashed line shows the Chebyshev density function.*

In order to find a numerical solution to this integral equation, we assume that the optimal μ can be approximated by

$$\mu(t) \cong \sum_{k=0}^{N_\mu} a_k \frac{T_{2k}(t)}{\sqrt{1-t^2}}, \tag{2.14}$$

where T_{2k} is the Chebyshev polynomial of order $2k$. We consider only even functions in our expansion because we expect the density function to be even due to symmetry. This generalizes the Chebyshev density function $\mu(t) = \pi^{-1}(1-t^2)^{-1/2}$. We also tried more general expressions, replacing $\sqrt{1-t^2}$ with $(1-t^2)^{-\alpha}$, and found that $\alpha = 1/2$ was suitable.

Figure 2.6 shows density functions computed with the expression above. We computed the coefficients a_k by discrete least-squares, and the integral in (2.13) was approximated by Gaussian quadrature. We used $N_\mu = 9$ and 50 points to evaluate the residue in the least-squares process. With this choice of parameters, the residual was less than 10^{-7} in all computations.

In Figure 2.7 we show 21 nodes computed using (2.13) and (2.14) for several values of β. For large values of β the nodes are nearly equally spaced, and for small values they are approximately equal to Chebyshev extreme points. The optimal equipotentials obtained for $\beta = 0.1, 0.8, 2, 5$ are presented in Figure 2.8. For all these values of β, $[-1, 1]$ seems to be a level curve of the logarithmic potential.

As mentioned in section 2.2, in the presence of rounding errors the Lebesgue constant also plays a crucial role. Fortunately, for the optimal nodes computed numerically in this section, experiments suggest that the Lebesgue constant grows at a logarithmic rate. Figure 2.9 presents computed Lebesgue constants for different values of β on optimal nodes.

Figure 2.10 shows the convergence of the GRBF interpolation to the four functions used to illustrate the Runge phenomenon in section 2.1. Now all four functions can be approximated nearly to machine precision. The algorithm used to obtain these

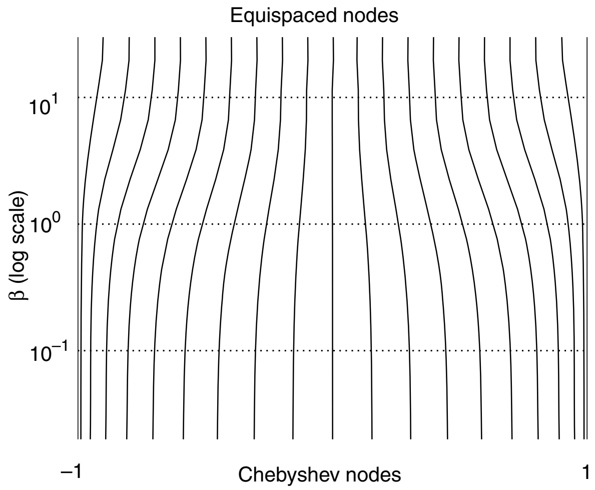


FIG. 2.7. *Node locations obtained using a density function computed by solving the integral equation* (2.13) *for* $N = 20$ *and several values of* β.

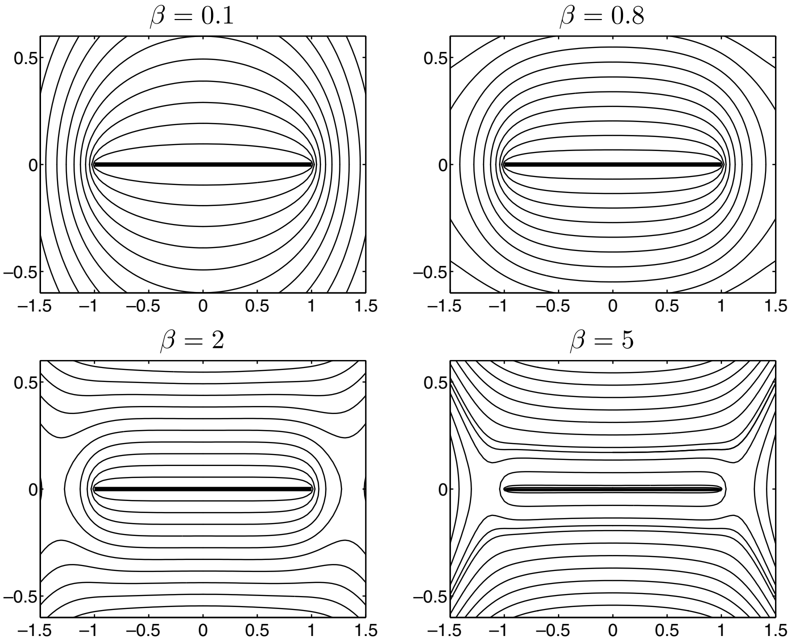


FIG. 2.8. *Contour maps of the logarithmic potential obtained with a numerically approximated optimal density function.*

data is presented in section 3. Notice that the convergence rates are determined by the singularities of the function being interpolated. Dashed lines in this figure mark the convergence rates predicted by (2.7). For instance, if $f(x) = 1/(1 + 25x^2)$ and $\beta = 0.8$, then m_ϵ is approximately the difference between the value of the potential in $[-1, 1]$ and the potential at $z = 0.2i$ (where f is singular), giving $m_\epsilon \cong 0.23$.

Notice that for $\beta = 2$ the equipotentials that enclose the interval $[-1, 1]$ are

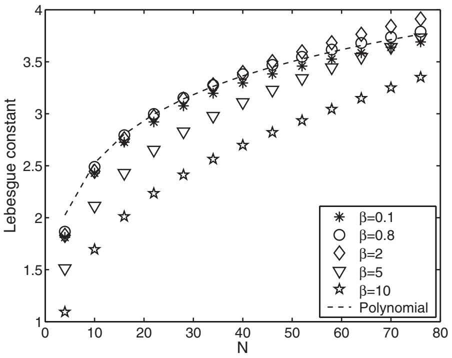
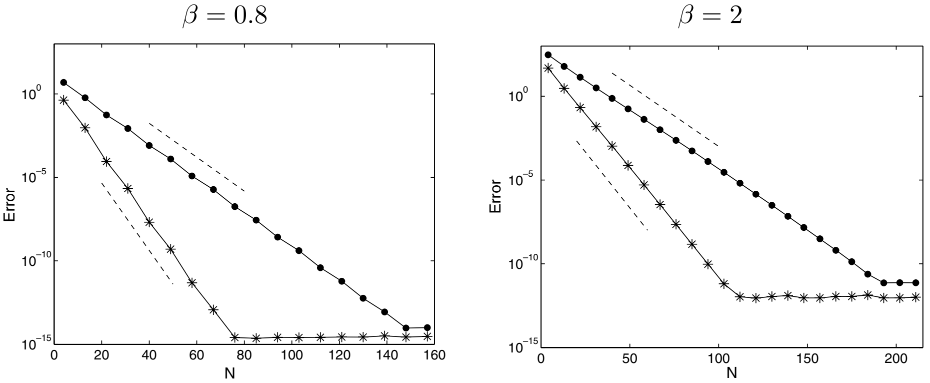


FIG. 2.9. *Lebesgue constant for different values of β and optimal node distribution.*


$\beta = 0.8$

$\beta = 2$

FIG. 2.10. *Maximum error of the interpolation process using optimal nodes. Left:* $f(x) = 1/(1+25x^2)$ *(*$\bullet$*) and* $f(x) = 1/(4+25x^2)$ *(*$*$*). Right:* $f(x) = 1/(x^2 - 1.8x + 0.82)$ *(*$\bullet$*) and* $f(x) = 1/(x^2 - 1.8x + 0.85)$ *(*$*$*). Dashed lines mark convergence rates predicted by* (2.7).

contained in a bounded region (Figure 2.8). This indicates that the convergence rate given by (2.7) is the same for *all* functions that have singularities outside this region. In polynomial interpolation, convergence to entire functions is much faster than to functions with finite singularities. This is not the case for GRBFs. With $\beta = 2$ we found that the rate of convergence of interpolants of $1/(1+4x^2)$, $1/(100+x^2)$, $\sin(x)$, and $|x+2|$ were all about the same. What these functions have in common is that they are analytic inside the smallest region that includes all equipotentials that enclose $[-1,1]$.

It is also worth noting that the one-parameter family μ_γ of node density functions proportional to $(1-t^2)^{-\gamma}$ [9] was used in [10] and [20] to cluster nodes near boundaries in RBF approximations. Although numerical results there showed improvement in accuracy, no clear criteria for choosing γ was provided in those papers. By using these node density functions and minimizing the residue in (2.13) with respect to γ, we found that optimal values of γ are approximately given by $\gamma \cong 0.5e^{-0.3\beta}$. We point out, however, that interpolations using these density functions may not converge if large values of N are required.

2.4. Location of centers. Up to this point we have assumed that the centers are uniformly distributed on $[-1,1]$. Here we briefly investigate the consequences of

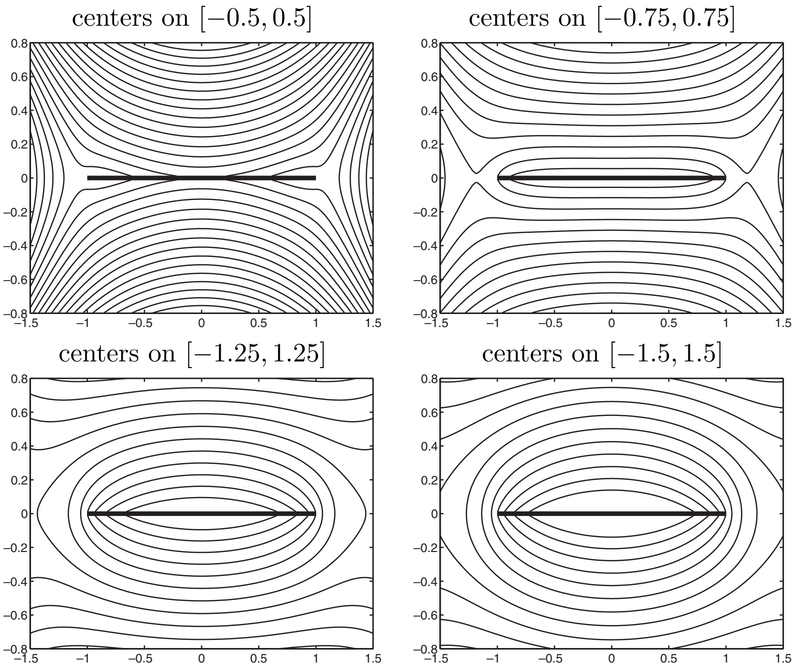


FIG. 2.11. *Equipotentials for $\beta = 2$ (compare with Figure 2.2). Uniformly distributed centers on interval specified above. Interpolation points are uniformly distributed on* $[-1, 1]$.

choosing centers ξ_k that are equispaced on the interval $[-L, L]$, where $L \neq 1$, and also discuss results where centers are not equally spaced. Taking centers outside the interval of approximation, as was suggested in [10, 17] to improve edge accuracy, is of practical interest.

For equispaced centers on $[-L, L]$, a straightforward modification of (2.2) gives

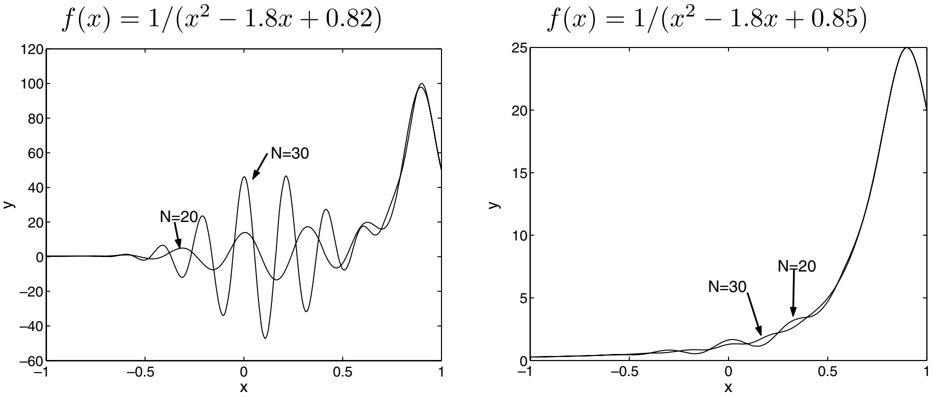
$$F(x) = e^{\frac{-N\beta}{4L}(x+L)^2} \sum_{k=0}^{N} \tilde{\lambda}_k e^{k\beta x},$$

where $\beta = 4L/Nc^2$. In this case the logarithmic potential becomes

$$u_\beta^L(z) = \frac{\beta}{4L} \operatorname{Re}\left[(z+L)^2\right] - \int_{-1}^{1} \log(|e^{\beta z} - e^{\beta t}|)\mu(t)dt.$$

Equipotentials for different values of L are presented in Figure 2.11. We considered equispaced interpolation nodes on $[-1, 1]$. Notice that if $L = 0.5$, there is no guarantee of convergence, as no equipotential encloses $[-1, 1]$. For $L =$0.75, 1.25, and 1.5, there are equipotentials enclosing this interval. The region where f is required to be smooth seems to increase with L. We also point out that the asymptotic behavior for small β, given in (2.8), holds independently of L, indicating that center location is irrelevant in the polynomial limit.

It is common practice to choose the same nodes for centers and interpolation. In Figure 2.12 we show the graphs of the GRBF interpolants, for $f(x) = 1/(x^2 - 1.8x + 0.82)$ and $f(x) = 1/(x^2 - 1.8x + 0.85)$, where both centers and interpolation nodes are Chebyshev points. These data suggest that interpolation with Chebyshev centers also


$f(x) = 1/(x^2 - 1.8x + 0.82)$

$f(x) = 1/(x^2 - 1.8x + 0.85)$

FIG. 2.12. *GRBF interpolation using Chebyshev points for centers and interpolation nodes, $\beta = 2$.*

suffers from the Runge phenomenon. These results are similar to the ones obtained in Figure 2.4 for equispaced centers. Notice that we cannot use the definition involving h for β if the centers are not equispaced; in this case we use the definition $\beta = 4/(Nc^2)$.

3. Algorithmic implications. It is well known that most RBF-based algorithms suffer from ill-conditioning. The interpolation matrix $[\phi(\|x_i - \xi_j\|)]$ in most conditions becomes ill-conditioned as the approximations get more accurate, to the extent that global interpolants are rarely computed for more than a couple of hundred nodes. Based on numerical and theoretical observations, in [22] Schaback states that for RBFs, "Either one goes for a small error and gets a bad sensitivity, or one wants a stable algorithm and has to take a comparably larger error." Several researchers have addressed this issue [4, 8, 15, 21]. In particular, Fornberg and Wright [12] recently presented a contour-integral approach that allows numerically stable computations of RBF interpolants for all values of the free parameter c, but this technique is expensive and has been applied only for experimental purposes.

For GRBFs with equispaced centers, (2.11) provides an explicit interpolation formula through the use of the cardinal functions L_k, so the difficulty of inverting the interpolation matrix can be avoided. This is equivalent to Lagrange polynomial interpolation.

Notice that the exponential term $e^{-\frac{N\beta}{4}((x+1)^2-(x_k+1)^2)}$ in (2.10) becomes very close to zero for certain values of x if N is large, affecting the accuracy of the approximations. A simple modification of (2.10) improves matters:

$$L_k(x) = \prod_{\substack{j=0 \\ j\neq k}}^{N} \frac{e^{-\frac{\beta}{4}\left((x+1)^2-(x_k+1)^2\right)}\left(e^{\beta x} - e^{\beta x_j}\right)}{\left(e^{\beta x_k} - e^{\beta x_j}\right)}. \tag{3.1}$$

The direct implementation of (3.1) together with (2.11) provides a simple algorithm for computing the GRBF interpolant for moderate values of N. In our experiments, effective computations were carried out up to $N = 300$. We shall next derive a more stable formula to handle larger problems.

In [1] Berrut and Trefethen point out the difficulties of using the standard Lagrange formula for practical computations and argue that the barycentric form of Lagrange interpolation should be the method of choice for polynomial interpolation.

For GRBFs we define the barycentric weights by

$$w_k = \left(\prod_{\substack{j=0 \\ j \neq k}}^{N} e^{-\frac{\beta}{4}(x_k+1)^2} \left(e^{\beta x_k} - e^{\beta x_j} \right) \right)^{-1}, \tag{3.2}$$

and thus we have that

$$L_k(x) = L(x) \frac{w_k}{e^{-\frac{\beta}{4}(x+1)^2} \left(e^{\beta x} - e^{\beta x_k} \right)} \qquad (x \neq x_k),$$

where

$$L(x) = \prod_{j=0}^{N} e^{-\frac{\beta}{4}(x+1)^2} \left(e^{\beta x} - e^{\beta x_j} \right).$$

Therefore, the GRBF interpolant can be written as

$$F(x) = L(x) \sum_{k=0}^{N} \frac{w_k}{e^{-\frac{\beta}{4}(x+1)^2} \left(e^{\beta x} - e^{\beta x_k} \right)} f(x_k). \tag{3.3}$$

For reasons of numerical stability, it is desirable to write L as a sum involving the barycentric weights. For polynomial interpolation this is done by considering that 1 can be exactly written in terms of interpolation formulas, since it is itself a polynomial. Unfortunately, a constant function is not exactly represented in terms of GRBFs. Nevertheless, this difficulty can be circumvented if we properly choose a function that belongs to the GRBF space. In our implementation, we consider the function

$$v(x) = \frac{1}{N} \sum_{k=0}^{N} e^{-\frac{N\beta}{4}(x-\xi_k)^2}.$$

Notice that in this case,

$$L(x) = \frac{v(x)}{\sum_{k=0}^{N} \frac{w_k}{e^{-\frac{\beta}{4}(x+1)^2} \left(e^{\beta x} - e^{\beta x_k} \right)} v(x_k)}.$$

Combining the last expression with (3.3) gives our *GRBF barycentric formula*:

$$F(x) = v(x) \frac{\sum_{k=0}^{N} \frac{w_k}{\left(e^{\beta x} - e^{\beta x_k} \right)} f(x_k)}{\sum_{k=0}^{N} \frac{w_k}{\left(e^{\beta x} - e^{\beta x_k} \right)} v(x_k)}. \tag{3.4}$$

As mentioned in [1], the fact that the weights w_k appear symmetrically in the denominator and in the numerator means that any common factor in all the weights may be canceled without affecting the value of F. In some cases it is necessary to rescale terms in (3.2) to avoid overflow. In our implementation we divided each term by $\prod_{j=1}^{N} |e^{\beta x_j} - e^{-\beta}|^{1/N}$.

In [13] Higham shows that for polynomials the barycentric formula is forward stable for any set of interpolation points with a small Lebesgue constant. Our numerical experiments suggest that the GRBF barycentric formula is also stable.

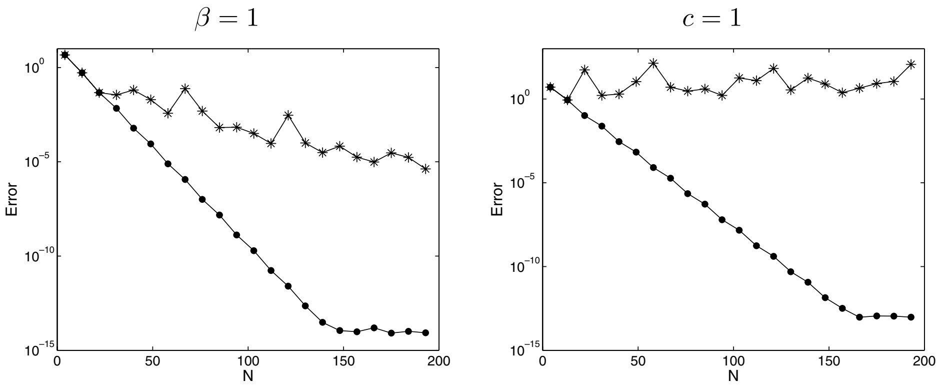


FIG. 3.1. *Maximum error of the interpolation of $f(x) = 1/(1+25x^2)$ using barycentric interpolation (●) and the standard RBF algorithm ($*$). Left: β fixed. Right: c fixed.*

Figure 2.10 was obtained using the barycentric formula. We point out that the direct inversion of the interpolation matrix becomes unstable even for moderate values of N. In Figure 3.1 we compare the convergence of the GRBF interpolant computed with the barycentric formula with that found by inverting the interpolation matrix (standard RBF algorithm). We first computed approximations with β fixed (left graph). Notice that for the standard implementation, convergence rate changes at a level around 10^{-2}, and the method becomes very inefficient for larger values of N. For the barycentric formula, on the other hand, convergence continues to machine precision. For these approximations we used nodes computed with an approximate optimal density function, as in section 2.3.

We also compared the algorithms for fixed c (right graph). In this instance we used Chebyshev nodes, as c constant implies that $\beta \to 0$ as N becomes large and approximations become polynomial. The performance of the standard algorithm is even worse in this case.

4. Final remarks. GRBFs using equally spaced centers are easily related to polynomials in a transformed variable through (2.2). This connection allows us to apply polynomial interpolation and potential theory to draw a number of precise conclusions about the convergence of GRBF interpolation. In particular, for a given interpolation node density, one can derive spectral convergence (or divergence) rates based on the singularity locations of the target function. Conversely, one can easily compute node densities for which analyticity of the function in $[-1, 1]$ is sufficient for convergence and for which the Lebesgue constant is controlled. Furthermore, the polynomial connection allows us to exploit barycentric Lagrange interpolation to construct a simple explicit interpolation algorithm that avoids the ill-conditioning of the interpolation matrix. We stress that the convergence illustrated in Figure 3.1 is made possible only through the use of *both* the stable nodes and the stable algorithm.

Numerical evidence suggests that other RBFs such as multiquadrics may also be susceptible to the Runge phenomenon and dependent on node location for numerically stable interpolations. Figure 4.1 shows graphs of multiquadric interpolants of two functions. We first considered the small β case (nearly polynomial) with the same function that caused the Runge phenomenon for GRBFs on equispaced nodes. The high oscillations of the interpolant at the ends of the interval indicates that this function also causes the Runge phenomenon for multiquadrics. The multiquadric

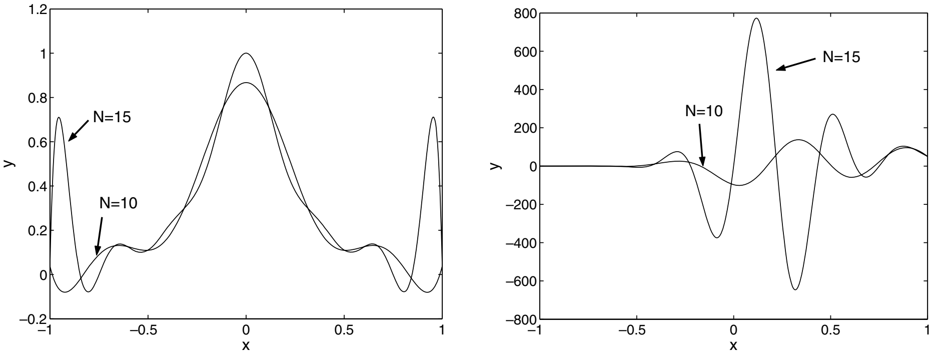


Fig. 4.1. *Runge phenomenon in multiquadric RBF interpolation. Left: Interpolation of $f(x) = 1/(1+25x^2)$ using equispaced nodes and $\beta = 0.1$. Right: Interpolation of $f(x) = 1/(x^2-1.8x+0.82)$ using Chebyshev nodes and $\beta = 2$.*

interpolant of $f(x) = 1/(x^2 - 1.8x + 0.82)$ with $\beta = 2$ and equispaced centers also presented spurious oscillations, as its GRBF counterpart did, when Chebyshev interpolation nodes were used.

Practical interest in RBF methods is fueled by their flexibility in the node and center locations and by their simple use in higher-dimensional approximation. The results of this paper do not extend immediately in either of those directions, except to a tensor-product situation of uniform center locations in a box. Still, we believe that the explicit GRBF interpolation algorithm, in particular, may be adaptable to selective resolution requirements and geometric flexibility.

REFERENCES

[1] J.-P. Berrut and L. N. Trefethen, *Barycentric Lagrange interpolation*, SIAM Rev., 46 (2004), pp. 501–517.

[2] M. Bozzini, L. Lenarduzzi, and R. Schaback, *Adaptive interpolation by scaled multiquadrics*, Adv. Comput. Math., 16 (2002), pp. 375–387.

[3] M. D. Buhmann, *Radial Basis Functions*, Cambridge University Press, Cambridge, UK, 2003.

[4] A. H.-D Cheng, M. A. Golberg, E. J. Kansa, and G. Zammito, *Exponential convergence and H-c multiquadric collocation method for partial differential equations*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 571–594.

[5] P. J. Davis, *Interpolation and Approximation*, Dover, New York, 1975.

[6] T. A. Driscoll and B. Fornberg, *Interpolation in the limit of increasingly flat radial basis functions*, Comput. Math. Appl., 43 (2002), pp. 413–422.

[7] G. E. Fasshauer, *Solving partial differential equations by collocation with radial basis functions*, in Surface Fitting and Multiresolution Methods, A. LeMéhauté, C. Rabut, and L. Schumaker, eds, Vanderbilt University Press, Nashville, TN, 1997, pp. 131–138.

[8] G. E. Fasshauer, *Solving differential equations with radial basis functions: Multilevel methods and smoothing*, Adv. Comput. Math., 11 (1999), pp. 139–159.

[9] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Cambridge University Press, New York, 1996.

[10] B. Fornberg, T. A. Driscoll, G. Wright, and R. Charles, *Observations on the behavior of radial basis function approximations near boundaries*, Comput. Math. Appl., 43 (2002), pp. 473–490.

[11] B. Fornberg and N. Flyer, *Accuracy of radial basis function interpolation and derivative approximations on 1-D infinite grids*, Adv. Comput. Math., 23 (2005), pp. 5–20.

[12] B. Fornberg and G. Wright, *Stable computation of multiquadric interpolants for all values of the shape parameter*, Comput. Math. Appl., 47 (2004), pp. 497–523.

[13] N. J. Higham, *The numerical stability of barycentric Lagrange interpolation*, IMA J. Numer. Anal., 24 (2004), pp. 547–556.

[14] E. J. Kansa, *Multiquadrics—A scattered data approximation scheme with applications to computational fluid dynamics* II. *Solutions to hyperbolic, parabolic, and elliptic partial differential equations*, Comput. Math. Appl., 19 (1990), pp. 147–161.

[15] E. J. Kansa and Y. C. Hon, *Circumventing the ill-conditioning problem with multiquadric radial basis functions: Applications to elliptic partial differential equations*, Comput. Math. Appl., 39 (2000), pp. 123–137.

[16] V. I. Krylov, *Approximate Calculation of Integrals*, A. H. Stroud, trans., Macmillan, New York, 1962.

[17] E. Larsson and B. Fornberg, *A numerical study of some radial basis function based solution methods for elliptic PDEs*, Comput. Math. Appl., 46 (2003), pp. 891–902.

[18] W. R. Madych, *Miscellaneous error bounds for multiquadric and related interpolators*, Comput. Math. Appl., 24 (1992), pp. 121–138.

[19] W. R. Madych and S. A. Nelson, *Bounds on multivariate polynomials and exponential error estimates for multiquadric interpolation*, J. Approx. Theory, 70 (1992), pp. 94–114.

[20] R. B. Platte and T. A. Driscoll, *Computing eigenmodes of elliptic operators using radial basis functions*, Comput. Math. Appl., 48 (2004), pp. 561–576.

[21] S. Rippa, *An algorithm for selecting a good value for the parameter c in radial basis function interpolation*, Adv. Comput. Math., 11 (1999), pp. 193–210.

[22] R. Schaback, *Error estimates and condition numbers for radial basis function interpolation*, Adv. Comput. Math., 3 (1995), pp. 251–264.

[23] L. N. Trefethen and J. A. C. Weideman, *Two results on polynomial interpolation in equally spaced points*, J. Approx. Theory, 65 (1991), pp. 247–260.

[24] J. A. C. Weideman and L. N. Trefethen, *The eigenvalues of second-order spectral differentiation matrices*, SIAM J. Numer. Anal., 25 (1988), pp. 1279–1298.

[25] H. Wendland, *Gaussian interpolation revisited*, in Trends in Approximation Theory, K. Kopotun, T. Lyche, and N. Neamtu, eds., Vanderbilt University Press, Nashville, TN, 2001, pp. 1–10.

[26] J. Yoon, *Spectral approximation orders of radial basis function interpolation on the Sobolev space*, SIAM J. Math. Anal., 33 (2001), pp. 946–958.

SIAM J. Numer. Anal.
Vol. 43, No. 2, pp. 767–786

ANALYSIS OF REGULARIZATION METHODS FOR THE SOLUTION OF ILL-POSED PROBLEMS INVOLVING DISCONTINUOUS OPERATORS*

F. FRÜHAUF†, O. SCHERZER†, and A. LEITÃO‡

Abstract. We consider a regularization concept for the solution of ill-posed operator equations, where the operator is composed of a continuous and a discontinuous operator. A particular application is level set regularization, where we develop a novel concept of minimizers. The proposed level set regularization is capable of handling changing topologies. A functional analytic framework explaining the splitting of topologies is given. The asymptotic limit of the level set regularization method is an evolution process, which is implemented numerically, and the quality of the proposed algorithm is demonstrated by solving an inverse source problem.

Key words. ill-posed problems, regularization methods, level set methods

AMS subject classifications. 65J20, 47A52

DOI. 10.1137/S0036142903430906

1. Introduction.

The goal of this paper is to analyze *regularization* models for the *stable* solution of *ill-posed* operator equations

$$F(P(\phi)) = y. \tag{1.1}$$

Here F is a continuous operator between Banach spaces X and Y, and P is a probably *discontinuous operator* into an admissible class $\mathcal{P} \subset X$. Classical results on convergence and stability of variational regularization principles for solving *nonlinear* ill-posed problems (see, e.g., [19, 20, 10]) in a Hilbert spaces setting such as

1. existence of a regularized solution,
2. stability of the regularized approximations,
3. approximation properties of the regularized solutions

are applicable if the operator P is

1. bounded and linear or
2. nonlinear, continuous, and weakly closed.

In this paper we particularly emphasize operator equations (1.1), where the operator P is discontinuous. Of particular interest for this paper is

$$P(t) := \begin{cases} 0 & \text{for} \quad t < 0, \\ 1 & \text{for} \quad t \geq 0. \end{cases} \tag{1.2}$$

*Received by the editors July 3, 2003; accepted for publication (in revised form) October 13, 2004; published electronically August 31, 2005.
http://www.siam.org/journals/sinum/43-2/43090.html

†Department of Computer Science, Universität Innsbruck, Technikerstraße 25, A-6020 Innsbruck, Austria (Florian.Frühauf@uibk.ac.at, Otmar.Scherzer@uibk.ac.at). The work of the first author was supported by the Tiroler Zukunftsstiftung. The work of the second author was partly supported by the FWF (Austrian Science Foundation) through grant Y-123 INF.

‡Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, c/o Kepler Universität Linz, A-4040 Linz, Austria (Antonio.Leitao@oeaw.ac.at). The work of this author was supported by the Austrian Academy of Sciences and CNPq via grant 305823/2003-5. This author is on leave from the Department of Mathematics, Federal University of St. Catarina, Brazil.

With P there is associated the *admissible class*

$$\mathcal{P} := \left\{ u : u = \chi_D, \text{ where } D \subseteq \Omega \text{ is measurable and } \mathcal{H}^{n-1}(\partial D) < \infty \right\}. \tag{1.3}$$

Here

1. $\mathcal{H}^{n-1}(\partial D)$ denotes the $n-1$-dimensional Hausdorff-measure of the boundary of D;
2. χ_D denotes the characteristic function of the set D.

We refer to a regularization approach involving this projection as a *level set regularization* since we recover the boundary of an object ∂D, which is a level set (for instance, with value 0) of a function ϕ. The idea of considering characteristic functions as level sets of higher-dimensional data has been used before in the context of *multiphase flow* (see, e.g., [18, 26, 8]) and segmentation (see, e.g., [7]). Level set methods have been used successively in many applications since the pioneering work of Osher and Sethian [22]. For solving inverse problems applications with level sets, we refer to Santosa [24] and Burger [5].

In this work we base our considerations on ideas from nonlinear convex semigroup theory (cf. Brezis [4]), which allows us to characterize the solution of an evolution process by implicit time steps of regularization models. Since our regularization models appear to be nonconvex, the theoretical results of nonlinear semigroup theory are not available. Simulating this approach, we show in this work that iterated regularization is well posed, and (aside from the lack of theoretical results) we can interpret the iterated regularized solutions as time instance of an evolution process.

Various other models fit in the general framework of this paper but are not particularly emphasized: for instance, for $a \in \mathbb{R}$ let us consider the projection operator

$$P^a(t) := \begin{cases} -a & \text{for} \quad t < -a, \\ t & \text{for} \quad -a \leq t \leq a, \\ a & \text{for} \quad t > a, \end{cases}$$

with the admissible class

$$\mathcal{P}_a := \left\{ u : u = P^a(\phi) \text{ with } \phi \in H^1(\Omega) \right\}. \tag{1.4}$$

The operator P^a ensures that the recovered functions are absolutely bounded by a. The operator

$$P^+(t) := \exp(t)$$

with the *admissible class*

$$\mathcal{P}_+ := \left\{ u : 0 < u = P^+(\phi) \right\} \tag{1.5}$$

can be used to guarantee *nonnegativity.* Depending on the operator P, we actually solve a constraint optimization problem. With P_+, P_a, P we guarantee that the solution is in the corresponding admissible class.

The outline of this paper is as follows. In section 2 we introduce the concept of *level set regularization*, based on considerations in [24, 5, 17]. The level set regularization functionals derived in [17] are modified such that a convergence analysis becomes tractable (cf. section 2.1). That is, we show that each implicit time step is well defined. This a prerequisite step in showing that the corresponding gradient flow equation (cf. section 2.1) is well defined. To this end, we introduce a novel concept of a minimizer of regularization functionals involving discontinuous operators (cf. section 2.2). A convergence analysis is presented in section 2.3. The problem of numerical minimization is discussed in section 3, and in section 4 a relation to iterative regularization is considered. Finally, numerical examples are presented in section 5.

2. Analysis of level set regularization. In the following we pose the general assumptions which we assume to hold throughout this paper:

1. $\Omega \subseteq \mathbb{R}^n$ is bounded with $\partial\Omega$ piecewise C^1 (see, e.g., [2]).
2. The operator $F : L^1(\Omega) \to Y$ is continuous and Fréchet-differentiable. Y is a Banach space.
3. $\varepsilon, \alpha, \beta$ denote positive parameters.
4. We use the following notation:
   (i) $\to$ denotes strong convergence.
   (ii) $\overset{(*)}{\rightharpoonup}$ denotes weak(*) convergence.
   (iii) $L^p(\Omega)$ denotes the space of measurable p-times-integrable functions.
   (iv) $W^{1,p}(\Omega)$ denotes the Sobolev space of one time weakly differentiable functions where the function and its derivative are in L^p; in particular we set $H^1 = W^{1,2}$.
   (v) $\mathtt{BV}(\Omega)$ denotes the space of functions of *bounded variation*.
5. We assume that (1.1) has a solution; i.e., there exists a $z \in \mathcal{P}$ satisfying $F(z) = y$ and a function $\phi \in H^1(\Omega)$ satisfying $|\nabla\phi| \neq 0$ in a neighborhood of $\{\phi = 0\}$ and $P(\phi) = z$. If $z = \chi_A$ and $\emptyset \neq A$, then we let

$$\phi = -d_{\overline{A}} + d_{\overline{CA}},$$

where $d_{\overline{A}}$ and $d_{\overline{CA}}$ denote the distance functions from $\overline{A}$, and $\overline{CA}$, respectively. Since $d_{\overline{A}}$ and $d_{\overline{CA}}$ are uniformly Lipschitz-continuous (see, e.g., [9]), they are in $L^\infty(\Omega)$. Moreover, $|\nabla d_{\overline{A}}| \leq 1$ and $|\nabla d_{\overline{CA}}| \leq 1$ (see again, e.g., [9]). In particular this shows that $d_{\overline{A}}, d_{\overline{CA}} \in W^{1,\infty}(\Omega) \subseteq H^1(\Omega)$. Thus $z \in \mathcal{P}$ if $z = \chi_A$ and A satisfies that the closure of the interior of A is the closure of A.

We consider the unconstrained inverse problem of solving (1.1) with

$$\begin{aligned} P : H^1(\Omega) &\to \mathcal{P}\ , \\ \phi &\mapsto \frac{1}{2} + \frac{1}{2}\mathrm{sgn}(\phi) =: \frac{1}{2} + \frac{1}{2}\left\{ \begin{array}{rl} 1 & \text{for } \phi \geq 0, \\ -1 & \text{for } \phi < 0. \end{array} \right. \end{aligned}$$

The standard form of Tikhonov regularization for solving (1.1) consists of minimizing the functional

$$\mathcal{F}_\alpha(\phi) := \|F(P(\phi)) - y^\delta\|_Y^2 + \alpha\|\phi - \phi_0\|_{H^1(\Omega)}^2 \tag{2.1}$$

over $H^1(\Omega)$. Actually, we understand the minimizer ϕ_α of this functional as

$$\phi_\alpha = \lim_{\varepsilon\to 0+} \phi_{\varepsilon,\alpha},$$

where the limit is understood in an appropriate sense (weak, weak* convergence) and $\phi_{\varepsilon,\alpha}$ minimizes the functional over $H^1(\Omega)$:

$$\mathcal{F}_{\varepsilon,\alpha}(\phi) := \|F(P_\varepsilon(\phi)) - y^\delta\|_Y^2 + \alpha\|\phi - \phi_0\|_{H^1(\Omega)}^2, \tag{2.2}$$

where we use

$$P_\varepsilon(\phi) := \begin{cases} 0 & \text{for} \quad \phi < -\varepsilon, \\ 1 + \frac{\phi}{\varepsilon} & \text{for} \quad \phi \in [-\varepsilon, 0], \\ 1 & \text{for} \quad \phi > 0, \end{cases}$$

for approximating P as $\varepsilon \to 0^+$. In this case we define

$$P'(t) := \lim_{\varepsilon \to 0+} P'_\varepsilon(t) = \delta(t).$$

Here and in the following, $\delta(t)$ denotes the one-dimensional δ-distribution.

Taking into account that

$$\|P_\varepsilon(\phi_k) - P_\varepsilon(\phi)\|_{L^1(\Omega)} \leq \frac{1}{\varepsilon}\sqrt{\text{meas}(\Omega)}\|\phi_k - \phi\|_{L^2(\Omega)},$$

the proof of existence of a minimizer of the functional $\mathcal{F}_{\varepsilon,\alpha}$ is similar to the proof of existence of regularized solutions of Tikhonov functionals for approximately minimizing nonlinear ill-posed problems in [11, 25] (see also [10]).

THEOREM 2.1. *For any $\phi_0 \in H^1(\Omega)$ the functional $\mathcal{F}_{\varepsilon,\alpha}$ (cf. (2.2)) attains a minimizer $\phi_{\varepsilon,\alpha}$ in $H^1(\Omega)$.*

2.1. Towards an analysis of level set regularization techniques. In the following we outline the difficulties in performing a rigorous analysis for the functional $\mathcal{F}_\alpha$, defined in (2.1).

1. $\phi_{\varepsilon,\alpha}$ satisfies

$$\|P(\phi_{\varepsilon,\alpha})\|_{L^\infty} \leq 1 \quad \text{and} \quad \|\phi_{\varepsilon,\alpha} - \phi_0\|_{H^1(\Omega)} < \infty.$$

Since $L^\infty(\Omega)$ is the dual of $L^1(\Omega)$, i.e., $L^{1*}(\Omega) = L^\infty(\Omega)$, we find that there exists a subsequence $\{\phi_{\varepsilon_k,\alpha_k}\}_{k\in\mathbb{N}}$ such that

$$\phi_{\varepsilon_k,\alpha_k} \rightharpoonup \phi \text{ in } H^1(\Omega) \quad \text{and} \quad P(\phi_{\varepsilon_k,\alpha_k}) \overset{*}{\rightharpoonup} z \text{ in } L^\infty(\Omega).$$

There is no analytical evidence for $z \in \mathcal{P}$; i.e., it may not be in the range of the operator P.

2. To overcome this difficulty let us assume that the sequence $\{\phi_{\varepsilon_k,\alpha_k}\}_{k\in\mathbb{N}}$ satisfies the condition that the Hausdorff-measure of the boundary of the set

$$\{x : \phi_{\varepsilon_k,\alpha_k}(x) \geq 0\}$$

is uniformly bounded. Then the bounded variation seminorm of $P(\phi_{\varepsilon_k,\alpha_k})$ is uniformly bounded, and consequently $P(\phi_{\varepsilon_k,\alpha_k})$ has a convergent subsequence in $L^1(\Omega)$, showing that z is admissible.

This suggests that we incorporate in the functional (2.1) as an additional regularization term the bounded variation seminorm of $P(\phi)$, penalizing the length of the zero level set of ϕ. Actually in design problems the necessity of incorporating such a term is well documented in [14, 15, 16]. This leads to the following modified regularization method of minimizing

$$\mathcal{G}_\alpha(\phi) := \|F(P(\phi)) - y^\delta\|_Y^2 + 2\beta\alpha|P(\phi)|_{\mathtt{BV}} + \alpha\|\phi - \phi_0\|_{H^1(\Omega)}^2. \tag{2.3}$$

In order to guarantee existence of a minimizer of $\mathcal{G}_\alpha$ we introduce a novel concept of a minimizer in the next subsection.

2.2. Minimizing concept.

Definition 2.2. 1. *A* pair *of functions*

$$(z, \phi) \in L^\infty(\Omega) \times H^1(\Omega)$$

is called admissible

(i) *if there exists a sequence $\{\phi_k\}_{k\in\mathbb{N}}$ in $H^1(\Omega)$ such that $\phi_k \to \phi$ with respect to the $L^2(\Omega)$-norm and*

(ii) *if there exists a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ of positive numbers converging to zero such that*

$$P_{\varepsilon_k}(\phi_k) \to z \textit{ in } L^1(\Omega).$$

2. *A minimizer of $\mathcal{G}_\alpha$ is considered any admissible pair of functions (z, ϕ) minimizing*

$$\mathcal{G}_\alpha(z, \phi) = \|F(z) - y^\delta\|_Y^2 + \alpha\rho(z, \phi) \tag{2.4}$$

over all admissible pairs. Here

$$\rho(z, \phi) := \inf \liminf_{k\to\infty} \left\{ 2\beta |P_{\varepsilon_k}(\phi_k)|_{\mathtt{BV}} + \|\phi_k - \phi_0\|_{H^1(\Omega)}^2 \right\}, \tag{2.5}$$

where the infimum is taken with respect to all sequences $\{\varepsilon_k\}_{k\in\mathbb{N}}$ satisfying item 1(ii) *and $\{\phi_k\}_{k\in\mathbb{N}}$ satisfying item* 1(i).

A generalized minimizer of $\mathcal{G}_\alpha(\phi)$ is a minimizer of $\mathcal{G}_\alpha(z, \phi)$ on the set of admissible pairs.

The following lemma shows that the functional ρ is coercive on the set of admissible pairs.

Lemma 2.3. *For each (z, ϕ) admissible,*

$$2\beta |z|_{\mathtt{BV}} + \|\phi - \phi_0\|_{H^1(\Omega)}^2 \le \rho(z, \phi).$$

Proof. Let (z, ϕ) be an admissible pair; then there exists sequences $\{\varepsilon_k\}_{k\in\mathbb{N}}$ and $\{\phi_k\}_{k\in\mathbb{N}}$ satisfying items 1(i) and 1(ii) and

$$\rho(z, \phi) = \lim_{k\to\infty} 2\beta |P_{\varepsilon_k}(\phi_k)|_{\mathtt{BV}} + \|\phi_k - \phi_0\|_{H^1(\Omega)}^2.$$

By the weak lower semicontinuity of the $\mathtt{BV}$ and H^1-norms, it follows that

$$\|\phi - \phi_0\|_{H^1(\Omega)}^2 \le \liminf_{k\in\mathbb{N}} \|\phi_k - \phi_0\|_{H^1(\Omega)}^2,$$
$$|z|_{\mathtt{BV}} \le \liminf_{k\in\mathbb{N}} |P_{\varepsilon_k}(\phi_k)|_{\mathtt{BV}},$$

which proves the assertion. □

The definition of $\rho(z, \phi)$ is impractical, since it is defined via a relaxation procedure. The following arguments allow an explicit characterization of this functional. From several experiments which we outline below, we conjecture the following characterization of the functional $\rho(z, \phi)$.

Conjecture 2.4. *We define*

$$\Phi_+ = \{x \in \Omega : \phi(x) > 0\} \quad \textit{and} \quad \Phi_- = \{x \in \Omega : \phi(x) < 0\}$$

and

$$C\Phi = \Omega \backslash (\Phi_+ \cup \Phi_-).$$

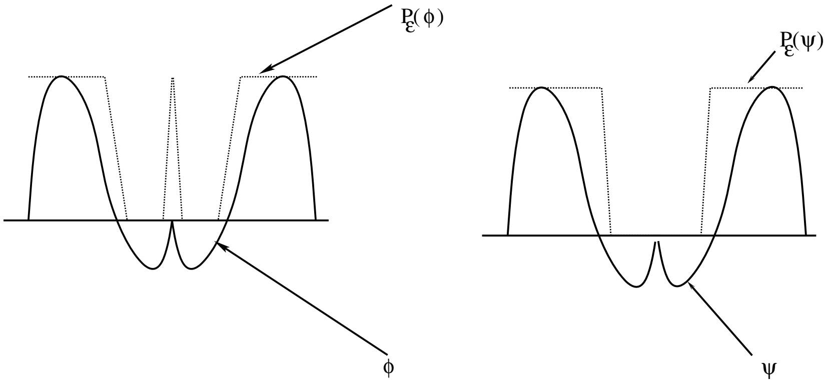


FIG. 2.1. *(Left)* $n = 1$*: The functions* ϕ *and* $P_\varepsilon(\phi)$*:* $|P_\varepsilon(\phi)|_{\mathtt{BV}} = 4$*. (Right) A slight perturbation:* ψ *and* $P_\varepsilon(\psi)$*:* $|P_\varepsilon(\psi)|_{\mathtt{BV}} = 2$*.*

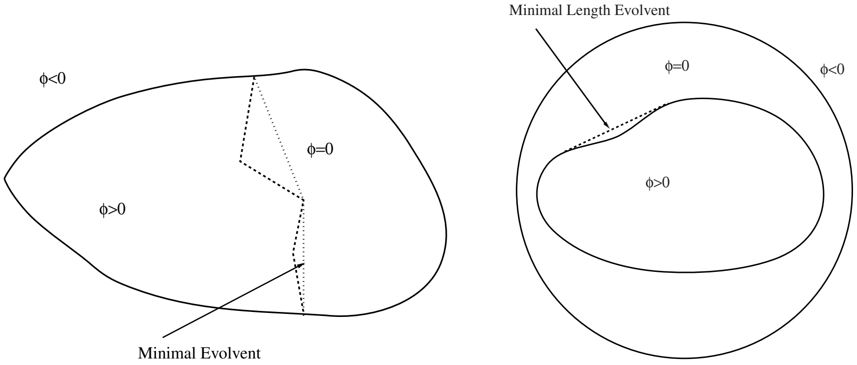


FIG. 2.2. *The minimal evolvent in* $C\Phi$*.*

(i) *If* $\partial\Phi_+ \cap \Omega = \partial\Phi_- \cap \Omega$*, then*

$$\begin{aligned}\rho(z, \phi) &= 2\beta\mathcal{H}^{n-1}(\partial\Phi_- \cap \Omega) + \|\phi - \phi_0\|^2_{H^1(\Omega)}\\ &= 2\beta\mathcal{H}^{n-1}(\partial\Phi_+ \cap \Omega) + \|\phi - \phi_0\|^2_{H^1(\Omega)}\ .\end{aligned}$$

(ii) *If the* n*-dimensional Lebesgue measure* $\lambda^n(C\Phi) > 0$*, then* z *is not uniquely identified; in particular,* z *can attain all values in* $[0, 1]$ *in* $C\Phi$*. We conjecture that*

$$\inf_{z\ admissible} \rho(z, \phi) = 2\beta\mathcal{H}^{n-1}(S) + \|\phi - \phi_0\|^2_{H^1(\Omega)}.$$

The problem consists of finding the surface S *of minimal* $n-1$*-dimensional Hausdorff-measure, which is contained in* $C\Phi$ *and divides* Ω *in two sets. One set completely contains* Φ_+ *and the other set contains* Φ_- *(cf. Figures* 2.1 *and* 2.2*).*

Intuitively the conjecture is quite obvious. Assuming the conjecture to be true, we are further led to conjecture that the functional ρ is independent of the choice of the approximation P_ε. Thus any other approximation of P with Lipschitz-continuous functions P_ε approximating the δ-distribution is suitable as well.

Remark 2.5. For $\phi \in H^1(\Omega)$, where $\{\phi = 0\}$ is a set of positive Lebesgue measure (cf. Figure 2.3), it is possible to find sequences $\{\phi_k\}_{k\in\mathbb{N}}$ and $\{\tilde{\phi}_k\}_{k\in\mathbb{N}}$, which converge

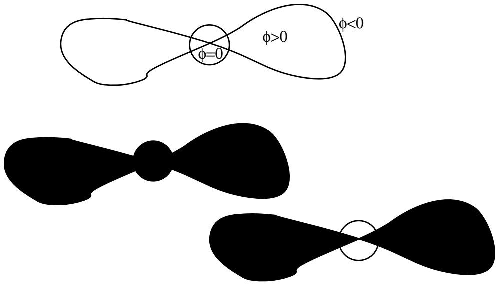


FIG. 2.3. Top: *The level set function has critical values (i.e.,* $|\nabla\phi| = 0$ *in a circle).* Bottom: *Two possible functions* z *and* $\tilde{z}$*. The black value corresponds to a value of* $z = 1$*.*

strongly to ϕ in $L^2(\Omega)$. However, the limits of the projections are different; i.e., $z = \lim_{k\to\infty} P_{\varepsilon_k}(\phi_k) \neq \tilde{z} = \lim_{k\to\infty} P_{\varepsilon_k}(\tilde{\phi}_k)$; cf. Figure 2.3. In such a situation we have $\rho(z,\phi) \neq \rho(\tilde{z},\phi)$.

In the following we summarize some properties of the functional ρ.

LEMMA 2.6. *The functional* ρ *satisfies*

$$\rho(z,\phi) \leq \liminf_{n\in\mathbb{N}} \rho(z_n,\phi_n)$$

if $z_n \to z$ *in* $L^1(\Omega)$ *and if* $\phi_n \rightharpoonup \phi$ *in* $H^1(\Omega)$ *and* (z_n,ϕ_n) *is admissible.*

Proof. From the definition of the functional ρ and Lemma 2.3 it follows that the functional ρ is a Γ^--limit (see, e.g., [3]), and thus we conclude that it is weak lower semicontinuous. □

Remark 2.7. Suppose for the moment that P is a continuous operator, in which case we can set $P_\varepsilon := P$. Then the admissible class is just the set of pairs (z,ϕ) satisfying $P(\phi) = z$. This is just another formulation of constraint optimization. In our context P is discontinuous, and therefore we consider the more general concept of admissible pairs.

Example 2.8. Let $\phi \in H^1(\Omega)$ satisfying $|\nabla\phi| > 0$ in a neighborhood of $\{\phi = 0\}$.

1. Let $\phi_k = \phi \in H^1(\Omega)$ and let $z = P(\phi_k)$. Since for any sequence $\varepsilon_k \to 0$

$$P_{\varepsilon_k}(\phi_k) \to z \quad \text{in } L^1(\Omega),$$

it follows that (z,ϕ) is admissible.

2. Let $\phi \in H^1(\Omega)$ and define $\phi_k = \frac{1}{k}\phi$. Then there is a sequence $\varepsilon_k \to 0$ with

$$P_{\varepsilon_k}(\phi_k) \to z \quad \text{in } L^1(\Omega).$$

Consequently, $(z,0)$ is admissible.

The consequence of the second item is striking. Suppose that $\phi_0 = 0$ and that there exists a minimizer $\phi_\alpha \neq 0$ of (2.3). Then for any $k \in \mathbb{N}$

$$\mathcal{G}_\alpha(\phi_\alpha/k) < \mathcal{G}_\alpha(\phi)\,,$$

showing that a minimizer of $\mathcal{G}_\alpha$ is not attained in a common setting. However, the pair $(z = P(\phi_\alpha), 0)$ is admissible and can be considered as the generalized solution.

Note that in this example we consider only functions $\phi \in H^1(\Omega)$ without critical points along the zero level set.

2.3. Well-posedness and convergence analysis.

THEOREM 2.9 (well-posedness). *Both the functional $\mathcal{G}_\alpha$ and the functional*

$$\tilde{\mathcal{G}_\alpha}(z,\phi) := \|F(z) - y^\delta\|_Y^2 + 2\beta\alpha|z|_{\mathtt{BV}} + \alpha\|\phi - \phi_0\|_{H^1(\Omega)}^2$$

attain minimizers on the set of admissible pairs.

Proof. 1. Since $(0,0)$ is admissible, the set of admissible pairs is not empty.

2. Suppose that $\{(z_k,\phi_k)\}_{k\in\mathbb{N}}$ is a sequence of admissible pairs such that

$$\mathcal{G}_\alpha(z_k,\phi_k) \to \inf \mathcal{G}_\alpha \le \mathcal{G}_\alpha(0,0) < \infty.$$

From Lemma 2.3 it follows that $\{(z_k,\phi_k)\}_{k\in\mathbb{N}}$ is uniformly bounded in $\mathtt{BV} \times H^1(\Omega)$. By the Sobolev embedding theorem there exists a subsequence, denoted again by $\{\phi_k\}_{k\in\mathbb{N}}$, such that

$$\begin{gathered}\phi_k \rightharpoonup \phi \text{ in } H^1(\Omega) \quad \text{and} \quad \phi_k \to \phi \text{ in } L^2(\Omega),\\ z_k \to z \text{ in } L^1(\Omega), \qquad 2\beta|z|_{\mathtt{BV}} \le \rho(z,\phi) \le \liminf_{k\to\infty} \rho(z_k,\phi_k).\end{gathered}$$

Since ρ is weakly lower semicontinuous (cf. Lemma 2.6) it follows that

$$\begin{aligned}\inf \mathcal{G}_\alpha &= \lim_{k\to\infty} \mathcal{G}_\alpha(z_k,\phi_k)\\ &= \lim_{k\to\infty} \left\{\|F(z_k) - y^\delta\|_Y^2 + \alpha\rho(z_k,\phi_k)\right\}\\ &\ge \|F(z) - y^\delta\|_Y^2 + \alpha\rho(z,\phi)\\ &= \mathcal{G}_\alpha(z,\phi).\end{aligned} \tag{2.6}$$

3. It remains to prove that (z,ϕ) is admissible. For k fixed, since (z_k,ϕ_k) is admissible, there exists a sequence $\{\varepsilon_{k,l}\}_{l\in\mathbb{N}}$ of positive numbers and a sequence $\{\phi_{k,l}\}_{l\in\mathbb{N}}$ in $H^1(\Omega)$ such that

$$\phi_{k,l} \to_{l\to\infty} \phi_k \text{ in } L^2(\Omega), \quad P_{\varepsilon_{k,l}}(\phi_{k,l}) \to_{l\to\infty} z_k \text{ in } L^1(\Omega).$$

Thus there exists an index $l(k) \in \mathbb{N}$ such that

(i) $\varepsilon_{k,l(k)} < \frac{1}{2}\varepsilon_{k-1,l(k-1)}$;

(ii) $\|\phi_{k,l(k)} - \phi_k\|_{L^2(\Omega)} \le \frac{1}{k}$;

(iii) $\|P_{\varepsilon_{k,l(k)}}(\phi_{k,l(k)}) - z_k\|_{L^1(\Omega)} \le \frac{1}{k}$.

Define

$$\psi_k := \phi_{k,l(k)} \quad \text{and} \quad \eta_k := \varepsilon_{k,l(k)}.$$

Then, since

$$\psi_k \to \phi \text{ in } L^2(\Omega) \quad \text{and} \quad P_{\eta_k}(\psi_k) \to z \text{ in } L^1(\Omega),$$

we see that (z,ϕ) is admissible.

The proof of existence of a minimizer of $\tilde{\mathcal{G}_\alpha}$ is analogous to that for $\mathcal{G}_\alpha$ and is thus omitted. □

We have shown that for any positive parameters α,β the functionals $\mathcal{G}_\alpha$ and $\tilde{\mathcal{G}_\alpha}$ both attain a minimizer.

In what follows we denote by (z_α,ϕ_α) a minimizer of $\mathcal{G}_\alpha$.

In the following we summarize some convergence results for regularized minimizers, which are based on the existence of a *minimum norm solution.*

THEOREM 2.10 (existence of a minimum norm solution). *Under the general assumptions of this paper there exists a minimum norm solution* $(z^\dagger, \phi^\dagger)$*, that is, an admissible pair of functions that satisfies*

1. $F(z^\dagger) = y$,
2. $\rho(z^\dagger, \phi^\dagger) = ms := \inf\{\rho(z,\phi) : (z,\phi) \text{ admissible and } F(z) = y\}$.

Proof. 1. According to assumption 5 in section 2, there exists a function $\tilde{z} \in \mathcal{P}$ and a function $\tilde{\phi} \in H^1(\Omega)$ such that $P(\tilde{\phi}) = \tilde{z}$ and $F(\tilde{z}) = y$. Then the pair $(\tilde{z}, \tilde{\phi})$ is admissible for the sequence $\tilde{\phi}_k = \tilde{\phi}$, because $P_{\varepsilon_k}(\tilde{\phi}_k) \to \tilde{z}$ converges in $L^1(\Omega)$ for every sequence $\varepsilon_k \to 0$ due to the fact that P_{ε_k} is a convolution of P with a δ-distribution; i.e., $P_{\varepsilon_k} = P * \delta_k$. Thus the set of admissible pairs with $F(z) = y$ is not empty.

2. Suppose that $\{(z_k, \phi_k)\}_{k\in\mathbb{N}}$ is a sequence of admissible pairs with $F(z_k) = y$ such that

$$\rho(z_k, \phi_k) \to \mathrm{ms} \le \rho(\tilde{z}, \tilde{\phi}) < \infty.$$

From Lemma 2.3 it follows that the sequences $\{\phi_k\}_{k\in\mathbb{N}}$ and $\{z_k\}_{k\in\mathbb{N}}$ are uniformly bounded in $H^1(\Omega)$ and $\mathtt{BV}(\Omega)$, respectively. Thus there exists subsequences, again denoted by $\{\phi_k\}_{k\in\mathbb{N}}$ and $\{z_k\}_{k\in\mathbb{N}}$, such that

$$\phi_k \to \phi^\dagger \text{ in } L^2(\Omega), \qquad z_k \to z^\dagger \text{ in } L^1(\Omega).$$

Since ρ is weakly lower semicontinuous, it follows that

$$\mathrm{ms} = \lim_{k\to\infty} \rho(z_k, \phi_k) \ge \rho(z^\dagger, \phi^\dagger).$$

Since F is continuous on $L^1(\Omega)$, $F(z^\dagger) = \lim_{k\to\infty} F(z_k) = y$. Analogous to the proof of Theorem 2.9 it follows that $(z^\dagger, \phi^\dagger)$ is admissible and therefore a minimal norm solution. □

Below, we summarize a stability and convergence result. The proof uses classical techniques from the analysis of Tikhonov-type regularization methods (e.g., see [11, 25, 1, 10, 21]) and thus is omitted.

THEOREM 2.11 (convergence and stability). *Let* $\|y^\delta - y\|_Y \le \delta$. *If* $\alpha = \alpha(\delta)$ *satisfies*

$$\lim_{\delta\to 0} \alpha(\delta) = 0 \quad \textit{and} \quad \lim_{\delta\to 0} \frac{\delta^2}{\alpha(\delta)} = 0,$$

then, for a sequence $\{\delta_k\}_{k\in\mathbb{N}}$ *converging to* 0*, there exists a sequence* $\{\alpha_k := \alpha(\delta_k)\}_{k\in\mathbb{N}}$ *such that* $(z_{\alpha_k}, \phi_{\alpha_k})$ *converges in* $L^1(\Omega) \times L^2(\Omega)$ *to a minimal norm solution.*

3. Numerical solution.

We consider a stabilized functional

$$\mathcal{G}_{\varepsilon,\alpha}(\phi) := \|F(P_\varepsilon(\phi)) - y^\delta\|_Y^2 + 2\beta\alpha|P_\varepsilon(\phi)|_{\mathtt{BV}} + \alpha\|\phi - \phi_0\|_{H^1(\Omega)}^2. \tag{3.1}$$

This functional is well posed, as the following lemma shows.

LEMMA 3.1. *For any* $\phi_0 \in H^1(\Omega)$ *the functional* (3.1) *attains a minimizer.*

Proof. The proof is similar to that of Theorem 2.1, taking into account that, for any sequence $\{\phi_k\}_{k\in\mathbb{N}}$ converging weakly to ϕ in the $H^1(\Omega)$-norm, there exists a strongly convergent subsequence in $L^2(\Omega)$. Denoting the subsequence again by $\{\phi_k\}_{k\in\mathbb{N}}$, we find

1.

$$\|P_\varepsilon(\phi_k) - P_\varepsilon(\phi)\|_{L^1(\Omega)} \le \frac{1}{\varepsilon}\sqrt{\mathrm{meas}(\Omega)}\|\phi_k - \phi\|_{L^2(\Omega)} \to 0.$$

2. Therefore

$$|P_\varepsilon(\phi)|_{\mathtt{BV}} \le \liminf_{k\to\infty} |P_\varepsilon(\phi_k)|_{\mathtt{BV}}.$$

Now, the assertion can be proved as for Theorem 2.1. □

In the following we show that for $\varepsilon \to 0$ the minimizer of $\mathcal{G}_{\varepsilon,\alpha}$ approximates a minimizer of $\mathcal{G}_\alpha$; i.e., it approximates an admissible pair.

THEOREM 3.2. *Let $\phi_{\varepsilon,\alpha}$ be a minimizer of $\mathcal{G}_{\varepsilon,\alpha}$. Then for $\varepsilon_k \to 0$ there exists a convergent subsequence $(P_{\varepsilon_k}(\phi_{\varepsilon_k,\alpha}), \phi_{\varepsilon_k,\alpha}) \to (\tilde{z}, \tilde{\phi})$ in $L^1(\Omega) \times L^2(\Omega)$, and the limit minimizes $\mathcal{G}_\alpha$ in the set of admissible pairs.*

Proof. 1. The infimum of $\mathcal{G}_\alpha$ is attained (cf. Theorem 2.9); i.e., there exists (z_α, ϕ_α) minimizing $\mathcal{G}_\alpha$ over all admissible pairs. In particular, taking into account the definition of admissible pairs, there exists a sequence $\{\varepsilon_k\}_{k\in\mathbb{N}}$ of positive numbers converging to zero and a corresponding sequence $\{\phi_k\}_{k\in\mathbb{N}}$ in $H^1(\Omega)$ satisfying

$$\begin{aligned}(P_{\varepsilon_k}(\phi_k), \phi_k) &\to (z_\alpha, \phi_\alpha) \quad \text{in } L^1(\Omega)\times L^2(\Omega),\\ \rho(z_\alpha, \phi_\alpha) &= \lim_{k\to\infty} \left\{ 2\beta |P_{\varepsilon_k}(\phi_k)|_{\mathtt{BV}} + \|\phi_k - \phi_0\|^2_{H^1(\Omega)} \right\}.\end{aligned}$$

2. Let ϕ_{ε_k} be a minimizer of $\mathcal{G}_{\varepsilon_k,\alpha}$. The sequence $\{\phi_{\varepsilon_k}\}_{k\in\mathbb{N}}$ is uniformly bounded in $H^1(\Omega)$. Thus it has a weakly convergent subsequence (which is again denoted by the same indices), and the weak limit is denoted $\tilde{\phi}$. Moreover, $\{P_{\varepsilon_k}(\phi_{\varepsilon_k})\}_{k\in\mathbb{N}}$ is uniformly bounded in $\mathtt{BV}(\Omega)$. Thus, by the compact Sobolev embedding theorem there exists a subsequence $\{\phi_{\varepsilon_k}\}_{k\in\mathbb{N}}$ (again denoted with the same indices) satisfying

$$\phi_{\varepsilon_k} \to \tilde{\phi} \text{ in } L^2(\Omega) \quad \text{and} \quad P_{\varepsilon_k}(\phi_{\varepsilon_k}) \to \tilde{z} \text{ in } L^1(\Omega).$$

Thus $(\tilde{z}, \tilde{\phi}) \in \mathcal{P} \times H^1(\Omega)$ is admissible.

3. From the definition of ρ and the continuity of $F : L^1(\Omega) \to Y$ it follows that

$$\begin{aligned}\|F(\tilde{z}) - y^\delta\|^2_Y &= \lim_{k\to\infty} \|F(P_{\varepsilon_k}(\phi_{\varepsilon_k})) - y^\delta\|^2_Y,\\ \rho(\tilde{z}, \tilde{\phi}) &\le \liminf_{k\to\infty} \left\{ 2\beta |P_{\varepsilon_k}(\phi_{\varepsilon_k})|_{\mathtt{BV}} + \|\phi_{\varepsilon_k} - \phi_0\|^2_{H^1(\Omega)} \right\}.\end{aligned}$$

This shows that

$$\begin{aligned}\mathcal{G}_\alpha(\tilde{z}, \tilde{\phi}) &\le \liminf_{k\to\infty} \mathcal{G}_{\varepsilon_k,\alpha}(\phi_{\varepsilon_k})\\ &\le \liminf_{k\to\infty} \mathcal{G}_{\varepsilon_k,\alpha}(\phi_k)\\ &= \|F(z_\alpha) - y^\delta\|^2_Y + \alpha\rho(z_\alpha, \phi_\alpha)\\ &= \inf \mathcal{G}_\alpha.\end{aligned}$$

Therefore the infimum of $\mathcal{G}_\alpha$ is attained at $(\tilde{z}, \tilde{\phi})$. □

Theorem 3.2 justifies using the functionals $\mathcal{G}_{\varepsilon,\alpha}$ for approximation of the minimizer of $\mathcal{G}_\alpha$. In contrast to the minimizer of $\mathcal{G}_{\varepsilon,\alpha}$, which is a function in $H^1(\Omega)$, the minimizer of $\mathcal{G}_\alpha$ is an admissible pair (z_α, ϕ_α). Recall that the function z_α is not uniquely defined by ϕ_α if it attains critical values in a neighborhood of the zero level set (cf. Remark 2.5).

For numerical purposes it is convenient to derive the optimality conditions of a minimizer of this functional. To this end we consider the functional $\mathcal{G}_{\varepsilon,\alpha}$ with $Y = L^2(\partial\Omega)$.

Since $P_\varepsilon'(\phi)$ is self-adjoint, we can write the formal optimality condition for a minimizer of the functional $\mathcal{G}_{\varepsilon,\alpha}$ as follows:

$$\alpha(\Delta - I)(\phi - \phi_0) = R_{\varepsilon,\alpha,\beta}(\phi), \tag{3.2}$$

where

$$R_{\varepsilon,\alpha,\beta}(\phi) = P_\varepsilon'(\phi)F'(P_\varepsilon(\phi))^*(F(P_\varepsilon(\phi)) - y^\delta) - \beta\alpha P_\varepsilon'(\phi)\nabla \cdot \left(\frac{\nabla P_\varepsilon(\phi)}{|\nabla P_\varepsilon(\phi)|}\right).$$

4. Iterative regularization and the relation to dynamic level set methods. For $n = 1$ set $\mathcal{G}_\alpha^{(1)}(z,\phi) = \mathcal{G}_\alpha(z,\phi)$ (cf. (2.4)). Iterative regularization consists of minimizing the family of functionals

$$\mathcal{G}_\alpha^{(n)}(z,\phi) = \|F(z) - y^\delta\|_Y^2 + \alpha\rho^{(n)}(z,\phi), \tag{4.1}$$

where $\rho^{(n)}$ is the functional ρ (as defined in (2.5)) with ϕ_0 replaced by ϕ_{n-1}. The minimizer of $\mathcal{G}_\alpha^{(n)}(z,\phi)$ is denoted by ϕ_n.

Proceeding as before, we find that ϕ_n can be realized by solving the formal optimality condition

$$\alpha(\Delta - I)(\phi - \phi_{n-1}) = R_{\varepsilon,\alpha,\beta}(\phi). \tag{4.2}$$

Identifying $\alpha = 1/\Delta t$, $t_n = n\Delta t$, and $\phi_n = \phi(t_n)$, $n = 0, 1, \dots,$ we find

$$(\Delta - I)\left(\frac{\phi(t_n) - \phi(t_{n-1})}{\Delta t}\right) = R_{\varepsilon,1/\Delta t,\beta}(\phi(t_n)). \tag{4.3}$$

Considering Δt as a time discretization and using $\beta = b_\Delta \Delta t$, we find that in a formal sense the iterative regularized solution ϕ_n is a solution of an implicit time step for the dynamic system

$$(\Delta - I)\left(\frac{\partial\phi(t)}{\partial t}\right) = R_{\varepsilon,1/\Delta t,b_\Delta\Delta t}(\phi(t)). \tag{4.4}$$

In our numerical experiments we have calculated the solution of the dynamic system (4.4).

Each time step requires solving (4.3). Then $\phi(t_n)$ in (4.3) can be solved with a fixed point iteration: setting $\phi(t_{n-1}) = \phi^{(0)}$, we get $\phi(t_n) = \lim_{k\to\infty}\phi^{(k)}$

$$(\Delta - I)\left(\frac{\phi^{(k+1)} - \phi^{(0)}}{\Delta t}\right) = R_{\varepsilon,1/\Delta t,b_\Delta\Delta t}(\phi^{(k)})\ . \tag{4.5}$$

In our numerical experiments we observed that the iteration starts oscillating after the first iteration (cf. Figure 4.1). This behavior becomes transparent by noting that the H^1-seminorm typically dominates the L^2-norm in the quadratic regularization term. The H^1-seminorm difference of the regularized solution and $\phi^{(0)}$ is small if it is just shifted up or down. In numerical experiments it is observed that the first iteration almost corresponds to a horizontal shift of $\phi^{(0)}$ such that the residual functional is minimized (cf. Figure 4.2), and also the further iterations are again nearly horizontally shifted versions of $\phi^{(0)}$ (cf. Figure 4.3).

In almost all test examples the residual $\|F(P_\varepsilon(\phi^{(k)})) - y^\delta\|^2$ is oscillating in a way dependent on k (cf. Figure 4.2) and smallest for $k = 1$.

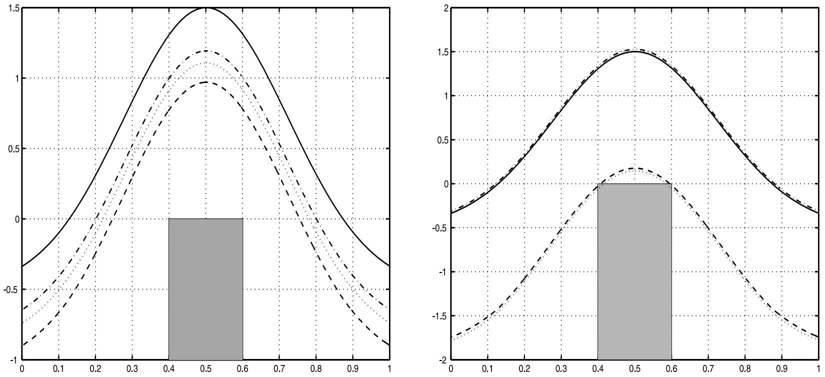


FIG. 4.1. *The functions $\phi^{(0)}$ (solid line), $\phi^{(1)}$ (dashed line), $\phi^{(2)}$ (dash-dot line), and $\phi^{(3)}$ (dotted line). To recover is the interval $[0.4, 0.6]$, which is displayed by the grey rectangle. The first iteration is the best. In the right picture α is smaller than in the left picture.*

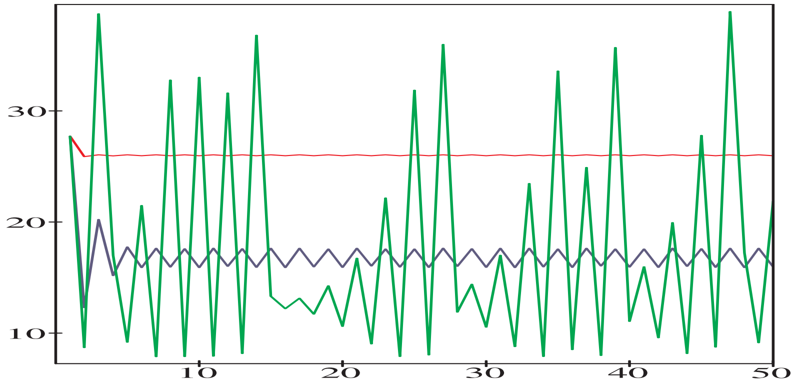


FIG. 4.2. *Decay of the residual $\|F(P_\varepsilon(\phi^{(k)})) - y^\delta\|_Y^2$ dependent on the number of iterations (residual evaluated for the first experiment—noise-free data in section 5). After the first iteration the fixed-point iteration stagnates.*

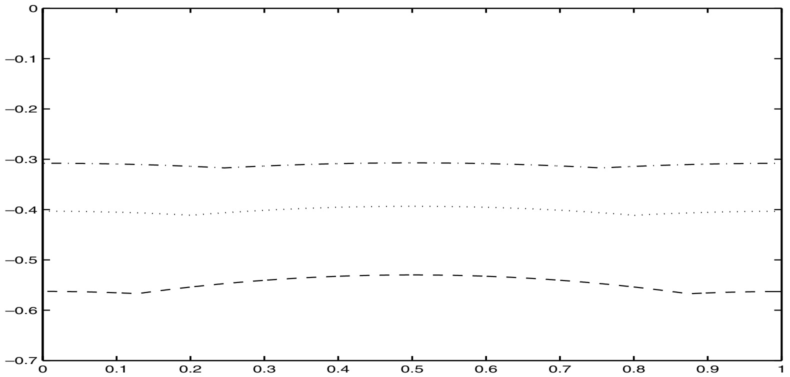


FIG. 4.3. *The differences between $\phi^{(0)}$ and the functions $\phi^{(1)}$ (dashed line), $\phi^{(2)}$ (dash-dot line), and $\phi^{(3)}$ (dotted line) from the left picture of Figure 4.1.*

The above consideration justifies our restricting attention to the approximate solution of the dynamic system (4.2), where in each time step only one iteration step of (4.5) is used; i.e., we use an explicit Euler method for solving the evolution process. In this case numerical instabilities may occur by dividing by small absolute values of the gradient in the differential $\nabla \cdot \left(\frac{\nabla P_\varepsilon(\phi)}{|\nabla P_\varepsilon(\phi)|}\right)$. Thus, for numerical purposes it is convenient to introduce a small positive number h and replace the differential by

$$\nabla \cdot \left(\frac{\nabla P_\varepsilon(\phi)}{\sqrt{|\nabla P_\varepsilon(\phi)|^2 + h^2}} \right) .$$

Usually semiimplicit iteration schemes require a less restrictive time marching. (This approach is commonly referred to as Dziuk's method.) The implementation would require solving

$$\begin{aligned}(\Delta - I)\left(\frac{\phi^{(k+1)} - \phi^{(0)}}{\Delta t}\right) &= P_\varepsilon'(\phi^{(k)})F'(P_\varepsilon(\phi^{(k)}))^*(F(P_\varepsilon(\phi^{(k)})) - y^\delta) \\ &\quad - b_\Delta P_\varepsilon'(\phi^{(k)})\nabla \cdot \left(\frac{\nabla P_\varepsilon(\phi^{(k+1)})}{\sqrt{|\nabla P_\varepsilon(\phi^{(k)})|^2 + h^2}} \right) .\end{aligned} \tag{4.6}$$

In implementation of this approach the difficulty arises that the function in front of $\nabla \cdot \left(\nabla P_\varepsilon(\phi^{(k+1)})/\sqrt{|\nabla P_\varepsilon(\phi^{(k)})|^2 + h^2}\right)$ vanishes outside of a neighborhood of the zero level set, which makes it almost impossible to implement this scheme efficiently.

5. Numerical experiments. In this section we shall consider an inverse potential problem of recovering the shape of a domain D using the knowledge of its (constant) density and the measurements of the Cauchy data of the corresponding potential on the boundary of a fixed Lipschitz domain $\Omega \subset \mathbb{R}^2$, which contains $\overline{D}$. This is the same problem as considered by Hettlich and Rundell [12], which used iterative methods for recovering a single star-shaped object.

To achieve an analogous problem, a certain definition of the operator F is necessary:

$$\begin{aligned} F : L^2(\Omega) &\to L^2(\partial\Omega), \\ \chi_D &\to \ F(\chi_D). \end{aligned}$$

This is possible because we consider only characteristic functions χ_D. The $L^2(\Omega)$-norm is then equivalent to the $L^1(\Omega)$-norm of χ_D. Therefore the necessary properties are retained.

The problem introduced above can mathematically be described as follows:

$$\Delta u = \chi_D \ \text{ in } \Omega, \qquad u|_{\partial\Omega} = 0, \tag{5.1}$$

where χ_D is the characteristic function of the domain $D \subset \Omega$, which has to be reconstructed. Since $\chi_D \in L^2(\Omega)$, the Dirichlet boundary value problem in (5.1) has a unique solution, the potential $u \in H^2(\Omega) \cap H_0^1(\Omega)$. Here $H_0^1(\Omega)$ is defined as the closure with respect to $H^1(\Omega)$ of functions in $C^\infty(\Omega)$ with compact support in Ω.

The inverse problem that we are concerned with consists of determining the shape of D from measurements of the Neumann trace of u at $\partial\Omega$, i.e., from $[\partial u/\partial \nu]_{\partial\Omega}$, where ν represents the outer normal vector to $\partial\Omega$.

Notice that this problem can be considered in the framework of an inverse problem for the *Dirichlet-to-Neumann map.* For given $h \in L^2(\Omega)$, the Dirichlet-to-Neumann operator maps a Dirichlet boundary datum onto the Neumann trace of the potential, i.e., $\Lambda : H^{1/2}(\partial\Omega) \to H^{-1/2}(\partial\Omega)$, $\Lambda(\varphi) := [\partial\tilde{u}/\partial\nu]_{\partial\Omega}$, where $\tilde{u}$ solves

$$\Delta\tilde{u} = h \text{ in } \Omega, \qquad \tilde{u}|_{\partial\Omega} = \varphi\,.$$

The inverse problem for the Λ operator consists of determining the unknown parameter (i.e., the function h) from different pairs of Dirichlet Neumann boundary data. The general case with $h \in L^2(\Omega)$ has already been considered by many authors, including [6, 23], who introduced numerical methods based on Tikhonov regularization, and [12], who used iterative regularization methods.

Hettlich and Rundell [12] observe that, in the particular case $h = \chi_D$, one pair of Dirichlet–Neumann measurement data furnishes as much information as the full Dirichlet–Neumann operator; i.e., it is sufficient to consider only one pair of Cauchy data for the inverse problem. Therefore, no further information on D can be gained by using various pairs of Dirichlet–Neumann data, since we can always reduce the reconstruction problem to the homogeneous Dirichlet case.

For the particular case $h = \chi_D$, it has been observed by Hettlich and Rundell [12] that the Cauchy data may not furnish enough information to reconstruct the boundary of D, e.g., if D is not simply connected. On the other hand, Isakov observed in [13] that star-like domains D are uniquely determined by their potentials.

The inverse potential problem is discussed within the general framework introduced in section 1. In particular, we allow domains that consist of a number of connected inclusions. For this general class we do not have unique identifiability, and we restrict our attention to "minimum-norm solutions." Recall that in this case a minimum-norm solution is a level set function ϕ, where $P(\phi)$ determines the inclusion. A minimum-norm solution satisfies the requirement that it minimize the functional $\rho(z, \phi)$ in the class of level set functions such that the corresponding Neumann boundary values $\frac{\partial u}{\partial \nu}$ fit the data y^δ.

5.1. The level set regularization algorithm. In the following we describe the level set regularization algorithm. This method is comparable with the Landweber iteration as proposed by Hettlich and Rundell [12]. In our context the operator F' can be considered as an approximation of the *domain derivative operator* for multiple connected domains (cf. Figure 5.1).

The complexity of our algorithm is as follows: at each iteration of the level set method, three elliptic boundary value problems are solved (two of Dirichlet type and one of Neumann type).

In Figure 5.1 the iteration procedure for the solution of the formal optimality condition (3.2) is outlined. The algorithm can be implemented using finite element codes (as we did) or finite difference methods for the solution of partial differential equations.

5.2. Reconstruction of a density function with non–simply connected support. In this first experiment we consider the inverse problem of reconstructing the right-hand-side χ_D in (5.1) from the knowledge of a single pair of boundary data $(u, \Lambda u) = (0, y^\delta)$ at $\partial\Omega$. In the examples considered below we always use the squared domain $\Omega = (0,1)^2 \subset \mathbb{R}^2$. Additionally, $\chi_D \in L^2(\Omega)$ is the characteristic function as represented in Figure 5.2.

The overdetermined boundary measurement data y^δ for solving the inverse problem is obtained by solving the elliptic boundary value problem in (5.1). Notice that χ_D corresponds to the characteristic function of a not-connected proper subset of Ω.

1. Evaluate the residual $r_k := F(P_\varepsilon(\phi_k)) - y^\delta = \frac{\partial u_k}{\partial \nu} - y^\delta$, where u_k solves

$$\Delta u_k = P_\varepsilon(\phi_k) \ \texttt{in} \ \Omega, \qquad u_k|_{\partial\Omega} = 0.$$

2. Evaluate $v_k := F'(P_\varepsilon(\phi_k))^*(r_k) \in L^2(\Omega)$, solving

$$\Delta v_k = 0 \ \texttt{in} \ \Omega, \quad v_k|_{\partial\Omega} = r_k\,.$$

3. Evaluate $w_k \in H^1(\Omega)$, satisfying

$$(I - \Delta) w_k = -P'_\varepsilon(\phi_k)\, v_k + \beta\alpha P'_\varepsilon(\phi_k) \nabla \cdot \left(\frac{\nabla P_\varepsilon(\phi_k)}{|\nabla P_\varepsilon(\phi_k)|} \right) \ \texttt{in} \ \Omega,$$
$$\frac{\partial w_k}{\partial \nu}|_{\partial\Omega} = 0.$$

4. Update the level set function $\phi_{k+1} = \phi_k + \frac{1}{\alpha}\, w_k\,.$

FIG. 5.1. *Implementation of a single iteration step for minimizing the level set regularization.*

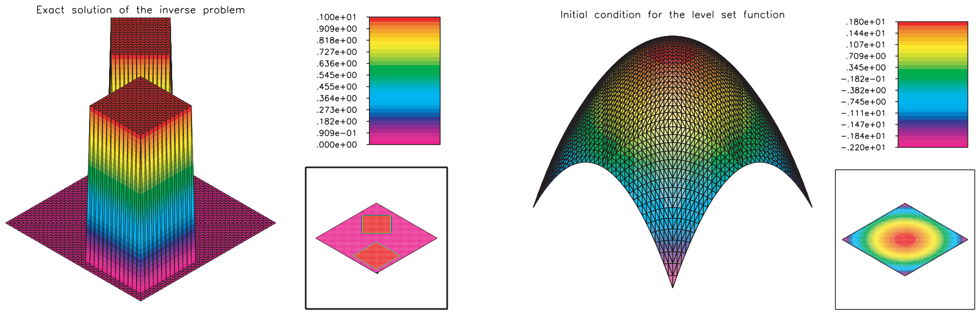


FIG. 5.2. *The picture on the left-hand side shows the coefficient to be reconstructed. In the other picture, the initial condition for the level set regularization method is given.*

The initial condition for the level set function is shown in Figure 5.2. In order to avoid inverse crimes, the direct problem (5.1) is solved on an adaptively refined grid with 8.807 nodes (three levels of adaptive refinement). Alternatively, in the numerical implementation of the level set method, all boundary value problems are solved on a uniformly refined grid with 2.113 nodes.

When the data is given exactly, we tested the iterative level set regularization without the additional regularization term $|P_\varepsilon(\phi_k)|_{\mathtt{BV}}$, i.e., $\beta = 0$.

In all computed experiments we use the operator P_ε defined in section 2 with $\varepsilon = 1/8$. This seems to be compatible with the size of our mesh, since the diameter of the triangles in the uniform grid (used in the finite element method) is approximately $\sqrt{2}/32$.

In Figure 5.3 we present the evolution of the level set function for given exact data for the first 3000 iterative steps. As one can see in this figure, the original level set splits into two convex components after approximately 800 iterations. After 1000 iterations, the level set function still changes but very slowly. We performed similar tests for different initial conditions and observed that, after 1000 iterations, the corresponding pictures look very much alike.

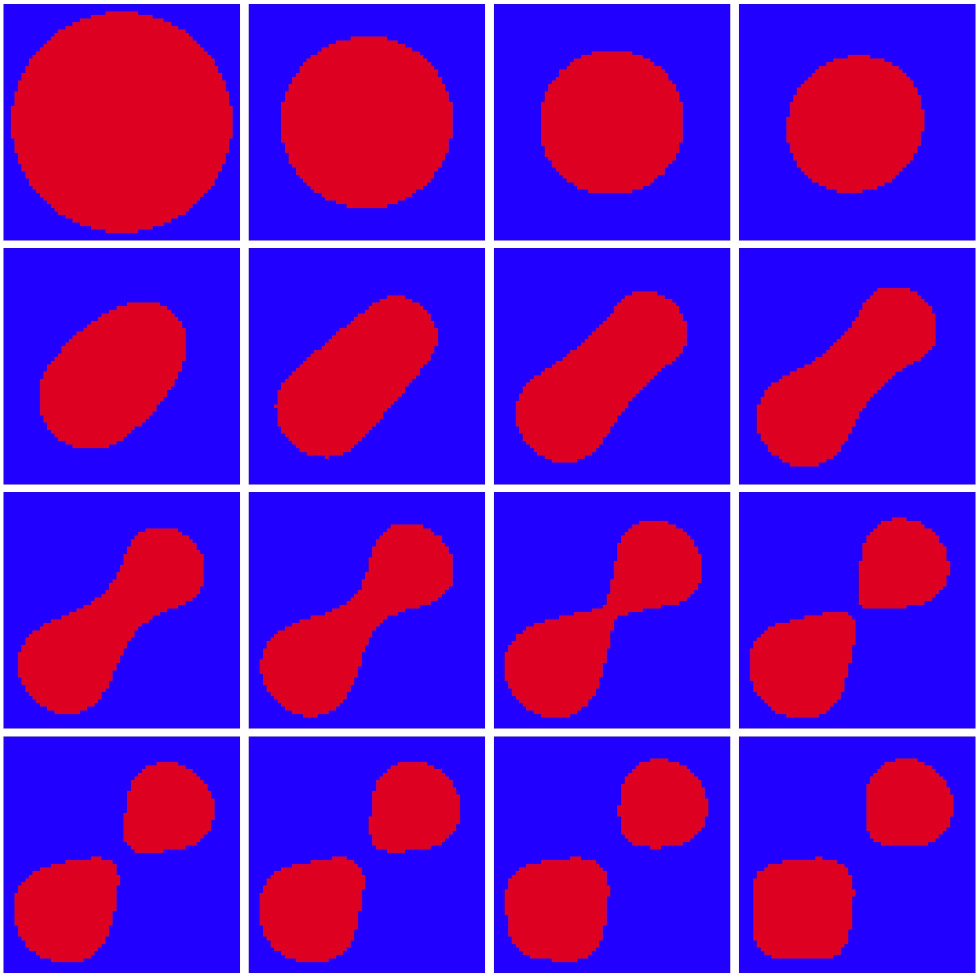


FIG. 5.3. *Level set evolution for exact data. Plots after (grouped by row)* $0, 1, 2, 10$; $100, 200, 300, 400$; $500, 600, 700, 800$; $900, 1000, 2000, 3000$ *iterative steps.*

For the second part of this experiment, the density function to be reconstructed is still the one shown in Figure 5.2. This time, however, we add randomly generated noise to the data y^δ used in the first part of the experiment.

The exact boundary data y^δ is shown in Figure 5.4 as the square-dotted (blue) line. We consider actually two distinct sets of perturbed data. For the first experiment we add to the exact data a white noise of 10% (in the l_∞-norm). For the second experiment we use a noise level of 50%. Both sets of inaccurate data are plotted in Figure 5.4 and correspond to the solid (red) line.

As in the noise-free experiment, care was taken to avoid inverse crimes. The choice of the parameter ε (operator P_ε) also follows the same criteria as before. However, since we are now dealing with noisy data, we have to develop a strategy for choosing the regularization parameter β. For this proposal we opted for the fit-to-data strategy; i.e., $\beta\alpha$ is chosen such that the regularization term (see Figure 5.1) has the same order as the noise level.

The corresponding results generated by the level set method were surprisingly stable, as one can observe in Figures 5.5 and 5.6. In the first case (noise level of 10%)

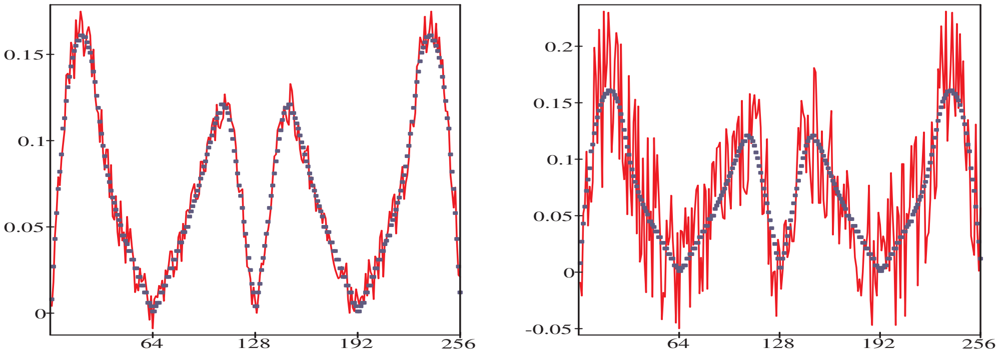


FIG. 5.4. *The square-dotted (blue) line represents the exact data* y^δ*; the solid (red) line represents the perturbed data. The noise level corresponds to* 10% *at the left-hand side and* 50% *at the right-hand side.*

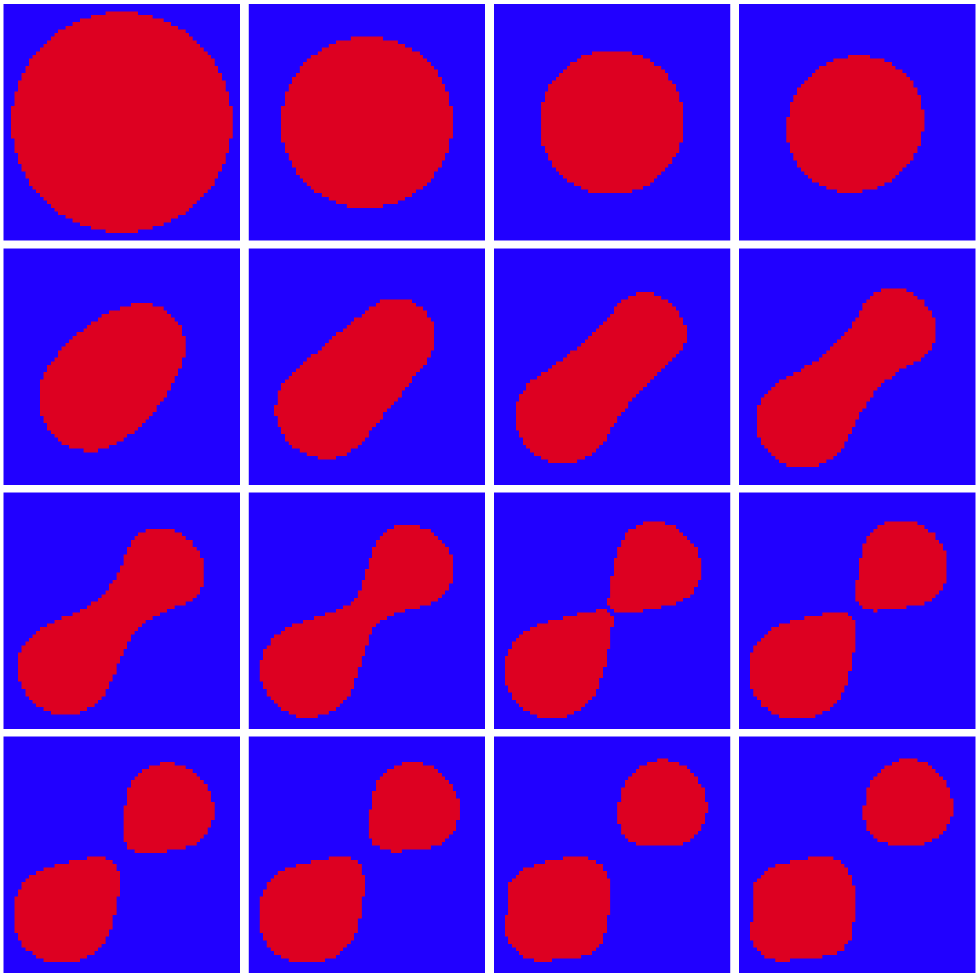


FIG. 5.5. *Level set evolution for inaccurate data; noise level of* 10%. *Plots after* $0, 1, 2, 10$; $100, 200, 300, 400$; $500, 600, 700, 800$; $900, 1000, 2000, 3000$ *iterative steps.*

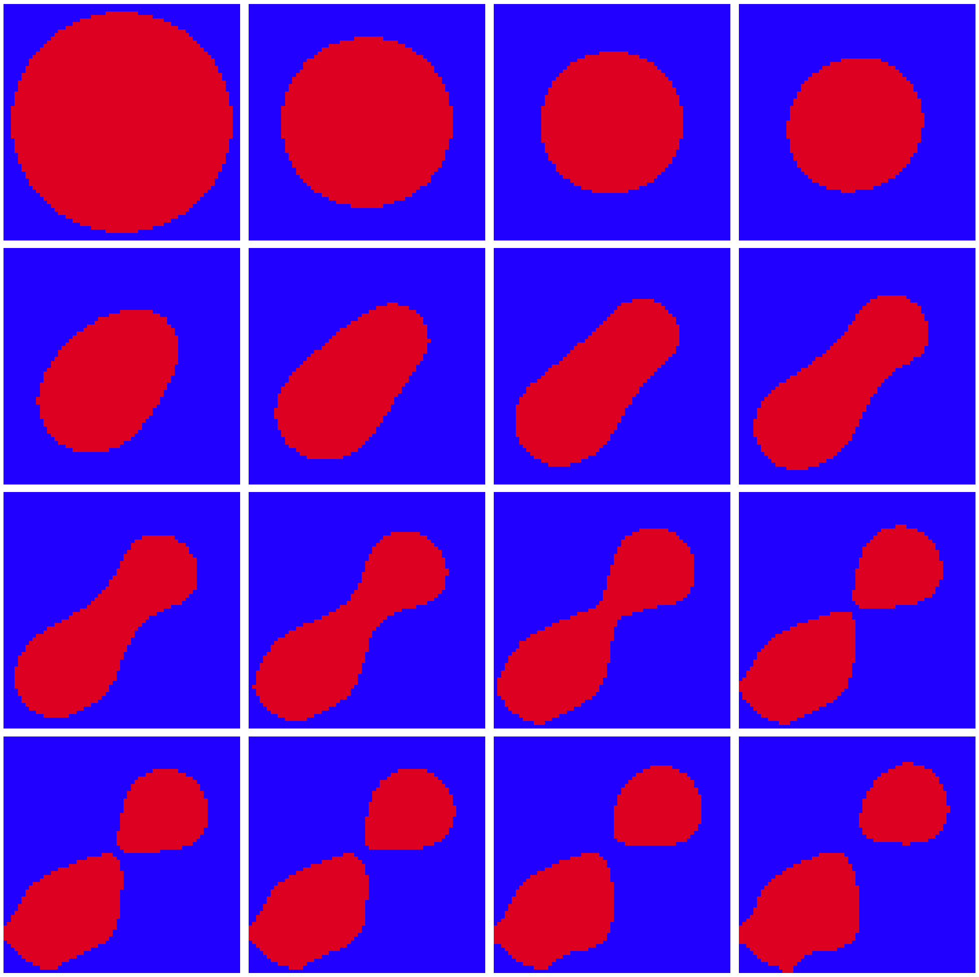


FIG. 5.6. *Level set evolution for inaccurate data; noise level of* 50%. *Plots after* $0, 1, 2, 10$; $100, 200, 300, 400$; $500, 600, 700, 800$; $900, 1000, 1300, 1600$ *iterative steps.*

the results are comparable with the previous experiment, where exact data was available. In the second case (noise level of 50%) we are not able to precisely recover the shape of the set D, corresponding to the characteristic function shown in Figure 5.2. However, we are still able to identify the number of connected components of D, as well as their relative positions inside the domain Ω.

5.3. Reconstruction of a density function with nonconvex support. In this second experiment we consider the problem of reconstructing the density function shown in Figure 5.7. The main goal now is to investigate the difficulty of the level set method in recovering nonconvex domains. The domain Ω is the same used in subsection 5.2, and again we aim to reconstruct the density function in (5.1) from boundary measurements.

As in the first part of the previous experiment, the data is given almost exactly, and the velocity w_k is again obtained by solving the boundary value problem with $\beta = 0$. The evolution of the level set function is shown in Figure 5.8.

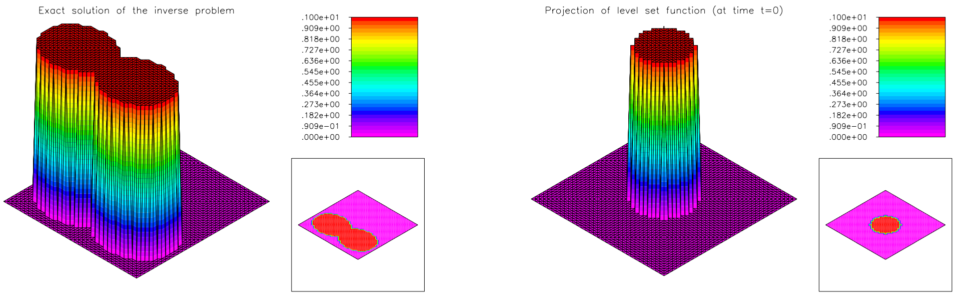


FIG. 5.7. *(Left) The coefficient to be reconstructed. (Right) The (projection of the) initial condition for the level set regularization method.*

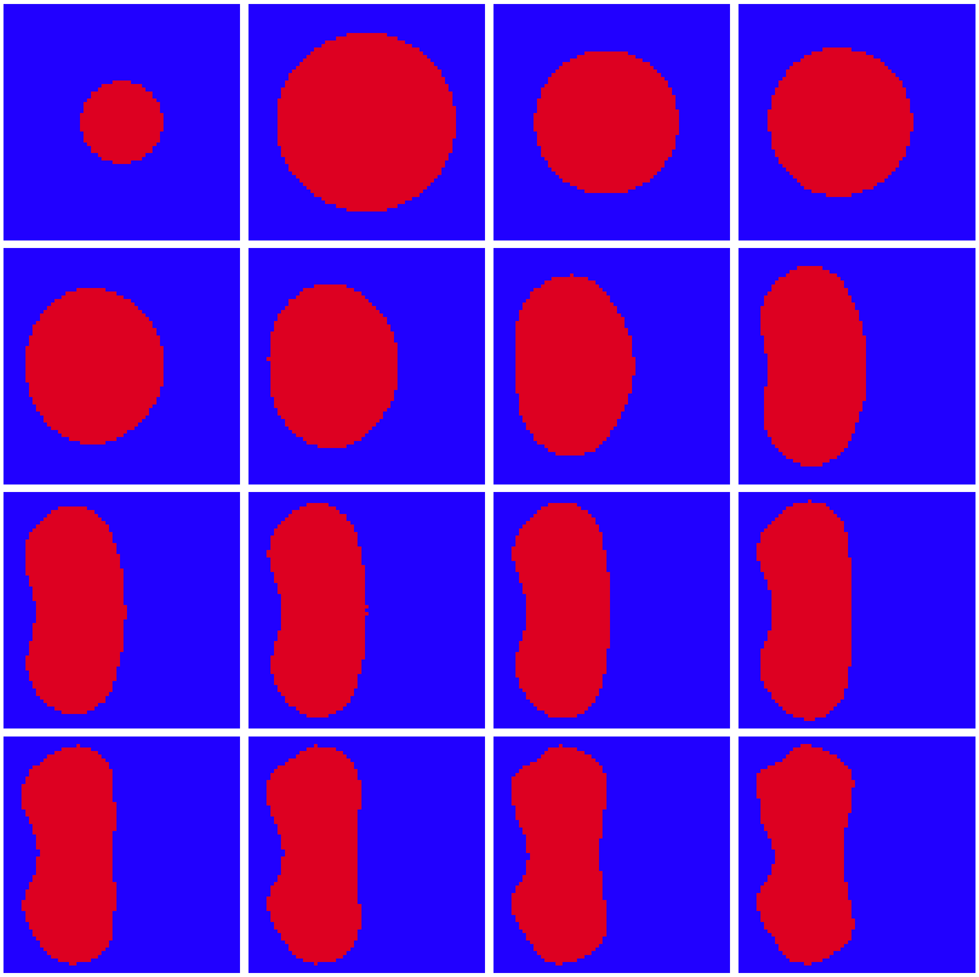


FIG. 5.8. *Level set evolution for the second experiment. Plots after* $0, 1, 5, 20;$ $50, 100, 200, 400;$ $600, 800, 1000, 2000;$ $5000, 10000, 20000, 50000$ *iterative steps.*

Remark 5.1. *The effect of parameter changes.* In our numerical observations we observed that in numerical simulations the minimizer is not severely affected by the choice of $\beta\alpha$ and can, in fact, be neglected.

REFERENCES

[1] R. Acar and C. R. Vogel, *Analysis of bounded variation penalty methods for ill-posed problems*, Inverse Problems, 10 (1994), pp. 1217–1229.

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

[3] L. Ambrosio, *Geometric evolution problems, distance function and viscosity solutions*, in Calculus of Variations and Partial Differential Equations, L. Ambrosio and N. Dancer, eds., Springer, New York, 1999, pp. 5–94.

[4] H. Brezis, *Operateurs Maximaux Monotones et semi-groupes de contractions dans les espaces de Hilbert*, North–Holland, Amsterdam, 1973.

[5] M. Burger, *A level set method for inverse problems*, Inverse Problems, 17 (2001), pp. 1327–1355.

[6] H. S. Cabayan and G. G. Belford, *On computing a stable least squares solution to the inverse problem for a planar Newtonian potential*, SIAM J. Appl. Math., 20 (1971), pp. 51–61.

[7] T. Chan, J. Shen, and L. Vese, *Variational PDE models in image processing*, Notices Amer. Math. Soc., 50 (2003), pp. 14–26.

[8] S. Chen, B. Merriman, S. Osher, and P. Smereka, *A simple level set method for solving Stefan problems*, J. Comput. Phys., 135 (1997), pp. 8–29.

[9] M. Delfour and J.-P. Zolesio, *Shape analysis via oriented distance functions*, J. Funct. Anal., 123 (1994), pp. 129–201.

[10] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[11] H. W. Engl, K. Kunisch, and A. Neubauer, *Convergence rates for Tikhonov regularization of nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 523–540.

[12] F. Hettlich and W. Rundell, *Iterative methods for the reconstruction of an inverse potential problem*, Inverse Problems, 12 (1996), pp. 251–266.

[13] V. Isakov, *Inverse Source Problems*, AMS, Providence, RI, 1990.

[14] R. Kohn and G. Strang, *Optimal design and relaxation of variational problems.* I, Comm. Pure Appl. Math., 39 (1986), pp. 113–137.

[15] R. Kohn and G. Strang, *Optimal design and relaxation of variational problems.* II, Comm. Pure Appl. Math., 39 (1986), pp. 139–182.

[16] R. Kohn and G. Strang, *Optimal design and relaxation of variational problems.* III, Comm. Pure Appl. Math., 39 (1986), pp. 353–377.

[17] A. Leitão and O. Scherzer, *On the relation between constraint regularization, level sets, and shape optimization*, Inverse Problems, 19 (2003), pp. L1–L11.

[18] B. Merriman, J. Bence, and S. Osher, *Motion of multiple functions: A level set approach*, J. Comput. Phys., 112 (1994), pp. 334–363.

[19] V. A. Morozov, *Methods for Solving Incorrectly Posed Problems*, Springer-Verlag, New York, Berlin, Heidelberg, 1984.

[20] V. A. Morozov, *Regularization Methods for Ill-Posed Problems*, CRC Press, Boca Raton, FL, 1993.

[21] M. Z. Nashed and O. Scherzer, *Least squares and bounded variation regularization with nondifferentiable functional*, Numer. Funct. Anal. Optim., 19 (1998), pp. 873–901.

[22] S. Osher and J. A. Sethian, *Fronts propagating with curvature-dependent speed: Algorithms based on Hamilton–Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[23] W. Ring, *Identification of a core from boundary data*, SIAM J. Appl. Math., 55 (1995), pp. 677–706.

[24] F. Santosa, *A level-set approach for inverse problems involving obstacles*, ESAIM Control Optim. Calc. Var., 1 (1995/96), pp. 17–33.

[25] T. I. Seidman and C. R. Vogel, *Well-posedness and convergence of some regularization methods for nonlinear ill-posed problems*, Inverse Problems, 5 (1989), pp. 227–238.

[26] H.-K. Zhao, T. Chan, B. Merriman, and S. Osher, *A variational level set approach to multiphase motion*, J. Comput. Phys., 127 (1996), pp. 179–195.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 787–795

GAUSSIAN INTERVAL QUADRATURE FORMULAE FOR TCHEBYCHEFF SYSTEMS[*]

BORISLAV BOJANOV[†] AND PETAR PETROV[†]

Abstract. For any given system of continuously differentiable functions $\{u_k\}_{k=1}^{2n}$ that constitute an extended Tchebycheff system of order 2 on $[a, b]$ we prove the existence and uniqueness of the Gaussian interval quadrature formula based on n weighted integrals over nonoverlapping subintervals of $[a, b]$ of preassigned lengths. This supplies an analogue of the result of Krein about canonical representation of linear positive functionals.

Key words. Gaussian quadrature, Tchebycheff systems, interval analysis, polynomials

AMS subject classifications. 65D32, 65D30, 41A55

DOI. 10.1137/040606521

1. Introduction.

The classical Gauss quadrature formula was extended over the years in various directions. The present paper concerns an interesting development of the subject, initiated by Krein in [5]. He proved that for any given system $U_{2n} := \{u_1, u_2, \dots, u_{2n}\}$ of continuous functions that form a Tchebycheff system on $[a, b]$, there exists a unique set of points $x_1 < \cdots < x_n$ in $[a, b]$ such that the interpolatory quadrature formula

$$\int_a^b f(t)\,dt \approx \sum_{k=1}^{n} C_k f(x_k)$$

is exact for every generalized polynomial

$$u(t) = c_1 u_1(t) + \cdots + c_{2n} u_{2n}(t)$$

with real coefficients $\{c_j\}$. Actually, his result is slightly stronger and covers the canonical representation of a general linear positive functional $L[f]$ on $C[a, b]$. Simplified proofs of Krein's result can be found in [4] and [1].

A significant effort was made in the last few decades in extending the Gauss formula to other natural types of data, in addition to the standard one of sampling function values. Recently, we proved in [2], [3] the existence and uniqueness of a formula of the form

$$\int_a^b f(t)\,dt \approx \sum_{k=1}^{n} A_k \frac{1}{h_k} \int_{x_k}^{x_k+h_k} f(t)\,dt$$

of highest degree of precision with respect to the class of algebraic polynomials, for any fixed system of lengths $\{h_k\}$, $h_1 + \cdots + h_n \le b - a$. The problem stayed open for quite a long time (see [7], [9], [10], [11], [12] for previous results). In [8], Milovanović and Cvetković showed that the proof from [3] can be modified to cover the case of

[*] Received by the editors April 8, 2004; accepted for publication (in revised form) November 23, 2004; published electronically August 31, 2005. This research was supported by the Bulgarian Ministry of Science under contract MM-1402/2004.

http://www.siam.org/journals/sinum/43-2/60652.html

[†] Department of Mathematics, University of Sofia, Blvd. James Boucher 5, 1164 Sofia, Bulgaria (boris@fmi.uni-sofia.bg, peynov@fmi.uni-sofia.bg).

Jacobi weight function. The purpose of this paper is to go further and show that the result also holds for Tchebycheff systems and thus supplies an interval analogue of Krein's result. In case all lengths $\{h_k\}$ are supposed equal to zero, the quadrature reduces to Krein's canonical representation.

Throughout this paper we assume that n is a natural number, $[a,b]$ is a fixed finite interval, and $U_{2n} := \{u_1, u_2, \dots, u_{2n}\}$ is a given system of continuously differentiable functions which constitute an extended Tchebycheff (ET) system of order 2 on $[a,b]$. Recall that the system U_{2n} constitutes an ET system of order 2 on $[a,b]$ if any nonzero generalized polynomial u with respect to U_{2n} possesses no more than $2n-1$ zeros in $[a,b]$ (counting twice every common zero of u and u') (see [4, Chapter II]). We shall denote by $\mathcal{U}_{2n}$ the linear space of all generalized polynomials with respect to the system U_{2n}, that is,

$$\mathcal{U}_{2n} := \text{span}\,\{u_1, u_2, \dots, u_{2n}\}.$$

Further, assume that $\mu(t)$ is a given integrable function on $[a,b]$, which is continuous and strictly positive on (a,b), and denote by $L[f]$ the integral

$$L[f] := \int_a^b \mu(t) f(t)\,dt.$$

We are going to give a canonical representation of $L[f]$ in $\mathcal{U}_{2n}$ of the form

$$L[f] = \sum_{k=1}^{n} a_k \frac{1}{I_k} I_k[f]\,,$$

where

$$I_k[g] := \int_{x_k}^{x_k+h_k} \mu(t) g(t)\,dt, \qquad I_k := I_k[1].$$

Here we assume that

$$\Delta_k := [x_k, x_k + h_k]$$

are n nonoverlapping subintervals of $[a,b]$ of preassigned lengths $|\Delta_k| := h_k \geq 0$.

In case $h_k = 0$ the quantity $I_k[f]/I_k$ is defined by continuity, that is,

$$\frac{1}{I_k} I_k[f]\Big|_{h_k=0} := \lim_{h_k \to 0} \frac{1}{I_k} I_k[f] = f(x_k).$$

Keeping the notation from [3], we introduce the set H of admissible lengths $\mathbf{h} = (h_1, \dots, h_n)$,

$$H := \left\{ \mathbf{h} \in \mathbb{R}^n \,:\, h_k \geq 0,\, k = 1, \dots, n,\, \sum_{k=1}^{n} h_k < b - a \right\},$$

the associated set

$$D = D(\mathbf{h}) := \{\mathbf{x} \in \mathbb{R}^n \,:\, a < x_1 \leq x_1 + h_1 < \cdots < x_n \leq x_n + h_n < b\}$$

of admissible nodes, and its closure

$$\bar{D} = \bar{D}(\mathbf{h}) := \{\mathbf{x} \in \mathbb{R}^n \,:\, a \leq x_1 \leq x_1 + h_1 \leq \cdots \leq x_n \leq x_n + h_n \leq b\}.$$

Let us denote by ∂D the boundary of D. We prove the following.

THEOREM 1. *Let $U_{2n} = \{u_1, \dots, u_{2n}\}$ be any ET system of order 2 of continuously differentiable functions on $[a,b]$ and let μ be an integrable function on $[a,b]$ which is continuous and positive on (a,b). Then, for every given set of numbers $\mathbf{h} \in H$ there exists a unique set of nodes $\mathbf{x} \in D(\mathbf{h})$ such that*

$$L[f] = \sum_{k=1}^{n} a_k \frac{1}{I_k} \int_{x_k}^{x_k+h_k} \mu(t) f(t)\, dt \tag{1}$$

for every f from the space $\mathcal{U}_{2n}$.

We shall call (1) a Gaussian formula.

Note that Theorem 1 also holds in the trivial case $h_1 + \cdots + h_n = b - a$ since the uniqueness of the best coefficients $a_k = I_k$, $k = 1, \dots, n$, can be easily verified.

The proof of Theorem 1 is given in section 3 and is based on some auxiliary results contained in section 2.

2. Auxiliary results. The lemmas in this section hold under weakened conditions on the space $\mathcal{U}_{2n}$ and the weight μ. It is enough to assume that the Tchebycheff system U_{2n} consists of continuous functions on $[a,b]$ and $\mu(t)$ is any given continuous weight function on (a,b) such that $\int_\alpha^\beta \mu(t)\, dt > 0$ for every $a \le \alpha < \beta \le b$.

We begin with two lemmas concerning interpolation properties of the Tchebycheff space $\mathcal{U}_{2n}$.

LEMMA 1. *Let*

$$a \le t_1 \le t_1 + h_1 \le \cdots \le t_{2n} \le t_{2n} + h_{2n} \le b$$

be given points such that $t_i < t_{i+1}$ in case $h_i = h_{i+1} = 0$. Then the interpolation problem

$$\left(\int_{t_k}^{t_k+h_k} \mu(t)\, dt \right)^{-1} \cdot \int_{t_k}^{t_k+h_k} \mu(t) u(t)\, dt = f_k\,, \quad k = 1, \dots, 2n, \tag{2}$$

is uniquely solvable in $\mathcal{U}_{2n}$ for any values $f_k \in \mathbb{R}$.

Proof. It is sufficient to show that the corresponding homogeneous interpolation problem admits only the zero solution $u \equiv 0$. In order to do this, suppose that $u \in \mathcal{U}_{2n}$ satisfies conditions (2) with $f_k = 0$. Note that in case $h_k = 0$ the corresponding homogeneous condition reduces to $u(t_k) = 0$, and if $h_k > 0$, it implies that $u(t)$ must change sign on $(t_k, t_k + h_k)$. Then u must have at least one zero in every interval $[t_k, t_k + h_k]$, i.e., at least $2n$ zeros in total. But this means that $u \equiv 0$, since $\mathcal{U}_{2n}$ is a Tchebycheff space, and the proof is complete. □

LEMMA 2. *Let $\mathbf{h} \in H$ be fixed. Then for every $\mathbf{x} \in \partial D$ there exists a generalized polynomial $u \in \mathcal{U}_{2n}$ which is positive on $[a,b] \setminus \cup_{k=1}^n \Delta_k$ and*

$$\frac{I_k[u]}{I_k} = 0\,, \quad \textit{for } k = 1, \dots, n. \tag{3}$$

Proof. We consider only the case $x_j + h_j = x_{j+1}$ for some $j \in \{1, \dots, n-1\}$, $h_j > 0$, $h_{j+1} > 0$, $x_k + h_k < x_{k+1}$ for all $k \ne j$. The other cases follow similarly.

Any generalized polynomial w from $\mathcal{U}_{2n}$,

$$w = c_1 u_1 + \cdots + c_{2n} u_{2n}\,,$$

is completely determined by its coefficients $\{c_i\}$. We shall look for a generalized polynomial w satisfying the conditions (3) and, in addition,

$$w(x_k + h_k) = 0 \quad \text{for every} \quad k \neq j, j+1.$$

If some $h_k = 0$, we interpret the conditions $I_k[w]/I_k = 0$, $w(x_k + h_k) = 0$ as "x_k is a double zero of w" (see [4, Chapter I, Theorems 5.1 and 5.2] for handling "double zeros" of Tchebycheff systems). Thus, we have imposed $2n-2$ zero conditions on w. Now, if one of the end points, say a, is not among the prescribed nodes $x_1, \ldots, x_n$, we define one more condition, $w(a) = 0$. If $a = x_1$ and $b = x_n + h_n$, then we delete the condition $w(x_1 + h_1) = 0$, choose two distinct points ξ_1 and ξ_2 outside any of the subintervals Δ_k, $k = 1, \ldots, n$, and impose the conditions "w has a double zero at ξ_1," $w(\xi_2) = 1$. In this way, we have a system of $2n$ linear equations with respect to the coefficients of w. It is easy to see that the corresponding homogeneous system admits only the zero solution since any of the $2n$ homogeneous equations leads to a zero of w. Thus the system of conditions imposed on u determine it uniquely. All zeros of w lie in the intervals $\{\Delta_i\}_{i=1}^n$ and also at ξ_1. Constructing similarly another polynomial $v \in \mathcal{U}_{2n}$ with different point ξ_1, we conclude that $u = w + v$ is positive outside $\{\Delta_i\}_{i=1}^n$ and satisfies (3). The lemma is proved. □

The next observation is a simple consequence from known properties of the Tchebycheff systems.

Lemma 3. *For all* $\mathbf{h} \in H$*, the coefficients* a_k *of the Gaussian formula* (1) *are uniformly bounded.*

Proof. It is a well-known fact (see, for example, [6, Chapter II, Theorem 1.4]) that every Tchebycheff space contains a strictly positive function. Let $\tilde{u} \in \mathcal{U}_{2n}$ and $\tilde{u}(t) > 0$ on $[a, b]$. Since (1) holds for $\tilde{u}$ and the coefficients a_k are strictly positive (see [2, Theorem 1]), then

$$a_k < \frac{\int_a^b \mu(x)\tilde{u}(x)\,dx}{\min_{\,x\in[a,b]}\tilde{u}(x)}, \quad k = 1, \ldots, n. \qquad \square$$

Further, we prove a lemma similar to Lemma 2 from [3] , which asserts that the node subintervals of any Gaussian formula are uniformly distant from each other and from the end points of the interval $[a, b]$.

For each $0 < \varepsilon < b - a$ we define

$$H_\varepsilon := \left\{ \mathbf{h} \in H \, : \, \sum_{k=1}^n h_k \leq b - a - \varepsilon \right\}. \qquad \square$$

Lemma 4. *Let* $0 < \varepsilon < b - a$*. Then there exists an* $\varepsilon_0 \in (0, \varepsilon)$ *such that all* $\mathbf{h} \in H_\varepsilon$ *and the corresponding nodes* $\mathbf{x} \in D(\mathbf{h})$ *which define a Gaussian formula* (1) *satisfy the conditions*

$$(4) \quad x_1 - a > \varepsilon_0, \quad x_{j+1} - x_j - h_j > \varepsilon_0, \; j = 1, \ldots, n-1, \quad b - x_n - h_n > \varepsilon_0.$$

Proof. Assume the contrary. Then for each $0 < \delta < \varepsilon$ there exist numbers $\mathbf{h}_\delta \in H_\varepsilon$ and nodes $\mathbf{x}_\delta \in D(\mathbf{h}_\delta)$ which define a Gaussian quadrature formula (1) with coefficients $a_{k,\delta}$, but at least one of the conditions in (4) (with $\varepsilon_0 = \delta$) is violated.

Since the sequences $\{\mathbf{h}_\delta\}$ and $\{\mathbf{x}_\delta\}$ are bounded, and according to Lemma 3 the corresponding coefficients $a_{k,\delta}$ are bounded, there exist subsequences $\{\mathbf{h}^{(i)}\}_{i=1}^\infty$,

$\{\mathbf{x}^{(i)}\}_{i=1}^{\infty}$, and $\{a_k^{(i)}\}_{i=1}^{\infty}$, with $\mathbf{x}^{(i)} \in D(\mathbf{h}^{(i)})$, which converge to certain $\mathbf{h}^{(0)} \in H_\varepsilon$, $\mathbf{x}^{(0)} \in \partial D(\mathbf{h}^{(0)})$, and $a_k^{(0)}$, respectively. They define a Gaussian formula

$$L[u] = \sum_{k=1}^{n} a_k^{(0)} \frac{1}{I_k^{(0)}} I_k^{(0)}[u] \tag{5}$$

(with $I_k^{(0)}$ and $I_k^{(0)}[u]$, defined by $x_k^{(0)}$ and $h_k^{(0)}$). In addition, since $\mathbf{x}^{(0)} \in \partial D(\mathbf{h}^{(0)})$, at least one of the following equalities holds:

$$x_1^{(0)} = a, \quad x_{j+1}^{(0)} = x_j^{(0)} + h_j^{(0)},\ j = 1, \dots, n-1, \quad x_n^{(0)} + h_n^{(0)} = b.$$

Using Lemma 2, in any of the above cases we can construct a generalized polynomial $u \in \mathcal{U}_{2n}$ for which the right-hand side of (5) is zero, but the integral in the left-hand side is positive. This leads to a contradiction, and the proof is complete. □

3. Proof of Theorem 1. Let $\mathbf{h} \in H_\varepsilon$ be a fixed vector. First we shall find a system of equations for the nodes $\mathbf{x} \in D(\mathbf{h})$ ensuring the Gaussian property of formula (1). Lemma 4 implies that the nodes of any Gaussian formula (1) belong to the set $D_{\varepsilon_0}(\mathbf{h})$ defined by (4). Let us choose arbitrary points

$$a < s_1 < \cdots < s_n < a + \varepsilon_0/2 \quad \text{and} \quad b - \varepsilon_0/2 < t_1 < \cdots < t_n < b.$$

According to Lemma 1, for each $i \in \{1, \dots, n\}$, there exist a unique pair of generalized polynomials p_i, q_i from $\mathcal{U}_{2n}$ such that

$$p_i(s_k) = 0\,, \quad \frac{1}{I_k} I_k[p_i] = \delta_{ik}\,, \quad i, k = 1, \dots, n\,, \tag{6}$$

and

$$q_i(t_k) = 0\,, \quad \frac{1}{I_k} I_k[q_i] = \delta_{ik}\,, \quad i, k = 1, \dots, n. \tag{7}$$

It is easily seen that $\{p_i, q_i\}_{i=1}^{n}$ are linearly independent and thus form a basis in $\mathcal{U}_{2n}$. Thus, the interval formula (1) is Gaussian if and only if

$$a_i = L[p_i] = L[q_i]\,, \quad i = 1, \dots, n\,,$$

and the latter is equivalent to the system

$$\Psi_i(\mathbf{x}) := \Psi_i(\mathbf{x}, \mathbf{h}) := L[p_i - q_i] = 0\,, \quad i = 1, \dots, n. \tag{8}$$

Let

$$p_i(t) = \sum_{m=1}^{2n} \alpha_{im} u_m(t), \quad q_i(t) = \sum_{m=1}^{2n} \beta_{im} u_m(t).$$

We have to show that the system (8) possesses a unique solution in $D_{\varepsilon_0}(\mathbf{h})$. To this aim, we prove first that the Jacobian of (8) is distinct from zero and has a constant sign at any solution of (8). In order to compute the elements of the Jacobian matrix, we note that, in view of Lemma 1, the matrix D (which does not depend on i) of the linear system (6) with respect to the coefficients $\{\alpha_{im}\}$ possesses a nonzero determinant.

Moreover, by the implicit function theorem, the coefficients are differentiable functions of x_j and

$$\frac{\partial \alpha_{im}}{\partial x_j} = -C_{ij}\,\frac{\det D_{jm}}{\det D},$$

where D_{jm} is the matrix obtained from D by replacing the $(n+j)$th element of its mth column by 1 and the other elements of the column by zero. The constant C_{ij} is given by

$$C_{ij} = \frac{\partial}{\partial x_j}\left\{\frac{I_j[p_i]}{I_j}\right\}.$$

But, by Cramer's rule,

$$\alpha_{im} = \frac{\det D_{im}}{\det D}.$$

Therefore

$$\frac{\partial \alpha_{im}}{\partial x_j} = -\frac{\partial}{\partial x_j}\left\{\frac{I_j[p_i]}{I_j}\right\}\alpha_{jm}.$$

This implies the relation

$$\frac{\partial p_i(t)}{\partial x_j} = -\frac{\partial}{\partial x_j}\left\{\frac{I_j[p_i]}{I_j}\right\}p_j(t).$$

Similarly we obtain

$$\frac{\partial q_i(t)}{\partial x_j} = -\frac{\partial}{\partial x_j}\left\{\frac{I_j[q_i]}{I_j}\right\}q_j(t).$$

Using the last relations and the equality

$$L[p_i] = L[q_i] = a_i,$$

we easily compute the elements of the Jacobian matrix

$$J := J(\mathbf{x},\mathbf{h}) := \left\{\frac{\partial \Psi_i}{\partial x_j}\right\}_{i,j=1}^{n}.$$

We have

$$\begin{aligned}\frac{\partial \Psi_i}{\partial x_j} &= \frac{\partial}{\partial x_j}L[p_i(t)-q_i(t)]\\ &= L\left[\frac{\partial}{\partial x_j}(p_i(t)-q_i(t))\right]\\ &= -\frac{\partial}{\partial x_j}\left\{\frac{I_j[p_i]}{I_j}\right\}L[p_j] + \frac{\partial}{\partial x_j}\left\{\frac{I_j[q_i]}{I_j}\right\}L[q_j]\\ &= a_j d_j[r_i]\,,\end{aligned}$$

where $r_i := q_i - p_i$ and

$$d_j[g] := \frac{\partial}{\partial x_j}\left\{\frac{I_j[g]}{I_j}\right\}.$$

Thus,

$$\det J(\mathbf{x},\mathbf{h}) = a_1 \cdots a_n \det J_1(\mathbf{x},\mathbf{h}) \,,$$

J_1 being the matrix

$$J_1(\mathbf{x},\mathbf{h}) := \begin{bmatrix} d_1[r_1] & d_2[r_1] & \cdots & d_n[r_1] \\ d_1[r_2] & d_2[r_2] & \cdots & d_n[r_2] \\ \vdots & \vdots & \vdots & \vdots \\ d_1[r_n] & d_2[r_n] & \cdots & d_n[r_n] \end{bmatrix}.$$

Besides, the coefficients of any Gaussian formula are positive (see [2]). Thus $a_j > 0$ at any solution $\mathbf{x}$ of (8). Next we examine the sign of $J_1(\mathbf{y},\mathbf{v})$ in the set

$$E_{\varepsilon_0} := \{(\mathbf{y},\mathbf{v}) \,:\, \mathbf{v} \in H \,,\ \mathbf{y} \in D_{\varepsilon_0}(\mathbf{v})\}.$$

E_{ε_0} is a bounded connected set with nonempty interior which contains all the points $(\mathbf{x},\mathbf{h})$ corresponding to any Gaussian formula (1).

Now we show that $\det J_1(\mathbf{y},\mathbf{v}) \neq 0$ at every point $(\mathbf{y},\mathbf{v}) \in E_{\varepsilon_0}$. Assume the contrary, i.e., $\det J_1(\mathbf{x},\mathbf{h}) = 0$ for some $(\mathbf{x},\mathbf{h}) \in E_{\varepsilon_0}$. Then there exists a linear dependence between the rows of $J_1(\mathbf{x},\mathbf{h})$, and thus there exists a nonzero vector $(b_1,\ldots,b_n)$ such that

$$\text{(9)} \qquad d_j[r] = \frac{\partial}{\partial x_j}\left\{\frac{I_j[r]}{I_j}\right\} = 0\,, \quad j = 1,\ldots,n\,,$$

where $r := \sum_{i=1}^{n} b_i r_i$. In case $h_j = 0$ the last condition reduces to $r'(x_j) = 0$. Besides, by (6) and (7), $r(x_j) = 0$ and thus r has a double zero at x_j. If $h_j > 0$, then $I_j > 0$ and (9) leads to

$$\left(\frac{\partial}{\partial x_j} I_j[r]\right) I_j - I_j[r] \frac{\partial}{\partial x_j} I_j = 0.$$

On the other hand, (6) and (7) yield

$$I_j[r] = 0\,, \quad j = 1,\ldots,n.$$

Therefore, if $h_j > 0$, (9) reduces to

$$\frac{\partial}{\partial x_j} I_j[r] = 0.$$

Performing the differentiation we obtain

$$\frac{\partial}{\partial x_j} I_j[r] = \mu(x_j + h_j) r(x_j + h_j) - \mu(x_j) r(x_j) = 0.$$

Since $I_j[r] = 0$ and $\mu(t) > 0$, the function $r(t)$ must have at least two zeros in $[x_j, x_j + h_j]$ (counting the multiplicities up to 2). And this holds for every $j = 1,\ldots,n$. Therefore r has at least $2n$ zeros in (a,b), and hence $r \equiv 0$. This was the point at which we used that $\mathcal{U}_{2n}$ is an ET space of order 2. Then

$$r = \sum_{i=1}^{n} b_i (q_i - p_i) \equiv 0,$$

and thus $\sum_{i=1}^{n} b_i p_i$ must vanish at $s_1, \ldots, s_n, t_1, \ldots, t_n$. This means that $b_1 = \cdots = b_n = 0$, which is a contradiction. Therefore $\det J_1(\mathbf{y}, \mathbf{v})$ does not vanish in E_{ε_0}, and consequently $\det J(\mathbf{x}, \mathbf{h}) \neq 0$ (and even has a constant sign) at any solution $\mathbf{x}$ of system (8).

Now we are ready to complete the proof of Theorem 1. The existence of (1) was proved in [2] in a more general case. It can also be derived easily from the implicit function theorem. Indeed, for any given $\mathbf{h} \in H_\varepsilon$ we consider the family of lengths $\alpha\mathbf{h}$, parameterized by α, $0 \leq \alpha \leq 1$. According to Krein's theorem (see [5]), the Gaussian quadrature exists for $\alpha = 0$. We shall use this fact to extend the solution $(\mathbf{x}(\alpha), \alpha\mathbf{h})$ to any $0 \leq \alpha \leq 1$. And this can be done by the implicit function theorem since the Jacobian J is different from zero at any solution $(\mathbf{x}(\alpha), \alpha\mathbf{h})$ of (8) (since $\alpha\mathbf{h} \in H_\varepsilon$ for $0 \leq \alpha \leq 1$).

It remains to prove that the Gaussian quadrature is unique. Assume the contrary, that is, assume that system (8) has two distinct solutions $\mathbf{x}$ and $\mathbf{y}$. Consider the unique extensions $\mathbf{x}(\alpha)$, $\mathbf{y}(\alpha)$ of these solutions for α going back from 1 to 0. We have $\mathbf{x}(1) = \mathbf{x}$, $\mathbf{y}(1) = \mathbf{y}$, and $\mathbf{x}(1) \neq \mathbf{y}(1)$. On the other hand, by Krein's theorem $\mathbf{x}(0) = \mathbf{y}(0)$. Let us set

$$\alpha_0 := \max\{\alpha \in [0,1) : \mathbf{x}(\alpha) = \mathbf{y}(\alpha)\}.$$

Then, there must be two different extensions of $\mathbf{x}(\alpha_0) = \mathbf{y}(\alpha_0)$ in a neighborhood of α_0, which is a contradiction to the implicit function theorem. This ends the proof. $\square$

It is worth mentioning explicitly the important particular case of interval quadrature formula of Gauss–Christoffel type.

Corollary 1. *Let μ be any integrable function on $[a, b]$ that is continuous and positive on (a, b). Then, for every given set of nonnegative numbers $\mathbf{h}$ satisfying the condition*

$$h_1 + \cdots + h_n < b - a,$$

there exists a unique set of nodes $\mathbf{x} \in D(\mathbf{h})$ such that

$$\int_a^b \mu(t) f(t)\, dt = \sum_{k=1}^{n} a_k \frac{1}{I_k} \int_{x_k}^{x_k + h_k} \mu(t) f(t)\, dt$$

for every algebraic polynomial of degree less than or equal to $2n - 1$.

Acknowledgments. This work was accomplished while the second author was visiting Linköping University. He would like to thank Professor Lars-Erik Andersson for the invitation and the Swedish Institute for the scholarship.

REFERENCES

[1] B. Bojanov, *A note on the Hobby-Rice and Gauss-Krein theorems*, East J. Approx., 4 (1998), pp. 371–377.

[2] B. Bojanov and P. Petrov, *Gaussian interval quadrature formula*, Numer. Math., 87 (2001), pp. 625–643.

[3] B. Bojanov and P. Petrov, *Uniqueness of the Gaussian interval quadrature formula*, Numer. Math., 95 (2003), pp. 53–62.

[4] S. Karlin and W. J. Studden, *Tchebycheff Systems: With Applications in Analysis and Statistics*, Pure Appl. Math. XV, Wiley-Interscience, New York, 1966.

[5] M. Krein, *The ideas of P. L. Tchebycheff and A. A. Markov in the theory of limiting values of integrals and their further developments*, Uspehi Matem. Nauk (N.S.), 6 (1951), pp. 3–120 (in Russian); Amer. Math. Soc. Transl. (2), 12 (1959), pp. 1–121 (in English).

[6] M. G. Krein and A. A. Nudelman, *The Markov Moment Problem and Extremal Problems*, Nauka, Moscow, 1973 (in Russian); Transl. Math. Monogr. 50, AMS, Providence, RI, 1977 (in English).

[7] A. L. Kuz'mina, *Interval quadrature formulas with multiple nodal intervals*, Izv. Vyssh. Uchebn. Zaved. Mat., 7 (1980), pp. 39–44 (in Russian).

[8] G. Milovanović and A. Cvetković, *Uniqueness and computation of Gaussian interval quadrature formula for Jacobi weight function*, Numer. Math., 99 (2004), pp. 141–162.

[9] V. P. Motornyi, *On the best interval quadrature formula in the class of functions with bounded rth derivative*, East J. Approx., 4 (1998), pp. 459–478.

[10] M. Omladič, S. Pahor, and A. Suhadolc, *On a new type of quadrature formulas*, Numer. Math., 25 (1975/76), pp. 421–426.

[11] Fr. Pittnauer and M. Reimer, *Interpolation mit Intervallfunktionalen*, Math. Z., 146 (1976), pp. 7–15.

[12] Fr. Pittnauer and M. Reimer, *Intervallfunktionale vom Gauss-Legendre-Type*, Math. Nachr., 87 (1979), pp. 239–248.

SIAM J. NUMER. ANAL.
Vol. 43, No. 2, pp. 796–824

NUMERICAL SOLUTIONS TO COMPRESSIBLE FLOWS IN A NOZZLE WITH VARIABLE CROSS-SECTION*

DIETMAR KRÖNER† AND MAI DUC THANH‡

Abstract. Compressible flows in a nozzle can be modeled by the gas dynamics equations in one-dimensional space with source terms. It turns out that along stationary waves, the entropy is conserved. Investigating properties of the system leads us to the determination of stationary waves. Relying on this analysis, we construct a numerical scheme which takes into account the use of stationary waves. Our scheme is shown to be capable of maintaining equilibrium states. This demonstrates the efficiency of the new scheme over classical ones, which usually give unsatisfactory results when reducing the refinement of the mesh-size. Moreover, our scheme converges much faster than the classical ones in most cases.

Key words. gas dynamics, nozzle, conservation law, weak solution, stationary wave, modified Lax–Friedrichs scheme

AMS subject classifications. Primary, 35L65, 74XX; Secondary, 65BXX, 76N10, 76L05

DOI. 10.1137/040607460

1. Introduction.

In this paper we consider the following one-dimensional space gas dynamics equations that describe the evolution of a gas flow in a nozzle with cross-sectional area $a = a(x) > 0, x \in \boldsymbol{R}$:

$$\begin{aligned} &\partial_t(a\rho) + \partial_x(a\rho u) \;=\; 0, \\ &\partial_t(a\rho u) + \partial_x(a(\rho u^2 + p)) \;=\; p\partial_x a, \\ &\partial_t(a\rho e) + \partial_x(au(\rho e + p)) = 0, \quad x \in \boldsymbol{R},\, t > 0. \end{aligned} \tag{1.1}$$

As usual, the thermodynamical variables $\varepsilon, \rho, v = 1/\rho, p, T, S$ are the internal energy, density, specific volume, pressure, absolute temperature, and entropy, respectively; u is the velocity; and $e = \varepsilon + u^2/2$ is the total energy. The system (1.1) has the form of a system of conservation laws with source terms. It is nonconservative due to the effect of the geometry.

An efficient way to deal with system (1.1) is to supplement it with an additional trivial equation,

$$\partial_t a = 0 \tag{1.2}$$

(see [24, 21]), and then treat a (nontrivially) as an unknown. This step drives out the obstacle of the nonconservativeness in producing another characteristic field which turns out to be linearly degenerate. However, the system is then nonstrictly hyperbolic. Consequently, this causes the ill-posedness of the initial value problem. The

*Received by the editors April 28, 2004; accepted for publication (in revised form) April 15, 2005; published electronically September 15, 2005.

http://www.siam.org/journals/sinum/43-2/60746.html

†Institute of Applied Mathematics, University of Freiburg, Hermann-Herder Str. 10, 79104 Freiburg, Germany (dietmar@mathematik.uni-freiburg.de).

‡Faculty of Applied Science, University of Technology, 268 Ly Thuong Kiet Street, District 10, Ho Chi Minh City, Vietnam. Current address: Department of Mathematics, International University–National University Ho Chi Minh City, 6th Street, Linh Trung, Thu Duc Disctrict, Ho Chi Minh City, Vietnam (mdthanh@hcmiu.edu.vn).

This author was supported by a postdoctoral research fellowship from the Deutschen Forschungsgemeinschaft in the Graduiertenkolleg *Nichtlineare Differentialgleichungen: Modellierung, Theorie, Numerik, Visualisierung*, University of Freiburg.

reader is referred to the related works in [24, 14, 15, 9, 22, 1]. In [22], the Riemann problem for isentropic flows with discontinuous cross-section area was solved.

Due to the geometry, system (1.1) has the form of conservation laws with source terms. There is another familiar form of systems of conservation laws with sources: the model of shallow water equations. There have recently been lots of contribution studies of hyperbolic systems of conservation laws with source terms. A well-balanced scheme was proposed for a single conservation law by [12, 13]. In [23], the author proposed modified Riemann problems for the Godunov method. Recently, well-balanced schemes aimed at dealing with conservation laws with sources were constructed in [6, 11, 4, 5, 2, 16]. In particular, for a single conservation law with sources, the method in [4, 5] actually gives a high accuracy, and seems to be efficient at capturing stationary solutions. The computations and comparisons with two-dimensional averaged computing were carried out in [1].

Our main purpose in this paper is to construct a new scheme for the system of conservation laws with source terms (1.1) which is capable of maintaining equilibrium states. The idea was originally proposed by Greenberg and Leroux [12] (see also [13]) for scalar conservation laws. Here, we first provide some background and investigate properties of the system (1.1). The motivation is given and the notion of stationary contact waves for ideal isentropic gas flows [22] is reviewed and is formulated for general nonisentropic fluid flows. It turns out that the entropy is conserved across any stationary solution. Based on this step, we propose a new scheme which is capable of capturing stationary contact waves whenever available. The new scheme is constructed from the following arguments: first we take care of the effect of the source terms which produce equilibrium states resulted by stationary contact waves; second, using any standard finite difference scheme for gas dynamics equations (without source terms) in one-dimensional space, we take into account all the available equilibrium states. Our method can be seen as being composed of two steps, in which the first one deals with stationary waves, which are independent of time. Thus, we can argue that the first step provides an *immediate response* that gives us equilibrium states. Therefore, the first step does not delay the evolution of the fluid described by system (1.1), roughly speaking. Since the first step provides us with the exact equilibrium states produced from the source terms, our method has advantage over the standard finite difference schemes for gas dynamics equations with a usual discretization of the source terms, referred in the following as classical schemes. This explains the reason why the new scheme converges much faster than the classical schemes as seen from our test cases. Although our method can be applied to any standard finite difference scheme, for simplicity, in the test cases we just take the Lax–Friedrichs scheme which has the numerical flux of the form

$$g^{\mathrm{L}}(U,V) := \frac{1}{2}(f(U)+f(V)) - \frac{1}{2\lambda}(V-U). \tag{1.3}$$

Here, we use the definition of numerical fluxes as the one in standard books of conservation laws (see [18, 10], for example).

Our scheme is determined for a broad class of fluids which satisfy

$$2p_\rho(\rho,S) + \rho p_{\rho\rho}(\rho,S) \; > \; 0 \tag{1.4}$$

in the domain under consideration. The condition (1.3) is fulfilled by the *stiffened gas* equation of state (see Menikoff and Plohr [25])

$$p = (\gamma-1)\rho(\varepsilon-\varepsilon_\infty) - \gamma p_\infty, \quad \gamma > 1, \tag{1.5}$$

where $\varepsilon_\infty, p_\infty$ are constants, depending on the material under consideration, and by *van der Waals fluid* equations of states in a certain domain where the system may not be genuinely nonlinear, but necessarily strictly hyperbolic. Notice that a stiffened gas equation of state reduces to an ideal polytropic gas when $p_\infty = 0$ and $\varepsilon_\infty = 0$. And, as mentioned in [25], for some materials, p_∞ can be quite large.

To demonstrate the efficiency of our scheme, we consider test cases for polytropic gases which have equations of state of the form

$$p = (\gamma - 1)\rho\varepsilon, \quad \gamma > 1, \tag{1.6}$$

for nonisentropic gases and

$$p = \kappa\rho^\gamma, \quad \kappa > 0, \gamma > 1, \tag{1.7}$$

for isentropic gases, which for simplicity, we take $\kappa = 1$. Precisely, the test cases include the following: two tests for stationary waves for a nonisentropic polytropic gas (1.6) and for an isentropic ideal gas (1.7); seven tests for nonstationary waves for an isentropic ideal gas (1.7); and two tests for nonstationary waves for a nonisentropic polytropic gas (1.6). We note that the seven test cases for the isentropic gas (1.7) cover all the possibilities of the location of left- and right-hand states of the Riemann problem: either the left- and right-hand states lying in the same phase or in different phases. Here, we call a *phase* the biggest region of the phase domain where the system remains strictly hyperbolic. As shown in [22], there are exactly three such phases.

The paper is organized as follows. Section 2 provides a simple way of modeling of fluid flows in a nozzle. Although the model has been known for a long time, to our knowledge the complete derivation is not available in the literature. Therefore, we want to review it for the sake of completeness. We assume the nozzle is smooth with small variation so that the flow through it would have some symmetry and relatively uniform properties. We note here that for fluid flows in an arbitrary nozzle, and in particular a nozzle with discontinuous cross-sections, we can consider the same model, but the derivation presented here may not be applied. In section 3 we will investigate the properties of general fluid flows, and we will give the definition of elementary waves. We will show that the entropy is conserved across any stationary wave. This will help us construct the new scheme by solving the jump relations of a system as if it were of isentropic gases. Section 4 will deal with the analysis of stationary waves which yields their determination at the end. In section 5 we will construct the numerical scheme, relying on the use of stationary waves. Section 6 will show the evidence of the advantage of the new scheme over the classical ones. Here we compare the CPU times of the two schemes: The modified Lax–Friedrichs scheme, which takes the Lax–Friedrichs scheme together with a usual cell average discretization of the source term, and our new one.

The model of fluid flows in a nozzle (1.1) is closely related to the model of multiphase flows which has a vast domain of applications. Formally, the previous one can be obtained from the last one by restricting it to a single fluid. Both systems share common features such as the lack of strict hyperbolicity, systems of conservation laws with sources, ill-posedness, etc. A two-phase mixture theory for reactive granular materials was introduced by [3]. We note that a theory of multiphase flows was established in the book [8]. An overview of the ill-posedness of the two-fluid model was observed in [7]. In recent research [17], the authors investigated properties of the two-fluid model which lacks hyperbolicity. See also the references therein.

Several important properties of the numerical scheme presented in this paper concerning the *minimum principle* of the numerical entropy and the *nonnegativity* of the numerical density can be found in [20]. A well-balanced scheme for the nonreactive version of the Baer–Nunziato (BN) model of two-phase fluids which aim to maintain equilibrium states has been under study in [19].

2. A simple way of modeling. The model (1.1) is well known and one could easily find it in the literature. However, we do not know whether it was derived in a complete form. The derivation in the stationary case can be found in the standard book of Zucrow and Hoffman [26]. Here, for the sake of completeness, we present a simple way of modeling the physical phenomenon. To validate our analysis below, *we restrict our attention to a given smooth nozzle with small variation.*

Let the Euler equations in a two-dimensional tube with height $a(x) > 0$ be given:

$$\begin{aligned} &\partial_t \rho + \partial_x(\rho u) + \partial_y(\rho v) \;=\; 0,\\ &\partial_t(\rho u) + \partial_x(\rho u^2 + p) + \partial_y(\rho uv) \;=\; 0,\\ &\partial_t(\rho v) + \partial_x(\rho uv) + \partial_y(\rho v^2 + p) \;=\; 0,\\ &\partial_t(\rho e) + \partial_x(u(\rho e + p)) + \partial_y(v(\rho e + p)) = 0, \quad x, y \in \boldsymbol{R},\ t > 0. \end{aligned} \tag{2.1}$$

Let us assume that the nozzle is put in the x-direction (see Figure 2.1). First, we consider the first and the second equations. Integrating the first and the second equations of (2.1) in y we obtain, respectively,

$$\begin{aligned} &\int_0^{a(x)} [\partial_t \rho + \partial_x(\rho u) + \partial_y(\rho v)]\, dy \;=\; 0,\\ &\int_0^{a(x)} \big[\partial_t(\rho u) + \partial_x(\rho u^2 + p) + \partial_y(\rho uv)\big]\, dy \;=\; 0, \quad x \in \boldsymbol{R},\ t > 0, \end{aligned}$$

or

$$\begin{aligned} &\partial_t\bigg(\int_0^{a(x)} \rho(x,y,t)\, dy\bigg) + \int_0^{a(x)} \partial_x(\rho u)\, dy + \int_0^{a(x)} \partial_y(\rho v)\, dy \;=\; 0,\\ &\partial_t\bigg(\int_0^{a(x)} \rho u(x,y,t)\, dy\bigg) + \int_0^{a(x)} \partial_x(\rho u^2 + p)\, dy + \int_0^{a(x)} \partial_y(\rho uv)\, dy \;=\; 0. \end{aligned} \tag{2.2}$$

Clearly,

$$\begin{aligned} &\int_0^{a(x)} \partial_y(\rho v)\, dy = \rho v(x, a(x), t) - \rho v(x, 0, t),\\ &\int_0^{a(x)} \partial_y(\rho uv)\, dy = \rho uv(x, a(x), t) - \rho uv(x, 0, t). \end{aligned} \tag{2.3}$$

In addition, at the boundary, particles are moving along the boundary and thus their trajectories follow the shape of the boundary. Thus, the particle velocity (u, v) at each point on the boundary, which has the coordinate $(x, y = a(x))$, is the tangent of the nozzle at that point. In other words, it holds that

$$v(x, a(x), t) = u(x, a(x), t) \tan\alpha,$$

where α is the angle between the tangent of the graph of $a(x)$ and the x-axis. Since

$$\tan\alpha = a'(x),$$

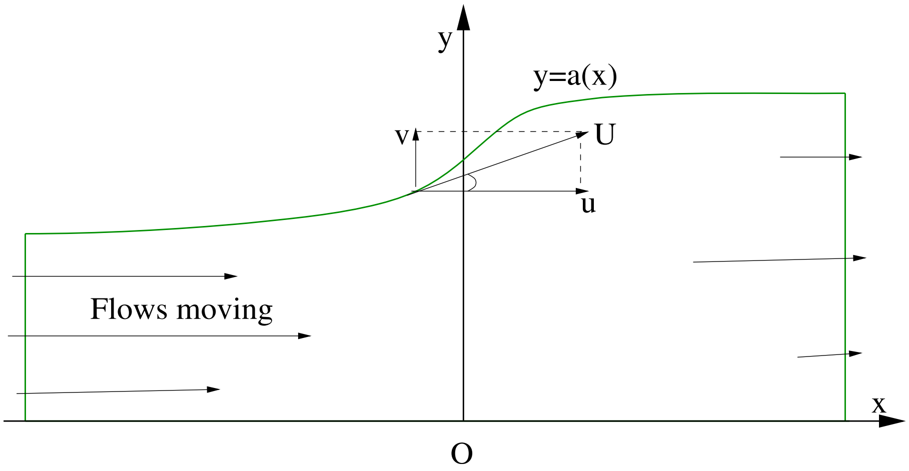


FIG. 2.1. *Flows moving inside a nozzle.*

we have from the last two equations

$$v(x, a(x), t) = a'(x)u(x, a(x), t). \tag{2.4}$$

Similarly,

$$v(x, y = 0, t) = 0, \tag{2.5}$$

since the boundary of the nozzle at the bottom $y = 0$ is parallel to the x-direction. It is derived from (2.3) to (2.5) that

$$\begin{aligned} &\int_0^{a(x)} \partial_y(\rho v)\, dy = a'(x)\rho u(x, a(x), t),\\ &\int_0^{a(x)} \partial_y(\rho uv)\, dy = a'(x)\rho u^2(x, a(x), t). \end{aligned} \tag{2.6}$$

On the other hand, we have

$$\begin{aligned} \partial_x\left(\int_0^{a(x)} \rho u(x,y,t)\, dy\right) &= \int_0^{a(x)} \partial_x(\rho u)(x,y,t)\, dy + a'(x)\rho u(x, a(x), t),\\ \partial_x\left(\int_0^{a(x)} \rho u^2(x,y,t)\, dy\right) &= \int_0^{a(x)} \partial_x(\rho u^2)(x,y,t)\, dy + a'(x)\rho u^2(x, a(x), t),\\ \partial_x\left(\int_0^{a(x)} p(x,y,t)\, dy\right) &= \int_0^{a(x)} \partial_x p(x,y,t)\, dy + a'(x)p(x, a(x), t). \end{aligned} \tag{2.7}$$

From (2.2) and (2.7), we obtain

$$\begin{aligned} &\partial_t\left(\int_0^{a(x)} \rho(x,y,t)\, dy\right) + \partial_x\left(\int_0^{a(x)} \rho u(x,y,t)\, dy\right) = 0,\\ &\partial_t\left(\int_0^{a(x)} \rho u(x,y,t)\, dy\right) + \partial_x\left(\int_0^{a(x)} (\rho u^2 + p)(x,y,t)\, dy\right) = a'(x)p(x, a(x), t). \end{aligned} \tag{2.8}$$

To obtain the first two equations of (1.1) from (2.8), we make the following assumptions.

- The cross-section average density and pressure $\bar{\rho}, \bar{p}$ play the role of ρ, p, respectively, where, for example,

$$\bar{\rho}(x,t) := \frac{1}{a(x)} \int_0^{a(x)} \rho(x,y,t)\,dy, \quad x \in \boldsymbol{R}, t > 0.$$

- The u-component of the particle velocity is uniform in each cross-section, i.e.,

$$u(x,y,t) \;=\; u(x,y',t) := \bar{u}(x,t) \quad \forall y,y' \in (0,a(x)) \quad \forall x \in \boldsymbol{R}, t > 0.$$

- The average pressure $\bar{p}$ is approximated by $p(x,a(x),t)$.

We observe that these assumptions are quite limited, since the nozzle then should be smooth with small variation so that the flow through it would approximately have the symmetry and uniform properties. Since the system (1.1) serves only the component u of the velocity in the x-direction, we can ignore the third equation of (2.1). The third equation of (1.1) can be similarly derived as the first equation. The modeling is complete.

3. Basic properties and stationary waves. In this section we recall basic properties of system (1.1)–(1.2) and draw elementary conclusions for stationary waves which will be used in the next sections. Together with shock waves and rarefaction waves in genuinely nonlinear characteristic fields, and contact discontinuities in the linearly degenerate field of the usual gas dynamics equations, stationary waves will be defined as one kind of *elementary wave*. We will show that *the entropy remains constant across stationary waves*. This result enables us to determine stationary waves as in the isentropic case (see [22]).

Precisely, we are studying the following system:

$$\begin{aligned}
&\partial_t(a\rho) + \partial_x(a\rho u) \;=\; 0,\\
&\partial_t(a\rho u) + \partial_x(a(\rho u^2 + p)) \;=\; p\partial_x a,\\
&\partial_t(a\rho e) + \partial_x(au(\rho e + p)) = 0,\\
&\partial_t a = 0, \quad x \in \boldsymbol{R},\, t > 0.
\end{aligned} \tag{3.1}$$

In what follows, the density and the entropy are chosen as independent thermodynamics variables. Note that other thermodynamics variables can always be expressed in terms of any two thermodynamics variables. Thus, we can write $p = p(\rho, S), \varepsilon = \varepsilon(\rho, S)$, etc. Let us be given a smooth solution $U(x,t) = (\rho(x,t), u(x,t), S(x,t), a(x))$ of the system (4.1). A straightforward calculation shows that the smooth solution U satisfies the following system of conservation laws in nonconservative form:

$$\begin{aligned}
&\rho_t + u\rho_x + \rho u_x + \frac{u\rho}{a} a_x \;=\; 0,\\
&u_t + uu_x + \frac{1}{\rho} p_x \;=\; 0,\\
&S_t + uS_x \;=\; 0,\\
&a_t = 0, \quad x \in \boldsymbol{R},\, t > 0.
\end{aligned} \tag{3.2}$$

The Jacobian matrix of the system (3.2) is given by

$$B(U) = \begin{pmatrix} u & \rho & 0 & \frac{u\rho}{a} \\ \frac{p_\rho}{\rho} & u & \frac{p_S}{\rho} & 0 \\ 0 & 0 & u & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.3}$$

The characteristic equation is given by

$$\lambda(u-\lambda)[(u-\lambda)^2 - p_\rho] = 0. \tag{3.4}$$

Note. In fact, (3.4) is obtained by multiplying a factor λ, which represents an additional linearly degenerate field, to the characteristic equation of the usual gas dynamics equations.

Thus, we obtain four real eigenvalues,

$$\lambda_0 = 0, \quad \lambda_1 = u - c, \quad \lambda_2 = u, \quad \lambda_3 = u + c, \tag{3.5}$$

where c is the *local sound speed*

$$c = \sqrt{p_\rho} \geq 0.$$

The associated eigenvectors of $B(U)$ can be chosen as

$$r_0 = \begin{pmatrix} u^2\rho \\ -up_\rho \\ 0 \\ a(p_\rho - u^2) \end{pmatrix}, \quad r_1 = \begin{pmatrix} \rho \\ -c \\ 0 \\ 0 \end{pmatrix}, \quad r_2 = \begin{pmatrix} -p_S \\ 0 \\ p_\rho \\ 0 \end{pmatrix}, \quad r_3 = \begin{pmatrix} \rho \\ c \\ 0 \\ 0 \end{pmatrix}.$$

In the following we will assume that the local sound speed is positive:

$$c = \sqrt{p_\rho} > 0.$$

The system (3.1) is in general *not strictly hyperbolic*, since the characteristic field associated with λ_0 may coincide with any other field. The phase domain is in fact divided into four regions in which the system is strictly hyperbolic:

$$\begin{aligned} G_1 &= \{U : \lambda_1(U) < \lambda_2(U) < \lambda_3(U) < \lambda_0(U)\}, \\ G_2 &= \{U : \lambda_1(U) < \lambda_2(U) < \lambda_0(U) < \lambda_3(U)\}, \\ G_3 &= \{U : \lambda_1(U) < \lambda_0(U) < \lambda_2(U) < \lambda_3(U)\}, \\ G_4 &= \{U : \lambda_0(U) < \lambda_1(U) < \lambda_2(U) < \lambda_3(U)\}. \end{aligned} \tag{3.6}$$

In what follows, we will call each of these regions a *phase*. The phases are separated by three isolated surfaces on which the system fails to be strictly hyperbolic:

$$\begin{aligned} \Sigma_+ &= \{U : \lambda_1(U) = \lambda_0(U)\}, \\ \Sigma_0 &= \{U : \lambda_2(U) = \lambda_0(U)\}, \\ \Sigma_- &= \{U : \lambda_3(U) = \lambda_0(U)\}. \end{aligned} \tag{3.7}$$

We will refer to each of these surfaces a *hyperbolic boundary*.

Clearly, the 0- and the 2-characteristic fields are linearly degenerate. And under the condition that

$$\frac{\partial c}{\partial \rho} \geq 0,$$

one can easily check that the 1- and the 3-characteristic fields are genuinely nonlinear.

A stationary smooth solution U of (1.1) is a time-independent smooth solution. Thus, the derivative in t in (1.1) can be omitted. Stationary solutions of the initial value problem for (1.1) are, therefore, the ones for the following ordinary differential equations:

$$\begin{aligned} &(a\rho u)' = 0, \\ &(a(\rho u^2 + p))' = pa', \\ &(au(\rho e + p))' = 0, \end{aligned} \tag{3.8}$$

with the smooth initial data

$$U(x,0) = U_0(x), \tag{3.9}$$

where we denote

$$\frac{d(.)}{dx} = (.)'.$$

The following lemma will help us in computing stationary contacts.

LEMMA 3.1. *The system* (3.8) *for smooth solutions is equivalent to*

$$\begin{aligned} (a\rho u)' &= 0, \\ uu' + \frac{p'}{\rho} &= 0, \\ S' &= 0. \end{aligned} \tag{3.10}$$

In other words, the entropy is conserved across any stationary smooth solution of the initial value problem for (1.1).

Consequently, the nonisentropic system is reduced to the isentropic case for stationary solutions. Therefore, we can define a *stationary contact of* (3.1) as the limit of sequences of stationary smooth solutions in a similar way of the isentropic gases (see [22]).

Proof. Let us be given initial data U_0. The first equation of (4.8) can be expressed as

$$a\rho u = C$$

for some constant C dependent only on U_0. Therefore, the second equation can be written as

$$(C \cdot u + a \cdot p)' = p \cdot a'$$

or

$$C \cdot u' + a \cdot p' = 0.$$

Restoring C from its original definition, we obtain

$$a\rho\biggl(uu' + \frac{p'}{\rho}\biggr) = 0.$$

Since the cross-section is assumed to be positive, the last equation can be written as the second equation of (3.10). Besides, the third equation of (3.8) can be written as

$$C \cdot \biggl(e + \frac{p}{\rho}\biggr)' = 0. \tag{3.11}$$

Recall the thermodynamics identity that

$$TdS = d\varepsilon + pdv, \quad v = \frac{1}{\rho}.$$

Since we are considering stationary waves, i.e., solutions independent of time, the thermodynamics identity applied to this kind of wave gives

$$TS' = \varepsilon' + pv'. \tag{3.12}$$

It is derived from the definition of the total energy and from (3.12) that (3.11) can be written as

$$\varepsilon' + uu' + (pv)' = 0$$

or

$$TS' + \frac{p'}{\rho} + uu' = 0.$$

In view of the second equation of (3.10) we just established, the last equation takes the simpler form

$$S' = \frac{dS}{dx} = 0.$$

Thus, we have demonstrated that system (3.8) is equivalent to (3.10).

Furthermore, since we are considering stationary waves, we have

$$dS = \frac{dS}{dx}dx + \frac{dS}{dt}dt = 0, \tag{3.13}$$

which implies the rest of Lemma 3.1. □

Let us now discuss the notion of shock waves of the system (3.1). A shock wave from a left-hand state $U_- = (\rho_-, u_-, S_-, a_-)$ to a right-hand state $U_+ = (\rho_+, u_+, S_+, a_+)$, which propagates with a speed $\bar{\lambda}$, should satisfy the Rankine–Hugoniot relation for the last equation of (3.1), which yields

$$\bar{\lambda}[a] \; = \; 0, \tag{3.14}$$

where $[a] = a_+ - a_-$. This implies that either

- the component a must be constant, or
- the shock speed is equal to zero (*stationary contact*).

Assume first that this shock corresponds to a constant level of the cross-section a. The right-hand side of the system (3.1) thus vanishes, and system (3.1) is reduced to the usual form of conservation laws. It is not difficult to check that the left- and right-hand states of the shock are constraint by the Rankine–Hugoniot relations for the usual form of gas dynamics equations:

$$\begin{aligned}
&\partial_t \rho + \partial_x(\rho u) \; = \; 0,\\
&\partial_t(\rho u) + \partial_x(\rho u^2 + p) \; = \; 0,\\
&\partial_t(\rho e) + \partial_x(u(\rho e + p)) = 0, \quad x \in \boldsymbol{R},\, t > 0.
\end{aligned} \tag{3.15}$$

Assume next that the shock speed $\bar{\lambda}$ is zero. Then, we define this kind of shock wave as the ones provided by Lemma 3.1. Let us introduce the *specific enthalpy*

$$h = \varepsilon + pv, \tag{3.16}$$

which satisfies

$$dh = T\,dS + v\,dp.$$

In view of Lemma 3.1, the left- and right-hand states of the shock are constraint by the jump relations

$$\begin{aligned} &S = S_- = S_+,\\ &p = p(\rho, S_-),\\ &[a\rho u] = 0,\\ &\Big[\frac{u^2}{2} + h(\rho, S_-)\Big] = 0, \end{aligned} \tag{3.17}$$

where h is the specific enthalpy (3.16). Observe in this case that

$$\frac{\partial}{\partial\rho} h(\rho, S_-) = v \frac{\partial}{\partial\rho} p(\rho, S_-).$$

For rarefaction waves of (3.1), which are smooth solutions, it is not difficult to see that they are the ones of the usual gas dynamics equations (3.15). Similar arguments can be applied for contact discontinuities. Finally, we can define the elementary waves of (3.1) as follows.

DEFINITION 3.2. *Elementary waves for the system* (3.1) *are the following ones.*

- *If* $a = \text{const}$, *all the elementary waves* (*shock waves, rarefaction waves, contact discontinuities*) *are the ones of the usual gas dynamics equations* (3.15).
- *The stationary contacts which have zero propagation speed and are given by* (3.17).

4. Equilibrium states. We will describe in this section how to get the exact equilibrium states, i.e., the two left- and right-hand states of a stationary contact. In other words, we search for an exact solution of (3.17) for given data U_- and given cross-sections on the left and right of the stationary contact. This exact solution will be used in the second part for the construction of the new scheme.

Let us be given a state $U_- = (\rho_-, u_-, a_-)$ with a level of cross-section a_- and another level cross-section a_+. As seen in the previous section, a state $U_+ = (\rho_+, u_+, a_+)$ with the level cross-section a_+ which can be connected with U_- via a stationary wave is determined by the system

$$\begin{aligned} &S = S_- = S_+,\\ &p = p(\rho, S_-),\\ &[a\rho u] = 0,\\ &\Big[\frac{u^2}{2} + h(\rho, S_-)\Big] = 0, \end{aligned} \tag{4.1}$$

where h is the specific enthalpy (3.16), which satisfies

$$\frac{\partial}{\partial\rho} h(\rho, S_-) = v \frac{\partial}{\partial\rho} p(\rho, S_-). \tag{4.2}$$

To solve system (4.1) for $\rho_+ = \rho$, we solve the equation

$$\Phi(\rho) := (u_-^2 + 2h(\rho_-, S_-))\rho^2 - 2\rho^2 h(\rho, S_-) \ = \ \Big(\frac{a_- u_- \rho_-}{a_+}\Big)^2. \tag{4.3}$$

To this end, we investigate properties of the function Φ. We have due to (4.2)

$$\begin{aligned}\Phi'(\rho) &= 2(u_-^2 + 2h(\rho_-, S_-))\rho - 4\rho h(\rho, S_-) - 2\rho^2 h_\rho(\rho, S_-)\\ &= 2(u_-^2 + 2h(\rho_-, S_-))\rho - 4\rho h(\rho, S_-) - 2\rho p_\rho(\rho, S_-)\\ &= 2\rho\Big(u_-^2 + 2\int_\rho^{\rho_-} h_\rho(\tau, S_-)\,d\tau - p_\rho(\rho, S_-)\Big)\\ &= 2\rho\Big(u_-^2 + 2\int_\rho^{\rho_-} \frac{p_\rho(\tau, S_-)}{\tau}\,d\tau - p_\rho(\rho, S_-)\Big),\end{aligned}$$

which has the same sign as

$$\Psi(\rho) := \frac{\Phi'(\rho)}{2\rho} = u_-^2 + 2\int_\rho^{\rho_-} \frac{p_\rho(\tau, S_-)}{\tau}\,d\tau - p_\rho(\rho, S_-). \tag{4.4}$$

Assumption (1.4) implies

$$\begin{aligned}\Psi'(\rho) &:= -(2h_\rho(\rho, S_-) + p_{\rho\rho}(\rho, S_-))\\ &= -\frac{1}{\rho}\Big(2p_\rho(\rho, S_-) + \rho p_{\rho\rho}(\rho, S_-)\Big) \ < \ 0 \quad \forall\,\rho.\end{aligned} \tag{4.5}$$

In addition, we can always have

$$\begin{aligned}&\Psi(\rho) > 0 \qquad \text{as} \quad \rho \to 0,\\ &\Psi(\rho) < 0 \qquad \text{for large } \rho,\end{aligned} \tag{4.6}$$

for example, under the assumptions that

$$\begin{aligned}&p_\rho(\rho = 0, S_-) = 0,\\ &p_\rho(\rho, S_-) \to +\infty \quad \text{as} \quad \rho \to +\infty.\end{aligned} \tag{4.7}$$

It is, therefore, derived from (4.4) to (4.6) that there exists exactly one value $\rho = \rho_{\max}$ such that

$$\begin{aligned}&\Phi'(\rho) > 0, \quad \rho < \rho_{\max},\\ &\Phi'(\rho) < 0, \quad \rho > \rho_{\max},\\ &\Phi'(\rho) = 0, \quad \rho = \rho_{\max}.\end{aligned} \tag{4.8}$$

Observe that

$$\begin{aligned}&\Phi(\rho = 0) = 0,\\ &\Phi(\rho) \to -\infty \quad \text{as} \quad \rho \to +\infty,\end{aligned} \tag{4.9}$$

and that the right-hand side of (4.3) is nonnegative. It is thus derived from (4.8) to (4.9) that (4.3) has a solution iff

$$\Phi(\rho_{\max}) \geq \Big(\frac{a_- u_- \rho_-}{a_+}\Big)^2 \tag{4.10}$$

or, equivalently,

$$a_+ \geq a_{\min}(U_-) := \frac{a_- u_- \rho_-}{\sqrt{\Phi(\rho_{\max})}}. \tag{4.11}$$

In this case, we can easily see that (4.3), and, therefore, system (4.1), has two roots, denoted by $\varphi_1(U_-, a_+) \leq \varphi_2(U_-, a_+)$, which coincide iff $a_+ = a_{\min}(U_-)$. The fact that $a = a_-$ still gives us a solution of (4.1) and, therefore, of (4.3), a_- has to satisfy

$$a_{\min}(U_-) \leq a_-. \tag{4.12}$$

For polytropic ideal gases (1.6), it is not difficult to check that $a_{\min}(U_-) = a_-$ iff the state U_- belongs to the hyperbolic boundaries (3.7) (see Lemma 2.3 of [22]). Therefore, we can conclude from (4.11) to (4.12) that *the capturing stationary waves can always be done when a_+ is closed to a_-*, which holds for nozzles with small sudden changes.

In order to select one solution $\varphi_1(U_-, a_+)$ or $\varphi_2(U_-, a_+)$ of (5.1), we will use the following admissibility criterion. Observe that for a given state U_-, the last equation of (4.1) also determines a curve $u = u(\rho)$ in the plan (ρ, u), called *stationary curve*, and, therefore, the third equation of (4.1) implies the component a can be expressed as a function $a = a(\rho)$ of the variable ρ along this curve. Let us impose an additional admissibility criterion (see [14, 22]).

Admissibility Criterion 4.1. *Along the stationary curve in the (ρ, u)-plan between left- and right-hand states of any stationary wave, the component a obtained from* (4.1) *and expressed as a function of ρ has to be monotone in ρ.*

We know by the previous section that entropy is constant along stationary curves. Thus, we arrive at the following lemma as in the case of isentropic gases [22], and, therefore, we omit the proof.

Lemma 4.1. *Admissibility Criterion* 4.1 *is equivalent to the statement that any stationary wave has to remain in the closure of only one phase.*

Remark 1. Without such an admissibility condition, the Riemann problem for piecewise constant cross-sections may admit a one-parameter family of solutions; when this condition is used, the Riemann problem may have at most three solutions (see [14, 22, 9]). Here, it is interesting that the approximate solutions given by our scheme and the chosen standard Lax–Friedrichs scheme will both converge to the same exact solution in all the tests (see section 6).

Remark 2. It is derived from (4.1) that, under Admissibility Criterion 4.1, the equilibrium states depend continuously on the cross-section component in the sense that when the left- and the right-hand cross-sections $a_\pm$ of a stationary wave are closed, the corresponding states $U_\pm$ remain closed as well. That implies the continuous dependence on the data of our numerical scheme defined in the next section when the nozzle is a continuous function of $x \in \boldsymbol{R}$. If the nozzle is discontinuous, this conclusion may not be held. In fact, let one state U_0 of the stationary wave with fixed distinct cross-sections $a_\pm$ belong to the hyperbolic boundaries. Then, a small perturbation to this state U_0 would cause a considerable difference. For example, let a small $\varepsilon > 0$ and $U_0 \in \Sigma_+$ be given (see (3.7)). We consider the stationary wave with the cross-sections $a_\pm$ from any fixed left-hand state $U_1 \in B(U_0, \varepsilon) \cap G_3$, and the stationary wave from any fixed left-hand state $U_2 \in B(U_0, \varepsilon) \cap G_4$, where $B(U_0, \varepsilon)$ is the open ball with center U_0 and radius ε. The stationary wave from U_1 remains in G_3 while the one from U_2 belongs to G_4 in view of Lemma 4.1. As a consequence, the structures of solutions (e.g., the number and the order of waves) are thus different as observed from the constructions of the Riemann solutions in [22, 9].

Since a is positive, the function $a = a(\rho)$ is decreasing (increasing) in ρ between ρ_- and ρ_+ if the function

$$\left(\frac{a_- u_- \rho_-}{a(\rho)}\right)^2 = (u_-^2 + 2h(\rho_-, S_-))\rho^2 - 2\rho^2 h(\rho, S_-) \ = \Phi(\rho) \tag{4.13}$$

is increasing (decreasing, respectively) in ρ between ρ_- and ρ_+. Hence, Admissibility Criterion 4.1 selects only one root $\rho = \varphi_i(U_-, a_+)$, $i \in \{1, 2\}$, of (4.1) such that the function Φ is monotone between ρ_- and $\varphi_i(U_-, a_+)$. That means ρ_- and $\varphi_i(U_-, a_+)$ must be located in the same side with respect to $\rho_{\max}$ obtained from (4.8), i.e.,

$$(\varphi_i(U_-, a_+) - \rho_{\max})(\rho_- - \rho_{\max}) \ \geq \ 0. \tag{4.14}$$

Let us now discuss the geometrical sense of the condition (4.14). Let $u = u(\rho)$ be the stationary curve defined by the last equation of (4.1). Then, since $a(\rho)\rho u(\rho) = a_-\rho_- u_-$ for all ρ, we have

$$\begin{aligned} &u(\rho_{\max}) \cdot u_- \geq 0, \\ &u(\rho_{\max})^2 - p_\rho(\rho_{\max}) = \Psi(\rho_{\max}) = 0. \end{aligned} \tag{4.15}$$

The relations (4.15) imply that the point $U_{\max} := (\rho_{\max}, u(\rho_{\max}), S_-)$ belongs to the hyperbolic boundaries. More precisely,

$$\begin{aligned} U_{\max} &\in \Sigma_+ \quad \text{for} \quad u_- > 0, \\ U_{\max} &\in \Sigma_0 \quad \text{for} \quad u_- = 0, \\ U_{\max} &\in \Sigma_- \quad \text{for} \quad u_- < 0. \end{aligned} \tag{4.16}$$

Now, the condition (4.14) simply implies that the *stationary contact wave remains in the same phase* whenever U_- belongs to one of the phase domains $G_i, i = 1, 2, 3, 4$, or the hyperbolic boundary Σ_0. When U_- belongs to the hyperbolic boundaries $\Sigma_\pm$, in the following scheme, we can avoid the situation of multiple choices by skipping to the next space-step, which means we do not use stationary jump at U_-. Another way, following our experience in the construction of composite wave curves of the Riemann problem [22], is the following:

- choose $\varphi_1(U_-, a_+)$ when $U_- \in \Sigma_+$, choose $\varphi_2(U_-, a_+)$ when $U_- \in \Sigma_-$ for forward construction, i.e., when U_- is the left-hand state and we look for a right-hand state of a stationary contact;
- vice versa for backward construction: given a right-hand state U_+ and another cross-section level a_-, we look for a left-hand state U_- of a stationary contact; then, we choose $\varphi_2(U_+, a_-)$ when $U_+ \in \Sigma_+$, and choose $\varphi_1(U_+, a_-)$ when $U_+ \in \Sigma_-$.

5. Numerical schemes. In this section we will present a new scheme for approximating solutions of the system (1.1), relying on the arguments in the previous sections. Given a uniform time step Δt and a special mesh size Δx, setting $x_j = j\Delta x, j \in \mathbf{Z}$, and $t_n = n\Delta t, n \in \mathbf{N}$, we denote by U_j^n in what follows the approximation of the values $U(x_j, t_n)$ of the exact solution $U = (a\rho, a\rho u, a\rho e)$ of (1.1).

Set

$$\lambda = \frac{\Delta t}{\Delta x}.$$

Let us take any standard finite difference scheme for gas dynamics equations with the numerical flux g^{C}. The *classical scheme* is of the form

$$U_j^{n+1} = U_j^n - \lambda\Big(g^{\mathrm{C}}\big(U_j^n, U_{j+1}^n\big) - g^{\mathrm{C}}\big(U_{j-1}^n, U_j^n\big)\Big) + \frac{\lambda}{2}\big(0, p_j^n(a_{j+1} - a_{j-1}), 0\big), \tag{5.1}$$

where p_j^n is given by suitable equations of state, and $a_j := a(x_j)$. The *modified Lax–Friedrichs scheme* is the one of (5.1) with the Lax–Friedrichs numerical flux:

$$\begin{aligned} &g^{\mathrm{C}}(U,V) := \frac{1}{2}(f(U) + f(V)) - \frac{1}{2\lambda}(V - U),\\ &U := (a\rho, a\rho u, a\rho e), \quad f(U) := (a\rho u, a(\rho u^2 + p), au(\rho e + p)). \end{aligned} \tag{5.2}$$

The constant λ is also required to satisfy the so-called *CFL stability condition*

$$\lambda \max_U |f'(U)| \le 1. \tag{5.3}$$

The motivation of our study in the following new scheme is to take into account the effect of stationary waves. The method was originally developed in [12, 13], and extended to more general problems in [4, 5, 2]. It can be described by two steps.

- First, at each grid node $x_j, j \in \mathbf{Z}$, we determine two stationary waves of (1.1) in which one stationary wave arrives at x_j with the cross-section level a_j from the given left-hand state U_{j-1}^n (with $a = a_{j-1}$) by a right-hand state, denoted by $U_{j-1,+}^n$, and another stationary wave arrives at x_j with the cross-section level a_j from the given right-hand state U_{j+1}^n (with $a = a_{j+1}$) by a left-hand state, denoted by $U_{j+1,-}^n$.
- Second, taking the Lax–Friedrichs scheme for the usual gas dynamics equations (3.15) for computing U_j^{n+1} at time $t = (n+1)h$, we substitute U_{j+1}^n by $U_{j+1,-}^n$ and U_{j-1}^n by $U_{j-1,+}^n$.

Precisely, the new scheme is defined by

$$U_j^{n+1} = U_j^n - \lambda\big(g^{\mathrm{N}}(U_j^n, U_{j+1,-}^n) - g^{\mathrm{N}}(U_{j-1,+}^n, U_j^n)\big), \tag{5.4}$$

where $g^{\mathrm{N}}(U,V)$ can be any standard numerical flux for gas dynamics equations, and $U_{j+1,-}^n, U_{j-1,+}^n$ are given shortly below. In the next section devoted to numerical tests, we take the Lax–Friedrichs numerical flux:

$$\begin{aligned} &g^{\mathrm{N}}(U,V) := g^{\mathrm{C}}(U,V) = \frac{1}{2}\big(f(U) + f(V)\big) - \frac{1}{2\lambda}(V - U),\\ &U := (\rho, \rho u, \rho e), \quad f(U) := \big(\rho u, (\rho u^2 + p), u(\rho e + p)\big). \end{aligned}$$

In the scheme (5.4), the states

$$U_{j+1,-}^n = (\rho, \rho u, \rho e)_{j+1,-}^n, \quad U_{j-1,+}^n = (\rho, \rho u, \rho e)_{j-1,+}^n$$

are defined by observing that the entropy is constant across each stationary jump, and by computing $\rho_{j+1,-}^n, u_{j+1,-}^n$ from the equations

$$\begin{aligned} &a_{j+1}^n \rho_{j+1}^n u_{j+1}^n = a_j^n \rho_{j+1,-}^n u_{j+1,-}^n,\\ &\frac{(u_{j+1}^n)^2}{2} + h(\rho_{j+1}^n) = \frac{(u_{j+1,-}^n)^2}{2} + h(\rho_{j+1,-}^n), \end{aligned} \tag{5.5}$$

and computing $\rho^n_{j-1,+}, u^n_{j-1,+}$ from the equations

$$\begin{aligned} a^n_{j-1}\rho^n_{j-1}u^n_{j-1} &= a^n_j\rho^n_{j-1,+}u^n_{j-1,+}, \\ \frac{(u^n_{j-1})^2}{2} + h(\rho^n_{j-1}) &= \frac{(u^n_{j-1,+})^2}{2} + h(\rho^n_{j-1,+}). \end{aligned} \tag{5.6}$$

Remember that we have for stationary solutions

$$\begin{aligned} a^n_{j+1}\rho^n_{j+1}u^n_{j+1} &= a^n_j\rho^n_j u^n_j, \\ \frac{(u^n_{j+1})^2}{2} + h(\rho^n_{j+1}) &= \frac{(u^n_j)^2}{2} + h(\rho^n_j). \end{aligned} \tag{5.7}$$

Therefore, the definition of $U^n_{j+1,-}, U^n_{j-1,+}$ in (5.5) and (5.6), respectively, implies that in the stationary case the unique solutions of (5.5) and (5.6) will be

$$\begin{aligned} \rho^n_{j+1,-} &= \rho^n_j, \quad u^n_{j+1,-} = u^n_j, \\ \rho^n_{j-1,+} &= \rho^n_j, \quad u^n_{j-1,+} = u^n_j, \end{aligned} \tag{5.8}$$

i.e.,

$$U^n_{j+1,-} = U^n_j, \quad U^n_{j-1,+} = U^n_j,$$

and, therefore (see (5.4)),

$$U^{n+1}_j = U^n_j. \tag{5.9}$$

This means that we exactly recover the stationary solution.

In a domain where a is a constant, it is easy to verify from (5.5) and (5.6) that

$$U^n_{j+1,-} = U^n_{j+1}, \quad U^n_{j-1,+} = U^n_{j-1}, \tag{5.10}$$

which implies that in this case (a constant) the scheme (5.4) as well as the modified scheme (5.1) both reduce to the chosen standard scheme for gas dynamics equations (3.15) without source term effects.

The convergence of numerical approximations given by this method was established in [12, 4] for (scalar) single conservation laws in one-dimensional space, and in [5] for single conservation laws in multidimensional space.

6. Test cases. In this section we will provide some test cases to demonstrate the efficiency of our new scheme (5.4) by using MATLAB. To compare between two kinds of schemes, we take the standard Lax–Friedrichs numerical flux. We first compute solutions by using the modified Lax–Friedrichs scheme (5.1)–(5.2) and the new scheme (5.4). Then, we compare the numerical solutions with the corresponding exact solutions, which were obtained in [22] for the isentropic gases. We provide the exact Riemann solutions for the nonisentropic test cases only for special data. In that case, the exact Riemann solution can be easily computed.

The first subsection, consisting of two test cases, is devoted to computing stationary waves for both nonisentropic polytropic and isentropic ideal gases (1.6), (1.7), respectively. The second subsection consists of seven test cases of nonstationary waves for isentropic gases (1.7). The third subsection consists of test cases of nonstationary waves for nonisentropic polytropic gases (1.6). For all test cases, the exact solutions are available (see also [22]) and we will compute the error and the corresponding CPU times. The notation $U_h^{\mathrm{C}}, U_h^{\mathrm{N}}$ refer to Lax–Friedrichs solutions obtained by (5.1)–(5.2) and the new scheme (5.4), respectively.

Solutions $U(x,t)$ of the Riemann problem for system (1.1) will be computed for

$$x \in [-1,1], \quad t = 0.2.$$

We note the left- and right-hand states of the Riemann problem by U_L, U_R, respectively.

6.1. Stationary contacts. In this subsection, our new scheme will be shown to maintain equilibrium states resulted by stationary waves.

6.1.1. Test case 1. Nonisentropic polytropic ideal gases. Let us denote $U = (\rho, u, p, a)$. The Riemann initial data

$$U_L = (3.4718, -2.5923, 5.7118, 1), \quad U_R = (2, -3, 2.639, 1.5), \quad CFL = 0.5$$

can be easily verified from the relations (3.17) to correspond to a stationary contact.

The two columns on the left of Figure 6.1 shows an approximation with a visible distinction from describing a stationary contact for the modified Lax–Friedrichs scheme (5.1)–(5.2) after 156 s of CPU time with 1000 mesh-points. Meanwhile, the two columns on the right of Figure 6.1 show an immediate recovery of the stationary wave by our new scheme after 112 s of CPU time with 400 mesh-points.

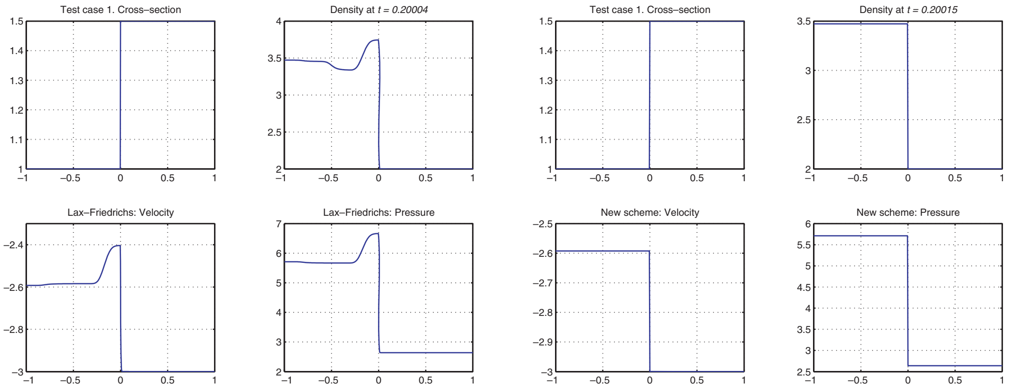


FIG. 6.1. *Test case* 1*: A stationary contact approximated by the two schemes. Left four plots: Lax–Friedrichs. Right four plots: New scheme.*

6.1.2. Test case 2. Isentropic ideal gases. We denote $U = (\rho, u, a)$ and take

$$U_L = (3.4718, -2.5923, 1), \quad U_R = (2, -3, 1.5), \quad CFL = 0.5.$$

Description. The solution is just a stationary wave from U_L to U_R; see Figure 6.2 (see [22]).

(6.1)

N	$\lVert U_h^{\mathrm{C}} - U \rVert_{L^1}$	CPU time (s)
500	0.0522	44
1000	0.0538	141
2000	0.0550	473

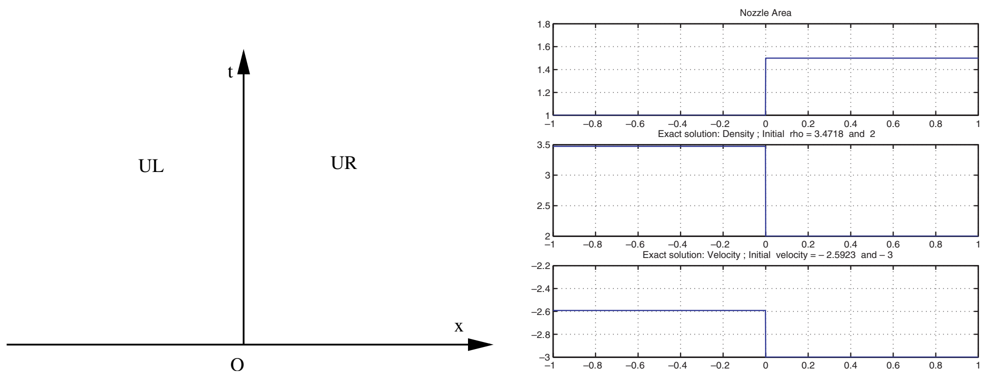


FIG. 6.2. *Test case 1: Exact Riemann solution at $t = 0.2$.*

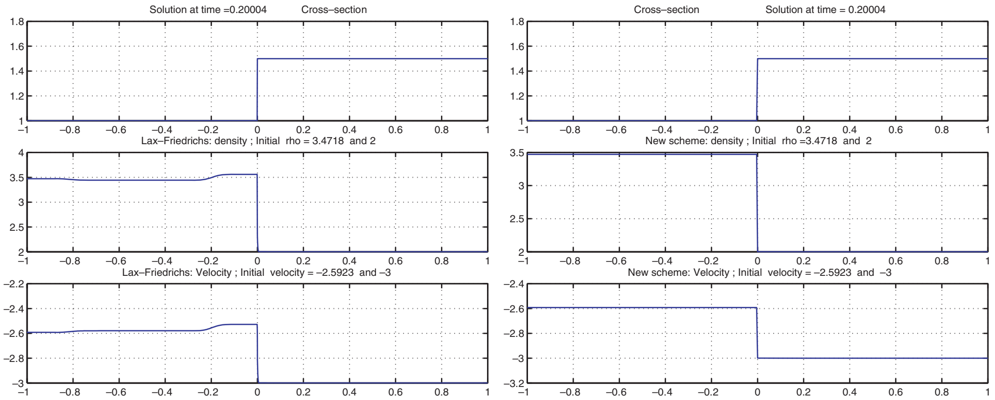


FIG. 6.3. *Test case 2: Isentropic gases: A stationary contact approximated by the two schemes.*

Looking at table (6.1), we can see that the Lax–Friedrichs scheme gives unsatisfactory results: while *decreasing the mesh-size, the errors increase*. In this case the numerical results show that our scheme can maintain the equilibrium states, i.e., the states before and after a stationary jump, while there is a visible oscillation for the modified Lax–Friedrichs scheme (see Figure 6.3).

(6.2)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
500	0.000016	91

6.2. Nonstationary waves for isentropic gases. To provide the comparison of CPU time between the two schemes, we need exact solutions of the Riemann problem which were all constructed in [22] for isentropic ideal gases, where the pressure is given by

$$p = \kappa\, \rho^{\gamma}, \qquad \kappa > 0,\, 1 < \gamma < 5/3. \tag{6.3}$$

The exact solutions of the Riemann problem will be used as reference solutions to compare the CPU times of convergence between the modified Lax–Friedrichs scheme and the new scheme.

As seen in section 3, there are three regions separated by two curves, in which the system is strictly hyperbolic, and the order of eigenvalues in each region is different

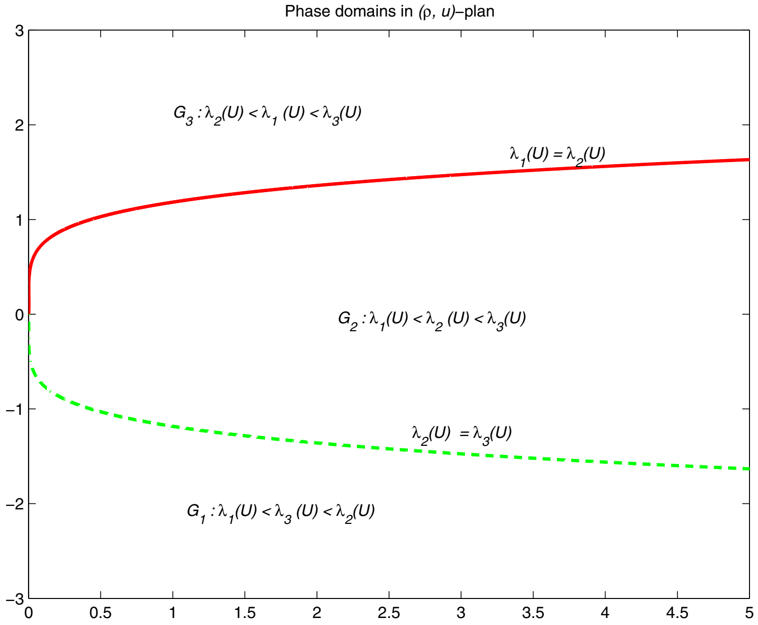


FIG. 6.4. *Hyperbolic boundaries and the three phases.*

(see Figure 6.4). Precisely, let us define

$$\begin{aligned} G_1 &:= \{(\rho,u) \ : \ \lambda_1(\rho,u) \ < \ \lambda_3(\rho,u) \ < \ \lambda_2(\rho,u)\},\\ G_2 &:= \{(\rho,u) \ : \ \lambda_1(\rho,u) \ < \ \lambda_2(\rho,u) \ < \ \lambda_3(\rho,u)\}, \qquad (6.4)\\ G_3 &:= \{(\rho,u) \ : \ \lambda_2(\rho,u) \ < \ \lambda_1(\rho,u) \ < \ \lambda_3(\rho,u)\}. \end{aligned}$$

The tests in this subsection consist of seven cases which cover all possible locations of the left- and right-hand states of the Riemann data with respect to the hyperbolic boundaries.

- Test case 3: $U_L \in G_1, \quad U_R \in G_1.$
- Test case 4: $U_L \in G_1, \quad U_R \in G_2.$
- Test case 5: $U_L \in G_2, \quad U_R \in G_1.$
- Test case 6: $U_L \in G_2, \quad U_R \in G_2.$
- Test case 7: $U_L \in G_2, \quad U_R \in G_3.$
- Test case 8: $U_L \in G_3, \quad U_R \in G_2.$
- Test case 9: $U_L \in G_3, \quad U_R \in G_3.$

6.2.1. Test case 3.

$$U_L = (0.5, -1, 1) \in G_1, \quad U_R = (2, -3, 1.5) \in G_1, \quad CFL = 0.5.$$

Description. The solution is a 1-shock from U_L to a state U_1, followed by a 3-rarefaction wave from U_1 to a state U_2, then followed by a stationary contact from U_2 to U_R. All these states belong to the same phase G_1 (see Figures 6.5 and 6.6).

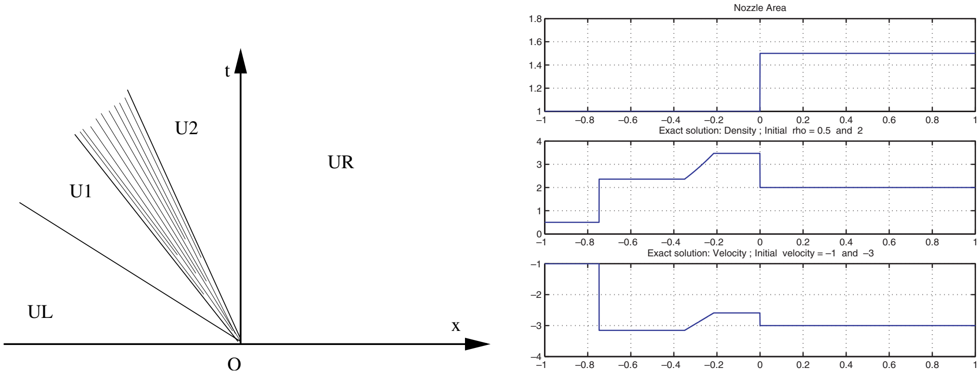


FIG. 6.5. *Test case 3: Exact Riemann solution at* $t = 0.2$.

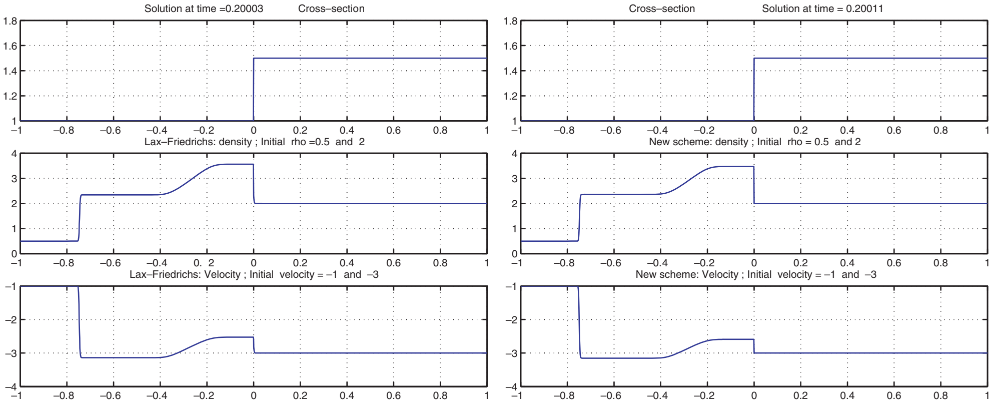


FIG. 6.6. Test case 3: *Numerical solutions by schemes with* 2000 *mesh-points.*

(6.5)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0864	138
2000	0.0684	495
4000	0.0626	2262 = 37.71 mn

(6.6)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0600	342
2000	0.0349	1403 = 23.39 mn

6.2.2. Test case 4.

$$U_L = (1, -3, 1) \in G_1, \quad U_R = (5, 0.5, 1.5) \in G_2, \quad CFL = 0.75.$$

Description. The states U_L, U_R belong to the different phases. The solution is a 1-rarefaction wave from U_L to a state U_1, followed by a 3-rarefaction wave from U_1 to a state U_2 on the hyperbolicity boundary, then jumps to a state U_3 by a stationary contact, followed by a 3-rarefaction wave from U_3 to U_R (see Figures 6.7 and 6.8).

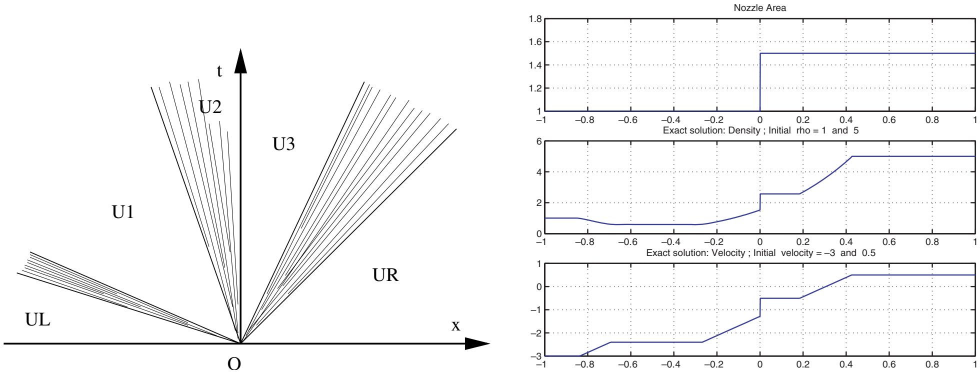


FIG. 6.7. *Test case* 4*: Exact Riemann solution at* $t = 0.2$.

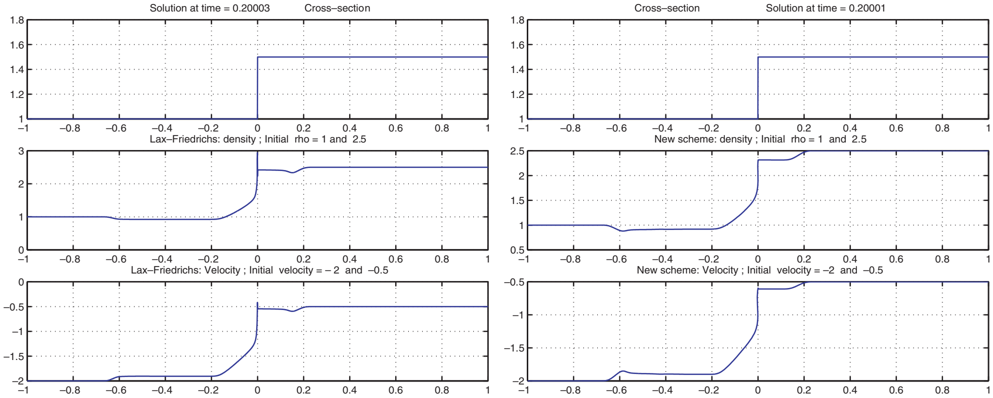


FIG. 6.8. *Test case* 4*: Numerical solutions: L-F with* 4000 *mesh-points. New scheme with* 3000 *mesh-points.*

(6.7)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
1500	0.1302	176
2000	0.1149	304
4000	0.0984	1426 = 23.7683 mn

(6.8)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
1500	0.1222	473
2000	0.0826	838.5960 = 13.9766 mn

6.2.3. Test case 5.

$$U_L = (4, -1, 1) \in G_2, \quad U_R = (2, -3, 1.5) \in G_1, \quad CFL = 0.5.$$

Description. The solution is a 1-shock from U_L to a state U_1, followed by a 3-shock from U_1 to a state U_2, then followed by a stationary contact from U_2 to U_R. All these states belong to the same phase (see Figures 6.9 and 6.10).

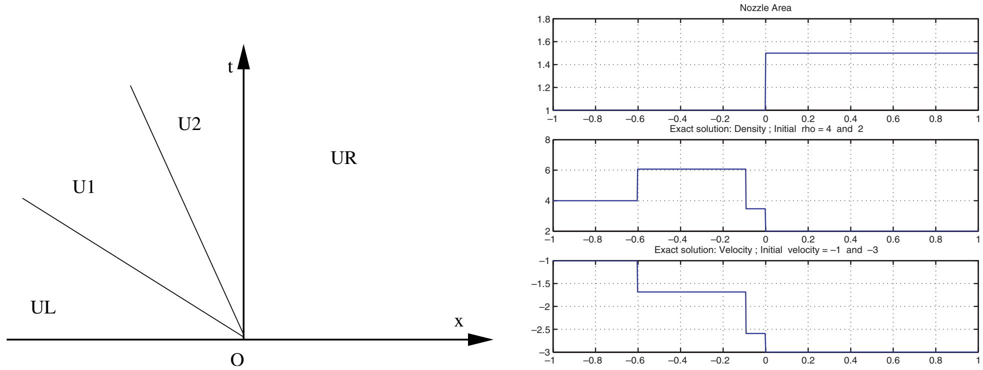


FIG. 6.9. *Test case* 5*: Exact Riemann solution at* $t = 0.2$.

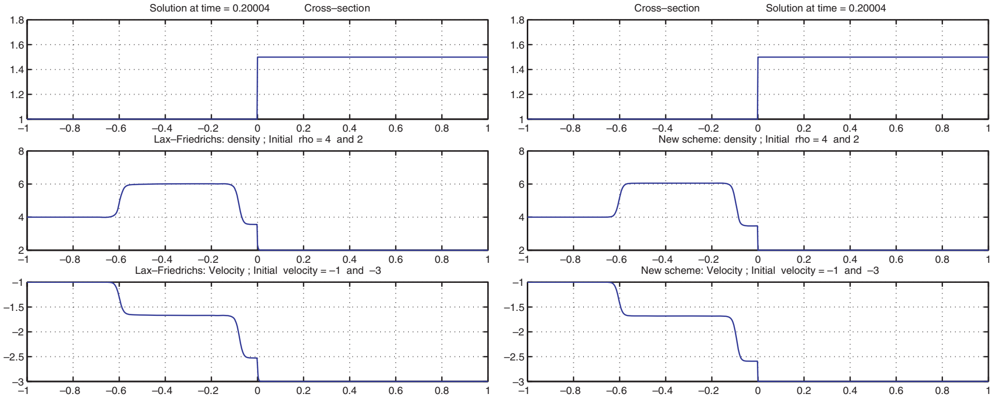


FIG. 6.10. *Test case* 5*: Numerical solutions by schemes with* 1000 *mesh-points.*

(6.9)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
1000	0.1463	127
2000	0.1051	483
4000	0.0877	$2.1128e{+}003 = 35.2133$ mn

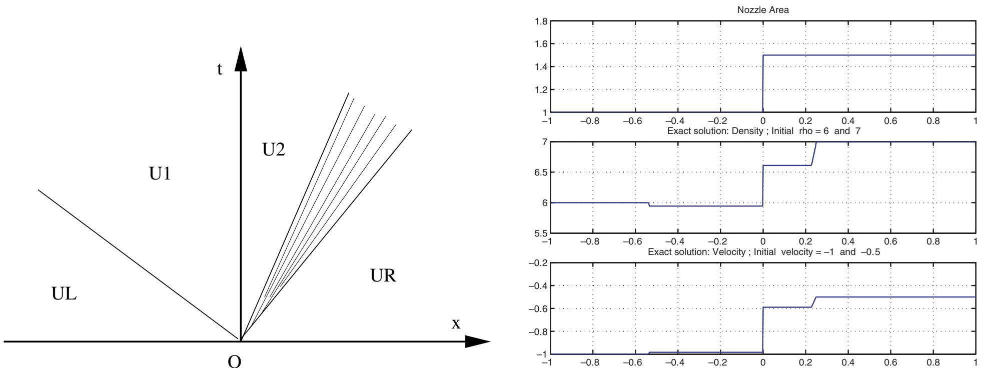
(6.10)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
700	0.1101	166
720	0.1072	167
1000	0.0801	331
2000	0.0439	$1.3422e{+}003 = 22.3700$ mn

6.2.4. Test case 6.

$$U_L = (6, -1, 1) \in G_2, \quad U_R = (7, -.5, 1.5) \in G_2, \quad CFL = 0.75.$$

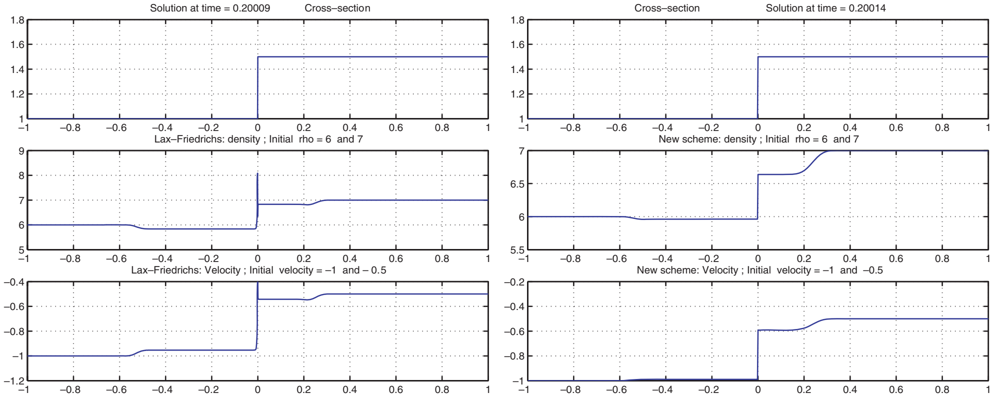
Description. The solution begins with a 1-shock from U_L to a state U_1, followed by a stationary wave from U_1 to U_2, then followed by a 3-rarefaction wave from U_2 to U_R. Both states U_L, U_R belong to the same phase (see Figures 6.11 and 6.12).


Nozzle Area

Exact solution: Density ; Initial rho = 6 and 7

Exact solution: Velocity ; Initial velocity = −1 and −0.5

FIG. 6.11. *Test case* 6*: Exact Riemann solution.*


Solution at time = 0.20009 Cross–section

Cross–section Solution at time = 0.20014

Lax–Friedrichs: density ; Initial rho = 6 and 7

New scheme: density ; Initial rho = 6 and 7

Lax–Friedrichs: Velocity ; Initial velocity = −1 and − 0.5

New scheme: Velocity ; Initial velocity = −1 and −0.5

FIG. 6.12. *Test case* 6*: Numerical solutions: Lax–Friedrichs scheme with* 2000 *mesh-points. New scheme with* 1000 *mesh-points,* $CFL = 0.5$.

(6.11)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
500	0.1378	17
1000	0.1322	53
2000	0.1297	196

(6.12)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
500	0.0309	35
1000	0.0284	130

If we take $CFL = 0.5$, then we have

(6.13)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
2000	0.1414	306

(6.14)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0302	207

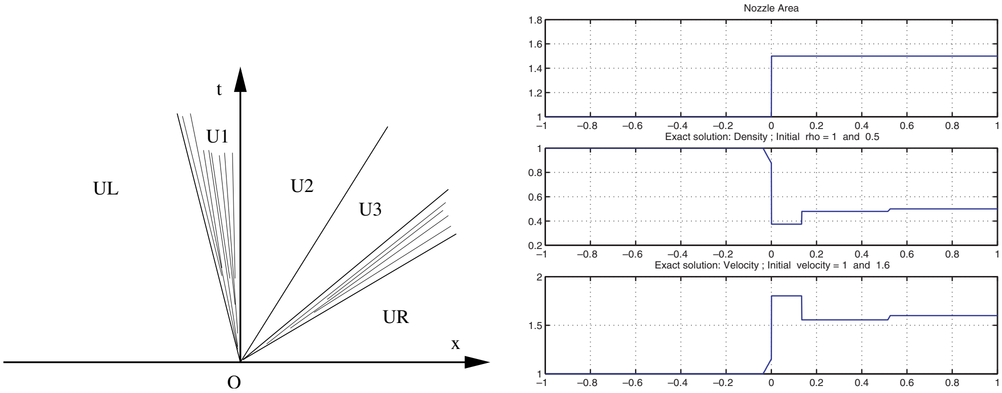


FIG. 6.13. *Test case 7: Exact Riemann solution.*

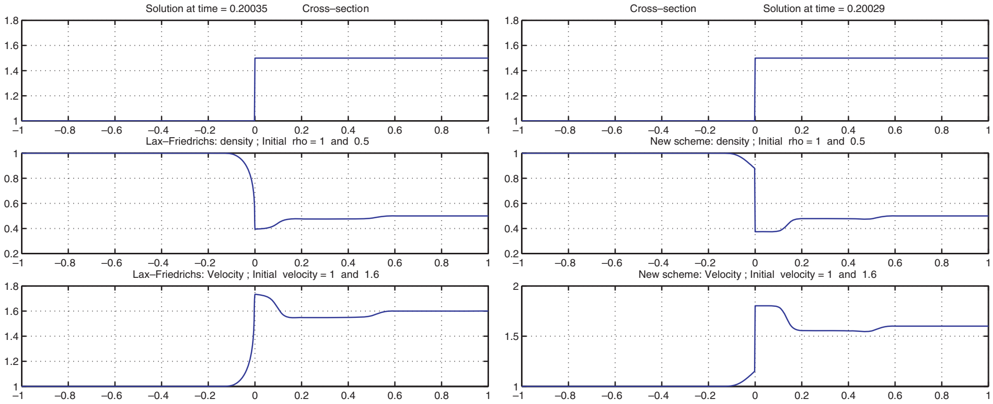


FIG. 6.14. *Test case 7: Numerical solutions with* 1000 *mesh-points,* $CFL = 0.5$.

6.2.5. Test case 7.

$$U_L = (1, 1, 1) \in G_2, \quad U_R = (0.5, 1.6, 1.5) \in G_3, \quad CFL = 0.75.$$

Description. U_L, U_R belong to different phases. The solution starts with a 1-rarefaction wave from U_L to a state U_1 belonging to the hyperbolicity boundary, then followed by a stationary contact from U_1 to a state U_2, followed by a 1-shock from U_2 to a state U_3, then followed by a 3-rarefaction wave from U_3 to U_R (see Figures 6.13 and 6.14).

(6.15)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
500	0.0528	17
1000	0.0470	78
2000	0.0348	289

(6.16)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
500	0.0246	52
1000	0.0151	196

With $CFL = 0.5$

(6.17)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0489	87
2000	0.0354	311

(6.18)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0174	209

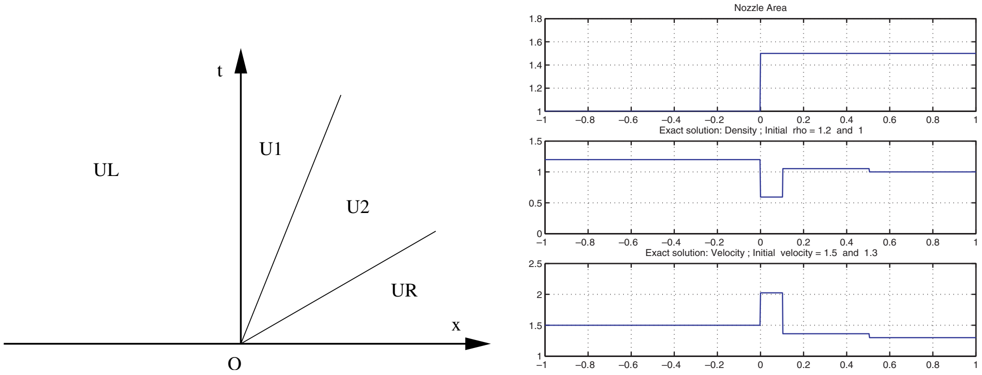


Fig. 6.15. *Test case 8: Exact Riemann solution.*

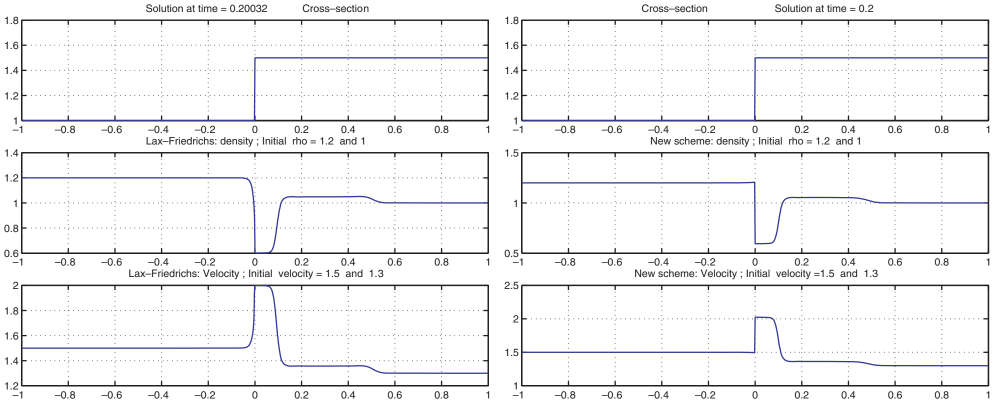


Fig. 6.16. *Test case 8: Numerical solutions with* 1000 *mesh-points.*

6.2.6. Test case 8.

$$U_L = (1, 1.5, 1) \in G_3, \quad U_R = (1, 1.2, 1.5) \in G_2, \quad CFL = 0.5.$$

Description. U_L, U_R belong to different phases. The solution starts by a stationary contact from U_L to a state U_1, followed by a 1-shock from U_1 to a state U_2, then followed by a 3-shock from U_2 to U_R (see Figures 6.15 and 6.16).

(6.19)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0295	90
2000	0.0198	335
4000	0.0153	1476 = 24.6 mn

(6.20)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0170	226
1500	0.0114	501 =8.35 mn
2000	0.0086	919 = 15.32 mn

6.2.7. Test case 9.

$$U_L = (1, 2, 1) \in G_3, \quad U_R = (0.8, 1.8, 1.5) \in G_3, \quad CFL = 0.5.$$

Description. The solution starts with a stationary contact from U_L to a state U_1, followed by a 1-shock from U_1 to s state U_2, then followed by a 3-shock wave from U_2 to U_R (see Figures 6.17 and 6.18).

(6.21)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0192	108
2000	0.0122	369

(6.22)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
1000	0.0148	249
1200	0.0127	348
1500	0.0105	556

In this case, both schemes give good results.

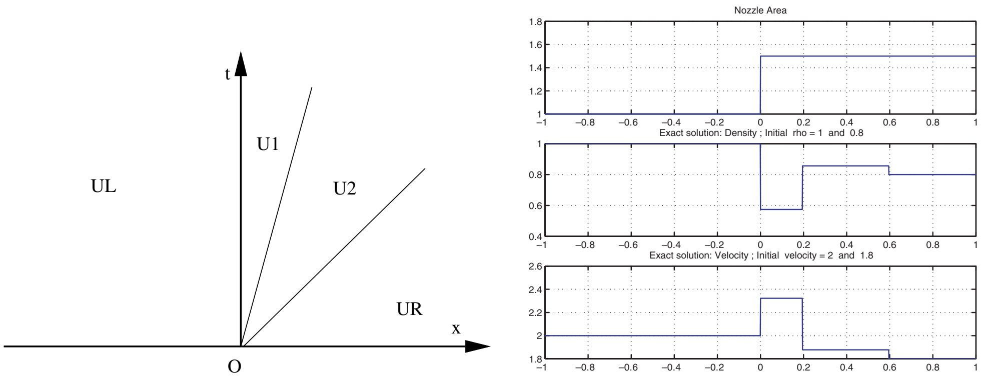


FIG. 6.17. *Test case 9: Exact Riemann solution.*

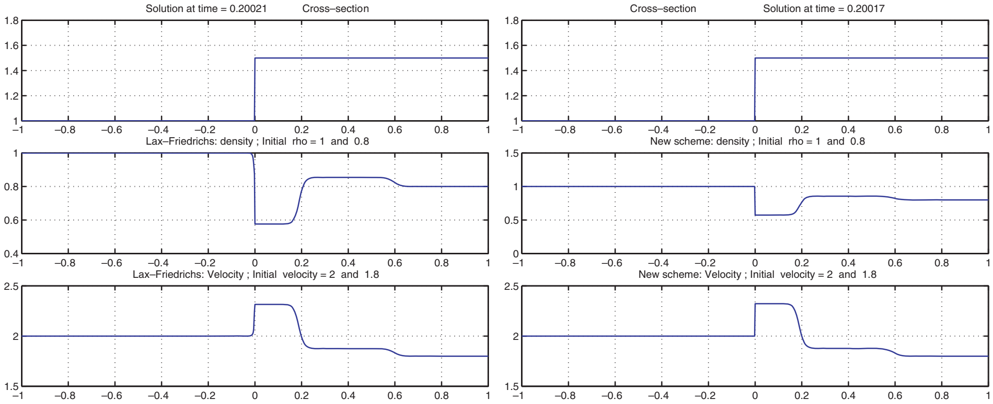


FIG. 6.18. *Test case 9: Numerical solutions with* 1000 *mesh-points.*

6.3. Nonstationary waves for nonisentropic gases. Denote in this subsection for the tests

$$U = (\rho, u, p, a).$$

It is not difficult to construct concrete Riemann solution of (1.1). In this subsection we will present two tests for nonisentropic waves in the case of nonisentropic polytropic ideal gases (1.6). In all the following figures addressed to the comparison between the two schemes, the two columns on the left are the plotting of numerical solutions from the modified Lax–Friedrichs scheme (5.1)–(5.2), and the two columns on the right are the plotting of numerical solutions from our scheme (5.4).

6.3.1. Test case 10. Riemann data

$$U_L = (4.7, 0.7452, 8.7042, 1), \quad U_R = (4.9, 0.4131, 9.2549, 1.5).$$

Set

$$\begin{aligned} U_1 &:= (4.7607, 0.7229, 8.8867, 1),\\ U_2 &:= (5.0542, 0.4539, 9.6630, 1.5),\\ U_3 &:= (5, 0.4539, 9.6630, 1.5). \end{aligned}$$

Description. The solution is a 1-shock wave from U_L to the state U_1, followed by a stationary contact from U_1 to a state U_2, followed by a contact discontinuity from U_2 to the state U_3, then followed by a 3-shock wave from U_3 to U_R (see Figures 6.19 and 6.20).

(6.23)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
1000	0.1783	86
2000	0.1676	321

(6.24)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
700	0.0968	109
1000	0.0967	210

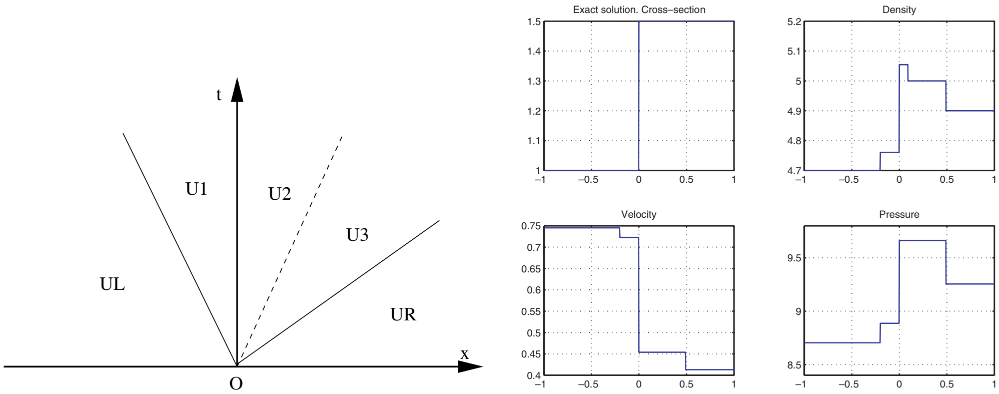


FIG. 6.19. *Test case* 10*: Exact Riemann solution.*

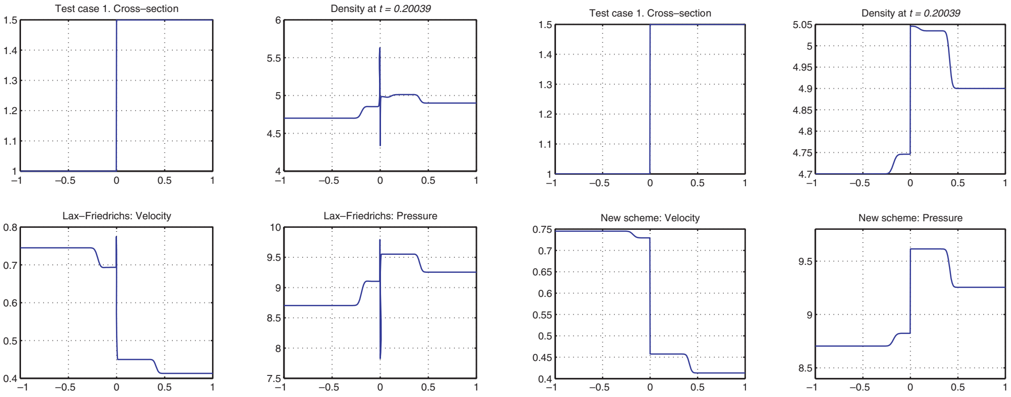


FIG. 6.20. *Test case* 10*: Numerical solutions with* 1000 *mesh-points. Left four plots: Lax–Friedrichs. Right four plots: New scheme.*

6.3.2. Test case 11. Riemann data

$$U_L = (5.5, -0.44, 11.9281, 1), \quad U_R = (6, -0.3512, 12.4326, 1.5).$$

Set

$$\begin{aligned} U_1 &:= (5.8000, -0.4945, 12.2445, 1), \\ U_2 &:= (5.9855, -0.4945, 12.2445, 1), \\ U_3 &:= (6.1340, -0.3217, 12.6719, 1.5). \end{aligned}$$

Description. The solution is a 1-shock wave from U_L to the state U_1 followed by a contact discontinuity from U_1 to the state U_2, followed by a stationary contact from U_2 to the state U_3, then followed by a 3-shock wave from U_3 to U_R (see Figures 6.21 and 6.22).

(6.25)

N	$\|\|U_h^{\mathrm{C}} - U\|\|_{L^1}$	CPU time (s)
1000	0.2799	84
2000	0.2783	293
3000	0.2783	723

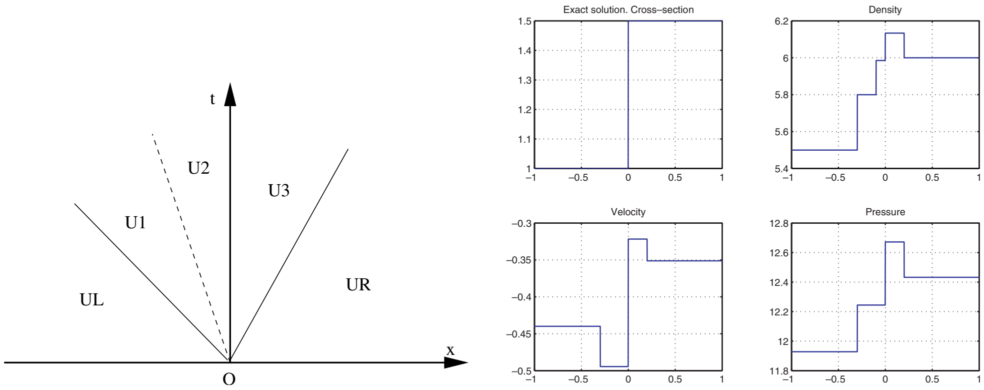


FIG. 6.21. *Test case* 11*: Exact Riemann solution.*

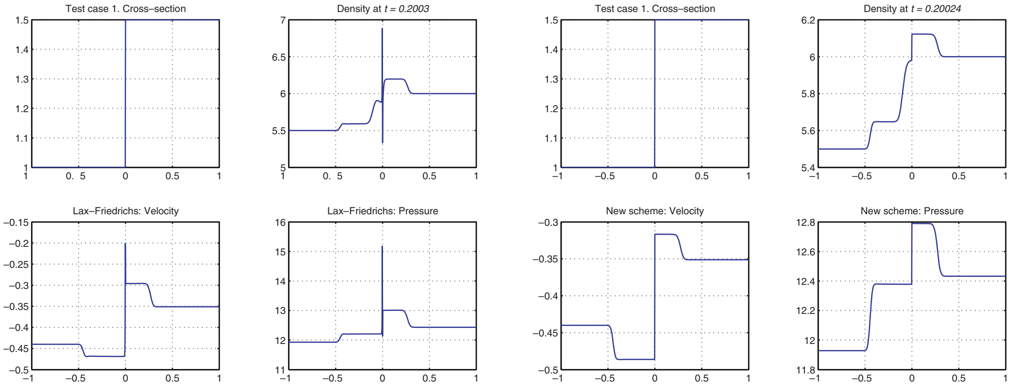


FIG. 6.22. *Test case* 11*: Numerical solutions with* 1000 *mesh-points. Left four plots: Lax–Friedrichs. Right four plots: New scheme.*

(6.26)

N	$\|\|U_h^{\mathrm{N}} - U\|\|_{L^1}$	CPU time (s)
700	0.2329	155
1000	0.2322	309

Acknowledgment. The authors would like to thank the referees for their constructive comments.

REFERENCES

[1] N. ANDRIANOV AND G. WARNECKE, *On the solution to the Riemann problem for the compressible duct flow*, SIAM J. Appl. Math., 64 (2004), pp. 878–901.

[2] E. AUDUSSE, F. BOUCHUT, M.-O. BRISTEAU, R. KLEIN, AND B. PERTHAME, *A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows,* SIAM J. Sci. Comput., 25 (2004), pp. 2050–2065.

[3] M.R. BAER AND J.W. NUNZIATO, *A two-phase mixture theory for the deflagration-to-detonation transition (DDT) in reactive granular materials,* Int. J. Multi-Phase Flow, 12 (1986), pp. 861–889.

[4] R. Botchorishvili, B. Perthame, and A. Vasseur, *Equilibrium schemes for scalar conservation laws with stiff sources,* Math. Comp., 72 (2003), pp. 131–157.

[5] R. Botchorishvili and O. Pironneau, *Finite volume schemes with equilibrium type discretization of source terms for scalar conservation laws,* J. Comput. Phys., 187 (2003), pp. 391–427.

[6] F. Boutchut, *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws, and Well-Balanced Schemes for Sources,* Front. Math., Birkhäuser, Basel, 2004.

[7] T.N. Dinh, R.R. Nourgaliev, and T.G. Theofanous, *Understanding the ill-posed twofluid model,* in The 10th International Topical Meeting on Nuclear Reactor Thermal Hydraulics (NERETH-10), 2003.

[8] D.A. Drew and S.L. Passman, *Theory of Multicomponent Fluids,* Springer-Verlag, New York, 1999.

[9] P. Goatin and P.G. LeFloch, *The Riemann problem for a class of resonant nonlinear systems of balance laws,* Ann. Inst. H. Poincaré Anal. Non Linéaire, 21 (2004), pp. 881–902.

[10] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws,* Springer-Verlag, New York, 1996.

[11] L. Gosse, *A well-balanced flux-vector splitting scheme designed for hyperbolic systems of conservation laws with source terms,* Comput. Math. Appl., 39 (2000), pp. 135–159.

[12] J.M. Greenberg and A.Y. Leroux, *A well-balanced scheme for the numerical processing of source terms in hyperbolic equations,* SIAM J. Numer. Anal., 33 (1996), pp. 1–16.

[13] J.M. Greenberg, A.Y. Leroux, R. Baraille, and A. Noussair, *Analysis and approximation of conservation laws with source terms,* SIAM J. Numer. Anal., 34 (1997), pp. 1980–2007.

[14] E. Isaacson and B. Temple, *Nonlinear resonance in systems of conservation laws,* SIAM J. Appl. Math., 52 (1992), pp. 1260–1278.

[15] E. Isaacson and B. Temple, *Convergence of the* 2×2 *Godunov method for a general resonant nonlinear balance law,* SIAM J. Appl. Math., 55 (1995), pp. 625–640.

[16] S. Jin and X. Wen, *An efficient method for computing hyperbolic systems with geometrical source terms having concentrations,* J. Comput. Math., 22 (2004), pp. 230–249.

[17] B.L. Keyfitz, R. Sander, and M. Sever, *Lack of hyperbolicity in the two-fluid model for two-phase incompressible flow,* Discrete Contin. Dyn. Syst. Ser. B, 3 (2003), pp. 541–563.

[18] D. Kröner, *Numerical Schemes for Conservation Laws,* John Wiley & Sons, Stuttgart, 1997.

[19] D. Kröner and M.D. Thanh, *A Well-Balanced Scheme for the Baer–Nunziato Model of Two-Phase Fluids,* in preparation.

[20] D. Kröner, P.G. LeFloch, and M.D. Thanh, *The Model of Fluid Flows in a Nozzle with Variable Cross-Section: Stability, Numerics and Entropy*, in preparation.

[21] P.G. LeFloch, *Shock Waves for Nonlinear Hyperbolic Systems in Nonconservative Form,* Preprint 593, Inst. Math. Appl., Minneapolis, 1989.

[22] P.G. LeFloch and M.D. Thanh, *The Riemann problem for fluid flows in a nozzle with discontinuous cross-section,* Commun. Math. Sci., 1 (2003), pp. 763–797.

[23] R.J. LeVeque, *Balancing source terms and flux gradients in high-resolution Godunov methods: The quasi-steady wave-propagation algorithm,* J. Comput. Phys., 146 (1998), pp. 346–365.

[24] D. Marchesin and P.J. Paes-Leme, *A Riemann problem in gas dynamics with bifurcation. Hyperbolic partial differential equations* III, Comput. Math. Appl. (Part A), 12 (1986), pp. 433–455.

[25] R. Menikoff and B. Plohr, *The riemann problem for fluid flow of real materials,* Rev. Modern Phys., 61 (1989), pp. 75–130.

[26] M.J. Zucrow and J.D. Hoffman, *Gas Dynamics,* John Wiley & Sons, New York, 1977.

SIAM J. Numer. Anal.
Vol. 43, No. 2, pp. 825–859

AN ERROR ANALYSIS OF CONSERVATIVE SPACE-TIME MESH REFINEMENT METHODS FOR THE ONE-DIMENSIONAL WAVE EQUATION[*]

PATRICK JOLY[†] AND JERÓNIMO RODRÍGUEZ[†]

Abstract. We study two space-time mesh refinement methods as the one introduced in [F. Collino, T. Fouquet, and P. Joly, *Numer. Math.*, 95 (2003), pp. 197–221]. The stability of such methods is guaranteed by construction through the conservation of a discrete energy. In this paper, we show the L^2 convergence of these schemes and provide optimal error estimates. The proof is based on energy techniques and bootstrap arguments. Our results are validated with numerical simulations and compared with results from plane wave analysis [F. Collino, T. Fouquet, and P. Joly, *Numer. Math.*, 95 (2003), pp. 223–251].

Key words. mesh refinement, local time stepping, error estimates, stability, energy conservation, wave equation

AMS subject classifications. 65M06, 65M12, 65M55, 35L05

DOI. 10.1137/040603437

1. Introduction.

For the numerical solution of time-dependent wave propagation problems, in which one often has to deal with complex geometries in diffraction problems, it is natural to try to use local mesh refinements with nonmatching grids. One initial idea consists in using only spatial refinement (see [1] for acoustic waves, [8] and [22] for Maxwell's equations). However, with explicit schemes, when a uniform time step is used, it is the finest mesh that imposes the time step because of the stability condition. There are two problems with this. First, the computational cost is increased. Second, the ratio $c\Delta t/h$ (where h is the space step size) in the coarser grid will be much smaller than its optimal value. With standard numerical schemes (such as Yee's scheme for Maxwell's equations) this generates dispersion errors. To avoid these problems, it is useful to be able to work with a local time step in order to keep the ratio $c\Delta t/h$ constant (or almost constant) in the whole computational domain.

The use of local time stepping raises new practical and theoretical problems, especially for hyperbolic equations, that are much more delicate than those raised by a simple spatial refinement.

The solutions suggested in the electromagnetic literature are primarily based on interpolation techniques (in time and/or in space) especially designed to guarantee the consistency of the scheme at the coarse grid/fine grid interface (see [20], [18], [24], [9]). Unfortunately, the resulting schemes appear to be very difficult to analyze and may suffer from some instability phenomena [11]. Another possible solution for local time stepping is to use a domain decomposition approach such as that recently developed in [14]. However, the stability and convergence analysis of these techniques remain to be completed.

[*] Received by the editors January 21, 2004; accepted for publication (in revised form) January 18, 2005; published electronically September 15, 2005. This work was supported by a research contract between EDF/DER and INRIA.

http://www.siam.org/journals/sinum/43-2/60343.html

[†] INRIA Rocquencourt, BP105 Le Chesnay, France (Patrick.Joly@inria.fr, Jeronimo.Rodriguez@inria.fr).

It seems that very few papers in the mathematical literature have been devoted to space-time mesh refinement for the specific case of Maxwell's equations (and more generally for linear wave propagation problems). However, these questions have been treated in many articles in the conservation laws community during the 1980s. Let us mention, for instance, the work of Osher and Sanders [23] based on finite volume methods or, closer to what we are doing here, the works of Berger and her coauthors [3, 4, 7, 5] on finite differences schemes. These works are devoted to various space-time mesh refinement techniques for first order hyperbolic systems. These techniques are mainly based on interpolation-type procedures and concern both overlapping and nonoverlapping grids (see [7, 5] for a general presentation). In [3], Berger has developed a stability analysis of such methods in the cases of the one-dimensional (1D) linear advection equations using the GKS theory [19, 15]. She was able to establish the results in the case where dissipative interior schemes (typically Lax–Wendroff scheme) were used or when conservative (typically leap-frog) schemes are used provided that overlapping grids are considered. However, it is also mentioned in [3] that using leap-frog-type schemes and nonoverlapping grids may lead to instability, as has already been mentioned for Maxwell's equations.

Recently, we developed alternative solutions to these interpolation procedures that we call conservative space-time mesh refinement methods. These methods, originally invented for the 1D wave equation, have been developed for Maxwell's equations [13] and recently extended to the elastodynamic equations [2]. A general presentation of these kinds of methods can be found in [16]. The main ideas and properties of these methods have been treated in more detail in [10] for the model problem of the 1D wave equation written as the first order system

$$(1.1)\qquad \begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial v}{\partial x} = 0, & \dfrac{\partial v}{\partial t} + \dfrac{\partial u}{\partial x} = 0, \qquad x \in \mathbb{R}, \qquad t > 0, \\ u(x,0) = u_0(x), & v(x,0) = v_0(x), \end{cases}$$

when one uses the FDTD Yee [26] scheme as the reference interior scheme in each subdomain

$$(1.2)\qquad \begin{cases} \dfrac{u_j^{n+1} - u_j^n}{\Delta t} + \dfrac{v_{j+\frac{1}{2}}^{n+\frac{1}{2}} - v_{j-\frac{1}{2}}^{n+\frac{1}{2}}}{h} = 0, \quad j \in \mathbb{Z}, \quad n \geq 0, \\ \dfrac{v_{j+\frac{1}{2}}^{n+\frac{1}{2}} - v_{j+\frac{1}{2}}^{n-\frac{1}{2}}}{\Delta t} + \dfrac{u_{j+1}^n - u_j^n}{h} = 0, \quad j \in \mathbb{Z}, \quad n \geq 0. \end{cases}$$

In particular the construction of the scheme and stability analysis based on energy conservation properties is presented in [10]. An important fact is that the stability CFL condition, namely, in the 1D case,

$$(1.3)\qquad \alpha = \frac{\Delta t}{h} < 1,$$

is not affected by the mesh refinement process. A plane wave analysis for measuring the accuracy of the method is detailed in [12]. The present article, whose purpose is essentially theoretical, is the sequel to [12]. Our goal is to derive optimal error estimates and to validate them through numerical tests. More precisely, we present a convergence analysis for two different conservative space-time mesh refinement schemes introduced in [11].

We mention that in [4] and more recently in [6] the authors have constructed space-time mesh refinement schemes devoted to the conservation of a discrete equivalent of the integral of the solution of a first order hyperbolic system. Such schemes are also called conservative schemes in the conservation laws community, but they are not necessarily stable (the integral of the solution is not a norm!). However, such conservation properties are highly desirable for the approximation of solutions with shocks.

The outline of the rest of the article is as follows. In section 2, we define our two grids model problem, present the two mesh refinement schemes (I and II) and recall the main stability theorems. In section 3, we state our main convergence Theorem 3.1. Section 4, devoted to the proof of this theorem, is the main section of this article. We think that one of the contributions of the present paper is precisely the proof that appears rather nonstandard, although based on energy techniques. Finally, in section 5, the theoretical results are compared to numerical ones and those obtained in [12] by Fourier-like techniques. This analysis did not result in rigorous error estimates, as did that of Theorem 3.1, but permitted us to predict the order of convergence that we prove in the present paper. The other interest of the energy proof we develop here is that it can be generalized to spatially variable coefficients and higher space dimensions with only purely technical additional difficulties.

2. Conservative space-time mesh refinement schemes.

We recall the construction of the method presented in [11]. In order to solve system (1.1) with a local space-time mesh refinement, the computational domain is split into two half-spaces, $\Omega_c = \{x < 0\}$ and $\Omega_f = \{x > 0\}$. Denoting by (u_c, v_c) and (u_f, v_f) the restrictions of (u, v) to Ω_c and Ω_f respectively, problem (1.1) can be rewritten as a transmission problem through the interface $x = 0$ as follows:

$$\left\{\begin{aligned} &\frac{\partial u_c}{\partial t} + \frac{\partial v_c}{\partial x} = 0,\\ &\frac{\partial v_c}{\partial t} + \frac{\partial u_c}{\partial x} = 0, \end{aligned}\right. \quad \text{in } \Omega_c, \tag{2.1}$$

$$\left\{\begin{aligned} &\frac{\partial u_f}{\partial t} + \frac{\partial v_f}{\partial x} = 0,\\ &\frac{\partial v_f}{\partial t} + \frac{\partial u_f}{\partial x} = 0, \end{aligned}\right. \quad \text{in } \Omega_f, \tag{2.2}$$

coupled by the interface conditions

$$u_c(0,t) = u_f(0,t), \tag{2.3}$$

$$v_c(0,t) = v_f(0,t), \tag{2.4}$$

to obtain a solution of the global problem.

2.1. The interior scheme.

Assume that we have a mesh with step size $(2h, 2\Delta t)$ for Ω_c and a mesh with step size $(h, \Delta t)$ for Ω_f. It is important to note that the ratio of the time step to the space step is the same in both domains. With the obvious notation, the unknowns of our scheme will be the following:

- for the coarse grid,

$$u_{2j}^{2n}, \quad j \le 0, \quad n \ge 0, \qquad v_{2j+1}^{2n+1}, \quad j \le -1, \quad n \ge 0;$$

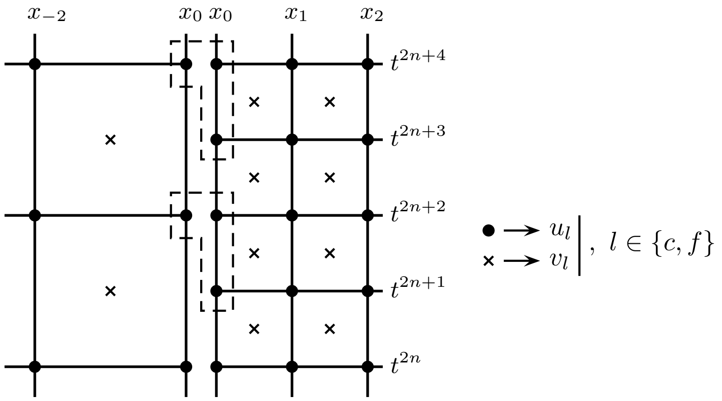


FIG. 2.1. *Distribution of the unknowns over* Γ.

- for the fine grid,

$$u_j^n, \qquad j \geq 0, \qquad n \geq 0, \qquad\qquad v_{j+\frac{1}{2}}^{n+\frac{1}{2}}, \qquad j \geq 0, \qquad n \geq 0.$$

At the interior of each subdomain, the standard Yee scheme [26, 25] is considered. The discrete equations in the coarse and in the fine grids between the instants t^{2n} and t^{2n+2} are

$$\text{(2.5)} \quad \begin{cases} \dfrac{(u_c)_{2j}^{2n+2} - (u_c)_{2j}^{2n}}{2\Delta t} + \dfrac{(v_c)_{2j+1}^{2n+1} - (v_c)_{2j-1}^{2n+1}}{2h} = 0, \quad j \leq -1, \quad n \geq 0, \\ \dfrac{(v_c)_{2j+1}^{2n+1} - (v_c)_{2j+1}^{2n-1}}{2\Delta t} + \dfrac{(u_c)_{2j+2}^{2n} - (u_c)_{2j}^{2n}}{2h} = 0, \quad j \leq -1, \quad n \geq 0, \end{cases}$$

$$\text{(2.6)} \quad \begin{cases} \dfrac{(u_f)_j^{2n+1} - (u_f)_j^{2n}}{\Delta t} + \dfrac{(v_f)_{j+\frac{1}{2}}^{2n+\frac{1}{2}} - (v_f)_{j-\frac{1}{2}}^{2n+\frac{1}{2}}}{h} = 0, \quad j \geq 1, \quad n \geq 0, \\ \dfrac{(v_f)_{j+\frac{1}{2}}^{2n+\frac{1}{2}} - (v_f)_{j+\frac{1}{2}}^{2n-\frac{1}{2}}}{\Delta t} + \dfrac{(u_f)_{j+1}^{2n} - (u_f)_j^{2n}}{h} = 0, \quad j \geq 0, \quad n \geq 0, \end{cases} \\ \begin{cases} \dfrac{(u_f)_j^{2n+2} - (u_f)_j^{2n+1}}{\Delta t} + \dfrac{(v_f)_{j+\frac{1}{2}}^{2n+\frac{3}{2}} - (v_f)_{j-\frac{1}{2}}^{2n+\frac{3}{2}}}{h} = 0, \quad j \geq 1, \quad n \geq 0, \\ \dfrac{(v_f)_{j+\frac{1}{2}}^{2n+\frac{3}{2}} - (v_f)_{j+\frac{1}{2}}^{2n+\frac{1}{2}}}{\Delta t} + \dfrac{(u_f)_{j+1}^{2n+1} - (u_f)_j^{2n+1}}{h} = 0, \quad j \geq 0, \quad n \geq 0, \end{cases}$$

completed with discrete initial conditions.

$$\text{(2.7)} \quad \begin{array}{llll} (u_c)_{2j}^0, & j \leq 0, & (v_c)_{2j+1}^1, & j \leq -1\,, \\ (u_f)_j^0, & j \geq 0, & (v_f)_{j+\frac{1}{2}}^{\frac{1}{2}}, & j \geq 0\,. \end{array}$$

As is shown in Figure 2.1, two values of the solution are allowed along the interface $\Gamma = \{x = 0\}$ at the even time steps. The continuity of the unknown u is imposed in a weak way: this seems to be useful for guaranteeing the stability of the scheme [11].

2.2. The discrete transmission conditions. For coupling (2.5) and (2.6), the idea is to approximate the transmission conditions (2.3) and (2.4) in such a way that the stability of the method is ensured a priori. A simple way to do that is to impose a discrete version of the energy conservation property

$$E(t) = E(0), \qquad \text{where} \quad E(t) = \frac{1}{2}\int_{\mathbb{R}} u(x,t)^2 \mathrm{d}x + \frac{1}{2}\int_{\mathbb{R}} v(x,t)^2 \mathrm{d}x,$$

satisfied by the exact solution of (1.1). In our case, it is natural to define the total discrete energy only at the even instants by

$$E^{2n} = E_c^{2n} + E_f^{2n}, \tag{2.8}$$

where E_c^{2n} and E_f^{2n} are, respectively, the coarse grid and fine grid energies:

$$E_c^{2n} = \frac{1}{2}\left(\sum_{j\leq -1} |(u_c)_{2j}^{2n}|^2\; 2h + \sum_{j\leq -1} (v_c)_{2j+1}^{2n+1}(v_c)_{2j+1}^{2n-1}\; 2h + |(u_c)_0^{2n}|^2\; h\right),$$

$$E_f^{n} = \frac{1}{2}\left(\sum_{j\geq 1} |(u_f)_{j}^{n}|^2\; h + \sum_{j\geq 0} (v_f)_{j+\frac{1}{2}}^{n+\frac{1}{2}}(v_f)_{j+\frac{1}{2}}^{n-\frac{1}{2}}\; h + |(u_f)_0^{n}|^2\; \frac{h}{2}\right).$$

This is the most natural extension to the "two grids" scheme of the discrete energy which is conserved with the Yee scheme on a single grid. The idea pursued in [11] is to impose the conservation of E^{2n}. To state the main result of [11], it is useful to introduce the following "discrete traces" of u_c, v_c, u_f, and v_f:

$$\begin{cases} (U_c)_0^{2n+1} &= \dfrac{1}{2}\left((u_c)_0^{2n+2} + (u_c)_0^{2n}\right), \\[2ex] (V_c)_0^{2n+1} &= (v_c)_{-1}^{2n+1} - h\dfrac{(u_c)_0^{2n+2} - (u_c)_0^{2n}}{2\Delta t}, \end{cases} \tag{2.9}$$

$$\begin{cases} (U_f)_0^{n+\frac{1}{2}} &= \dfrac{1}{2}\left((u_f)_0^{n+1} + (u_f)_0^{n}\right), \\[2ex] (V_f)_0^{n+\frac{1}{2}} &= (v_f)_{\frac{1}{2}}^{n+\frac{1}{2}} + \dfrac{h}{2}\dfrac{(u_f)_0^{n+1} - (u_f)_0^{n}}{\Delta t}. \end{cases} \tag{2.10}$$

THEOREM 2.1. *Consider a solution of* (2.5) *and* (2.6), *the discrete energy* (2.8) *is conserved if and only if*

$$\frac{1}{2}\left((U_f)_0^{2n+\frac{1}{2}}(V_f)_0^{2n+\frac{1}{2}} + (U_f)_0^{2n+\frac{3}{2}}(V_f)_0^{2n+\frac{3}{2}}\right) = (U_c)_0^{2n+1}(V_c)_0^{2n+1}. \tag{2.11}$$

Let us come back to the approximation of the continuity conditions (2.3) and (2.4). We first remark that, assuming the discrete unknowns have been computed up to time t^{2n}, the interior scheme given by (2.5) and (2.6) permits us to obtain all the unknowns up to time t^{2n+2} except the three following values:

$$(u_c)_0^{2n+2}, \qquad (u_f)_0^{2n+1}, \qquad (u_f)_0^{2n+2}.$$

So, three additional (linear) equations consistent with the transmission conditions (2.3) and (2.4) and compatible with (2.11) should be added. A first natural choice consists in imposing the following discrete continuity conditions:

$$(2.12)\qquad \begin{cases} (V_f)_0^{2n+\frac{1}{2}} = (V_f)_0^{2n+\frac{3}{2}} = (V_c)_0^{2n+1}, \\ \dfrac{1}{2}\left((U_f)_0^{2n+\frac{1}{2}} + (U_f)_0^{2n+\frac{3}{2}}\right) = (U_c)_0^{2n+1}. \end{cases}$$

The scheme given by (2.5), (2.6), and (2.12) will be called scheme I. The continuity of u is imposed once and that of v twice. A second possible choice is given by

$$(2.13)\qquad \begin{cases} (U_f)_0^{2n+\frac{1}{2}} = (U_f)_0^{2n+\frac{3}{2}} = (U_c)_0^{2n+1}, \\ \dfrac{1}{2}\left((V_f)_0^{2n+\frac{1}{2}} + (V_f)_0^{2n+\frac{3}{2}}\right) = (V_c)_0^{2n+1}, \end{cases}$$

which gives us scheme II. Unlike in the previous scheme, the continuity of u is written twice and that of v once.

Remark 2.1. Using the two first equations of (2.13) we can easily prove that $(u_f)_0^{2n+2} = (u_f)_0^{2n}$ and so the discrete trace of u in the fine side at the even instants is always the same. As a consequence, this scheme cannot be L^∞-convergent. However the numerical simulations will show us that scheme II gives a "good" approximation of the solution "except at the interface," and the L^2-convergence will be proven in section 4.

3. Error analysis: The main results.

Notation. We first introduce some notation for discrete sequences. In order to show the L^2-stability and convergence of both schemes, some discrete norms and spaces must be introduced. Let us define the discrete coarse and fine L^2 spaces for u,

$$(3.1)\qquad \begin{aligned} L^2_{c,u} &= \Big\{ u_{c,h} = \{(u_c)_{2j}\}_{j\le 0} \text{ such that } \sum_{j\le 0} |(u_c)_{2j}|^2 < +\infty \Big\}, \\ L^2_{f,u} &= \Big\{ u_{f,h} = \{(u_f)_{j}\}_{j\ge 0} \text{ such that } \sum_{j\ge 0} |(u_f)_{j}|^2 < +\infty \Big\}, \end{aligned}$$

and their natural Hilbert norms,

$$(3.2)\qquad \begin{aligned} \|u_{c,h}\|^2 &= \sum_{j\le -1} |(u_c)_{2j}|^2 2h + |(u_c)_0|^2\, h, \\ \|u_{f,h}\|^2 &= \sum_{j\ge 1} |(u_f)_{j}|^2 h + |(u_f)_0|^2\, \frac{h}{2}. \end{aligned}$$

In the same way, we have the discrete coarse and fine L^2 spaces for v,

$$(3.3)\qquad \begin{aligned} L^2_{c,v} &= \Big\{ v_{c,h} = \{(v_c)_{2j+1}\}_{j\le -1} \text{ such that } \sum_{j\le -1} |(v_c)_{2j+1}|^2 < +\infty \Big\}, \\ L^2_{f,v} &= \Big\{ v_{f,h} = \{(v_f)_{j+\frac{1}{2}}\}_{j\ge 0} \text{ such that } \sum_{j\ge 0} |(v_f)_{j+\frac{1}{2}}|^2 < +\infty \Big\}, \end{aligned}$$

and the norms,

$$\|v_{c,h}\|^2 = \sum_{j\leq -1} |(v_c)_{2j+1}|^2 \; 2h, \qquad \|v_{f,h}\|^2 = \sum_{j\geq 1} |(v_f)_{j+\frac{1}{2}}|^2 \; h. \tag{3.4}$$

It is immediate to check that our schemes are well posed in these spaces; i.e., as soon as the discrete initial data $u^0_{c,h}, v^1_{c,h}, u^0_{f,h}$, and $v^{\frac{1}{2}}_{f,h}$ belong, respectively, to $L^2_{c,u}$, $L^2_{c,v}$, $L^2_{f,u}$, and $L^2_{f,v}$, then the discrete solution is such that

$$\left(u^{2n}_{c,h}, v^{2n+1}_{c,h}, u^n_{f,h}, v^{n+\frac{1}{2}}_{f,h}\right) \in L^2_{c,u} \times L^2_{c,v} \times L^2_{f,u} \times L^2_{f,v}.$$

For the convergence study, we assume that our solution is at least continuous and we introduce the pointwise exact values

$$\tilde{u}^s_r = u(rh, s\Delta t), \quad \tilde{v}^s_r = v(rh, s\Delta t), \quad (r,s) \in \mathbb{R} \times \mathbb{R}^+.$$

We will assume that the corresponding sequences $\tilde{u}^{2n}_{c,h}, \tilde{v}^{2n+1}_{c,h}, \tilde{u}^n_{f,h}$, and $\tilde{v}^{n+\frac{1}{2}}_{c,h}$ belong, respectively, to $L^2_{c,u}$, $L^2_{c,v}$, $L^2_{f,u}$, and $L^2_{f,v}$. We then define the pointwise errors:

$$\left| \begin{array}{ll} \left(e^u_{c,h}\right)^{2n} = \tilde{u}^{2n}_{c,h} - u^{2n}_{c,h}, & \left(e^u_{f,h}\right)^n = \tilde{u}^n_{f,h} - u^n_{f,h}, \\[2ex] \left(e^v_{c,h}\right)^{2n+1} = \tilde{v}^{2n}_{c,h} - v^{2n}_{c,h}, & \left(e^v_{f,h}\right)^{n+\frac{1}{2}} = \tilde{v}^{n+\frac{1}{2}}_{f,h} - v^{n+\frac{1}{2}}_{f,h}. \end{array} \right.$$

We shall denote by a *superscript h* sequences in both discrete space and time. More precisely, we set

$$\begin{cases} u^h_c = \left(u^{2n}_{c,h}\right)_{n\geq 0}, & u^h_f = \left(u^n_{f,h}\right)_{n\geq 0}, \qquad u^h = \left(u^h_c, u^h_f\right), \\[2ex] v^h_c = \left(v^{2n+1}_{c,h}\right)_{n\geq 0}, & v^h_f = \left(v^{n+\frac{1}{2}}_{f,h}\right)_{n\geq 0}, \qquad v^h = \left(v^h_c, v^h_f\right), \end{cases}$$

for the discrete solutions and, in the same way,

$$\begin{cases} e^{u,h}_c = \left(\left(e^u_{c,h}\right)^{2n}\right)_{n\geq 0}, & e^{u,h}_f = \left(\left(e^u_{f,h}\right)^n\right)_{n\geq 0}, \quad e^{u,h} \; (\equiv u - u^h) = \left(e^{u,h}_c, e^{u,h}_f\right), \\[2ex] e^{v,h}_c = \left(\left(e^v_{c,h}\right)^{2n+1}\right)_{n\geq 0}, & e^{v,h}_f = \left(\left(e^v_{f,h}\right)^{n+\frac{1}{2}}\right)_{n\geq 0}, \quad e^{v,h} \; (\equiv v - v^h) = \left(e^{v,h}_c, e^{v,h}_f\right), \end{cases}$$

for the errors.

For a given $T > 0$, we can introduce the discrete $L^\infty(0,T;L^2)$ norms (that we define here for the errors $e^{u,h} = u - u^h$ and $e^{v,h} = v - v^h$)

$$\begin{cases} \|e^{u,h}\|^*_{\infty,2,T} = \displaystyle\sup_{t^{2n+\frac{3}{2}}\leq T} \left(\, \|\left(e^u_{c,h}\right)^{2n}\| + \|\left(e^u_{f,h}\right)^{2n}\| \, \right), \\[3ex] \|e^{u,h}\|_{\infty,2,T} = \|e^{u,h}\|^*_{\infty,2,T} + \displaystyle\sup_{t^{2n+\frac{3}{2}}\leq T} \|\left(e^u_{f,h}\right)^{2n+1}\|, \\[3ex] \|e^{v,h}\|_{\infty,2,T} = \displaystyle\sup_{t^{2n+\frac{3}{2}}\leq T} \left(\, \|\left(e^v_{c,h}\right)^{2n+1}\| + \|\left(e^v_{f,h}\right)^{2n+\frac{1}{2}}\| + \|\left(e^v_{f,h}\right)^{2n+\frac{3}{2}}\| \, \right). \end{cases} \tag{3.5}$$

Remark 3.1. Note that $\|u^h\|^*_{\infty,2,T}$ is only a seminorm since the odd instants are not concerned. The complete norm is $\|u^h\|_{\infty,2,T}$. The interest of the introduction of the seminorm $\|u^h\|^*_{\infty,2,T}$ will appear in the proof of Theorem 3.1 (cf. section 4).

We also need to introduce some notation for norms in spaces of continuous functions. For any $a > 0$ and any integer $k \geq 0$, we shall denote

$$\|f\|_{\mathcal{C}^k_{a,T}} = \sup_{x\in[-a,a],\ 0\leq t\leq T} \ \sup_{i+j\leq k} \left| \frac{\partial^{i+j} f}{\partial x^i \partial t^j}(x,t) \right| \quad \forall\ f \in \mathcal{C}^k([-a,a]\times[O,T]), \tag{3.6}$$

$$\|(u,v)\|_{\mathcal{C}^k_{a,T}} = \|u\|_{\mathcal{C}^k_{a,T}} + \|v\|_{\mathcal{C}^k_{a,T}} \quad \forall\ (u,v) \in \mathcal{C}^k([-a,a]\times[O,T])^2. \tag{3.7}$$

It will also be useful to introduce the class of functions

$$\mathbf{C}^\infty_{a,T} = \{(u,v) \in \mathcal{C}^\infty([-a,a]\times[O,T])^2 \text{ such that (3.9) holds}\}. \tag{3.8}$$

Property (3.9) expresses in some sense that the successive derivatives of the functions do not increase too quickly with the order of derivation. More precisely, that

$$|||(u,v)|||_{\mathbf{C}^\infty_{a,T}} \equiv \sup_{k\geq 0} |||(u,v)|||_{\mathcal{C}^k_{a,T}} < +\infty, \tag{3.9}$$

where, for each integer k,

$$|||(u,v)|||_{\mathcal{C}^k_{a,T}} = \|(u,v)\|^{\frac{1}{2^k}}_{\mathcal{C}^k_{a,T}} \prod_{j=1}^{k} \|(u,v)\|^{\frac{1}{2^j}}_{\mathcal{C}^{j+1}_{a,T}}. \tag{3.10}$$

In what follows, if (u,v) is defined for $x \in \mathbb{R}, t \geq 0$, we shall say that $(u,v) \in \mathbf{C}^\infty_{a,T}$ if its restriction to $[-a,a]\times[O,T]$ belongs to $\mathbf{C}^\infty_{a,T}$.

Remark 3.2. The introduction of the set $\mathbf{C}^\infty_{a,T}$ as well as the "norms" $|||(\cdot,\cdot)|||_{\mathcal{C}^k_{a,T}}$ and $|||(\cdot,\cdot)|||_{\mathbf{C}^\infty_{a,T}}$ is rather surprising in such a simple context as the 1D wave equation, and we are not sure that it is really necessary (see also the comments that follow the statement of the theorem at the end of this section). However, these notions will naturally appear in the proof of the theorem. It is interesting to note the following here:

- From the remark that

$$\forall\ k \geq 1, \quad \sum_{j=1}^{k} \frac{1}{2^j} + \frac{1}{2^k} = 1, \tag{3.11}$$

it follows that the maps $(u,v) \to |||(u,v)|||_{\mathcal{C}^k_{a,T}}$ and $(u,v) \to ||(u,v)|||_{\mathcal{C}^\infty_{a,T}}$ are homogeneous of degree 1, and that the set $\mathbf{C}^\infty_{a,T}$ is a cone. It also possesses a *homogeneity property.* Let (u,v) be defined for all real x and all positive t. For any real $\lambda > 0$, we set

$$(u_\lambda(x,t), v_\lambda(x,t)) = (u(\lambda x, \lambda t), v(\lambda x, \lambda t));$$

then

$$(u,v) \in \mathbf{C}^\infty_{a,T} \Longrightarrow (u_\lambda, v_\lambda) \in \mathbf{C}^\infty_{a/\lambda, T/\lambda}.$$

It suffices to remark that if $\hat{\lambda} = \max(1, \lambda)$, then

$$|||(u_\lambda, v_\lambda)|||_{\mathcal{C}^k_{a/\lambda, T/\lambda}} \leq \left(\hat{\lambda}^{\frac{1}{2^k}} \prod_{j=1}^{k} \hat{\lambda}^{\frac{j+1}{2^j}} \right) |||(u,v)|||_{\mathcal{C}^k_{a,T}},$$

and that $\lim_{k \to +\infty} \hat{\lambda}^{\frac{1}{2^k}} = 1$, $\prod_{j=1}^{+\infty} \hat{\lambda}^{\frac{j+1}{2^j}} < +\infty$. Note also that $|||(u,v)|||_{\mathcal{C}^k_{a,T}} = 0 \Longrightarrow (u,v) = 0$.

- However, $|||(\cdot,\cdot)|||_{\mathcal{C}^k_{a,T}}$ and $|||(\cdot,\cdot)|||_{\mathcal{C}^\infty_{a,T}}$ are not norms since they do not satisfy the triangular inequality. As a consequence, it is not so clear that the set $\mathbf{C}^\infty_{a,T}$ is a vector set (this point may be interesting but not central to this paper). Clearly, the fact that u belongs to $\mathbf{C}^\infty_{a,T}$ implies that the successive derivatives of u must not increase too quickly with the order of derivation. It is easy to see that $\mathbf{C}^\infty_{a,T}$ contains some well-known functional spaces such as the *Gevrey* spaces $\mathcal{G}^s_{a,T}$, $s \geq 1$ that can be defined as (see [21] for instance)

$$\mathcal{G}^s_{a,T} = \left\{ (u,v) \in \mathcal{C}^\infty_{a,T} \ / \ \exists \ (\mathcal{C}, \gamma) \ / \ \|(u,v)\|_{\mathcal{C}^j_{a,T}} \leq \mathcal{C} \ \gamma^j \ (j!)^s \right\}. \tag{3.12}$$

The set $\mathcal{G}^1_{a,T}$ is made up of analytic functions while the set $\mathcal{G}^s_{a,T}$ for $s > 1$ contains functions with compact support such as

$$f(x,t) = e^{\frac{1}{(x-t)^2 - \alpha^2}} \chi_{[-\alpha,\alpha]},$$

which is easily shown to belong to $\mathcal{G}^3_{a,T}$.

Let us introduce the spaces

$$H^k(\mathbb{R}) = \left\{ f \in L^2(\mathbb{R}) \text{ such that } \partial_x^{\tilde{k}} f \in L^2(\mathbb{R}), \ 0 \leq \tilde{k} \leq k \right\},$$

equipped by the natural norm (in particular $H^0(\mathbb{R}) = L^2(\mathbb{R})$). Let us also define for any Hilbert space H

$$W^{k,\infty}([0,T], H) = \left\{ w : [0,T] \mapsto H, \text{ such that } \sup_{t \in [0,T]} \|\partial_t^{\tilde{k}} w(t)\|_H \leq \infty, \ 0 \leq \tilde{k} \leq k \right\},$$

also equipped by the natural norm. In that way, we set the space

$$E = W^{3,\infty}([0,T], L^2(\mathbb{R})) \cap W^{0,\infty}([0,T], H^3(\mathbb{R})),$$

and we introduce the following notation:

$$\|f\|_E = \max\left\{ \|f\|_{W^{3,\infty}([0,T],L^2(\mathbb{R}))}, \|f\|_{W^{0,\infty}([0,T],H^3(\mathbb{R}))} \right\} \quad \forall \ f \in E,$$

$$\|(u,v)\|_E = \|u\|_E + \|v\|_E \qquad \forall \ (u,v) \in E^2.$$

For technical reasons we will assume that the initial conditions of problem (1.1) are such that

$$(u_0, v_0) \in \left(H^3(\mathbb{R})\right)^2, \quad \text{supp}(u_0, v_0) \cap \{0\} = \varnothing, \tag{3.13}$$

so that the exact solution $(u,v) \in E^2$. In particular, this implies that

$$\tilde{u}^{2n}_{c,h} \in L^2_{c,u}, \quad \tilde{v}^{2n+1}_{c,h} \in L^2_{c,v}, \quad u^n_{f,h} \in L^2_{f,u}, \quad v^{n+\frac{1}{2}}_{f,h} \in L^2_{f,v} \tag{3.14}$$

for all $n \in \mathbb{N}$. Let us also assume that the discrete initial conditions (2.7) satisfy

$$u^0_{c,h} \in L^2_{c,u}, \quad v^1_{c,h} \in L^2_{c,v}, \quad u^0_{f,h} \in L^2_{f,u}, \quad v^{\frac{1}{2}}_{f,h} \in L^2_{f,v}, \tag{3.15}$$

and that they are a good approximation of the exact initial conditions, for example,

$$\begin{aligned}
(u_c)^0_{2j} &= \frac{1}{2h}\int_{x_{2j-1}}^{x_{2j+1}} u_0(x)\mathrm{d}x,\\
(v_c)^1_{2j+1} &= \frac{1}{2h}\int_{x_{2j}}^{x_{2j+2}} v_0(x)\mathrm{d}x - \Delta t\frac{(u_c)^0_{2j+2}-(u_c)^0_{2j}}{2h},\\
(u_f)^0_j &= \frac{1}{h}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u_0(x)\mathrm{d}x,\\
(v_f)^{\frac{1}{2}}_{j+\frac{1}{2}} &= \frac{1}{h}\int_{x_j}^{x_{j+1}} v_0(x)\mathrm{d}x - \frac{\Delta t}{2}\frac{(u_f)^0_{j+1}-(u_f)^0_j}{h},
\end{aligned} \tag{3.16}$$

or

$$\begin{aligned}
(u_c)^0_{2j} &= u_0(2jh), & (v_c)^1_{2j+1} &= v_0((2j+1)h) - \Delta t\frac{(u_c)^0_{2j+2}-(u_c)^0_{2j}}{2h},\\
(u_f)^0_j &= u_0(jh), & (v_f)^{\frac{1}{2}}_{j+\frac{1}{2}} &= v_0((n+\tfrac{1}{2})h) - \frac{\Delta t}{2}\frac{(u_f)^0_{j+1}-(u_f)^0_j}{h}.
\end{aligned} \tag{3.17}$$

Statement of the main results.

THEOREM 3.1. *Assume that the discretization parameters h and Δt are related by the strict CFL condition* (1.3) *and that the initial condition (u_0, v_0) of equations* (1.1) *satisfies* (3.13) (*so that the exact solution (u, v) belongs to E^2*). *Let us consider initial data given by* (3.16) *or* (3.17). *Then the following hold:*

(i) *The discrete solutions (u^h, v^h) given by the schemes* I *and* II *satisfy the error estimates*

$$\left|\begin{aligned}
\|u-u^h\|_{\infty,2,T} + \|v-v^h\|_{\infty,2,T} &\le \mathcal{C}\ (1-\alpha^2)^{-1}\ T\ h^{\frac{1}{2}}\ \|(u,v)\|_{\mathcal{C}^1_{a,T}}\\
&\quad + \mathcal{C}\ (1-\alpha^2)^{-1}\ (1+T)\ h^2\ \|(u,v)\|_E.
\end{aligned}\right. \tag{3.18}$$

(ii) *If moreover $(u, v) \in \mathcal{C}^{k+1}([-a, a] \times [0, T])$ for some real $a > 0$ and integer $k \ge 1$, the discrete solution given by the scheme* I *satisfies*

$$\left|\begin{aligned}
\|u-u^h\|_{\infty,2,T} + \|v-v^h\|_{\infty,2,T} &\le \mathcal{C}\ (1-\alpha^2)^{-1} T\ h^{\left(\frac{3}{2}-\frac{1}{2^k}\right)}\ |||(u,v)|||_{\mathcal{C}^k_{a,T}}\\
&\quad + \mathcal{C}\ (1-\alpha^2)^{-1}\ (1+T)\ h^2\ \|(u,v)\|_E.
\end{aligned}\right. \tag{3.19}$$

(iii) *If, finally, $(u, v) \in E^2 \cap \mathbf{C}^\infty_{a,T}$, the discrete solution given by the scheme* I *satisfies*

$$\left|\begin{aligned}
\|u-u^h\|_{\infty,2,T} + \|v-v^h\|_{\infty,2,T} &\le \mathcal{C}\ (1-\alpha^2)^{-1}\ T\ h^{\frac{3}{2}}\ |||(u,v)|||_{\mathbf{C}^\infty_{a,T}}\\
&\quad + \mathcal{C}\ (1-\alpha^2)^{-1}\ (1+T)\ h^2\ \|(u,v)\|_E.
\end{aligned}\right. \tag{3.20}$$

Let us complete this theorem by the following comments:

- This theorem expresses the fact that, in the $L^\infty(0,T;L^2)$ norm, scheme I is of order 3/2 while scheme II is of order 1/2. In (3.18), (3.19), and (3.20), the right-hand side is the sum of two terms: the first one measures the error introduced by the transmission scheme while the second one (of second order) is due to the interior scheme.
- The coefficients appearing in the right-hand side of estimates (3.18) through (3.20) blow up when α goes to 1. This is coherent with what one observes numerically: the two schemes are not strongly convergent for $\alpha = 1$. Nevertheless, as has been explained in [12], good results are obtained with the values of α close to 1 (see also section 5).
- One can define the discrete $L^\infty(0,T;L^\infty)$ norm of the error, namely, $\|u - u^h\|_{\infty,T}$ and $\|v-v^h\|_{\infty,T}$, by replacing in definition (3.5) the discrete L^2 norms $\|u_{f,h}\|$, $\|v_{f,h}\|$, $\|u_{c,h}\|$, and $\|v_{c,h}\|$ by the discrete L^∞ norms

$$|u_{f,h}|_\infty = \sup_{j\geq 0}|u_j|, \quad |v_{f,h}|_\infty = \sup_{j\geq 0}|v_{j+\frac{1}{2}}|,$$
$$|u_{c,h}|_\infty = \sup_{j\leq 0}|u_{2j}|, \quad |v_{f,h}|_\infty = \sup_{j\leq -1}|v_{2j+1}|.$$

From the obvious inequality (see also Lemma 4.5)

$$\|u - u^h\|_{\infty,T} \leq \frac{C}{\sqrt{h}}\,\|u - u^h\|_{\infty,2,T}, \tag{3.21}$$

we deduce that

$$\begin{cases} \|u - u^h\|_{\infty,T} + \|v - u^h\|_{\infty,T} = O(h) & \text{with scheme I.} \\ \|u - u^h\|_{\infty,T} + \|v - u^h\|_{\infty,T} = O(1) & \text{with scheme II.} \end{cases}$$

- From Remark 2.1, we already know that scheme II cannot be convergent in the (discrete) $L^\infty(0,T;L^\infty)$ space. As a consequence, using (3.21), we deduce that the error estimate (3.18) is sharp (see also section 5). We also conjecture that the $O(h^{3/2})$ estimate for scheme I is optimal. This is more or less implicit in the plane wave analysis (see [12] and section 5.2) and in good agreement with the numerical results.
- If our results are optimal in terms of powers of h, it is not clear that it is the case concerning the required regularity of the solution, for instance, that we need the C^∞ regularity to obtain the $O(h^{3/2})$ error estimate with scheme I. However, the Fourier analysis (see [12]) does suggest that, at least, time regularity is needed. Moreover, the "norms" $|||(u,v)|||_{\mathcal{C}^k_{\alpha,T}}$ naturally appears in the proof of the theorem (see section 3.1).
- If the solution of the continuous problem is regular enough, scheme I is of order $h^{\frac{3}{2}}$ in the $L^\infty(0,T;L^2)$ norm. As is strongly suggested by the plane wave reflection-transmission analysis (see [12]) as well as numerical results (cf. section 5), we conjecture that scheme I (resp., scheme II) provides $O(h^2)$ (resp., $O(h)$) errors when these are measured in space regions that do not contain a neighborhood of the origin. The proof of such a result remains an open question for us.
- We have considered here the case of the 1D wave equation with constant coefficients. However, it is not difficult to see that the proofs of section 4 (based on energy methods) can be adapted to the case of the 1D wave equation with spatially variable coefficients.

Remark 3.3. The hypotheses demanded in Theorem 3.1 can be rewritten in terms of the regularity of the initial condition. In this way we have the following:

- Estimation (3.19) is satisfied since the initial condition satisfies (3.13) and belongs to $\mathcal{C}^{k+1}(-a-T, a+T)$.
- We define the class of functions

  $$\mathbf{C}_b^\infty = \{(u,v) \in \mathcal{C}^\infty([-b,b])^2 \text{ such that (3.22) holds}\},$$

  where

  $$|||(u,v)|||_{\mathbf{C}_b^\infty} \equiv \sup_{k\geq 0} |||(u,v)|||_{\mathcal{C}_b^k} < +\infty, \tag{3.22}$$

  with

  $$|||(u,v)|||_{\mathcal{C}_b^k} = \|(u,v)\|_{\mathcal{C}_b^k}^{\frac{1}{2^k}} \prod_{j=1}^{k} \|(u,v)\|_{\mathcal{C}_b^{j+1}}^{\frac{1}{2^j}},$$

  and

  $$\|f\|_{\mathcal{C}_b^k} = \sup_{x\in[-b,b]} \sup_{i+j\leq k} \left| \frac{\partial^{i+j} f}{\partial x^i \partial t^j}(x,t) \right| \quad \forall\, f \in \mathcal{C}^k([-b,b]).$$

  Then the hypothesis demanded in statement (iii) is satisfied since $(u_0, v_0) \in \mathbf{C}_{a+T}^\infty$.

4. Proof of the error estimates.

4.1. The equations satisfied by the errors. The first step of the proof consists, of course, in writing the scheme satisfied by the errors. This is also the opportunity to define some useful notation. We introduce "discrete traces" for the exact solution,

$$\begin{cases} (\widetilde{U}_c)_0^{2n+1} = \dfrac{1}{2}\left(\tilde{u}_0^{2n+2} + \tilde{u}_0^{2n}\right), \\[2ex] (\widetilde{V}_c)_0^{2n+1} = \tilde{v}_{-1}^{2n+1} - h\dfrac{\tilde{u}_0^{2n+2} - \tilde{u}_0^{2n}}{2\Delta t}, \end{cases} \tag{4.1}$$

$$\begin{cases} (\widetilde{U}_f)_0^{n+\frac{1}{2}} = \dfrac{1}{2}\left(\tilde{u}_0^{n+1} + \tilde{u}_0^{n}\right), \\[2ex] (\widetilde{V}_f)_0^{n+\frac{1}{2}} = \tilde{v}_{\frac{1}{2}}^{n+\frac{1}{2}} + \dfrac{h}{2}\dfrac{\tilde{u}_0^{n+1} - \tilde{u}_0^{n}}{\Delta t}, \end{cases} \tag{4.2}$$

and for the error,

$$\begin{cases} \left(e_c^U\right)_0^{2n+1} = \dfrac{1}{2}\left(\left(e_c^u\right)_0^{2n+2} + \left(e_c^u\right)_0^{2n}\right), \\[2ex] \left(e_c^V\right)_0^{2n+1} = \left(e_c^v\right)_{-1}^{2n+1} - h\dfrac{\left(e_c^u\right)_0^{2n+2} - \left(e_c^u\right)_0^{2n}}{2\Delta t}, \end{cases} \tag{4.3}$$

$$\begin{cases} \left(e_f^U\right)_0^{n+\frac{1}{2}} = \dfrac{1}{2}\left(\left(e_f^u\right)_0^{n+1} + \left(e_f^u\right)_0^{n}\right), \\[2ex] \left(e_f^V\right)_0^{n+\frac{1}{2}} = \left(e_f^v\right)_{\frac{1}{2}}^{n+\frac{1}{2}} + \dfrac{h}{2}\dfrac{\left(e_f^u\right)_0^{n+1} - \left(e_f^u\right)_0^{n}}{\Delta t}. \end{cases} \tag{4.4}$$

The interior equations satisfied by the error are

$$\tag{4.5} \begin{cases} \dfrac{\left(e_c^u\right)_{2j}^{2n+2} - \left(e_c^u\right)_{2j}^{2n}}{2\Delta t} + \dfrac{\left(e_c^v\right)_{2j+1}^{2n+1} - \left(e_c^v\right)_{2j-1}^{2n+1}}{2h} = \left(\eta_c^u\right)_{2j}^{2n+1}, \quad j \leq -1, \\ \dfrac{\left(e_c^v\right)_{2j+1}^{2n+1} - \left(e_c^v\right)_{2j+1}^{2n-1}}{2\Delta t} + \dfrac{\left(e_c^u\right)_{2j+2}^{2n} - \left(e_c^u\right)_{2j}^{2n}}{2h} = \left(\eta_c^v\right)_{2j+1}^{2n}, \quad j \leq -1, \end{cases}$$

in the coarse grid, and

$$\tag{4.6} \begin{array}{l} \begin{cases} \dfrac{\left(e_f^u\right)_j^{2n+1} - \left(e_f^u\right)_j^{2n}}{\Delta t} + \dfrac{\left(e_f^v\right)_{j+\frac12}^{2n+\frac12} - \left(e_f^v\right)_{j-\frac12}^{2n+\frac12}}{h} = \left(\eta_f^u\right)_j^{2n+\frac12}, \quad j \geq 1, \\ \dfrac{\left(e_f^v\right)_{j+\frac12}^{2n+\frac12} - \left(e_f^v\right)_{j+\frac12}^{2n-\frac12}}{\Delta t} + \dfrac{\left(e_f^u\right)_{j+1}^{2n} - \left(e_f^u\right)_j^{2n}}{h} = \left(\eta_f^v\right)_{j+\frac12}^{2n}, \quad j \geq 0, \end{cases} \\ \begin{cases} \dfrac{\left(e_f^u\right)_j^{2n+2} - \left(e_f^u\right)_j^{2n+1}}{\Delta t} + \dfrac{\left(e_f^v\right)_{j+\frac12}^{2n+\frac32} - \left(e_f^v\right)_{j-\frac12}^{2n+\frac32}}{h} = \left(\eta_f^u\right)_j^{2n+\frac32}, \quad j \geq 1, \\ \dfrac{\left(e_f^v\right)_{j+\frac12}^{2n+\frac32} - \left(e_f^v\right)_{j+\frac12}^{2n+\frac12}}{\Delta t} + \dfrac{\left(e_f^u\right)_{j+1}^{2n+1} - \left(e_f^u\right)_j^{2n+1}}{h} = \left(\eta_f^v\right)_{j+\frac12}^{2n+1}, \quad j \geq 0, \end{cases} \end{array}$$

in the fine grid. In (4.5) and (4.6), the terms on the right-hand side are classical *interior truncation errors* that are completely defined from the exact solution: they are nothing but the quantities in the left-hand sides of (4.5) and (4.6) after the substitution

$$\left(e_c^u\right)_{2j}^{2n}, \left(e_c^v\right)_{2j+1}^{2n+1}, \left(e_f^u\right)_j^n, \left(e_f^v\right)_{j+\frac12}^{n+\frac12} \longrightarrow (\tilde u_c)_{2j}^{2n}, (\tilde v_c)_{2j+1}^{2n+1}, (\tilde u_f)_j^n, (\tilde v_f)_{j+\frac12}^{n+\frac12}.$$

The equations on the interface are

$$\tag{4.7} \begin{cases} \left(e_c^V\right)_0^{2n+1} - \left(e_f^V\right)_0^{2n+\frac12} = \left(\varepsilon_r^v\right)^{2n+\frac12}, \\ \left(e_c^V\right)_0^{2n+1} - \left(e_f^V\right)_0^{2n+\frac32} = \left(\varepsilon_r^v\right)^{2n+\frac32}, \\ \left(e_c^U\right)_0^{2n+1} - \dfrac12\left(\left(e_f^U\right)_0^{2n+\frac12} + \left(e_f^U\right)_0^{2n+\frac32}\right) = \left(\varepsilon_r^u\right)^{2n+1} \end{cases}$$

for scheme I, and

$$\tag{4.8} \begin{cases} \left(e_c^U\right)_0^{2n+1} - \left(e_f^U\right)_0^{2n+\frac12} = \left(\widetilde\varepsilon_r^u\right)^{2n+\frac12}, \\ \left(e_c^U\right)_0^{2n+1} - \left(e_f^U\right)_0^{2n+\frac32} = \left(\widetilde\varepsilon_r^u\right)^{2n+\frac32}, \\ \left(e_c^V\right)_0^{2n+1} - \dfrac12\left(\left(e_f^V\right)_0^{2n+\frac12} + \left(e_f^V\right)_0^{2n+\frac32}\right) = \left(\widetilde\varepsilon_r^v\right)^{2n+1} \end{cases}$$

for scheme II. The quantities on the right-hand side of (4.7) (resp., (4.8)) are the *interface truncation errors* for scheme I (resp., scheme II). Once again, they are completely defined by the exact solution, as the terms in the left-hand side of (4.7) (resp., (4.8)) in which we have made the substitution

$$\left(e_c^U\right)_0^{2n+1}, \left(e_c^V\right)_0^{2n+1}, \left(e_f^U\right)_0^{n+\frac12}, \left(e_f^V\right)_0^{n+\frac12} \longrightarrow (\widetilde U_c)_0^{2n+1}, (\widetilde V_c)_0^{2n+1}, (\widetilde U_f)_0^{n+\frac12}, (\widetilde V_f)_0^{n+\frac12}.$$

4.2. Outline of the proof. It is clear that, by exploiting the linearity of the equations, the error can be separated into two parts:

- the error due to the *interior truncation errors* and *initial conditions*;
- the error due to the *interface truncation errors.*

Let us formalize this setting:

$$\delta^h := \Big\{ \big(e_c^u\big)^0_{2j}, \big(e_c^v\big)^1_{2j+1}, \big(e_f^u\big)^0_j, \big(e_f^v\big)^{\frac{1}{2}}_{j+\frac{1}{2}} \Big\}, \qquad \text{the } \textit{initial errors,}$$

$$\eta^h := \Big\{ \big(\eta_c^u\big)^{2n}_{2j}, \big(\eta_c^v\big)^{2n+1}_{2j+1}, \big(\eta_f^u\big)^{n+\frac{1}{2}}_j, \big(\eta_f^v\big)^{n}_{j+\frac{1}{2}} \Big\}, \qquad \text{the } \textit{interior truncation errors,}$$

$$\varepsilon^h_I := \big\{ \big(\varepsilon_r^u\big), \big(\varepsilon_r^v\big) \ \big\}, \quad \varepsilon^h_{II} := \big\{ \big(\widetilde{\varepsilon}_r^u\big), \big(\widetilde{\varepsilon}_r^v\big) \big\}, \qquad \text{the } \textit{interface truncation errors.}$$

It is clear that if $(\delta^h, \eta^h, \varepsilon^h)$ are known, the sequences $(e^{u,h}, e^{v,h})$ are completely characterized by the interior equations and the transmission scheme I or II. In this way, we define two maps $\Phi_l, l = I, II$,

$$\big(\delta^h, \eta^h, \varepsilon^h_l\big) \xrightarrow{\Phi_l} (e^{u,h}, e^{v,h}),$$

which are linear. In particular

$$\Big(\big(e_c^{u,h}\big), \big(e_c^{v,h}\big), \big(e_f^{u,h}\big), \big(e_f^{v,h}\big) \Big) = \Phi_l(\delta^h, \eta^h, 0) + \Phi_l\big(0, 0, \varepsilon^h_l\big).$$

- The estimate of $\Phi_l(\delta^h, \eta^h, 0)$, due to the *interior truncation errors* and the approximations of *initial data*, does not really depend on the *transmission scheme*, provided that this scheme is conservative in the sense of Theorem 2.1, which is the case for schemes I and II. As a consequence of the centered nature of the scheme, this *error* is $O(h^2)$ provided that the initial conditions are approximated to $O(h^2)$. The precise result is the following.

  Proposition 4.1. *Let h and Δt be constants such that*

  $$\alpha := \frac{\Delta t}{h} < 1$$

  and let $T > 0$. Assume that the initial condition (u_0, v_0) of equations (1.1) *satisfies* (3.13) *(so that the exact solution (u, v) belongs to E^2). Let us consider initial data given by* (3.16) *or* (3.17). *Then, $(e^{u,h}, e^{v,h}) = \Phi_l(\delta^h, \eta^h, 0)$ satisfies the following estimate:*

  $$\|e^{u,h}\|_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T} \leq \mathcal{C}(1-\alpha^2)^{-1}\,(1+T)\,h^2\|(u,v)\|_E. \tag{4.9}$$

  The analysis of this error is very similar to that (rather standard) of the "pure" Yee scheme (i.e., without any mesh refinement) and, as this point is not central to this paper, we have chosen not to give the proof here. We refer to the reader to [17] for more details.

- The estimate of $\Phi_l(0, 0, \varepsilon^h_l)$ does depend on the *transmission scheme.* The analysis, presented in sections 4.3 and 4.4, is much less classical and consists in two main steps:
  - For both schemes I and II, a direct analysis combining the use of *energy techniques* (as for the stability analysis), *consistency estimates* for the

transmission conditions (*globally* in $O(h)$—see Lemma 4.3), and the use of a *discrete trace inequality* (which results in the loss of one half-power of h—see Lemma 4.5) permits us to show an $O(\sqrt{h})$ estimate for both schemes I and II. This is Lemma 4.2. The proof stops here for scheme II.

– For scheme I, one can use a *bootstrap* argument to improve iteratively the obtained rate of convergence: the estimate that will lead to (3.19) is proved by *induction* on k (this is Lemma 4.8) and that leading to (3.20) by passing to the limit when $k \to +\infty$ (this is Lemma 4.9). This demands a closer look at the structure of the *transmission truncation error* ε_I that has some properties that the error ε_{II} does not.

In summary, estimate (3.18) is obtained by combining Proposition 4.1 with Lemma 4.2, (3.19) is obtained by regrouping Proposition 4.1 with Lemma 4.8, and (3.20) is obtained by regrouping Proposition 4.1 with Lemma 4.9.

4.3. Proof of the $\mathcal{O}(\sqrt{h})$ estimates. Provided that similar techniques can be applied to the analysis of scheme II, only the estimate for scheme I will be proven. The main difference between the two proofs will be pointed out in Remark 4.1.

What we are going to derive here is the equivalent of estimate (3.18) for $\Phi_I(0,0,\varepsilon_l)$. For the sake of simplicity, we shall still denote in this section

$$\Phi_I\big(0,0,\varepsilon_l^h\big) = \Big(\big(e_c^{u,h}\big),\big(e_c^{v,h}\big),\big(e_f^{u,h}\big),\big(e_f^{v,h}\big)\Big).$$

We shall also use the notation

$$e^{u,h} = \Big(\big(e_c^{u,h}\big),\big(e_f^{u,h}\big)\Big), \quad e^{v,h} = \Big(\big(e_c^{v,h}\big),\big(e_f^{v,h}\big)\Big)$$

and refer to definition (3.5) for the discrete norms. Throughout this section we will use only the last two norms of (3.5). The first one (that we call norm-star) will be useful for the proof of (3.19). The estimate we want to prove here is the following.

LEMMA 4.2. *If the solution of the continuous problem* (1.1) *belongs to* $\mathcal{C}^1_{a,T}$ *for* $a > 0$*, then*

$$\|e^{u,h}\|_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T} \;\leq\; \mathcal{C}\,(1-\alpha^2)^{-1}\,T\,h^{\frac{1}{2}}\,\|(u,v)\|_{\mathcal{C}^1_{a,T}}. \tag{4.10}$$

The rest of this section is devoted to the proof of this lemma. By definition,

$$\big(e_c^{u,h}\big),\big(e_c^{v,h}\big),\big(e_f^{u,h}\big), \text{ and } \big(e_f^{v,h}\big)$$

satisfy the *homogeneous* interior equations (4.5) and (4.6) (i.e., with *zero* right-hand sides), and we recall below the equations at the interface,

$$\begin{cases} \big(e_c^V\big)_0^{2n+1} - \big(e_f^V\big)_0^{2n+\frac{1}{2}} = (\varepsilon_r^v)^{2n+\frac{1}{2}}, \\ \big(e_c^V\big)_0^{2n+1} - \big(e_f^V\big)_0^{2n+\frac{3}{2}} = \big(\varepsilon_r^v\big)^{2n+\frac{3}{2}}, \\ \big(e_c^U\big)_0^{2n+1} - \dfrac{1}{2}\Big(\big(e_f^U\big)_0^{2n+\frac{1}{2}} + \big(e_f^U\big)_0^{2n+\frac{3}{2}}\Big) = (\varepsilon_r^u)^{2n+1}, \end{cases} \tag{4.11}$$

where the *interface truncation errors* are given by (we indicate the order of magnitude of each term obtained by a Taylor expansion)

$$\big(\varepsilon_r^v\big)^{2n+\frac{1}{2}} = (\widetilde{V}_c)_0^{2n+1} - (\widetilde{V}_f)_0^{2n+\frac{1}{2}} = \mathcal{O}(\Delta t),$$

$$\big(\varepsilon_r^v\big)^{2n+\frac{3}{2}} = (\widetilde{V}_c)_0^{2n+1} - (\widetilde{V}_f)_0^{2n+\frac{3}{2}} = \mathcal{O}(\Delta t),$$

$$\big(\varepsilon_r^u\big)^{2n+1} = (\widetilde{U}_c)_0^{2n+1} - \tfrac{1}{2}\Big((\widetilde{U}_f)_0^{2n+\frac{1}{2}} + (\widetilde{U}_f)_0^{2n+\frac{3}{2}}\Big) = \mathcal{O}(\Delta t^2),$$

where the quantities $(\widetilde{U}_c)_0^{2n+1}, (\widetilde{V}_c)_0^{2n+1}, (\widetilde{V}_f)_0^{2n+\frac{1}{2}}, (\widetilde{V}_f)_0^{2n+\frac{1}{2}}$ are defined from the exact solution (u,v) by ((4.1) and (4.2)). The following lemma, whose immediate proof is omitted here (it is based on a simple Taylor expansion), gives us the magnitude of these quantities.

LEMMA 4.3. *Assume that, for some $a > 0$, u and v belong to $\mathcal{C}^1_{a,T}$. Then, provided that $\alpha \leq 1$,*

$$\text{(4.12)} \quad \sup_{t^{2n}\leq T} |\big(\varepsilon_r^v\big)^{n+\frac{1}{2}}| \leq \mathcal{C}\ h\ \|(u,v)\|_{\mathcal{C}^1_{a,T}}, \qquad \sup_{t^{2n}\leq T} |\big(\varepsilon_r^u\big)^{2n+1}| \leq \mathcal{C}\ \alpha\ h\ \|u\|_{\mathcal{C}^1_{a,T}}.$$

If moreover $(u,v) \in \mathcal{C}^2_{a,T}$, then

$$\sup_{t^{2n}\leq T} |(\varepsilon_r^u)^{2n+1}| \leq \mathcal{C}\ \alpha^2\ h^2\ \|u\|_{\mathcal{C}^2_{a,T}}, \qquad \sup_{t^{2n}\leq T} |\big(\varepsilon_r^v\big)^{2n+\frac{1}{2}} + \big(\varepsilon_r^v\big)^{2n+\frac{3}{2}}| \leq \mathcal{C}\ h^2\ \|(u,v)\|_{\mathcal{C}^2_{a,T}}.$$
(4.13)

Next, as for to the discrete energy (2.8), we introduce the discrete energy of the error at even instants:

$$\text{(4.14)} \qquad \mathcal{E}^{2n} = \mathcal{E}_c^{2n} + \mathcal{E}_f^{2n},$$

$$\text{where} \quad \left| \begin{aligned} \mathcal{E}_c^{2n} &= \frac{1}{2} \sum_{j\leq -1} |\big(e_c^u\big)_{2j}^{2n}|^2\ 2h + \frac{1}{2} \sum_{j\leq -1} \big(e_c^v\big)_{2j+1}^{2n+1} \big(e_c^v\big)_{2j+1}^{2n-1}\ 2h + \frac{1}{2}\ |\big(e_c^u\big)_0^{2n}|^2\ h, \\ \mathcal{E}_f^n &= \frac{1}{2} \sum_{j\geq 1} |\big(e_f^u\big)_j^n|^2\ h + \frac{1}{2} \sum_{j\geq 0} \big(e_f^v\big)_{j+\frac{1}{2}}^{n+\frac{1}{2}} \big(e_f^v\big)_{j+\frac{1}{2}}^{n-\frac{1}{2}}\ h + \frac{1}{2}\ |\big(e_f^u\big)_0^n|^2\ \frac{h}{2}. \end{aligned} \right.$$

Our goal will be to obtain an estimate for $\sqrt{\mathcal{E}^{2n}}$. This will provide an L^2-estimate for the error with the aid of the following lemma.

LEMMA 4.4. *Assume that* (1.3) *holds and that*

$$\Big(\big(e_{f,h}^u\big)^n, \big(e_{f,h}^v\big)^{n+\frac{1}{2}}\Big) \in L^2_{f,u} \times L^2_{f,v} \quad \text{ and } \quad \Big(\big(e_{c,h}^u\big)^{2n}, \big(e_{c,h}^v\big)^{2n+1}\Big) \in L^2_{c,u} \times L^2_{c,v}.$$

Then, there exists a positive constant $\mathcal{C}$ independent of Δt, h, and α such that for any $n > 0$

$$\text{(4.15)} \quad \|\big(e_{c,h}^u\big)^{2n}\|^2 + \|\big(e_{c,h}^v\big)^{2n+1}\|^2 + \|\big(e_{c,h}^v\big)^{2n-1}\|^2 \leq \mathcal{C}\ (1-\alpha^2)^{-1}\ \mathcal{E}_c^{2n},$$

$$\text{(4.16)} \quad \|\big(e_{f,h}^u\big)^{2n}\|^2 + \|\big(e_{f,h}^v\big)^{2n+\frac{1}{2}}\|^2 + \|\big(e_{f,h}^v\big)^{2n-\frac{1}{2}}\|^2 \leq \mathcal{C}\ (1-\alpha^2)^{-1}\ \mathcal{E}_f^{2n},$$

$$\text{(4.17)} \quad \|\big(e_{f,h}^u\big)^{2n+1}\|^2 \leq \mathcal{C}\ (1-\alpha^2)^{-1}\ \big(\mathcal{E}^{2n} + \mathcal{E}^{2n+2}\big) + \mathcal{C}\ h\ |\big(\varepsilon_r^u\big)^{2n+1}|^2.$$

Proof. We first prove (4.15). Using the identity $4ab = (a+b)^2 - (a-b)^2$, we obtain

$$\text{(4.18)} \quad \left| \begin{aligned} \mathcal{E}_c^{2n} = \frac{1}{2} \left(\sum_{j\leq -1} |\big(e_c^u\big)_{2j}^{2n}|^2\ 2h + |\big(e_c^u\big)_0^{2n}|^2\ h \right) &+ \frac{1}{2} \sum_{j\leq -1} \left| \frac{\big(e_c^v\big)_{2j+1}^{2n+1} + \big(e_c^v\big)_{2j+1}^{2n-1}}{2} \right|^2\ 2h \\ &- \frac{1}{2} \sum_{j\leq -1} \left| \frac{\big(e_c^v\big)_{2j+1}^{2n+1} - \big(e_c^v\big)_{2j+1}^{2n-1}}{2} \right|^2\ 2h. \end{aligned} \right.$$

Using the second equation of scheme (4.5), we observe that

$$\left|\frac{\left(e_c^v\right)_{2j+1}^{2n+1}-\left(e_c^v\right)_{2j+1}^{2n-1}}{2}\right|^2=\frac{\alpha^2}{4}\,\left|\left(e_c^u\right)_{2j+2}^{2n}-\left(e_c^u\right)_{2j}^{2n}\right|^2\leq\frac{\alpha^2}{2}\,\left(\left|\left(e_c^u\right)_{2j+2}^{2n}\right|^2+\left|\left(e_c^u\right)_{2j}^{2n}\right|^2\right).$$

We use this in (4.18) to deduce that

$$\mathcal{E}_c^{2n}\geq\frac{1}{2}\left\|\frac{\left(e_{c,h}^v\right)^{2n-1}+\left(e_{c,h}^v\right)^{2n+1}}{2}\right\|^2+\frac{1-\alpha^2}{2}\left\|\left(e_{c,h}^u\right)^{2n}\right\|^2, \tag{4.19}$$

which implies in particular

$$\|\left(e_{c,h}^u\right)^{2n}\|^2\leq 2\,(1-\alpha^2)^{-1}\,\mathcal{E}_c^{2n}. \tag{4.20}$$

Next, we remark that the second equality of (4.5) can be rewritten as

$$\left(e_c^v\right)_{2j+1}^{2n\pm1}=\frac{\left(e_c^v\right)_{2j+1}^{2n+1}+\left(e_c^v\right)_{2j+1}^{2n-1}}{2}\mp\frac{\Delta t}{2h}\left(\left(e_c^u\right)_{2j+2}^{2n}-\left(e_c^u\right)_{2j}^{2n}\right).$$

Then, using the inequality $(a+b)^2\leq 2(a^2+b^2)$ twice, we obtain

$$\|\left(e_{c,h}^v\right)^{2n\pm1}\|^2\leq 2\,\|\frac{\left(e_{c,h}^v\right)^{2n-1}+\left(e_{c,h}^v\right)^{2n+1}}{2}\|^2+2\alpha^2\,\|\left(e_{c,h}^u\right)^{2n}\|^2.$$

Using now (4.19) and (4.20), we deduce the existence of a constant C such that

$$\|\left(e_{c,h}^u\right)^{2n}\|^2+\|\left(e_{c,h}^v\right)^{2n+1}\|^2+\|\left(e_{c,h}^v\right)^{2n-1}\|^2\leq\mathcal{C}(1-\alpha^2)^{-1}\mathcal{E}_c^{2n}, \tag{4.21}$$

and (4.15) is proven.

If we use techniques similar to those used above, it is easy (we omit the details) to show that, for any integer k,

$$\|\left(e_{f,h}^u\right)^{k}\|^2+\|\left(e_{f,h}^v\right)^{k+\frac{1}{2}}\|^2+\|\left(e_{f,h}^v\right)^{k-\frac{1}{2}}\|^2\leq\mathcal{C}(1-\alpha^2)^{-1}\mathcal{E}_f^{k},$$

which gives (4.16) for $k=2n$.

If we take $k=2n+1$, we are only able to bound $\|(e_{f,h}^u)^{2n+1}\|$ in terms of $\mathcal{E}_f^{2n+1}$ but not in terms of $\mathcal{E}^{2n}$ and $\mathcal{E}^{2n+2}$ (the conserved energy).

In order to do so, we use the first equality of (4.6) for $j\geq 1$ to obtain

$$|\left(e_f^u\right)_j^{2n+1}|\leq|\left(e_f^u\right)_j^{2n}|+\alpha\left(|\left(e_f^v\right)_{j+\frac{1}{2}}^{2n+\frac{1}{2}}|+|\left(e_f^v\right)_{j-\frac{1}{2}}^{2n+\frac{1}{2}}|\right).$$

Therefore, using $(a+b)^2\leq 2(a^2+b^2)$ once again,

$$\|\left(e_{f,h}^u\right)^{2n+1}\|^2\leq\mathcal{C}\left(\|\left(e_{f,h}^u\right)^{2n}\|^2+\|\left(e_{f,h}^v\right)^{2n+\frac{1}{2}}\|^2\right)+\frac{h}{2}\,|\left(e_f^u\right)_0^{2n}|^2,$$

which yields, with the aid of (4.16),

$$\|\left(e_{f,h}^u\right)^{2n+1}\|^2\leq\mathcal{C}\left(1-\alpha^2\right)^{-1}\mathcal{E}_f^{2n}+\frac{h}{2}|\left(e_f^u\right)_0^{2n+1}|^2. \tag{4.22}$$

To conclude we use the third equality of (4.7) rewritten as

$$\left(e_f^u\right)_0^{2n+1} = \left(e_c^u\right)_0^{2n+2} + \left(e_c^u\right)_0^{2n} - \frac{1}{2}\left(\left(e_f^u\right)_0^{2n+2} + \left(e_f^u\right)_0^{2n}\right) - 2(\varepsilon_r^u)^{2n+1}.$$

This obviously implies that

$$|\left(e_f^u\right)_0^{2n+1}|^2 \leq \frac{\mathcal{C}}{h}(\ \|\left(e_{c,h}^u\right)^{2n+2}\|^2 + \|\left(e_{c,h}^u\right)^{2n}\|^2 + \|\left(e_{f,h}^u\right)^{2n+2}\|^2 + \|\left(e_{f,h}^u\right)^{2n}\|^2) + |\left(\varepsilon_r^u\right)^{2n+1}|^2,$$

which yields, using this time (4.15) and (4.16), that

$$\frac{h}{2}\ |\left(e_f^u\right)_0^{2n+1}|^2 \leq \mathcal{C}\ (1-\alpha^2)^{-1}\ \left(\mathcal{E}^{2n} + \mathcal{E}^{2n+2}\right) + \mathcal{C}\ h\ |\left(\varepsilon_r^u\right)^{2n+1}|^2. \tag{4.23}$$

Finally, (4.17) is a direct consequence of (4.22) and (4.23). □

For the estimation of the energy, we use the following identity which is the equivalent for the error of the estimate of Theorem 2.1 for the discrete solution:

$$\left|\begin{aligned} \frac{1}{2\Delta t}\left(\mathcal{E}^{2n+2} - \mathcal{E}^{2n}\right) = {} & \frac{1}{2}\left(\left(e_f^U\right)_0^{2n+\frac{1}{2}}\left(e_f^V\right)_0^{2n+\frac{1}{2}} + \left(e_f^U\right)_0^{2n+\frac{3}{2}}\left(e_f^V\right)_0^{2n+\frac{3}{2}}\right) \\ & - \left(e_c^U\right)_0^{2n+1}\left(e_c^V\right)_0^{2n+1}. \end{aligned}\right. \tag{4.24}$$

The rest of the proof consists in deducing from (4.24) an appropriate estimate for the energy $\mathcal{E}^{2n}$ and then applying Lemma 4.3 to get an estimate of the error. In order to do this, we reorganize the three terms using (4.7) in order to exhibit the consistency errors (see Remark 4.1)

$$\left|\begin{aligned} \frac{1}{2\Delta t}\left(\mathcal{E}^{2n+2} - \mathcal{E}^{2n}\right) = {} & -\frac{1}{2}\left(\left(e_f^U\right)_0^{2n+\frac{1}{2}}\left(\varepsilon_r^v\right)^{2n+\frac{1}{2}} + \left(e_f^U\right)_0^{2n+\frac{3}{2}}\left(\varepsilon_r^v\right)^{2n+\frac{3}{2}}\right) \\ & - \left(\varepsilon_r^u\right)^{2n+1}\left(e_c^V\right)_0^{2n+1}. \end{aligned}\right. \tag{4.25}$$

According to Lemma 4.3, the quantities $(\varepsilon_r^v)^{2n+\frac{1}{2}}$, $(\varepsilon_r^v)^{2n+\frac{3}{2}}$, and $(\varepsilon_r^u)^{2n+1}$ are small. In order to bound the other terms appearing on the right-hand side of (4.25) (and defined in (4.3) and (4.4)) by a function of the error norm, we use a discrete trace lemma.

LEMMA 4.5. *Under the hypotheses of Lemma* 4.4*, there exists a positive constant* $\mathcal{C}$ *independent of* Δt *and* h *such that, for each* $n > 0$*,*

$$\begin{aligned} &|\left(e_f^u\right)_0^n| \leq \tfrac{\sqrt{2}}{\sqrt{h}}\|\left(e_{f,h}^u\right)^n\|, && |\left(e_f^v\right)_0^{n+\frac{1}{2}}| \leq \tfrac{1}{\sqrt{h}}\|\left(e_{f,h}^v\right)^{n+\frac{1}{2}}\|, \\ &|\left(e_c^u\right)_0^{2n}| \leq \tfrac{1}{\sqrt{h}}\|\left(e_{c,h}^u\right)^{2n}\|, && |\left(e_c^v\right)_0^{2n+1}| \leq \tfrac{1}{\sqrt{2h}}\|\left(e_{c,h}^v\right)^{2n+1}\|. \end{aligned} \tag{4.26}$$

Proof. The result is trivial. □

For simplicity of exposition, it is useful to introduce a local measure of the error on the time interval $I_{2n+1} = [t^{2n}, t^{2n+2}]$. Thus we set

$$\left\{\begin{aligned} \|e^{u,h}\|_{h,I_{2n+1}}^2 = {} & \|\left(e_{c,h}^u\right)^{2n}\|^2 + \|\left(e_{c,h}^u\right)^{2n+2}\|^2 \\ & + \|\left(e_{f,h}^u\right)^{2n}\|^2 + \|\left(e_{f,h}^u\right)^{2n+1}\|^2 + \|\left(e_{f,h}^u\right)^{2n+2}\|^2, \\ \|e^{v,h}\|_{h,I_{2n+1}}^2 = {} & \|\left(e_{c,h}^v\right)^{2n+1}\|^2 + \|\left(e_{f,h}^v\right)^{2n+\frac{1}{2}}\|^2 + \|\left(e_{f,h}^v\right)^{2n+\frac{3}{2}}\|^2. \end{aligned}\right. \tag{4.27}$$

Note that these are the quantities appearing in the left-hand sides of inequalities (4.15) through (4.17) of Lemma 4.4. By definition of the discrete $L^\infty(0,T;L^2)$ norm, we have, for $t^{2n+2} \leq T$,

$$\left\{\begin{array}{ll} \|e^{u,h}\|_{h,I_{2n+1}} \leq C\ \|e^{u,h}\|_{\infty,2,T}, & \|e^{v,h}\|_{h,I_{2n+1}} \leq C\ \|e^{v,h}\|_{\infty,2,T}, \\ \|e^{u,h}\|_{\infty,2,T} \leq \displaystyle\sup_{t^{2n+2}\leq T} \|e^{u,h}\|_{h,I_{2n+1}}, & \|e^{v,h}\|_{\infty,2,T} \leq \displaystyle\sup_{t^{2n+2}\leq T} \|e^{v,h}\|_{h,I_{2n+1}}. \end{array}\right. \tag{4.28}$$

Using elementary manipulations on expression (4.25) and Lemma 4.5, the following inequality can be obtained (note that the factor $1/\alpha$ appearing below comes from the right-hand side of the second equation of (4.3)):

$$\left| \begin{array}{l} \displaystyle\frac{\mathcal{E}^{2n+2}-\mathcal{E}^{2n}}{2\Delta t} \leq \frac{\mathcal{C}}{\sqrt{h}}\ \|e^{u,h}\|_{h,I_{2n+1}}\ \left\{|(\varepsilon_r^v)^{2n+\frac{1}{2}}| + |(\varepsilon_r^v)^{2n+\frac{3}{2}}|\right\} \\ \qquad\qquad\qquad \displaystyle + \frac{\mathcal{C}}{\alpha\sqrt{h}}\ \left\{\|e^{u,h}\|_{h,I_{2n+1}} + \alpha\|e^{v,h}\|_{h,I_{2n+1}}\right\}\ |(\varepsilon_r^u)^{2n+1}|. \end{array}\right.$$

Then, by Lemma 4.3 (inequalities (4.12)) and (4.28)

$$\frac{\mathcal{E}^{2n+2}-\mathcal{E}^{2n}}{2\Delta t} \leq \mathcal{C}\ \sqrt{h}\ \|(u,v)\|_{\mathcal{C}^1_{a,T}} \left(\|e^{u,h}\|_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T}\right). \tag{4.29}$$

Adding the above inequalities from $n = 0$ to $m-1$, for any integer $m > 1$, we obtain ($\mathcal{E}_0 = 0$)

$$\mathcal{E}^{2m} \leq \mathcal{C}\ t^{2m}\ \sqrt{h}\ \|(u,v)\|_{\mathcal{C}^1_{a,T}} \left\{\|e^{u,h}\|_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T}\right\}.$$

Now, using Lemma 4.4, we can write, for $t^{2n+2} \leq T$,

$$\left| \begin{array}{l} \|e^{u,h}\|^2_{h,I_{2n+1}} + \|e^{v,h}\|^2_{h,I_{2n+1}} \leq \mathcal{C}\ (1-\alpha^2)^{-1} \left(\mathcal{E}^{2n} + \mathcal{E}^{2n+2}\right) + \mathcal{C}\ h\ |(\varepsilon_r^u)^{2n+1}|^2 \\ \quad \leq \mathcal{C}\ (1-\alpha^2)^{-1}\ T\ \sqrt{h}\ \|(u,v)\|_{\mathcal{C}^1_{a,T}} \left\{\|e^{u,h}\|_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T}\right\} + \mathcal{C}\ h^3\ \|(u,v)\|^2_{\mathcal{C}^1_{a,T}}. \end{array}\right.$$

Therefore, taking the supremum over $t^{2n+2} \leq T$, using (4.28) and classical manipulations based on Young's inequality, one proves the final estimate (4.10) (we omit the details).

Remark 4.1. Let us give some details concerning the derivation of (4.25). We start from the identities

$$\left(e_f^U\right)_0^{2n+\frac{1}{2}} \left(e_f^V\right)_0^{2n+\frac{1}{2}} = \left(e_f^U\right)_0^{2n+\frac{1}{2}} \left[\left(e_f^V\right)_0^{2n+\frac{1}{2}} - \left(e_c^V\right)_0^{2n+1}\right] + \left(e_f^U\right)_0^{2n+\frac{1}{2}} \left(e_c^V\right)_0^{2n+1},$$

$$\left(e_f^U\right)_0^{2n+\frac{3}{2}} \left(e_f^V\right)_0^{2n+\frac{3}{2}} = \left(e_f^U\right)_0^{2n+\frac{3}{2}} \left[\left(e_f^V\right)_0^{2n+\frac{3}{2}} - \left(e_c^V\right)_0^{2n+1}\right] + \left(e_f^U\right)_0^{2n+\frac{3}{2}} \left(e_c^V\right)_0^{2n+1}.$$

After summation, we obtain, using (4.11),

$$\begin{array}{l} \displaystyle\frac{1}{2}\left\{\left(e_f^U\right)_0^{2n+\frac{1}{2}} \left(e_f^V\right)_0^{2n+\frac{1}{2}} + \left(e_f^U\right)_0^{2n+\frac{3}{2}} \left(e_f^V\right)_0^{2n+\frac{3}{2}}\right\} = \left(e_c^V\right)_0^{2n+1} \left[\frac{\left(e_f^U\right)_0^{2n+\frac{1}{2}} + \left(e_f^U\right)_0^{2n+\frac{3}{2}}}{2}\right] \\ \qquad\qquad\qquad\qquad \displaystyle - \frac{1}{2}\ \left(e_f^U\right)_0^{2n+\frac{1}{2}} \left(\varepsilon_r^v\right)_0^{2n+\frac{1}{2}} - \frac{1}{2}\ \left(e_f^U\right)_0^{2n+\frac{3}{2}} \left(\varepsilon_r^v\right)_0^{2n+\frac{3}{2}}. \end{array} \tag{4.30}$$

On the other hand, one has the identity

$$\left|\ \begin{aligned} \left(e_c^U\right)_0^{2n+1}\left(e_c^V\right)_0^{2n+1} = \ & \left[\left(e_c^U\right)_0^{2n+1} - \frac{1}{2}\ \left\{\left(e_f^U\right)_0^{2n+\frac{3}{2}} + \left(e_f^U\right)_0^{2n+\frac{3}{2}}\right\}\right]\left(e_c^V\right)_0^{2n+1} \\ & + \frac{1}{2}\ \left[\left(e_f^U\right)_0^{2n+\frac{3}{2}} + \left(e_f^U\right)_0^{2n+\frac{3}{2}}\right]\left(e_c^V\right)_0^{2n+1}, \end{aligned}\right.$$

that is to say, with the aid of (4.11),

$$\left(e_c^U\right)_0^{2n+1}\left(e_c^V\right)_0^{2n+1} = \left(e_c^V\right)_0^{2n+1}\left(\varepsilon_r^u\right)_0^{2n+1} + \frac{1}{2}\left[\left(e_f^U\right)_0^{2n+\frac{3}{2}} + \left(e_f^U\right)_0^{2n+\frac{3}{2}}\right]\left(e_c^V\right)_0^{2n+1}. \tag{4.31}$$

Finally, (4.25) is obtained as the difference between (4.30) and (4.31).

If scheme II is used, we have

$$\frac{\mathcal{E}^{2n+2} - \mathcal{E}^{2n}}{2\Delta t} = -\frac{1}{2}\left(\left(e_f^V\right)_0^{2n+\frac{1}{2}}\left(\widetilde{\varepsilon}_r^u\right)^{2n+\frac{1}{2}} + \left(e_f^V\right)_0^{2n+\frac{3}{2}}\left(\widetilde{\varepsilon}_r^u\right)^{2n+\frac{3}{2}}\right) - \left(\widetilde{\varepsilon}_r^v\right)^{2n+1}\left(e_c^U\right)_0^{2n+1}, \tag{4.32}$$

and the proof of estimate (3.18) is similar to that presented for scheme I.

4.4. Proof of estimate (3.19). Apart from the results of section 4.4.1, which are essentially generalizations of estimate (3.18), what we do in this section is valid only when we use scheme I to do the coupling. The real novelty in the proof will appear in section 4.4.2. The main differences between scheme I and scheme II will be explained in Remark 4.3.

4.4.1. Estimate of coarse discrete derivatives. Our goal in this paragraph is to derive estimates similar to (3.18) for which we shall call the successive *coarse time discrete derivatives* of the error $(e^{u,h}, e^{v,h})$. The proof is essentially a repetition of the proof of section 4.3, but the statement of the precise result requires some notation.

We define the *coarse discrete derivative* operator D by defining its action on a sequence $w^h = (w)_s^t$ (where t and s are integers or integers plus one half—negative indices t are allowed):

$$(Dw)_s^{t+1} := \frac{(w)_s^{t+2} - (w)_s^t}{2\Delta t}.$$

We also define D^m as the mth successive power of D:

$$D^m w^h = D\ (D^{m-1} w^h).$$

We shall now write the numerical scheme satisfied by the mth discrete derivative of $e^{u,h}$ and $e^{v,h}$ (these sequences are implicitly extended by 0 for negative times). Note that the sequences $D^m e^{u,h}$ and $D^m e^{v,h}$ are naturally defined on a grid shifted by $m\Delta t$ (the initial grid is supposed to contain negative discrete instants), so that the fine grids differ depending whether m is odd or even. As the scheme (4.5), (4.6), and (4.7) is "invariant" under a translation by $2\Delta t$, it is easy to see that the *odd* discrete coarse derivative of the sequences $(e_c^{u,h}), (e_c^{v,h}), (e_f^{u,h})$, and $(e_f^{v,h})$, namely,

$$\left(D^{2q+1} e_c^{u,h}\right), \left(D^{2q+1} e_c^{v,h}\right), \left(D^{2q+1} e_f^{u,h}\right), \text{ and } \left(D^{2q+1} e_f^{v,h}\right),$$

satisfies a similar but different set of equations. More precisely, at the instants at which the odd discrete derivative are defined, the only change concerns the coarse grid and corresponds to the substitutions $e_{c,h}^u \leftrightarrow D^{2q+1} e_{c,h}^v$ and $e_{c,h}^v \leftrightarrow D^{2q+1} e_{c,h}^u$.

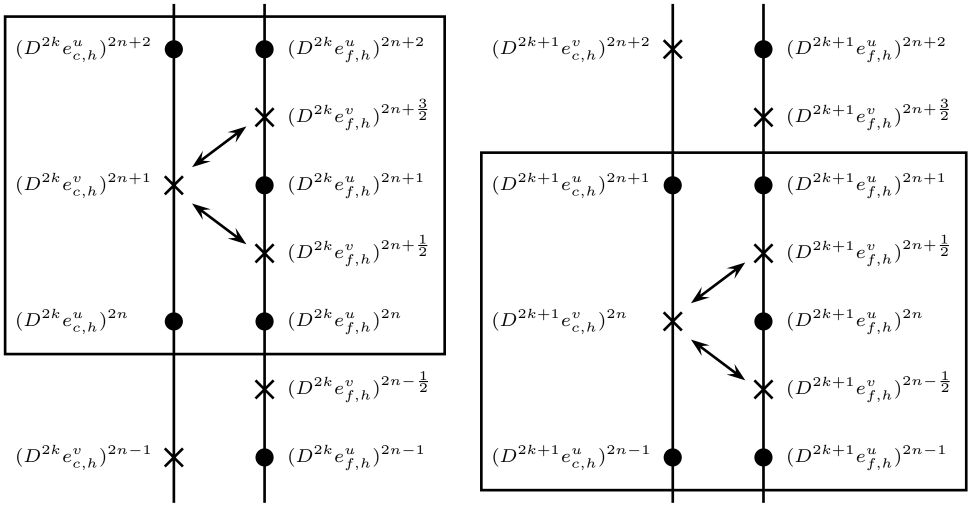

$(D^{2k}e^u_{c,h})^{2n+2}$
$(D^{2k}e^u_{f,h})^{2n+2}$
$(D^{2k}e^v_{f,h})^{2n+\frac{3}{2}}$
$(D^{2k}e^v_{c,h})^{2n+1}$
$(D^{2k}e^u_{f,h})^{2n+1}$
$(D^{2k}e^v_{f,h})^{2n+\frac{1}{2}}$
$(D^{2k}e^u_{c,h})^{2n}$
$(D^{2k}e^u_{f,h})^{2n}$
$(D^{2k}e^v_{f,h})^{2n-\frac{1}{2}}$
$(D^{2k}e^v_{c,h})^{2n-1}$
$(D^{2k}e^u_{f,h})^{2n-1}$
$(D^{2k+1}e^v_{c,h})^{2n+2}$
$(D^{2k+1}e^u_{f,h})^{2n+2}$
$(D^{2k+1}e^v_{f,h})^{2n+\frac{3}{2}}$
$(D^{2k+1}e^u_{c,h})^{2n+1}$
$(D^{2k+1}e^u_{f,h})^{2n+1}$
$(D^{2k+1}e^v_{f,h})^{2n+\frac{1}{2}}$
$(D^{2k+1}e^v_{c,h})^{2n}$
$(D^{2k+1}e^u_{f,h})^{2n}$
$(D^{2k+1}e^v_{f,h})^{2n-\frac{1}{2}}$
$(D^{2k+1}e^u_{c,h})^{2n-1}$
$(D^{2k+1}e^u_{f,h})^{2n-1}$

FIG. 4.1. *Time distribution of the unknowns.*

Other than this change, the scheme for the time intervals $[t^{2n-1}, t^{2n+1}]$ for the odd coarse discrete derivative is the same as that satisfied by $e^{u,h}$ and $e^{v,h}$ in the time intervals $[t^{2n}, t^{2n+2}]$. This is illustrated by Figure 4.1 in which the arrows represent the discrete transmission conditions. We shall find two types of schemes that are easily deduced from each other. In order to avoid repetition, it is useful to introduce

$$\overline{m} = 1 \quad \text{if } m \text{ is odd }, \qquad \overline{m} = 0 \quad \text{if } m \text{ is even.}$$

Then the equations of the scheme in a characteristic interval $[t^{2n-\overline{m}}, t^{2n+2-\overline{m}}]$ are

$$\left\{\begin{aligned} &\frac{\left(D^m e^u_c\right)^{2n+2-\overline{m}}_{2j} - \left(D^m e^u_c\right)^{2n-\overline{m}}_{2j}}{2\Delta t} + \frac{\left(D^m e^v_c\right)^{2n+1-\overline{m}}_{2j+1} - \left(D^m e^v_c\right)^{2n+1-\overline{m}}_{2j-1}}{2h} = 0,\\ &\frac{\left(D^m e^v_c\right)^{2n+1-\overline{m}}_{2j+1} - \left(D^m e^v_c\right)^{2n-1-\overline{m}}_{2j+1}}{2\Delta t} + \frac{\left(D^m e^u_c\right)^{2n-\overline{m}}_{2j+2} - \left(D^m e^u_c\right)^{2n-\overline{m}}_{2j}}{2h} \quad = 0 \end{aligned}\right. \tag{4.33}$$

on the coarse grid (i.e., for $j \leq -1$), and

$$\begin{gathered} \left\{\begin{aligned} &\frac{\left(D^m e^u_f\right)^{2n+1}_{j} - \left(D^m e^u_f\right)^{2n}_{j}}{\Delta t} + \frac{\left(D^m e^v_f\right)^{2n+\frac{1}{2}}_{j+\frac{1}{2}} - \left(D^m e^v_f\right)^{2n+\frac{1}{2}}_{j-\frac{1}{2}}}{h} = 0, \quad j \geq 1,\\ &\frac{\left(D^m e^v_f\right)^{2n+\frac{1}{2}}_{j+\frac{1}{2}} - \left(D^m e^v_f\right)^{2n-\frac{1}{2}}_{j+\frac{1}{2}}}{\Delta t} + \frac{\left(D^m e^u_f\right)^{2n}_{j+1} - \left(D^m e^u_f\right)^{2n}_{j}}{h} \quad = 0, \quad j \geq 0, \end{aligned}\right. \\ \left\{\begin{aligned} &\frac{\left(D^m e^u_f\right)^{2n+2}_{j} - \left(D^m e^u_f\right)^{2n+1}_{j}}{\Delta t} + \frac{\left(D^m e^v_f\right)^{2n+\frac{3}{2}}_{j+\frac{1}{2}} - \left(D^m e^v_f\right)^{2n+\frac{3}{2}}_{j-\frac{1}{2}}}{h} = 0, \quad j \geq 1,\\ &\frac{\left(D^m e^v_f\right)^{2n+\frac{3}{2}}_{j+\frac{1}{2}} - \left(D^m e^v_f\right)^{2n+\frac{1}{2}}_{j+\frac{1}{2}}}{\Delta t} + \frac{\left(D^m e^u_f\right)^{2n+1}_{j+1} - \left(D^m e^u_f\right)^{2n+1}_{j}}{h} = 0, \quad j \geq 0, \end{aligned}\right. \end{gathered} \tag{4.34}$$

on the fine grid. Finally the discrete transmission conditions read

$$
(4.35)\quad \begin{cases}
\big(D^m e_c^V\big)_0^{2n+1-\overline{m}} - \big(D^m e_f^V\big)_0^{2n+\frac{1}{2}-\overline{m}} = \big(D^m \varepsilon_r^v\big)^{2n+\frac{1}{2}-\overline{m}}, \\
\big(D^m e_c^V\big)_0^{2n+1-\overline{m}} - \big(D^m e_f^V\big)_0^{2n+\frac{3}{2}-\overline{m}} = \big(D^m \varepsilon_r^v\big)^{2n+\frac{3}{2}-\overline{m}}, \\
\big(D^m e_c^U\big)_0^{2n+1-\overline{m}} - \dfrac{1}{2}\Big(\big(D^m e_f^U\big)_0^{2n+\frac{1}{2}-\overline{m}} + \big(D^m e_f^U\big)_0^{2n+\frac{3}{2}-\overline{m}}\Big) = \big(D^m \varepsilon_r^u\big)^{2n+1-\overline{m}}.
\end{cases}
$$

If we assume that the exact solution (u, v) belongs to $\mathcal{C}_{a,T}^{m+1}$, using a Taylor–Lagrange expansion, it is easy to show that the truncation errors appearing on the right-hand side of (4.35) satisfy (this is the analog of (4.12) of Lemma 4.3)

$$
(4.36)\quad \sup_{t^{2n}\le T} \big|\big(D^m \varepsilon_r^u\big)^{2n+1}\big| \le \mathcal{C}\ \alpha\ h\ \|u\|_{\mathcal{C}_{a,T}^{m+1}}, \quad \sup_{t^{2n}\le T} \big|\big(D^m \varepsilon_r^v\big)^{n+\frac{1}{2}}\big| \le \mathcal{C}\ h\ \|(u,v)\|_{\mathcal{C}_{a,T}^{m+1}}.
$$

If we define the discrete $L^\infty(0, T; L^2)$ norms of the discrete coarse derivatives as

$$
(4.37)\quad \left|\begin{aligned}
&\|D^m e^{u,h}\|_{\infty,2,T}^* = \sup_{t^{2n+\frac{3}{2}}\le T} \Big(\ \big\|\big(D^m e_{c,h}^u\big)^{2n-\overline{m}}\big\| + \big\|\big(D^m e_{f,h}^u\big)^{2n-\overline{m}}\big\|\Big) \\
&\|D^m e^{u,h}\|_{\infty,2,T} = \|D^m e^{u,h}\|_{\infty,2,T}^* + \sup_{t^{2n+\frac{3}{2}}\le T} \ \big\|\big(D^m e_{f,h}^u\big)^{2n+1-\overline{m}}\big\| \\
&\|D^m e^{v,h}\|_{\infty,2,T} = \sup_{t^{2n+\frac{3}{2}}\le T} \Big(\ \big\|\big(D^m e_{c,h}^v\big)^{2n+1-\overline{m}}\big\| \\
&\qquad\qquad + \big\|\big(D^m e_{f,h}^v\big)^{2n+\frac{1}{2}-\overline{m}}\big\| + \big\|\big(D^m e_{f,h}^v\big)^{2n+\frac{3}{2}-\overline{m}}\big\|\ \Big),
\end{aligned}\right.
$$

we can apply the proof of section 4.3 to prove the following lemma.

LEMMA 4.6. *If the solution of the continuous problem* (1.1) *belongs to* $\mathcal{C}_{a,T}^{m+1}$ *for* $a > 0$*, then*

$$
(4.38)\quad \|D^m e^{u,h}\|_{\infty,2,T} + \|D^m e^{v,h}\|_{\infty,2,T} \le \mathcal{C}\ (1-\alpha^2)^{-1}\ T\ h^{\frac{1}{2}}\ \big\|\big(\partial_t^m u, \partial_t^m v\big)\big\|_{\mathcal{C}_{a,T}^1}.
$$

4.4.2. The bootstrap argument.

Step 1*: Derivation of an* $\mathcal{O}(h)$ *estimate.* In order to improve estimate (4.10), we note that equations (4.11) can be rewritten as

$$
(4.39)\quad \begin{cases}
\big(e_f^V\big)_0^{2n+\frac{3}{2}} - \big(e_f^V\big)_0^{2n+\frac{1}{2}} = \big(\varepsilon_r^v\big)^{2n+\frac{1}{2}} - \big(\varepsilon_r^v\big)^{2n+\frac{3}{2}} = \mathcal{O}(\Delta t) \\
2(e_c^V)_0^{2n+1} - \Big(\big(e_f^V\big)_0^{2n+\frac{3}{2}} + \big(e_f^V\big)_0^{2n+\frac{1}{2}}\Big) = \big(\varepsilon_r^v\big)^{2n+\frac{1}{2}} + \big(\varepsilon_r^v\big)^{2n+\frac{3}{2}} = \mathcal{O}(\Delta t^2) \\
\big(e_c^U\big)_0^{2n+1} - \dfrac{1}{2}\Big(\big(e_f^U\big)_0^{2n+\frac{1}{2}} + \big(e_f^U\big)_0^{2n+\frac{3}{2}}\Big) = \big(\varepsilon_r^u\big)^{2n+1} = \mathcal{O}(\Delta t^2).
\end{cases}
$$

Now, we reorganize the right-hand side of (4.24) in a clever way so that the left-hand sides of (4.39) appear as (see Remark 4.2)

$$
(4.40)\quad \left|\begin{aligned}
\frac{\mathcal{E}^{2n+2} - \mathcal{E}^{2n}}{2\Delta t} = &- \frac{1}{2}\ \big(\varepsilon_r^u\big)^{2n+1}\Big(\big(e_f^V\big)_0^{2n+\frac{1}{2}} + \big(e_f^V\big)_0^{2n+\frac{3}{2}}\Big) \\
&- \frac{1}{2}\ \Big(\big(\varepsilon_r^v\big)^{2n+\frac{1}{2}} + \big(\varepsilon_r^v\big)^{2n+\frac{3}{2}}\Big)\big(e_c^U\big)_0^{2n+1} \\
&+ \frac{1}{4}\ \Big(\big(\varepsilon_r^v\big)^{2n+\frac{3}{2}} - \big(\varepsilon_r^v\big)^{2n+\frac{1}{2}}\Big)\Big(\big(e_f^U\big)_0^{2n+\frac{3}{2}} - \big(e_f^U\big)_0^{2n+\frac{1}{2}}\Big).
\end{aligned}\right.
$$

Let us analyze the equality in detail. In order to bound the terms appearing on the right-hand side by an L^2 norm of the error, we first point out that from (4.3) and (4.4) it follows that

$$\left\{\begin{aligned} \big(e_f^V\big)_0^{2n+\frac{1}{2}} + \big(e_f^V\big)_0^{2n+\frac{3}{2}} &= \big(e_f^v\big)_{\frac{1}{2}}^{2n+\frac{1}{2}} + \big(e_f^v\big)_{\frac{1}{2}}^{2n+\frac{3}{2}} + h\frac{\big(e_f^u\big)_0^{2n+2} - \big(e_f^u\big)_0^{2n}}{2\Delta t},\\ \big(e_c^U\big)_0^{2n+1} &= \frac{1}{2}\left(\big(e_c^u\big)_0^{2n+2} + \big(e_c^u\big)_0^{2n}\right),\\ \big(e_f^U\big)_0^{2n+\frac{3}{2}} - \big(e_f^U\big)_0^{2n+\frac{1}{2}} &= \frac{1}{2}\left(\big(e_f^u\big)_0^{2n+2} - \big(e_f^u\big)_0^{2n}\right). \end{aligned}\right. \tag{4.41}$$

It is important to note that the quantity $(e_f^u)_0^{2n+1}$ does not appear in expression (4.40). This fact allows us to work with the norm $\|e^{u,h}\|^*_{\infty,2,T}$, which uses only the even time steps of the error of u. This will permit us to use only inequalities (4.15) and (4.16) of Lemma 4.4. Obtaining an estimate using the norm-star yields a similar estimate for the other norm because of the following lemma.

LEMMA 4.7. *Assume the hypothesis of Lemma* 4.4. *Then*

$$\sup_{t^{2n+\frac{3}{2}}\le T} \|\big(e_{f,h}^u\big)^{2n+1}\|_{\infty,2,T} \le \mathcal{C}\left(\|e^{u,h}\|^*_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T} + h^{\frac{5}{2}}\|(u,v)\|_{\mathcal{C}^2_{a,T}}\right). \tag{4.42}$$

Proof. Using the first equality of (4.6) for $j \ge 1$ we get

$$\big(e_f^u\big)_j^{2n+1} = \big(e_f^u\big)_j^{2n} - \alpha\left(\big(e_f^v\big)_{j+\frac{1}{2}}^{2n+\frac{1}{2}} - \big(e_f^v\big)_{j-\frac{1}{2}}^{2n+\frac{1}{2}}\right),$$

and in this way we also get

$$\|\big(e_{f,h}^u\big)^{2n+1}\|^2 \le \mathcal{C}\left(\|\big(e_{f,h}^u\big)^{2n}\|^2 + \|\big(e_{f,h}^v\big)^{2n+\frac{1}{2}}\|^2\right) + \frac{h}{2}|\big(e_f^u\big)_0^{2n+1}|^2. \tag{4.43}$$

In order to estimate the last term we use the last equation in (4.7) to obtain

$$\big(e_f^u\big)_0^{2n+1} = -2\big(\varepsilon_r^u\big)^{2n+1} + \big(e_c^u\big)_0^{2n} + \big(e_c^u\big)_0^{2n+2} - \left(\big(e_f^u\big)_0^{2n} + \big(e_f^u\big)_0^{2n+2}\right)/2.$$

Thus it is clear that

$$\begin{aligned}\frac{h}{2}|\big(e_f^u\big)_0^{2n+1}|^2 \le \mathcal{C}\Big(&\|\big(e_{f,h}^u\big)^{2n}\|^2 + \|\big(e_{f,h}^u\big)^{2n+2}\|^2 + \|\big(e_{c,h}^u\big)^{2n}\|^2\\ &+ \|\big(e_{c,h}^u\big)^{2n+2}\|^2 + h|\big(\varepsilon_r^u\big)^{2n+1}|^2\Big).\end{aligned}$$

Introducing this inequality in (4.43) and using (4.13) of Lemma 4.3 we obtain

$$\|\big(e_{f,h}^u\big)^{2n+1}\|^2 \le \mathcal{C}\left(\|e^{u,h}\|^{*\,2}_{\infty,2,T} + \|e^{v,h}\|^2_{\infty,2,T} + h^5\|(u,v)\|^2_{\mathcal{C}^2_{a,T}}\right),$$

which easily implies (4.42). □

Returning to (4.40), we use successively the trace Lemma 4.5 and the inequalities

(4.13) of Lemma 4.3 to obtain

$$\left|\begin{aligned}&\frac{1}{2}\left|\big(e_f^V\big)_0^{2n+\frac{1}{2}}+\big(e_f^V\big)_0^{2n+\frac{3}{2}}\right|\left|\big(\varepsilon_r^u\big)^{2n+1}\right|+\frac{1}{2}\left|\big(e_c^U\big)_0^{2n+1}\right|\left|\big(\varepsilon_r^v\big)^{2n+\frac{1}{2}}+\big(\varepsilon_r^v\big)^{2n+\frac{3}{2}}\right|\\&\qquad\le\frac{\mathcal{C}}{\alpha\sqrt{h}}\left(\|e^{u,h}\|^*_{\infty,2,T}+\alpha\|e^{v,h}\|_{\infty,2,T}\right)|\big(\varepsilon_r^u\big)^{2n+1}|\\&\qquad\quad+\frac{\mathcal{C}}{\sqrt{h}}\left(|\big(\varepsilon_r^v\big)^{2n+\frac{1}{2}}+\|e^{u,h}\|^*_{\infty,2,T}\big(\varepsilon_r^v\big)^{2n+\frac{3}{2}}|\right)\|e^{u,h}\|^*_{\infty,2,T}\\&\qquad\le\mathcal{C}\;h^{\frac{3}{2}}\;\|(u,v)\|_{\mathcal{C}^2_{a,T}}\left(\|e^{u,h}\|^*_{\infty,2,T}+\|e^{v,h}\|_{\infty,2,T}\right).\end{aligned}\right. \tag{4.44}$$

The third term of the right-hand side of (4.40) is more complicated to treat, and this is where we need the result on *coarse discrete derivatives.* Indeed, we can write, using estimate (4.12) of Lemma 4.3 and the discrete trace Lemma 4.5,

$$\left|\begin{aligned}&\frac{1}{4}|\big(\varepsilon_r^v\big)^{2n+\frac{1}{2}}-\big(\varepsilon_r^v\big)^{2n+\frac{3}{2}}|\;|\big(e_f^U\big)_0^{2n+\frac{3}{2}}-\big(e_f^U\big)_0^{2n+\frac{1}{2}}|\\&\qquad=\frac{\Delta t}{4}|\big(\varepsilon_r^v\big)^{2n+\frac{1}{2}}-\big(\varepsilon_r^v\big)^{2n+\frac{3}{2}}|\;|\big(De_f^u\big)_0^{2n+1}|\\&\qquad\le\;\mathcal{C}\;h^{\frac{3}{2}}\;\|(u,v)\|_{\mathcal{C}^1_{a,T}}\;\|De^{u,h}\|^*_{\infty,2,T}.\end{aligned}\right. \tag{4.45}$$

Substituting (4.44) and (4.45) into (4.40) and using inequalities (4.10) and (4.38) (for $m=1$) from Lemmas 4.2 and 4.6, we finally obtain

$$\frac{\mathcal{E}^{2n+2}-\mathcal{E}^{2n}}{2\Delta t}\le\;\mathcal{C}\;(1-\alpha^2)^{-1}\;T\;h^2\;\|(u,v)\|_{\mathcal{C}^2_{a,T}}\;\|(u,v)\|_{\mathcal{C}^1_{a,T}}. \tag{4.46}$$

As a consequence, after summation over n, we have

$$\mathcal{E}^{2n}\le\mathcal{C}\;(1-\alpha^2)^{-1}\;T^2\;h^2\;\|(u,v)\|_{\mathcal{C}^2_{a,T}}\;\|(u,v)\|_{\mathcal{C}^1_{a,T}},$$

and similar computations to those of the previous section lead to

$$\|e^{u,h}\|^*_{\infty,h,T}+\|e^{v,h}\|_{\infty,h,T}\le\mathcal{C}_1\;(1-\alpha^2)^{-1}\;T\;h\;\|(u,v)\|^{\frac{1}{2}}_{\mathcal{C}^2_{a,T}}\;\|(u,v)\|^{\frac{1}{2}}_{\mathcal{C}^1_{a,T}}. \tag{4.47}$$

That is, using (4.47), (4.42), and Proposition 4.1, we obtain estimate (3.19) for $k=1$; i.e., the scheme is of order h.

To initiate the recurrence that will be the object of step 2 of the proof, we shall also need similar estimates for the successive discrete coarse derivatives of the error $(e^{u,h},e^{v,h})$ (where we assume more regularity for the exact solution). Such estimates are easily obtained along the same lines as (4.47). Clearly, if the exact solution belongs to $\mathcal{C}^{2+m}_{a,T}$, we have that

$$\begin{aligned}&\sup_{t^{2n}\le T}|\big(D^m\varepsilon_r^u\big)^{2n+1-\overline{m}}|\;\le\;\mathcal{C}\;\alpha^2\;h^2\;\|\partial_t^k u\|_{\mathcal{C}^2_{a,T}},\\&\sup_{t^{2n}\le T}|\big(D^m\varepsilon_r^v\big)^{2n+\frac{1}{2}-\overline{m}}+\big(D^m\varepsilon_r^v\big)^{2n+\frac{3}{2}-\overline{m}}|\;\le\;\mathcal{C}\;h^2\;\|\big(\partial_t^k u,\partial_t^k v\big)\|_{\mathcal{C}^2_{a,T}},\end{aligned} \tag{4.48}$$

and so we can apply in the case of the mth discrete coarse derivatives $(D^m e^{u,h},D^m e^{v,h})$ a proof similar to that used for proving (4.47) and obtain

$$\begin{aligned}&\|D^m e^{u,h}\|^*_{\infty,2,T}+\|D^m e^{v,h}\|_{\infty,2,T}\le\mathcal{C}\;(1-\alpha^2)^{-1}\;T\;h\\&\qquad\qquad\times\|\big(\partial_t^m u,\partial_t^m v\big)\|^{\frac{1}{2}}_{\mathcal{C}^2_{a,T}}\|\big(\partial_t^m u,\partial_t^m v\big)\|^{\frac{1}{2}}_{\mathcal{C}^1_{a,T}}.\end{aligned} \tag{4.49}$$

Step 2: The recurrence proof. Assume the following by induction.

Assumption $\mathcal{R}_k$. If (u,v) belongs to $\mathcal{C}^{m+k+1}_{a,T}$, for $m \geq 0$,

$$\left|\begin{aligned}&\|D^m e^{u,h}\|^*_{\infty,h,T} + \|D^m e^{v,h}\|_{\infty,h,T} \leq \mathcal{C}_k\ (1-\alpha^2)^{-1}\ T\ h^{p_k}\\ &\qquad\qquad\qquad \times \|(\partial_t^m u, \partial_t^m v)\|^{\frac{1}{2^k}}_{\mathcal{C}^k_{a,T}} \prod_{j=1}^{k} \|(\partial_t^m u, \partial_t^m v)\|^{\frac{1}{2^j}}_{\mathcal{C}^{j+1}_{a,T}}.\end{aligned}\right.$$

In what follows, the constant $\mathcal{C}$ should change from one line to another, but it is always independent of k. From (4.40), using (4.44) and (4.45) we obtain

$$\left|\begin{aligned}&\frac{\mathcal{E}^{2n+2}-\mathcal{E}^{2n}}{2\Delta t} \leq \mathcal{C}\ h^{\frac{3}{2}}\ \|(u,v)\|_{\mathcal{C}^2_{a,T}}\ \Big(\ \|e^{u,h}\|^*_{\infty,h,T} + \|e^{v,h}\|_{\infty,h,T}\Big)\\ &\qquad\qquad + \mathcal{C}\ h^{\frac{3}{2}} \|(u,v)\|_{\mathcal{C}^1_{a,T}}\ \Big(\ \|De^{u,h}\|^*_{\infty,h,T} + \|De^{v,h}\|_{\infty,h,T}\Big).\end{aligned}\right. \tag{4.50}$$

We assume that $(u,v) \in \mathcal{C}^{k+2}_{a,T}$. Using Assumption $\mathcal{R}_k$ for $m=0$ and for $m=1$, we obtain

$$\left|\begin{aligned}\frac{\mathcal{E}^{2n+2}-\mathcal{E}^{2n}}{2\Delta t} \leq \mathcal{C}\ \mathcal{C}_k\ (1-\alpha^2)^{-1}\ T\ h^{p_k+\frac{3}{2}} \Bigg(&\|(u,v)\|_{\mathcal{C}^2_{a,T}}\ \|(u,v)\|^{\frac{1}{2^k}}_{\mathcal{C}^k_{a,T}} \prod_{j=1}^{k} \|(u,v)\|^{\frac{1}{2^j}}_{\mathcal{C}^{j+1}_{a,T}}\\ &+ \|(u,v)\|_{\mathcal{C}^1_{a,T}}\ \|(u,v)\|^{\frac{1}{2^k}}_{\mathcal{C}^{k+1}_{a,T}} \prod_{j=1}^{k} \|(u,v)\|^{\frac{1}{2^j}}_{\mathcal{C}^{j+2}_{a,T}}\Bigg).\end{aligned}\right.$$

From the inequalities

$$\|(u,v)\|_{\mathcal{C}^2_{a,T}} \geq \|(u,v)\|_{\mathcal{C}^1_{a,T}},\ \|(u,v)\|_{\mathcal{C}^k_{a,T}} \leq \|(u,v)\|_{\mathcal{C}^{k+1}_{a,T}} \text{ and } \|(u,v)\|_{\mathcal{C}^{j+1}_{a,T}} \leq \|(u,v)\|_{\mathcal{C}^{j+2}_{a,T}}\ ,$$

we deduce that

$$\frac{\mathcal{E}^{2n+2}-\mathcal{E}^{2n}}{2\Delta t} \leq \mathcal{C}\ \mathcal{C}_k\ (1-\alpha^2)^{-1}\ T\ h^{p_k+\frac{3}{2}} \left(\ \|(u,v)\|_{\mathcal{C}^2_{a,T}}\ \|(u,v)\|^{\frac{1}{2^k}}_{\mathcal{C}^{k+1}_{a,T}} \prod_{j=1}^{k} \|(u,v)\|^{\frac{1}{2^j}}_{\mathcal{C}^{j+2}_{a,T}}\right).$$

Summing over n, we obtain after some manipulations (including a shift of index in the product)

$$\mathcal{E}^{2n} \leq \mathcal{C}\ \mathcal{C}_k\ (1-\alpha^2)^{-1}\ T^2\ h^{p_k+\frac{3}{2}} \left(\ \|(u,v)\|^{\frac{1}{2^k}}_{\mathcal{C}^{k+1}_{a,T}} \prod_{j=1}^{k+1} \|(u,v)\|^{\frac{1}{2^{j-1}}}_{\mathcal{C}^{j+1}_{a,T}}\right).$$

To conclude, it suffices to use once again (4.15) and (4.16) of Lemma 4.4, which gives

$$\|e^{u,h}\|^*_{\infty,h,T} + \|e^{v,h}\|_{\infty,h,T} \leq (\mathcal{C}\ \mathcal{C}_k)^{\frac{1}{2}}\ (1-\alpha^2)^{-1}\ T\ h^{\frac{p_k}{2}+\frac{3}{4}} \left(\ \|(u,v)\|^{\frac{1}{2^{k+1}}}_{\mathcal{C}^{k+1}_{a,T}} \prod_{j=1}^{k+1} \|(u,v)\|^{\frac{1}{2^j}}_{\mathcal{C}^{j+1}_{a,T}}\right),$$

which is what we wanted to prove with

$$p_{k+1} = \frac{p_k}{2} + \frac{3}{4}, \quad C_{k+1} = (\mathcal{C}\ \mathcal{C}_k)^{\frac{1}{2}}.$$

In order to finish the recurrence proof we should obtain similar estimates for the mth discrete coarse derivative of the error assuming more regularity of the solution. Suppose that $(u,v) \in \mathcal{C}^{m+k+2}_{a,T}$. Using similar techniques to those used in the present section (and using $\mathcal{R}_k$ with m and $m+1$) it is easy to show that

$$\begin{aligned}\|D^m e^{u,h}\|^*_{\infty,h,T} + \|D^m e^{v,h}\|_{\infty,h,T} \leq \mathcal{C}_k \ (1-\alpha^2)^{-1} \ T \ h^{p_{k+1}} \\ \times \|\big(\partial_t^m u, \partial_t^m v\big)\|^{\frac{1}{2^{k+1}}}_{\mathcal{C}^{k+1}_{a,T}} \prod_{j=1}^{k+1} \|\big(\partial_t^m u, \partial_t^m v\big)\|^{\frac{1}{2^j}}_{\mathcal{C}^{j+1}_{a,T}}.\end{aligned}$$

Since the result of Step 1, namely, estimates (4.47) and (4.49), is nothing but Assumption $\mathcal{R}_1$, we see easily that

$$p_k = \frac{3}{2} - \frac{1}{2^k}, \quad \mathcal{C}_k = \mathcal{C} \ \left(\mathcal{C}_1/\mathcal{C}\right)^{\frac{1}{2^{k-1}}} .$$

As the sequence $\mathcal{C}_k$ is convergent, it is in particular bounded. Finally we obtain the estimate

$$\|e^{u,h}\|^*_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T} \leq \mathcal{C} \ (1-\alpha^2)^{-1} \ T \ h^{\left(\frac{3}{2}-\frac{1}{2^k}\right)} \ |||(u,v)|||_{\mathcal{C}^k_{a,T}} \tag{4.51}$$

and using Lemma 4.7 we complete the proof of the following lemma.

Lemma 4.8. *If the solution of the continuous problem* (1.1) *belongs to* $\mathcal{C}^k_{a,T}$ *for* $a > 0$*, then*

$$\|e^{u,h}\|_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T} \leq \mathcal{C} \ (1-\alpha^2)^{-1} \ \ T \ h^{\left(\frac{3}{2}-\frac{1}{2^k}\right)} \ |||(u,v)|||_{\mathcal{C}^k_{a,T}} + \mathcal{C} \ h^{\frac{5}{2}} \ \|(u,v)\|_{\mathcal{C}^2_{a,T}}. \tag{4.52}$$

Step 3*: Case* $(u,v) \in \mathbf{C}^{\infty}_{a,T}$. Under this hypothesis, and using the previous part of the proof, we have that

$$\begin{aligned}\|e^{u,h}\|^*_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T} &\leq \mathcal{C} \ (1-\alpha^2)^{-1} \ T \ h^{\left(\frac{3}{2}-\frac{1}{2^k}\right)} \ |||(u,v)|||_{\mathcal{C}^k_{a,T}} \\ &\leq \mathcal{C} \ (1-\alpha^2)^{-1} \ T \ h^{\left(\frac{3}{2}-\frac{1}{2^k}\right)} \ |||(u,v)|||_{\mathbf{C}^{\infty}_{a,T}}\end{aligned}$$

for all $k \in \mathbb{N}$. Passing to the limit when $k \to +\infty$ and using Lemma 4.7, we get the following.

Lemma 4.9. *If the solution of the continuous problem* (1.1) *belongs to* $\mathbf{C}^{\infty}_{a,T}$ *for* $a > 0$*, then*

$$\|e^{u,h}\|_{\infty,2,T} + \|e^{v,h}\|_{\infty,2,T} \leq \mathcal{C} \ (1-\alpha^2)^{-1} \ T \ h^{\frac{3}{2}} \ |||(u,v)|||_{\mathbf{C}^{\infty}_{a,T}} + \mathcal{C} \ h^{\frac{5}{2}} \ \|(u,v)\|_{\mathcal{C}^2_{a,T}}. \tag{4.53}$$

Remark 4.2. To see how (4.40) can be derived, we start from the two following identities:

$$\begin{aligned}\big(e_f^U\big)_0^{2n+1\pm\frac{1}{2}} \big(e_f^V\big)_0^{2n+1\pm\frac{1}{2}} = \big(e_f^V\big)_0^{2n+1\pm\frac{1}{2}} &\left[\frac{\big(e_f^U\big)_0^{2n+\frac{1}{2}} + \big(e_f^U\big)_0^{2n+\frac{3}{2}}}{2} - \big(e_c^U\big)_0^{2n+1}\right] \\ + \big(e_f^V\big)_0^{2n+1\pm\frac{1}{2}} &\left[\big(e_c^U\big)_0^{2n+1} \pm \frac{\big(e_f^U\big)_0^{2n+\frac{3}{2}} - \big(e_f^U\big)_0^{2n+\frac{1}{2}}}{2}\right].\end{aligned}$$

Adding and identifying the consistency errors we obtain

$$\begin{aligned}\frac{1}{2}\left(\left(e_f^U\right)_0^{2n+\frac{1}{2}}\left(e_f^V\right)_0^{2n+\frac{1}{2}}+\left(e_f^U\right)_0^{2n+\frac{3}{2}}\left(e_f^V\right)_0^{2n+\frac{3}{2}}\right)=&-\frac{1}{2}(\varepsilon_r^u)^{2n+1}\left(\left(e_f^V\right)_0^{2n+\frac{1}{2}}+\left(e_f^V\right)_0^{2n+\frac{3}{2}}\right)\\&+\frac{1}{2}\left(e_c^U\right)_0^{2n+1}\left(\left(e_f^V\right)_0^{2n+\frac{3}{2}}+\left(e_f^V\right)_0^{2n+\frac{1}{2}}\right)\\&+\frac{1}{4}\left(\left(\varepsilon_r^v\right)^{2n+\frac{3}{2}}-\left(\varepsilon_r^v\right)^{2n+\frac{1}{2}}\right)\left(\left(e_f^U\right)_0^{2n+\frac{3}{2}}-\left(e_f^U\right)_0^{2n+\frac{1}{2}}\right).\end{aligned}$$

Finally subtracting the term $(e_c^U)_0^{2n+1}(e_c^V)_0^{2n+1}$ we obtain the desired expression.

Remark 4.3. Let us explain why we cannot use the same proof for scheme II. Following the same steps as for scheme I we rewrite the transmission conditions (4.8) as

$$(4.54)\quad\begin{cases}\left(e_f^U\right)_0^{2n+\frac{3}{2}}-\left(e_f^U\right)_0^{2n+\frac{1}{2}}=(\tilde{\varepsilon}_r^u)^{2n+\frac{1}{2}}-\left(\tilde{\varepsilon}_r^u\right)^{2n+\frac{3}{2}}=\mathcal{O}(\Delta t),\\ 2\left(e_c^U\right)_0^{2n+1}-\left(\left(e_f^U\right)_0^{2n+\frac{3}{2}}+\left(e_f^U\right)_0^{2n+\frac{1}{2}}\right)=\left(\tilde{\varepsilon}_r^u\right)^{2n+\frac{1}{2}}+\left(\tilde{\varepsilon}_r^u\right)^{2n+\frac{3}{2}}=\mathcal{O}(\Delta t^2),\\ \left(e_c^V\right)_0^{2n+1}-\frac{1}{2}\left(\left(e_f^V\right)_0^{2n+\frac{1}{2}}+\left(e_f^V\right)_0^{2n+\frac{3}{2}}\right)=\left(\tilde{\varepsilon}_r^v\right)^{2n+1}=\mathcal{O}(\Delta t^2)\end{cases}$$

in order to have two consistency errors of order two and only one of first order. We rewrite (4.32) using this last truncation errors obtaining

$$(4.55)\quad\left|\begin{aligned}\frac{1}{2\Delta t}\left(\mathcal{E}^{2n+2}-\mathcal{E}^{2n}\right)=&-\frac{1}{2}\left(\left(e_f^U\right)_0^{2n+\frac{1}{2}}+\left(e_f^U\right)_0^{2n+\frac{3}{2}}\right)\left(\widetilde{\varepsilon}_r^v\right)^{2n+1}\\&-\frac{1}{2}\left(e_c^V\right)_0^{2n+1}\left(\left(\widetilde{\varepsilon}_r^u\right)^{2n+\frac{1}{2}}+\left(\widetilde{\varepsilon}_r^u\right)^{2n+\frac{3}{2}}\right)\\&+\frac{1}{4}\left(\left(\widetilde{\varepsilon}_r^u\right)^{2n+\frac{1}{2}}-\left(\widetilde{\varepsilon}_r^u\right)^{2n+\frac{3}{2}}\right)\left(\left(e_f^V\right)_0^{2n+\frac{3}{2}}-\left(e_f^V\right)_0^{2n+\frac{1}{2}}\right),\end{aligned}\right.$$

which is an analog of (4.40) for scheme II. The term that is most complicated to treat is, as for scheme I, the last one. However, its expression in this case is not as easy as for scheme I because

$$(e_f^V)_0^{2n+\frac{3}{2}}-(e_f^V)_0^{2n+\frac{1}{2}}=(e_f^v)_{\frac{1}{2}}^{n+\frac{3}{2}}-(e_f^v)_{\frac{1}{2}}^{n+\frac{1}{2}}+\frac{h}{2}\frac{(e_f^u)_0^{2n+2}-2(e_f^u)_0^{2n+1}+(e_f^u)_0^{2n}}{\Delta t}.$$

In fact we cannot use similar arguments as before to estimate the last term. We recall that for scheme I we had an simpler expression (third equation of (4.41)).

5. Comparison between theory and numerics.

5. Comparison between theory and numerics. The results obtained in section 3 are compared in subsection 5.2 to those obtained in [12] using the Fourier technique (see subsection 5.1 for a brief recap of these results) and with some numerical results in subsection 5.3.

5.1. Fourier analysis results. This study is based on the behavior of plane wave solutions in the presence of a space-time mesh refinement. More precisely, we study the reflection and transmission through the artificial interface between the coarse grid and the fine grid of an incident wave in Ω_c, of amplitude 1 and frequency ω. One reflected wave and one transmitted wave each of frequency ω and of amplitude R_c (reflection coefficient) and T_c (transmission coefficient), respectively, are generated in the coarse and fine grids, respectively. Due to the aliasing phenomena (namely,

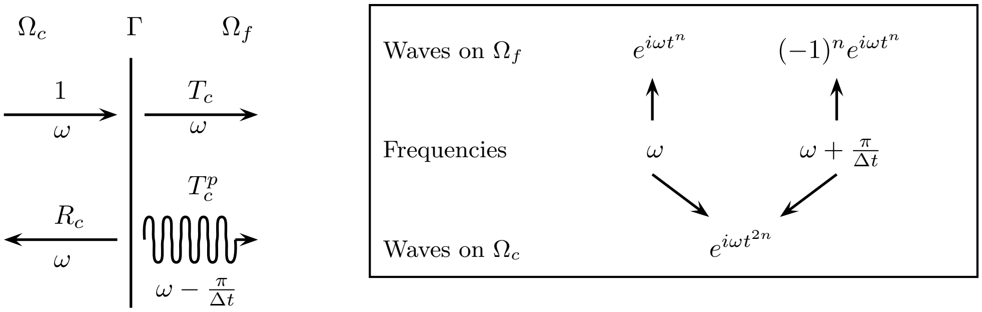


FIG. 5.1. *Schematic representation of the aliasing phenomena.*

that the frequencies ω and $\omega + \frac{\pi}{\Delta t}$ coincide in the coarse grid but are distinct in the fine grid—see Figure 5.1) a parasitic transmitted wave of frequency $\omega + \frac{\pi}{\Delta t}$ is generated in the fine grid. If ωh is small enough and α is less than 1, this parasitic wave is highly oscillatory in space (with space frequency $\frac{\pi}{h}$) and evanescent (i.e., decaying exponentially with distance from the interface) with a penetration depth $l(h, \alpha)$ which satisfies

$$l(h, \alpha) = \frac{h}{2\text{argch}(1/\alpha)} + \mathcal{O}(h^3). \tag{5.1}$$

The amplitude of this parasitic wave at the interface is given by a coefficient T_c^p, the parasitic transmission coefficient. The particular solution one looks for is thus given by the following expressions for u (the wave numbers k_c and k_f appearing in the formula below depend on α and h and are deduced from the dispersion relation on each grids—see [12] for more details):

$$\begin{cases} (u_c)_{2j}^{2n} = e^{i(k_c x_{2j} - \omega t^{2n})} + R_c \; e^{i(-k_c x_{2j} - \omega t^{2n})}, & j \leq 0, \\ (u_f)_j^n = T_c \; e^{i(k_f x_j - \omega t^n)} + T_c^p \; (-1)^{j+n} \; e^{-i\omega t^n} \; e^{-\frac{x_j}{l(h,\alpha)}}, & j \geq 0. \end{cases}$$

The expressions for v are similar. The unknown coefficients R_c, T_c, and T_c^p can be determined from the coupling equations (2.12) or (2.13). As the interface $x = 0$ is purely artificial, if we consider the continuous case, we should find

$$R_c = 0, \quad T_c = 1, \quad T_c^p = 0,$$

the *physical values* of the parameters. In the discrete case the coefficients R_c, T_c, and T_c^p depend only on ωh and α, and, for fixed $0 < \alpha < 1$, their Taylor expansions for small ωh are given by

$$\begin{aligned} R_c(\omega h, \alpha) &= \frac{1}{16} \big(\alpha^2 - 3\big) (\omega h)^2 + \mathcal{O}\big((\omega h)^3\big), \\ T_c(\omega h, \alpha) &= 1 - \frac{3}{16} \big(\alpha^2 + 1\big) (\omega h)^2 + \mathcal{O}\big((\omega h)^3\big), \\ T_c^p(\omega h, \alpha) &= \frac{1}{2} \frac{i\alpha^2}{\sqrt{1 - \alpha^2}} (\omega h) + \mathcal{O}\big((\omega h)^3\big) \end{aligned} \tag{5.2}$$

if scheme I is used and

$$\begin{aligned} R_c(\omega h,\alpha) &= -\frac{i\sqrt{1-\alpha^2}}{4}(\omega h)+\mathcal{O}(\omega^2h^2),\\ T_c(\omega h,\alpha) &= 1-\frac{i\sqrt{1-\alpha^2}}{4}(\omega h)+\mathcal{O}(\omega^2h^2),\\ T_c^p(\omega h,\alpha) &= -1+\frac{i\sqrt{1-\alpha^2}}{4}(\omega h)+\mathcal{O}(\omega^2h^2) \end{aligned} \tag{5.3}$$

if scheme II is used.

Remark 5.1. When one considers the limit case $\alpha = 1$, one can show that the parasitic wave becomes propagative ($l(h,\alpha)$ tends to infinity when α tends to one) and we have for both schemes

$$R_c = 0, \quad T_c = 1+\mathcal{O}(\omega^2h^2), \quad T_c^p = -1+\mathcal{O}(\omega^2h^2).$$

5.2. Comparing our results with the Fourier analysis results. Let us summarize the main information provided by the Fourier analysis. First of all we consider scheme I with $\alpha < 1$. In this case we have the following:

- The discrete reflection and transmission coefficients are second order approximations of the physical ones. The contribution to the error is of order two in the L^∞ and L^p, $p > 1$, norms.
- The L^∞ norm of the error coming from the parasitic transmitted wave is approximately (for h small enough)

$$\frac{\alpha^2\omega h}{\sqrt{1-\alpha^2}}, \tag{5.4}$$

  that is, the method should be first order accurate for this norm. Due to the exponential decay of this wave, the L^p error is approximately

$$\left(\int_0^\infty \frac{(\alpha^2\omega h)^p}{(1-\alpha^2)^{\frac{p}{2}}}\, e^{-\frac{2px\,\mathrm{argch}(\alpha^{-1})}{h}}\,\mathrm{d}x\right)^{\frac{1}{p}} \approx \frac{h^{\frac{p+1}{p}}}{(1-\alpha)^{\frac{p+1}{2p}}} \quad (\alpha \longrightarrow 1). \tag{5.5}$$

  Taking $p = 2$ we remark that the order of convergence is in conformity with that given by Theorem 3.1 and the comments following it. It seems that the Fourier techniques allow us to obtain a sharper estimate for the dependence on $1-\alpha$ when α goes to 1. The best estimate is obtained in the L^1 norm, where the error should be of order two.
- We also remark that this last error is localized at the artificial interface. In effect, computing the L^∞ and L^p, $p \geq 1$, norms on the complement of a neighborhood of the artificial interface, this error is exponentially decreasing with the space step. The method should be of order two in these norms.

Concerning scheme II with $\alpha < 1$, the Fourier analysis allows us to make the following conclusions:

- The error coming from the reflected and (not parasitic) transmitted waves is of order one in the L^∞ and L^p, $p \geq 1$, norms.
- The amplitude of the parasitic transmitted wave does not go to zero when the discretization parameters go to zero. This means that the method does

not converge in the L^∞ norm. Due to the exponential decay of this wave, a simple computation allows us to estimate its L^p error,

$$\frac{h^{\frac{1}{p}}}{\operatorname{argch}^{\frac{1}{p}}(\alpha^{-1})} \approx \frac{h^{\frac{1}{p}}}{(1-\alpha)^{\frac{1}{2p}}} \quad (\alpha \longrightarrow 1).$$

Again, the results obtained for $p = 2$ are coherent with those of Theorem 3.1 and the comments given later on. The L^1 error should be of order one. The estimate for the dependence on $1 - \alpha$ seems to be more precise with the Fourier techniques.

- As for scheme I, this last error is localized. If we consider the L^∞ and $L^p,\ p \geq 1$, norms on the complement of a neighborhood of the point $x = 0$, the method should be of first order, because the error provided by the transmitted parasitic wave is exponentially decreasing.

For $\alpha = 1$, the amplitude of the transmitted parasitic wave does not tend to zero and the wave is not evanescent. Neither method should be strongly convergent.

5.3. Numerical results. In this section we will obtain numerically the orders of convergence of schemes I and II for several norms to compare them with the theoretical ones provided by Theorem 3.1 and the Fourier analysis. We consider the 1D wave equation

$$(5.6)\quad \begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial v}{\partial x} = 0, \\ \dfrac{\partial v}{\partial t} + \dfrac{\partial u}{\partial x} = 0, \end{cases} \quad (x,t) \in \mathbb{R} \times \mathbb{R}^+, \qquad \begin{cases} u(x,t=0) = u_0(\frac{x-x_0}{L}), \\ v(x,t=0) = 0, \end{cases} \quad x \in \mathbb{R},$$

where $x_0 = -0.25$, $L = 0.25$, and

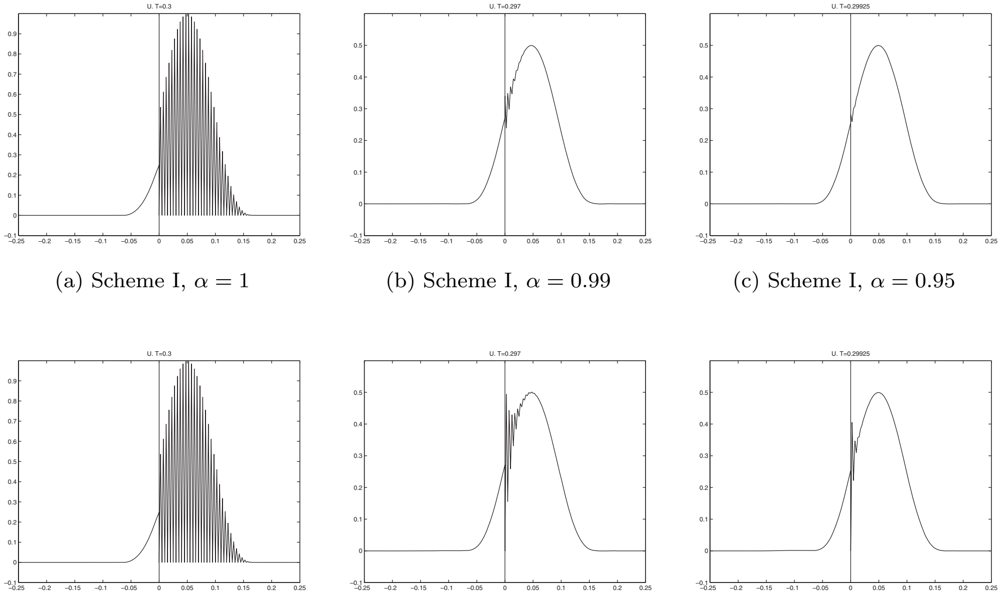
$$u_0(x) \quad = \quad \begin{cases} 256(x-1/2)^4(x+1/2)^4 & \text{if } x \in [-1/2, 1/2], \\ 0 & \text{otherwise.} \end{cases}$$

The exact solution of the problem is given by

$$u(x,t) = \frac{1}{2}u_0\left((x - x_0 - t)/L\right) + \frac{1}{2}u_0\left((x - x_0 + t)/L\right),$$

$$v(x,t) = \frac{1}{2}u_0\left((x - x_0 - t)/L\right) - \frac{1}{2}u_0\left((x - x_0 + t)/L\right).$$

The computational domain for the numerical resolution of the equations is the interval $\Omega = [-0.5, 0.5]$. We use transparent boundary conditions to simulate the unbounded domain. A space step of size h is used in $\Omega_c = [-0.5, 0]$ and of size $h/2$ in $\Omega_f = [0, 0.5]$. We recall that both schemes also depend on the parameter $\alpha = \Delta t/h$ that we must choose in the interval $(0,1)$ to ensure the stability of the method. In practice, it is interesting to choose α to be as large as possible to reduce the computational costs. The problem is that all the error and stability estimates given in section 3 blow up when α tends to 1. In Figures 5.2 and 5.3 we can note this phenomenon as well. For $\alpha = 1$ both schemes give us similar results. A high frequency wave appears when the waves cross the artificial boundary (see Figures 5.2(a) and 5.2(d)). Even so, the method seems to be L^2 stable. Taking $\alpha < 1$ most of oscillatory parasitic waves become evanescent and we obtain a good solution if we remove the behavior near



(a) Scheme I, $\alpha = 1$ (b) Scheme I, $\alpha = 0.99$ (c) Scheme I, $\alpha = 0.95$

(d) Scheme II, $\alpha = 1$ (e) Scheme II, $\alpha = 0.99$ (f) Scheme II, $\alpha = 0.95$

FIG. 5.2. *Dependence on α of u^h. $T \approx 0.3$. Zoom around $x = 0$.*

$x = 0$. The penetration depth of the transmitted parasitic wave increases as α goes to the limit value 1. As we can see in Figures 5.2(c) and 5.2(f), $\alpha = 0.95$ is sufficient to obtain a good result. We can also see that the amplitude of the parasitic wave is higher for scheme II than for scheme I (see Figure 5.3). In particular, scheme II does not converge in the L^∞ norm (see Remark 2.1).

In order to measure the error between the exact solution (u, v) and the numerical solution (u^h, v^h) we consider the discrete equivalent of the norms

$$\begin{aligned} &L_t^\infty([0,T], L_x^p(\Omega)), \qquad && L_t^\infty([0,T], L_x^\infty(\Omega)), \\ &L_t^\infty([0,T], L_x^p(\Omega^\star)), \qquad && L_t^\infty([0,T], L_x^\infty(\Omega^\star)) \end{aligned} \tag{5.7}$$

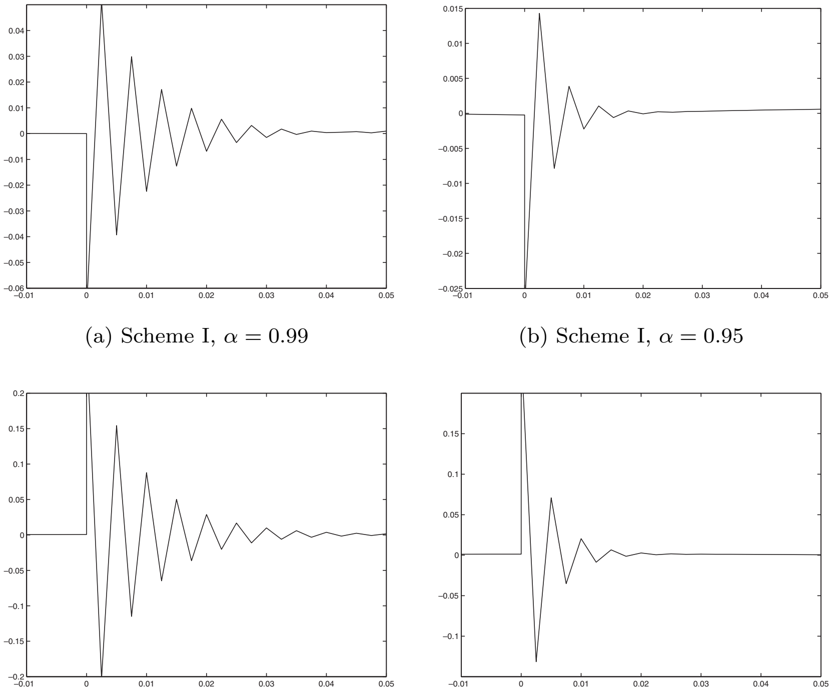
(with $\Omega^\star = \Omega \setminus [0, 0.1]$, and $p \in \mathbb{N}$), which we will denote by

$$\|e^h\|_{\infty,p,T}, \qquad \|e_h\|_{\infty,T}, \qquad \|e_h\|^\star_{\infty,p,T}, \qquad \|e_h\|^\star_{\infty,T}.$$

Let us assume that the error has approximately the form

$$\|e_h\| \approx \mathcal{C}(u, v)(1 - \alpha)^{-k_2} h^{k_1}. \tag{5.8}$$

Fixing a value of the parameter α and computing the different norms at $T = 0.5$ (where the wave has already crossed the artificial interface) for $h = 0.005$, $0.00\hat{3}$, 0.0025, and 0.002 we obtain the value of k_1 for each norm. The results for $\alpha = 0.95$, using the discrete $L_t^\infty([0,T], L_x^1(\Omega))$, $L_t^\infty([0,T], L_x^2(\Omega))$, and $L_t^\infty([0,T], L_x^\infty(\Omega))$ norms, are plotted in Figures 5.4(a), 5.5(a), and 5.6(a), respectively. We can see that the slopes are coherent with the estimates predicted in section 3 and subsection 5.2. As we pointed



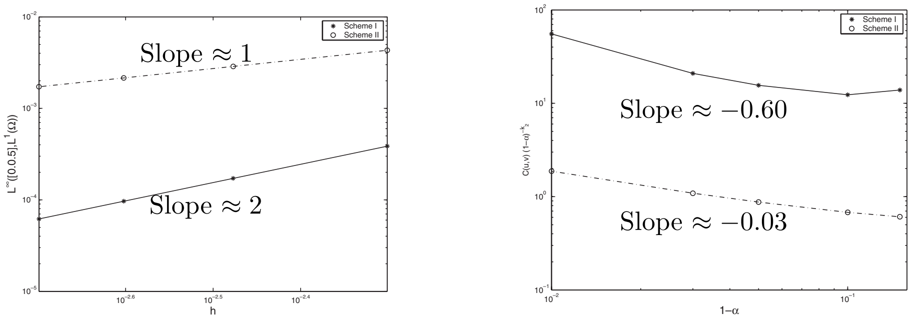
(a) Scheme I, $\alpha = 0.99$

(b) Scheme I, $\alpha = 0.95$

(c) Scheme II, $\alpha = 0.99$

(d) Scheme II, $\alpha = 0.95$

FIG. 5.3. *Dependence on α of $u^h - u$. $T \approx 0.3$. Zoom around $x = 0$.*



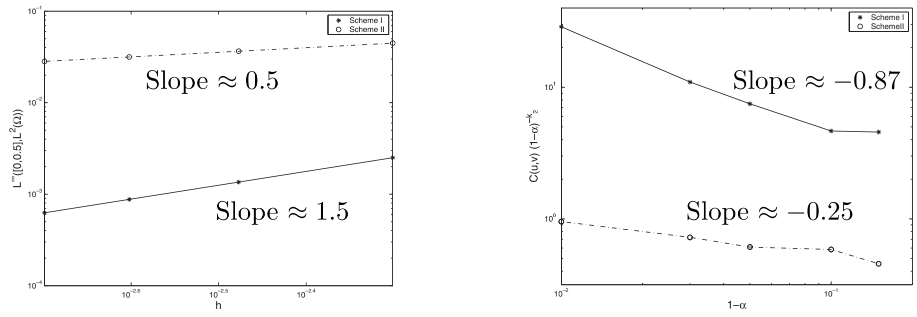
(a) Norm of the error

(b) Dependence on $1 - \alpha$

FIG. 5.4. *$L^\infty(0, T, L^1(\Omega))$ norm with $\alpha = 0.95$.*

out in section 5.2, the most important part of the error is localized at the interface as we can see in Figure 5.7, where the $L_t^\infty([0, T], L_x^2(\Omega^\star))$, and $L_t^\infty([0, T], L_x^p(\Omega^\star))$ errors have been computed.

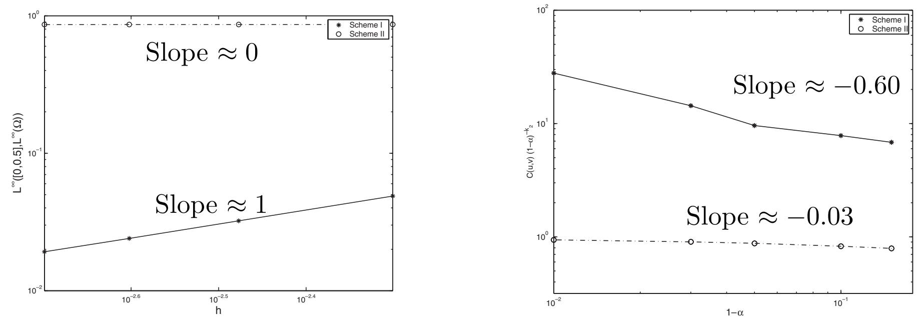
The same computations have been done with $\alpha = 0.85$, 0.9, 0.97, and 0.99 and the same orders of convergence have been obtained (see Table 5.1). This allows us to compute the dependence on α of the method when this parameter goes to 1. In Figures 5.4(b), 5.5(b), and 5.6(b), we represent the results obtained for the $L_t^\infty([0, T], L_x^1(\Omega))$, $L_t^\infty([0, T], L_x^2(\Omega))$, and $L_t^\infty([0, T], L_x^\infty(\Omega))$ norms that are in cor-



(a) Norm of the error

(b) Dependence on $1-\alpha$

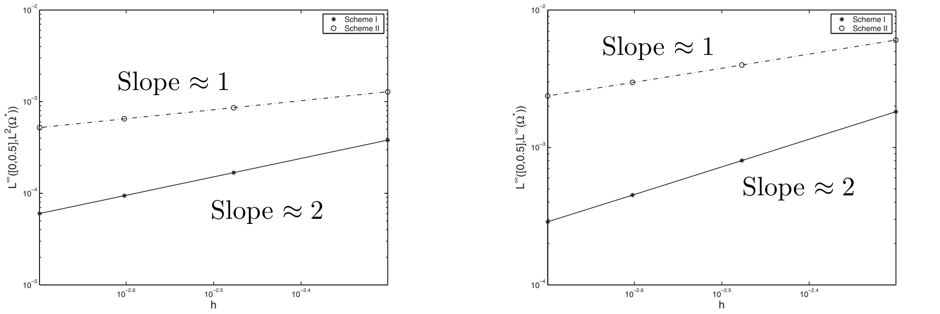
FIG. 5.5. *$L^\infty(0, T, L^2(\Omega))$ norm with $\alpha = 0.95$.*



(a) Norm of the error

(b) Dependence on $1-\alpha$

FIG. 5.6. *$L^\infty(0, T, L^\infty(\Omega))$ norm with $\alpha = 0.95$.*



(a) $L^\infty(0, T, L^2(\Omega^*))$ norm of the error

(b) $L^\infty(0, T, L^\infty(\Omega^*))$ norm of the error

FIG. 5.7. *Convergence of both schemes with $\alpha = 0.95$.*

respondence with the Fourier analysis (that gives sharper estimates) and, in consequence, also with Theorem 3.1. We have also noted that the hypotheses demanded in Theorem 3.1 concerning the smoothness of the exact solution of (1.1) are in practice too strong. We have observed the same rates of convergence for initial conditions that are $\mathcal{C}_2(\mathbb{R} \times [0, T])^2$.

Table 5.1
Observed orders of convergence.

	$\|e_h\|_{L^1}$	$\|e_h\|_{L^2}$	$\|e_h\|_{L^\infty}$	$\|e_h\|_{L^2_*}$	$\|e_h\|_{L^\infty_*}$
Scheme I	2	1.5	1	2	2
Scheme II	1	0.5	n.c.	1	1

REFERENCES

[1] A. Bamberger, R. Glowinski, and Q. H. Tran, *A domain decomposition method for the acoustic wave equation with discontinuous coefficients and grid change*, SIAM J. Numer. Anal., 34 (1997), pp. 603–639.

[2] E. Bécache, P. Joly, and J. Rodríguez, *Space-time mesh refinement for elastodynamics: Numerical results*, Comput. Methods. Appl. Mech. Engrg., 194 (2005), pp. 355–366.

[3] M. J. Berger, *Stability of interfaces with mesh refinement*, Math. Comp., 45 (1985), pp. 301–318.

[4] M. J. Berger, *On conservation at grid interfaces*, SIAM J. Numer. Anal., 24 (1987), pp. 967–984.

[5] M. J. Berger and P. Colella, *Local adaptative mesh refinement for shock hydrodynamics*, J. Comput. Phys., 82 (1989), pp. 64–84.

[6] M. J. Berger and R. J. LeVeque, *Adaptive mesh refinement using wave-propagation algorithms for hyperbolic systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2298–2316.

[7] M. J. Berger and J. Oliger, *Adaptive mesh refinement for hyperbolic partial differential equations*, J. Comput. Phys., 53 (1984), pp. 484–512.

[8] A. Buffa, F. Ben Belgacem, and Y. Maday, *The mortar finite element method for the 3D Maxwell equations: First results*, SIAM J. Numer. Anal., 39 (2001), pp. 880–901.

[9] M. W. Chevalier and R. J. Luebbers, *FDTD local grid with material traverse*, IEEE Transactions on Antennas and Propagation, 45 (1997), pp. 411–421.

[10] F. Collino, T. Fouquet, and P. Joly, *Une méthode de raffinement de maillage espace-temps pour le système de Maxwell en dimension un*, C. R. Acad. Sci. Paris, 328 (1999), pp. 263–268.

[11] F. Collino, T. Fouquet, and P. Joly, *A conservative space-time mesh refinement method for the 1-D wave equation. Part* I: *Construction*, Numer. Math., 95 (2003), pp. 197–221.

[12] F. Collino, T. Fouquet, and P. Joly, *A conservative space-time mesh refinement method for the 1-D wave equation. Part* II: *Analysis*, Numer. Math., 95 (2003), pp. 223–251.

[13] T. Fouquet, *Raffinement de maillage spatio-temporel pour les équations de Maxwell*, Ph.D. thesis, Université Paris IX Dauphine, 2000.

[14] M. Gander, L. Halpern, and F. Nataf, *Optimal Schwarz waveform relaxation for the one dimensional wave equation*, SIAM J. Numer. Anal., 41 (2003), pp. 1643–1681.

[15] B. Gustafsson, H.-O. Kreiss, and A. Sundström, *Stability theory of difference approximations for mixed initial boundary value problems.* II, Math. Comp., 26 (1972), pp. 649–686.

[16] P. Joly, *Variational methods for time-dependent wave propagation problems*, in Topics in Computational Wave Propagation. Direct and Inverse Problems., Lect. Notes Comput. Sci. Engrg. 31, Springer, Berlin, 2003, pp. 201–264.

[17] P. Joly and J. Rodríguez, l^2 *Error Analysis of Two Space-Time Mesh Refinement Schemes for the 1d Wave Equation*, Tech. Report, INRIA, to appear.

[18] I. S. Kim and W. J. R. Hoefer, *A local mesh refinement algorithm for the time-domain finite-difference method to solve Maxwell's equations*, IEEE Trans. Microwave Theory Tech., 38 (1990), pp. 812–815.

[19] H.-O. Kreiss, *Stability theory for difference approximations of mixed initial boundary value problems.* I, Math. Comp., 22 (1968), pp. 703–714.

[20] K. S. Kunz and L. Simpson, *A technique for increasing the resolution of finite-difference solutions to the maxwell equations*, IEEE Trans. Electromagn. Compat., EMC-23 (1981), pp. 419–422.

[21] R. Manfrin and F. Tonin, *On the Gevrey regularity for weakly hyperbolic equations with space-time degeneration of Oleinik type*, Rend. Mat. Appl. (7), 16 (1996), pp. 203–231.

[22] P. Monk, *Sub-gridding FDTD schemes*, ACES J., 11 (1996), pp. 37–46.

[23] S. Osher and R. Sanders, *Numerical approximations to nonlinear conservation laws with locally varying time and space grids*, Math. Comp., 41 (1983), pp. 321–336.

[24] D. T. Prescott, and N. V. Shuley, *A method for incorporating different sized cells into the*

finite-difference time-domain analysis technique, IEEE Microwave Guided Wave Lett., 2 (1992), pp. 434–436.

[25] A. Taflove, *Computational Electrodynamics, The Finite-Difference Time Domain Method*, Artech House, London, 1995.

[26] K. S. Yee, *Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media*, IEEE Trans. Antennas and Propagation, 14 (1966), pp. 302–307.

SIAM J. Numer. Anal.
Vol. 43, No. 2, pp. 860–890

FINITE VOLUME METHODS FOR DOMAIN DECOMPOSITION ON NONMATCHING GRIDS WITH ARBITRARY INTERFACE CONDITIONS*

L. SAAS†, I. FAILLE†, F. NATAF‡, and F. WILLIEN†

Abstract. We are interested in a robust and accurate domain decomposition method with arbitrary interface conditions on nonmatching multiblock grids using a finite volume discretization. To take into account the nonmatching grids at the interface, we introduce transmission operators, Dirichlet–Neumann interface conditions, and arbitrary equivalent interface conditions (for example, Robin interface conditions). Under a compatibility assumption on the transmission operators, we prove the equivalence between the different types of interface conditions and the well posedness of the local and global problems. Then two error estimates are proven in terms of the discrete H^1-norm: the first in $O(h)^{1/2}$ with transmission operators based on piecewise constant functions and the second in $O(h)$ (as in the matching case) with transmission operators using a piecewise linear rebuilding. In conclusion, numerical results are presented in confirmation of the theory.

Key words. nonmatching grids, finite volume, domain decomposition methods, optimized Schwarz method

AMS subject classifications. 65N12, 65N55

DOI. 10.1137/S0036142903434059

1. Introduction.

Our aim is to develop numerical methods which combine a finite volume method [21] and a domain decomposition [1] algorithm on nonmatching multiblock grids with arbitrary interface conditions. The goal is to obtain a numerical scheme as accurate as the finite volume scheme on matching grids ($O(h)$ in discrete H^1-norm). Finite volume schemes are used for their conservation property and are well adapted to deal with convective terms. A particular case of arbitrary interface conditions is the Robin interface conditions ($\frac{\partial}{\partial n}+\alpha$) which were introduced in domain decomposition methods in [18]. They enable the use of nonoverlapping subdomains and speed up the convergence of iterative domain decomposition algorithms [20] and [17]. Nonmatching grids make grid generation much easier and faster, since it is then a parallel task, and enable sliding blocks.

Some domain decomposition methods use the same numerical schemes in each subdomain as in the matching case. They define transmission operators and interface conditions to impose weak continuity of the primary unknown and its normal derivative across the nonmatching interface. The first introduced, the mortar method [7], is based on a finite element discretization [12] and introduces a mortar space at the interface to define interface conditions of Dirichlet–Neumann-type (mortar interface conditions). An extension to mixed finite elements with different Dirichlet–Neumann interface conditions is done in [4]. In [13], the mortar method is extended to the finite volume method after formulating the discretization as a finite element method using the dual mesh. In all these methods, error estimates are proven (order 1 ($O(h)$) in H^1-norm with P^1 finite element, with RT0 mixed finite element, or with vertex cen-

*Received by the editors August 26, 2003; accepted for publication (in revised form) February 18, 2005; published electronically September 15, 2005.

http://www.siam.org/journals/sinum/43-2/43405.html

†Institut Français du Pétrole, 1 et 4 avenue de Bois Préau, 92852 Rueil-Malmaison cedex, France (laurent.saas@wanadoo.fr, isabelle.faille@ifp.fr, francoise.willien@ifp.fr).

‡CMAP, Ecole Polytechnique, 91128 Palaiseau cedex, France (nataf@cmap.polytechnique.fr).

tered finite volume), but the use of Robin interface conditions does not seem possible. In [3] a finite volume scheme with Robin interface conditions is proposed (see [5] for the mixed finite element case). The error estimate depends on the Robin interface conditions ($\frac{\partial .}{\partial n} + \alpha\ .$): if $\alpha = O(1/h)^{\gamma}$, the error estimate is in $O(h)^{1-\gamma/2}$, and optimized Robin interface coefficients with respect to the convergence rate of the iterative domain decomposition algorithm [20], $\alpha_{opt} = O(1/h)^{1/2}$, give an error of $O(h)^{3/4}$ (less than that in the matching case). Moreover the Robin coefficients must be constant along the interface. These methods enable the use of Robin interface conditions but not arbitrary ones involving, for instance, second-order tangential derivatives.

Other domain decomposition methods for nonmatching grids consist of introducing a matching subgrid at the interface. They use classical interface conditions on this subgrid but need a modified numerical scheme in each subdomain. In the two-point flux approximation (TPFA) method, the intersection grids at the interface are introduced and classical Robin conditions are used. This method has already been used in oil engineering applications [6] and has been studied in [11] and [10]. A two-point scheme is used on the subgrid leading to an error estimate in $O(h)^{1/2}$ [10] because of the lack of consistency of the flux (a more general error estimate is performed depending on the number of nonmatching elements). In [8] the previous method is extended using interpolation near the subgrid to ensure a consistent flux approximation: stability and an error estimate are not proven but it seems to be in $O(h)$. In [2] a multipoint flux approximation (MPFA) is used near the nonmatching interface to ensure weak continuity of the principal unknown and of its normal derivative, but no theoretical results are proven.

All those methods either do not use arbitrary interface conditions or do not have finite volume accuracy or a mathematical basis. In this paper, we present and study, for a second-order elliptic problem, a domain decomposition method on nonmatching multiblock grids using finite volume and arbitrary interface conditions. For instance, the interface conditions used in the domain decomposition method could involve the discrete Steklov–Poincaré operator (see [20]), Robin interface conditions as in [5] or in [19], or second-order tangential derivatives as in [17], [16]. The reason for using such interface conditions is that they lead to faster domain decomposition methods.

The paper is organized as follows. In section 2, we introduce the domain decomposition formulation of a problem and its discretization by a finite volume scheme. Section 2.3 gives the finite volume discretization on the nonmatching grids. In section 3, we define transmission operators, Dirichlet–Neumann interface conditions, and equivalent arbitrary interface conditions. Section 4 deals with the wellposedness of the global and local problems. We give two examples of transmission operators in section 5. In section 6 error estimates are performed. The first transmission operators are piecewise constant projections and lead to an error estimate in $O(h^{1/2})$. The second transmission operators are based on the linear rebuilding of [4] and lead to an error estimate in $O(h)$ as in the matching case. In section 7 numerical results are shown. Finally, in section 8 we give our conclusion.

2. Formulation of the problem.

2.1. Domain decomposition.

Let Ω be a bounded domain in $\mathbb{R}^d$ $(d = 2, 3)$, $\eta > 0$ and $\operatorname{div}(\vec{a}) \geq 0$ with $\vec{a} \in C^1(\Omega)$. We consider the following second-order elliptic problem:

$$\eta p - \Delta p + \operatorname{div}(\vec{a} p) = f \text{ in } \Omega, \tag{2.1}$$

$$p = g \text{ on } \partial\Omega, \tag{2.2}$$

where f and g are given data. The assumption $\eta > 0$ is made throughout the paper. It may probably be relaxed but at the expense of longer and more tedious proofs. The domain Ω is decomposed into N ($N \geq 2$) nonoverlapping subdomains Ω_i, and we introduce the set $\mathcal{I} = \{i \in \mathbb{N} \mid 1 \leq i \leq N\}$ such that $\overline{\Omega} = \cup_{i\in\mathcal{I}}\overline{\Omega_i}$ (with $\Omega_i \cap \Omega_j = \emptyset$). We also introduce for all $i \in \mathcal{I}$, $\mathcal{V}_i = \{j \in \mathcal{I} \mid i \neq j \text{ and } \dim(\overline{\Omega}_i \cap \overline{\Omega}_j) = d-1\}$ and for all $j \in \mathcal{V}_i$, we denote $\Gamma_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$ (we have the following property: $i \in \mathcal{V}_j \iff j \in \mathcal{V}_i$). At the continuous level, for all $i \in \mathcal{I}$, for all $j \in \mathcal{V}_i$ given arbitrary positive definite interface operators $\tilde{S}^i_j$ acting on functions living on the interface between two subdomains Ω_i and Ω_j, the above problem (2.1)–(2.2) can be reformulated as the following domain decomposition problem with arbitrary continuous interface conditions (2.5): for all $i \in \mathcal{I}$

$$\eta p_i - \Delta p_i + \operatorname{div}(\vec{a}p_i) = f \quad \text{in } \Omega_i, \tag{2.3}$$

$$p_i = g \quad \text{on } \partial\Omega \cap \partial\Omega_i, \tag{2.4}$$

$$\frac{\partial p_i}{\partial n_i} + \tilde{S}^i_j(p_i) = -\frac{\partial p_j}{\partial n_j} + \tilde{S}^i_j(p_j) \text{ on } \Gamma_{ij} \quad \forall j \in \mathcal{V}_i. \tag{2.5}$$

A simple iterative method for solving the above domain decomposition method is the additive Schwarz method:

$$\eta p_i^{n+1} - \Delta p_i^{n+1} + \operatorname{div}(\vec{a}p_i^{n+1}) = f \quad \text{in } \Omega, \tag{2.6}$$

$$p_i^{n+1} = g \quad \text{on } \partial\Omega \cap \partial\Omega_i, \tag{2.7}$$

$$\frac{\partial p_i^{n+1}}{\partial n_i} + \tilde{S}^i_j(p_i^{n+1}) = -\frac{\partial p_j^n}{\partial n_j} + \tilde{S}^i_j(p_j^n) \text{ on } \Gamma_{ij} \quad \forall j \in \mathcal{V}_i. \tag{2.8}$$

The wellposedness and convergence of the above problems have been studied in [18] for Robin interface conditions ($\tilde{S}^j_i(p) = \alpha^j_i p$ with $\alpha^j_i > 0$). It is also possible to use (2.6)–(2.8) as a preconditioner for Krylov-type methods; see, for example, [1], [22]. In this paper, we present a finite volume counterpart of problem (2.1)–(2.2) on nonmatching multiblock grids with discrete arbitrary interface conditions. We want to solve it with an iterative domain decomposition algorithm of the same type as (2.6)–(2.8). In the next subsection, we give the classical finite volume that we use inside a subdomain Ω_i.

2.2. Finite volume discretization.

(2.1)–(2.2) are discretized using a cell centered finite volume scheme in each subdomain [21]. We choose this scheme as an example but other schemes would be possible as well.

2.2.1. Mesh and definition.

For $i \in \mathcal{I}$, let $\mathcal{T}_i$ be a set of closed polygonal subsets associated with Ω_i such that $\Omega_i = \cup_{K\in\mathcal{T}_i} K$. We shall use the following notation for all $i \in \mathcal{I}$.

- $\mathcal{E}_{\Omega_i}$ is the set of faces of $\mathcal{T}_i$.
- $\mathcal{E}_{iD}$ is the set of faces such that $\partial\Omega_i \cap \partial\Omega = \cup_{\epsilon\in\mathcal{E}_{iD}}\epsilon$ (let us recall that a Dirichlet boundary condition will be imposed on $\partial\Omega_i \cap \partial\Omega$).
- $\mathcal{E}_i$ is the set of faces such that $\partial\Omega_i \backslash \partial\Omega = \cup_{\epsilon\in\mathcal{E}_i}\epsilon$ (let us recall that a Dirichlet–Neumann or an arbitrary boundary condition will be imposed on $\partial\Omega_i \backslash \partial\Omega$).
- For all $j \in \mathcal{V}_i$, $\mathcal{E}_{i\to j} = \{\epsilon \in \mathcal{E}_i | \epsilon \cap \Gamma_{ij} \neq \emptyset\}$, the grid of Γ_{ij} for the subdomain Ω_i, $\mathcal{E}_i = \cup_{j\in\mathcal{V}_i}\mathcal{E}_{i\to j}$ (because the grids are nonmatching at the interface Γ_{ij}, we have $\mathcal{E}_{i\to j} \neq \mathcal{E}_{j\to i}$).
- For all $j \in \mathcal{V}_i$, $\mathcal{E}_{ij} = \{\epsilon_i \cap \epsilon_j | \epsilon_i \in \mathcal{E}_{i\to j} \text{ and } \epsilon_j \in \mathcal{E}_{j\to i}\}$, the subgrid intersection of $\mathcal{E}_{i\to j}$ and $\mathcal{E}_{j\to i}$. For any $\epsilon \in \mathcal{E}_{ij}$, there exists a unique pair $(K_i, K_j) \in \mathcal{T}_i \times \mathcal{T}_j$ such that $\epsilon = K_i \cap K_j$. We shall use the notation $K_i = K_i(\epsilon)$ and $K_j = K_j(\epsilon)$.

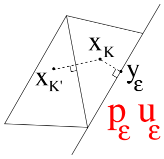


FIG. 1. *Assumption* 1.

- $\forall K \in \mathcal{T}_i$,

  $\mathcal{E}(K)$ denotes the set of faces of K.

  $\mathcal{E}_{iD}(K) = \mathcal{E}(K) \cap \mathcal{E}_{iD}$ is the set of faces of K which are on $\partial\Omega_i \cap \partial\Omega$.

  $\mathcal{E}_i(K) = \mathcal{E}(K) \cap \mathcal{E}_i$ is the set of faces of K which are on $\partial\Omega_i \backslash \partial\Omega$.

  $\mathcal{N}_i(K) = \{K' \in \mathcal{T}_i | K \cap K' \in \mathcal{E}_{\Omega_i}\}$ is the set of the control cells adjacent to K in Ω_i.

  $\mathcal{N}_{ij}(K) = \{K' \in \mathcal{T}_j | K \cap K' \in \mathcal{E}_{ij}\}$ is the set of the control cells adjacent to K which belongs to Ω_j.

  $\mathcal{N}(K) = \mathcal{N}_i(K) \cup (\cup_{j \in \mathcal{V}_i} \mathcal{N}_{ij}(K))$ and we note for all $K' \in \mathcal{N}(K)$, $[K, K'] = K \cap K'$.

We make the following geometrical assumptions on the meshes (see Figure 1).

Assumption 1. For all $i \in \mathcal{I}$, $\mathcal{T}_i$ is a finite volume admissible mesh, i.e., $\mathcal{T}_i$ is a set of closed subsets of dimension d such that

- for any $(K, K') \in \mathcal{T}_i^2$ with $K \neq K'$, one has either $K \cap K' \in \mathcal{E}_{\Omega_i}$ or $\dim(K \cap K') < d - 1$
- there exist points $(y_\epsilon)_{\epsilon \in \mathcal{E}_{\Omega_i}}$ on the faces and points $(x_K)_{K \in \mathcal{T}_i}$ inside the cells such that
  - for any adjacent cells K and K', the straight line $[x_K, x_{K'}]$ is perpendicular to the face $[K, K']$ and $[x_K, x_{K'}] \cap [K, K'] = \{y_{[K,K']}\}$
  - for any face $\epsilon \in \mathcal{E}_{iD}$, let $K(\epsilon) \in \mathcal{T}_i$ be such that $\epsilon \subset K$: then the straight line $[x_{K(\epsilon)}, y_\epsilon]$ is perpendicular to ϵ
- for all $i \in \mathcal{I}$, no face intersects both $\partial\Omega_i \backslash \partial\Omega$ and $\partial\Omega_i \cap \partial\Omega$ ($\mathcal{E}_i \cap \mathcal{E}_{iD} = \emptyset$)
- for all $j \in \mathcal{V}_i$, $\Gamma_{ij} = \partial\Omega_i \cap \partial\Omega_j$ can be written as the union of faces of $\mathcal{E}_{i \to j}$ and of $\mathcal{E}_{j \to i}$

Remark 2.1. As a result, $\bigcup_{i \in \mathcal{I}} \partial\Omega_i \cap \partial\Omega$ can be written as a union of (whole) faces. The same holds for $\partial\Omega_i \backslash \partial\Omega$. For all $j \in \mathcal{V}_i$ and for all $k \in \mathcal{V}_i$ ($j \neq k$), we have $\mathcal{E}_{i \to j} \cap \mathcal{E}_{i \to k} = \emptyset$. We note $\mathcal{T} = \cup_{i \in \mathcal{I}} \mathcal{T}_i$ and $h = \max_{i \in \mathcal{I}, K \in \mathcal{T}_i} diam(K)$ its mesh size.

2.2.2. Cell centered finite volume scheme in the subdomains.

Let $i \in \mathcal{I}$; we shall use the primary unknowns $(p_K^i)_{K \in \mathcal{T}_i}$ which aim at being approximations of $p(x_K)$ and for $\epsilon \in \mathcal{E}_i$, $(p_\epsilon^i,\, u_\epsilon^i)$ which aim at being approximations of

$$(p(y_\epsilon),\, \partial p / \partial n_i(y_\epsilon)).$$

The scheme is obtained by integrating (2.6) over each control volume $K \in \mathcal{T}_i$:

$$\int_K \eta p - \int_{\partial K} \frac{\partial p}{\partial \vec{n}_K} + \int_{\partial K} \vec{a} \cdot \vec{n}_K p = \int_K f, \tag{2.9}$$

where $\vec{n}_K$ is the outward normal on ∂K of K. If we introduce F_K such that

$$S_K = \frac{1}{meas(K)} \int_K f - F_K = O(diam(K)). \tag{2.10}$$

(2.9) is discretized by

$$\begin{aligned}&\eta\, meas(K)p_K^i - \sum_{K'\in\mathcal{N}_i(K)} [u_{K,K'}^i meas([K,K']) + v_{K,K'}^i] - \sum_{\epsilon\in\mathcal{E}_{iD}(K)\cup\mathcal{E}_i(K)} u_\epsilon^i meas(\epsilon)\\ &\quad + \sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} v_{[K,K']}^{i,j} + \sum_{\epsilon\in\mathcal{E}_{iD}(K)} v_{\epsilon,K^+}^i = F_K meas(K),\end{aligned} \tag{2.11}$$

where $meas(\cdot)$ denotes the Lebesgue measure and u^i and v^i are defined in what follows. For an interior face $[K,K'] \in \mathcal{E}_{\Omega_i}\backslash(\mathcal{E}_{iD}\cup\mathcal{E}_i)$, we define

$$u_{K,K'}^i = \frac{p_{K'}^i - p_K^i}{d(x_{K'},x_K)}, \tag{2.12}$$

where $d(x,y)$ is the Euclidean distance between x and y. For a face $\epsilon\in\mathcal{E}_{iD}(K)$ on the boundary $\partial\Omega_i\cap\Omega$, the Dirichlet boundary condition (2.7) is taken into account by using

$$u_\epsilon^i = \frac{g_\epsilon^i - p_K^i}{d(y_\epsilon,x_K)}, \tag{2.13}$$

where $g_\epsilon^i = g(y_\epsilon)$. On each interface face $\epsilon\in\mathcal{E}_i(K)$, we have

$$u_\epsilon^i = \frac{p_\epsilon^i - p_K^i}{d(y_\epsilon,x_K)}. \tag{2.14}$$

As for the convective terms, for any adjacent cells K, K', let

$$v_{K,K'}^i = a_{[K,K']}p_{[K,K']^+}^i, \tag{2.15}$$

where $a_{[K,K']} = \int_{[K,K']}\vec{a}\cdot\vec{n}_K$. If $a_{[K,K']}\geq 0$, then $p_{[K,K']^+}^i = p_K^i$; else $p_{[K,K']^+}^i = p_{K'}^i$ (upwind scheme). Note that (2.15) will even be used for cells adjacent to the interfaces between subdomains. We have the useful properties that $u_{K,K'}^i = -u_{K',K}^i$, $a_{[K,K']} = -a_{[K',K]}$, and $p_{[K,K']^+}^i = p_{[K',K]^+}^i$. For the discretization of the convective term on $\epsilon\in\mathcal{E}_{iD}(K)$, we introduce $a_{\epsilon,K} = \int_\epsilon \vec{a}\cdot\vec{n}_K$, and if $a_{\epsilon,K}\geq 0$, then $p_{\epsilon,K^+}^i = p_K^i$; else $p_{\epsilon,K^+}^i = g(y_\epsilon)$. We define

$$v_{\epsilon,K^+}^i = a_{\epsilon,K}p_{\epsilon,K^+}^i. \tag{2.16}$$

When there is no domain decomposition, this scheme has been analyzed in [21] in the more general case of discontinuous coefficients, and it is proven to be of order 1 for a discrete H^1-norm.

In order to define the domain decomposition discretization scheme, we shall define in section 2.3 the convective fluxes on the interface and the matching conditions for the diffusive fluxes.

To simplify the notation, for all $i\in\mathcal{I}$, we associate with any discrete values $(p_\epsilon^i)_{\epsilon\in\mathcal{E}_i}$, p_i its natural piecewise constant extrapolation in $P^0(\mathcal{E}_i)$:

$$\begin{aligned} p_i:\quad &\partial\Omega_i\backslash\partial\Omega \longrightarrow \mathbb{R},\\ &y \longmapsto p_i(y) = p_\epsilon^i \quad \text{if } y\in\epsilon\subset\mathcal{E}_i\end{aligned}$$

and also for all $j \in \mathcal{V}_i$, p_i^j the restriction of p_i to $\partial\Omega_i \cap \partial\Omega_j$.

We define the spaces of discretization and the associated discrete norms for all $i \in \mathcal{I}$:

$$\begin{aligned} P^0(\mathcal{E}_{i\to j}) &= \{p : \Gamma_{ij} \longrightarrow \mathbb{R} \mid p \text{ is constant on all } \epsilon \in \mathcal{E}_{i\to j}\}, \\ P^0(\mathcal{E}_i) &= \{p \text{ is constant on all } \epsilon \in \mathcal{E}_i\}, \\ Z_i &= \{p : \Omega_i \longrightarrow \mathbb{R} \mid \forall K \in \mathcal{T}_i, p \text{ is constant on } K\}, \\ Z &= \prod_{i\in\mathcal{I}} Z_i. \end{aligned}$$

The discrete L^2-norms and H^1-norms are defined by the following: for all $p \in Z_i$,

$$\|p\|^2_{L^2(\Omega_i)} = \sum_{K\in\mathcal{T}_i} (p_K)^2 meas(K)$$

and for all $q \in P^0(\mathcal{E}_i)$,

$$\|q\|^2_{L^2(\partial\Omega_i)} = \sum_{\epsilon\in\mathcal{E}_{iD}\cup\mathcal{E}_i} (q_\epsilon)^2 meas(\epsilon).$$

Let $X_i = Z_i \times P^0(\mathcal{E}_i)$ be endowed with the following seminorms: for $\tilde{p} = (p, p_b) \in X_i$,

$$\begin{aligned} |\tilde{p}|^2_{1,\mathcal{T}_i} = \sum_{K\in\mathcal{T}_i} \Bigg[&\sum_{K'\in\mathcal{N}_i(K)} \frac{(p_K - p_{K'})^2}{d(x_K, x_{K'})} meas([K,K']) \\ &+ \sum_{\epsilon\in\mathcal{E}_{iD}(K)} \frac{(p_K)^2}{d(x_K, y_\epsilon)} meas(\epsilon) + \sum_{\epsilon\in\mathcal{E}_i(K)} \frac{(p_{b_\epsilon} - p_K)^2}{d(x_K, y_\epsilon)} meas(\epsilon) \Bigg] \end{aligned}$$

and

$$\|\tilde{p}\|_{L^2(\partial\Omega_i)} = \Big(\sum_{\epsilon\in\mathcal{E}_i} |p_{b\epsilon}|^2 \, meas(\epsilon) \Big)^{1/2}.$$

In what follows, we shall drop the subscript b.

For all $p = (p_i)_{i\in\mathcal{I}} \in Z$, let $\|p\|^2_{L^2(\Omega)} = \sum_{i=1}^N \|p_i\|^2_{L^2(\Omega_i)}$. Let the product space $X = \Pi_{i\in\mathcal{I}} X_i$ be endowed with the following seminorm: for $\tilde{p} = (\tilde{p}_i)_{i\in\mathcal{I}} \in X$, $|\tilde{p}|^2_{1,\mathcal{T}} = \sum_{i=1}^N |\tilde{p}_i|^2_{1,\mathcal{T}_i}$.

2.3. Finite volume scheme on the interfaces.

We first give the discretization of the convective flux through the interface on the nonmatching grids. Then in order to enforce the weak continuity of the primary unknown p and of its normal derivative (denoted by u) through the interface on the nonmatching grids, we introduce transmission operators.

Convective flux. The convective flux on the nonmatching grids is discretized as in the matching case by an upwind scheme [21]. For all $K \in \mathcal{T}_i$, for all $j \in \mathcal{V}_i$, for all $K' \in \mathcal{N}_{ij}(K)$, we define

$$v^{i,j}_{[K,K']} = a_{[K,K']} p^{i,j}_{[K,K']^+}, \tag{2.17}$$

where $a_{[K,K']} = \int_{[K,K']} \vec{a} \cdot \vec{n}_K$ and $p^{i,j}_{[K,K']^+} = p^i_K$ if $a_{[K,K']} \geq 0$, and $p^{i,j}_{[K,K']^+} = p^j_K$ otherwise.

Transmission operators and Dirichlet–Neumann interface conditions. In order to enforce a weak continuity across the interface Γ_{ij} of the principal unknown p and of its normal derivative (denoted by u), we introduce the linear transmission operators $Q_i^j : P^0(\mathcal{E}_{j\to i}) \longrightarrow P^0(\mathcal{E}_{i\to j})$. They satisfy the following compatibility condition.

Assumption 2. For all $i \in \mathcal{I}$, $j \in \mathcal{V}_i$ and $u \in P^0(\mathcal{E}_{i\to j})$ and $v \in P^0(\mathcal{E}_{j\to i})$,

$$\langle Q_j^i(u), v\rangle_{L^2(\Gamma_{ij})} = \langle u, Q_i^j(v)\rangle_{L^2(\Gamma_{ij})}$$

As in mortar methods [7], we consider that one subdomain enforces the weak continuity of the primary unknown which is interpreted as the Dirichlet interface condition. This subdomain is called the master. The other subdomain enforces the weak continuity of the normal derivative which corresponds to a Neumann interface condition and is called the slave. The conditions are interpreted and considered in the following as Dirichlet–Neumann interface conditions. We introduce the following definition.

DEFINITION 2.2. *For all $i \in \mathcal{I}$, $\mathcal{V}_i$ is partitioned as $\mathcal{V}_i = \mathcal{S}_i \cup \mathcal{M}_i$ such that $\mathcal{S}_i \cap \mathcal{M}_i = \emptyset$ and for all $j \in \mathcal{I}$, $i \in \mathcal{S}_j \Longleftrightarrow j \in \mathcal{M}_i$.*

We define for all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$, $p_i, u_i \in P^0(\mathcal{E}_{j\to i})$, and $p_j, u_j \in P^0(\mathcal{E}_{i\to j})$ the Dirichlet–Neumann interface conditions on Γ_{ij} by

$$p_j = Q_j^i(p_i), \tag{2.18}$$

$$u_i = Q_i^j(-u_j), \tag{2.19}$$

where, if $j \in \mathcal{S}_i$, subdomain Ω_i is the master of subdomain Ω_j which is the slave. On each interface between two subdomains, one side has to be the master and the other one the slave. This choice is arbitrary. In what follows, when there is no ambiguity, we shall drop the superscripts in the above equalities. Two types of transmission operators will be given in section 5.

3. Interface boundary conditions.

To define more general interface conditions that have an impact on the convergence rate of the iterative domain decomposition algorithm (2.6)–(2.8) but not on the converged solution, we introduce the linear interface operator $S_{i,j} : P^0(\mathcal{E}_{i\to j}) \longrightarrow P^0(\mathcal{E}_{i\to j})$. The interface operators satisfy the following hypothesis.

Assumption 3. For all $i \in \mathcal{I}$, $j \in \mathcal{V}_i$, $i \neq j$, $S_{i,j}$ is positive definite (for all $u \in P^0(\mathcal{E}_{i\to j})$, $u \neq 0$, $\langle S_{i,j}(u), u\rangle_{L^2(\Gamma_{ij})} > 0$).

Assumption 3 implies that $S_{i,j}$ is invertible. For example $S_{i,j}$ can be

- the discrete Steklov–Poincaré operator (see [20]);
- a diagonal operator ($S_{i,j} = diag(\alpha_\epsilon)$, with $\alpha_\epsilon > 0$ a constant Robin coefficient on $\epsilon \in \mathcal{E}_{i\to j}$);
- optimized Robin interface coefficients of order 0 ($S_{i,j} = diag(\alpha_\epsilon^{opt})$, with $\alpha_\epsilon^{opt} > 0$ Robin optimized coefficients on $\epsilon \in \mathcal{E}_{i\to j}$) (see [19]);
- the discretization of $(\alpha - \frac{\partial}{\partial\tau}(\beta\frac{\partial}{\partial\tau}))$, where τ is the tangential vector of the interface ($S_{i,j}$ is a tridiagonal operator with $\alpha_\epsilon^{opt} > 0$ and β_ϵ^{opt} for $\epsilon \in \mathcal{E}_{i\to j}$) (see [17], [16]).

In what follows, when there is no ambiguity, we shall denote $S_{i,j}$ by S_i.

3.1. Dirichlet–Neumann and arbitrary interface conditions.

The arbitrary interface conditions are defined by the following: for all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$,

$$u_i + Q_i(S_jQ_j(p_i)) = Q_i(-u_j + S_j(p_j)) \quad \text{on } \Gamma_{i,j}, \tag{3.1}$$

$$Q_j(S_i^{-1}Q_i(u_j)) + p_j = Q_j(-S_i^{-1}(u_i) + p_i) \quad \text{on } \Gamma_{i,j}. \tag{3.2}$$

Theorem 3.1 proves that conditions (2.19) and (2.18) and conditions (3.1) and (3.2) are equivalent.

THEOREM 3.1. *Assume that for all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$, the transmission operators Q_i^j satisfy Assumption 2 and that the interface operators $S_{i,j}$ satisfy Assumption 3; then we have the following equivalence: for all $(p_i, u_i) \in P^0(\mathcal{E}_{i\to j}) \times P^0(\mathcal{E}_{i\to j})$ and for all $(p_j, u_j) \in P^0(\mathcal{E}_{j\to i}) \times P^0(\mathcal{E}_{j\to i})$,*

$$\begin{cases} u_i + Q_i(S_{j,i}Q_j(p_i)) & = & Q_i(-u_j + S_{j,i}(p_j)) \\ Q_j(S_{i,j}^{-1}Q_i(u_j)) + p_j & = & Q_j(-S_{i,j}^{-1}(u_i) + p_i) \end{cases} \iff \begin{cases} u_i & = & Q_i(-u_j), \\ p_j & = & Q_j(p_i). \end{cases}$$

Proof. ($\Rightarrow$) We introduce the auxiliary variables

$$\begin{cases} \delta p_j = Q_j(p_i) - p_j, \\ \delta u_i = Q_i(u_j) + u_i. \end{cases}$$

Then conditions (3.1) and (3.2) are rewritten with δp_j and δu_i as

$$\begin{cases} Q_i(S_j(\delta p_j)) + \delta u_i & = 0, \\ -Q_j(S_i^{-1}(\delta u_i)) + \delta p_j & = 0. \end{cases}$$

By multiplying the first equation by $S_i^{-1}(\delta u_i)$ and the second by $S_j(\delta p_j)$, integrating over Γ_{ij}, and summing we obtain

$$\begin{aligned} &\int_{\Gamma_{ij}} Q_i(S_j(\delta p_j))S_i^{-1}(\delta u_i) - \int_{\Gamma_{ij}} Q_j(S_i^{-1}(\delta u_i))S_j(\delta p_j) \\ &\quad + \int_{\Gamma_{ij}} (\delta u_i)S_i^{-1}(\delta u_i) + \int_{\Gamma_{ij}} (\delta p_j)S_j(\delta p_j) = 0. \end{aligned} \tag{3.3}$$

From Assumption 2, we deduce

$$\int_{\Gamma_{ij}} Q_i(S_j(\delta p_j))S_i^{-1}(\delta u_i) = \int_{\Gamma_{ij}} S_j(\delta p_j)Q_jS_i^{-1}(\delta u_i).$$

Consequently (3.3) becomes

$$\int_{\Gamma_{ij}} \delta u_i S_i^{-1}(\delta u_i) + \int_{\Gamma_{ij}} \delta p_j S_j(\delta p_j) = 0.$$

S_i and S_j satisfying Assumption 3, we have $\delta u_i = \delta p_j = 0$, and thus $u_i = Q_i(-u_j)$ and $p_j = Q_j(p_i)$.

($\Leftarrow$) Equation (2.18) yields

$$p_j^i = Q_j^i(p_i^j) \Rightarrow S_{j,i}(p_j^i) = S_{j,i}Q_j^i(p_i^j) \Rightarrow Q_i^j(S_{j,i}(p_j^i)) = Q_i^j(S_{j,i}Q_j^i(p_i^j)); \tag{3.4}$$

then combining (3.4) and (2.19) we get (3.1). Proceeding in the same way with (2.19), we get (3.2). ∎

4. Wellposedness. Before proving that the global problem defined by (2.11)–(2.19) is well posed, we state the following lemma.

LEMMA 4.1. *In each subdomain Ω_i $(i \in \mathcal{I})$, if $(p, u_\epsilon) \in X_i \times P^0(\mathcal{E}_i)$ satisfy* (2.11)–(2.16)*, then the following estimate holds:*

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_i}\Bigg(\eta\, meas(K)(p_K^i)^2+\frac{1}{2}\sum_{K'\in\mathcal{N}_i(K)}\frac{(p_K^i-p_{K'}^i)^2}{d(x_K,x_{K'})}meas([K,K'])\\
&\quad+\sum_{\epsilon\in\mathcal{E}_{iD}(K)}\frac{(p_K^i)^2}{d(y_\epsilon,x_K)}meas(\epsilon)+\sum_{\epsilon\in\mathcal{E}_i(K)}d(y_\epsilon,x_K)(u_\epsilon^i)^2 meas(\epsilon)-\sum_{\epsilon\in\mathcal{E}_i(K)}u_\epsilon^i p_\epsilon^i meas(\epsilon)\\
&\quad+\frac{1}{2}\sum_{\epsilon\in\mathcal{E}(K)\setminus\mathcal{E}_i(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_K\right](p_K^i)^2+\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)}v^{i,j}_{[K,K']}p_K^i\Bigg)\\
&\quad\le\sum_{K\in\mathcal{T}_i}F_K p_K^i meas(K)+\sum_{\epsilon\in\mathcal{E}_{iD}}\frac{p_K^i g_\epsilon^i}{d(y_\epsilon,x_K)}meas(\epsilon)-\sum_{\epsilon\in\mathcal{E}_{iD}}1_{[a_{\epsilon,K}\le 0]}\;a_{\epsilon,K}g_\epsilon^i p_K^i.
\end{aligned}$$

Remark 4.2. For any K such that $\mathcal{E}(K)\subset[\mathcal{E}_{\Omega_i}\setminus(\mathcal{E}_i\cup\mathcal{E}_{iD})]$, thanks to div $(\vec{a})\ge 0$, we have

$$\sum_{\epsilon\in\mathcal{E}(K)}\int_\epsilon \vec{a}\cdot\vec{n}_{K(\epsilon)}(p_{K(\epsilon)}^i)^2=\int_K(\mathrm{div}(\vec{a}))(p_{K(\epsilon)}^i)^2\ge 0.$$

Proof. The summation over $K\in\mathcal{T}_i$ in (2.11) multiplied by p_K^i yields

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_i}\Bigg(\eta\, meas(K)(p_K^i)^2-\sum_{K'\in\mathcal{N}_i(K)}\left[u_{K,K'}^i p_K^i meas([K,K'])+v^i_{[K,K']^+}p_K^i\right]\\
&\quad-\sum_{\epsilon\in\mathcal{E}_{iD}(K)\cup\mathcal{E}_i(K)}u_\epsilon^i p_K^i meas(\epsilon)+\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)}v^{i,j}_{[K,K']}p_K^i+\sum_{\epsilon\in\mathcal{E}_{iD}(K)}v^i_{\epsilon,K^+}p_K^i\Bigg)\\
&\qquad=\sum_{K\in\mathcal{T}_i}F_K p_K^i meas(K).
\end{aligned}\tag{4.1}$$

Taking into account (2.12)–(2.16), (4.1) reads as

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_i}\Bigg(\eta\, meas(K)(p_K^i)^2+\sum_{K'\in\mathcal{N}_i(K)}\left[p_K^i\frac{p_K^i-p_{K'}^i}{d(x_K,x_{K'})}meas([K,K'])+a_{[K,K']}p^i_{[K,K']^+}p_K^i\right]\\
&\qquad+\sum_{\epsilon\in\mathcal{E}_{iD}(K)}p_K^i\frac{p_K^i-g_\epsilon^i}{d(y_\epsilon,x_K)}\;meas(\epsilon)+\sum_{\epsilon\in\mathcal{E}_i(K)}d(y_\epsilon,x_K)(u_\epsilon^i)^2 meas(\epsilon)\\
&\qquad+\sum_{\epsilon\in\mathcal{E}_{iD}(K)}a_{\epsilon,K}p^i_{\epsilon,K^+}p_K^i-\sum_{\epsilon\in\mathcal{E}_i(K)}u_\epsilon^i p_\epsilon^i meas(\epsilon)+\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)}v^{i,j}_{[K,K']}p_K^i\Bigg)\\
&=\sum_{K\in\mathcal{T}_i}F_K p_K^i meas(K).
\end{aligned}\tag{4.2}$$

The second term of (4.2) can be replaced by

$$\sum_{K'\in\mathcal{N}_i(K)}p_K^i\frac{p_K^i-p_{K'}^i}{d(x_K,x_{K'})}meas([K,K'])=\frac{1}{2}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_i(K)}\frac{(p_K^i-p_{K'}^i)^2}{d(x_K,x_{K'})}meas([K,K']).$$

The third term of (4.2) corresponds to the convective term inside the subdomain Ω_i, and proceeding as in [21], we obtain

$$\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_i(K)} a_{[K,K']}p^i_{[K,K']^+}p^i_K \geq \frac{1}{2}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}(K)\setminus(\mathcal{E}_{iD}(K)\cup\mathcal{E}_i(K))}\int_\epsilon \vec{a}\cdot\vec{n}_K|p^i_K|^2. \tag{4.3}$$

For the convection fluxes on the Dirichlet boundary $\partial\Omega_i\cap\partial\Omega$, we have

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_{iD}(K)} a_{\epsilon,K}p^i_{\epsilon,K^+}p^i_K\\
&\quad=\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_{iD}(K)} 1_{[a_{\epsilon,K}\geq 0]}\ a_{\epsilon,K}(p^i_K)^2+\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_{iD}(K)} 1_{[a_{\epsilon,K}\leq 0]}\ a_{\epsilon,K}g^i_\epsilon p^i_K\\
&\quad\geq\frac{1}{2}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_{iD}(K)} 1_{[a_{\epsilon,K}\geq 0]}\ a_{\epsilon,K}(p^i_K)^2+\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_{iD}(K)} 1_{[a_{\epsilon,K}\leq 0]}\ a_{\epsilon,K}g^i_\epsilon p^i_K\\
&\quad\geq\sum_{K\in\mathcal{T}_i}\Bigg(\frac{1}{2}\sum_{\epsilon\in\mathcal{E}_{iD}(K)}\left[\int_\epsilon\vec{a}\cdot\vec{n}_K\right](p^i_K)^2+\sum_{\epsilon\in\mathcal{E}_{iD}(K)} 1_{[a_{\epsilon,K}\leq 0]}\ a_{\epsilon,K}g^i_\epsilon p^i_K\Bigg). \qquad (4.4)
\end{aligned}$$

Thus combining (4.2)–(4.4), we get the inequality of Lemma 4.1. □

4.1. Global wellposedness.

THEOREM 4.3. *Under Assumptions* 2 *and* 3*, the finite volume discretization defined by* (2.11)–(2.19) *is well posed and if* $(g^i_\epsilon)_{\epsilon\in\mathcal{E}_{iD}}=0$*, there exists* $C\geq 0$ *such that its unique solution* $p\in X$ *satisfies*

$$\eta\|p\|^2_{L^2(\Omega)}+|p|^2_{1,\mathcal{T}}\leq C\|F\|^2_{L^2(\Omega)}. \tag{4.5}$$

Remark 4.4. The assumptions of Theorem 3.1 are satisfied, so the interface conditions (2.19) and (2.18) are equivalent to the interface conditions (3.1) and (3.2), which are easier to use. The global problem is defined also with (2.19) and (2.18) instead of (3.1) and (3.2).

Proof. The global system is a linear square, so we just have to prove that the solution of the global homogeneous problem ($(F_K)_{K\in\mathcal{T}_i}=0$ and $(g^i_\epsilon)_{\epsilon\in\mathcal{E}_{iD}}=0$ for all $\in\mathcal{I}$) is zero. We sum over $i\in\mathcal{I}$ the equality of Lemma 4.1 and with (2.17) for all $i\in\mathcal{I}$, for all $j\in\mathcal{V}_i$, it yields

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\Bigg(\eta\ meas(K)(p^i_K)^2+\frac{1}{2}\sum_{K'\in\mathcal{N}_i(K)}\frac{(p^i_K-p^i_{K'})^2}{d(x_K,x_{K'})}\ meas([K,K'])\\
&\quad+\sum_{\epsilon\in\mathcal{E}_{iD}(K)}\frac{(p^i_K)^2}{d(y_\epsilon,x_K)}meas(\epsilon)+\sum_{\epsilon\in\mathcal{E}_i(K)} d(y_\epsilon,x_K)(u^i_\epsilon)^2meas(\epsilon)\\
&\quad-\sum_{\epsilon\in\mathcal{E}_i(K)} u^i_\epsilon p^i_\epsilon meas(\epsilon)+\sum_{\epsilon\in\mathcal{E}(K)\setminus\mathcal{E}_i(K)}\left[\int_\epsilon\vec{a}\cdot\vec{n}_K\right](p^i_K)^2\\
&\quad+\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} a_{[K,K']}p^{i,j}_{[K,K']^+}p^i_K\Bigg)\leq\sum_{i\in\mathcal{I}}\Bigg(\sum_{K\in\mathcal{T}_i}F_Kp^i_Kmeas(K)\\
&\quad+\sum_{\epsilon\in\mathcal{E}_{iD}}\frac{p^i_Kg^i_\epsilon}{d(y_\epsilon,x_K)}meas(\epsilon)-\sum_{\epsilon\in\mathcal{E}_{iD}} 1_{[a_{\epsilon,K}\leq 0]}\ a_{\epsilon,K}g^i_\epsilon p^i_K\Bigg). \qquad (4.6)
\end{aligned}$$

The term of the diffusive fluxes could be rewritten as

$$\begin{aligned}
-\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} u^i_\epsilon p^i_\epsilon meas(\epsilon) &= -\sum_{i\in\mathcal{I}}\int_{\Gamma_i} u_i p_i = -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\int_{\Gamma_{ij}} u^j_i p^j_i \\
&= -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}} \left[u^j_i p^j_i + u^i_j p^i_j\right] \\
&= -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}} \left[u^j_i Q^j_i(p^i_j) - Q^i_j(u^j_i) p^i_j\right]
\end{aligned}$$

and using Assumption 2

$$= -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}} \left[u^j_i Q^j_i(p^i_j) - u^j_i Q^j_i(p^i_j)\right] = 0. \tag{4.7}$$

For the convective fluxes on the interface, we obtain

$$\begin{aligned}
T_1 &= \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} a_{[K,K']} p^{i,j}_{[K,K']^+} p^i_K \\
&= \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} \big(a_{[K,K']} p^{i,j}_{[K,K']^+} p^i_K + a_{[K',K]} p^{j,i}_{[K',K]^+} p^j_{K'}\big).
\end{aligned}$$

Since $p^{i,j}_{[K,K']^+} = p^{j,i}_{[K',K]^+}$ and $a_{[K,K']} = -a_{[K',K]}$, we write

$$T_1 = \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} a_{[K,K']} p^{i,j}_{[K,K']^+}(p^i_K - p^j_{K'}). \tag{4.8}$$

We introduce the downstream value $p^{i,j}_{[K,K']^-}$ to $[K,K']$ with respect to $\vec{a}$, i.e., if $a_{[K,K']} \leq 0$, then $p^{i,j}_{[K,K']^-} = p^i_K$ else $p^{i,j}_{[K,K']^-} = p^j_{K'}$. We transform the term $a_{[K,K']} p^{i,j}_{[K,K']^+}(p^i_K - p^j_{K'})$ as follows:

$$a_{[K,K']} p^{i,j}_{[K,K']^+}(p^i_K - p^j_{K'}) = |a_{[K,K']}| p^{i,j}_{[K,K']^+}(p^{i,j}_{[K,K']^+} - p^{i,j}_{[K,K']^-}).$$

Then (4.8) becomes

$$\begin{aligned}
T_1 &= \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} |a_{[K,K']}| p^{i,j}_{[K,K']^+}\big(p^{i,j}_{[K,K']^+} - p^{i,j}_{[K,K']^-}\big) \\
&= \frac{1}{4}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} |a_{[K,K']}|\big[\big(p^{i,j}_{[K,K']^+} - p^{i,j}_{[K,K']^-}\big)^2\big] \\
&\quad + \big(\big(p^{i,j}_{[K,K']^+}\big)^2 - \big(p^{i,j}_{[K,K']^-}\big)^2\big),
\end{aligned}$$

and then we write

$$\begin{aligned}
T_1 &\geq \frac{1}{4}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} |a_{[K,K']}|\big[\big(p^{i,j}_{[K,K']^+}\big)^2 - \big(p^{i,j}_{[K,K']^-}\big)^2\big] \\
&\geq \frac{1}{4}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} \big[a_{[K,K']}(p^i_K)^2 + a_{[K',K]}(p^j_{K'})^2\big] \\
&\geq \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} a_{[K,K']}(p^i_K)^2
\end{aligned}$$

$$\geq \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_{ij}(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_K\right](p_K^i)^2 = \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_K\right](p_K^i)^2.$$

Let us denote by T_2 the convective terms that are not on the interfaces. Using the same method for the convective term inside the subdomains and on the Dirichlet boundary as in [21], we have

$$T_2 \geq \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\left[\sum_{\epsilon\in\mathcal{E}(K)\backslash\mathcal{E}_i(K)}\int_\epsilon \vec{a}\cdot\vec{n}_{K(\epsilon)}\right](p_{K(\epsilon)}^i)^2 + \sum_{\epsilon\in\mathcal{E}_{iD}} 1_{[a_{\epsilon,K}\leq 0]}\; a_{\epsilon,K}g_\epsilon^i p_K^i.$$

Consequently for the convective term, we obtain thanks to $\mathrm{div}(\vec{a})\geq 0$

$$\begin{aligned}T_1+T_2 &\geq \sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_{K(\epsilon)}\right](p_{K(\epsilon)}^i)^2 + \sum_{\epsilon\in\mathcal{E}_{iD}} 1_{[a_{\epsilon,K}\leq 0]}\; a_{\epsilon,K}g_\epsilon^i p_K^i \\ &\geq \sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\int_K[\mathrm{div}(\vec{a})(p_{K(\epsilon)}^i)^2] + \sum_{\epsilon\in\mathcal{E}_{iD}} 1_{[a_{\epsilon,K}\leq 0]}\; a_{\epsilon,K}g_\epsilon^i p_K^i \\ &\geq \sum_{\epsilon\in\mathcal{E}_{iD}} 1_{[a_{\epsilon,K}\leq 0]}\; a_{\epsilon,K}g_\epsilon^i p_K^i.\end{aligned}$$

(4.6) leads to

$$\begin{aligned}&\sum_{i\in\mathcal{I}}\Bigg(\sum_{K\in\mathcal{T}_i}\Bigg(\eta\, meas(K)(p_K^i)^2 + \frac{1}{2}\sum_{K'\in\mathcal{N}_i(K)}\frac{(p_K^i-p_{K'}^i)^2}{d(x_K,x_{K'})}meas([K,K']) \\ &\quad + \sum_{\epsilon\in\mathcal{E}_{iD}(K)}\frac{(p_K^i)^2}{d(y_\epsilon,x_K)}meas(\epsilon) + \sum_{\epsilon\in\mathcal{E}_i(K)} d(y_\epsilon,x_K)(u_\epsilon^i)^2 meas(\epsilon)\Bigg)\Bigg) \\ &\quad \leq \sum_{i\in\mathcal{I}}\Bigg(\sum_{K\in\mathcal{T}_i}F_K p_K^i meas(K) + \sum_{\epsilon\in\mathcal{E}_{iD}}\frac{p_K^i g_\epsilon^i}{d(y_\epsilon,x_K)}meas(\epsilon) \\ &\quad - \sum_{\epsilon\in\mathcal{E}_{iD}} 1_{[a_{\epsilon,K}\leq 0]}\; a_{\epsilon,K}g_\epsilon^i p_K^i\Bigg).\end{aligned} \tag{4.9}$$

For the homogeneous problem ($(F_K)_{K\in\mathcal{T}_i}=0$ and $(g_\epsilon^i)_{\epsilon\in\mathcal{E}_{iD}}=0$), all the terms on the right-hand side of (4.9) are nonnegative. We conclude that for all $i\in\mathcal{I}$

$$p_K^i = 0\ \forall K\in\mathcal{T}_i \tag{4.10}$$

$$u_\epsilon^i = 0\ \forall \epsilon\in\mathcal{E}_i. \tag{4.11}$$

Using (2.14), (4.10), and (4.11), we obtain $p_\epsilon^i = 0$ for all $\epsilon\in\mathcal{E}_i$. Consequently the global system is invertible and the wellposedness of the scheme for the global domain is proven.

Moreover if $(g_\epsilon^i)_{\epsilon\in\mathcal{E}_{iD}}=0$ for all $i\in\mathcal{I}$ and using the Young inequality, for all $C_f>0$, we have for all $i\in\mathcal{I}$

$$\begin{aligned}\sum_{K\in\mathcal{T}_i}F_K p_K^i meas(K) &\leq \frac{1}{2C_f}\sum_{K\in\mathcal{T}_i}(p_K^i)^2 meas(K) + \frac{C_f}{2}\sum_{K\in\mathcal{T}_i}(F_K)^2 meas(K) \\ &\leq \frac{1}{2C_f}\|p_i\|^2_{L^2(\Omega_i)} + \frac{C_f}{2}\|F\|^2_{L^2(\Omega_i)}.\end{aligned} \tag{4.12}$$

Then (4.9)–(4.12) give

$$\left(\eta - \frac{1}{2C_f}\right)\sum_{i\in\mathcal{I}} \|p_i\|^2_{L^2(\Omega_i)} + \sum_{i\in\mathcal{I}}\Bigg(\frac{1}{2} \sum_{K'\in\mathcal{N}_i(K)} \frac{(p^i_K - p^i_{K'})^2}{d(x_K,x_{K'})} meas([K,K'])$$
$$+ \sum_{\epsilon\in\mathcal{E}_{iD}(K)} \frac{(p^i_K)^2}{d(y_\epsilon,x_K)} meas(\epsilon) + \sum_{\epsilon\in\mathcal{E}_i(K)} d(y_\epsilon,x_K)(u^i_\epsilon)^2 meas(\epsilon)\Bigg) \le \frac{C_f}{2}\sum_{i\in\mathcal{I}}\|F\|^2_{L^2(\Omega_i)}.$$

Taking $C_f > 1/\eta$ with (2.14) we obtain with $C \geq C_f$

$$\frac{\eta}{2}\|p\|^2_{L^2(\Omega)} + \frac{1}{2}|p|^2_{1,\mathcal{T}} \le C\|F\|^2_{L^2(\Omega)}. \qquad \square \tag{4.13}$$

4.2. Local wellposedness. We want to solve the global problem on Ω with an iterative domain decomposition method based on the computation of successive local boundary problems on Ω_i for $i \in \mathcal{I}$. Given $(R_\epsilon)_{\epsilon\in\mathcal{E}_i}$ and $(C_\epsilon)_{\epsilon\in\cup_j\mathcal{E}^+_{ij}}$ where, for all $j \in \mathcal{V}_i$, $\mathcal{E}^+_{ij}$ is the subset of $\mathcal{E}_{ij}$ where $a_\epsilon < 0$. The local problem on Ω_i is defined by the boundary conditions on Γ_{ij} for all $j \in \mathcal{V}_i$, for all $\epsilon \in \mathcal{E}_{i\to j}$:

$$u^{i,j}_\epsilon + [Q^j_i S_{j,i} Q^i_j(p^j_i)]_\epsilon = R_\epsilon \quad \text{if } j \in \mathcal{S}_i, \tag{4.14}$$

$$\left[Q^j_i S^{-1}_{j,i} Q^i_j(u^j_i)\right]_\epsilon + p^{i,j}_\epsilon = R_\epsilon \quad \text{if } j \in \mathcal{M}_i, \tag{4.15}$$

and by (2.11)–(2.16) where in (2.11), we have for all $j \in \mathcal{V}_i$, for any $K' \in \mathcal{N}_{ij}(K)$,

$$v^{i,j}_{[K,K']} = \begin{cases} a_\epsilon\, p^i_{K(\epsilon)} & \text{if } a_\epsilon \geq 0 \\ C_\epsilon & \text{if } a_\epsilon < 0. \end{cases} \tag{4.16}$$

where $\epsilon = [K,K']$. We have the following theorem

THEOREM 1. *Assume Assumptions* 2 *and* 3 *hold. Then, there exists a unique* $p^i \in X_i$ *and* $(v^{i,j}_\epsilon)_{\epsilon\in\mathcal{E}_{i\to j}}$ *satisfying the local problem on* Ω_i *defined by* (2.11)–(2.16) *and* (4.14)–(4.16).

Proof. The system in p^i and $v^{i,j}_\epsilon$ is linear and square, so it suffices to prove that the solution of the homogeneous local problem is zero. The homogeneous form of the equation of Lemma 4.1 is

(4.17)
$$\sum_{K\in\mathcal{T}_i}\Bigg(\eta\, meas(K)(p^i_K)^2 + \frac{1}{2}\sum_{K'\in\mathcal{N}_i(K)} \frac{(p^i_K - p^i_{K'})^2}{d(x_K,x_{K'})} d(x_K,x_{K'})\ meas([K,K'])$$
$$+\frac{1}{2}\sum_{\epsilon\in\mathcal{E}_{\Omega_i}(K)\setminus\mathcal{E}_i(K)} \left[\int_\epsilon \vec{a}\cdot\vec{n}_K\right](p^i_K)^2 + \sum_{\epsilon\in\mathcal{E}_{iD}(K)} \frac{(p^i_K)^2}{d(y_\epsilon,x_K)} meas(\epsilon)$$
$$+ \sum_{\epsilon\in\mathcal{E}_i(K)} d(y_\epsilon,x_K)(u^i_\epsilon)^2 meas(\epsilon) + \sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} v^{i,j}_{[K,K']} p^i_K\Bigg) \le \sum_{\epsilon\in\mathcal{E}_i} u^i_\epsilon p^i_\epsilon meas(\epsilon).$$

The homogeneous boundary conditions on $\mathcal{E}_{i\to j}$ are

$$\begin{aligned} p^j_i + Q^j_i(S^{-1}_{j,i}Q^i_j(u^j_i)) &= 0 \quad \forall j \in \mathcal{S}_i, \\ u^j_i + Q^j_i(S_{j,i}Q^i_j(p^j_i)) &= 0 \quad \forall j \in \mathcal{M}_i \end{aligned}$$

and give

$$\begin{aligned} p_i^j &= -Q_i^j(S_{j,i}^{-1}Q_j^i(u_i^j)) \quad \forall j \in \mathcal{S}_i,\\ u_i^j &= -Q_i^j(S_{j,i}Q_j^i(p_i^j)) \quad \forall j \in \mathcal{M}_i \end{aligned}$$

so the term on the interface reads

$$\begin{aligned} \sum_{\epsilon\in\mathcal{E}_i} u_\epsilon^i p_\epsilon^i meas(\epsilon) &= \int_{\Gamma_i} u_i p_i = \sum_{j\in\mathcal{V}_i}\int_{\Gamma_{ij}} u_i^j p_i^j = \sum_{j\in\mathcal{S}_i}\int_{\Gamma_{ij}} u_i^j p_i^j + \sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}} u_i^j p_i^j\\ &= -\sum_{j\in\mathcal{S}_i}\int_{\Gamma_{ij}} Q_i^j(S_{j,i}Q_j^i(p_i^j))p_i^j - \sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}} Q_i^j(S_{j,i}^{-1}Q_j^i(u_i^j))u_i^j\\ &= -\sum_{j\in\mathcal{S}_i}\int_{\Gamma_{ij}} S_{j,i}Q_j^i(p_i^j)Q_j^i(p_i^j) - \sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}} S_{j,i}^{-1}Q_j^i(u_i^j)Q_j^i(u_i^j) \le 0 \end{aligned}$$

because Q_i^j and Q_j^i satisfy Assumption 2 and $S_{j,i}$ satisfies Assumption 3. For the homogeneous convective term on the interface, we have

$$\begin{aligned} T_1 &= \sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} 1_{[a_{[K,K']}\ge 0]}a_{[K,K']}(p_K^i)^2\\ &\ge \frac{1}{2}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} 1_{[a_{[K,K']}\ge 0]}a_{[K,K']}(p_K^i)^2 \ge \frac{1}{2}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} a_{[K,K']}(p_K^i)^2\\ &\ge \frac{1}{2}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{K'\in\mathcal{N}_{ij}(K)}\int_{[K,K']}\vec{a}\cdot\vec{n}_K(p_K^i)^2 \ge \frac{1}{2}\sum_{j\in\mathcal{V}_i}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}\int_\epsilon \vec{a}\cdot\vec{n}_K(p_{K(\epsilon)}^i)^2. \end{aligned}$$

For the convective term inside the subdomains Ω_i and on the homogeneous Dirichlet boundary, we have

$$T_2 = \frac{1}{2}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_(K)\setminus\mathcal{E}_i(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_{K(\epsilon)}\right](p_K^i)^2.$$

So using $\operatorname{div}(\vec{a}) \ge 0$, we have

$$\begin{aligned} T_1 + T_2 &\ge \frac{1}{2}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_{K(\epsilon)}\right](p_K^i)^2\\ &= \sum_{K\in\mathcal{T}_i}\int_K[\operatorname{div}(\vec{a})(p_{K(\epsilon)}^i)^2] \ge 0. \end{aligned}$$

Then inequality (4.17) could be rewritten as

$$\begin{aligned} \sum_{K\in\mathcal{T}_i}\Bigg(\eta\, meas(K)(p_K^i)^2 &+ \frac{1}{2}\sum_{K'\in\mathcal{N}_i(K)}\frac{(p_K^i - p_{K'}^i)^2}{d(x_K,x_{K'})}d(x_K,x_{K'})\ meas([K,K'])\\ &+ \sum_{\epsilon\in\mathcal{E}_{iD}(K)}\frac{(p_K^i)^2}{d(y_\epsilon,x_K)}\ meas(\epsilon) + \sum_{\epsilon\in\mathcal{E}_i(K)} d(y_\epsilon,x_K)(u_\epsilon^i)^2\ meas(\epsilon)\Bigg) \le 0. \end{aligned} \tag{4.18}$$

All the left-hand-side terms of (4.18) are ≥ 0, so this implies

$$p_K^i = 0 \qquad \forall K \in \mathcal{T}_i, \tag{4.19}$$

$$u_\epsilon^i = 0 \qquad \forall \epsilon \in \mathcal{E}_i. \tag{4.20}$$

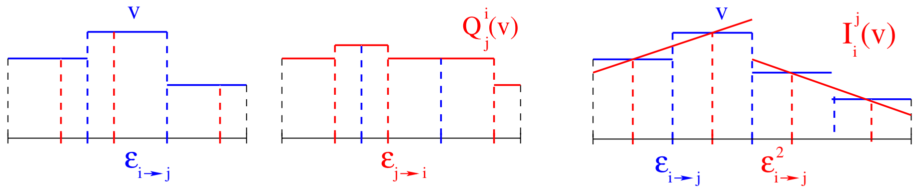


FIG. 2. *Projection on $P^0(\mathcal{E}_{j\to i})$ and interpolation on $P^1_d(\mathcal{E}^2_{i\to j})$ via the coarsened grid.*

Using (2.14), (4.19), and (4.20), we obtain $p^i_\epsilon = 0$ for all $\epsilon \in \mathcal{E}_i$. Then, from (4.16), we have $v^{i,j}_{[K,K']} = 0$ for $j \in \mathcal{V}_i$, $K' \in \mathcal{N}_{ij}(K)$. So the solution of the homogeneous problem on Ω_i is zero. □

5. Examples of transmission operators. We give two types of transmission operators which satisfy Assumption 2, but other transmission operators may be used. In the next section, an error estimate is performed for general transmission operators and then applied to the two types of transmission operators defined in the following.

5.1. Orthogonal L^2 projection on $P^0(\mathcal{E}_{i\to j})$. For $i \in \mathcal{I}$, $j \in \mathcal{V}_i$, let $P^C_{i,j}$ be the L^2 orthogonal projection onto $P^0(\mathcal{E}_{i\to j})$. The first type of operator is the restriction of $P^C_{i,j}$ to $P^0(\mathcal{E}_{j\to i})$. Let $u_j \in P^0(\mathcal{E}_{j\to i})$ and $\epsilon \in \mathcal{E}_{i\to j}$; it is defined by

$$[P^C_{i,j}(u_j)]_\epsilon = \frac{1}{meas(\epsilon)}\int_\epsilon u_j = \sum_{\epsilon'\in\mathcal{E}_j} \frac{meas(\epsilon\cap\epsilon')}{meas(\epsilon)} u_{j,\epsilon'}. \tag{5.1}$$

Assumption 2 is satisfied.

5.2. Transmission operators with linear rebuilding. The second type of operator, inspired by [4], uses a linear rebuilding process to have a more accurate transmission; see Figure 2. We shall use the following assumption.

Assumption 4. For all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$, there exists a coarsened grid $\mathcal{E}^2_{i\to j}$ of $\mathcal{E}_{i\to j}$ (with half the number of edges in two dimensions and one fourth the number of faces in three dimensions) such that

- if $d = 2$, for all $\epsilon \in \mathcal{E}^2_{i\to j}$, there exists $(\epsilon_1, \epsilon_2) \in (\mathcal{E}_{i\to j})^2$ such that $\epsilon = \epsilon_1 \cup \epsilon_2$ and $\bar\epsilon_1 \cap \bar\epsilon_2 \neq \emptyset$
- if $d = 3$, for all $\epsilon \in \mathcal{E}^2_{i\to j}$, there exists $(\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4) \in (\mathcal{E}_{i\to j})^4$ such that $\epsilon = \cup^4_{k=1}\epsilon_k$ and for all $1 \le k \le 4$, $card(\{l | 1 \le l \le 4,\ l \neq k,\ \text{and } \dim(\bar\epsilon_l \cap \bar\epsilon_k) = 1\}) = 2$.

For $d = 3$, some examples of mesh $\mathcal{E}_{i\to j}$ which satisfy Assumption 4 are corner point geometry mesh or quasi-uniform rectangular mesh. We introduce $P^1_d(\mathcal{E}^2_{i\to j})$, the space of discontinuous piecewise linear functions on $\mathcal{E}^2_{i\to j}$ if $d = 2$, or the space of discontinuous piecewise bilinear functions on $\mathcal{E}^2_{i\to j}$ if $d = 3$. The interpolation operator $I^j_i : P^0(\mathcal{E}_{i\to j}) \longrightarrow P^1_d(\mathcal{E}^2_{i\to j})$ is defined for all $p \in P^0(\mathcal{E}_{i\to j})$ by $[I^j_i(p)](y_\epsilon) = p(y_\epsilon)$, where y_ϵ is the barycenter of $\epsilon \in \mathcal{E}_{i\to j}$ and $(I^j_i)^T : P^1_d(\mathcal{E}^2_{i\to j}) \longrightarrow P^0(\mathcal{E}_{i\to j})$ is its transpose with respect to the $L^2(\Gamma_{ij})$ inner product. We also introduce $P^L_{i,j}$, the $L^2(\Gamma_{ij})$ orthogonal projection onto $P^1_d(\mathcal{E}^2_{i\to j})$. For $i \in \mathcal{I}$, $j \in \mathcal{S}_i$, the transmission operator is

$$Q^i_j = [P^C_{j,i|P^1_d(\mathcal{E}^2_{i\to j})}] I^j_i \qquad \text{(for the primary unknown)}. \tag{5.2}$$

In order to satisfy Assumption 2, we have

$$Q_i^j = (I_i^j)^T [P^L_{i,j|P^0(\mathcal{E}_{j\to i})}] \quad \text{(for the normal derivative).} \tag{5.3}$$

When the grids match, Q_i^j is the identity operator.

6. Error estimates. In this section, we study the accuracy of the finite volume scheme defined by (2.11)–(2.18) on nonmatching grids. The first subsection is an error estimate without taking into account the terms of the interface Γ_{ij}. The second subsection proves the error estimate in the general case with an additional assumption on transmission errors. In the other subsections we compute the transmission error in order to obtain the error estimate in the discrete H^1-norm. For the L^2 orthogonal projection on piecewise constant functions, we obtain an error in $O(h)^{1/2}$ in the general case and in $O(h)$ if the master sides are subgrids of the slave sides (subsection 5.1) and for the transmission operator with linear rebuilding (subsection 5.2), we get an error in $O(h)$. We denote the interface interpolation operators for all $i \in \mathcal{I}$, for all $j \in \mathcal{V}_i$ on Γ_{ij}, by

$$\begin{aligned} \mathcal{P}_i^j &: C^2(\overline{\Omega}) \longrightarrow P^0(\mathcal{E}_{i\to j}), \\ \mathcal{U}_i^j &: C^2(\overline{\Omega}) \longrightarrow P^0(\mathcal{E}_{i\to j}). \end{aligned}$$

We note $\mathcal{P}_\epsilon^{i,j}(p) = [\mathcal{P}_i^j(p)]_\epsilon$ and $\mathcal{U}_\epsilon^{i,j}(\frac{\partial f}{\partial n_i}) = [\mathcal{U}_i^j(\frac{\partial f}{\partial n_i})]_\epsilon$. These error interpolation operators allow us to introduce with flexibility an interpolated solution of p, the exact solution of (2.1)–(2.2), and of its normal derivative on Γ_{ij} (for all $i \in \mathcal{I}$, for all $j \in \mathcal{V}_i$) which is used to define the discrete error. The interpolated solution is defined by the following, for all $i \in \mathcal{I}$:

- $\tilde{p}_K^i = p(x_K)$ for any cell K in $\mathcal{T}_i$.
- $\tilde{p}_\epsilon^i = p(y_\epsilon)$ and $\tilde{u}_\epsilon^i = \frac{\partial p}{\partial n_i}(y_\epsilon)$ for any $\epsilon \in \mathcal{E}_{iD}$.
- For all $K \in \mathcal{T}_i$, for all $K' \in \mathcal{N}_i(K)$, if $a_{[K,K']} \geq 0$, then $\tilde{p}^i_{[K,K']+} = p(x_K)$; else $\tilde{p}^i_{[K,K']+} = p(x_{K'})$.
- For all $K \in \mathcal{T}_i$, for all $\epsilon \in \mathcal{E}_{iD}(K)$, if $a_{\epsilon,K} \geq 0$, then $\tilde{p}^i_{\epsilon,K+} = p(x_K)$; else $\tilde{p}^i_{\epsilon,K+} = p(y_\epsilon)$.
- For all $K \in \mathcal{T}_i$, for all $j \in \mathcal{V}_i$, for all $K \in \mathcal{N}_{ij}(K)$, if $a_{[K,K']} \geq 0$, then $\tilde{p}^{i,j}_{[K,K']+} = p(x_K)$; else $\tilde{p}^{i,j}_{[K,K']+} = p(x_{K'})$.

For the interface edges, the interpolated solution is adapted to the transmission operators:

- for $j \in \mathcal{V}_i$ and $\epsilon \in \mathcal{E}_{i\to j}$, $\tilde{p}_\epsilon^i = \mathcal{P}_i^j(p)_\epsilon$ and $\tilde{u}_\epsilon^i = \mathcal{U}_i^j(\frac{\partial p}{\partial n_i})_\epsilon$.

The discrete errors are defined by $e_K^i = p_K^i - \tilde{p}_K^i$ (for all $K \in \mathcal{T}_i$), $e_\epsilon^i = p_\epsilon^i - \tilde{p}_\epsilon^i$ (for all $\epsilon \in \mathcal{E}_{iD} \cup \mathcal{E}_i$), $q_\epsilon^i = u_\epsilon^i - \tilde{u}_\epsilon^i$ (for all $\epsilon \in \mathcal{E}_i \cup \mathcal{E}_{iD}$), $e^i_{[K,K']+} = p^i_{[K,K']+} - \tilde{p}^i_{[K,K']+}$ (for all $K \in \mathcal{T}_i$, for all $K' \in \mathcal{N}_i(K)$), $e^i_{\epsilon,K+} = p^i_{\epsilon,K+} - \tilde{p}^i_{\epsilon,K+}$ (for all $K \in \mathcal{T}_i$, for all $\epsilon \in \mathcal{E}_{iD}(K)$), and $e^{i,j}_{[K,K']+} = p^{i,j}_{[K,K']+}\tilde{p}^{i,j}_{[K,K']+}$ (for all $K \in \mathcal{T}_i$, for all $j \in \mathcal{V}_i$, for all $K' \in \mathcal{N}_{ij}(K)$). To simplify, we note for all $i \in \mathcal{I}$

- $R_K^i = \dfrac{1}{meas(K)} \displaystyle\int_K \eta(p_i - \tilde{p}_K^i)$ for all $K \in \mathcal{T}_i$,
- $R_{K,K'}^i = \dfrac{1}{meas([K,K'])} \displaystyle\int_{[K,K']} \frac{\partial p_i}{\partial n_i} - \frac{\tilde{p}^i_{K'} - \tilde{p}^i_K}{d(x_K, x_{K'})}$ for all $K \in \mathcal{T}_i$, for all $K' \in \mathcal{N}_i(K)$,
- $R_\epsilon^i = \dfrac{1}{meas(\epsilon)} \displaystyle\int_\epsilon \frac{\partial p_i}{\partial n_i} - \tilde{u}_\epsilon^i$ for all $\epsilon \in \mathcal{E}_i \cup \mathcal{E}_{iD}$,

- $r^i_{K,K'} = \dfrac{1}{meas([K,K'])}\displaystyle\int_{[K,K']} \vec{a}\cdot\vec{n}_K(p - \tilde{p}^i_{[K,K']^+})$ for all $K \in \mathcal{T}_i$ for all $K' \in \mathcal{N}_i$,
- $r^i_{K,\epsilon} = \dfrac{1}{meas(\epsilon)}\displaystyle\int_\epsilon \vec{a}\cdot\vec{n}_{K(\epsilon)}(p - \tilde{p}^i_{\epsilon,K^+})$ for all $K \in \mathcal{T}_i$, for all $\epsilon \in \mathcal{E}_{iD}(K)$,
- $r^{i,j}_{K,K'} = \dfrac{1}{meas([K,K'])}\displaystyle\int_{[K,K']} \vec{a}\cdot\vec{n}_K(p - \tilde{p}^{i,j}_{[K,K']^+})$ for all $K \in \mathcal{T}_i$, for all $j \in \mathcal{V}_i$, for all $K' \in \mathcal{N}_{ij}(K)$.

For $\epsilon \in \mathcal{E}_{i\to j}$, let us define

$$\begin{aligned}
T^i_\epsilon &= e^i_{K(\epsilon)} - e^i_\epsilon - d(x_K, y_\epsilon)q^i_\epsilon,\\
R^i_\epsilon &= \tfrac{1}{meas(\epsilon)}\textstyle\int_\epsilon \tfrac{\partial p}{\partial n_i} - \left[\mathcal{U}^j_i\left(\tfrac{\partial p}{\partial n}\right)\right]_\epsilon,\\
\Delta e^j_{i,\epsilon} &= \left[Q^j_i(\mathcal{P}^i_j(p)) - \mathcal{P}^j_i(p)\right]_\epsilon \quad \text{if } j \in \mathcal{M}_i,\\
\Delta q^j_{i,\epsilon} &= \left[Q^j_i(\mathcal{U}^i_j(\tfrac{\partial p}{\partial n_j})) + \mathcal{U}^j_i(\tfrac{\partial p}{\partial n_i})\right]_\epsilon \quad \text{if } j \in \mathcal{S}_i.
\end{aligned}$$

We introduce the global transmission errors:

- $R = \max_{i\in\mathcal{I},\epsilon\mathcal{E}_i} |R^i_\epsilon|$,
- $T = \max_{i\in\mathcal{I},\epsilon\mathcal{E}_i} |T^i_\epsilon|$,
- $T^2_d = \max_{i\in\mathcal{I},\epsilon\in\mathcal{E}_i} \frac{(T^i_\epsilon)^2}{d(x_{K(\epsilon)},y_\epsilon)}$,
- $\delta e = \max_{i\in\mathcal{I},j\in\mathcal{M}_i,\epsilon\in\mathcal{E}_i} |\Delta e^j_{i,\epsilon}|$,
- $\delta e^2_d = \max_{i\in\mathcal{I},j\in\mathcal{M}_i,\epsilon\in\mathcal{E}_i} \frac{(\Delta e^j_{i,\epsilon})^2}{d(x_{K(\epsilon)},y_\epsilon)}$,
- $\delta q = \max_{i\in\mathcal{I},j\in\mathcal{M}_i,\epsilon\in\mathcal{E}_j} |\Delta q^i_{j,\epsilon}|$.

6.1. Error analysis without the terms due to nonmatching grids.

In this section we give Lemma 6.1 in which we study the classical term of error without taking into account the term on the interface Γ_{ij} due to the nonmatching grids ($\mathcal{E}_{i\to j} \neq \mathcal{E}_{j\to i}$). This lemma will be used in the next subsection to estimate the interface error and thus the global error.

LEMMA 6.1. *Assuming that p the solution of* (2.1)–(2.2) *belongs to $C^2(\overline{\Omega})$ for each domain Ω_i $(i \in \mathcal{I})$, there exists $C > 0$ depending only on p, Ω_i, d, f, g, η such that*

(6.1)

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\Bigg(\frac{\eta}{2}meas(K)(e^i_K)^2 + \frac{1}{4}\sum_{K'\in\mathcal{N}_i(K)}\frac{(e^i_K - e^i_{K'})^2}{d(x_K,x_{K'})}meas([K,K'])\\
&\quad + \sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} a_{[K,K']}e^{i,j}_{[K,K']^+} + \frac{1}{2}\sum_{\epsilon\in\mathcal{E}_{iD}(K)}\frac{(e^i_K)^2}{d(x_K,y_\epsilon)}meas(\epsilon)\\
&\quad - \sum_{\epsilon\in\mathcal{E}_i(K)} q^i_\epsilon e^i_K meas(\epsilon) + \frac{1}{2}\sum_{\epsilon\in\mathcal{E}(K)\setminus\mathcal{E}_i(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_K\right](e^i_K)^2\Bigg)\\
&\quad \le Ch^2 + \sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\left(\sum_{\epsilon\in\mathcal{E}_i(K)} R^i_\epsilon e^i_K meas(\epsilon) + \sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} e^i_K r^{i,j}_{K,K'} meas([K,K'])\right).
\end{aligned}$$

The proof is based on standard arguments; see [3] or more generally [21].

6.2. Error analysis with interface terms.

In order to have as far as possible a common treatment of the error analysis for the two transmission operators we consider, we first derive a general estimate making the following extra assumption.

Assumption 5. There exist $\gamma_1 > 0$, $\gamma_2 > 0$, $\gamma_3 > 0$, $\gamma_4 > 0$, $\gamma_5 > 0$, and $\gamma_6 > 0$ such that $R = O(h)^{\gamma_1}$, $T = O(h)^{\gamma_2}$, $\delta e = O(h)^{\gamma_3}$, $\delta q = O(h)^{\gamma_4}$, $T_d^2 = O(h)^{\gamma_5}$, and $\delta e_d^2 = O(h)^{\gamma_6}$.

The values of the parameters γ depend on the transmission operator and will be analyzed in sections 6.3 and 6.4. To estimate the interface term due to the nonmatching grids, we need the following theorem which is proven in [15].

THEOREM 6.2. *For all $p \in X_i$, there exists $C(\Omega_i) > 0$ depending only on Ω_i and d such that*

$$\|p\|_{L^2(\partial\Omega_i)} \leq C(\Omega_i)(\|p\|_{L^2(\Omega_i)} + |p|_{1,\mathcal{T}_i}). \tag{6.2}$$

We shall also use the following formula (see [21]):

$$\sum_{K\in\mathcal{T}_i}\Big(\sum_{K'\in\mathcal{N}_i(K)} d(x_K, x_{K'})meas([K,K']) + \sum_{\epsilon\in\mathcal{E}_{iD}(K)\cup\mathcal{E}_i(K)} d(x_K, y_\epsilon)meas(\epsilon)\Big) = dmeas(\Omega_i). \tag{6.3}$$

We have the following theorem.

THEOREM 6.3. *Let us consider a family of admissible meshes $\mathcal{T}_i$ (for all $i \in \mathcal{I}$) which satisfy Assumption* 1. *We assume that the solution p of* (2.1)–(2.2) *is $C^2(\overline{\Omega})$ and that Assumptions* 2, 5, *and* 3 *are satisfied; then there exists $C > 0$ independent of h such that*

$$\big(\eta\|e\|^2_{L^2(\Omega)} + |e|^2_{1,\mathcal{T}}\big)^{1/2} \leq Ch^{\beta},$$

where $h = \max_{i\in\mathcal{I},K\in\mathcal{T}_i} diam(K)$ and $\beta = \frac{1}{2}\min(2, 2\gamma_1, \gamma_2+\gamma_3, 2\gamma_4, \gamma_5, \gamma_6)$.

Proof. Lemma 6.1 is satisfied for all $i \in \mathcal{I}$, so we sum over $i \in \mathcal{I}$ (6.1):

$$\begin{aligned}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\Bigg(&\frac{\eta}{2}meas(K)(e_K^i)^2 + \frac{1}{4}\sum_{K'\in\mathcal{N}_i(K)}\frac{(e_K^i - e_{K'}^i)^2}{d(x_K, x_{K'})}meas([K,K'])\\ &+\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} a_{[K,K']}e^{i,j}_{[K,K']^+} + \frac{1}{2}\sum_{\epsilon\in\mathcal{E}_{iD}(K)}\frac{(e_K^i)^2}{d(y_\epsilon, x_K)}meas(\epsilon)\\ &-\sum_{\epsilon\in\mathcal{E}_i(K)} q_\epsilon^i e_K^i meas(\epsilon) + \frac{1}{2}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_{\Omega_i}(K)\setminus\mathcal{E}_i(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_K\right](e_K^i)^2\Bigg)\\ \leq \sum_{i\in\mathcal{I}} C_i h^2 + \sum_{i\in\mathcal{I}}\Bigg(&\sum_{\epsilon\in\mathcal{E}_i} R_\epsilon^i e_{K(\epsilon)}^i meas(\epsilon) + \sum_{K\in\mathcal{T}_i}\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} r^{i,j}_{K,K'}e_K^i meas([K,K'])\Bigg).\end{aligned} \tag{6.4}$$

For the term $E_1 = \sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} R_\epsilon^i e_{K(\epsilon)}^i meas(\epsilon)$, we have used the Young inequality for all $C_6 > 0$:

$$\begin{aligned}E_1 &= \sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} R_\epsilon^i(e_{K(\epsilon)}^i - e_\epsilon^i)meas(\epsilon) + \sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} R_\epsilon^i e_\epsilon^i meas(\epsilon)\\ &\leq \sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} R_\epsilon^i e_\epsilon^i meas(\epsilon) + \frac{C_6}{2}\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} d(x_{K(\epsilon)}, y_\epsilon)(R_\epsilon^i)^2 meas(\epsilon)\\ &\quad+\frac{1}{2C_6}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}\frac{(e_\epsilon^i - e_K^i)^2}{d(x_K, y_\epsilon)}meas(\epsilon).\end{aligned} \tag{6.5}$$

The first right-hand-side term of (6.5) becomes, with the Young inequality for all $C_7 > 0$ and Theorem 6.2,

$$\begin{aligned} E_2 = \sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} R^i_\epsilon e^i_\epsilon meas(\epsilon) &\leq \frac{C_7}{2}\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i}(R^i_\epsilon)^2 meas(\epsilon) + \frac{1}{2C_7}\|e\|^2_{L^2(\partial\Omega_i)} \\ &\leq \frac{C_7}{2}\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i}(R^i_\epsilon)^2 meas(\epsilon) + \sum_{i\in\mathcal{I}}\frac{C(\Omega_i)}{2C_7}(\|e\|^2_{L^2(\Omega_i)} + |e|^2_{1,\mathcal{T}_i}). \end{aligned} \tag{6.6}$$

We estimate $E_3 = \sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} r^{i,j}_{K,K'} e^i_K meas([K,K'])$. For any positive constants C^1_c and C^2_c, we have

$$\begin{aligned} E_3 \leq &\frac{1}{2C^1_c}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)}\frac{(e^i_\epsilon - e^i_K)^2}{d(x_K,y_\epsilon)} meas([K,K']) \\ &+\frac{C^1_c}{2}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)} d(x_K,y_\epsilon)(r^{i,j}_{K,K'})^2 meas([K,K']) \\ &+\frac{1}{2C^2_c}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}(e^i_\epsilon)^2 meas(\epsilon) \\ &+\frac{C^2_c}{2}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)}(r^{i,j}_{K,K'})^2 meas([K,K']). \end{aligned}$$

By the Taylor expansion there exists $C > 0$ such that $|r^{i,j}_{K,K'}| \leq Ch$ so that together with (6.3), we have

$$\begin{aligned} E_3 \leq &\frac{1}{2C^1_c}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)}\frac{(e^i_\epsilon - e^i_K)^2}{d(x_K,y_\epsilon)} meas([K,K']) \\ &+\frac{C^1_c}{2}C^2h^2d\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i} meas(\Omega_j) \\ &+\frac{1}{2C^2_c}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}(e^i_\epsilon)^2 meas(\epsilon) + \frac{C^2_c}{2}C^2h^2\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i} meas(\Gamma_{ij}). \end{aligned}$$

With Theorem 6.2, we obtain

$$\begin{aligned} E_3 \leq &\frac{1}{2C^1_c}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{j\in\mathcal{V}_i}\sum_{K'\in\mathcal{N}_{ij}(K)}\frac{(e^i_\epsilon - e^i_K)^2}{d(x_K,y_\epsilon)} meas([K,K']) \\ &+\sum_{i\in\mathcal{I}}\frac{C(\Omega_i)}{2C^2_c}(\|e\|^2_{L^2(\Omega_i)} + |e|^2_{1,\mathcal{T}_i}) + \frac{C^1_c}{2}C^2h^2d\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i} meas(\Omega_j) \\ &+\frac{C_c}{2}C^2h^2\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{V}_i} meas(\Gamma_{ij}). \end{aligned} \tag{6.7}$$

For the term $T_1 = \sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)} e^i_K q^i_\epsilon meas(\epsilon)$, one has

$$\begin{aligned} T_1 &= -\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} q^i_\epsilon e^i_\epsilon meas(\epsilon) - \sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)} q^i_\epsilon(e^i_K - e^i_\epsilon) meas(\epsilon) \\ &= -\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i} q^i_\epsilon e^i_\epsilon meas(\epsilon) + \sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}\frac{(e^i_K - e^i_\epsilon)^2}{d(x_K,y_\epsilon)} meas(\epsilon) \end{aligned}$$

$$-\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}\frac{(e^i_K-e^i_\epsilon)T^i_\epsilon}{d(x_K,y_\epsilon)}meas(\epsilon).$$

Then by the Young inequality for all $C_8>0$,

$$\begin{aligned}T_1\geq{}&-\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i}q^i_\epsilon e^i_\epsilon meas(\epsilon)-\frac{C_8}{2}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}\frac{(T^i_\epsilon)^2}{d(x_K,y_\epsilon)}meas(\epsilon)\\&+\left(1-\frac{1}{2C_8}\right)\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}\frac{(e^i_K-e^i_\epsilon)^2}{d(x_K,y_\epsilon)}meas(\epsilon).\end{aligned}\tag{6.8}$$

Using the definition of the sets $\mathcal{M}_i$ and $\mathcal{S}_i$, the term $T_2=-\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i}q^i_\epsilon e^i_\epsilon meas(\epsilon)$ is rewritten as

$$T_2=-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}}\left[q^j_ie^j_i+q^i_je^i_j\right].\tag{6.9}$$

By definition we have on Γ_{ij} for all $i\in\mathcal{I}$ and for all $j\in\mathcal{V}_i$

$$e^{i,j}_\epsilon=p^{i,j}_\epsilon-\tilde{p}^{i,j}_\epsilon=p^{i,j}_\epsilon-\mathcal{P}^{i,j}_\epsilon(p).$$

Using the interface conditions (2.18), the discrete error on the primary unknown $e^{i,j}_\epsilon$ corresponds to, for all $j\in\mathcal{M}_i$,

$$\begin{aligned}e^{i,j}_\epsilon&=\left[Q^j_i(p^i_j)\right]_\epsilon-\mathcal{P}^{i,j}_\epsilon(p)=\left[Q^j_i\left(e^i_j+\mathcal{P}^i_j(p)\right)\right]_\epsilon-\mathcal{P}^{i,j}_\epsilon(p)\\&=\left[Q^j_i\left(e^i_j\right)\right]_\epsilon+\left[Q^j_i\left(\mathcal{P}^i_j(p)\right)\right]_\epsilon-\mathcal{P}^{i,j}_\epsilon(p)=\left[Q^j_i\left(e^i_j\right)\right]_\epsilon+\Delta e^{i,j}_\epsilon\end{aligned}\tag{6.10}$$

or, in compact form, $e^j_i=Q^j_i(e^i_j)+\Delta e^j_i$. In the same way, we have for all $j\in\mathcal{M}_i$

$$q^{j,i}_\epsilon=\left[Q^i_j\left(-q^j_i\right)\right]_\epsilon+\Delta q^{i,j}_\epsilon\tag{6.11}$$

or, in compact form, $q^i_j=Q^i_j(q^j_i)+\Delta q^i_j$. From (6.9)–(6.11) and Assumption 2, we have

$$T_2=-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}}\left[q^j_i\Delta e^j_i+\Delta q^j_ie^i_j\right].\tag{6.12}$$

By the Young inequality, for all $C_9>0$, and by definition of T^ϵ_i, we have

$$\begin{aligned}T_3={}&-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}}q^j_i\Delta e^j_i=-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{j\to i}}q^{i,j}_\epsilon\Delta e^{i,j}_\epsilon meas(\epsilon)\\\geq{}&-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{j\to i}}|T^i_\epsilon||\Delta e^{i,j}_\epsilon|meas(\epsilon)-\frac{1}{2C_9}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{j\to i}}\frac{(e^i_\epsilon-e^i_{K(\epsilon)})^2}{d(x_{K(\epsilon)},y_\epsilon)}meas(\epsilon)\\&-\frac{C_9}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{j\to i}}\frac{(\Delta e^{i,j}_\epsilon)^2}{d(x_{K(\epsilon)},y_\epsilon)}meas(\epsilon).\end{aligned}\tag{6.13}$$

Using Theorem 6.2 on Ω_i, the second term becomes, with the Young inequality for all $C_{10}>0$,

$$\begin{aligned}-\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}}\Delta q^j_ie^i_j\geq{}&-\frac{1}{2C_{10}}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\|e\|^2_{L^2(\partial\Omega_j)}\\&-\frac{C_{10}}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{i\to j}}(\Delta q^{j,i}_\epsilon)^2meas(\Gamma_{ij}).\end{aligned}$$

With Theorem 6.2, we get

$$(6.14)\qquad -\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\int_{\Gamma_{ij}}\Delta q_i^j e_j^i \ \geq\ \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\frac{C(\Omega_j)}{2C_{10}}(\|e\|^2_{L^2(\Omega_j)}+|e|^2_{1,\mathcal{T}_j}) - \frac{C_{10}}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{i\to j}}(\Delta q_\epsilon^{j,i})^2 meas(\Gamma_{ij}).$$

For the convective fluxes on the interfaces we proceed as in the proof of Theorem 4.3: we have

$$(6.15)\qquad \sum_{i\in\mathcal{I},j\in\mathcal{V}_i,K\in\mathcal{T}_i,K'\in\mathcal{N}_{ij}(K)} a_{[K,K']}e^{i,j}_{[K,K']^+}e^i_K \geq \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{\epsilon\in\mathcal{E}_i}\left[\int_\epsilon \vec{a}\cdot\vec{n}_{K(\epsilon)}\right](e^i_{K(\epsilon)})^2.$$

Let T be the convective terms of (6.4); thanks to (6.15) and $\operatorname{div}(\vec{a})\geq 0$, we have

$$T \geq \frac{1}{2}\sum_{i\in\mathcal{I}}\left(\sum_{\epsilon\in\mathcal{E}_i}\left[\int_\epsilon \vec{a}\cdot\vec{n}_{K(\epsilon)}\right](e^i_{K(\epsilon)})^2 + \sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}(K)\setminus\mathcal{E}_i(K)}\left[\int_\epsilon \vec{a}\cdot\vec{n}_K\right](e^i_K)^2\right)$$

$$(6.16)\qquad \geq \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{K\in\mathcal{T}_i}\int_K[\operatorname{div}(\vec{a})](e^i_{K(\epsilon)})^2 \geq 0.$$

So we obtain for the global error estimate, thanks to (6.4)–(6.6), (6.8), (6.13), (6.14), and (6.16),

$$\begin{aligned}
&\sum_{i\in\mathcal{I}}\left[\sum_{K\in\mathcal{T}_i}\left(\frac{\eta}{2}meas(K)(e^i_K)^2 + \frac{1}{4}\sum_{K'\in\mathcal{N}_i(K)}\frac{(e^i_K-e^i_{K'})^2}{d(x_K,x_{K'})}meas([K,K'])\right.\right.\\
&\quad+\left(1-\frac{1}{2C_6}-\frac{1}{2C_8}-\frac{1}{2C_9}\right)\sum_{\epsilon\in\mathcal{E}_i(K)}\frac{(e^i_\epsilon-e^i_K)^2}{d(y_\epsilon,x_K)}meas(\epsilon)\\
&\quad\left.\left.+\frac{1}{2}\sum_{\epsilon\in\mathcal{E}_{iD}(K)}\frac{(e^i_K)^2}{d(y_\epsilon,x_K)}meas(\epsilon)\right)\right]-\sum_{i\in\mathcal{I}}C(\Omega_i)\left(\frac{1}{2C_7}+\frac{1}{2C_{10}}\right)(\|e\|^2_{L^2(\Omega_i)}+|e|^2_{1,\mathcal{T}_i})\\
&\leq \sum_{i\in\mathcal{I}}\left[C_ih^2+\frac{C_7}{2}\sum_{\epsilon\in\mathcal{E}_i}(R^i_\epsilon)^2meas(\epsilon)+\frac{C_6}{2}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}d(x_K,y_\epsilon)(R^i_\epsilon)^2meas(\epsilon)\right.\\
&\quad\left.+\frac{C_8}{2}\sum_{K\in\mathcal{T}_i}\sum_{\epsilon\in\mathcal{E}_i(K)}\frac{(T^i_\epsilon)^2}{d(x_K,y_\epsilon)}meas(\epsilon)\right]+\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{j\to i}}|T^i_\epsilon||\Delta e^{i,j}_\epsilon|meas(\epsilon)\\
&\quad+\frac{C_9}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{j\to i}}\frac{(\Delta e^{i,j}_\epsilon)^2}{d(x_{K(\epsilon)},y_\epsilon)}meas(\epsilon)+\frac{C_{10}}{2}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{M}_i}\sum_{\epsilon\in\mathcal{E}_{i\to j}}(\Delta q^{i,j}_\epsilon)^2meas(\epsilon).
\end{aligned}$$

We conclude the proof of Theorem 6.3 by using Assumption 5. □

We are now interested in the consistency of the diffusive fluxes. For all $\epsilon\in\mathcal{E}_i(K)$, using (2.14) and the Taylor expansion, we get

$$\begin{aligned}
T^i_\epsilon &= e^i_K - e^i_\epsilon + d(x_K,y_\epsilon)q^i_\epsilon = (p^i_K-\tilde{p}^i_K)-(p^i_\epsilon-\tilde{p}^i_\epsilon)+d(x_K,y_\epsilon)(u^i_\epsilon-\tilde{u}^i_\epsilon)\\
&= -\tilde{p}^i_K+\tilde{p}^i_\epsilon-d(x_K,y_\epsilon)\tilde{u}^i_\epsilon\\
&= -p(y_\epsilon)+d(x_K,y_\epsilon)\frac{\partial p}{\partial n_i}(y_\epsilon)+\tilde{p}^i_\epsilon-d(x_K,y_\epsilon)\tilde{u}^i_\epsilon+O(d(x_K,y_\epsilon))^2
\end{aligned}$$

$$= [\mathcal{P}^i_\epsilon(p) - p(y_\epsilon)] + d(x_K, y_\epsilon)\left[\frac{\partial p}{\partial n_i}(y_\epsilon) - \mathcal{U}^i_\epsilon\left(\frac{\partial p}{\partial n_i}\right)\right] + O(d(x_K, y_\epsilon))^2$$

$$= [\mathcal{P}^i_\epsilon(p) - p(y_\epsilon)] + d(x_K, y_\epsilon)\left[\frac{\partial p}{\partial n_i}(y_\epsilon) - \mathcal{U}^i_\epsilon\left(\frac{\partial p}{\partial n_i}\right)\right] + O(d(x_K, y_\epsilon))^2. \tag{6.17}$$

6.3. Error estimate for orthogonal L^2 projection on $P^0(\mathcal{E}_{i\to j})$. In order to give the error estimate for the transmission operators defined by (5.1), we define the interpolation error operator for all $f \in C^2(\overline{\Omega})$, $i \in \mathcal{I}$, and $j \in \mathcal{V}_i$, $\epsilon \in \mathcal{E}_{i\to j}$ as

$$\mathcal{P}^{i,j}_\epsilon(f) = [P^C_{i,j}(f)]_\epsilon = \frac{1}{meas(\epsilon)}\int_\epsilon f,$$

$$\mathcal{U}^{i,j}_\epsilon\left(\frac{\partial f}{\partial n_i}\right) = \left[Q^j_i\left(\frac{\partial f}{\partial n_i}\right)\right]_\epsilon = \left[P^C_{i,j}\left(\frac{\partial f}{\partial n_i}\right)\right]_\epsilon = \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial f}{\partial n_i}.$$

We have to estimate the transmission error of Assumption 5 due to the transmission operator (5.1) to state the following theorem.

THEOREM 6.4. *We assume that the solution p of* (2.1)–(2.2) *is $C^2(\overline{\Omega})$. Let us consider a family of admissible meshes $\mathcal{T}_i$ (for all $i \in \mathcal{I}$) which satisfy Assumption 1.*

We assume that the transmission operators are defined by (5.1)*, that the interface operators satisfy Assumption* 3*, and that $C > 0$ independent of the mesh such that*

$$\forall i \in \mathcal{I}, \forall \epsilon \in \mathcal{E}_i \quad diam(\epsilon) \le C(d(x_{K(\epsilon)}, y_\epsilon)); \tag{6.18}$$

then there exists $C_1 > 0$ such that

$$\left(\eta\|e\|^2_{L^2(\Omega)} + |e|^2_{1,\mathcal{T}}\right)^{1/2} \le C_1 h^{1/2}.$$

Moreover if for all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$, $\mathcal{E}_{i\to j}$ is a subgrid of $\mathcal{E}_{j\to i}$ (the grids of the masters are subgrids of the slaves) and $(y_\epsilon)_{\epsilon\in\mathcal{E}_i}$ are the barycenter of ϵ and there exists $C > 0$ independent of the mesh such that

$$\forall i \in \mathcal{I}, \forall \epsilon \in \mathcal{E}_i \; diam(\epsilon) \le C(d(x_{K(\epsilon)}, y_\epsilon))^{1/2}, \tag{6.19}$$

then there exists $C_2 > 0$ such that

$$\left(\eta\|e\|^2_{L^2(\Omega)} + |e|^2_{1,\mathcal{T}}\right)^{1/2} \le C_2 h.$$

Remark 6.5. The assumption that the master sides are subgrids of the slave sides allows both a weaker assumption on the mesh ((6.19) is weaker than (6.18)) and a better error estimate. This result is assessed in the numerical tests; see Figure 6.

Proof. For all $i \in \mathcal{I}$, for all $\epsilon \in \mathcal{E}_i$, we have

$$R^i_\epsilon = \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial p}{\partial n_i} - \left[\mathcal{U}_i\left(\frac{\partial p}{\partial n_i}\right)\right]_\epsilon = 0.$$

Then $R = 0$ and we can take any $\gamma_1 > 0$. Using (6.17), we get

$$\begin{aligned} T^i_\epsilon &= \left[p(y_\epsilon) - \frac{1}{meas(\epsilon)}\int_\epsilon p\right] + d(x_{K(\epsilon)}, y_\epsilon)\left[\frac{\partial p_i}{\partial n_i}(y_\epsilon) - \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial p}{\partial n_i}\right] \\ &\quad - O(d(x_{K(\epsilon)}, y_\epsilon))^2 \\ &= [O(diam(\epsilon))] + d(x_{K(\epsilon)}, y_\epsilon)O(diam(\epsilon)) - O(d(x_{K(\epsilon)}, y_\epsilon))^2 = O(h). \end{aligned}$$

It implies that $\gamma_2 = 1$. For all $\epsilon \in \mathcal{E}_{i\to j}$, we write

$$\begin{aligned}\Delta e^j_{i,\epsilon} &= \left[Q^j_i\big(\mathcal{P}^i_j(p)\big) - \mathcal{P}^j_i(p)\right]_\epsilon = \left[P^C_{i,j}P^C_{j,i}(p) - P^C_{i,j}(p)\right]_\epsilon = \left[P^C_{i,j}\big(P^C_{j,i} - I\big)(p)\right]_\epsilon \\ &= \frac{1}{meas(\epsilon)}\int_\epsilon \big(P^C_{j,i} - I\big)(p) = O\big(\max_{\epsilon'\in\mathcal{E}_{j\to i}} diam(\epsilon')\big) = O(h).\end{aligned}$$

Hence we get $\gamma_3 = 1$. For all $\epsilon \in \mathcal{E}_{j\to i}$, we have

$$\begin{aligned}\Delta q^i_{j,\epsilon} &= \left[Q^i_j\left(\mathcal{U}^j_i\left(\frac{\partial p}{\partial n_i}\right)\right) + \mathcal{U}^i_j\left(\frac{\partial p}{\partial n_j}\right)\right]_\epsilon = \left[P^C_{j,i}P^C_{i,j}\left(\frac{\partial p}{\partial n_i}\right) + P^C_{j,i}\left(\frac{\partial p}{\partial n_j}\right)\right]_\epsilon \\ &= \left[P^C_{j,i}(P^C_{i,j} - I)\left(\frac{\partial p}{\partial n_j}\right)\right]_\epsilon \\ &= \frac{1}{meas(\epsilon)}\int_\epsilon (P^C_{i,j} - I)\left(\frac{\partial p}{\partial n_j}\right) = O\left(\max_{\epsilon'\in\mathcal{E}_{i\to j}} diam(\epsilon')\right) = O(h).\end{aligned}$$

It gives $\gamma_4 = 1$. Using (6.18), we have

$$\begin{aligned}\frac{(T^i_\epsilon)^2}{d(x_{K(\epsilon)}, y_\epsilon)} &= \frac{O(diam(\epsilon))^2}{d(x_{K(\epsilon)}, y_\epsilon)} + d(x_{K(\epsilon)}, y_\epsilon)O(diam(\epsilon))^2 + O(d(x_{K(\epsilon)}, y_\epsilon))^3 \\ &+O(diam(\epsilon)) + O(d(x_K, y_\epsilon))O(diam(\epsilon))^2 + O(d(x_{K(\epsilon)}, y_\epsilon))^2 O(diam(\epsilon)) = O(h).\end{aligned}$$

We have for $T^2_d = O(h)$, so we have $\gamma_5 = 1$. Using (6.18), we obtain the following estimate:

$$\frac{(\Delta e^{i,j}_\epsilon)^2}{d(x_{K(\epsilon)}, y_\epsilon)} = \frac{O(diam(\epsilon))^2}{d(x_{K(\epsilon)}, y_\epsilon)} = O(h).$$

We get $\delta e^2_d = O$ so that $\gamma_6 = 1$. Consequently due to $\gamma_6 = 1$ we obtain $\frac{1}{2}\min(2, 2\gamma_1, \gamma_2 + \gamma_3, 2\gamma_4, \gamma_5, \gamma_6) = \frac{1}{2}$. This proves the first part of Theorem 6.4.

As for the second part of Theorem 6.4, we assume that for all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$, $\mathcal{E}_{i\to j}$ is a subgrid of $\mathcal{E}_{j\to i}$ (the grids of the masters are subgrids of the slaves). This assumption implies that for all $j \in \mathcal{S}_i$, $P^C_{i,j} = P^C_{i,j}P^C_{j,i}$ and consequently

$$\Delta e^{i,j}_\epsilon = [P^C_{i,j}(P^C_{j,i} - I)(p)]_\epsilon = 0.$$

We get $\delta e = 0$ and $\delta e^2_d = 0$. For all $\epsilon \in \mathcal{E}_i$, y_ϵ is the barycenter of ϵ and p is $C^2(\overline{\Omega})$, so we get $p(\epsilon) - \frac{1}{meas(\epsilon)}\int_\epsilon p = O(diam(\epsilon))^2$, and consequently the consistency error on the fluxes becomes

$$\begin{aligned}T^i_\epsilon &= [p(y_\epsilon) - \mathcal{P}^i_\epsilon(p)] + d(x_{K(\epsilon)}, y_\epsilon)\left[\frac{\partial p_i}{\partial n_i}(y_\epsilon) - \mathcal{U}^i_\epsilon\left(\frac{\partial p}{\partial n_i}\right)\right] - O(d(x_{K(\epsilon)}, y_\epsilon))^2 \\ &= \left[p(y_\epsilon) - \frac{1}{meas(\epsilon)}\int_\epsilon p\right] + d(x_{K(\epsilon)}, y_\epsilon)\left[\frac{\partial p_i}{\partial n_i}(y_\epsilon) - \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial p}{\partial n_i}\right] - O(h)^2 \\ &= [O(diam(\epsilon))^2] + d(x_{K(\epsilon)}, y_\epsilon)O(diam(\epsilon)) - O(h)^2 = O(h)^2;\end{aligned}$$

then $\gamma_2 = 2$. For T^2_d, we have using (6.19)

$$\frac{(T^i_\epsilon)^2}{d(x_{K(\epsilon)}, y_\epsilon)} = \frac{O(diam(\epsilon))^4}{d(x_{K(\epsilon)}, y_\epsilon)} + d(x_{K(\epsilon)}, y_\epsilon)O(diam(\epsilon))^2 + O(d(x_{K(\epsilon)}, y_\epsilon))^3 = O(h^2).$$

Then γ_1 and γ_4 are unchanged, $\gamma_2 = 2$, $\gamma_5 = 2$, and we can take any $\gamma_3 > 0$ and $\gamma_6 > 0$ (for example, $\gamma_3 = \gamma_6 = 2$), which gives

$$\frac{1}{2}\min(2, 2\gamma_1, \gamma_2 + \gamma_3, 2\gamma_4, \gamma_5, \gamma_6) = 1. \qquad \square$$

6.4. Error estimate for transmission operators with linear rebuilding. For the error estimate with transmission operators defined by (5.3) and (5.2), we assume the following geometrical additional assumption is satisfied.

Assumption 6. For all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$, the grid of the master $\mathcal{E}_{i\to j}$ is Cartesian and there exists δ independent of the mesh such that for all $\epsilon \in \mathcal{E}^2_{i\to j}$ $\max_{\epsilon' \in \mathcal{E}_{i\to j} | \epsilon' \subset \epsilon} \text{meas}(\epsilon') \le \delta \min_{\epsilon' \in \mathcal{E}_{i\to j} | \epsilon' \subset \epsilon} \text{meas}(\epsilon')$.

We define a local operator $E^j_i = (I^j_i)^T - P^C_{i,j|P^0(\mathcal{E}_{j\to i})} = (I^j_i)^T - (I^j_i)^{-1}$ and prove the following lemma.

LEMMA 6.6. *For $d = \{2,3\}$ and under Assumption 6, for all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$, I^j_i is invertible, $(I^j_i)^{-1} = P^C_{i,j|P^1_d(\mathcal{E}_{i\to j})}$, $\ker(E^j_i) = P^0(\mathcal{E}^2_{i\to j})$, and there exists $C_E(\delta) > 0$ such that $\|E^j_i\|_{L^2(\Gamma_{ij})} < C_E(\delta)$.*

Proof. From the locality of the operators, it suffices to consider $\epsilon_1, \epsilon_2 \in (\mathcal{E}_{i\to j})^2$ such that $\epsilon = \epsilon_1 \cup \epsilon_2 \in \mathcal{E}^2_{i\to j}$ and the restriction $E^j_{i,\epsilon}$ of E^j_i from $P^0(\epsilon_1) \times P^0(\epsilon_2)$ to $P^1_d(\epsilon)$. Let

$$
\begin{aligned}
e_1(y) &= 1 \qquad \text{if } y \in \epsilon_1 \cup \epsilon_2,\\
e_2(y) &= \begin{cases} -meas(\epsilon_2) & \text{if } y \in \epsilon_1,\\ \ \ meas(\epsilon_1) & \text{if } y \in \epsilon_2,\end{cases}\\
f_1 &= I^j_i(e_1) = 1,\\
f_2 &= I^j_i(e_2) = 2y - (meas(\epsilon_2) - meas(\epsilon_1))/2,
\end{aligned}
$$

and $\tilde{e}_i = e_i / \|e_i\|_{L^2(\epsilon)}$, $\tilde{f}_i = f_i / \|f_i\|_{L^2(\epsilon)}$ for $i = 1, 2$. The bases $(\tilde{e}_1, \tilde{e}_2)$ of $P^0(\epsilon_1) \times P^0(\epsilon_2)$ and $(\tilde{f}_1, \tilde{f}_2)$ of $P^1_d(\epsilon)$ are orthonormal. In these bases, the matrix of $E^j_{i,\epsilon}$ is

$$
\frac{\|f_2\|_{L^2}}{\|e_2\|_{L^2}} \left[\begin{array}{cc} 0 & 0 \\ 0 & 1 - \dfrac{(meas(\epsilon_1) + meas(\epsilon_2))^2}{3meas(\epsilon_1)meas(\epsilon_2)} \end{array} \right].
$$

Since $1 - \frac{(meas(\epsilon_1)+meas(\epsilon_2))^2}{3meas(\epsilon_1)meas(\epsilon_2)} < 0$, we have $\ker(E^j_{i,\epsilon}) = \mathbb{R}e_1 = P^0(\epsilon)$. By Assumption 6, we have that there exists $C(\delta) > 0$ independent of ϵ such that $\|E^j_{i,\epsilon}\|_{L^2(\epsilon)} \le C(\delta)$. Lemma 6.6 holds locally for $E^j_{i,\epsilon}$ and therefore globally for E^j_i.

For $d = 3$, the proof is very similar. Let us just indicate that we consider in local coordinates four faces $\epsilon_1 = [-h_1, 0] \times [-k_1, 0], \epsilon_2 = [-h_1, 0] \times [0, k_2], \epsilon_3 = [0, h_2] \times [-k_1, 0], \epsilon_4 = [0, h_2] \times [0, k_2]$ such that $\cup^4_{i=1} \epsilon_i \in \mathcal{E}^2_{i\to j}$. It is convenient to introduce (e_1, e_2, e_3, e_4) a local basis of the space of piecewise constant functions $\prod^4_{i=1} P^0(\epsilon_i)$:

$$
\begin{array}{rclcrcl}
e_1(y,z) &=& 1 \text{ on } E = \cup^4_{i=1}\epsilon_i, & & e_2(y,z) &=& \begin{cases} -h_2 & \text{on } \epsilon_1,\\ -h_2 & \text{on } \epsilon_2,\\ \ \ h_1 & \text{on } \epsilon_3,\\ \ \ h_1 & \text{on } \epsilon_4,\end{cases}\\
e_3(y,z) &=& \begin{cases} -k_2 & \text{on } \epsilon_1,\\ \ \ k_1 & \text{on } \epsilon_2,\\ -k_2 & \text{on } \epsilon_3,\\ \ \ k_1 & \text{on } \epsilon_4\end{cases} & \text{and} & e_4(y,z) &=& \begin{cases} \ \ h_2k_2 & \text{on } \epsilon_1,\\ -k_1h_2 & \text{on } \epsilon_2,\\ -k_2h_1 & \text{on } \epsilon_3,\\ \ \ h_1k_1 & \text{on } \epsilon_4.\end{cases}
\end{array}
$$

As in the case $d = 2$, let $f_i = I^j_i(e_i)$, $\tilde{e}_i = e_i/\|e_i\|_{L^2}$, and $\tilde{f}_i = f_i / \|f_i\|_{L^2}$ for $i = 1, \ldots, 4$. The theorem is proved by considering the matrix of the restriction of E^j_i in the orthonormal bases $(\tilde{e}_i)_{i=1,\ldots,4}$ and $(\tilde{f}_i)_{i=1,\ldots,4}$. □

We now define the transmission interpolations operator by

$$\mathcal{P}_\epsilon^{i,j}(f) = P_{i,j}^C(f) = \frac{1}{meas(\epsilon)}\int_\epsilon f \quad \mathcal{U}_\epsilon^{i,j}\left(\frac{\partial f}{\partial n_i}\right) = P_{i,j}^C\left(\frac{\partial f}{\partial n_i}\right) = \frac{1}{meas(\epsilon)}\frac{\partial f}{\partial n_i}$$

for all $f \in C^2(\overline{\Omega})$, for all $i \in \mathcal{I}$, for all $j \in \mathcal{M}_i$, for any $\epsilon \in \mathcal{E}_{i\to j}$ and by

$$\mathcal{P}_\epsilon^{i,j}(f) = (I_i^j)^{-1}P_{i,j}^L(f) = \frac{1}{meas(\epsilon)}P_{i,j}^L(f) \qquad \mathcal{U}_\epsilon^{i,j}\left(\frac{\partial f}{\partial n_i}\right) = (I_i^j)^T P_{i,j}^L P_{j,i}^C\left(\frac{\partial f}{\partial n_i}\right)$$

for all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$.

THEOREM 6.7. *We assume that the solution p of* (2.1)–(2.2) *is $C^2(\overline{\Omega})$. Let us consider a family of admissible meshes for all $i \in \mathcal{I}$ $\mathcal{T}_i$ which satisfies Assumptions* 1–4 *and* 6. *Assume that for all $\epsilon \in \mathcal{E}_i$ y_ϵ is the barycenter of ϵ and that there exists $C' > 0$, independent of the family of meshes, such that*

$$\forall i \in \mathcal{I}, \forall \epsilon \in \mathcal{E}_i \quad diam(\epsilon) \le C'\, d(y_\epsilon, x_K)^{1/2}. \tag{6.20}$$

We assume that the transmission operators are defined by (5.3) *and* (5.2) *and that the interface operators satisfy Assumption* 3*; then there exists $C > 0$ such that*

$$\left(\eta\|e\|^2_{L^2(\Omega)} + |e|^2_{1,\mathcal{T}}\right)^{1/2} \le Ch.$$

Proof. For all $i \in \mathcal{I}$, for all $j \in \mathcal{M}_i$, for all $\epsilon \in \mathcal{E}_{i\to j}$, we have

$$R_\epsilon^i = \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial p}{\partial n_i} - \left[\mathcal{U}_i\left(\frac{\partial p}{\partial n_i}\right)\right]_\epsilon = 0,$$

and using the fact that y_ϵ is the barycenter of $\epsilon \in \mathcal{E}_i$

$$\begin{aligned} T_\epsilon^i &= [p(y_\epsilon) - P_{i,j}^C(p)] + d(x_{K(\epsilon)}, y_\epsilon)\left[\frac{\partial p}{\partial n_i}(y_\epsilon) - P_{i,j}^C\left(\frac{\partial p}{\partial n_i}\right)\right] - O(d(x_{K(\epsilon)}, y_\epsilon))^2 \\ &= O(diam(\epsilon))^2 + d(x_{K(\epsilon)}, y_\epsilon)O(diam(\epsilon)) - O(d(x_{K(\epsilon)}, y_\epsilon))^2 = O(h)^2. \end{aligned}$$

We introduce P_{ij}^C, the $L^2(\Gamma_{ij})$ orthogonal projection on $P^0(\mathcal{E}^2_{i\to j})$. Then we have

$$\begin{aligned} R_\epsilon^i &= \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial p}{\partial n_i} - \left[\mathcal{U}_i^j\left(\frac{\partial p}{\partial n_i}\right)\right]_\epsilon \\ &= \left[P_{i,j}^C\left(\frac{\partial p}{\partial n_i}\right) - (I_i^j)^{-1}P_{i,j}^L P_{j,i}^C\left(\frac{\partial p}{\partial n_i}\right)\right]_\epsilon \\ &\quad + \left[(I_i^j)^{-1}P_{i,j}^L P_{j,i}^C\left(\frac{\partial p}{\partial n_i}\right) - (I_i^j)^T P_{i,j}^L P_{j,i}^C\left(\frac{\partial p}{\partial n_i}\right)\right]_\epsilon. \end{aligned}$$

Thanks to Lemma 6.6, we have $(I_i^j)^{-1} = P^C_{i,j|P^1_d(\mathcal{E}^2_{i\to j})}$ and by definition $E_i^j = (I_i^j)^{-1} - (I_i^j)^T$ so that

$$R_\epsilon^i = \left[P_{i,j}^C\left(\frac{\partial p}{\partial n_i} - P_{i,j}^L P_{j,i}^C\left(\frac{\partial p}{\partial n_i}\right)\right)\right]_\epsilon + \left[E_i^j P_{i,j}^L P_{j,i}^C\left(\frac{\partial p}{\partial n_i}\right)\right]_\epsilon.$$

Since $\ker(E_i^j) = P^0(\mathcal{E}^2_{i\to j})$ and $P^C_{ij} P^L_{i,j} P^C_{j,i}(\frac{\partial p}{\partial n_i}) \in P^0(\mathcal{E}^2_{i\to j})$ we get

$$\begin{aligned}
R^i_\epsilon &= \left[P^C_{i,j}\left(\left(\frac{\partial p}{\partial n_i}\right) - P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right) + P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right) - P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right)\right]_\epsilon \\
&\quad + \left[E_i^j\left[P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right) - P^C_{ij}P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right]\right]_\epsilon \\
&= \left[P^C_{i,j}\left[(I - P^C_{j,i})\left(\frac{\partial p}{\partial n_i}\right)\right]\right]_\epsilon + \left[P^C_{i,j}\left[(I - P^L_{i,j})P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right]\right]_\epsilon \\
&\quad + \left[E_i^j\left[(I - P^C_{ij})P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right]\right]_\epsilon \\
&= O(\max_{\epsilon_j \in \mathcal{E}_{j\to i}} diam(\epsilon_j)) + \left[P^C_{i,j}\left[(I - P^L_{i,j})P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right]\right]_\epsilon \\
&\quad + \left[E_i^j\left[(I - P^C_{ij})P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right]\right]_\epsilon .
\end{aligned}$$

For all $i \in \mathcal{I}$, for all $j \in \mathcal{S}_i$, for all $\epsilon \in \mathcal{E}_{i\to j}$, we denote by ϵ_2 the unique $\epsilon_2 \in \mathcal{E}_{i\to j}$ such that $\epsilon \subset \epsilon_2$. The second term gives

$$\begin{aligned}
\left[P^C_{i,j}\left[(I - P^L_{i,j})P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right]\right]_\epsilon &= \frac{1}{meas(\epsilon)}\int_\epsilon \left[(I - P^L_{i,j})P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right] \\
&\le \frac{1}{(meas(\epsilon))^{1/2}}\left\|(I - P^L_{i,j})P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right\|_{L^2(\epsilon_2)} \\
&\le \left(\frac{meas(\epsilon_2)}{meas(\epsilon)}\right)^{1/2} O(diam(\epsilon_2))^2 \\
&\le O(diam(\epsilon))^2 \text{ (thanks to Assumption 6).}
\end{aligned}$$

By the definition of E_i^j, the third term becomes

$$\begin{aligned}
&\left[E_i^j\left[(I - P^C_{ij})P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right]\right]_\epsilon \\
&\quad = \frac{1}{meas(\epsilon)}\int_\epsilon E_i^j\left[(I - P^C_{ij})P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right] \\
&\quad \le \frac{1}{(meas(\epsilon))^{1/2}}\left\|E_i^j[(I - P^C_{ij})P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right\|_{L^2(\epsilon_2)} \\
&\quad \le \frac{1}{(meas(\epsilon))^{1/2}}\|E_i^j\|_{L^2(\epsilon_2)}\left\|(I - P^C_{ij})P^L_{i,j}P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right\|_{L^2(\epsilon_2)} \\
&\quad \le \left(\frac{meas(\epsilon_2)}{meas(\epsilon)}\right)^{1/2} C_E(\delta)O(diam(\epsilon_2)) = O(h).
\end{aligned}$$

The last line comes from Lemma 6.6. Thus we have $R^i_\epsilon = O(diam(\epsilon))$.

From (6.17), we have

$$\begin{aligned}
T^i_\epsilon &= [p(y_\epsilon) - \mathcal{P}^i_\epsilon(p)] + d(x_{K(\epsilon)}, y_\epsilon)\left[\frac{\partial p_i}{\partial n_i}(y_\epsilon) - \mathcal{U}^i_\epsilon\left(\frac{\partial p}{\partial n_i}\right)\right] - O(d(x_{K(\epsilon)}, y_\epsilon))^2 \\
&= [p(y_\epsilon) - \mathcal{P}^i_\epsilon(p)] + d(x_{K(\epsilon)}, y_\epsilon)
\end{aligned}$$

$$\times \left[\frac{\partial p_i}{\partial n_i}(y_\epsilon) - \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial p}{\partial n_i} + \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial p}{\partial n_i} - \mathcal{U}^i_\epsilon\left(\frac{\partial p}{\partial n_i}\right)\right] - O(h)^2$$

$$= [p(y_\epsilon) - (I^j_i)^{-1}P^L_{i,j}(p)] + d(x_{K(\epsilon)}, y_\epsilon)\left[\frac{\partial p_i}{\partial n_i}(y_\epsilon) - \frac{1}{meas(\epsilon)}\int_\epsilon \frac{\partial p}{\partial n_i} + R^i_\epsilon\right] - O(h)^2;$$

by assumption y_ϵ is the barycenter of $\epsilon \in \mathcal{E}_i$, so

$$= [p(y_\epsilon) - (I^j_i)^{-1}P^L_{i,j}(p)] + d(x_{K(\epsilon)}, y_\epsilon)[R^i_\epsilon + O(diam(\epsilon))^2] - O(h)^2;$$

using Lemma 6.6, we get

$$= [P^C_{i,j}[I - P^L_{i,j}](p)]_\epsilon + d(x_{K(\epsilon)}, y_\epsilon)O(diam(\epsilon)) - O(h)^2$$
$$= O(diam(\epsilon))^2 + d(x_{K(\epsilon)}, y_\epsilon)O(diam(\epsilon)) - O(d(x_{K(\epsilon)}, y_\epsilon))^2 = O(h)^2.$$

Thus for all $i \in \mathcal{I}$, for all $\epsilon \in \mathcal{E}_i$, we have shown that $R^i_\epsilon = O(h)$ and $T^i_\epsilon = O(h)^2$ and consequently $\gamma_1 = 1$ and $\gamma_2 = 2$. Now we estimate δe, δq, T^2_d, and δe^2_d only for all $i \in \mathcal{I}$, for all $j \in \mathcal{M}_i$, for all $\epsilon \in \mathcal{E}_{i\to j}$:

$$\Delta e^{i,j}_\epsilon = \left[Q^j_i(\mathcal{P}^i_j(f)) - \mathcal{P}^j_i(f)\right]_\epsilon = [P^C_{i,j}I^i_jP^C_{j,i}(p)] - P^C_{i,j}I^i_jP^C_{j,i}(p) = 0 \tag{6.21}$$

and similarly

$$\Delta q^{j,i}_\epsilon = \left[(I^i_j)^T P^L_{j,i}\left(P^C_{j,i}\left(\frac{\partial p}{\partial n_i}\right)\right) + (I^i_j)^T P^L_{j,i}\left(P^C_{j,i}\left(\frac{\partial p}{\partial n_j}\right)\right)\right]_\epsilon = 0. \tag{6.22}$$

(6.21) and (6.22) allow us to take any γ_3, $\gamma_4 > 0$, for example, $\gamma_3 = 2$ and $\gamma_4 = 2$. With (6.21) and (6.20), we estimate γ_5:

$$\frac{(T^i_\epsilon)^2}{d(x_{K(\epsilon)}, y_\epsilon)} = \frac{1}{d(x_{K(\epsilon)}, y_\epsilon)}[O(diam(\epsilon))^2 + d(x_{K(\epsilon)}, y_\epsilon)O(diam(\epsilon)) - O(d(x_{K(\epsilon)}, y_\epsilon))^2]^2$$
$$= \frac{1}{d(x_{K(\epsilon)}, y_\epsilon)}[O(diam(\epsilon))^2 + d(x_{K(\epsilon)}, y_\epsilon)O(h)]^2$$
$$= \frac{O(diam(\epsilon))^4}{d(x_{K(\epsilon)}, y_\epsilon)} + O(h)O(diam(\epsilon))^2 + d(x_{K(\epsilon)}, y_\epsilon)O(h)^2 = O(h)^2.$$

Hence we have $\gamma_5 = 2$. (6.21) implies that $\frac{(\Delta e^{i,j}_\epsilon)^2}{d(x_{K(\epsilon)}, y_\epsilon)} = 0$, and so we can take, for example, $\gamma_6 = 2$. We conclude that $\min(2, 2\gamma_1, \gamma_2 + \gamma_3, 2\gamma_4, \gamma_5, \gamma_6) = 2$. □

7. Numerical results. We consider the domain $\Omega =]0,1[\times]0,1[$ with the problem defined by $\eta = 1$, $f(x,y) = x^3y^2 - 6x^2y^2 - 2x^3 + (1 + x^2 + y^2)\sin(xy)$ and $g(x,y) = x^3y^2 + \sin(xy)$. The associated analytical solution is $p(x,y) = x^3y^2 + \sin(xy)$. We have tested a two-domain decomposition and a four-domain decomposition with a corner in order to compare the following methods on nonmatching grids: TPFA [9] (or [11]), Ceres (like TPFA but a linear interpolation is done to have a consistent flux on the interface; see [24], [8], which tested only for a two-domain decomposition because of an implementation too complex when there is a corner), New Cement [3], and our two methods associated with the transmission operators (5.1) (method Constant) and (5.3) and (5.2) (method Linear). In [3], the interface conditions (3.1) and (3.2) are replaced by a more symmetric form:

$$u_i + S_i(p_i) = Q_i(-u_j + S_i(p_j)) \qquad \text{on } \Gamma_{i,j}, \tag{7.1}$$

$$u_j + S_j(p_j) = Q_j(-u_i + S_j(p_i)) \qquad \text{on } \Gamma_{i,j} \tag{7.2}$$

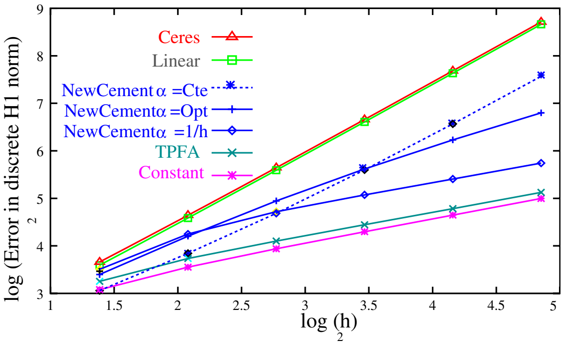


FIG. 3. *Two domains: log of the error H^1-norm vs. log of the mesh size (h).*

with the peculiarity that the solution depends on the choice of the Robin parameters in $S_{i,j}$. We study the dependency of the error estimate in discrete H^1-norm and of the number of GMRES iterations as a function of the mesh size and for Robin interface conditions corresponding to different diagonal interface operators: $S_i = \text{diag}(\alpha^i_\epsilon)$ with for all $\epsilon \in \mathcal{E}_i$ $\alpha^i_\epsilon = 1$ or $\alpha^i_\epsilon = \alpha^i_{opt,\epsilon} = O(1/\sqrt{h_i})$ or $\alpha^i_\epsilon = 1/h_i$ with h_i the size of the mesh of Ω_i.

7.1. Domain decomposition solver. The algorithm that we used is defined in detail in [23]. When no interface grid is a subgrid of the other, the Robin interface conditions (3.1) and (3.2) introduce a dependency of the interface conditions on Ω_i with the grid of the neighboring subdomain Ω_j. To cancel this dependency a subgrid based on $\mathcal{E}_{i\to j}$ and $\mathcal{E}_{j\to i}$ and new unknowns are introduced on Γ_{ij} to write arbitrary interface conditions equivalent to interface conditions (3.1) and (3.2). As a result, we get a new finite volume discretization equivalent to (2.11)–(2.13), (3.1), and (3.2), which is locally and globally well posed, but an additional interface band linear system has to be solved [23] at each domain decomposition iteration. Then the problem is formulated as a substructuring method and is solved by a GMRES algorithm.

7.2. Two-subdomain decomposition. For the decomposition into two subdomains, we use two regular Cartesian meshes on $\Omega = \Omega_1 \cup \Omega_2 =]0,1[\times]0,1[$. Figure 4 shows the solution and the error for the different methods. We see that the error is located on the interface and on the Dirichlet boundary. Figure 3 gives the asymptotic behavior of the log of the error as a function of the log of the mesh size h for the different interface operators. We observe that the solution and the error do not depend on the interface operators except for the method New Cement. This is in agreement with Theorem 3.1. For Ceres and Linear we obtain an order of 1.8 (higher than what is expected, maybe because we use regular Cartesian meshes in each subdomain). For New Cement, as predicted by theory, the order depends on the choice of α^i_ϵ ($\alpha^i_\epsilon = 1$, order 1.8; $\alpha^i_{opt,\epsilon}$, order 0.8; and $\alpha^i_\epsilon = 1/h_i$, order 0.5). For TPFA and Constant we obtain an order of 0.5 as expected.

7.3. Four-subdomain decomposition with a corner. We cut the domain $\Omega = \cup^4_{i=1}\Omega_i =]0,1[\times]0,1[$ into four subdomains with a corner. The method Ceres is not tested in this decomposition because the scheme does not extend easily to this situation. In Figure 7.3, we observe the H^1-norm of the error for the different methods and for the different interface operators. As in the decomposition into two subdomains only the error and the solution for New Cement depend on the interface

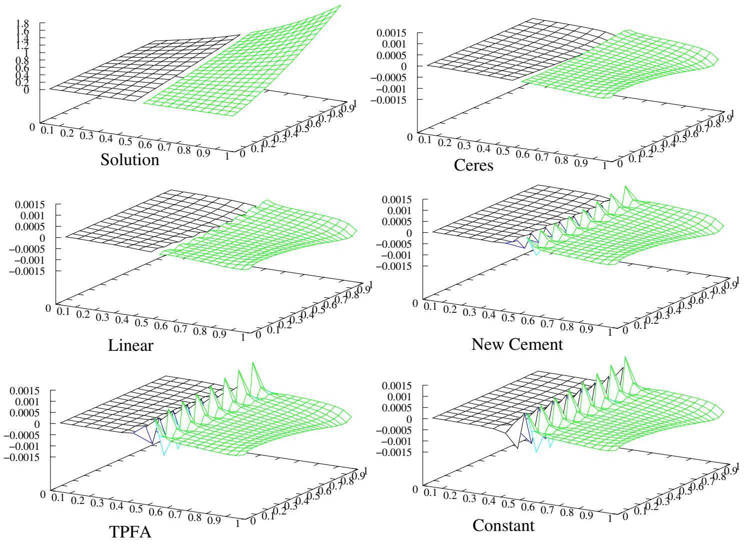


FIG. 4. *Solution $x^3y^2 + sin(xy)$ and the error for different methods for two subdomains.*

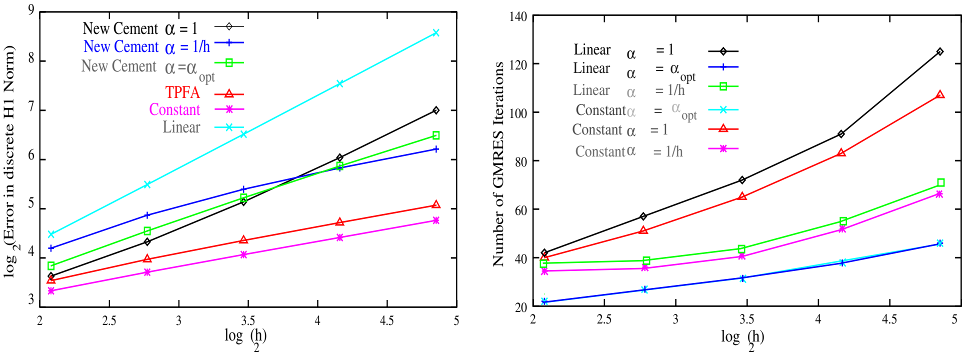


FIG. 5. *Left: Log of the error H^1-norm vs. log of the mesh size (h) for four subdomains. Right: Number of GMRES iterations vs. mesh size for different Robin coefficients for four subdomains.*

operators. For Linear we have an error order of 1.5. This is better than the error in $O(h)$ proved in Theorem 6.7. This is most probably due to the regular meshes in each subdomain which yield a superconvergence effect. For New Cement the order is 1.3 with $\alpha_\epsilon^i = 1$, 0.9 with $\alpha_{opt,\epsilon}^i$, and 0.6 with $\alpha_\epsilon^i = 1/h_i$. For TPFA and Constant we see an order of 0.5. As an illustration of the second error estimate in Theorem 6.4, we show Figure 6, which illustrates the need to have the master grid as a subgrid of the slave grid in order to have an error of order $O(h)$ when using only piecewise constant projections. Figure 7.3 represents the number of GMRES iterations for the different interface operators as a function of the log of the mesh size, and we note that the number of GMRES iterations is always the lowest for $\alpha_i = \alpha_{opt,\epsilon}^i$ as expected.

8. Conclusion. We have presented a finite volume method for a domain decomposition with nonmatching grids that allows for the use of arbitrary interface conditions. This last feature is important for a fast convergence of the iterative domain decomposition algorithm. Contrary to [5] or [3] the discrete solution does not

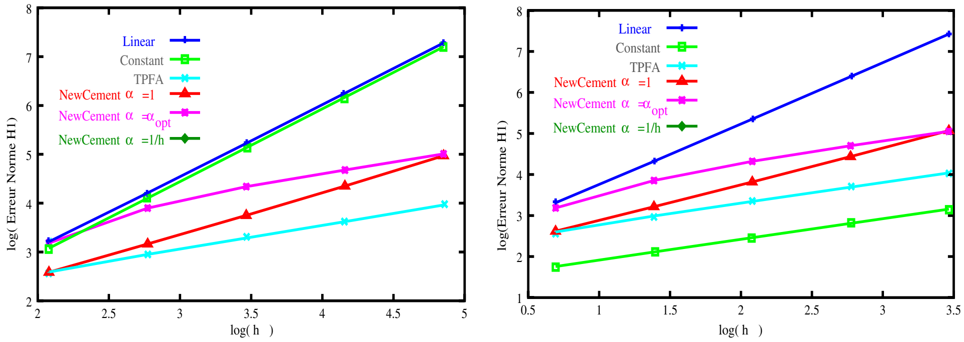


FIG. 6. *Piecewise constant projection. Left: the master grid is a subgrid of the slave. Right: the slave grid is a subgrid of the master.*

depend on the interface conditions. In practical computations in porous media flow, the diffusion operator is not a scalar nor a smooth function. It is typical to have jumps in the coefficients of diffusion of four orders of magnitude that do not match across the interface. In order to still have good results, the proposed method has to be enhanced by the introduction of a thin subdomain at each interface; see [14]. Numerical results are very satisfactory but the corresponding analysis still demands further investigation.

REFERENCES

[1] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford University Press, Oxford, 1999.

[2] I. Aavatsmark, E. Reiso, and R. Teigland, *Control-volume discretization method for quadrilateral grids with faults and local refinements*, Comput. Geosci., 5 (2001), pp. 1–23.

[3] Y. Achdou, C. Japhet, F. Nataf, and Y. Maday, *A new cement to glue non-conforming grids with Robin interface conditions: The finite volume case*, Numer. Math., 92 (2002), pp. 593–620.

[4] T. Arbogast, L.C. Lawrence, M.F. Wheeler, and I. Yotov, *Mixed finite element methods on nonmatching multiblock grids*, SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.

[5] T. Arbogast and I. Yotov, *A non-mortar mixed finite element method for elliptic problems on non-matching multiblock grids*, Comput. Methods Appl. Mech. Engrg., 149 (1997), pp. 255–265.

[6] R. Belmouhoud, *Modélisation tridimensionnelle de la genèse des bassins sédimentaires*, Ph.D. thesis, Ecole nationale supérieure des Mines de Paris, 1996.

[7] C. Bernardi, Y. Maday, and A. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, H. Brezis and J.L. Lions, eds., Pitman Res. Notes Math. Ser., Longman, Harlow, UK, 1989, pp. 13–51.

[8] C. Bernier, E. Candus, I. Faille, and F. Nataf, *Maillages non coincidents pour la modélisation des écoulements en milieux poreux*, Technical report 54 332, IFP, 2000.

[9] R. Cautrés, R. Herbin, and F. Hubert, *Non matching finite volume grids and the non overlapping Schwarz algorithm*, in Proceedings of the 13th International Conference on Domain Decomposition Methods, N. Debit, M. Garbey, R. Hoppe, J. Periaux, D. Keyes, and Y. Kutnnetsov, eds., 2000, pp. 213–219.

[10] R. Cautrés, R. Herbin, and F. Hubert, *The Lions domain decomposition algorithm on non matching cell-centered finite volume meshes*, IMA J. Numer. Anal., 24 (2004), pp. 465–490.

[11] R. Cautrés, R. Herbin, and F. Hubert, *Finite volume scheme on non matching grids. Application to domain decomposition methods*, in Finite Volume for Complex Applications III: Problems and Perspectives, Laboratoire d'Analyse, Topologie et Probabilités CNRS, Marseille, 2002, pp. 141–148.

[12] P.G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 17–351.

[13] R.E. Ewing, R.D. Lazarov, T. Lin, and Y. Lin, *Domain decomposition capabilities for the mortar finite volume element methods*, in 11th International Conference on Domain Decomposition Methods in Science and Engineering, C.H. Lai, P.E. Bjorstad, M. Cross, and O. Widlund, eds., DDM. Org., Augsburg 1999, pp. 220–227.

[14] I. Faille, F. Nataf, L. Saas, and F. Willien, *Finite volume methods on non-matching grids with arbitrary interface conditions and highly heterogeneous media*, in Domain Decomposition Methods in Science and Engineering, Lecture Notes in Comput. Sci. Eng. 40, Springer, Berlin, 2004, pp. 243–250.

[15] T. Gallouët, R. Herbin, and M.H. Vignal, *Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions*, J. Numer. Anal., 37 (2000), pp. 1935–1972.

[16] M. J. Gander, F. Magoulès, and F. Nataf, *Optimized Schwarz methods without overlap for the Helmholtz equation*, SIAM J. Sci. Comput., 24 (2002), pp. 38–60.

[17] C. Japhet and F. Nataf, *The best interface conditions for domain decomposition methods: Absorbing boundary conditions*, in Absorbing Boundaries and Layers, Domain Decomposition Methods. Applications to Large Scale Computations, L. Tourrette and L. Halpern eds., Nova Science Publishers, Huntington, NY, 2001, pp. 348–373.

[18] P.L. Lions, *On the Schwarz alternating method* III*: A variant for nonoverlapping subdomains*, in 3rd International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1989, pp. 202–203.

[19] F. Nataf, *Interface connections in domain decomposition methods*, in Modern Methods in Scientific Computing and Applications, NATO Science Ser. II Math. Phys. Chem. 75, Kluwer, Dordrecht, 2001, pp. 323–364.

[20] F. Nataf, F. Rogier, and E. de Sturler, *Domain decomposition methods for fluid dynamics*, in Navier–Stokes Equations and Related Nonlinear Problems, Plenum Press, New York, 1995, pp. 367–376.

[21] R. Herbin, R. Eymard, and T. Gallouët, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 2000, pp. 723–1020.

[22] Y. Saad, *Iterative Methods for Sparse Linear Systems*, 2nd ed., SIAM, Philadelphia, PA, 2003.

[23] L. Saas, I. Faille, F. Nataf, and F. Willien, *Décomposition de domaine pour volumes finis sur maillages non conformes avec conditions d'interface de robin optimisées,* Technical report 56947, IFP, 2002.

[24] F. Willien, *Extension au* 3*d des méthodes de décomposition de domaine appliquées à la modélisation des bassins sédimentaires*, Technical report 45 547, Rapport IFP, 1999.

SIAM J. Numer. Anal.
Vol. 43, No. 2, pp. 891–907

AN EFFICIENT AND STABLE METHOD FOR COMPUTING MULTIPLE SADDLE POINTS WITH SYMMETRIES[*]

ZHI-QIANG WANG[†] AND JIANXIN ZHOU[‡]

Abstract. In this paper, an efficient and stable numerical algorithm for computing multiple saddle points with symmetries is developed by modifying the local minimax method established in [Y. Li and J. Zhou, *SIAM J. Sci. Comput.* 23 (2001), pp. 840–865; Y. Li and J. Zhou, *SIAM J. Sci. Comput.*, 24 (2002), pp. 840–865]. First an invariant space is defined in a more general sense and a principle of invariant criticality is proved for the generalization. Then the orthogonal projection to the invariant space is used to preserve the invariance and to reduce computational error across iterations. Simple averaging formulas are used for the orthogonal projections. Numerical computations of examples with various symmetries, of which some can and others cannot be characterized by a compact group of linear isomorphisms, are carried out to confirm the theory and to illustrate applications. The mathematical features of various problems demonstrated in these examples fall into two categories: nodal solutions of saddle-point type with large Morse indices and nonradial positive solutions via symmetry breaking in radially symmetric elliptic problems. The new numerical algorithm generates these rather unstable solutions in an efficient and stable way. The existence of many unstable solutions and their behavior found in this paper remain to be investigated.

Key words. multiple saddle points, Morse index, symmetry, invariance, local minimax method, semilinear PDE

AMS subject classifications. 58E05, 58E30, 35A40, 35A65

DOI. 10.1137/S0036142903416626

1. Introduction.

Let H be a Hilbert space and let $J : H \to \mathbb{R}$ be Fréchet differentiable; let J' be its Fréchet derivative and let ∇J be the gradient and J'' its second Fréchet derivative if it exists. A point $u^* \in H$ is a *critical point* of J if u^* solves the Euler–Lagrange equation $J'(u^*) = 0$. A critical point u^* is *nondegenerate* if $J''(u^*)$ is invertible; otherwise u^* is *degenerate*. According to the Morse theory, the *Morse index* (MI) of a critical point u^* of J is the maximal dimension of a subspace of H on which the operator $J''(u^*)$ is negative definite. The first candidates for a critical point are the local extrema to which the classical critical point theory was devoted in calculus of variations [22]. Most conventional numerical algorithms focus on finding such stable solutions. Critical points that are not local extrema are *unstable* and called *saddle points*. Because of unstable nature, saddle points are too elusive to be numerically captured.

It is interesting for both theory and applications to develop *efficient* and *stable* numerical algorithms for finding multiple saddle points. Minimax principle is one of the most popular approaches in critical point theory. However, most minimax theorems in the literature (see [1], [2], [3], [6], [16], [18], [19], [20], [21], [24]), such as the mountain pass, various linking and saddle point theorems, require one to solve a two-level *global* optimization problem and, therefore, are not useful for algorithm implementation.

[*] Received by the editors March 14, 2003; accepted for publication (in revised form) April 15, 2005; published electronically September 15, 2005.
http://www.siam.org/journals/sinum/43-2/41662.html

[†] Department of Mathematics and Statistics, Utah State University, Logan, UT 84322 (wang@math.usu.edu).

[‡] Department of Mathematics, Texas A&M University, College Station, TX 77843 (jzhou@math.tamu.edu).

Efforts for numerically computing saddle points have been made in [7] for MI = 1 and in [10] for MI = 2 which were motivated by theoretical (global minimax) characterizations of saddle points in [1] and [23], respectively. Inspired by [7], [10], and an idea in [9], [17], a local minimax method (LMM) was developed in [12], [13] and many multiple solutions were numerically computed for a class of semilinear elliptic equations. Its convergence results are obtained in [13]. Several results in instability analysis of saddle points are established in [14], [25].

Let us briefly recall LMM. Its basic idea is to define a local peak selection and a solution set. Let $L \subset H$ be a closed subspace, called a support to the critical point u^* to be found, $S_{L^\perp} = \{v \in L^\perp : \|v\| = 1\}$ and denote $[L, v] = \{tv + v_L : t \in \mathbb{R}, v_L \in L\}\,\forall v \in S_{L^\perp}$. A set-valued mapping $P\colon S_{L^\perp} \to 2^H$ is called a *peak mapping* of J if $P(v)$ is the set of all local maximum points of J in $[L, v]$. A single-valued mapping $p\colon S_{L^\perp} \to H$ is called a *peak selection* if $p(v) \in P(v)\,\forall v \in S_{L^\perp}$. Let $v \in S_{L^\perp}$ be a point. p is said to be a *local peak selection* of J with respect to (w.r.t.) L at v if a peak selection p is locally defined near v.

A local minimax theorem which characterizes a saddle point as a local minimax solution has been established in [12]. It states the following: let a solution set be defined by

$$\mathcal{M} = \{p(v) : v \in S_{L^\perp}\}; \tag{1.1}$$

then a point $u^* = p(v^*) \in \mathcal{M}$ is a saddle point of J if p is continuous at v^*, $p(v^*) \notin L$, and

$$v^* = \arg \underset{v \in S_{L^\perp}}{\text{loc-min}}\, J(p(v)) = \arg \underset{v \in S_{L^\perp}}{\text{loc-min}}\ \underset{u \in [L,v]}{\text{loc-max}}\, J(u).$$

It becomes a local minimization problem on the solution set $\mathcal{M}$, which can be numerically approximated by, e.g., the steepest descent method.

A numerical LMM.

Step 1. Given positive ε and λ, let $n-1$ critical points $w_1, w_2, \ldots, w_{n-1}$ of J be given, of which w_{n-1} has the highest critical value. Set the support $L = \operatorname{span}\{w_1, w_2, \ldots, w_{n-1}\}$. Let $v_n^1 \in S_{L^\perp}$ be an ascent direction at w_{n-1}. Let $t_0^0 = 1$ and $v_L^0 = w_{n-1}$ and set $k = 0$.

Step 2. Use the initial guess $w = t_0^k v^k + v_L^k$; solve for

$$w^k \equiv p(v^k) \equiv t_0^k v^k + v_L^k = \arg \max_{w \in [L, v^k]} J(w) \text{ and denote } t_0^k v^k + v_L^k = w^k \equiv p(v^k).$$

Step 3. Compute the steepest descent vector $d^k = -\nabla J(w^k)$.

Step 4. If $\|d^k\| \le \varepsilon$, then output $w_n = w^k$, stop; else go to Step 5.

Step 5. Set $v^k(s) = \frac{v^k + sd^k}{\|v^k + sd^k\|}$ and find

$$s^k = \max\left\{\frac{\lambda}{2^m} \,\middle|\, m \in \mathbb{N}, J\Big(p\Big(v^k\Big(\frac{\lambda}{2^m}\Big)\Big)\Big) - J(w^k) \le -\frac{t_0^*}{2}\|d^k\| \left\|v^k\Big(\frac{\lambda}{2^m}\Big) - v^k\right\|\right\}.$$

Initial guess $u = t_0^k v^k(\frac{\lambda}{2^m}) + v_L^k$, where t_0^k and v_L^k are found in Step 2, is used to find the local maximum point $p(v^k(\frac{\lambda}{2^m}))$ in $[L, v^k(\frac{\lambda}{2^m})]$, similar to Step 2.

Step 6. Set $v^{k+1} = v^k(s^k)$ and update $k = k + 1$, then go to Step 2.

The subspace L containing critical points previously found serves as a support to a saddle point u^* to be found at a higher critical level. A support to u^* is said to be *sufficient* if it contains all critical points below u^*'s critical level. When MI of u^* gets larger, the dimension of L grows larger [25]. Solving the local maximization problem in the space $[L, v^k]$ in Step 2 of LMM becomes more expensive. Since $u^* = p(v^*)$ and $v^* = \arg\max_{v \in S_{L^\perp}} J(p(v))$, the solution set $\mathcal{M}$ is a stable set, i.e., when J is restricted to $\mathcal{M}$, u^* is a stable solution.

Symmetries exist in many natural phenomena, such as in crystals, elementary particle physics, symmetry of the Schrödinger equation for the atomic nucleus and the electron shell w.r.t. permutations and rotations, energy conservation law for systems which are invariant w.r.t. time translation, etc. Symmetry properties are usually studied by group actions in mathematics. Symmetries described by compact group actions in variational problems have been used in the literature to prove the existence of *multiple critical points*, typically in the Ljusternik–Schnirelman theory (see recent results in [5], [15], [11], and others). It is known that symmetries in a nonlinear variational problem can lead to the existence of more solutions of saddle type and can also cause degeneracy. In this paper, we study the impact of the presence of symmetries on LMM in finding multiple saddle points. By modifying LMM, we shall develop an efficient and stable numerical algorithm for computing multiple critical points with general symmetries. Consider a semilinear elliptic BVP and its energy function J,

$$\begin{cases} \Delta u(x) + F(u(x), x) = 0, & x \in \Omega \subset R^n, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \qquad J(u) = \int_\Omega \left\{ \frac{1}{2}|\nabla u(x)|^2 - f(u(x), x) \right\} dx,$$

where $f(t, x) = \int_0^t F(\tau, x) d\tau$ satisfies some standard conditions. When $L = \{0\}$, the solution set becomes $\mathcal{M} = \{t_u u : u \in H, \|u\| = 1, t_u \neq 0, \langle \nabla J(t_u u), u\rangle = 0\}$, which is called the Nehari manifold. Under some standard conditions, it can be shown that $\mathcal{M}$ is smooth, or the peak selection p is continuously differentiable [12].

Several researchers have tried to use certain symmetry of a problem to capture a solution of higher MI. For example, odd symmetry is used in [7] to capture sign-changing solutions (MI = 2) by a minimization on the Nehari manifold with an odd symmetric initial guess. When a negative gradient-type minimax algorithm is used, the symmetry is inherited but not enforced, and the sequence generated by the algorithm will get close to a saddle point. However, when computational error builds up or $\|\nabla J(u)\|$ becomes small, computational error will dominate and break the symmetry of $\nabla J(u)$. Eventually the symmetry of the sequence collapses. Therefore, the sequence will not stop near a sign-changing solution; instead it will slide down to a positive solution (MI = 1), unless a forcing stop action is taken. Thus such an approximation is unstable and no convergence can be established. Even rotational symmetry is considered in [8] to capture sign-changing solutions (MI = 3) by a high-linking algorithm. Action to preserve the symmetry is taken, so the algorithm is stable.

In this paper, we consider more general symmetries and try to establish some mathematical justifications. In particular, we are concerned with not only preserving the symmetry but also reducing computational error across iterations. There are at least three motivations for one to use symmetries to define an invariant space H_I in computing a saddle point:

(1) a sufficient support is available in H, but one wants to reduce its dimension by using a sufficient support L in H_I to enhance the efficiency of the algorithm;
(2) no sufficient support is available in H; one has to use the symmetries to find a sufficient support L in H_I;
(3) to use symmetries to bypass the degeneracy of a problem.

To find a critical point u^* at a higher critical level by LMM, one needs to know if the support L is sufficient or not. If the answer is yes, one can expect a stable convergence even without using any symmetry. If the answer is no, the minimization process will sooner or later find a slider and bypass u^*. The algorithm becomes unstable and fails to reach u^*.

Assume one has identified symmetries of a problem and defined an invariant subspace H_I. Then one can restrict the problem in H_I, i.e., the support L contains critical points at lower critical level only in H_I. By doing so, the dimension of the support L can be greatly reduced. Since LMM with an insufficient support in H_I is unstable, we assume the support L in H_I is sufficient. When computational error builds up or $\|\nabla J(w^k)\|$ becomes small, computational error will dominate and break the symmetry of $\nabla J(w^k)$. That is, $\nabla J(w^k)$, eventually w^{k+1} goes outside of H_I. There are two possibilities. If the support L in H_I is also sufficient in H, there is no slider around u^*. Thus the symmetry has no effect on the algorithm and, therefore, the collapse of the symmetry has no effect either; the algorithm will still converge to u^*. If the support L in H_I is not sufficient in H, and w^{k+1} is outside H_I, the minimization process will sooner or later find a slider and then fail to reach u^*. A projection of $\nabla J(w^k)$ onto H_I will pull $\nabla J(w^k)$ and then w^{k+1} back to H_I and resolve the problem.

These types of projections into an invariant space have been used in the literature to preserve the symmetry, where computational error is not a concern. In this case any projection operator onto H_I will serve the purpose. However, computational error is a main concern in numerical computation, particularly in multilevel iterations for finding unstable saddle points at a higher critical level, which is rather sensitive to computational error. It is, therefore, a main concern of this paper. There are infinitely many projection operators onto H_I. Some of them are poor at handling computational error. Only the orthogonal projection operator onto H_I is the optimal one to handle computational error. Thus in this paper, we look for the orthogonal projection operator onto H_I. It is known that finding an orthogonal projection onto a subspace is equivalent to an infinite-dimensional minimization problem, which is very expensive. It is in particular very difficult, since there is no explicit expression for H_I. The average formula defined by the Haar integral has been used in the literature to project a point onto an invariant subspace. Here we expose a fact that the Haar integral operator is actually the orthogonal projection operator onto H_I. Implementation of this formula with LMM for numerical computations of multiple critical points with symmetries at higher critical levels will be discussed in detail by using typical numerical examples.

The numerical examples we choose also serve to reveal new phenomena in the corresponding mathematical problems. Here we are mainly interested in two directions: nodal solutions of saddle-point type for nonlinear elliptic problems and nonradial positive solutions in radially symmetric elliptic equations. In both cases, people expect that many unstable solutions exist, that these solutions should have large MIs, and that degeneracy occurs in general. Using LMM with symmetry we shall not only demonstrate solutions that are known to exist in theory but also exhibit many cases for which the existence is still open in theory. Some of these examples give surprising, new mathematical features and should shed light on the study of the nonlinear elliptic PDEs.

2. Invariant space and its orthogonal projection.

2.1. Invariant spaces and LMM in invariant spaces. In order to let LMM handle symmetry we need the concept of invariant spaces. The following is a more general one without reference to symmetry.

DEFINITION 2.1. *Let H be a Hilbert space and $J \in C^1(H,R)$. A closed subspace H_I of H is said to be a J-invariant space if for every $u \in H_I$ it holds that $\nabla J(u) \in H_I$.*

Along the line of the classical principle of symmetric criticality by Palais, we have the following principle of invariant criticality (PIC) without reference to symmetry.

THEOREM 2.1. *Let H be a Hilbert space, $J \in C^1(H,R)$ and H_I a J-invariant space. If $u^* \in H_I$ is a critical point of J restricted to H_I, then u^* is a critical point of J in H.*

Proof. That u^* is a critical point of J restricted to H_I implies $\langle \nabla J(u^*), v\rangle = 0\ \forall v \in H_I$, i.e., $\nabla J(u^*) \perp H_I$. On the other hand, H_I is a J-invariant space, i.e., $u^* \in H_I$ implies $\nabla J(u^*) \in H_I$. Therefore, $\nabla J(u^*) \in H_I \cap H_I^{\perp} = \{0\}$. □

Thus we can restrict solving the problem only in H_I. It is clear that H is trivially a J-invariant space. Since the smaller the dimension of H_I is, the smaller the MI of u^* relative to H_I is—which implies that the smaller the dimension of the support L [25] in H_I is, the more efficient and stable is the numerical computation—we always look for the smallest such J-invariant space. In this way, LMM can be used, in a much more efficient and stable way, to find multiple saddle points with certain symmetries at higher critical level.

Let L be a closed subspace of H_I and denote $S_{L^\perp} = \{v \in H_I : \|v\| = 1, v \perp L\}$. Let $k = 0$; started from a point $v^k \in S_{L^\perp}$ we have $p(v^k) = t^k v^k + v_L^k \in H_I$ for some $t^k \neq 0$ and $v_L^k \in L \subset H_I$. Then $d^k = -\nabla J(p(v^k)) \in H_I$ and $v^{k+1} = \frac{v^k + s^k d^k}{\|v^k + s^k d^k\|} \in H_I$, where $s^k > 0$ is the stepsize. If $v^0 \in H_I$ and $d^k = -\nabla J(p(v^k)) \neq 0$, we have [12]

$$J(p(v^{k+1})) - J(p(v^k)) < -\frac{t^k}{2}\|d^k\|\|v^{k+1} - v^k\|.$$

This brings us to the conclusion that, in theory, LMM is closed in an invariant space H_I and generates a strict minimizing sequence $\{p(v^k)\} = \{t^k v^k + v_L^k\}$ in $\mathcal{M} \cap H_I$, where $\mathcal{M}$ is the solution set defined in (1.1). The limit of the sequence is a critical point of J in H_I by the convergence results in [13] and thus a critical point of J in H by PIC. However, in numerical computation of the negative gradient $d^k = -\nabla J(p(v^k))$, it involves discretization, approximation, round-off error, etc. It generates numerical error and then breaks the invariance of d^k. To preserve the invariance, we use the decomposition $H = H_I \oplus H_I^c$ for some complement space H_I^c of H_I in H and

$$d^k = d_I^k + (d^k)_I^c, \quad d_I^k \in H_I, \quad (d^k)_I^c \in H_I^c.$$

Note that if $p(v^k) \in H_I$ and $d^k = -\nabla J(p(v^k))$ is computed exactly, we should have

$$d^k = (d^k)_I \in H_I \quad \text{and} \quad (d^k)_I^c = 0.$$

When numerical error is involved, we use $(d^k)_I$ to replace d^k in Step 3 of LMM, the updated point v^{k+1} is now in H_I, and the invariance is preserved.

The above decomposition for H needs to find a projection operator T from H to H_I. There are infinitely many such projection operators. If preserving the invariance is the only concern, by PIC, any one of them will serve the purpose. However, when error analysis is concerned, the case is different. Note that the term $(d^k)_I^c = d^k - (d^k)_I$ represents computational error. In numerical computation, we have to not only preserve the invariance, but also reduce error which will be carried to the next iteration. Among all those projection operators, some of them are very poor at dealing with error and there is the optimal one that minimizes the distance from d^k to H_I, the error term, that is, the orthogonal projection. In this case, the maximum invariant part of d^k has been carried to $(d^k)_I$ and $(d^k)_I^c$ becomes $(d^k)_I^\perp$, i.e.,

$$d^k = (d^k)_I + (d^k)_I^\perp \quad \text{and} \quad (d^k)_I \perp (d^k)_I^\perp .$$

For this reason, in this paper we will do our best to adopt the orthogonal projection.

EXAMPLE 2.1. *Let $X = R^2$, $X_2 = \{(0, x_2)^T\}$. Both $A = \left[\begin{smallmatrix} 0 & 0 \\ 100 & 1 \end{smallmatrix}\right]$ and $B = \left[\begin{smallmatrix} 0 & 0 \\ 0 & 1 \end{smallmatrix}\right]$ are projection operators of X onto X_2. Let $(\epsilon_1, \epsilon_2)^T$ represents computational error in computing $(0,1)^T \in X_2$; we get $u = (\epsilon_1, \epsilon_2 + 1)^T \notin X_2$. To do projection, we have*

$$Au = (0, 100\epsilon_1 + 1 + \epsilon_2)^T \in X_2 \quad \textit{and} \quad Bu = (0, 1 + \epsilon_2)^T \in X_2.$$

It is clear that A greatly enlarges error while B does not. As a matter of fact, where error analysis is concerned, B is the optimal operator that minimizes error, i.e., B is the orthogonal projection operator from $X \to X_2$ and Bu is the best approximation one can get.

2.2. Invariant spaces from symmetries. Invariant spaces appear naturally when the problems considered possess certain symmetry, e.g., when the function J is invariant w.r.t. certain symmetry. There are two usual ways in numerical computations to preserve symmetry. By dividing the domain into several subdomains along the axes of symmetry, one may solve the problem only on one subdomain with an additional continuity condition with the Neumann data across the cuts; the problem may become much harder to solve, but the size of the problem becomes smaller; one may also solve the problem on the entire domain, but use the solution data only on one subdomain. Then in either case, one can produce a solution on the entire domain according to the symmetry. Note that either way, it preserves the symmetry, but carries computational error with the solution to the next iteration. Indeed it forces computational error to be of the same symmetry. However, computational error is usually asymmetric and even random.

In contrast to the usual method, our new method separates a solution from computational error (at least the asymmetric part), and then carries only the solution, not the error, in iterations. Thus the advantage is clear that it makes the algorithm more efficient and more stable. Our numerical examples confirm the analysis.

Let us start with the most studied symmetries in the literature, i.e., those symmetries that can be characterized by a compact group of linear isomorphisms.

Let H be a Hilbert space and G be a compact group of linear isomorphisms of H, i.e., the map from $G \times H \to H$ evaluated by $(g, u) \to gu$ is continuous such that for each $h \in G$,

$$1 \cdot u = u, \quad (gh)u = g(hu), \quad u \to gu \quad \text{is linear}, \quad \|gu\| = \|u\|.$$

The identification of G as a subgroup of linear isomorphisms of H is called a representation of G, and we shall still use G to stand for a representation of it.

For a representation of G we may define its invariant subspace. A set $A \subset H$ is G-invariant if $g(A) = A \, \forall g \in G$. The *$G$-invariant subspace* of H is the subspace $H_G = \{u \in H : gu = u \, \forall g \in G\}$. Let $J \in C^1(H, R)$. J is said to be G-invariant if $J(gu) = J(u)$ for every $(g, u) \in G \times H$. A map $F : H \to H$ is G-equivariant if $g \circ F = F \circ g$ for every $g \in G$. Since $J \in C^1(H, R)$ is G-invariant implies that ∇J is G-equivariant, i.e., $\nabla J(gu) = g \nabla J(u) \; \forall u \in H$, when $u \in H_G$, we have $g \nabla J(u) = \nabla J(gu) = \nabla J(u) \; \forall g \in G$ or $\nabla J(u) = \nabla J(u)_G$.

The above definition for an invariant space separates the space H from the function J, which may have other applications. On the other hand, Definition 2.1 combines the space H with the function J in the definition of a J-invariant space, which serves precisely the purpose of applications in this research. It is clear that if J is G-invariant, then H_G is a J-invariant space as in Definition 2.1. Thus the following classical result follows from Theorem 2.1.

THEOREM 2.2 (principle of symmetric criticality (Palais, 1979)). *Let H be a Hilbert space and G a compact group of linear isomorphisms of H. If $J \in C^1(H, R)$ is G-invariant and if $u \in H_G$ is a critical point of J restricted to H_G, then u is a critical point of J.*

2.3. Orthogonal projections. Note that the invariant space H_G is usually infinite-dimensional and in particular does not have an explicit formula; finding the orthogonal projection of d^k onto H_G is very expensive and difficult. However, for many usual symmetries, the orthogonal projection onto H_G turns out to be simple algebraic computations. Let us first cite the following theorem.

THEOREM 2.3 (Haar, 1933). *Let G be a compact group and $C(G)$ the vector space of real-valued continuous functions on G. Then there exists a unique positive integral (the Haar integral) such that the map $C(G) \to R$ by $f \mapsto \int_G f(g)\,dg$ is*

(a) *linear, monotone, and normalized ($\int_G 1\,dg = 1$);*

(b) *left-invariant, i.e., $\int_G f(h^{-1}g)\,dg = \int_G f(g)\,dg\,\forall h \in G$.*

The Haar integral operator $\mathcal{H}$ from H to H_G defined by $\mathcal{H}u = \int_G gu\,dg\,\forall u \in H$ has been used in the literature as a projection from H onto H_G to preserve an invariance, where computational error is not a concern. When reducing computational error across iterations is a concern, we are interested mainly in the orthogonal projection onto H_G. Since for $u, v \in H$,

$$\begin{aligned}\langle \mathcal{H}u, \mathcal{H}v\rangle &= \int_G \langle gu, \mathcal{H}v\rangle\,dg = \int_G \langle u, g^*\mathcal{H}v\rangle\,dg\\ &= \int_G \langle u, \mathcal{H}v\rangle\,dg = \langle u, \mathcal{H}v\rangle,\end{aligned}$$

we have $\langle u-\mathcal{H}u, v\rangle = \langle u-\mathcal{H}u, \mathcal{H}v\rangle = 0\,\forall v \in H_G$, i.e., $\mathcal{H}$ turns out to be the orthogonal projection operator from H onto H_G and $u = \mathcal{H}u + (u - \mathcal{H}u)$ is the orthogonal direct sum.

2.4. Examples. We give some examples that will be used in our numerical computations.

EXAMPLE 2.2. *Let H be a Hilbert space with inner product $\langle , \rangle$ and G a finite group of linear isomorphisms of H with m elements. Then for each $u \in H$, the Haar integral operator on u is given by $\mathcal{H}u = u_G = \frac{1}{m}\sum_{g \in G} gu \in H_G$ and $u_G^\perp = u - u_G \in H_G^\perp$.*

EXAMPLE 2.3. *Let $\Omega \subset R^n$ $(n \geq 1)$ be a bounded open set and $H = W^{k,2}(\Omega)$, where $k \geq 0$ be the Sobolev space. Let $\mathcal{O}(n)$ be the set of all orthogonal matrices in $R^{n\times n}$. $\mathcal{O}(n)$ is a compact group. Let G be the set of all orthogonal matrices $g \in \mathcal{O}(n)$ such that $g(\Omega) = \Omega$. For each $g \in G$ and $u \in H$, let $gu(x) = u(gx)\,\forall x \in \Omega$; then G is a compact group of linear isomorphisms of H. The Haar integral operator $\mathcal{H}$ defines the orthogonal projection operator from H onto H_G. Indeed, if g is represented by an orthogonal matrix, then g is an isomorphism of H, or in other words, the inner* *product is g-invariant, i.e.,*

$$\begin{aligned}\langle gu, gv\rangle &= \int_\Omega \sum_{|\alpha|\le k} (D^\alpha u(gx))^T (D^\alpha v(gx))\,dx \\ &\quad (\textit{by substituting } y = gx \textit{ and } g(\Omega) = \Omega, g^T g = I) \\ &= \int_\Omega \sum_{|\alpha|\le k} (D^\alpha u(y))^T (g^T)^\alpha (g)^\alpha (D^\alpha v(y))|g|\,dy = \langle u, v\rangle \quad \forall u, v \in H.\end{aligned}$$

Example 2.4. *Let $\Omega \subset R^n$ be a bounded open domain. Assume that Ω is symmetric about the reflections w.r.t. the first $n-1$ axes. Let $H = W^{k,2}(\Omega)$. Define $g : H \to H$ by $(gu)(x_1, \dots, x_{n-1}, x_n) = -u(-x_1, \dots, -x_{n-1}, x_n)$. Then $G = \{id, g\} \cong \mathbb{Z}_2$.*

Example 2.5. *Let $\Omega \subset R^2$ be a bounded open domain. Let $m > 1$ be an integer. For each point $x = (r, \theta) \in \Omega$ denote $gx \equiv g(r, \theta) = (r, \theta + \frac{2\pi}{m})$ and $\bar h x \equiv \bar h(r, \theta) = (r, -\theta)$, i.e., for $\alpha = \frac{2\pi}{m}, g = \left[\begin{smallmatrix}\cos(\alpha) & -\sin(\alpha)\\ \sin(\alpha) & \cos(\alpha)\end{smallmatrix}\right]$ and $\bar h = \left[\begin{smallmatrix}1 & 0\\ 0 & -1\end{smallmatrix}\right]$. Let $g(\Omega) = \Omega$, $\bar h(\Omega) = \Omega$ and $H = W^{1,2}(\Omega)$ be the Sobolev space of functions on Ω with the inner product $\langle u, v\rangle = \int_\Omega (\nabla u \cdot \nabla v + uv)\,dx$. Let $\mathbf{eo} = \pm 1$ be fixed. For each $u \in H$, we define*

$$g(u)(x) = u(gx) \quad \textit{and} \quad h(u)(x) = \mathbf{eo}(\bar h x).$$

It is clear that g represents a rotation and h represents an even ($\mathbf{eo} = 1$) or an odd ($\mathbf{eo} = -1$) reflection. The inner product is invariant under both linear operators g and h. To see this let f denote the operator g or h and the matrix g or $\bar h$. We have $f^T f = I$, and thus

$$\begin{aligned}\langle fu, fv\rangle &= \int_\Omega (\nabla u(fx)^T f^T f \nabla v(fx) + u(fx)v(fx))\,dx \\ &\quad (\textit{substituting } y = fx \textit{ and note } f(\Omega) = \Omega, |f| = 1) \\ &= \int_\Omega (\nabla u(y)\nabla v(y) + u(y)v(y))|f|\,dy = \langle u, v\rangle.\end{aligned}$$

We have a finite group $G = \{g, g^2, \dots, g^m, hg, hg^2, \dots, hg^m\}$ of linear isomorphisms of H which has two generators g and h. The invariant subspace H_G of H is defined by

$$H_G = \{u \in H : g^i u = u \;\; \textit{and} \;\; hg^i u = u\, \forall i = 1, 2, \dots, m\}.$$

For each $u \in H$, the Haar integral operator on u is given by

$$\mathcal{H}u = u_G \equiv \frac{1}{2m}\left(\sum_{i=1}^m g^i u + \sum_{i=1}^m hg^i u\right) \in H_G \quad \textit{and} \quad u_G^\perp \equiv u - u_G \in H_G^\perp.$$

There are symmetries that cannot be defined by a compact group of linear isomorphisms of H, such as composite symmetries involving partially defined symmetries. We may identify those symmetries by using several projections, among which the first one is orthogonal with which computational error is expected to be minimized.

Example 2.6. *Let $\Omega = [-a, a] \times [-a, a]$ in R^2 and $H = H_0^1(\Omega)$. We are interested in finding a critical point u^* with the following symmetries. The profile of u^* is even symmetric about the line $y = -x$, even symmetric about the x-axis for points (x, y) with $0 \ge y \ge -x \ge -a$, and even symmetric about the y-axis for points (x, y) with*

$0 \geq -x \geq y \geq -a$. To define the invariant space, we combine two projection operators $T = T_2 \cdot T_1$, where

$$\begin{cases} (T_1u)(x,y) = \frac{1}{2}(u(x,y) + u(-y,-x)), \\ \qquad (x,y) \in \Omega, \end{cases} \quad (T_2u)(x,y) = \begin{cases} u(x,-y), & 0 \geq y \geq -x, \\ u(-x,y), & 0 \geq -x \geq y, \\ u(x,y) & \textit{otherwise.} \end{cases}$$

T_1 is an orthogonal projection from H to H_{T_1} with which computational error is minimized and T_2 is a projection from H_{T_1} to H_T. Note that T_2 can map H to outside of H. However, T_2 projects H_{T_1} into H. Thus $T = T_2 \cdot T_1$ projects H into $H_T \subset H$.

Remark 2.1. When certain symmetries cannot be described by a compact group of linear isomorphisms, such as the case in Example 2.6, it is difficult to analytically verify whether or not a point u having the symmetries will imply that $\nabla J(u)$ has the same symmetries. But it can be numerically checked as follows. Let T represent the projection operator onto H_I. The term $\nabla J(u) - T(\nabla J(u))$ represents the asymmetric part of $\nabla J(u)$ together with the computational error. Let ε represent the order of the computational error. If $\|\nabla J(u) - T(\nabla J(u))\| \approx \varepsilon \|\nabla J(u)\|$, it means that the asymmetric part is caused by the computational error, not the asymmetric part of $\nabla J(u)$. Thus $\nabla J(u)$ has the same symmetries. Otherwise $\nabla J(u)$ does not have the same symmetries.

3. Numerical examples. In this section, we present several typical examples to illustrate the theory and numerical algorithm. Each of these examples has its own feature in symmetry and other properties. These examples also exhibit two types of mathematical phenomena. The first is about symmetry breaking in terms of some parameters of the problems. In general, for the problems with the full radial symmetry in $\mathbf{R}^n$, there is always a radial solution for all parameters. When we vary the parameters the problems may or may not have nonradial solutions. If nonradial solutions appear, we say symmetry breaking occurs. We demonstrate this feature by using autonomous equations with the Dirichlet boundary condition (the Lane–Emden equation) on thin annular domains as well as for the Henon equation in ball domains. For the existence of these solutions, some have been proved theoretically and others are still open (see [5], [4] for more references of theoretical studies). These nonradial solutions tend to have higher energy and possess large MIs. The second is about nodal solutions (sign-changing solutions) for nonlinear Dirichlet problems. Again these solutions tend to have higher energy and large MIs. The existence of most of these solutions are still open (see [11] for more references of theoretical studies on nodal solutions). As these examples show our new numerical algorithm is very powerful in dealing with multiple saddle points with large MIs and is very efficient and stable in dealing with the presence of symmetries.

To find a saddle point of J, when H is replaced by a J-invariant space H_I in LMM, whether or not the algorithm is stable depends on whether or not the support L in H_I is sufficient. We will pay special attention to the case where $L = \{0\}$, the smallest possible support. So we have tried to explore the symmetries of a problem to the maximum. It is known that when $L = \{0\}$, our solution set coincides with the Nehari manifold in H_G defined by

$$\mathcal{M} = \{t_u u : u \in H_G, \|u\| = 1, t_u \neq 0, \langle \nabla J(t_u u), u \rangle = 0\}.$$

Under some standard conditions, it can be shown [12] that $\mathcal{M}$ is smooth and the peak selection p is unique and C^1, which gives us a great advantage to show the convergence of LMM [13].

Let $\Omega \subset R^2$ be a bounded open domain. Consider the Henon equation ($p = 3, q \geq 0$)

$$\Delta u(x) + |x|^q u^p(x) = 0, \quad x \in \Omega, \quad u(x) = 0, \quad x \in \partial\Omega. \tag{2.0}$$

When $q = 0$, (2.0) reduces to the Lane–Emden equation. When $q > 0$, due to an explicit dependency on x, the well-known Gidas–Ni–Nirenberg theorem on symmetry does not apply. Thus it is interesting to examine the symmetry breaking phenomena.

Though the abstract theory asserts the invariance under general conditions of the group action, we demonstrate as an example for the Henon equation how to verify this for certain symmetries. We show here if u has a symmetry, then $\nabla J(u)$ possesses the same symmetry, where J is the corresponding functional to the Henon equation. Let $G \subset \mathcal{O}(n)$ be a closed subgroup, and Ω is invariant under the action of G. Let $\lambda_1 < \lambda_2 < \lambda_3 < \cdots$ be the eigenvalues of $-\Delta$ with Dirichlet boundary condition and E_i the eigenspace corresponding to λ_i. Then for $u \in E_i$, $-\Delta u = \lambda_i u$. It follows from this that for $g \in G$ we also have $-\Delta gu = \lambda_i gu$. That is, E_i is invariant subspace under G. This means H_G, the invariant space of group G, and $H_G^\perp$ are both generated by G-invariant eigenspaces. Now it is easy to verify that

$$(\nabla J(u), v) = \int_\Omega \nabla(u - (-\Delta)^{-1}|x|^q u^p)\nabla v \, dx,$$

which implies $\nabla J(u) = u - (-\Delta)^{-1}|x|^q u^p$. Assume $u \in H_G$ is in the G-invariant subspace. Then $|x|^q u^p \in H_G$ and we have $\nabla J(u) \in H_G$ if and only if $w = (-\Delta)^{-1}|x|^q u^p \in H_G$. Let $w = w_G + w_G^\perp$ with $w_G \in H_G$ and $w_G^\perp \in H_G^\perp$. Since w solves for $-\Delta w = |x|^q u^p \in H_G$ we have $-\Delta w_G^\perp = 0$. Thus $w_G^\perp = 0$ and $w \in H_G$. This gives $\nabla J(u) \in H_G$.

In the examples we consider four types of domains:

(a) $\Omega_s = [-1, 1] \times [-1, 1]$, (c) $\Omega_d = \{(x, y) \in R^2 : x^2 + y^2 \leq 1\}$,
(b) $\Omega_r = [-1.5, 1.5] \times [-1, 1]$, (d) $\Omega_a = \{(x, y) \in R^2 : 0.7 \leq \sqrt{x^2 + y^2} \leq 1\}$.

Domains Ω_s and Ω_r are used to exploit the structure of nodal solutions which are saddle points with large MIs. Domains Ω_d and Ω_a are discs and annulars which will be used to demonstrate symmetry breaking phenomena and to find nonradial positive solutions while radial positive solutions always exist. Disks and annular domains are the most symmetric domains in R^2, i.e., they are invariant under the full $\mathcal{O}(2)$ symmetry. But the symmetry causes degeneracy, due to the fact that a rotation of a nonradial solution about any angle is still a solution. Thus the solutions are not isolated in the whole space H. When symmetry is properly used to define an invariant space H_I, a solution in H_I can be isolated.

It is known that on a disk, the Lane–Emden equation has a unique positive solution as the ground state which is radially symmetric and has a single peak; while the Henon equation, in addition to the radially symmetric positive solution which has the highest critical level among all other positive solutions if exist, may have a nonradially symmetric positive solution. The existence of a nonradial solution to the Henon equation depends on the parameter q, which is a typical symmetry breaking phenomenon. There may have multipeak positive solutions. The number of peaks that a positive solution may have depends on the parameter q. This may be considered a bifurcation problem with parameter q. Because of high critical level, the

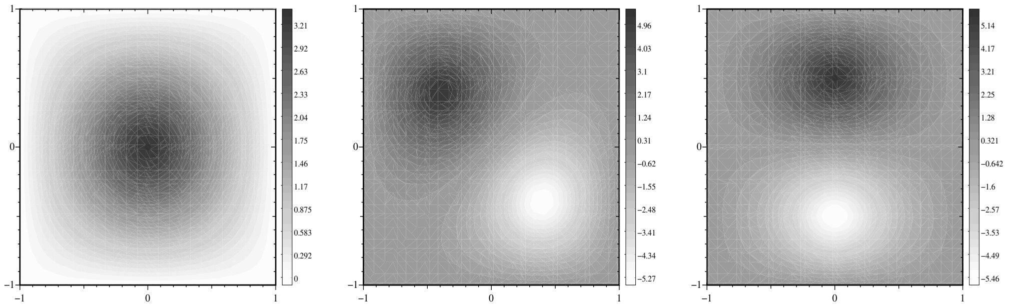


FIG. 1. *Solution contours to the Lane–Emden equation on* Ω_s, $J = 9.441$ (*left*), $J = 48.81$ (*center*), *and* $J = 53.58$ (*right*).

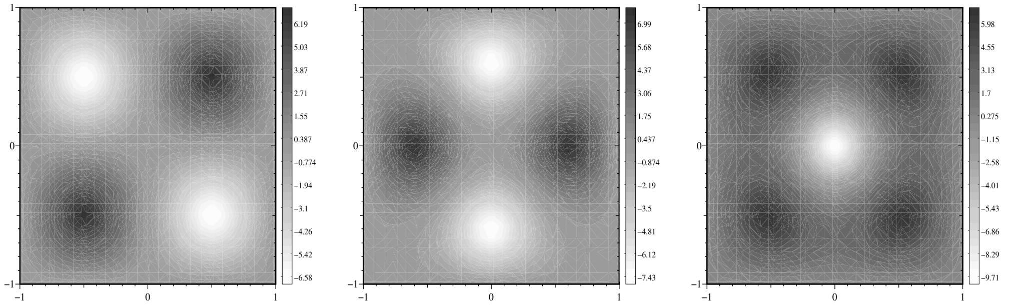


FIG. 2. *Solution contours to the Lane–Emden equation on* Ω_s, $J = 151.0$ (*left*), $J = 195.1$ (*center*), *and* $J = 233.0$ (*right*).

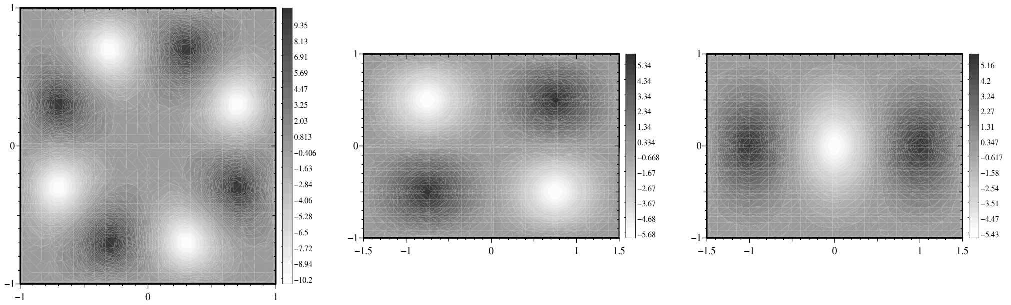


FIG. 3. *Contours of a solution to the Lane–Emden equation on* Ω_s, $J = 777.4$ (*left*). *Contours of two solutions to the Lane–Emden equation on* Ω_r, $J = 114.7$ (*center*) *and* $J = 80.34$ (*right*).

radially symmetric positive solution is the most unstable and, therefore, most elusive to capture among all positive solutions by LMM, since it will never get a sufficient support, unless one uses the radial symmetry to convert it into solving an ordinary differential equation. However, the method developed in this paper can easily capture this radially symmetric solution.

We point out that for a numerical computation using symmetry to be successful, it is important that its discretized mesh points must match the symmetry. In the numerical examples, $\varepsilon < 10^{-4}$ is used to terminate iterations, an initial guess v_0^1 is obtained from solving

$$\Delta v(x) = c(x), \quad x \in \Omega, \qquad \text{and} \qquad v(x) = 0, \quad x \in \partial\Omega,$$

where $c(x)$ is equal to -1 $(+1)$ if one wants $v(x)$ to be convex down (up) at x and is

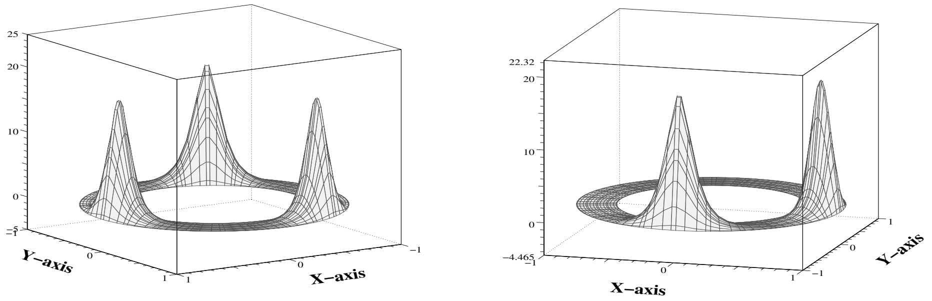


FIG. 4. *Two solutions to the Lane–Emden equation on Ω_a with $J = 870.7$ (left) and $J = 580.5$ (right).*

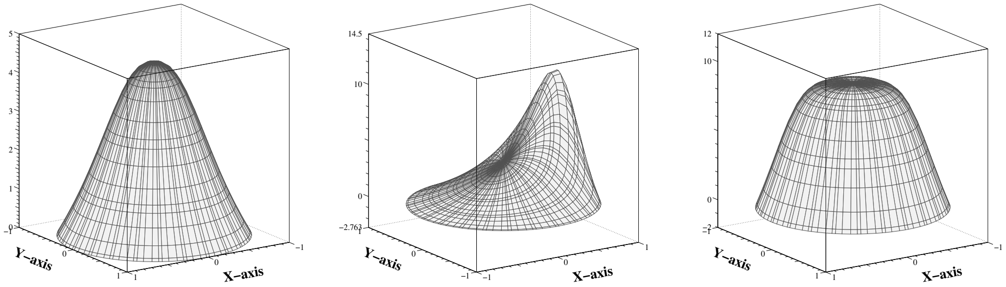


FIG. 5. *Solutions to the Henon equation on Ω_d with $q = 0.5, J = 21.5346$ (left), $q = 3, J = 114.4$ (center), and $q = 3, J = 171.7$ (right).*

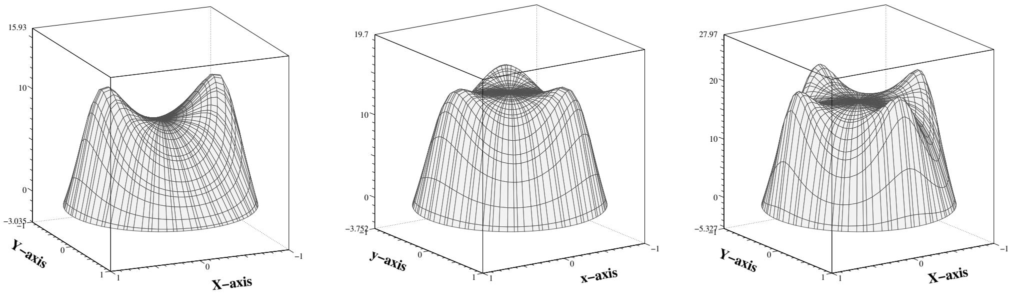


FIG. 6. *Solutions to the Henon equation on Ω_d with $q = 4, J = 285.9$ (left), $q = 6, J = 702.2$ (center), and $q = 9, J = 1815$ (right).*

equal to 0 if the profile of $v(x)$ is not of concern at x, and $L = \{0\}$, i.e., the smallest possible invariant subspace is used unless it is otherwise specified. In all the figures, J is the critical value. Figures 1, 2, and 3 are solutions to the Lane–Emden equation and we mainly want to demonstrate a variety of nodal solutions which are saddle points having large MIs. Figure 4 contains two positive solutions of the Lane-Emden equation on annular domains. Figures 5, 6, 7, 8, 9, and 10 are solutions to the Henon equation. For Figures 5 and 6, we explore the symmetry breaking phenomenon by showing that as the parameter q increases more and more nonradial positive solutions

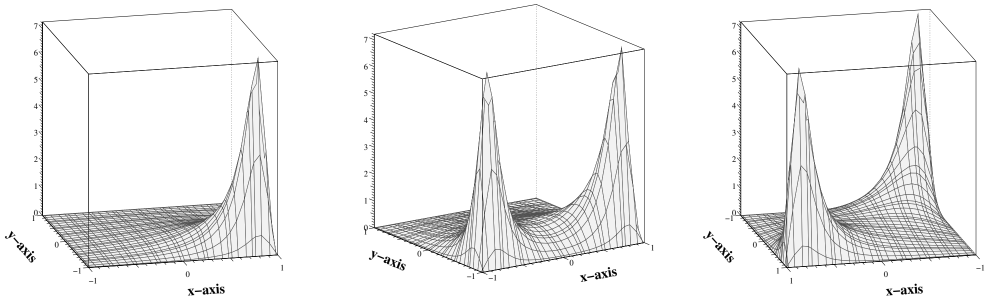


FIG. 7. *Solutions to the Henon equation on* Ω_s $(q = 9)$ *with* $J = 43.44$ (*left*), $J = 86.34$ (*center*), *and* $J = 86.61$ (*right*).

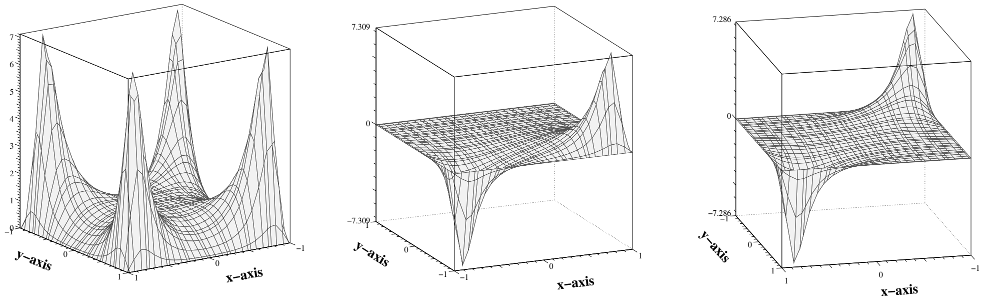


FIG. 8. *Solutions to the Henon equation on* Ω_s $(q = 9)$ *with* $J = 171.0$ (*left*), $J = 87.42$ (*center*), *and* $J = 87.15$ (*right*).

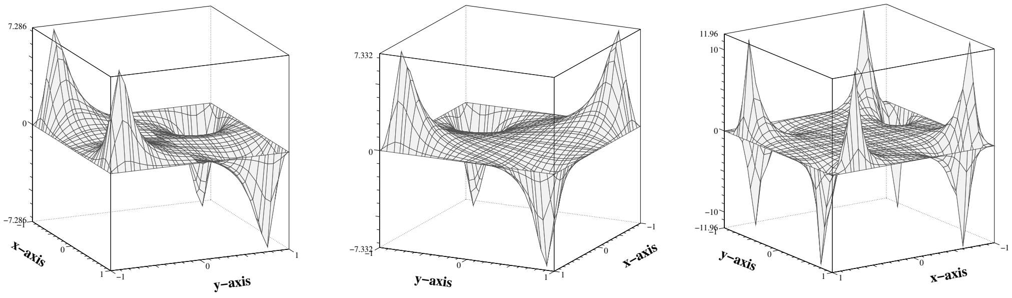


FIG. 9. *Solutions to the Henon equation on* Ω_s $(q = 9)$ *with* $J = 174.3$ (*left*), $J = 175.4$ (*center*), *and* $J = 1027$ (*right*).

appear which should also have large MIs. For Figures 7, 8, and 9, we show how to handle various even and/or odd symmetries. The symmetry in Figure 10 is partial and cannot be described by a compact group. The following list contains details on how we use symmetries to find each of the solutions, i.e., a formula is given in each case for us to compute $\mathcal{H}(d^k)$ and then to replace d^k in LMM.

(1) Cf. Figure 1 (left). No symmetry is needed or $(\mathcal{H}u)(x, y) = u(x, y)$. This is a mountain-pass solution with $\mathrm{MI} = 1$.

(2) Cf. Figure 1 (center). Either odd reflection about the line $y = x$ or odd re-

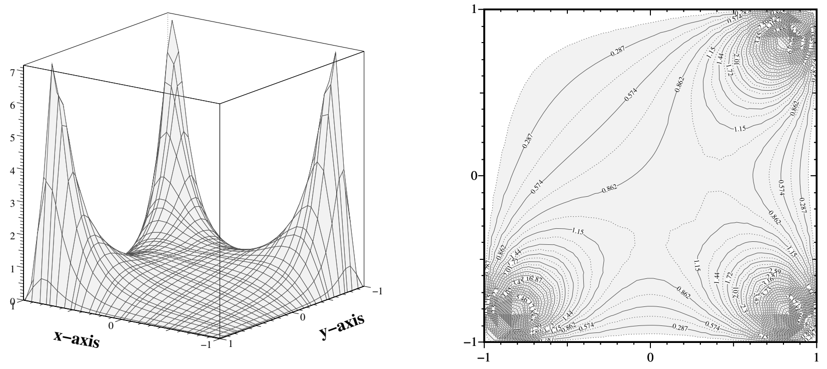


FIG. 10. *A solution to the Henon equation on Ω_s and its contours with $q = 9, J = 129.1$.*

flection about the origin. $(\mathcal{H}u)(x,y) = \frac{1}{2}(u(x,y) - u(-y,-x))$ or $(\mathcal{H}u)(x,y) = \frac{1}{2}(u(x,y) - u(-x,-y))$. This is a solution whose MI is at least 2 in H, and its MI is 1 when restricted in H_I.

(3) Cf. Figure 1 (right). Odd reflection about the x-axis. $(\mathcal{H}u)(x,y) = \frac{1}{2}(u(x,y) - u(x,-y))$.

(4) Cf. Figure 2 (left). Either odd reflections about the x-axis and the y-axis, or odd symmetry about the rotation by $\frac{\pi}{2}$. $(\mathcal{H}u)(x,y) = \frac{1}{4}(u(x,y) - u(-x,y) + u(-x,-y) - u(x,-y))$ or $(\mathcal{H}u)(\theta,r) = \frac{1}{4}(u(\theta,r) - u(\theta+\frac{\pi}{2},r) + u(\theta+\pi,r) - u(\theta+\frac{3\pi}{2},r))$. This solution has MI at least 4 in H though the MI $= 1$ in H_I.

(5) Cf. Figure 2 (center). Odd reflections about the lines $y = x$ and $y = -x$. $(\mathcal{H}u)(x,y) = \frac{1}{4}(u(x,y) - u(y,x) + u(-x,-y) - u(-y,-x))$.

(6) Cf. Figure 2 (right). Either even reflections about the x-axis and the y-axis, or even symmetry about the rotation by $\frac{\pi}{2}$. $L = \{u_1\}$ and u_1 is the solution in (1). $(\mathcal{H}u)(x,y) = \frac{1}{4}(u(x,y) + u(-x,y) + u(-x,-y) + u(x,-y))$ or $(\mathcal{H}u)(\theta,r) = \frac{1}{4}(u(\theta,r) + u(\theta+\frac{\pi}{2},r) + u(\theta+\pi,r) + u(\theta+\frac{3\pi}{2},r))$.

(7) Cf. Figure 3 (left). Odd reflections about the x-axis, the y-axis, the lines $y = x$ and $y = -x$. $\mathcal{H} = \mathcal{H}_4\mathcal{H}_3\mathcal{H}_2\mathcal{H}_1$, where $(\mathcal{H}_1u)(x,y) = \frac{1}{2}(u(x,y) - u(x,-y))$, $(\mathcal{H}_2u)(x,y) = \frac{1}{2}(u(x,y) - u(-x,y))$, $(\mathcal{H}_3u)(x,y) = \frac{1}{2}(u(x,y) - u(y,x))$, $(\mathcal{H}_4u)(x,y) = \frac{1}{2}(u(x,y) - u(-y,-x))$. Such a solution has a vary large MI in H and is too expensive to compute without using the symmetry.

(8) Cf. Figure 3 (center). Odd reflections about the x-axis and the y-axis. $(\mathcal{H}u)(x,y) = \frac{1}{4}(u(x,y) - u(-x,y) + u(-x,-y) - u(x,-y))$. Since the domain is a rectangle, the rotation by $\frac{\pi}{2}$ is not applicable. Contrast to Figure 2 (left).

(9) Cf. Figure 3 (right). Even reflections about the x-axis and the y-axis. $(\mathcal{H}u)(x,y) = \frac{1}{4}(u(x,y) + u(-x,y) + u(-x,-y) + u(x,-y))$. $L = \{u_1\}$, where u_1 is the single peak positive solution in the rectangle. This is the same symmetry used in the example shown in Figure 2 (right). It is interesting to compare this to Figure 2 (right). If we let the rectangle $[-1.5, 1.5] \times [-1, 1]$ change gradually to the square $[-1, 1] \times [-1, 1]$, e.g., $[-1.01, 1.01] \times [-1, 1]$, the solution remains of the same profile; it is an interesting observation. If we compare the critical values, profiles and symmetries of Figure 2 (left) and (right) with that of Figure 3 (center) and (right), we note that their sequential orders in critical level and profiles have changed drastically due to even a slight change in the geometry of the domain.

(10) Cf. Figure 4 (left). Rotation symmetry by $\frac{2\pi}{3}$. $(\mathcal{H}u)(\theta,r) = \frac{1}{3}(u(\theta,r) + u(\theta +$

$\frac{2\pi}{3}, r) + u(\theta + \frac{4\pi}{3}, r))$. Such a solution failed to be captured in [12] due to the fact that without using the symmetry, a sufficient support L in H contains infinitely many saddle points at lower critical level. We observe that the solution generated not only has the Z_3 symmetry it also has the additional symmetry of being even about the rotation by $\frac{2\pi}{3}$. Thus the solution has D_3 symmetry as well. With our new algorithm we may capture, in a stable way, solutions with D_k symmetry for any prime number k. The MIs of these solutions should be large depending upon the number of peaks k. This is related to the symmetry breaking phenomena for radially symmetric elliptic problems. The existence and qualitative behavior of these solutions can be found in [5] and the references therein. A rotation of the solution by any angle is still a solution. It is a degenerate case. However, adding an even symmetry about the x-axis can bypass the degeneracy.

(11) Cf. Figure 4 (right). Same symmetry as in Example 2.6. The existence of such a solution is still open. By Remark 2.1, we numerically checked that for each $v^k = p(u^k) \in H_I$ generated by LMM, $\nabla J(v^k) \in H_I$, i.e., the invariant subspace H_I is well defined. However, this problem is degenerate and also ill-conditioned in the sense that if we fix one peak and move another peak around, the change in J is almost invisible.

(12) Cf. Figure 5 (left). No symmetry is needed. This radial solution should be the unique positive solution to the Henon equation. A nonradial solution can be found for $q \geq 1$.

(13) Cf. Figure 5 (center). No symmetry is needed. However, adding an even symmetry about the x-axis will bypass the degeneracy. This is the least energy solution of the problem and this exhibits the phenomenon of symmetry breaking of ground state solutions (see [4]).

(14) Cf. Figure 5 (right). Even symmetry about the origin or rotation by π. $(\mathcal{H}u)(x,y) = \frac{1}{2}(u(x,y) + u(-x,-y))$. Such a radially symmetric solution is impossible to capture without using the symmetry since it has the highest critical value among all positive solutions and a sufficient support L in H needs to contain infinitely many positive solutions. A traditional way to find this solution is to use the radial symmetry to convert it into an ODE. On the other hand, this also shows that for small q (in this case $q = 3$) the radial solution is still the least energy solution in the class of even functions, and when we increase q to $q = 4$ as in the next example, the radial solution is going to lose its stability and the least energy solution becomes nonradial again.

(15) Cf. Figure 6 (left). Either even symmetry about the origin or even symmetry about the rotation by π. $(\mathcal{H}u)(x,y) = \frac{1}{2}(u(x,y) + u(-x,-y))$. Adding an even symmetry about the x-axis or y-axis will bypass the degeneracy. A symmetry about the rotation by $\frac{2\pi}{3}$ will generate the radially symmetric solution. From this example onward, the solutions demonstrated should have large MI depending upon the number of peaks of the solutions.

(16) Cf. Figure 6 (center). Symmetry about the rotation by $\frac{2\pi}{3}$. $(\mathcal{H}u)(\theta,r) = \frac{1}{3}(u(\theta,r) + u(\theta + \frac{2\pi}{3}, r) + u(\theta + \frac{4\pi}{3}, r))$. Adding an even symmetry about the x-axis will bypass the degeneracy. A symmetry about the rotation by $\frac{\pi}{2}$ will generate the radially symmetric solution. This implies the radial solution is the ground state in the class of functions invariant under Z_4, but is not the ground state in the class of Z_3-invariant functions.

(17) Cf. Figure 6 (right). Symmetry about the rotation by $\frac{\pi}{2}$. $(\mathcal{H}u)(\theta,r) = \frac{1}{4}(u(\theta,r)+$

$u(\theta+\frac{\pi}{4},r)+u(\theta+\frac{\pi}{2},r)+u(\theta+\frac{3\pi}{4},r))$. Adding an even symmetry about the x-axis or y-axis will bypass the degeneracy.

(18) Cf. Figure 7 (left). No symmetry is needed or $\mathcal{H}u(x,y)=u(x,y)$.

(19) Cf. Figure 7 (center). Even reflection about the y-axis. $(\mathcal{H}u)(x,y)=\frac{1}{2}(u(x,y)+u(-x,y))$.

(20) Cf. Figure 7 (right). Even reflection about the line $y=-x$. $(\mathcal{H}u)(x,y)=\frac{1}{2}(u(x,y)+u(-y,-x))$.

(21) Cf. Figure 8 (left). Even reflection about the x-axis and the y-axis. $(\mathcal{H}u)(x,y)=\frac{1}{4}(u(x,y)+u(-x,y)+u(-x,-y)+u(x,-y))$.

(22) Cf. Figure 8 (center). Odd reflection about the y-axis. $(\mathcal{H}u)(x,y)=\frac{1}{2}(u(x,y)-u(-x,y))$.

(23) Cf. Figure 8 (right). Odd reflection about the line $y=-x$. $(\mathcal{H}u)(x,y)=\frac{1}{2}(u(x,y)-u(-y,-x))$.

(24) Cf. Figure 9 (left). Even reflection about the y-axis and odd reflection about the x-axis.

(25) Cf. Figure 9 (center). Odd reflections about the x-axis and the y-axis. $(\mathcal{H}u)(x,y)=\frac{1}{4}(u(x,y)+u(-x,y)-u(-x,-y)-u(x,-y))$.

(26) Cf. Figure 9 (right). The same symmetry used in finding the solution in (7). Such a solution has a vary large MI in H and is too expensive to compute without using the symmetry.

(27) Cf. Figure 10. The same symmetry as in Example 2.6 and also in (11). It is an interesting example, since its symmetry is partial and cannot be described by a compact group. Theoretically the existence of such a solution is still open. However, we are able to follow Remark 2.1 to numerically verify the invariant subspace in the sense of Definition 2.1. For each $v^k=p(u^k)\in H_I$ generated in LMM, we find that the asymmetric part of $d^k=\nabla J(v^k)$ satisfies $\|\text{asymmetric part of } d^k\|_{H_0^1}\cong 2\times 10^{-4}\|d^k\|_{H_0^1}$. Thus we are very optimistic about the existence of this solution.

REFERENCES

[1] A. Ambrosetti and P. Rabinowitz, *Dual variational methods in critical point theory and applications,* J. Funct. Anal., 14 (1973), pp. 349–381.

[2] T. Bartsch and Z.-Q. Wang, *On the existence of sign-changing solutions for semilinear Dirichlet problems,* Topol. Methods Nonlinear Anal., 7 (1996), pp. 115–131.

[3] H. Brezis and L. Nirenberg, *Remarks on finding critical points,* Comm. Pure Appl. Math., 44 (1991), pp. 939–963.

[4] J. Byeon and Z.-Q. Wang, *On the Hénon Equation: Asymptotic Profile of Ground States,* preprint.

[5] F. Catrina and Z.-Q. Wang, *Nonlinear elliptic equations on expanding symmetric domains,* J. Differential Equations, 156 (1999), pp. 153–181.

[6] K.C. Chang, *Infinite Dimensional Morse Theory and Multiple Solution Problems,* Birkhäuser, Boston, 1993.

[7] Y.S. Choi and P.J. McKenna, *A mountain pass method for the numerical solution of semilinear elliptic problems,* Nonlinear Anal., 20 (1993), pp. 417–437.

[8] D. Costa, Z. Ding, and J. Neuberger, *A numerical investigation of sign-changing solutions to superlinear elliptic equations on symmetric domains,* J. Comput. Appl. Math., 131 (2001), pp. 299–319.

[9] W.Y. Ding and W.M. Ni, *On the existence of positive entire solutions of a semilinear elliptic equation,* Arch. Ration. Mech. Anal., 91 (1986), pp. 238–308.

[10] Z. Ding, D. Costa, and G. Chen, *A high linking method for sign changing solutions for semilinear elliptic equations,* Nonlinear Anal., 38 (1999), pp. 151–172.

[11] S. Li and Z.-Q. Wang, *Ljusternik–Schnirelman theory in partially ordered Hilbert spaces,* Trans. Amer. Math. Soc., 354 (2002), pp. 3207–3227.

[12] Y. Li and J. Zhou, *A minimax method for finding multiple critical points and its applications to semilinear PDEs,* SIAM J. Sci. Comput. 23 (2001), pp. 840–865.

[13] Y. Li and J. Zhou, *Convergence results of a minimax method for finding multiple critical points,* SIAM J. Sci. Comput., 24 (2002), pp. 865–885.

[14] Y. Li and J. Zhou, *Local characterizations of saddle points and their Morse indices,* in Control Nonlinear Distributed Parameter Systems, G. Chen, I. Lasiecka, and J. Zhou, eds., Lecture Notes in Pure and Appl. Math. 218, Marcel Dekker, New York, 2001, pp. 233–251.

[15] Z. Liu and J. Sun, *Invariant sets of descending flow in critical point theory with applications to nonlinear differential equations,* J. Differential Equations, 172 (2001), pp. 257–299.

[16] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems,* Springer-Verlag, New York, 1989.

[17] Z. Nehari, *On a class of nonlinear second-order differential equations,* Trans. Amer. Math. Soc., 95 (1960), pp. 101–123.

[18] W.M. Ni, *Some Aspects of Semilinear Elliptic Equations,* Department of Mathematics, National Tsing Hua University, Hsinchu, Taiwan, Republic of China, 1987.

[19] W.M. Ni, *Recent Progress in Semilinear Elliptic Equations,* RIMS Kokyuroku 679, Kyoto University, Kyoto, Japan, 1989, pp. 1–39.

[20] P. Rabinowitz, *Minimax Method in Critical Point Theory with Applications to Differential Equations,* CBMS Reg. Conf. Ser. Math. 65, AMS, Providence, RI, 1986.

[21] M. Schechter, *Linking Methods in Critical Point Theory,* Birkhäuser, Boston, 1999.

[22] M. Struwe, *Variational Methods,* Springer-Verlag, New York, 1996.

[23] Z.-Q. Wang, *On a superlinear elliptic equation,* Ann. Inst. H. Poincarè, 8 (1991), pp. 43–57.

[24] M. Willem, *Minimax Theorems,* Birkhäuser, Boston, 1996.

[25] J. Zhou, *Instability indices of saddle points by a local minimax method,* Math. Comp., 74 (2005), pp. 1391–1411.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 909–923

SHARP ERROR ESTIMATES FOR INTERPOLATORY APPROXIMATION ON CONVEX POLYTOPES*

ALLAL GUESSAB† AND GERHARD SCHMEISSER‡

Abstract. Let $\mathfrak{P}$ be a convex polytope in the d-dimensional Euclidean space. We consider an interpolation of a function f at the vertices of $\mathfrak{P}$ and compare it with the interpolation of f and its derivative at a fixed point $y \in \mathfrak{P}$. The two methods may be seen as multivariate analogues of an interpolation by secants and tangents, respectively. For twice continuously differentiable functions, we establish sharp error estimates with respect to a generalized L^p norm for $1 \leq p \leq \infty$. The case $p = 1$ is of special interest since it provides analogues of the midpoint rule and the trapezoidal rule for approximate integration over the polytope $\mathfrak{P}$. In the case where $\mathfrak{P}$ is a simplex and $p > 1$, this investigation covers recent results by S. Waldron [*SIAM J. Numer. Anal.*, 35 (1998), pp. 1191–1200] and by M. Stämpfle [*J. Approx. Theory*, 103 (2000), pp. 78–90].

Key words. interpolation on convex polytopes, sharp error estimates, approximation of functions, approximate integration, approximation of functionals

AMS subject classifications. 41A05, 41A20, 41A44, 41A63, 41A80, 65D05, 65D30

DOI. 10.1137/S0036142903435958

1. Introduction and notation.

Denote by $\mathcal{P}_1$ the class of all polynomials in d real variables of degree at most 1, also called the class of *affine functions* on $\mathbb{R}^d$. Let $\mathfrak{P} \subset \mathbb{R}^d$ be a convex polytope of positive measure with vertices $v_1, \ldots, v_n$, and let $B_1, \ldots, B_n$ be an associated system of continuous functions on $\mathfrak{P}$ with the following properties.

Nonnegativity. For $i = 1, \ldots, n$, we have

$$B_i(x) \geq 0 \qquad (x \in \mathfrak{P}). \tag{1.1}$$

Linear precision. For every $\lambda \in \mathcal{P}_1$, we have

$$\lambda(x) = \sum_{i=1}^{n} \lambda(v_i) B_i(x). \tag{1.2}$$

Warren [10] showed that $B_1, \ldots, B_n$ can be chosen as rational functions, which are uniquely determined if one requires that each B_i have minimal degree. Furthermore, for an arbitrary convex polytope, he presented an algorithm for constructing these functions $B_1, \ldots, B_n$ in a finite number of steps.

Since vertices of a convex polytope are extremal points, it is easily deduced from the "linear precision" that

$$B_i(v_j) = \delta_{ij} \qquad (i, j \in \{1, \ldots, n\}), \tag{1.3}$$

where we use Kronecker's delta. As a consequence of (1.2) and (1.3), the functions $B_1, \ldots, B_n$ are linearly independent and span an n-dimensional linear space $\mathcal{R}_n$ which contains $\mathcal{P}_1$ as a subspace.

*Received by the editors October 10, 2003; accepted for publication (in revised form) January 13, 2005; published electronically September 20, 2005.
http://www.siam.org/journals/sinum/43-3/43595.html

†Laboratoire de Mathématiques Appliquées de Pau, Université de Pau et de Pays de l'Adour, 64013 Pau, France (allal.guessab@univ-pau.fr).

‡Mathematical Institute, University of Erlangen-Nuremberg, 91054 Erlangen, Germany (schmeisser@mi.uni-erlangen.de).

By $C(\mathfrak{P})$, $C^1(\mathfrak{P})$, and $C^2(\mathfrak{P})$, we denote the spaces of functions which are defined on $\mathfrak{P}$ and are continuous, continuously differentiable, and twice continuously differentiable, respectively.

Next, let $\mathcal{L}$ be a positive linear functional on $C(\mathfrak{P})$. The positivity means that $\mathcal{L}(f) > 0$ for every nontrivial nonnegative function $f \in C(\mathfrak{P})$.

Examples of such functionals are weighted integrals

$$\mathcal{L}(f) := \int_{\mathfrak{P}} w(x)f(x)\mathrm{d}x \qquad \bigl(f \in C(\mathfrak{P})\bigr), \tag{1.4}$$

where w is integrable and positive on $\mathfrak{P}$ except for a set of measure zero.

For $f \in C(\mathfrak{P})$, we introduce

$$\|f\|_p := \bigl(\mathcal{L}(|f|^p)\bigr)^{1/p} \qquad (1 \le p < \infty) \tag{1.5}$$

and

$$\|f\|_\infty := \sup_{x\in\mathfrak{P}} |f(x)|\,, \tag{1.6}$$

which define norms on $C(\mathfrak{P})$. When $\mathcal{L}$ is given by (1.4) and $w = 1$, then $\|\cdot\|_p$ is the familiar L^p norm. For general $\mathcal{L}$, we may think of $\mathfrak{P}$ as being equipped with a mass distribution such that $\mathcal{L}(1)$ is the total mass of $\mathfrak{P}$. The possibility of having an arbitrary $\mathcal{L}$ is of interest mainly in our applications of the case $p = 1$ (see section 4). For this reason, we do not use a weighted supremum norm.

By $\|\cdot\|$, *without* any subscript, and by $\langle\cdot,\cdot\rangle$, we want to denote the Euclidean norm and the standard inner product in $\mathbb{R}^d$.

In this paper, we shall study the linear interpolation operator Λ^{v}, defined by

$$\Lambda^{\mathrm{v}}[f] := \sum_{i=1}^{n} f(v_i)B_i \qquad \bigl(f \in C(\mathfrak{P})\bigr), \tag{1.7}$$

which interpolates f at the vertices of $\mathfrak{P}$, and shall compare it with

$$\Lambda_y[f] := f(y) + Df(y)(\cdot - y) \qquad \bigl(f \in C^1(\mathfrak{P})\bigr), \tag{1.8}$$

where $y \in \mathfrak{P}$. Clearly, $\Lambda_y[f]$ interpolates f at y, and the same holds for the first derivative.

As regards our notation, we want to follow the convention that a superscript in roman type indicates an abbreviation for a word, while a subscript in italic type is a mathematical quantity. In particular, the superscript v shall always refer to interpolation at the *vertices*. Similarly, we shall use the superscripts sb for *smallest ball* and cm for *center of mass*.

2. Auxiliary results. For convenient reference, we first state some properties of the operators Λ_y and Λ^{v} as lemmas.

LEMMA 2.1. *For $y \in \mathfrak{P}$, the operator Λ_y has the following properties:*

(i) *It maps $C^1(\mathfrak{P})$ into $\mathcal{P}_1$.*

(ii) *It reproduces functions from $\mathcal{P}_1$.*

(iii) *It approximates convex functions from below.*

Proof. The properties (i) and (ii) are obvious. Property (iii) is a well-known fact about differentiable, convex functions; see [5, Theorem A, p. 98]. ◻

LEMMA 2.2. *The operator Λ^{v} has the following properties:*

(i) *It maps $C(\mathfrak{P})$ into $\mathcal{R}_n$.*
(ii) *It reproduces functions from $\mathcal{R}_n$.*
(iii) *It approximates convex functions from above.*
(iv) *If $f, g \in C(\mathfrak{P})$ and $f(v_i) \le g(v_i)$ for $i = 1, \dots, n$, then $\Lambda^{\mathrm{v}}[f] \le \Lambda^{\mathrm{v}}[g]$.*

Proof. Since $\{B_1, \dots, B_n\}$ is a basis of $\mathcal{R}_n$, the properties (i) and (ii) are obvious consequences of the definition of Λ^{v}.

Next, it follows from (1.2) that

$$x \,=\, \sum_{i=1}^{n} v_i B_i(x) \qquad (x \in \mathfrak{P}),$$

which is a representation of x as a convex combination of the vertices of $\mathfrak{P}$. Hence, for a convex function f,

$$f(x) \,=\, f\left(\sum_{i=1}^{n} v_i B_i(x)\right) \,\le\, \sum_{i=1}^{n} f(v_i) B_i(x) \,=\, \Lambda^{\mathrm{v}}[f](x) \qquad (x \in \mathfrak{P}),$$

and so statement (iii) is verified.

Finally, recalling (1.1), we see that, under the hypothesis of statement (iv),

$$\Lambda^{\mathrm{v}}[f] \,=\, \sum_{i=1}^{n} f(v_i) B_i \,\le\, \sum_{i=1}^{n} g(v_i) B_i \,=\, \Lambda^{\mathrm{v}}[g].$$

This completes the proof. □

It will turn out that the constants in our error estimates are determined by the interpolation error of the quadratic function $\|\cdot\|^2$. We therefore introduce the (nonnegative) functions

$$e_y \,:=\, \|\cdot\|^2 - \Lambda_y\left[\|\cdot\|^2\right], \tag{2.1}$$

where $y \in \mathfrak{P}$, and

$$e^{\mathrm{v}} \,:=\, \Lambda^{\mathrm{v}}\left[\|\cdot\|^2\right] - \|\cdot\|^2 \,=\, \sum_{i=1}^{n} \|v_i\|^2 B_i - \|\cdot\|^2. \tag{2.2}$$

Representations, interrelations, and estimates for these functions are stated in the following lemma.

LEMMA 2.3. *The functions e_y and e^{v} are nonnegative and vanish at the interpolation points of Λ_y and Λ^{v}, respectively. They satisfy the equations*

$$e_y = \|\cdot - y\|^2, \tag{2.3}$$

$$e^{\mathrm{v}} = \sum_{i=1}^{n} \|\cdot - v_i\|^2 B_i, \tag{2.4}$$

$$e^{\mathrm{v}} + e_y = \sum_{i=1}^{n} e_y(v_i) B_i. \tag{2.5}$$

Furthermore, denoting by

$$\mathfrak{B}^{\mathrm{sb}} \,=:\, \{x \in \mathbb{R}^d \,:\, \|x - x^{\mathrm{sb}}\| \le r^{\mathrm{sb}}\} \tag{2.6}$$

the smallest ball that contains $\mathfrak{P}$*, we have*

$$e^{\mathrm{v}}(x) \le (r^{\mathrm{sb}})^2 - \|x - x^{\mathrm{sb}}\|^2 \le (r^{\mathrm{sb}})^2 \tag{2.7}$$

for all $x \in \mathfrak{P}$.

For notational simplicity, we write

$$\Lambda^{\mathrm{sb}} := \Lambda_y \quad \text{and} \quad e^{\mathrm{sb}} := e_y \quad \text{if } y = x^{\mathrm{sb}}. \tag{2.8}$$

Proof. From the definition of the functions e_y and e^{v}, it is clear that they vanish at the interpolation points of Λ_y and Λ^{v}, respectively. Since $\|\cdot\|^2$ is a convex function, the statements (iii) of Lemmas 2.1 and 2.2 show that e_y and e^{v} are nonnegative.

Next, from the definition of e_y, we deduce that

$$e_y(x) = \|x\|^2 - \big(\|y\|^2 + 2\langle y, x - y\rangle\big) = \|x\|^2 + \|y\|^2 - 2\langle y, x\rangle = \|x - y\|^2,$$

which is (2.3).

Since $e^{\mathrm{v}} + e_y$ belongs to $\mathcal{R}_n$, statement (ii) of Lemma 2.2 shows that, for any $x \in \mathfrak{P}$, we have

$$e^{\mathrm{v}}(x) + e_y(x) = \sum_{i=1}^{n} \big(e^{\mathrm{v}}(v_i) + e_y(v_i)\big)\, B_i(x) = \sum_{i=1}^{n} e_y(v_i) B_i(x), \tag{2.9}$$

which is (2.5).

Substituting $y = x$ in (2.9) and using (2.3), we obtain (2.4).

For a proof of (2.7), we first note that $x^{\mathrm{sb}} \in \mathfrak{P}$, as a consequence of the convexity of $\mathfrak{P}$. Since

$$h^{\mathrm{sb}} := (r^{\mathrm{sb}})^2 - \|\cdot - x^{\mathrm{sb}}\|^2 \tag{2.10}$$

is nonnegative on $\mathfrak{P}$, while e^{v} vanishes at *all* the vertices of $\mathfrak{P}$, we clearly have

$$h^{\mathrm{sb}}(v_i) - e^{\mathrm{v}}(v_i) \ge 0 \qquad (i = 1, \ldots, n).$$

Therefore statement (iv) of Lemma 2.2 implies that $\Lambda^{\mathrm{v}}[h^{\mathrm{sb}} - e^{\mathrm{v}}] \ge 0$. Furthermore, using (2.3), (2.5), and the notation (2.8), we find that

$$h^{\mathrm{sb}} - e^{\mathrm{v}} = (r^{\mathrm{sb}})^2 - e^{\mathrm{sb}} - e^{\mathrm{v}} = (r^{\mathrm{sb}})^2 - \sum_{i=1}^{n} e^{\mathrm{sb}}(v_i) B_i, \tag{2.11}$$

which obviously belongs to $\mathcal{R}_n$. Hence statement (ii) of Lemma 2.2 allows us to conclude that

$$h^{\mathrm{sb}} - e^{\mathrm{v}} = \Lambda^{\mathrm{v}}\big[h^{\mathrm{sb}} - e^{\mathrm{v}}\big] \ge 0, \tag{2.12}$$

which gives (2.7) immediately. □

Remark 2.4. Inequality (2.7) is of interest for the following reason. As we shall see, the best constants in our error estimates for $\Lambda^{\mathrm{v}}[f]$ are determined by norms of e^{v}. If e^{v} is complicated, then we may use the simpler function (2.10) instead and obtain a constant which is possibly somewhat worse, but which may still be good enough for practical applications. In the case where $\mathfrak{P}$ is a simplex, it can even be shown that

$$\sup_{x \in \mathfrak{P}} e^{\mathrm{v}}(x) = \sup_{x \in \mathfrak{P}} h^{\mathrm{sb}}(x) = (r^{\mathrm{sb}})^2;$$

see [6, Lemma 4.2].

3. Approximation of functions. We are mainly interested in approximation of functions from $C^2(\mathfrak{P})$. However, in the case where $\mathcal{P}$ is a simplex, Stämpfle [6, Theorem 4.1, statements (i)–(iv)] also presented results for functions belonging to lower regularity classes. These statements extend to Λ^{v} by exactly the same arguments as in [6]. We only mention a result for a Lipschitz class which is more general than the one considered in [6].

For $\alpha \in (0,1]$ and $L > 0$, we write $f \in \mathrm{Lip}_L(\alpha, \mathfrak{P})$ and say that f satisfies a *Lipschitz condition* of *order* α with *Lipschitz constant* L on $\mathfrak{P}$ if $f \in C(\mathfrak{P})$ and

$$|f(x) - f(y)| \,\leq\, L\|x-y\|^{\alpha} \qquad (x, y \in \mathfrak{P}).$$

THEOREM 3.1. *Let* $f \in \mathrm{Lip}_L(\alpha, \mathfrak{P})$. *Then*

$$\big|f(x) - \Lambda^{\mathrm{v}}[f](x)\big| \,\leq\, L\left(e^{\mathrm{v}}(x)\right)^{\alpha/2} \qquad (x \in \mathfrak{P}) \tag{3.1}$$

and, for each $p \in [1, \infty]$,

$$\big\|f - \Lambda^{\mathrm{v}}[f]\big\|_p \,\leq\, L\,\big\|(e^{\mathrm{v}})^{\alpha/2}\big\|_p. \tag{3.2}$$

Proof. From (1.2) and the definition of Λ^{v}, it is clear that

$$f(x) - \Lambda^{\mathrm{v}}[f](x) \,=\, \sum_{i=1}^{n}\big(f(x) - f(v_i)\big)\,B_i(x),$$

and so, by the triangle inequality and the Lipschitz condition,

$$\big|f(x) - \Lambda^{\mathrm{v}}[f](x)\big| \,\leq\, L\sum_{i=1}^{n}\|x - v_i\|^{\alpha}B_i(x). \tag{3.3}$$

Next, using Hölder's inequality with $p := 2/\alpha$ and $q := 2/(2-\alpha)$, which is an admissible pair of exponents, and recalling (1.2) and (2.4), we find that

$$\begin{aligned}\sum_{i=1}^{n}\|x-v_i\|^{\alpha}B_i(x) &= \sum_{i=1}^{n}\|x-v_i\|^{\alpha}B_i(x)^{1/p}\cdot B_i(x)^{1/q}\\ &\leq \left(\sum_{i=1}^{n}\|x-v_i\|^{\alpha p}B_i(x)\right)^{1/p}\cdot\left(\sum_{i=1}^{n}B_i(x)\right)^{1/q}\\ &= \left(\sum_{i=1}^{n}\|x-v_i\|^{2}B_i(x)\right)^{\alpha/2} \,=\, \left(e^{\mathrm{v}}(x)\right)^{\alpha/2}.\end{aligned}$$

Combining this with (3.3), we obtain (3.1). Clearly, (3.2) is an immediate consequence of (3.1). □

For twice differentiable functions $f : \mathfrak{P} \to \mathbb{R}$, we denote by

$$H[f](x) \,:=\, \left(\frac{\partial^2 f}{\partial x_i \partial x_j}(x)\right)_{i,j=1,\ldots,d}$$

the Hessian matrix of f at x and introduce

$$\boldsymbol{|}D^2 f\boldsymbol{|} \,:=\, \sup_{x\in\mathfrak{P}}\,\sup_{\substack{y\in\mathbb{R}^d\\ \|y\|=1}}\left|y^{\top}H[f](x)y\right|, \tag{3.4}$$

agreeing that the elements of $\mathbb{R}^d$ are column vectors so that $y^\top$, which denotes the transpose of y, becomes a row vector. Clearly, $\left|D^2 f\right| = 0$ for $f \in \mathcal{P}_1$, and $\left|D^2 f\right| = 2\,|c|$ for $f = c\|\cdot\|^2$.

Subsequently, we shall often refer to the space

$$\mathcal{F}_2 := \left\{ f := \lambda + c\,\|\cdot\|^2 \;:\; \lambda \in \mathcal{P}_1,\, c \in \mathbb{R} \right\}. \tag{3.5}$$

The following theorem for Λ_y is not more than an easy exercise in calculus. We formulate it as a theorem only in order to compare it with the corresponding result for Λ^{v}.

Theorem 3.2. *Let* $f \in C^2(\mathfrak{P})$. *Then,*

$$\left|f(x) - \Lambda_y[f](x)\right| \le \frac{1}{2}\,\|x-y\|^2\,\left|D^2 f\right| \qquad (x, y \in \mathfrak{P}). \tag{3.6}$$

Furthermore, for each $p \in [1, \infty]$,

$$\left\|f - \Lambda_y[f]\right\|_p \le c_{y,p}\,\left|D^2 f\right|, \tag{3.7}$$

where

$$c_{y,p} := \frac{1}{2}\,\|e_y\|_p. \tag{3.8}$$

Both inequalities are sharp. Equality is attained for every $f \in \mathcal{F}_2$.

Proof. By the Taylor formula of order two, we have

$$f(x) - \Lambda_y[f](x) = \frac{1}{2}\,(x-y)^\top H[f]\big(y + \theta(x-y)\big)(x-y)$$

for some $\theta \in (0,1)$. Now the definition of $\left|D^2 f\right|$, given in (3.4), shows that (3.6) holds. Inequality (3.7) is an immediate consequence of (3.6). Finally, the case of equality is easily verified. □

Theorem 3.3. *Let* $f \in C^2(\mathfrak{P})$. *Then,*

$$\left|f(x) - \Lambda^{\mathrm{v}}[f](x)\right| \le \frac{1}{2}\,e^{\mathrm{v}}(x)\,\left|D^2 f\right| \qquad (x \in \mathfrak{P}). \tag{3.9}$$

Furthermore, for each $p \in [1, \infty]$,

$$\left\|f - \Lambda^{\mathrm{v}}[f]\right\|_p \le c_p^{\mathrm{v}}\,\left|D^2 f\right|, \tag{3.10}$$

where

$$c_p^{\mathrm{v}} := \frac{1}{2}\,\|e^{\mathrm{v}}\|_p. \tag{3.11}$$

Both inequalities are sharp. Equality is attained for every $f \in \mathcal{F}_2$.

Proof. Inequality (3.6) may be rewritten as

$$-\frac{1}{2}\,\|x-y\|^2\,\left|D^2 f\right| \le f(x) - \Lambda_y[f](x) \le \frac{1}{2}\,\|x-y\|^2\,\left|D^2 f\right| \quad (x, y \in \mathfrak{P}). \tag{3.12}$$

Next, from statement (iv) of Lemma 2.2, it follows that inequalities between continuous functions on $\mathfrak{P}$ are preserved when the operator Λ^{v} is applied on both sides.

Moreover, statement (i) of Lemma 2.1 together with statement (ii) of Lemma 2.2 shows that

$$\Lambda^{\mathrm{v}}\big[\Lambda_y[f]\big] \,=\, \Lambda_y[f].$$

Hence (3.12) implies that

$$-\frac{1}{2}\,\Lambda^{\mathrm{v}}\big[\|\cdot - y\|^2\big](x)\,\,\big|D^2 f\big| \,\le\, \Lambda^{\mathrm{v}}[f](x) - \Lambda_y[f](x) \,\le\, \frac{1}{2}\,\Lambda^{\mathrm{v}}\big[\|\cdot - y\|^2\big](x)\,\,\big|D^2 f\big|.$$

Now, taking $y = x$ and noting that $\Lambda_x[f](x) = f(x)$ and, by (2.4),

$$\Lambda^{\mathrm{v}}\big[\|\cdot - x\|^2\big](x) \,=\, \sum_{i=1}^{n} \|v_i - x\|^2 B_i(x) \,=\, e^{\mathrm{v}}(x),$$

we obtain

$$-\frac{1}{2}\,e^{\mathrm{v}}(x)\,\,\big|D^2 f\big| \,\le\, \Lambda^{\mathrm{v}}[f](x) - f(x) \,\le\, \frac{1}{2}\,e^{\mathrm{v}}(x)\,\,\big|D^2 f\big|,$$

which is equivalent to (3.9). Inequality (3.10) is an immediate consequence of (3.9). The statement on the occurrence of equality is easily verified by a calculation. ∎

Remark 3.4. Since Λ^{v} is a positive operator which reproduces affine functions, inequality (3.9) can also be deduced from [9, Theorem 1.4] in conjunction with the above Lemma 2.3.

The operator Λ_y has just one interpolation point, which is of multiplicity two. Such an interpolation can be described by $d+1$ scalar equations. The interpolation of the operator Λ^{v}, which has n simple interpolation points, can be described by n scalar equations. Since $n \ge d+1$, we may expect that the operator Λ^{v} is at least as precise as Λ_y. In the following proposition, we compare the constants (3.8) and (3.11) when $p = \infty$.

PROPOSITION 3.5. *For $p = \infty$, the constants* (3.8) *and* (3.11) *satisfy the relations*

$$c^{\mathrm{v}}_{\infty} \,\le\, c_{y,\infty} \qquad (y \in \mathfrak{P}) \tag{3.13}$$

and

$$\inf_{y\in\mathfrak{P}} c_{y,\infty} \,=\, \frac{(r^{\mathrm{sb}})^2}{2}, \tag{3.14}$$

the infimum being attained for $y = x^{\mathrm{sb}}$, where r^{sb} and x^{sb} specify the smallest ball $\mathfrak{B}^{\mathrm{sb}}$ which contains $\mathfrak{P}$, as introduced in (2.6).

If all the vertices of $\mathfrak{P}$ lie on the boundary of $\mathfrak{B}^{\mathrm{sb}}$, then

$$c^{\mathrm{v}}_{\infty} \,=\, \frac{(r^{\mathrm{sb}})^2}{2}. \tag{3.15}$$

Proof. It follows from (2.5) that

$$e^{\mathrm{v}}(x) \,\le\, \sum_{i=1}^{n} e_y(v_i) B_i(x) \,\le\, \max_{1\le i\le n} e_y(v_i) \qquad (x \in \mathfrak{P}), \tag{3.16}$$

which implies (3.13).

Since a convex function, defined on a convex set, attains its supremum at an extreme point (see, for example, [4, p. 91]), we have

$$\max_{1\le i\le n} e_y(v_i) \;=\; \sup_{x\in\mathfrak{P}} e_y(x) \;=\; 2c_{y,\infty}. \tag{3.17}$$

This shows that $c_{y,\infty}$ attains its smallest value at a point where

$$\phi(y) \;:=\; \max_{1\le i\le n} e_y(v_i) \;=\; \max_{1\le i\le n} \|y-v_i\|^2$$

attains its minimum. Clearly, this is the center of the smallest ball $\mathfrak{B}^{\rm sb}$ that contains $\mathfrak{P}$, and so

$$\min_{y\in\mathfrak{P}} \phi(y) \;=\; \phi(x^{\rm sb}) \;=\; (r^{\rm sb})^2.$$

Thus (3.14) is verified.

If *all* the vertices of $\mathfrak{P}$ lie on the boundary of $\mathfrak{B}^{\rm sb}$, then $\|x^{\rm sb}-v_i\| = r^{\rm sb}$ for $i=1,\dots,n$. Therefore, by (2.4),

$$e^{\rm v}(x^{\rm sb}) \;=\; \sum_{i=1}^n \|x^{\rm sb}-v_i\|^2 B_i(x^{\rm sb}) \;=\; (r^{\rm sb})^2 \sum_{i=1}^n B_i(x^{\rm sb}) \;=\; (r^{\rm sb})^2,$$

which shows that $c_\infty^{\rm v} \ge (r^{\rm sb})^2/2$. Combining this inequality with (3.13) and (3.14), we obtain (3.15). □

In the univariate case, where $\mathfrak{P}$ is an interval $[a,b]$, it is known and also seen from (3.15) that, for $y=(b+a)/2$, we have

$$c_\infty^{\rm v} \;=\; c_{y,\infty} \;=\; \frac{(b-a)^2}{8}.$$

Moreover, the mean value

$$\frac{1}{2}\left(\Lambda_y[f] + \Lambda^{\rm v}[f]\right) \qquad \left(y = \frac{a+b}{2}\right)$$

gives an approximation whose constant in the error bound is $(b-a)^2/16$. A generalization is given in the following proposition.

PROPOSITION 3.6. *Let $f\in C^2(\mathfrak{P})$. Then, for every $y\in\mathfrak{P}$ and $\alpha\in[0,1]$, we have*

$$\left\|f-\alpha\Lambda_y[f]-(1-\alpha)\Lambda^{\rm v}[f]\right\|_\infty \;\le\; c(\alpha,y)\;\left|D^2 f\right|, \tag{3.18}$$

where

$$c(\alpha,y) \;:=\; \frac{1}{2}\sup_{x\in\mathfrak{P}}\left(\alpha e_y(x) + (1-\alpha)e^{\rm v}(x)\right). \tag{3.19}$$

Furthermore,

$$\inf_{0\le\alpha\le1}\,\inf_{y\in\mathfrak{P}} c(\alpha,y) \;\le\; \frac{(r^{\rm sb})^2}{4} \;=\; c\left(\frac{1}{2},x^{\rm sb}\right), \tag{3.20}$$

where $r^{\rm sb}$ and $x^{\rm sb}$ are the radius and the center, respectively, of the smallest ball $\mathfrak{B}^{\rm sb}$ that contains $\mathfrak{P}$. Equality occurs in (3.20) *if all the vertices of $\mathfrak{P}$ lie on the boundary* *of* $\mathfrak{B}^{\mathrm{sb}}$. *In this case, inequality* (3.18) *is sharp when* $\alpha = 1/2$ *and* $y = x^{\mathrm{sb}}$, *and equality is attained for every function* $f \in \mathcal{F}_2$.

Proof. The estimates (3.6) and (3.9) may be rewritten as

$$-\frac{1}{2}e_y(x)\ \left|D^2 f\right|\ \le\ f(x) - \Lambda_y[f](x)\ \le\ \frac{1}{2}e_y(x)\ \left|D^2 f\right|$$

and

$$-\frac{1}{2}e^{\mathrm{v}}(x)\ \left|D^2 f\right|\ \le\ f(x) - \Lambda^{\mathrm{v}}[f](x)\ \le\ \frac{1}{2}e^{\mathrm{v}}(x)\ \left|D^2 f\right|.$$

Multiplying the first inequalities by α and the second by $1-\alpha$, and adding the results, we obtain

$$\left|f(x) - \alpha\Lambda_y[f](x) - (1-\alpha)\Lambda^{\mathrm{v}}[f](x)\right|\ \le\ \frac{1}{2}\left(\alpha e_y(x) + (1-\alpha)e^{\mathrm{v}}(x)\right)\ \left|D^2 f\right|.$$

This implies (3.18).

Next, using (2.5) and the notation (2.8), we find that

$$c\left(\frac{1}{2}, x^{\mathrm{sb}}\right)\ =\ \frac{1}{4}\sup_{x\in\mathfrak{P}}\left(e^{\mathrm{v}}(x) + e^{\mathrm{sb}}(x)\right)\ =\ \frac{1}{4}\sup_{x\in\mathfrak{P}}\sum_{i=1}^{n}\|x^{\mathrm{sb}} - v_i\|^2 B_i(x).$$

If v_j is a vertex on the boundary of $\mathfrak{B}^{\mathrm{sb}}$, then, by (1.1), (1.3), (2.11), and (2.12),

$$\sum_{i=1}^{n}\|x^{\mathrm{sb}} - v_i\|^2 B_i(v_j)\ =\ \|x^{\mathrm{sb}} - v_j\|^2\ =\ (r^{\mathrm{sb}})^2\ \ge\ \sum_{i=1}^{n}\|x^{\mathrm{sb}} - v_i\|^2 B_i(x)$$

for all $x \in \mathfrak{P}$. This shows that

$$\sup_{x\in\mathfrak{P}}\sum_{i=1}^{n}\|x^{\mathrm{sb}} - v_i\|^2 B_i(x)\ =\ (r^{\mathrm{sb}})^2$$

and completes the proof of (3.20).

Using (3.19), we deduce that

$$c(\alpha, y)\ \ge\ \frac{1-\alpha}{2}\sup_{x\in\mathfrak{P}} e^{\mathrm{v}}(x)\ =\ (1-\alpha)\,c^{\mathrm{v}}_{\infty}\ \ge\ \frac{c^{\mathrm{v}}_{\infty}}{2}\qquad \text{if } \alpha \in \left[0, \frac{1}{2}\right]$$

and, in conjunction with (3.14),

$$c(\alpha, y)\ \ge\ \frac{\alpha}{2}\sup_{x\in\mathfrak{P}} e_y(x)\ =\ \alpha\, c_{y,\infty}\ \ge\ \frac{(r^{\mathrm{sb}})^2}{4}\qquad \text{if } \alpha \in \left[\frac{1}{2}, 1\right].$$

Under the hypothesis that all the vertices of $\mathfrak{P}$ lie on the boundary of $\mathfrak{B}^{\mathrm{sb}}$, we know from Proposition 3.5 that

$$c^{\mathrm{v}}_{\infty}\ =\ \frac{(r^{\mathrm{sb}})^2}{2}.$$

Hence

$$c(\alpha, y)\ \ge\ \frac{(r^{\mathrm{sb}})^2}{4}\qquad (\alpha \in [0,1],\ y \in \mathfrak{P}),$$

which shows that equality occurs in (3.20).

Finally, we have to verify the statement on the occurrence of equality for functions f from the class $\mathcal{F}_2$. For this, it is clearly enough to consider the function $f := \|\cdot\|^2$ only.

Using the notation (2.8), we may rewrite (2.1) and (2.2) as

$$\begin{aligned} f(x) - \Lambda^{\mathrm{sb}}[f](x) &= e^{\mathrm{sb}}(x), \\ f(x) - \Lambda^{\mathrm{v}}[f](x) &= -e^{\mathrm{v}}(x). \end{aligned}$$

Therefore,

$$f(x) - \frac{1}{2}\,\Lambda^{\mathrm{sb}}[f](x) - \frac{1}{2}\,\Lambda^{\mathrm{v}}[f](x) \;=\; \frac{1}{2}\left(e^{\mathrm{sb}}(x) - e^{\mathrm{v}}(x)\right)$$

and consequently,

$$\left\| f - \frac{1}{2}\,\Lambda^{\mathrm{sb}}[f] - \frac{1}{2}\,\Lambda^{\mathrm{v}}[f]\right\|_\infty \;=\; \frac{1}{2}\,\sup_{x\in\mathfrak{P}} \left|e^{\mathrm{sb}}(x) - e^{\mathrm{v}}(x)\right|.$$

If all the vertices of $\mathfrak{P}$ lie on the boundary of $\mathfrak{B}^{\mathrm{sb}}$, then

$$\sup_{x\in\mathfrak{P}} \left|e^{\mathrm{sb}}(x) - e^{\mathrm{v}}(x)\right| \;\geq\; \left|e^{\mathrm{sb}}(x^{\mathrm{sb}}) - e^{\mathrm{v}}(x^{\mathrm{sb}})\right| \;=\; e^{\mathrm{v}}(x^{\mathrm{sb}}) \;=\; (r^{\mathrm{sb}})^2,$$

where the last equation follows from (2.4) and (1.2), and so

$$\left\| f - \frac{1}{2}\,\Lambda^{\mathrm{sb}}[f] - \frac{1}{2}\,\Lambda^{\mathrm{v}}[f]\right\|_\infty \;\geq\; \frac{(r^{\mathrm{sb}})^2}{2}.$$

On the other hand, (3.18) and (3.20) show that

$$\left\| f - \frac{1}{2}\,\Lambda^{\mathrm{sb}}[f] - \frac{1}{2}\,\Lambda^{\mathrm{v}}[f]\right\|_\infty \;\leq\; \frac{(r^{\mathrm{sb}})^2}{2}.$$

Hence equality occurs for $f = \|\cdot\|^2$. □

4. Approximation of linear functionals. In the case $p = 1$, Theorems 3.1–3.3 provide an approximation of $\mathcal{L}(f)$, defined in (1.4), by the values of f (and possibly of Df) at the interpolation points of Λ_y and Λ^{v}, respectively. Indeed, if Λ is any of the two operators Λ_y and Λ^{v}, and $I(f) := \mathcal{L}(\Lambda[f])$, then, using that $\mathcal{L}$ is linear and positive, we have

$$\left|\mathcal{L}(f) - I(f)\right| \;=\; \left|\mathcal{L}(f - \Lambda[f])\right| \;\leq\; \mathcal{L}\big(\,|f - \Lambda[f]|\big) \;=\; \left\|f - \Lambda[f]\right\|_1.$$

Let us now turn to details. Denoting by id the identity mapping on $\mathfrak{P}$ and observing that $\mathcal{L}(\mathrm{id})$ is a mapping from $\mathfrak{P}$ into $\mathbb{R}^d$, we shall consider the operators

$$I_y(f) \;:=\; \mathcal{L}(\Lambda_y[f]) \;=\; \mathcal{L}(1)\left[f(y) + Df(y)\left(\frac{\mathcal{L}(\mathrm{id})}{\mathcal{L}(1)} - y\right)\right] \tag{4.1}$$

and

$$I^{\mathrm{v}}(f) \;:=\; \mathcal{L}(\Lambda^{\mathrm{v}}[f]) \;=\; \sum_{i=1}^{n} f(v_i)\mathcal{L}(B_i). \tag{4.2}$$

In the case $p = 1$, the constants (3.8) and (3.11) can be expressed as

$$c_{y,1} \;=\; \frac{1}{2}\,\mathcal{L}(e_y) \qquad \text{and} \qquad c_1^{\mathrm{v}} \;=\; \frac{1}{2}\left(I^{\mathrm{v}}(e_y) - \mathcal{L}(e_y)\right). \tag{4.3}$$

Note that the last equation, which is deduced with the help of (2.5), is independent of y. Now Theorems 3.2 and 3.3 imply the following corollaries.

COROLLARY 4.1. *Let $f \in C^2(\mathfrak{P})$. Then, for any $y \in \mathfrak{P}$, we have*

$$|\mathcal{L}(f) - I_y(f)| \;\le\; \frac{\mathcal{L}(e_y)}{2}\;\big|D^2 f\big|.$$

Equality is attained for every $f \in \mathcal{F}_2$.

COROLLARY 4.2. *Let $f \in C^2(\mathfrak{P})$. Then, for any $y \in \mathfrak{P}$, we have*

$$|\mathcal{L}(f) - I^{\mathrm{v}}(f)| \;\le\; \frac{I^{\mathrm{v}}(e_y) - \mathcal{L}(e_y)}{2}\;\big|D^2 f\big|.$$

Equality is attained for every $f \in \mathcal{F}_2$.

Remark 4.3. The conclusions of Corollaries 4.1 and 4.2 can be refined when, in addition, f is known to be a convex function. In fact, in this case, we also have

$$I_y(f) \;\le\; \mathcal{L}(f) \;\le\; I^{\mathrm{v}}(f)$$

as a consequence of the statements (iii) of Lemmas 2.1 and 2.2.

The "cubature rule" $I^{\mathrm{v}}(f)$ may be seen as a multivariate analogue of the trapezoidal rule. As (4.1) shows, the "cubature rule" $I_y(f)$ simplifies and does not depend on Df if y is chosen as

$$x^{\mathrm{cm}} \;:=\; \frac{\mathcal{L}(\mathrm{id})}{\mathcal{L}(1)}.$$

In this case, $I_y(f)$ is a multivariate analogue of the midpoint rule.

The point x^{cm} will be called the *center of mass* of $\mathfrak{P}$ with respect to the functional $\mathcal{L}$. Note that x^{cm} always belongs to $\mathfrak{P}$. Indeed, if x^{cm} were outside $\mathfrak{P}$, then there would exist a separating hyperplane

$$\lambda(x) \;:=\; a + \langle b, x\rangle \;=\; 0,$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}^d$, such that $\lambda(x) > 0$ for $x \in \mathfrak{P}$ and $\lambda(x^{\mathrm{cm}}) < 0$. Since $\mathcal{L}$ is positive, we would have $\mathcal{L}(\lambda) > 0$. On the other hand, the linearity of $\mathcal{L}$ implies that

$$\mathcal{L}(\lambda) = a\mathcal{L}(1) + \langle b, \mathcal{L}(\mathrm{id})\rangle = a\mathcal{L}(1) + \langle b, \mathcal{L}(1)x^{\mathrm{cm}}\rangle = \mathcal{L}(1)\lambda(x^{\mathrm{cm}}) < 0,$$

which is a contradiction.

For notational simplicity, we now write

$$\Lambda^{\mathrm{cm}} := \Lambda_y, \quad I^{\mathrm{cm}} := I_y, \quad e^{\mathrm{cm}} := e_y, \quad c_p^{\mathrm{cm}} := c_{y,p} \qquad \text{if } y = x^{\mathrm{cm}}. \tag{4.4}$$

Since

$$e_y(x) = \|x - y\|^2 = \|x - x^{\mathrm{cm}}\|^2 + \|x^{\mathrm{cm}} - y\|^2 + 2\langle x - x^{\mathrm{cm}},\, x^{\mathrm{cm}} - y\rangle,$$

we find, using the definition of x^{cm}, that

$$c_{y,1} \;=\; \mathcal{L}(e_y) \;=\; \mathcal{L}(e^{\mathrm{cm}}) + \mathcal{L}(1)\|x^{\mathrm{cm}} - y\|^2.$$

This shows that the constant in the error estimate of Corollary 4.1 becomes smallest if and only if $y = x^{\mathrm{cm}}$.

Remark 4.4. It may be interesting to compare the operators I^{cm} and I^{v}. Recalling that c_1^{v} in (4.3) does not depend on y, we may take $y = x^{\mathrm{cm}}$. Then Corollaries 4.1 and 4.2 show that the quotient

$$\kappa := \frac{\mathcal{L}(e^{\mathrm{cm}})}{I^{\mathrm{v}}(e^{\mathrm{cm}})} \tag{4.5}$$

indicates which one of the two operators I^{cm} and I^{v} has the smaller constant in its error estimate. We see that $c_1^{\mathrm{cm}} < c_1^{\mathrm{v}}$ if and only if $\kappa \in (0, 1/2)$. Since, for convex functions, I^{v} approximates $\mathcal{L}$ from above, we always have $\kappa \in (0, 1)$. In all the standard examples considered by us, we found that $\kappa \in (0, 1/2)$. However, $\kappa \in [1/2, 1)$ will occur when $\mathcal{L}$ is of the form (1.4) and the weight function w is large near the vertices.

5. Examples. We illustrate our results by considering three special classes of convex polytopes for which interpolation and approximation problems have been studied in the literature.

5.1. Intervals (the univariate case). Let $d := 1$, $\mathfrak{P} := [a, b]$, and $\mathcal{L}(f) := \int_a^b f(x)\,\mathrm{d}x$. Then $x^{\mathrm{sb}} = x^{\mathrm{cm}} = \frac{1}{2}(a + b)$,

$$\Lambda^{\mathrm{cm}}[f](x) = f\left(\frac{a+b}{2}\right) + f'\left(\frac{a+b}{2}\right)\left(x - \frac{a+b}{2}\right),$$

and

$$\Lambda^{\mathrm{v}}[f](x) = \frac{b-x}{b-a}\,f(a) + \frac{x-a}{b-a}\,f(b).$$

Moreover, $\left|D^2 f\right| = \sup_{a \le x \le b} |f''(x)|$. For the constants (3.8) with $y = x^{\mathrm{cm}}$ and (3.11), we find that

$$c_p^{\mathrm{cm}} = \frac{1}{2}\left[\frac{(b-a)^{2p+1}}{2^{2p}(2p+1)}\right]^{1/p} \qquad (1 \le p < \infty)$$

and

$$c_p^{\mathrm{v}} = \frac{1}{2}\left[B(p+1, p+1)(b-a)^{2p+1}\right]^{1/p} \qquad (1 \le p < \infty),$$

where

$$B(s, t) := \int_0^1 x^{s-1}(1-x)^{t-1}\,\mathrm{d}x$$

is the beta function. Furthermore,

$$c_\infty^{\mathrm{cm}} = c_\infty^{\mathrm{v}} = \frac{(b-a)^2}{8}.$$

It can be shown that $c_p^{\mathrm{cm}} < c_p^{\mathrm{v}}$ for $1 \le p < \infty$. In particular, $c_1^{\mathrm{v}}/c_1^{\mathrm{cm}} = 2$, which expresses the well-known fact that the constant in the error term of the trapezoidal rule is twice as large as that of the midpoint rule.

5.2. Multidimensional simplices. Let $\mathfrak{S} \subset \mathbb{R}^d$ be a nondegenerate simplex with vertices $v_0, \dots, v_d$. The uniquely determined rational basis functions $B_0, \dots, B_d$ of minimal degree are the classical barycentric coordinates, which may be constructed as follows. Let $\lambda_i(x) = 0$ be the equation of a hyperplane that contains all the vertices of $\mathfrak{S}$ other than v_i. Then

$$B_i(x) \;=\; \frac{\lambda_i(x)}{\lambda_i(v_i)} \qquad (i = 0, \dots, d).$$

For $\mathcal{L}(f) := \int_{\mathfrak{S}} f(x)\,\mathrm{d}x$, we obtain

$$x^{\mathrm{cm}} \;=\; \frac{1}{|\mathfrak{S}|} \int_{\mathfrak{S}} x \,\mathrm{d}x \;=\; \frac{1}{d+1} \sum_{i=0}^{d} v_i,$$

where we write $|\mathfrak{S}|$ for the d-dimensional volume of $\mathfrak{S}$. This gives a representation of e^{cm} in terms of the vertices, which, via (4.4) and (3.17), leads us to

$$c_\infty^{\mathrm{cm}} \;=\; \frac{1}{2(d+1)^2} \max_{0 \le i \le d} \left\| \sum_{j=0}^{d} (v_i - v_j) \right\|^2 .$$

Since the basis functions B_i belong to $\mathcal{P}_1$, the function e^{v}, defined in (2.2), is now of the form $e^{\mathrm{v}} = \lambda - \|\cdot\|^2$, where $\lambda \in \mathcal{P}_1$. Therefore $e^{\mathrm{v}}(x) = 0$ is the equation of the uniquely defined sphere that contains all the vertices of $\mathfrak{S}$ (see, e.g., Stämpfle [6, Proposition 3.1]). Thus e^{v} can be represented as

$$e^{\mathrm{v}}(x) \;=\; \widehat{r}^2 - \|x - \widehat{x}\|^2$$

for some $\widehat{r} > 0$ and $\widehat{x} \in \mathbb{R}^d$.

The case of the approximation by Λ^{v} with respect to the norm $\|\cdot\|_\infty$ is covered by the papers of Waldron [8, Theorem 2.1] and Stämpfle [6, Theorem 4.1]; also see de Boor [1]. Clearly, $c_\infty^{\mathrm{v}} = \widehat{r}^2/2$ when $\widehat{x} \in \mathfrak{S}$. Otherwise, it can be shown that $c_\infty^{\mathrm{v}} = \frac{1}{2}(\widehat{r}^2 - \rho^2)$, where ρ is the distance of $\widehat{x}$ from $\mathfrak{S}$. Geometrically, $2c_\infty^{\mathrm{v}}$ may be interpreted as the square of the radius of the smallest ball that contains $\mathfrak{S}$ (see [6, Lemma 4.2]).

For the standard unit simplex of dimension $d \ge 2$, a straightforward calculation gives

$$c_\infty^{\mathrm{cm}} \;=\; \frac{d^2 + d - 1}{2(d+1)^2} \qquad \text{and} \qquad c_\infty^{\mathrm{v}} \;=\; \frac{d-1}{2d},$$

and so

$$\frac{c_\infty^{\mathrm{v}}}{c_\infty^{\mathrm{cm}}} \;=\; 1 - \frac{1}{d(d^2 + d - 1)} \;<\; 1.$$

For the calculation of the constants (4.3) for $y = x^{\mathrm{cm}}$, we first determine $\mathcal{L}(e^{\mathrm{cm}})$ with the help of a cubature rule which is exact for all polynomials of degree less than or equal to 2, taken from the book of Stroud [7, formula $T_n : 2$-2, p. 307]. This gives

$$\begin{aligned} \mathcal{L}(e^{\mathrm{cm}}) = \int_{\mathfrak{S}} e^{\mathrm{cm}}(x)\,\mathrm{d}x = \frac{(2-d)\,|\mathfrak{S}|}{(d+2)(d+1)} & \sum_{i=0}^{d} e^{\mathrm{cm}}(v_i) \\ + \frac{4\,|\mathfrak{S}|}{(d+2)(d+1)} & \sum_{0 \le i < j \le d} e^{\mathrm{cm}}(v_{ij}), \end{aligned}$$

where $v_{ij} = \frac{1}{2}(v_i + v_j)$. Simplifying the second sum by making use of the special form of e^{cm}, we arrive at

$$\mathcal{L}(e^{\mathrm{cm}}) = \frac{|\mathfrak{S}|}{(d+2)(d+1)} \sum_{i=0}^{d} e^{\mathrm{cm}}(v_i).$$

Since the basis functions $B_0, \ldots, B_d$ belong to $\mathcal{P}_1$, we conclude that

$$\mathcal{L}(B_i) \;=\; \mathcal{L}(1)B_i(x^{\mathrm{cm}}) \;=\; \frac{\mathcal{L}(1)}{d+1} \;=\; \frac{|\mathfrak{S}|}{d+1} \qquad (i = 0, \ldots, d)$$

and therefore

$$I^{\mathrm{v}}(e^{\mathrm{cm}}) = \frac{|\mathfrak{S}|}{d+1} \sum_{i=0}^{d} e^{\mathrm{cm}}(v_i).$$

Thus, by (4.3), the definition of e^{cm} in (4.4), and the representation in (2.3), we have

$$c_1^{\mathrm{cm}} \;=\; \frac{|\mathfrak{S}|}{2(d+2)(d+1)} \sum_{i=0}^{d} \|v_i - x^{\mathrm{cm}}\|^2$$

and

$$c_1^{\mathrm{v}} \;=\; \frac{|\mathfrak{S}|}{2(d+2)} \sum_{i=0}^{d} \|v_i - x^{\mathrm{cm}}\|^2.$$

These values for c_1^{cm} and c_1^{v} also follow from [3, Corollary 6.2, formulae (6.4) and (6.5)]. We see that $c_1^{\mathrm{v}} = (d+1)c_1^{\mathrm{cm}}$ and $\kappa = 1/(d+2)$ in (4.5).

5.3. Hyperrectangles. Let

$$\mathfrak{R} \;:=\; [a_1, b_1] \times \cdots \times [a_d, b_d]$$

be a rectangle in $\mathbb{R}^d$ with vertices

$$v_i \;:=\; (v_{i1}, \ldots, v_{id}) \qquad (i = 1, \ldots, 2^d).$$

To each vertex v_i, there corresponds exactly one vertex of maximal distance, which we call the *diametrically opposite* vertex and which we denote by $\overline{v}_i := (\overline{v}_{i1}, \ldots, \overline{v}_{id})$. Any two vertices v_i and v_j have at least one common component unless they are a pair of diametrically opposite vertices. Therefore

$$B_i(x) \;:=\; \prod_{j=1}^{d} \frac{x_j - \overline{v}_{ij}}{v_{ij} - \overline{v}_{ij}} \qquad (i = 1, \ldots, 2^d),$$

where $x = (x_1, \ldots, x_d)$, are the rational basis functions of smallest degree, spanning a polynomial space of dimension 2^d, which contains $\mathcal{P}_1$ as a subspace.

For $\mathcal{L}(f) := \int_{\mathfrak{R}} f(x)\,\mathrm{d}x$, the center of mass is

$$x^{\mathrm{cm}} \;=\; \frac{1}{2}(a_1 + b_1, \ldots, a_d + b_d).$$

With this, we find that

$$e^{\mathrm{cm}}(v_i) \;=\; \frac{1}{4}\sum_{i=1}^{d}(a_i-b_i)^2 \;=:\; (r^{\mathrm{cm}})^2 \qquad (i=1,\dots,2^d).$$

Therefore (2.5) implies that

$$e^{\mathrm{cm}}(x)+e^{\mathrm{v}}(x) \;=\; (r^{\mathrm{cm}})^2$$

for all x. Consequently,

$$\sup_{x\in\mathfrak{R}} e^{\mathrm{cm}}(x) \;=\; \sup_{x\in\mathfrak{R}} e^{\mathrm{v}}(x) \;=\; (r^{\mathrm{cm}})^2,$$

or equivalently,

$$c_\infty^{\mathrm{cm}} \;=\; c_\infty^{\mathrm{v}} \;=\; \frac{(r^{\mathrm{cm}})^2}{2}.$$

For determining the best constants in the case $p=1$, we first calculate

$$\mathcal{L}(e^{\mathrm{cm}}) \;=\; \int_{\mathfrak{R}} \|x-x^{\mathrm{cm}}\|^2\,\mathrm{d}x \;=\; \frac{(r^{\mathrm{cm}})^2}{3}\,|\mathfrak{R}|\,,$$

where $|\mathfrak{R}| = \prod_{i=1}^{d}(b_i-a_i)$, and note that

$$I^{\mathrm{v}}(e^{\mathrm{cm}}) \;=\; (r^{\mathrm{cm}})^2\,|\mathfrak{R}|\,.$$

Hence (4.3) with $y=x^{\mathrm{cm}}$ implies that

$$c_1^{\mathrm{cm}} = \frac{(r^{\mathrm{cm}})^2}{6}\,|\mathfrak{R}| \qquad \text{and} \qquad c_1^{\mathrm{v}} = \frac{(r^{\mathrm{cm}})^2}{3}\,|\mathfrak{R}|\,.$$

Thus, $c_1^{\mathrm{v}}/c_1^{\mathrm{cm}} = 2$ and $\kappa = 1/3$, as in the univariate case.

In the literature, analogues of the trapezoidal rule for hyperrectangles have been studied in the context of tensor product rules (see, e.g., [2, section 8.2]).

REFERENCES

[1] C. de Boor, *Error in Linear Interpolation at the Vertices of a Simplex*, annotated bibliography, 1998, available online at http://www.cs.wisc.edu/∼deboor/multiint/.

[2] F.-J. Delvos and W. Schempp, *Boolean Methods in Interpolation and Approximation,* Longman Scientific & Technical, Essex, UK, 1989.

[3] A. Guessab and G. Schmeisser, *Convexity results and sharp error estimates in approximate multivariate integration*, Math. Comp., 73 (2004), pp. 1365–1384.

[4] G. Hadley, *Nonlinear and Dynamic Programming,* Addison–Wesley, Reading, MA, 1964.

[5] A. W. Roberts and D. E. Varberg, *Convex Sets,* Academic Press, New York, 1973.

[6] M. Stämpfle, *Optimal estimates for the linear interpolation error on simplices,* J. Approx. Theory, 103 (2000), pp. 78–90.

[7] A. H. Stroud, *Approximate Calculation of Multiple Integrals,* Prentice–Hall, Englewood Cliffs, NJ, 1971.

[8] S. Waldron, *The error in linear interpolation at the vertices of a simplex,* SIAM J. Numer. Anal., 35 (1998), pp. 1191–1200.

[9] S. Waldron, *Sharp error estimates for multivariate positive linear operators,* in Approximation Theory IX (Proceedings of the 9th International Conference, Nashville, TN, 1998), C. K. Chui and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, TN, 1998, Vol. I, pp. 339–346.

[10] J. Warren, *Barycentric coordinates for convex polytopes,* Adv. Comput. Math., 6 (1996), pp. 97–108.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 924–948

REPRESENTATIONS OF RUNGE–KUTTA METHODS AND STRONG STABILITY PRESERVING METHODS*

INMACULADA HIGUERAS[†]

Abstract. Over the last few years a great effort has been made to develop monotone high order explicit Runge–Kutta methods by means of their Shu–Osher representations. In this context, the stepsize restriction to obtain numerical monotonicity is normally computed using the optimal representation. In this paper we extend the Shu–Osher representations for any Runge–Kutta method giving sufficient conditions for monotonicity. We show how optimal Shu–Osher representations can be constructed from the Butcher tableau of a Runge–Kutta method.

The optimum stepsize restriction for monotonicity is given by the radius of absolute monotonicity of the Runge–Kutta method [L. Ferracina and M. N. Spijker, *SIAM J. Numer. Anal.*, 42 (2004), pp. 1073–1093], and hence if this radius is zero, the method is not monotone. In the Shu–Osher representation, methods with zero radius require negative coefficients, and to deal with them, an extra associate problem is considered. In this paper we interpret these schemes as representations of perturbed Runge–Kutta methods. We extend the concept of radius of absolute monotonicity and give sufficient conditions for monotonicity. Optimal representations can be constructed from the Butcher tableau of a perturbed Runge–Kutta method.

Key words. Runge–Kutta, strong stability preserving, SSP, absolutely monotonic, radius of absolute monotonicity, CFL coefficient, representations

AMS subject classifications. 65L06, 65L05, 65M20

DOI. 10.1137/S0036142903427068

1. Introduction.

We consider initial value problems for ordinary differential systems (ODEs) of the form

$$\begin{aligned} \frac{d}{dt}u(t) &= f(u(t)), \qquad t \geq t_0, \\ u(t_0) &= u_0. \end{aligned} \tag{1.1}$$

We assume that $t_0 \in \mathbb{R}$, $u_0 \in \mathbb{R}^m$, and f is a function from $\mathbb{R}^m$ to $\mathbb{R}^m$ such that for each $t_0 \in \mathbb{R}$ and $u_0 \in \mathbb{R}^m$ the problem (1.1) has a unique solution $u : [t_0, \infty) \to \mathbb{R}^m$. We assume too that $\| \cdot \| : \mathbb{R}^m \to \mathbb{R}$ is a convex functional (e.g., a norm, a seminorm, etc.) such that for any $t_0 \in \mathbb{R}$ and any solution $u(t)$ to (1.1) we have

$$\|u(t)\| \leq \|u(t_0)\| \qquad \forall\, t \geq t_0. \tag{1.2}$$

On $(f, \| \cdot \|)$ we impose the condition

$$\left\| y + \frac{1}{\rho} f(y) \right\| \leq \|y\| \qquad \forall\, y \in \mathbb{R}^m \tag{1.3}$$

for some fixed $\rho > 0$, and we denote this class of problems by $\mathcal{F}(\rho)$. From [10, Lemma II.5.1], we get that if (1.3) holds, condition

$$\|y + \tau f(y)\| \leq \|y\|, \qquad 0 \leq \tau \leq \frac{1}{\rho}, \tag{1.4}$$

*Received by the editors May 1, 2003; accepted for publication (in revised form) January 6, 2005; published electronically September 20, 2005. This research was supported by the Ministerio de Ciencia y Tecnología, Project BFM2001-2188.

http://www.siam.org/journals/sinum/43-3/42706.html

[†]Departamento de Matemática e Informática, Universidad Pública de Navarra 31006, Pamplona, Navarra, Spain (higueras@unavarra.es).

also holds. Denoting by D_+ the right-hand derivative, for $u(t)$, the solution of (1.1), we obtain [1, Lemma 1.5.3]

$$D_+\|u(t)\| = \lim_{\tau\to 0^+} \frac{\|u(t)+\tau u'(t)\| - \|u(t)\|}{\tau} = \lim_{\tau\to 0^+} \frac{\|u(t)+\tau f(u(t))\| - \|u(t)\|}{\tau} \le 0.$$

Hence assumption (1.3) with $\rho > 0$ gives the monotonicity inequality (1.2) for the solution of the ODE (1.1). Later on, we will also use the inequality

$$\|y\| \le \|y + \tau f(y)\| \qquad \forall\, \tau < 0,\ \forall\, y \in \mathbb{R}^m. \tag{1.5}$$

From [10, Lemma II.5.1], it can be concluded that condition (1.3) implies condition (1.5). In the rest of the paper we assume that f in (1.1) satisfies $(f, \|\cdot\|) \in \mathcal{F}(\rho)$.

Given the initial value problem (1.1), a common class of one-step methods to solve it numerically is the Runge–Kutta (RK) methods. An s-stage RK method is defined by an $s \times s$ real matrix $\mathcal{A}$ and a real vector $b \in \mathbb{R}^s$; we will refer to it as $(\mathcal{A}, b)$. From u_n, the numerical approximation of the solution $u(t)$ at $t = t_n$, we obtain u_{n+1}, the numerical approximation of the solution at $t_{n+1} = t_n + h$ from

$$u_{n+1} = u_n + h \sum_{i=1}^{s} b_i f(U_i), \tag{1.6}$$

$$U_i = u_n + h \sum_{j=1}^{s} a_{ij} f(U_j). \tag{1.7}$$

The internal stage U_i approximates $u(t_n + c_i h)$, where $c_i = \sum_{j=1}^{s} a_{ij}$. Furthermore, for many methods $c_i \ge 0$, $i = 1, \dots, s$, and thus $t_n + c_i h \ge t_n$.

If we solve numerically an ODE (1.1) with $(f, \|\cdot\|) \in \mathcal{F}(\rho)$ with an RK method, a natural requirement for the internal stages and the numerical solution is

$$\|U_i\| \le \|u_n\|, \quad i = 1, \dots, s, \quad \|u_{n+1}\| \le \|u_n\|, \tag{1.8}$$

for all $n \ge 0$, probably under a stepsize restriction $h \le \Delta t_{MAX}$.

1.1. SSP methods. Over the last few years ([13], [14], [5], [11], [16], [6]; see [4], [15] for reviews on this topic) a great effort has been done to develop high order methods satisfying (1.8) when the forward Euler discretization of (1.1) satisfies (1.8),

$$\|u_n + h f(u_n)\| \le \|u_n\| \qquad \text{for } h \le \Delta t_{FE}. \tag{1.9}$$

These methods are called strong stability preserving methods (SSP methods). The class of ODEs considered in this context arise from a method-of-lines approximation of hyperbolic conservation laws. A simple numerical example given in [4] shows that the use of non-SSP methods for the time discretization of these ODEs has the potential to produce an undesirable overshoot.

As the forward Euler method has the drawback of a low order of accuracy, higher order SSP methods are of great interest. The idea in [14], [13] is to derive conditions (1.8) from condition (1.9) for the forward Euler method by means of convex combinations of it. In [13], Shu and Osher write explicit s-stage RK methods as

$$\begin{aligned} u^{(1)} &= u_n, \\ u^{(i)} &= \sum_{k=1}^{i-1} (\alpha_{ik} u^{(k)} + h\, \beta_{ik}\, f(u^{(k)})), \qquad i = 2, \dots, s+1, \\ u_{n+1} &= u^{(s+1)}, \end{aligned} \tag{1.10}$$

where $\alpha_{ik} \geq 0$ for all i, j, and $\sum_{k=1}^{i-1} \alpha_{ik} = 1$, $i = 2, \ldots, s+1$. It is also imposed that

$$\beta_{i,j} = 0 \qquad \text{whenever} \qquad \alpha_{ij} = 0. \tag{1.11}$$

Proceeding in this way, the new method will also be strongly stable, perhaps with a modified stepsize restriction

$$h \leq c\,\Delta t_{FE}. \tag{1.12}$$

The coefficient c in (1.12) is known as a CFL coefficient. Convex combinations of the forward Euler method are obtained in (1.10) if $\beta_{ij} \geq 0$. In this case, the CFL coefficient is given by

$$c = \min_{ik} \frac{\alpha_{ik}}{\beta_{ik}}. \tag{1.13}$$

The representation of RK methods in the form (1.10) is not unique [16]. Different representations give rise to different values for the CFL coefficient (1.13).

If some $\beta_{ij} < 0$, then f is replaced by an associated operator $\tilde{f}$ corresponding to stepping backward in time. The requirement for $\tilde{f}$ is that it approximate the same spatial derivatives as f, and that (1.8) hold with the same stepsize restriction for the explicit Euler scheme solved backwards, $u_{n+1} = u_n - h\tilde{f}(u_n)$; i.e.,

$$\|u_n - h\tilde{f}(u_n)\| \leq \|u_n\| \qquad \text{for } h \leq \Delta t_{FE}. \tag{1.14}$$

In this case [13], the RK method is SSP under the stepsize restriction (1.12) with

$$c = \min_{ik} \frac{\alpha_{ik}}{|\beta_{ik}|}. \tag{1.15}$$

For example, the classical fourth order four-stage RK method can be written [13] as

$$\begin{aligned}
u^{(1)} &= u^{(0)},\\
u^{(2)} &= u^{(1)} + \frac{1}{2}hf(u^{(1)}),\\
u^{(3)} &= \frac{1}{2}u^{(1)} - \frac{1}{4}h\tilde{f}(u^{(1)}) + \frac{1}{2}u^{(2)} + \frac{1}{2}hf(u^{(2)}),\\
u^{(4)} &= \frac{1}{9}u^{(1)} - \frac{1}{9}h\tilde{f}(u^{(1)}) + \frac{2}{9}u^{(2)} - \frac{1}{3}h\tilde{f}(u^{(2)}) + \frac{2}{3}u^{(3)} + hf(u^{(3)}),\\
u^{(5)} &= \frac{1}{3}u^{(2)} + \frac{1}{6}hf(u^{(2)}) + \frac{1}{3}u^{(3)} + \frac{1}{3}u^{(4)} + \frac{1}{6}hf(u^{(4)}).
\end{aligned} \tag{1.16}$$

The CFL coefficient given by (1.15) is $c = 2/3$. A better definition of $u^{(4)}$ allows us to raise the CFL coefficient to $c = 137/200$ [13].

A particular class of implicit RK methods is also considered in [4, section 6.2]:

$$\begin{aligned}
u^{(1)} &= u_n,\\
u^{(i)} &= \sum_{k=1}^{i-1} \alpha_{ik}u^{(k)} + h\,\beta_i\,f(u^{(i)}), \qquad i = 2, \ldots, s+1,\\
u_{n+1} &= u^{(s+1)},
\end{aligned} \tag{1.17}$$

with $\alpha_{ik} \geq 0$ and $\sum_{k=1}^{i-1} \alpha_{ik} = 1$. In this case, monotonicity, or more precisely unconditional monotonicity, is assumed on the implicit Euler method; i.e., for

$$u_{n+1} = u_n + h\, f(u_{n+1}),$$

it holds that $\|u_{n+1}\| \leq \|u_n\|$ for any stepsize h; i.e.,

$$\|u_{n+1}\| \leq \|u_{n+1} - h\, f(u_{n+1})\| \qquad \forall\, h > 0. \tag{1.18}$$

At this point we would like to establish a link between the class of problems considered in this paper and SSP methods. If we compare the condition (1.3) (see also (1.4)) imposed on the problem, with the condition (1.9) assumed in the SSP context for the explicit Euler method, we get that $\Delta t_{FE} = 1/\rho$; i.e., the stepsize restriction for the forward Euler method is determined by the class of problem considered. Furthermore, condition (1.18) imposed for implicit problems in the SSP context is precisely (1.5). Remember that (1.3) implies (1.5).

Therefore the class of problems considered in the SSP context is also $(f, \|\cdot\|) \in \mathcal{F}(\rho)$. For these problems, the explicit Euler method is monotone under stepsize restriction $h \leq 1/\rho$, whereas the implicit Euler method is unconditionally monotone.

1.2. Radius of absolute monotonicity. In the context of contractive RK methods, the concept of radius of absolute monotonicity plays an important role [7]. Recently, Ferracina and Spijker [2] have proved that this concept is also relevant for monotone methods. We remember the definitions of absolute monotonicity and radius of absolute monotonicity.

Definition 1.1 (see [7, Definition 2.4]). *An s-stage RK method with coefficients $(\mathcal{A}, b)$ is said to be absolutely monotonic at a given point $\xi \leq 0$ if $I - \xi\mathcal{A}$ is nonsingular, the stability function $\phi(\xi) = 1 + \xi\, b^t(I - \xi\mathcal{A})^{-1}e \geq 0$, $\mathcal{A}(\xi) = \mathcal{A}(I - \xi\mathcal{A})^{-1} \geq 0$, $b(\xi)^t = b^t(I - \xi\mathcal{A})^{-1} \geq 0$, and $e(\xi) = (I - \xi\mathcal{A})^{-1}e \geq 0$, where $e = (1, 1, \ldots, 1) \in \mathbb{R}^s$, and the vector inequalities are understood componentwise. Further, the method is said to be absolutely monotonic on a given set $\Omega \subset \mathbb{R}$ if it is absolutely monotonic at each $\xi \in \Omega$. The radius of absolute monotonicity $R(\mathcal{A}, b)$ is defined by*

$$R(\mathcal{A}, b) = \sup\{r \mid r \geq 0 \text{ and } (\mathcal{A}, b) \text{ is absolutely monotonic on } [-r, 0]\}. \tag{1.19}$$

If there is no $r > 0$ such that $(\mathcal{A}, b)$ is absolutely monotonic on $[-r, 0]$, we set $R(\mathcal{A}, b) = 0$.

Observe that the four conditions on $\phi(\xi)$, $\mathcal{A}(\xi)$, $b(\xi)$, and $e(\xi)$ in Definition 1.1 can be written in compact form as

$$\begin{pmatrix} (I - \xi\mathcal{A})^{-1} & 0 \\ \xi\, b^t(I - \xi\mathcal{A})^{-1} & 1 \end{pmatrix} \begin{pmatrix} e \\ 1 \end{pmatrix} \geq 0, \qquad \begin{pmatrix} (I - \xi\mathcal{A})^{-1} & 0 \\ \xi\, b^t(I - \xi\mathcal{A})^{-1} & 1 \end{pmatrix} \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix} \geq 0. \tag{1.20}$$

Denoting the coefficients of an RK method by

$$\mathbb{A} = \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 0 \end{pmatrix},$$

we can write (1.20) as

$$(I - \xi\mathbb{A})^{-1}\mathbb{A} \geq 0, \qquad (I - \xi\mathbb{A})^{-1}e \geq 0, \tag{1.21}$$

where now $e = (1, 1, \ldots, 1)^t \in \mathbb{R}^{s+1}$. In the following we will denote $R(\mathcal{A}, b)$ in (1.19) by $R(\mathbb{A})$.

1.3. Scope of the paper. Recently, Ferracina and Spijker [2] have established a link between the radius of absolute stability and the stepsize restriction (1.12). More precisely, in [2] it is proved that if (1.9) holds, irreducible RK methods are monotone with stepsize restriction $h \leq c\,\Delta t_{FE}$ if and only if $c \leq R(\mathbb{A})$. In other words, the optimal CFL coefficient in (1.12) is $R(\mathbb{A})$. As the CFL coefficient in (1.13) is normally obtained using a numerical optimal representation, a natural question that arises is whether, given the Butcher tableau of an explicit RK method, there exists a representation such that the CFL coefficient (1.13) is the maximum one $R(\mathbb{A})$. On the other hand, representations (1.10) and (1.17) are given for explicit methods and a class of implicit methods, respectively, whereas $R(\mathbb{A})$ is defined for general RK methods. This suggests that more general representations can be considered, and from them a notion of CFL coefficient like the one in (1.13) can be computed. In this paper we study these issues. The rest of the paper is organized as follows.

In section 2 we define a kind of representation of the RK methods that contains schemes (1.10) and (1.17) as particular cases, and we extend the CFL coefficient (1.13). We also prove that if $R(\mathbb{A}) > 0$, it is possible to obtain an optimal representation such that the CFL coefficient computed from it is equal to $R(\mathbb{A})$.

Section 3 is devoted to RK methods $\mathbb{A}$ with $R(\mathbb{A}) = 0$. In the Shu–Osher representation, methods with negative coefficients β_{ij} imply that $R(\mathbb{A}) = 0$, and to deal with them, an extra associate problem is considered. In this paper we interpret these schemes as representations of perturbed RK methods. An extension of the CFL coefficient (1.15) is defined. We also extend the concept of radius of absolute monotonicity and give sufficient conditions for monotonicity. Optimal representations can be constructed from the Butcher tableau of a perturbed RK method.

The paper ends with some conclusions and some open questions for future work.

2. Representations of an RK method. We consider methods of the form

$$U = \alpha \otimes u_n + (\Lambda \otimes I)U + h(\Gamma \otimes I)F(U), \tag{2.1}$$

where $\alpha \in \mathbb{R}^{s+1}$, Λ and Γ are $(s+1) \times (s+1)$ matrices such that $\Lambda e + \alpha = e$, the matrix $I - \Lambda$ is invertible, and the last column in Γ, Λ is zero. We have defined $e = (1, \ldots, 1)^t \in \mathbb{R}^{s+1}$, $U = (U_1^t, \ldots, U_s^t, u_{n+1}^t)^t \in \mathbb{R}^{(s+1)m}$, and $F(U) = (f(U_1)^t, \ldots, f(U_s)^t, 0)^t \in \mathbb{R}^{(s+1)m}$. The symbol $\otimes$ denotes the Kronecker product (see, e.g., [9, section 12.1]).

Method (1.10) is a particular case of (2.1), with $\alpha = (1, 0, \ldots, 0)^t$ and

(2.2)
$$\Lambda = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \alpha_{21} & 0 & \ddots & & \vdots \\ \alpha_{31} & \alpha_{32} & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \alpha_{s+1,1} & \alpha_{s+1,2} & \cdots & \alpha_{s+1,s} & 0 \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ \beta_{21} & 0 & \ddots & & \vdots \\ \beta_{31} & \beta_{32} & 0 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ \beta_{s+1,1} & \beta_{s+1,2} & \cdots & \beta_{s+1,s} & 0 \end{pmatrix}.$$

Method (1.17) is also a particular case of (2.1) with $\alpha = (\alpha_{21}, \alpha_{31}, \ldots, \alpha_{s+1,1}, 0)^t$ and

(2.3)
$$\Lambda = \begin{pmatrix} 0 & 0 & \cdots & \cdots & \cdots & 0 \\ \alpha_{32} & 0 & \ddots & & & \vdots \\ \alpha_{42} & \alpha_{43} & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & 0 \\ \alpha_{s+1,2} & \alpha_{s+1,3} & \cdots & \alpha_{s+1,s} & 0 & 0 \\ 0 & \cdots & \cdots & 0 & 1 & 0 \end{pmatrix}, \quad \Gamma = \begin{pmatrix} \beta_2 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & \beta_3 & \ddots & & & \vdots \\ 0 & \ddots & \beta_4 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & \beta_{s+1} & 0 \\ 0 & \cdots & \cdots & 0 & 0 & 0 \end{pmatrix}.$$

Remark 1. Observe that in (2.2), $U_1 = u_n$, and hence the elements $\alpha_{21}, \ldots, \alpha_{s+1,1}$ in Λ can appear either in Λ or in α.

In compact form, (1.6)–(1.7) can be written as

$$U = e \otimes u_n + h\, (\mathbb{A} \otimes I)\, F(U). \tag{2.4}$$

It is straightforward to obtain that (2.1) can be written as

$$U = e \otimes u_n + h\, (((I - \Lambda)^{-1}\Gamma) \otimes I)\, F(U).$$

Comparing this expression with (2.4), we obtain that (2.1) is an RK method with $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$, and for this reason we will refer to (2.1) as a representation of the RK method $\mathbb{A}$. On the other hand, if the RK coefficient matrix $\mathbb{A}$ can be factorized as $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$, we can write (2.4) in the form of (2.1). Observe that such a representation is always possible. For example, we can always take a trivial one with $\Lambda = 0$ and $\Gamma = \mathbb{A}$. More interesting representations can also be given.

Example 1. For the classical four-stage order four method we can write $\mathbb{A} = (I - \Lambda_i)^{-1}\Gamma_i$, $i = 1, 2$, with

(2.5)
$$\Lambda_1 = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ 1/2 & 1/2 & 0 & & \\ 1/9 & 2/9 & 2/3 & 0 & \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}, \qquad \Gamma_1 = \begin{pmatrix} 0 & & & & \\ 1/2 & 0 & & & \\ -1/4 & 1/2 & 0 & & \\ -1/9 & -1/3 & 1 & 0 & \\ 0 & 1/6 & 0 & 1/6 & 0 \end{pmatrix},$$

(2.6)
$$\Lambda_2 = \begin{pmatrix} 0 & & & & \\ 1/3 & 0 & & & \\ 2/9 & 1/3 & 0 & & \\ 2/27 & 2/9 & 2/3 & 0 & \\ 40/243 & 22/81 & 4/27 & 1/9 & 0 \end{pmatrix}, \qquad \Gamma_2 = \begin{pmatrix} 0 & & & & \\ 1/2 & 0 & & & \\ -1/6 & 1/2 & 0 & & \\ -1/9 & -1/3 & 1 & 0 & \\ 5/162 & 7/27 & 2/9 & 1/6 & 0 \end{pmatrix}.$$

Observe that in both cases, $\Lambda \geq 0$ but Γ contains negative elements.

Once we have defined the representations (2.1), the next step is to extend the CFL coefficient (1.13) that gives the stepsize restriction for monotonicity.

PROPOSITION 2.1. *Consider a method of the form* (2.1) *such that* $\Lambda\, e \leq e$, $\Lambda \geq 0$, $\Gamma \geq 0$, *and*

$$\Lambda - c\,\Gamma \geq 0 \qquad \textit{for some } c > 0. \tag{2.7}$$

Then for

$$h \leq c\,\frac{1}{\rho}, \tag{2.8}$$

it holds that $\|U_i\| \leq \|u_n\|$, $i = 1, \dots, s$, $\|u_{n+1}\| \leq \|u_n\|$.

Proof. The proof is the same as that for Proposition 3.9 below, with $\tilde{\Lambda} = 0$ and $\tilde{\Gamma} = 0$. □

Remark 2.

1. Observe that (2.7) implies condition (1.11).
2. Observe that the maximum c in (2.7) is

   $$c = \min_{ij} \frac{\alpha_{ij}}{\beta_{ij}}. \tag{2.9}$$

   As a particular case we get the CFL coefficient (1.13) obtained in the SSP context.
3. As $\alpha + \Lambda\, e = e$, with $\alpha \geq 0$, $\Lambda \geq 0$, the proof is also valid for a convex functional $\| \cdot \|$.

Observe that, whereas the CFL coefficient (1.13) was defined only for explicit methods, the CFL coefficient c in (2.7) can also be computed for implicit methods.

Example 2. We consider the methods of the form (2.1) with $\alpha = (1, 0, 0)^t$ and

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} 0 & 0 & 0 \\ 1/4 & 1/4 & 0 \\ 1/4 & 1/4 & 0 \end{pmatrix}. \tag{2.10}$$

It corresponds to the RK method $\mathbb{A}$ with

$$\mathbb{A} = \begin{pmatrix} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \end{pmatrix}. \tag{2.11}$$

We get $\Lambda - c\Gamma = 0$ for $c = 2$. Hence Proposition 2.1 gives that the method is monotone under stepsize restriction (2.8) with $c = 2$.

The next goal is to establish a relationship between representations of an RK method $\mathbb{A}$, $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$, the CFL coefficient c in (2.7), and $R(\mathbb{A})$. Although from the study done by Ferracina and Spijker [2] we can conclude that, for irreducible RK methods, for the CFL coefficient c dictated by (2.7) we have $c \leq R(\mathbb{A})$, this fact can easily be obtained directly for arbitrary RK schemes.

Proposition 2.2. *Consider an RK method* $\mathbb{A}$. *Assume that it can be written in terms of* Λ *and* Γ *with*

$$\Lambda\, e \leq e, \qquad \Gamma \geq 0, \qquad \Lambda - c\,\Gamma \geq 0,$$

and the matrix $I - (\Lambda - c\,\Gamma)$ *invertible for some coefficient* $c \geq 0$. *Then the method is absolutely monotonic at* $-c$.

Proof. The proof is the same as that for Proposition 3.11 below, with $\tilde{\Lambda} = 0$ and $\tilde{\Gamma} = 0$. □

In particular, if for an RK method it holds that $R(\mathbb{A}) = 0$, then it is not possible to get a representation with $\Lambda \geq 0$, $\Gamma \geq 0$, $\alpha \geq 0$, and $\Lambda - c\Gamma \geq 0$ for some $c > 0$.

An algebraic criterion for obtaining $R(\mathbb{A}) > 0$ for irreducible methods is given in [7, Theorem 4.2]. See [1, Definitions 4.4.1 and 4.4.3] for the definition of reducible methods, and [7] for a complete study of order barriers and stage order barriers when $R(\mathbb{A}) > 0$. Taking into account the extension we plan to do in section 3, we are going to obtain algebraic criteria for obtaining $R(\mathbb{A}) > 0$ for any reducible or irreducible RK methods. To proceed we need the following definition.

DEFINITION 2.3. *For matrix $F = (f_{ij})$ we define its incidence matrix* $\mathrm{Inc}(F) = (g_{ij})$ *by $g_{ij} = 1$ if $f_{ij} \neq 0$ and by $g_{ij} = 0$ if $f_{ij} = 0$.*

We give an algebraic criterion for getting $R(\mathbb{A}) > 0$.

PROPOSITION 2.4. *Consider an RK method with coefficients $\mathbb{A}$. Then $R(\mathbb{A}) > 0$ if and only if $\mathbb{A} \geq 0$ and*

$$\mathrm{Inc}(\mathbb{A}^2) \leq \mathrm{Inc}(\mathbb{A}). \tag{2.12}$$

Proof. See the proof of Proposition 3.7 below for $\tilde{\mathbb{A}} = 0$. □

Recall that Proposition 2.4 is not exactly Theorem 4.2 in [7], as we have not assumed the irreducibility of the RK method. In the following lemma we study some of the conditions involved in Proposition 2.4 for irreducible methods.

LEMMA 2.5. *Consider an irreducible RK method $\mathbb{A}$ with $\mathbb{A} \geq 0$. Then $b > 0$ if and only if* $\mathrm{Inc}(b^t\mathcal{A}) \leq \mathrm{Inc}(b^t)$.

Proof. If $b > 0$, then trivially $\mathrm{Inc}(b^t\mathcal{A}) \leq \mathrm{Inc}(b^t)$. We assume now that $\mathrm{Inc}(b^t\mathcal{A}) \leq \mathrm{Inc}(b^t)$. We consider the sets $S = \{j \mid b_j = 0\}$ and $T = \{j \mid b_j \neq 0\}$. If there is an index j such that $b_j = 0$, then $S \neq \emptyset$. Condition $\mathrm{Inc}(b^t\mathcal{A}) \leq \mathrm{Inc}(b^t)$ gives $\sum_{i=1}^{s} b_i a_{ij} = 0$ for all $j \in S$, and hence, as $\mathbb{A} \geq 0$, we get $b_i a_{ij} = 0$, $i = 1, \dots, s$, for all $j \in S$. Thus we obtain $a_{ij} = 0$ for all $i \in T$, $j \in S$. This implies that the method is DJ-reducible [1, Definition 4.4.1], that contradicts the irreducibility assumption, and therefore $b > 0$. □

With this lemma, we get the result in [7] for irreducible methods.

COROLLARY 2.6 (see [7, Theorem 4.2]). *For an irreducible coefficient scheme $\mathbb{A}$ we have $R(\mathbb{A}) > 0$ if and only if $\mathcal{A} \geq 0$, $b > 0$, and* $\mathrm{Inc}(\mathcal{A}^2) \leq \mathrm{Inc}(\mathcal{A})$.

Proof. In terms of $(\mathcal{A}, b)$, the conditions in Proposition 2.4 are $\mathcal{A} \geq 0$, $b \geq 0$, $\mathrm{Inc}(\mathcal{A}^2) \leq \mathrm{Inc}(\mathcal{A})$, $\mathrm{Inc}(b^t\mathcal{A}) \leq \mathrm{Inc}(b^t)$. By Lemma 2.5, for irreducible methods, conditions $b > 0$ and $\mathrm{Inc}(b^t\mathcal{A}) \leq \mathrm{Inc}(b^t)$ are equivalent. □

The algebraic criterion in Proposition 2.4 is extremely useful to determine whether, for a given method, the condition $R(\mathbb{A}) > 0$ holds. For example, for the classical four-stage order four method it is straightforward to prove that condition $\mathrm{Inc}(\mathbb{A}^2) \leq \mathrm{Inc}(\mathbb{A})$ does not hold and thus $R(\mathbb{A}) = 0$.

So far we have obtained that, given a representation, the CFL coefficient c computed in (2.7) satisfies $c \leq R(\mathbb{A})$. Remember that the representation of an RK method is not unique, and hence the CFL coefficient computed in (2.7) depends on the representation available. Next we study whether, given the Butcher tableau of an RK method $\mathbb{A}$ with radius $R(\mathbb{A})$, there is a representation such that the CFL coefficient computed from (2.7) is equal to $R(\mathbb{A})$. We will refer to such representations as optimal ones. We remark that our study for optimal representations is done for a given method, and it differs from the study done in [16], where an optimization process over the CFL coefficient within classes of explicit methods, with a given number of stages and order, is done.

PROPOSITION 2.7. *We consider an RK method $\mathbb{A}$. If $0 < r = R(\mathbb{A}) < \infty$, then there exist matrices Λ and Γ such that $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$ with $\Lambda \geq 0$, $\Gamma \geq 0$, $\Lambda e \leq e$, $I - (\Lambda - r\Gamma)$ invertible, and $\Lambda - r\Gamma \geq 0$.*

Proof. See the proof of Proposition 3.7 below for $\tilde{\mathbb{A}} = 0$. □

The proof of the above result is constructive and can be used to get, for a given method, an optimal representation. We remark that we can construct a representation such that $\Lambda - r\Gamma = 0$, namely,

$$\Lambda = r\,\mathbb{A}(I + r\mathbb{A})^{-1}, \qquad \Gamma = \mathbb{A} - \Lambda\mathbb{A}, \tag{2.13}$$

and $\alpha = e - \Lambda e$.

Example 3. Consider the method

$$\begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & 0 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array} .$$

For this method, $R(\mathbb{A}) = 1$. From (2.13) we compute

$$\Lambda = r\,\mathbb{A}(I + r\mathbb{A})^{-1} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{2}{3} & 0 \end{pmatrix}, \qquad \Gamma = \mathbb{A} - \Lambda\mathbb{A} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & \frac{1}{4} & 0 & 0 \\ 0 & 0 & \frac{2}{3} & 0 \end{pmatrix},$$

and $\alpha = e - \Lambda e = (1, 0, 3/4, 1/3)^t$. This is essentially the representation given in [16, Table A1]. See Remark 1.

Example 4. For the method (2.11) the radius of absolute monotonicity is $R(\mathbb{A}) = 2$. The representation obtained from (2.13) is (2.10).

In Proposition 2.1 we have given conditions to ensure conditional monotonicity. The rest of the section is devoted to unconditional monotonicity issues.

The concept of M-matrix will be used. We remember that a matrix A is said to be an M-matrix if A is nonsingular, $A^{-1} \geq 0$, and the off-diagonal elements of A are nonpositive. In the following results, Γ_* denotes the matrix Γ, where the last column has been changed to $(0, \dots, 0, 1)^t$.

PROPOSITION 2.8. *Consider a representation* (2.1) *such that* Γ_* *is invertible,* $\Gamma_*^{-1}\alpha \geq 0$, *and* $\Gamma_*^{-1}(I - \Lambda)$ *is an* M*-matrix. Then for any stepsize it holds that* $\|U_i\| \leq \|u_n\|$, $i = 1, \dots, s$, $\|u_{n+1}\| \leq \|u_n\|$.

Proof. The proof is similar to the one done in [7, Theorem 6.1]. We remark that it is also valid for a convex functional $\|\cdot\|$. □

For RK methods, unconditional monotonicity is obtained when $R(\mathbb{A}) = \infty$. If we define

$$\mathbb{A}_* = \begin{pmatrix} \mathcal{A} & 0 \\ b^t & 1 \end{pmatrix},$$

and assume that the matrix $\mathbb{A}_*$ is nonsingular, then $R(\mathbb{A}) = \infty$ if and only if $\mathbb{A}_*^{-1}$ is an M-matrix and $\mathbb{A}_*^{-1}e \geq 0$ [7, Lemma 4.5].

Observe that if $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$, then $\mathbb{A}_* = (I - \Lambda)^{-1}\Gamma_*$. Hence it is straightforward to prove that if the conditions of Proposition 2.8 hold, then $R(\mathbb{A}) = \infty$.

If $R(\mathbb{A}) = \infty$, an optimal representation can be given trivially,

$$\Gamma = \mathbb{A}, \qquad \Lambda = 0. \tag{2.14}$$

Unfortunately methods with $R(\mathbb{A}) = \infty$ have at most order one [7, Theorem 8.3], and therefore they are not of great interest.

Remark 3. Some of the results contained in this section have also been obtained independently by Ferracina and Spijker [3]. The main difference is how the case of unconditional monotonicity is handled. We briefly summarize the analogies and differences between both papers. Formula (2.1) in [3] is essentially our representation (2.1). The relationship between the matrices in [3, formula (2.2)] and our matrices Λ and Γ is

$$\Lambda = \begin{pmatrix} L_0 & 0 \\ L_1 & 0 \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} M_0 & 0 \\ M_1 & 0 \end{pmatrix}.$$

The regularity assumption for the matrix $I - \Lambda$ is equivalent to the regularity of $I - L_0$ in [3], our relationship $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$ is expression (2.5) in [3], and our conditions $\Lambda \geq 0$ and $\Lambda\, e \leq e$ are assumption (2.8) in [3]. There are two differences between the definition of the coefficient c in formula (2.9) in [3] and our CFL coefficient c in (2.7): at this point we do not consider unconditional monotonicity, and in $\Lambda - c\Gamma \geq 0$ we do not make distinctions between diagonal and nondiagonal elements. However, for explicit methods both definitions are the essentially the same. Part I in Theorem 2.2 in [3] is contained in the introduction of section 2. Our Propositions 2.1 and 2.8 on conditional and unconditional monotonicity are part II in Theorem 2.2 in [3]. We remark that Theorem 2.2 in [3] is proved simultaneously for $R(\mathbb{A}) < \infty$ and $R(\mathbb{A}) = \infty$, whereas we have obtained the results separately with different techniques. With regard to optimal representations, if $R(\mathbb{A}) < \infty$, the ones given in (3.2b) in [3, Theorem 3.4] are the same as the ones given in this paper (see (2.13) in Proposition 2.7). If $R(\mathbb{A}) = \infty$, we have given the trivial representation (2.14) that fulfills our requirements, whereas in [3] another representation is given, as it is intended to obtain the CFL coefficient ∞ in formula (2.9) in [3]. Finally, irreducibility of the RK scheme is imposed in Theorem 3.4 in [3], whereas in this paper this condition is not assumed. A detailed reading of its proof gives that irreducibility is required to prove parts I and II, where some maximal properties of CFL coefficients obtained from representations are stated. Parts I and II in [3, Theorem 3.4] are closely related with the fact that for irreducible RK methods the maximum CFL coefficient for monotonicity is $R(\mathbb{A})$ [2].

3. Perturbed RK methods. In this section we consider RK methods $\mathbb{A}$ with $R(\mathbb{A}) = 0$. In this case (see Proposition 2.1), it is not possible to obtain a representation (2.1) with positive CFL coefficient (2.7). When $R(\mathbb{A}) = 0$, we can try to perturb the method so that the new numerical solution $u_n^{(p)}$ satisfies the monotonicity condition

$$\|U_i^{(p)}\| \leq \|u_n^{(p)}\|, \quad i = 1, \dots, s, \qquad \|u_{n+1}^{(p)}\| \leq \|u_n^{(p)}\|.$$

To do so, we consider an auxiliary problem

$$\frac{d}{dt}u(t) = -\tilde{f}(u(t)), \qquad t \geq t_0, \tag{3.1}$$

and assume that, like the original problem, $(-\tilde{f}, \|\cdot\|) \in \mathcal{F}(\rho)$; i.e.,

$$\left\| y - \frac{1}{\rho}\,\tilde{f}(y) \right\| \leq \|y\| \qquad \forall\, y \in \mathbb{R}^m. \tag{3.2}$$

Given an RK method $\mathbb{A}$, we define the perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$

$$U^{(p)} = e \otimes u_n^{(p)} + h(\mathbb{A} \otimes I)F(U^{(p)}) + h(\tilde{\mathbb{A}} \otimes I)\left(F(U^{(p)}) - \tilde{F}(U^{(p)})\right), \tag{3.3}$$

with $\tilde{\mathbb{A}}$ an $(s+1) \times (s+1)$ matrix with the last column equal to $(0, \ldots, 0)^t$, and $\tilde{F}(U) = (\tilde{f}(U_1)^t, \ldots, \tilde{f}(U_s)^t, 0^t) \in \mathbb{R}^{(s+1)m}$. As far as we know, this is a new concept.

Assuming stability of the RK method $\mathbb{A}$, it is easy to prove that if the perturbations are small enough, then the perturbed method $(\mathbb{A}, \tilde{\mathbb{A}})$ will retain the order of the unperturbed problem. The standard stability bounds [8, p. 32] give

$$\|u_n^{(p)} - u_n\| \leq M\varepsilon,$$

provided that

$$h\|F(U) - \tilde{F}(U)\| \leq \varepsilon. \tag{3.4}$$

Hence, if the method $\mathbb{A}$ has order p, we have

$$\|u_n^{(p)} - u(t_n)\| \leq \|u_n^{(p)} - u_n\| + \|u_n - u(t_n)\| \leq M\varepsilon + Ch^p,$$

and therefore we obtain order p whenever $M\varepsilon = \vartheta(h^p)$. In the following we drop the exponent (p) for the perturbed solution.

In the context of SSP methods, where the ODE comes from a semidiscretization of the hyperbolic conservation laws, the auxiliary problem (3.1) is obtained with another semidiscretization such that (3.2) holds. For example [13, p. 444], for $u_t = u_x$ we can semidiscretize u_x as

$$u_x(x_j, t) \approx \frac{u(x_j + \Delta x, t) - u(x_j, t)}{\Delta x} := f(u)_j$$

and

$$u_x(x_j, t) \approx \frac{u(x_j, t) - u(x_j - \Delta x, t)}{\Delta x} := \tilde{f}(u)_j,$$

obtaining, respectively, properties (1.3) and (3.2) for $\rho = 1/\Delta x$. In this case (3.4) holds with ε depending on the space discretization step Δx.

3.1. The radius of absolute monotonicity for perturbed RK methods. The radius of absolute monotonicity had an important role in monotonicity. In this section we extend this concept to perturbed RK methods $(\mathbb{A}, \tilde{\mathbb{A}})$ and analyze some of its properties. For a better understanding of that extension, we briefly show how conditions in Definition 1.1 arise in [7].

In [7] the scalar linear problems $u' = \lambda u$, $u' = \lambda(t)u$ and the vectorial linear problem $u' = L(t)u(t)$, with $L(t)$ an $m \times m$ matrix, are considered. An RK method $\mathbb{A}$ for these problems gives, respectively, $U = \phi(h\lambda)u_n$ with $\phi(z) = (I_{s+1} - z\mathbb{A})^{-1}e$, $U = K(\operatorname{diag}(h\lambda_1, \ldots, h\lambda_s, 0))\, u_n$ with $K(Z) = (I_{s+1} - \mathbb{A}Z)^{-1}e$ and Z a diagonal matrix, and $U = \mathbb{K}(\operatorname{diag}(hL_1, \ldots, hL_s, 0)) \otimes u_n$ with

$$\mathbb{K}(\mathbb{Z}) = \left(I_{(s+1)\cdot m} - (\mathbb{A} \otimes I_m)\mathbb{Z}\right)^{-1}(e \otimes I_m)$$

and $\mathbb{Z}$ a block diagonal matrix. The concepts of absolute monotonicity at a given point $\xi \in \mathbb{R}$ for ϕ, K, and $\mathbb{K}$ are given in [7, p. 487]. Roughly speaking, they mean the nonnegativity of all coefficients of the Taylor expansion of ϕ, K, and $\mathbb{K}$ about $z=\xi$, $Z=\xi I_{s+1}$, or $\mathbb{Z}=\xi I_{(s+1)\cdot m}$, respectively. In particular, for $\mathbb{Z}=\xi I_{(s+1)\cdot m}+\mathbb{W}$ with $\mathbb{W}$ sufficiently close to zero, we obtain

$$\mathbb{K}(\mathbb{Z})=\left[I_{(s+1)\cdot m}-(\mathsf{A}(\xi)\otimes I_m)\mathbb{W}\right]^{-1}\mathsf{e}(\xi)\otimes I_m$$

with $\mathsf{A}(\xi)=(I_{s+1}-\xi\mathbb{A})^{-1}\mathbb{A}$ and $\mathsf{e}(\xi)=(I_{s+1}-\xi\mathbb{A})^{-1}e$. Thus with $\mathsf{A}(\xi)\geq 0$ and $\mathsf{e}(\xi)\geq 0$ (see (1.21)), we obtain the absolute monotonicity of $\mathbb{K}$ at $\xi I_{(s+1)\cdot m}$.

Similarly, if we consider now the perturbed problems, $u'=-\tilde{\lambda}u$, $u'=-\tilde{\lambda}(t)u$, and $u'=-\tilde{L}(t)u(t)$, with $\tilde{L}(t)$ an $m\times m$ matrix, a perturbed RK method $(\mathbb{A},\tilde{\mathbb{A}})$ gives, respectively, $U=\phi(h\lambda,h\tilde{\lambda})u_n$ with $\phi(z,\tilde{z})=(I_{s+1}-z(\mathbb{A}+\tilde{\mathbb{A}})-\tilde{z}\tilde{\mathbb{A}})^{-1}e$, $U=K\left(\operatorname{diag}(h\lambda_1,\ldots,h\lambda_s,0),\operatorname{diag}(h\tilde{\lambda}_1,\ldots,h\tilde{\lambda}_s,0)\right)u_n$ with

$$K(Z,\tilde{Z})=(I_{s+1}-(\mathbb{A}+\tilde{\mathbb{A}})Z-\tilde{\mathbb{A}}\tilde{Z})^{-1}e$$

and $Z,\tilde{Z}$ diagonal matrices, and

$$U=\mathbb{K}\left(\operatorname{diag}(hL_1,\ldots,hL_s,0),\operatorname{diag}(h\tilde{L}_1,\ldots,h\tilde{L}_s,0)\right)\otimes u_n$$

with

$$\mathbb{K}(\mathbb{Z},\tilde{\mathbb{Z}})=(I_{(s+1)\cdot m}-((\mathbb{A}+\tilde{\mathbb{A}})\otimes I_m)\mathbb{Z}-(\tilde{\mathbb{A}}\otimes I_m)\tilde{\mathbb{Z}})^{-1}(e\otimes I_m)$$

and $\mathbb{Z},\tilde{\mathbb{Z}}$ block diagonal matrices. For $\mathbb{Z}=\xi I_{(s+1)\cdot m}+\mathbb{W}$, $\tilde{\mathbb{Z}}=\xi I_{(s+1)\cdot m}+\tilde{\mathbb{W}}$ with $\mathbb{W},\tilde{\mathbb{W}}$ sufficiently close to zero, we obtain

$$\mathbb{K}(\mathbb{Z},\tilde{\mathbb{Z}})=[I_{(s+1)\cdot m}-(\mathsf{A}(\xi)\otimes I_m)\mathbb{W}-(\tilde{\mathsf{A}}(\xi)\otimes I_m)\tilde{\mathbb{W}}]^{-1}\left(\mathsf{e}(\xi)\otimes I_m\right)$$

with $\mathsf{A}(\xi)$, $\tilde{\mathsf{A}}(\xi)$, and $\mathsf{e}(\xi)$ defined by (3.5)–(3.7) below. Thus with $\mathsf{A}(\xi)\geq 0$, $\tilde{\mathsf{A}}(\xi)\geq 0$, $\mathsf{e}(\xi)\geq 0$, we obtain that all the coefficients in the Taylor expansion of $\mathbb{K}$ at $\xi I_{(s+1)\cdot m}$ are nonnegative. This analysis leads us to the extension of Definition 1.1 as follows.

Definition 3.1. *An s-stage perturbed RK method $(\mathbb{A},\tilde{\mathbb{A}})$ is said to be absolutely monotonic at a given point $\xi\leq 0$ if $I-\xi(\mathbb{A}+2\tilde{\mathbb{A}})$ is invertible and*

$$\mathsf{A}(\xi)=(I-\xi(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}(\mathbb{A}+\tilde{\mathbb{A}})\geq 0, \tag{3.5}$$

$$\tilde{\mathsf{A}}(\xi)=(I-\xi(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}}\geq 0, \tag{3.6}$$

$$\mathsf{e}(\xi)=(I-\xi(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}e\geq 0. \tag{3.7}$$

Further, the perturbed method is said to be absolutely monotonic on a given set $\Omega\in\mathbb{R}$ if it is absolutely monotonic at each $\xi\in\Omega$. The radius of absolute monotonicity $R(\mathbb{A},\tilde{\mathbb{A}})$ is defined by

$$R(\mathbb{A},\tilde{\mathbb{A}})=\sup\{r\mid r\geq 0 \textit{ and } (\mathbb{A},\tilde{\mathbb{A}}) \textit{ is absolutely monotonic on } [-r,0]\}.$$

If there is no $r>0$ such that $(\mathbb{A},\tilde{\mathbb{A}})$ is absolutely monotonic on $[-r,0]$, we set $R(\mathbb{A},\tilde{\mathbb{A}})=0$.

We go deeper into the concept of radius of absolute monotonicity. In [7, Lemma 4.4] it is proved that for checking the absolute monotonicity of an RK method $\mathbb{A}$ on a given interval $[-r,0]$ it is sufficient to consider the left endpoint $-r$ only. Our next goal is to prove that a similar result is also true for perturbed RK methods $(\mathbb{A},\tilde{\mathbb{A}})$. We begin with some technical lemmas.

Lemma 3.2. *Consider an order m matrix $B \geq 0$ such that for a given $r \geq 0$ the matrix $I + rB$ is nonsingular and $B(I+rB)^{-1} \geq 0$. Then $I - \xi B$ is nonsingular for all $\xi \in [-2r, 0]$.*

Proof. As $B(I+rB)^{-1} \geq 0$ and $B \geq 0$, then $(I+rB)^{-1} = I - rB(I+rB)^{-1}$ is an M-matrix, and hence [9, section 15.2], denoting by $\sigma(C)$ the spectral radius of the matrix C, we have $\sigma\left(rB(I+rB)^{-1}\right) < 1$. Therefore $\det\left(\lambda I - rB(I+rB)^{-1}\right) \neq 0$ for all λ with $|\lambda| \geq 1$. For $\xi \neq -r$, $r \neq 0$, we have that

$$\det\left(I - (\xi + r)B(I+rB)^{-1}\right) = \frac{(\xi+r)^m}{r^m} \det\left(\frac{r}{\xi + r} I - rB(I+rB)^{-1}\right),$$

obtaining $\det\left(I - (\xi+r)B(I+rB)^{-1}\right) \neq 0$ for all ξ with $|\xi + r| \leq r$. Finally, as $I - \xi B = \left(I - (\xi+r)B(I+rB)^{-1}\right) \cdot (I+rB)$, we can write

$$\det(I - \xi B) = \det\left(I - (\xi+r)B(I+rB)^{-1}\right) \cdot \det(I + rB)$$

to obtain that $I - \xi B$ is nonsingular for all $\xi \in [-2r, 0]$. □

For the next result we recall the concept of absolutely monotonic function [7, Definition 2.1]. A function $\psi(z)$ is said to be absolutely monotonic at a given point $\xi \in \mathbb{R}$ if $(d^k\psi/dz^k)(\xi) \geq 0$, $k = 0, 1, \dots$. For matrices or vectors whose components are functions, we will say that they are absolutely monotonic at a given point ξ if each element or component is absolutely monotonic at ξ.

Lemma 3.3. *Consider the functions $\mathsf{A}(\xi)$, $\tilde{\mathsf{A}}(\xi)$, and $\mathsf{e}(\xi)$, defined by* (3.5), (3.6), *and* (3.7), *respectively. Then the perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ is absolutely monotonic at ξ_0 if and only if the functions $\mathsf{A}(\xi)$, $\tilde{\mathsf{A}}(\xi)$, and $\mathsf{e}(\xi)$ are absolutely monotonic at ξ_0.*

Proof. Recall that from the definition the perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ is absolutely monotonic at ξ_0 if and only if $\mathsf{A}(\xi_0) \geq 0$, $\tilde{\mathsf{A}}(\xi_0) \geq 0$, and $\mathsf{e}(\xi_0) \geq 0$. The if part is trivial. For the only if part we simply have to observe that $(d^k\mathsf{A}/d\xi^k)(\xi_0) = k!\,[\mathsf{A}(\xi_0) + \tilde{\mathsf{A}}(\xi_0)]^k \mathsf{A}(\xi_0)$, and similarly for $(d^k\tilde{\mathsf{A}}/d\xi^k)(\xi_0)$ and $(d^k\mathsf{e}/d\xi^k)(\xi_0)$. Thus from $\mathsf{A}(\xi_0) \geq 0$, $\tilde{\mathsf{A}}(\xi_0) \geq 0$, and $\mathsf{e}(\xi_0) \geq 0$ we obtain the absolute monotonicity of $\mathsf{A}(\xi)$, $\tilde{\mathsf{A}}(\xi)$, and $\mathsf{e}(\xi)$ at ξ_0. □

We are in position to prove for the perturbed RK methods an analogous result to Lemma 4.4 in [7].

Proposition 3.4. *Consider a perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ and a real positive number r. Then $R(\mathbb{A}, \tilde{\mathbb{A}}) \geq r$ if and only if $(\mathbb{A}, \tilde{\mathbb{A}})$ is absolutely monotonic at $\xi = -r$ and $\mathbb{A} + \tilde{\mathbb{A}} \geq 0$, $\tilde{\mathbb{A}} \geq 0$.*

Proof. 1. We begin by assuming that $R(\mathbb{A}, \tilde{\mathbb{A}}) \geq r$. Then $(\mathbb{A}, \tilde{\mathbb{A}})$ is absolutely monotonic on $(-r, 0]$, and hence by Lemma 3.3 the functions $\mathsf{A}(\xi)$, $\tilde{\mathsf{A}}(\xi)$, and $\mathsf{e}(\xi)$ are absolutely monotonic on $(-r, 0]$. We can now apply componentwise Lemma 3.6 in [7] to get that they are also absolutely monotonic on $[-r, 0]$. In particular, the method is absolutely monotonic at $\xi = 0$, and hence $\mathbb{A} + \tilde{\mathbb{A}} \geq 0$ and $\tilde{\mathbb{A}} \geq 0$.

2. We assume now that $(\mathbb{A}, \tilde{\mathbb{A}})$ is absolutely monotonic at $\xi = -r$ and $\mathbb{A} + \tilde{\mathbb{A}} \geq 0$, $\tilde{\mathbb{A}} \geq 0$. Using Lemma 3.2 for $B = \mathbb{A} + 2\tilde{\mathbb{A}}$, we get that $I - \xi(\mathbb{A} + 2\tilde{\mathbb{A}})$ is nonsingular for all $\xi \in [-r, 0]$, and hence the functions $\mathsf{A}(\xi)$, $\tilde{\mathsf{A}}(\xi)$, and $\mathsf{e}(\xi)$ are well defined for $\xi \in [-r, 0]$. We can apply componentwise Lemma 3.1 in [7] to obtain that these functions are absolutely monotonic on $[-r, 0]$, and hence, by Lemma 3.3, $R(\mathbb{A}, \tilde{\mathbb{A}}) \geq r$. □

Example 5. For the classical fourth order RK method we consider the perturbed RK method

$$U_1 = u_n,$$
$$U_2 = u_n + \frac{1}{2} h f(U_1),$$

$$
\begin{aligned}
U_3 &= u_n + \frac{1}{2}hf(U_2) + \frac{1}{4}h[f(U_1) - \tilde{f}(U_1)],\\
U_4 &= u_n + hf(U_3) + \frac{5}{18}h[f(U_1) - \tilde{f}(U_1)] + \frac{1}{3}h[f(U_2) - \tilde{f}(U_2)],\\
u_{n+1} &= u_n + \frac{1}{6}hf(U_1) + \frac{1}{3}hf(U_2) + \frac{1}{3}hf(U_3) + \frac{1}{6}hf(U_4) + \frac{19}{108}h[f(U_1) - \tilde{f}(U_1)]\\
&\quad + \frac{1}{9}h[f(U_2) - \tilde{f}(U_2)].
\end{aligned}
$$

The coefficient matrices $(\mathbb{A}, \tilde{\mathbb{A}})$ are

$$
\mathbb{A} = \begin{pmatrix} 0 & & & & \\ 1/2 & 0 & & & \\ 0 & 1/2 & 0 & & \\ 0 & 0 & 1 & 0 & \\ 1/6 & 1/3 & 1/3 & 1/6 & 0 \end{pmatrix}, \qquad \tilde{\mathbb{A}} = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 1/4 & 0 & 0 & & \\ 5/18 & 1/3 & 0 & 0 & \\ 19/108 & 1/9 & 0 & 0 & 0 \end{pmatrix}. \tag{3.8}
$$

It can be checked that for this perturbed RK method $R(\mathbb{A}, \tilde{\mathbb{A}}) = 2/3$.

We are in position to obtain a conditional monotonicity result for the perturbed method.

THEOREM 3.5. *Assume that the perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ is absolutely monotonic at $-r$. Then for*

$$h \le r\,\frac{1}{\rho}$$

it holds that $\|U_i\| \le \|u_n\|$, $i = 1, \dots, s$, $\|u_{n+1}\| \le \|u_n\|$.

Proof. The perturbed RK method can be written as

$$U = e \otimes u_n + h((\mathbb{A} + \tilde{\mathbb{A}}) \otimes I)F(U) - h(\tilde{\mathbb{A}} \otimes I)\tilde{F}(U). \tag{3.9}$$

Observe that the conditions on the problems imply, for $h \le r/\rho$, that

$$\left\| U_i + \frac{h}{r}F(U_i) \right\| \le \|U_i\|, \qquad \left\| U_i - \frac{h}{r}\tilde{F}(U_i) \right\| \le \|U_i\|. \tag{3.10}$$

In (3.9) we add to both sides $r((\mathbb{A} + 2\tilde{\mathbb{A}}) \otimes I)U$, obtaining

$$
\begin{aligned}
&(I + r((\mathbb{A} + 2\tilde{\mathbb{A}}) \otimes I))\, U\\
&= e \otimes u_n + r\,((\mathbb{A} + \tilde{\mathbb{A}}) \otimes I) \left(U + \frac{h}{r}F(U) \right) + r\,(\tilde{\mathbb{A}} \otimes I) \left(U - \frac{h}{r}\tilde{F}(U) \right),
\end{aligned}
$$

or equivalently

$$U = \mathsf{e}(-r) \otimes u_n + r\,(\mathsf{A}(-r) \otimes I) \left(U + \frac{h}{r}F(U) \right) + r\,(\tilde{\mathsf{A}}(-r)) \otimes I) \left(U - \frac{h}{r}\tilde{F}(U) \right),$$

where $\mathsf{e}(\xi)$, $\mathsf{A}(\xi)$, and $\tilde{\mathsf{A}}(\xi)$ are given by (3.5)–(3.7). If we take norms, the conditions on $\mathsf{e}(-r)$, $\mathsf{A}(-r)$, and $\tilde{\mathsf{A}}(-r)$ imply

$$
\begin{aligned}
[\|U\|] &\le \mathsf{e}(-r) \otimes \|u_n\| + r\,(\mathsf{A}(-r) \otimes I) \left[\left\| U + \frac{h}{r}F(U) \right\| \right]\\
&\quad + r\,(\tilde{\mathsf{A}}(-r) \otimes I) \left[\left\| U - \frac{h}{r}\tilde{F}(U) \right\| \right],
\end{aligned} \tag{3.11}
$$

where $[\|U\|]^t = (\|U_1\|, \dots, \|U_s\|, \|u_{n+1}\|)^t \in \mathbb{R}^{s+1}$. Conditions (3.10) now imply

$$[\|U\|] \le \mathsf{e}(-r) \otimes \|u_n\| + r\,((\mathsf{A}(-r) + \tilde{\mathsf{A}}(-r)) \otimes I)[\|U\|],$$

and hence

$$((I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1} \otimes I)\,[\|U\|] \le ((I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1} e) \otimes \|u_n\|, \tag{3.12}$$

where we have used that $I - r(\mathsf{A}(-r) + \tilde{\mathsf{A}}(-r)) = (I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}$. We simply have to multiply (3.12) by $I + r(\mathbb{A} + 2\tilde{\mathbb{A}}) \ge 0$ to get $\|U_i\| \le \|u_n\|$, $i = 1, \dots, s+1$. Observe that $U_{s+1} = u_{n+1}$, and thus we obtain the desired result. $\square$

Remark 4. As $\mathsf{e}(-r) + \mathsf{A}(-r)\,e + \tilde{\mathsf{A}}(-r)\,e = e$, with $\mathsf{e}(-r) \ge 0$, $\mathsf{A}(-r) \ge 0$, and $\tilde{\mathsf{A}}(-r) \ge 0$, inequality (3.11) is also valid for a convex functional $\|\cdot\|$.

The above result gives us monotonicity under the stepsize restriction

$$h \le R(\mathbb{A}, \tilde{\mathbb{A}})\,\frac{1}{\rho}. \tag{3.13}$$

Hence, given a method $\mathbb{A}$ with $R(\mathbb{A}) = 0$, we should find $\tilde{\mathbb{A}}$ such that $R(\mathbb{A}, \tilde{\mathbb{A}}) > 0$, because in this case we have a positive stepsize restriction for monotonicity.

We prove the following auxiliary lemma.

LEMMA 3.6. *Consider matrices $A = (a_{ij})$ and $B = (b_{ij})$ such that $A \ge 0$, $B \ge 0$, and* $\mathrm{Inc}(BA) \le \mathrm{Inc}(A)$. *Then* $\mathrm{Inc}(B^k A) \le \mathrm{Inc}(A)$ *for all $k \ge 2$.*

Proof. We prove it by induction. For $k = 1$ the statement is true. We assume that it is also true for k, and we will prove it for $k+1$. If $(B^{k+1}A)_{ij} \ne 0$, then $b_{il}(B^k A)_{lj} \ne 0$ for some l, and hence $b_{il} \ne 0$ and $(B^k A)_{lj} \ne 0$. From $(B^k A)_{lj} \ne 0$ and $\mathrm{Inc}(B^k A) \le \mathrm{Inc}(A)$, we get that $a_{lj} \ne 0$. Now from $b_{il} \ne 0$ and $a_{lj} \ne 0$ we obtain that $(BA)_{ij} \ne 0$. Finally, from $\mathrm{Inc}(BA) \le \mathrm{Inc}(A)$ we obtain that $a_{ij} \ne 0$. $\square$

We are in position to give a criterion for getting $R(\mathbb{A}, \tilde{\mathbb{A}}) > 0$.

PROPOSITION 3.7. *We have $R(\mathbb{A}, \tilde{\mathbb{A}}) > 0$ if and only if $\mathbb{A} + \tilde{\mathbb{A}} \ge 0$, $\tilde{\mathbb{A}} \ge 0$, and*

$$\mathrm{Inc}((\mathbb{A} + 2\tilde{\mathbb{A}})(\mathbb{A} + \tilde{\mathbb{A}})) \le \mathrm{Inc}(\mathbb{A} + \tilde{\mathbb{A}}), \tag{3.14}$$

$$\mathrm{Inc}((\mathbb{A} + 2\tilde{\mathbb{A}})\tilde{\mathbb{A}}) \le \mathrm{Inc}(\tilde{\mathbb{A}}). \tag{3.15}$$

Proof. The proof is similar to that of Theorem 4.2 in [7]. For real ξ close to zero, the matrix $(I - \xi(\mathbb{A} + 2\tilde{\mathbb{A}}))$ is nonsingular and

$$(I - \xi(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1} = I + \xi(\mathbb{A} + 2\tilde{\mathbb{A}}) + \xi^2(\mathbb{A} + 2\tilde{\mathbb{A}})^3 + \cdots,$$

and hence

$$(I - \xi(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}(\mathbb{A} + \tilde{\mathbb{A}}) = \mathbb{A} + \tilde{\mathbb{A}} + \xi(\mathbb{A} + 2\tilde{\mathbb{A}})(\mathbb{A} + \tilde{\mathbb{A}}) + \cdots, \tag{3.16}$$

$$(I - \xi(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}} = \tilde{\mathbb{A}} + \xi(\mathbb{A} + 2\tilde{\mathbb{A}})\tilde{\mathbb{A}} + \cdots. \tag{3.17}$$

From (3.16) and (3.17) we see that $\mathbb{A} + \tilde{\mathbb{A}} \ge 0$, $\tilde{\mathbb{A}} \ge 0$, and (3.14)–(3.15) are necessary conditions for (3.5) and (3.6) to hold on a left neighborhood of $\xi = 0$. To see that they are also sufficient we use Lemma 3.6 with $B = \mathbb{A} + 2\tilde{\mathbb{A}}$, $A = \mathbb{A} + \tilde{\mathbb{A}}$, and $A = \tilde{\mathbb{A}}$ to state that for $k \ge 2$ we have

$$\mathrm{Inc}((\mathbb{A} + 2\tilde{\mathbb{A}})^k(\mathbb{A} + \tilde{\mathbb{A}})) \le \mathrm{Inc}(\mathbb{A} + \tilde{\mathbb{A}}), \qquad \mathrm{Inc}((\mathbb{A} + 2\tilde{\mathbb{A}})^k\tilde{\mathbb{A}}) \le \mathrm{Inc}(\tilde{\mathbb{A}}).$$

Hence as $\mathbb{A}+\tilde{\mathbb{A}} \geq 0$ and $\tilde{\mathbb{A}} \geq 0$, in (3.16) and (3.17) we get (3.5) and (3.6), respectively. Observe that inequality (3.7) always holds for r close to zero. □

Given a method $\mathbb{A}$ with $R(\mathbb{A}) = 0$, we want to find $\tilde{\mathbb{A}}$ such that $R(\mathbb{A}, \tilde{\mathbb{A}}) > 0$. In terms of computation and storage it is expensive to deal with $f(U_i)$ and $\tilde{f}(U_i)$. Furthermore, as it is pointed out in [12, p. 978], the differences $f(U_i) - \tilde{f}(U_i)$ contribute to artificial dissipation and smearing. Therefore the matrix $\tilde{\mathbb{A}}$ should have as few nonzero columns as possible, although, on the other hand, with more nonzero columns, perhaps better CFL coefficients can be found.

Example 6. Consider the family of four-stage order four methods with $w \neq 0$,

$$\mathbb{A} = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 1/2 & 0 & \ddots & & 0 \\ 1/2 - 1/(6w) & 1/(6w) & 0 & \ddots & 0 \\ 0 & 1-3w & 3w & 0 & 0 \\ 1/6 & 2/3 - w & w & 1/6 & 0 \end{pmatrix}. \tag{3.18}$$

As we should have $\mathbb{A}+\tilde{\mathbb{A}} \geq 0$, whenever $a_{ij} < 0$ we require that $\tilde{a}_{ij} \neq 0$. Hence the sign pattern of the coefficient matrix $\mathbb{A}$ determines the compulsory nonzero elements in $\tilde{\mathbb{A}}$. Furthermore, from conditions (3.14) we obtain that $\tilde{a}_{31} \neq 0$ implies that $\tilde{a}_{41} \neq 0$; $\tilde{a}_{32} \neq 0$ implies that $\tilde{a}_{42} \neq 0$; $\tilde{a}_{41} \neq 0$ implies that $\tilde{a}_{51} \neq 0$; $\tilde{a}_{42} \neq 0$ implies that $\tilde{a}_{52} \neq 0$; and $\tilde{a}_{43} \neq 0$ implies that $\tilde{a}_{53} \neq 0$. For $w < 0$ or $w > 1/3$ from (3.15) we get that $\tilde{a}_{42} \neq 0$ implies that $\tilde{a}_{41} \neq 0$. Hence, we conclude the following.

w	Compulsory nonzero elements	Implied nonzero elements
$w < 0$	$\tilde{a}_{32}, \tilde{a}_{43}, \tilde{a}_{53}$	$\tilde{a}_{41}, \tilde{a}_{42}, \tilde{a}_{51}, \tilde{a}_{52}$
$0 < w < \frac{1}{3}$	$\tilde{a}_{31}$	$\tilde{a}_{41}, \tilde{a}_{51}$
$w = \frac{1}{3}$		$\tilde{a}_{31}, \tilde{a}_{41}, \tilde{a}_{51}, \tilde{a}_{42}, \tilde{a}_{52}$
$\frac{1}{3} < w < \frac{2}{3}$	$\tilde{a}_{42}$	$\tilde{a}_{41}, \tilde{a}_{51}, \tilde{a}_{52}$
$w = \frac{2}{3}$	$\tilde{a}_{42}$	$\tilde{a}_{41}, \tilde{a}_{51}, \tilde{a}_{52}$
$\frac{2}{3} < w$	$\tilde{a}_{42}, \tilde{a}_{52}$	$\tilde{a}_{41}, \tilde{a}_{51}$

Observe that for $0 < w < 1/3$ in $\tilde{\mathbb{A}}$ only one column is required, whereas for the rest we must consider two or three columns. Recall that the classical fourth order four-stage method is obtained for $w = 1/3$.

3.2. Representations of perturbed RK methods. We consider now perturbed methods of the form

$$U = \alpha \otimes u_n + ((\Lambda + \tilde{\Lambda}) \otimes I)U + h(\Gamma \otimes I)F(U) - h(\tilde{\Gamma} \otimes I)\tilde{F}(U), \tag{3.19}$$

with $\alpha \in \mathbb{R}^{s+1}$; Λ, $\tilde{\Lambda}$, Γ, and $\tilde{\Gamma}$ are $(s+1) \times (s+1)$ matrices such that $(\Lambda + \tilde{\Lambda})\, e + \alpha = e$, the matrix $I - (\Lambda + \tilde{\Lambda})$ is invertible, and the last column in Λ, $\tilde{\Lambda}$, Γ, and $\tilde{\Gamma}$ is zero. The elements in $\tilde{\Lambda}$, $\tilde{\Gamma}$ will be denoted by $\tilde{\alpha}_{ij}$ and $\tilde{\beta}_{ij}$, respectively. The elements in $\Lambda + \tilde{\Lambda}$ will be denoted by λ_{ij}.

We remark that in (3.19) we have considered the matrix $\Lambda + \tilde{\Lambda}$. We have used this notation because later on this matrix will be split into Λ and $\tilde{\Lambda}$.

It is straightforward to get that (3.19) can be written as

$$U = e \otimes u_n + h\left(((I-(\Lambda+\tilde{\Lambda}))^{-1}\Gamma)\otimes I\right)F(U) - h\left(((I-(\Lambda+\tilde{\Lambda}))^{-1}\tilde{\Gamma})\otimes I\right)\tilde{F}(U).$$

Comparing this expression with (3.3), we obtain that (3.19) is a perturbed RK method with

$$\mathbb{A} = (I-(\Lambda+\tilde{\Lambda}))^{-1}(\Gamma-\tilde{\Gamma}), \qquad \tilde{\mathbb{A}} = (I-(\Lambda+\tilde{\Lambda}))^{-1}\tilde{\Gamma}. \tag{3.20}$$

Example 7. Method (1.16) is a perturbed RK method (3.19) with

$$\Lambda+\tilde{\Lambda} = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ 1/2 & 1/2 & 0 & & \\ 1/9 & 2/9 & 2/3 & 0 & \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}, \tag{3.21}$$

$\alpha = (1,0,0,0,0)^t$, and

$$\Gamma = \begin{pmatrix} 0 & & & & \\ 1/2 & 0 & & & \\ 0 & 1/2 & 0 & & \\ 0 & 0 & 1 & 0 & \\ 0 & 1/6 & 0 & 1/6 & 0 \end{pmatrix}, \qquad \tilde{\Gamma} = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 1/4 & 0 & 0 & & \\ 1/9 & 1/3 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.22}$$

Later on we will split (3.21) into Λ and $\tilde{\Lambda}$. Observe that in $\Gamma = (\beta_{ij})$ and $\tilde{\Gamma} = (\tilde{\beta}_{ij})$ whenever $\beta_{ij} \neq 0$, then $\tilde{\beta}_{ij} = 0$, and whenever $\tilde{\beta}_{ij} \neq 0$, then $\beta_{ij} = 0$.

From (3.20) we obtain the matrices $\mathbb{A}$ and $\tilde{\mathbb{A}}$ in (3.8). The matrices Λ_1 and Γ_1 in the factorization (2.5) are $\Lambda_1 = \Lambda + \tilde{\Lambda}$ and $\Gamma_1 = \Gamma - \tilde{\Gamma}$, with $\Lambda + \tilde{\Lambda}$, Γ, and $\tilde{\Gamma}$ from (3.21)–(3.22).

For some results in the rest of the paper, the following technical lemma will be useful.

LEMMA 3.8. *If a matrix $B \in \mathbb{R}^{m\times m}$ satisfies $B \geq 0$, $B\,e \leq e$ and $I - B$ is invertible, then $(I-B)^{-1} \geq 0$.*

Proof. As $B\,e \leq e$, by the Gershgorin theorem, the spectral radius $\rho(B)$ satisfies $\rho(B) \leq 1$. As $B \geq 0$, the matrix B has a real eigenvalue r, equal to the spectral radius of B [9, Theorem 15.5.1]. Thus if $\rho(B) = 1$, the matrix B would have the eigenvalue $r = 1$, which contradicts the invertibility of $I - B$. Hence $\rho(B) < 1$. In this case, the matrix $I - B$ is an M-matrix [9, Theorem 15.2.2]; i.e., $(I-B)^{-1} \geq 0$. □

We study stepsize restrictions to get monotonicity for the schemes (3.19). We get an extension of the CFL coefficient (1.15) for the Shu–Osher representations (1.10).

PROPOSITION 3.9. *Consider a method of the form* (3.19) *such that $(\Lambda+\tilde{\Lambda})e \leq e$, $\Lambda \geq 0$, $\tilde{\Lambda} \geq 0$, $\Gamma \geq 0$, $\tilde{\Gamma} \geq 0$, and*

$$\Lambda - c\,\Gamma \geq 0, \qquad \tilde{\Lambda} - c\,\tilde{\Gamma} \geq 0 \qquad \textit{for some } c > 0. \tag{3.23}$$

Then for

$$h \leq c\,\frac{1}{\rho} \tag{3.24}$$

it holds that $\|U_i\| \leq \|u_n\|$, $i = 1,\dots,s$, $\|u_{n+1}\| \leq \|u_n\|$.

Proof. Conditions on the problems imply that

$$\left\| U_j + h\frac{\beta_{ij}}{\alpha_{ij}} F(U_j) \right\| \le \|U_j\| \qquad \text{for } h\,\frac{\beta_{ij}}{\alpha_{ij}} \le \frac{1}{\rho},$$
$$\left\| U_j - h\frac{\tilde\beta_{ij}}{\tilde\alpha_{ij}} \tilde F(U_j) \right\| \le \|U_j\| \qquad \text{for } h\,\frac{\tilde\beta_{ij}}{\tilde\alpha_{ij}} \le \frac{1}{\rho}.$$

From (3.19), using that $\alpha \ge 0$, $\Lambda \ge 0$, $\tilde\Lambda \ge 0$, $\Gamma \ge 0$, and $\tilde\Gamma \ge 0$, we immediately obtain

$$\|U_i\| \le \alpha_i \|u_n\| + \sum_{j=1}^{s} \alpha_{ij} \left\| U_j + h\frac{\beta_{ij}}{\alpha_{ij}} F(U_j) \right\| + \sum_{j=1}^{s} \tilde\alpha_{ij} \left\| U_j - h\frac{\tilde\beta_{ij}}{\tilde\alpha_{ij}} \tilde F(U_j) \right\|$$
$$\le \alpha_i \|u_n\| + \sum_{j=1}^{s} (\alpha_{ij} + \tilde\alpha_{ij}) \|U_j\|, \tag{3.25}$$

for h satisfying (3.24), or in vectorial form

$$[\|U\|] \le \alpha \otimes \|u_n\| + ((\Lambda + \tilde\Lambda) \otimes I)\, [\|U\|]\,,$$

i.e.,

$$((I - (\Lambda + \tilde\Lambda)) \otimes I)\, [\|U\|] \le \alpha \otimes \|u_n\|.$$

We can apply Lemma 3.8 to $\Lambda + \tilde\Lambda$ to obtain that $(I - (\Lambda + \tilde\Lambda))^{-1} \ge 0$, and thus we obtain the desired result. □

Remark 5.

1. Observe that from (3.23) we get that $\alpha_{ij} = 0$ implies $\beta_{ij} = 0$, and $\tilde\alpha_{ij} = 0$ implies $\tilde\beta_{ij} = 0$.
2. Observe that the maximum c in (3.23) is
$$c = \min_{ij} \left\{ \frac{\alpha_{ij}}{\beta_{ij}}, \frac{\tilde\alpha_{ij}}{\tilde\beta_{ij}} \right\}. \tag{3.26}$$
3. Given any factorization $\mathbb{A} = (I - \Lambda)^{-1}\Gamma$, where Γ has positive and negative coefficients, we can consider the sign splitting $\Gamma = \Gamma_+ - \Gamma_-$, with $\Gamma_+ \ge 0$ and $\Gamma_- \ge 0$. According to this splitting we take the corresponding terms in Λ to split it into $\Lambda = \Lambda_+ - \Lambda_-$. This is essentially the way of proceeding in the SSP context. It turns out that (3.26) is precisely the CFL coefficient (1.15) obtained in the SSP context.
4. As $\alpha + \Lambda\, e + \tilde\Lambda\, e = e$, with $\alpha \ge 0$, $\Lambda \ge 0$, and $\tilde\Lambda \ge 0$, the first inequality in (3.25) is also valid for a convex functional $\|\cdot\|$.

Example 8. We consider the representation (3.21)–(3.22) in Example 7 for the perturbed four-stage order four RK method. To split $\Lambda + \tilde\Lambda = (\lambda_{ij})$ in (3.21) we follow the zeros pattern in Γ and $\tilde\Gamma$ in (3.22); i.e., if $\beta_{ij} \ne 0$, we take $\alpha_{ij} = \lambda_{ij}$ and $\tilde\alpha_{ij} = 0$, and in a similar way if $\tilde\beta_{ij} \ne 0$. In this way we obtain

$$\Lambda = \begin{pmatrix} 0 & & & & \\ 1 & 0 & & & \\ 0 & 1/2 & 0 & & \\ 0 & 0 & 2/3 & 0 & \\ 0 & 1/3 & 1/3 & 1/3 & 0 \end{pmatrix}, \qquad \tilde\Lambda = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 1/2 & 0 & 0 & & \\ 1/9 & 2/9 & 0 & 0 & \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

It can be checked that condition (3.23) holds for $c = 2/3$.

Example 9. The numerically optimized four-stage order four RK method in [5, formula (3.4)] has

$$\Lambda+\tilde{\Lambda}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0\\ \frac{649}{1600} & \frac{951}{1600} & 0 & 0 & 0\\ \frac{53989}{2500000} & \frac{4806213}{20000000} & \frac{23619}{32000} & 0 & 0\\ \frac{1}{5} & \frac{6127}{30000} & \frac{7873}{30000} & \frac{1}{3} & 0 \end{pmatrix},$$

$\alpha=(1,0,0,0,0)^t$, and

$$\Gamma=\begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ \frac{1}{2} & 0 & 0 & 0 & 0\\ 0 & \frac{5000}{7873} & 0 & 0 & 0\\ 0 & 0 & \frac{7873}{10000} & 0 & 0\\ \frac{1}{10} & \frac{1}{6} & 0 & \frac{1}{6} & 0 \end{pmatrix},\qquad \tilde{\Gamma}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ \frac{10890423}{25193600} & 0 & 0 & 0 & 0\\ \frac{102261}{5000000} & \frac{5121}{20000} & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Following the zeros pattern in Γ and $\tilde{\Gamma}$, we can split $\Lambda+\tilde{\Lambda}$ into

$$\Lambda=\begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ 1 & 0 & 0 & 0 & 0\\ 0 & \frac{951}{1600} & 0 & 0 & 0\\ 0 & 0 & \frac{23619}{32000} & 0 & 0\\ \frac{1}{5} & \frac{6127}{30000} & \frac{7873}{30000} & \frac{1}{3} & 0 \end{pmatrix},\qquad \tilde{\Lambda}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ \frac{649}{1600} & 0 & 0 & 0 & 0\\ \frac{53989}{2500000} & \frac{4806213}{20000000} & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

It can be checked that condition (3.23) holds for $c=\frac{7487223}{8000000}\approx 0.936$. If we compute the matrices $\mathbb{A}$ and $\tilde{\mathbb{A}}$ from (3.20), we obtain a perturbed RK method $(\mathbb{A},\tilde{\mathbb{A}})$ with

$$\mathbb{A}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ \frac{1}{2} & 0 & 0 & 0 & 0\\ -\frac{2127}{15746} & \frac{5000}{7873} & 0 & 0 & 0\\ 0 & \frac{2127}{10000} & \frac{7873}{10000} & 0 & 0\\ \frac{1}{6} & \frac{12127}{30000} & \frac{7873}{30000} & \frac{1}{6} & 0 \end{pmatrix},\qquad \tilde{\mathbb{A}}=\begin{pmatrix} 0 & 0 & 0 & 0 & 0\\ 0 & 0 & 0 & 0 & 0\\ \frac{10890423}{25193600} & 0 & 0 & 0 & 0\\ \frac{869139357}{2560000000} & \frac{5121}{20000} & 0 & 0 & 0\\ \frac{580124399}{2560000000} & \frac{1707}{20000} & 0 & 0 & 0 \end{pmatrix}.$$

The matrix $\mathbb{A}$ is the four-stage order four method in the family (3.18) for $w=\frac{7873}{30000}$. It can be computed that $R(\mathbb{A},\tilde{\mathbb{A}})=\frac{7487223}{8000000}$. Observe that, for this value of w, the coefficient a_{31} in $\mathbb{A}$ is negative, and therefore, although $\tilde{\mathbb{A}}$ has two nontrivial columns, only the elements $\tilde{a}_{31}$, $\tilde{a}_{41}$, and $\tilde{a}_{51}$ in $\tilde{\mathbb{A}}$ are required in order to get $R(\mathbb{A},\tilde{\mathbb{A}})>0$.

Under some conditions on Λ and Γ, Proposition 3.9 ensures monotonicity under the stepsize restriction

$$h\le c\,\frac{1}{\rho}, \tag{3.27}$$

where c is such that $\Lambda - c\Gamma \geq 0$ and $\tilde{\Lambda} - c\tilde{\Gamma} \geq 0$. If we compare the stepsize restriction (3.27) with the stepsize restriction (3.13) obtained in terms of $R(\mathbb{A}, \tilde{\mathbb{A}})$, it is natural to wonder about the relationship between c in (3.27) and $R(\mathbb{A}, \tilde{\mathbb{A}})$. We begin expressing the concept of absolute monotonicity for the perturbed RK method in terms of the representation (3.19). The proof is omitted because it is straightforward.

LEMMA 3.10. *Consider a perturbed RK method* $(\mathbb{A}, \tilde{\mathbb{A}})$ *such that it can be factorized as* $\mathbb{A} = (I - (\Lambda + \tilde{\Lambda}))^{-1}(\Gamma - \tilde{\Gamma})$, $\tilde{\mathbb{A}} = (I - (\Lambda + \tilde{\Lambda}))^{-1}\tilde{\Gamma}$. *Then the method is absolutely monotonic at* ξ *if and only if*

1. *the matrix* $I - (\Lambda + \tilde{\Lambda} + \xi\,(\Gamma + \tilde{\Gamma}))$ *is invertible,*
2. *the following inequalities hold:*

$$\begin{aligned}
&(I - (\Lambda + \tilde{\Lambda} + \xi\,(\Gamma + \tilde{\Gamma}))^{-1}\Gamma \geq 0,\\
(3.28)\qquad &(I - (\Lambda + \tilde{\Lambda} + \xi\,(\Gamma + \tilde{\Gamma}))^{-1}\tilde{\Gamma} \geq 0,\\
&(I - (\Lambda + \tilde{\Lambda} + \xi\,(\Gamma + \tilde{\Gamma})))^{-1}(I - (\Lambda + \tilde{\Lambda}))\,e \geq 0.
\end{aligned}$$

The following result relates conditions in Proposition 3.9 with absolute monotonicity at a given point.

PROPOSITION 3.11. *Consider a perturbed RK method* $(\mathbb{A}, \tilde{\mathbb{A}})$. *Assume that it can be written in terms of* Λ, $\tilde{\Lambda}$, Γ, *and* $\tilde{\Gamma}$ *with*

$$(I - (\Lambda + \tilde{\Lambda}))\,e \geq 0, \qquad \Gamma \geq 0, \qquad \tilde{\Gamma} \geq 0, \qquad \Lambda + \tilde{\Lambda} - c\,(\Gamma + \tilde{\Gamma}) \geq 0,$$

and $I - (\Lambda + \tilde{\Lambda} - c\,(\Gamma + \tilde{\Gamma}))$ *invertible for some coefficient* $c \geq 0$. *Then the method is absolutely monotonic at* $-c$.

Proof. We will get the absolute monotonicity at $-c$ from Lemma 3.10. Some of the conditions imposed give part 1 in Lemma 3.10, and hence we simply have to check part 2. As $c\,(\Gamma + \tilde{\Gamma}) \geq 0$ and $(\Lambda + \tilde{\Lambda})\,e \leq e$, we obtain

$$(\Lambda + \tilde{\Lambda} - c\,(\Gamma + \tilde{\Gamma}))\,e \leq (\Lambda + \tilde{\Lambda})\,e \leq e.$$

This inequality, together with the assumption $\Lambda + \tilde{\Lambda} - c\,(\Gamma + \tilde{\Gamma}) \geq 0$ and the regularity of the matrix $I - (\Lambda + \tilde{\Lambda} - c\,(\Gamma + \tilde{\Gamma}))$, allows us to apply Lemma 3.8 with $B = \Lambda + \tilde{\Lambda} - c\,(\Gamma + \tilde{\Gamma})$, obtaining

$$(I - (\Lambda + \tilde{\Lambda} - c\,(\Gamma + \tilde{\Gamma})))^{-1} \geq 0.$$

In this way, as $\Gamma \geq 0$, $\tilde{\Gamma} \geq 0$, and $(I - (\Lambda + \tilde{\Lambda}))\,e \geq 0$, we get (3.28) and thus the desired result. □

From Proposition 3.11 we get that the CFL coefficient (3.23) obtained from a representation satisfies $c \leq R(\mathbb{A}, \tilde{\mathbb{A}})$. The next step is to study whether, given a perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$, there exists a representation such that the CFL coefficient (3.23) satisfies $c = R(\mathbb{A}, \tilde{\mathbb{A}})$. We will refer to these representations as optimal ones.

3.3. Optimal perturbed representations. In the next result we prove that optimal representations exist, and we show how to construct them.

PROPOSITION 3.12. *We consider a perturbed RK method* $(\mathbb{A}, \tilde{\mathbb{A}})$. *If* $r = R(\mathbb{A}, \tilde{\mathbb{A}}) > 0$, *then there exist matrices* $\Lambda, \tilde{\Lambda}, \Gamma, \tilde{\Gamma}$ *such that* $\mathbb{A} = (I - (\Lambda + \tilde{\Lambda}))^{-1}(\Gamma - \tilde{\Gamma})$, $\tilde{\mathbb{A}} = (I - (\Lambda + \tilde{\Lambda}))^{-1}\tilde{\Gamma}$, *with* $\Lambda + \tilde{\Lambda} \geq 0$, $\Gamma \geq 0$, $\tilde{\Gamma} \geq 0$, *and* $(I - (\Lambda + \tilde{\Lambda}))e \geq 0$, $I - (\Lambda + \tilde{\Lambda} - r(\Gamma + \tilde{\Gamma}))$ *invertible and* $\Lambda - r\Gamma \geq 0$, $\tilde{\Lambda} - r\tilde{\Gamma} \geq 0$.

Proof. Remember that $R(\mathbb{A}, \tilde{\mathbb{A}}) > 0$ implies that $\mathbb{A} + \tilde{\mathbb{A}} \geq 0$ and $\tilde{\mathbb{A}} \geq 0$. As $(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1} = I - r(\mathbb{A} + 2\tilde{\mathbb{A}})(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}$, we can write

$$\begin{aligned}
&\mathbb{A}+\tilde{\mathbb{A}}-r(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}(\mathbb{A}+\tilde{\mathbb{A}})=(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}(\mathbb{A}+\tilde{\mathbb{A}})\geq 0,\\
&\tilde{\mathbb{A}}-r(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}}=(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}}\geq 0,\\
&e-r(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}e=(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}e\geq 0.
\end{aligned}$$

Thus

$$r(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}(\mathbb{A}+\tilde{\mathbb{A}})\leq\mathbb{A}+\tilde{\mathbb{A}}, \tag{3.29}$$

$$r(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}}\leq\tilde{\mathbb{A}}, \tag{3.30}$$

$$r(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}e\leq e. \tag{3.31}$$

As $r>0$, conditions (3.5)–(3.6) give

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\geq 0.$$

Thus, taking this together with (3.31), we get

$$0\leq r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\leq\mathbb{E},$$

with $\mathbb{E}$ the matrix whose elements are all equal to one. We can take matrices $\Lambda+\tilde{\Lambda}$ such that

$$0\leq r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\leq\Lambda+\tilde{\Lambda}\leq\mathbb{E}. \tag{3.32}$$

If we multiply (3.32) by e, we obtain

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}e\leq(\Lambda+\tilde{\Lambda})\,e\leq s\,e.$$

As $e\leq s\,e$, inequality (3.31) gives

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}e\leq e\leq s\,e,$$

and hence we can impose for $(\Lambda+\tilde{\Lambda})e$ the condition

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}e\leq(\Lambda+\tilde{\Lambda})\,e\leq e. \tag{3.33}$$

If we now multiply (3.32) by $\mathbb{A}+\tilde{\mathbb{A}}\geq 0$, we obtain

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}(\mathbb{A}+\tilde{\mathbb{A}})\leq(\Lambda+\tilde{\Lambda})(\mathbb{A}+\tilde{\mathbb{A}})\ \leq\mathbb{E}\,(\mathbb{A}+\tilde{\mathbb{A}}).$$

As $\mathbb{A}+\tilde{\mathbb{A}}\leq\mathbb{E}\,(\mathbb{A}+\tilde{\mathbb{A}})$, inequality (3.29) gives

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}(\mathbb{A}+\tilde{\mathbb{A}})\leq\mathbb{A}+\tilde{\mathbb{A}}\ \leq\mathbb{E}\,(\mathbb{A}+\tilde{\mathbb{A}}),$$

and therefore we can impose for $(\Lambda+\tilde{\Lambda})(\mathbb{A}+\tilde{\mathbb{A}})$ the condition

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}(\mathbb{A}+\tilde{\mathbb{A}})\leq(\Lambda+\tilde{\Lambda})(\mathbb{A}+\tilde{\mathbb{A}})\leq\mathbb{A}+\tilde{\mathbb{A}}. \tag{3.34}$$

Finally, if we multiply (3.32) by $\tilde{\mathbb{A}}\geq 0$, we obtain

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}}\leq(\Lambda+\tilde{\Lambda})\tilde{\mathbb{A}}\ \leq\mathbb{E}\,\tilde{\mathbb{A}}.$$

As $\tilde{\mathbb{A}}\leq\mathbb{E}\,\tilde{\mathbb{A}}$, inequality (3.30) gives

$$r\,(\mathbb{A}+2\tilde{\mathbb{A}})(I+r(\mathbb{A}+2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}}\leq\tilde{\mathbb{A}}\ \leq\mathbb{E}\,\tilde{\mathbb{A}}.$$

We can impose for $(\Lambda + \tilde{\Lambda})\tilde{\mathbb{A}}$ the condition

$$r\,(\mathbb{A} + 2\tilde{\mathbb{A}})(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}} \leq (\Lambda + \tilde{\Lambda})\tilde{\mathbb{A}} \leq \tilde{\mathbb{A}}. \tag{3.35}$$

It is possible to impose conditions (3.32)–(3.35). For example, we can take

$$\Lambda = r\,(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}(\mathbb{A} + \tilde{\mathbb{A}}), \qquad \tilde{\Lambda} = r\,(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}}. \tag{3.36}$$

In this case, the matrix $I - (\Lambda + \tilde{\Lambda})$ is invertible. Some other choices of $\Lambda + \tilde{\Lambda}$ are possible. Once that $\Lambda + \tilde{\Lambda}$ has been chosen satisfying (3.32)–(3.35), with $I - (\Lambda + \tilde{\Lambda})$ invertible, we define

$$\Gamma = \mathbb{A} + \tilde{\mathbb{A}} - (\Lambda + \tilde{\Lambda})(\mathbb{A} + \tilde{\mathbb{A}}), \qquad \tilde{\Gamma} = \tilde{\mathbb{A}} - (\Lambda + \tilde{\Lambda})\tilde{\mathbb{A}}, \qquad \alpha = e - (\Lambda + \tilde{\Lambda})e. \tag{3.37}$$

Respectively from (3.34), (3.35), and (3.33) we get $\Gamma \geq 0$, $\tilde{\Gamma} \geq 0$, and $(\Lambda + \tilde{\Lambda})e \leq e$.

In order to prove that $\Lambda + \tilde{\Lambda} - r(\Gamma + \tilde{\Gamma}) \geq 0$, we compute

$$(\Lambda + \tilde{\Lambda} - r(\Gamma + \tilde{\Gamma}))(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1} = \Lambda + \tilde{\Lambda} - r(\mathbb{A} + 2\tilde{\mathbb{A}})(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1} \geq 0,$$

where we have used (3.32). If we right-multiply this expression by $I + r(\mathbb{A} + 2\tilde{\mathbb{A}}) \geq 0$, we obtain that

$$\Lambda + \tilde{\Lambda} - r(\Gamma + \tilde{\Gamma}) \geq 0. \tag{3.38}$$

If we choose Λ and $\tilde{\Lambda}$ from (3.36), we obtain $\Lambda - r\Gamma = 0$, $\tilde{\Lambda} - r\tilde{\Gamma} = 0$. Otherwise, we have to prove that

$$\Lambda - r\Gamma \geq 0 \quad \text{and} \quad \tilde{\Lambda} - r\tilde{\Gamma} \geq 0. \tag{3.39}$$

Observe that so far the different conditions imposed and the definitions of Γ, $\tilde{\Gamma}$, and α depend on $\Lambda + \tilde{\Lambda}$. Thus, once the matrix $\Lambda + \tilde{\Lambda}$ has been determined, we have to obtain Λ and $\tilde{\Lambda}$ such that (3.39) holds. We proceed componentwise as follows. We take, for example, $\alpha_{ij} = r\beta_{ij}$; in this way $(\Lambda - r\Gamma)_{ij} = 0$, and using (3.38), we have

$$\tilde{\alpha}_{ij} - r\tilde{\beta}_{ij} = \alpha_{ij} + \tilde{\alpha}_{ij} - r\tilde{\beta}_{ij} - r\beta_{ij} \geq 0.$$

We can alternatively take $\tilde{\alpha}_{ij} = r\tilde{\beta}_{ij}$ to get $(\tilde{\Lambda} - r\tilde{\Gamma})_{ij} = 0$ and $\alpha_{ij} - r\beta_{ij} \geq 0$.

Finally, remember that $r = R(\mathbb{A}, \tilde{\mathbb{A}})$ implies the invertibility of $I + r(\mathbb{A} + 2\tilde{\mathbb{A}})$, and hence by Lemma 3.10 the invertibility of $I - (\Lambda - r\Gamma)$. □

Remark 6.

1. We can construct matrices Λ, $\tilde{\Lambda}$, Γ, and $\tilde{\Gamma}$ such that $\Lambda - r\Gamma = 0$ and $\tilde{\Lambda} - r\tilde{\Gamma} = 0$, namely Λ and $\tilde{\Lambda}$ from (3.36), and Γ, $\tilde{\Gamma}$, and α from (3.37).
2. Inequality (3.32) has been used to prove (3.38). However, when a representation is constructed, if we ensure (3.38), we can drop (3.32).

Example 10. We consider again the classical four-stage fourth order RK method and its perturbation (3.8) with $R(\mathbb{A}, \tilde{\mathbb{A}}) = 2/3$ (see Example 5). According to Proposition 3.12, if we construct Λ and $\tilde{\Lambda}$ from (3.36), and Γ, $\tilde{\Gamma}$, and α from (3.37), we obtain Λ, $\tilde{\Lambda}$, Γ, and $\tilde{\Gamma}$ such that $\Lambda - 2/3\Gamma = \tilde{\Lambda} - 2/3\tilde{\Gamma} = 0$, and thus the CFL coefficient (3.23) is also 2/3. Simple computations give that

$$\Lambda = \begin{pmatrix} 0 & & & & \\ 1/3 & 0 & & & \\ 1/18 & 1/3 & 0 & & \\ 0 & 0 & 2/3 & 0 & \\ 5/54 & 2/9 & 4/27 & 1/9 & 0 \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} 0 & & & & \\ 1/2 & 0 & & & \\ 1/12 & 1/2 & 0 & & \\ 0 & 0 & 1 & 0 & \\ 5/36 & 1/3 & 2/9 & 1/6 & 0 \end{pmatrix},$$

$$\tilde{\Lambda} = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 1/6 & 0 & 0 & & \\ 2/27 & 2/9 & 0 & 0 & \\ 35/486 & 4/81 & 0 & 0 & 0 \end{pmatrix}, \qquad \tilde{\Gamma} = \begin{pmatrix} 0 & & & & \\ 0 & 0 & & & \\ 1/4 & 0 & 0 & & \\ 1/9 & 1/3 & 0 & 0 & \\ 35/324 & 2/27 & 0 & 0 & 0 \end{pmatrix},$$

and $\alpha = (1, 2/3, 4/9, 1/27, 74/243)^t$. From $\mathbb{A} = (I - (\Lambda + \tilde{\Lambda}))^{-1}(\Gamma - \tilde{\Gamma})$, we get the factorization given in (2.6), i.e., $\Lambda_2 = \Lambda + \tilde{\Lambda}$ and $\Gamma_2 = \Gamma - \tilde{\Gamma}$. However, from the implementation point of view, such a representation is not optimal as it requires one to store $f(u_n)$, $f(U_2)$, $f(U_3)$, $f(U_4)$, $\tilde{f}(u_n)$, and $\tilde{f}(U_2)$ until the computation of u_{n+1}. The computational cost can be decreased if the matrices Γ and $\tilde{\Gamma}$ have as many zeros as possible and the nonzeros appear in optimal positions. We will try to impose $\tilde{\beta}_{ij} = 0$ whenever $\beta_{ij} \neq 0$, and vice versa: $\beta_{ij} = 0$ whenever $\tilde{\beta}_{ij} \neq 0$.

Denoting the elements of $\Lambda + \tilde{\Lambda}$ by λ_{ij}, if we compute inequality (3.34), we obtain that $\lambda_{43} = 2/3$ and $\lambda_{42} = 2/9$. Next we construct $\Gamma = (\beta_{ij})$ and $\tilde{\Gamma} = (\beta_{ij})$ from (3.37). As $\tilde{\beta}_{31} \neq 0$, we impose $\beta_{31} = 0$, which gives $\lambda_{32} = 1/2$. We can get $\tilde{\beta}_{52} = 0$ if we set $\lambda_{54} = 1/3$. Similarly, we can obtain $\tilde{\beta}_{51} = 0$, imposing $\lambda_{53} = 1/3$. Finally we can get $\beta_{51} = 0$, imposing $\lambda_{52} = 1/3$. If we impose $\alpha = (1, 0, 0, 0, 0)^t$, we get $\lambda_{21} = 1$, $\lambda_{31} = 1/2$, $\lambda_{41} = 1/9$, and $\lambda_{51} = 0$. So far we have obtained the matrix $\Lambda + \tilde{\Lambda}$ in (3.21), and the matrices Γ and $\tilde{\Gamma}$ in (3.22) in Example 7.

They satisfy (3.38), although condition (3.32) does not hold (see Remark 6). We can take into account the zeros distribution in Γ and $\tilde{\Gamma}$ to split out $\Lambda + \tilde{\Lambda}$ into Λ and $\tilde{\Lambda}$ (see Example 8). We have that $\Lambda - 2/3\Gamma \geq 0$, $\tilde{\Lambda} - 2/3\tilde{\Gamma} \geq 0$. This is the representation (1.16) taken from [13].

We finish the section showing how the theory developed in this paper applies to the schemes in [12].

Example 11. In [12], explicit RK schemes are represented in terms of coefficient matrices $\Lambda = (\alpha_{ij}) \geq 0$ and $\Gamma = (\beta_{ij})$. The matrix Γ is such that the nonzero coefficients β_{ik} for a given k are all of the same sign. If a sign splitting is done for Γ, i.e., $\Gamma = \Gamma_+ - \Gamma_-$, the RK method is applied as follows:

$$U = \alpha \otimes u_n + h(\Lambda \otimes I)U + h(\Gamma_+ \otimes I)F(U) - h(\Gamma_- \otimes I)\tilde{F}(U), \tag{3.40}$$

where $\alpha = (1, 0, \ldots, 0)^t$. The same sign splitting can be done in Λ, $\Lambda = \Lambda_+ - \Lambda_-$, and then the CFL coefficient in the SSP context (1.15) is the same as that given in expression (3.23) of Proposition 3.9 (see Remark 5). Scheme (3.40) can also be written as

$$U = e \otimes u_n + h(\mathbb{A}_+ \otimes I)F(U) - h(\mathbb{A}_- \otimes I)\tilde{F}(U), \tag{3.41}$$

where $\mathbb{A}_+ = (I - \Lambda)^{-1}\Gamma_+$ and $\mathbb{A}_- = (I - \Lambda)^{-1}\Gamma_-$. Observe that by Lemma 3.8 we have $\mathbb{A}_+ \geq 0$ and $\mathbb{A}_- \geq 0$. Observe too that (3.41) can also be formulated as a perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$, with $\mathbb{A} = \mathbb{A}_+ - \mathbb{A}_-$ and $\tilde{\mathbb{A}} = \mathbb{A}_-$. The sign distribution of the matrix Γ gives that $\mathbb{A}_+$ and $\mathbb{A}_-$ are a sign splitting of $\mathbb{A}$. As has been pointed out above, the coefficient c computed from (3.23) (or (1.15) from the SSP context) satisfies $c \leq R(\mathbb{A}, \mathbb{A}_-)$.

Three optimal fifth order methods, denoted by SSP(7,5), SSP(8,5), and SSP(9,5), are constructed in [12]. For these methods it has been checked that the CFL coefficient c given in [12], computed from (1.15), is the radius of absolute monotonicity of the perturbed method $(\mathbb{A}, \mathbb{A}_-)$, i.e., $c = R(\mathbb{A}, \mathbb{A}_-)$. We remark that for these methods only one column in the matrix $\mathbb{A}_-$ is different from zero. This is enough to obtain schemes with $R(\mathbb{A}, \mathbb{A}_-) > 0$.

4. Conclusions and future work. In this paper we have extended the Shu–Osher representations to general RK methods, and we have related the stepsize restrictions for monotonicity with the radius of absolute monotonicity giving optimal representations. The case of Shu–Osher representations with positive coefficients corresponds to the case $R(\mathbb{A}) > 0$. Optimal representations are given by

$$\Lambda = r(I - r\mathbb{A})^{-1}\mathbb{A}, \qquad \Gamma = (I - \Lambda)\mathbb{A}, \qquad \alpha = (I - \Lambda)e,$$

with $r = R(\mathbb{A})$. In this case $\Lambda - r\Gamma = 0$.

The case of Shu–Osher representations with negative coefficients corresponds to the case $R(\mathbb{A}) = 0$. To deal with this case, we have interpreted the numerical integration as perturbations of the original RK method $\mathbb{A}$ with some coefficients $\tilde{\mathbb{A}}$, and we have referred to $(\mathbb{A}, \tilde{\mathbb{A}})$ as perturbed RK methods. These perturbed RK methods have perturbed representations.

A new definition of radius of absolute monotonicity for the perturbed RK method $(\mathbb{A}, \tilde{\mathbb{A}})$ has been given, and some of its properties have been investigated. In this way it is possible to have a positive generalized radius $R(\mathbb{A}, \tilde{\mathbb{A}}) > 0$ if RK methods with $R(\mathbb{A}) = 0$ are used. For $R(\mathbb{A}, \tilde{\mathbb{A}}) > 0$, we obtain monotonicity under nontrivial stepsize restrictions. Optimal representations are given by

$$\begin{aligned}
&\Lambda = r\,(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}(\mathbb{A} + \tilde{\mathbb{A}}), &&\tilde{\Lambda} = r\,(I + r(\mathbb{A} + 2\tilde{\mathbb{A}}))^{-1}\tilde{\mathbb{A}},\\
&\Gamma = \mathbb{A} + \tilde{\mathbb{A}} - (\Lambda + \tilde{\Lambda})(\mathbb{A} + \tilde{\mathbb{A}}), &&\tilde{\Gamma} = \tilde{\mathbb{A}} - (\Lambda + \tilde{\Lambda})\tilde{\mathbb{A}},\\
&\alpha = e - (\Lambda + \tilde{\Lambda})e,
\end{aligned}$$

with $r = R(\mathbb{A}, \tilde{\mathbb{A}})$. In this case $\Lambda - r\Gamma = 0$ and $\tilde{\Lambda} - r\tilde{\Gamma} = 0$.

With regard to the new concept $R(\mathbb{A}, \tilde{\mathbb{A}})$, at least two questions remain open for future work. The first one concerns the possibility of getting $R(\mathbb{A}, \tilde{\mathbb{A}}) = \infty$ and hence unconditional monotonicity for high order methods. The second one is how to construct $\tilde{\mathbb{A}}$ for a given RK method $\mathbb{A}$ so that $R(\mathbb{A}, \tilde{\mathbb{A}})$ is as large as possible.

Acknowledgment. The author thanks the anonymous referees for their valuable remarks and comments on the paper.

REFERENCES

[1] K. Dekker and J. G. Verwer, *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations*, CWI Monogr. 2, North–Holland, Amsterdam, 1984.

[2] L. Ferracina and M. N. Spijker, *Stepsize restrictions for the total-variation-diminishing property in general Runge–Kutta methods*, SIAM J. Numer. Anal., 42 (2004), pp. 1073–1093.

[3] L. Ferracina and M. N. Spijker, *An extension and analysis of the Shu–Osher representation of Runge–Kutta methods*, Math. Comp., 74 (2005), pp. 201–219.

[4] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[5] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge–Kutta schemes*, Math. Comp., 67 (1998), pp. 73–85.

[6] S. Gottlieb and L. J. Gottlieb, *Strong stability preserving properties of Runge–Kutta time discretization methods for linear constant coefficient operators*, J. Sci. Comput., 18 (2003), pp. 83–109.

[7] J. F. B. M. Kraaijevanger, *Contractivity of Runge–Kutta methods*, BIT, 31 (1991), pp. 482–528.

[8] J. D. Lambert, *Numerical Methods for Ordinary Differential Systems*, Wiley, New York, 1991.

[9] P. Lancaster and M. Tismenetsky, *The Theory of Matrices*, Academic Press, San Diego, CA, 1985.

[10] R. H. Martin, Jr., *Nonlinear Operators and Differential Equations in Banach Spaces*, Wiley, New York, 1976.

[11] S. J. Ruuth and R. J. Spiteri, *Two barriers on strong stability preserving time discretization methods*, J. Sci. Comput., 17 (2002), pp. 211–220.

[12] S. J. Ruuth and R. J. Spiteri, *High-order strong-stability-preserving Runge–Kutta methods with downwind-biased spatial discretizations*, SIAM J. Numer. Anal., 42 (2004), pp. 974–996.

[13] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.

[14] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 1073–1084.

[15] C.-W. Shu, *A survey of strong stability preserving high order time discretizations*, in Collected Lectures on the Preservation of Stability under Discretization, D. Estep and T. Tavener, eds., Proc. Appl. Math. 109, SIAM, Philadelphia, 2002, pp. 51–65.

[16] R. J. Spiteri and S. J. Ruuth, *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal., 40 (2002), pp. 469–491.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 949–969

A STABILIZED MIXED FINITE ELEMENT METHOD FOR ELLIPTIC SYSTEMS OF FIRST ORDER*

MANFRED DOBROWOLSKI[†] AND MANUEL VILLEGAS[‡]

Abstract. A quasilinear elliptic equation of second order can be split into a first order system in various ways. We present and analyze a stabilized finite element method for the system, which is well suited for any of these possible splittings. Under minimal assumptions on the continuous solution, existence and (nearly) optimal convergence in L^∞ of the discrete solutions is established. This result holds for any choice of the stabilization parameter $\omega > 0$. Moreover, the paper presents a framework for investigating other mixed methods for unsymmetric first order systems.

Key words. finite elements, elliptic systems

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/S0036142902408829

1. Introduction.

Let $\Omega \subset \mathbb{R}^n$, $n = 2, 3$, be a bounded domain. Consider the first order system with unknown functions $u = (u^1, \ldots, u^n)$ and p,

$$u^\alpha = b^\alpha(x, p, \nabla p) \quad \text{in } \Omega, \quad \alpha = 1, \ldots, n, \tag{1.1}$$

$$\partial_i c_i(x, p, u) = g(x, p, \nabla p) \quad \text{in } \Omega, \quad p = 0 \quad \text{on } \partial\Omega. \tag{1.2}$$

Here and in the following we use the usual summation convention for small Latin ($i, j = 0, \ldots, n$, $\partial_0 := 1$) and Greek ($\alpha, \beta = 1, \ldots, n$) indices. It is assumed that the functions b^α, g, c_i are sufficiently smooth with respect to the arguments u, p, and $q = \nabla p$. The system is equivalent to the second order equation

$$\partial_i c_i(x, p, b(x, p, \nabla p)) = g(x, p, \nabla p), \tag{1.3}$$

which is quasilinear; i.e., it is linear in the second derivatives of p. The coefficient function of the main part of (1.3) is given by

$$a_{ij} = c_i^\alpha b_j^\alpha, \quad i, j = 1, \ldots, n, \tag{1.4}$$

with

$$c_i^\alpha(x, p, u) = \frac{\partial}{\partial u^\alpha} c_i(x, p, u), \quad \alpha = 1, \ldots, n,$$

$$b_j^\alpha(x, p, q) = \frac{\partial}{\partial q^j} b^\alpha(x, p, q), \quad j = 1, \ldots, n.$$

*Received by the editors May 30, 2002; accepted for publication (in revised form) August 13, 2003; published electronically September 20, 2005. This research was supported by the Deutsche Akademische Austauschdienst (DAAD).

http://www.siam.org/journals/sinum/43-3/40882.html

[†] Institut für Angewandte Mathematik und Statistik, Universität Würzburg, Am Hubland, 97074 Würzburg, Germany (dobro@mathematik.uni-wuerzburg.de).

[‡] Unidad Académica de Matemáticas, Universidad del Valle, A.A. 25360 Cali, Colombia (mvillegas @univalle.edu.co).

Now we can formulate our existence and ellipticity condition, as follows.

ASSUMPTION A. *There exists a solution* $(u, p) \in C^0(\overline{\Omega})^n \times C^1(\overline{\Omega})$ *of* (1.1), (1.2) *such that the coefficient function*

$$a_{ij}(x) := a_{ij}(x, u(x), p(x), \nabla p(x))$$

in (1.4) *satisfies*

$$m_0|\xi|^2 \leq \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j \leq M_0|\xi|^2 \quad \forall \xi \in \mathbb{R}^n \tag{1.5}$$

with constants $m_0, M_0 > 0$ *independent of* $x \in \Omega$.

Note that (1.5) is a purely local condition. Hence, no growth conditions for the functions c_i^α, b_j^α are required.

Our next condition is concerned with the linearized operator of (1.1), (1.2). Writing

$$b_0^\alpha = \frac{\partial}{\partial p} b^\alpha, \quad c_i^0 = \frac{\partial}{\partial p} c_i,$$

$$g_0 = \frac{\partial}{\partial p} g, \quad g_i = \frac{\partial}{\partial q_i} g \quad (g = g(x, p, q)),$$

we define the linear first order operators

$$Bq(x) = (b_i^\alpha \partial_i q(x))_{\alpha=1,\ldots,n,}$$

$$C_i(v(x), q(x)) = c_i^0 q(x) + c_i^\alpha v^\alpha(x),$$

$$Dq(x) = g_i \partial_i q(x).$$

Note that b_i^α, c_i^α are evaluated in $p, \nabla p$ or p, u. The linearized operator of (1.1), (1.2) can be written in weak form. For

$$(v, q), (\varphi, \psi) \in L^2(\Omega)^n \times H_0^{1,2}(\Omega)$$

we define the bilinear form

$$a((v, q), (\varphi, \psi)) = (v, \varphi) - (Bq, \varphi) - (C_i(v, q), \partial_i \psi) - (Dq, \psi).$$

Obviously, the operator L corresponding to $a(\cdot, \cdot)$ is a continuous mapping

$$L : L^2(\Omega)^n \times H_0^{1,2}(\Omega) \to (L^2(\Omega)^n \times H_0^{1,2}(\Omega))'.$$

Here and below we use the usual Lebesgue and Sobolev spaces with norms

$$\|v\|_{H^{m,p}(\Omega)} = \|v\|_{m,p} = \left(\sum_{i=0}^{m} \|\nabla^i v\|_p^p \right)^{1/p}.$$

Theorem 1.1. *Assume that condition* A *holds. Then the operator L satisfies Fredholm's alternative: Either the problem*

$$L(v, q) = (f, g) \tag{1.6}$$

is uniquely solvable for all $(f, g) \in (L^2(\Omega)^n \times H_0^{1,2}(\Omega))'$ or $\lambda = 0$ is an eigenvalue of L.

This theorem is a simple consequence of the ellipticity condition (1.5) and will be proved in section 2. Usually, the second alternative in Theorem 1.1 signifies that (u, p) lies in a singularity of the bifurcation diagram of (1.1), (1.2). In this case, however, the system cannot be discretized by a direct finite element method. Hence, we state our next condition.

Assumption B. *$\lambda = 0$ is not an eigenvalue of the linearized operator L.*

In view of the fact that L can also be written as a second order elliptic operator, H^2-regularity can be assumed in the following form.

Assumption C. *For $(f, g) \in H^{1,2}(\Omega)^n \times L^2(\Omega)$ the solution of* (1.6) *satisfies $(v, q) \in H^{1,2}(\Omega)^n \times H^{2,2}(\Omega)$ and*

$$||v||_{1,2} + ||q||_{2,2} \leq c\{||f||_{1,2} + ||g||_2\}.$$

Let us describe our finite element method for approximating the system (1.1), (1.2). Let $S^m \subset C^0(\overline{\Omega})$ be a standard Lagrangian finite element space (see [9]) containing the polynomials of degree $\leq m$ on each element Λ of a quasi-regular subdivision Π of $\overline{\Omega}$, possibly isoperimetrically modified at the boundary $\partial\Omega$. It is assumed that each element Λ is contained in a ball of radius h and contains a ball of radius ch. Here and in the following the generic constant c does not depend on the mesh parameter h. Set $S_0^m = S^m \cap H_0^{1,2}$. The discrete solution (u_h, p_h) will be defined in the space $X_h \times Y_h$ with $X_h = (S^l)^n$, $Y_h = S_0^m$. In view of the fact that u is coupled with ∇p, we assume that $l \geq m - 1$. $(u_h, p_h) \in X_h \times Y_h$ is defined by

$$(u_h, \varphi_h) = (b(\cdot, p_h, \nabla p_h), \varphi_h) \quad \forall \varphi_h \in X_h, \tag{1.7}$$

$$\begin{aligned}(\partial_i c_i(\cdot, p_h, u_h), \psi_h) + \omega(u_h - b(\cdot, p_h, \nabla p_h), B(p_h)\psi_h)& \\ = (g(\cdot, p_h, \nabla p_h), \psi_h) \quad \forall \psi_h \in Y_h.&\end{aligned} \tag{1.8}$$

Here, $\omega > 0$ is the (uncritical) *stabilization parameter* and

$$B(p_h)\psi_h(x) = (b_i^\alpha(x, p_h(x), \nabla p_h(x))\partial_i \psi_h(x))_{\alpha=1,\ldots,n}$$

according to our notation introduced above. Note that, in contrast to stabilized mixed methods for the Stokes equation (see [7]), the stabilizing term is conforming; i.e., the solution (u, p) of the continuous system satisfies (1.7), (1.8) as well.

Now we can state the main result of this paper.

Theorem 1.2. *Assume that Assumptions* A, B, C *hold. Assume further that the solution (u, p) of* (1.1), (1.2) *is in the space $H^{m,\infty}(\Omega)^n \times H^{m+1,\infty}(\Omega)$, and that the stabilization parameter $\omega > 0$ is chosen independent of h. Then there exists a $h_0 > 0$ such that for $0 < h \leq h_0$ the discrete system* (1.7), (1.8) *has a local solution in a neighborhood of (u, p) which satisfies the error estimate*

$$h||u - u_h||_\infty + ||p - p_h||_\infty \leq ch^{m+1-\epsilon}\{||u||_{m,\infty} + ||p||_{m+1,\infty}\}$$

for all $\epsilon > 0$ with a constant $c = c(\epsilon)$.

The proof of this theorem will be given in the next sections.

Since our results are also new in the linear case, we study the corresponding linear problem

$$(u,\varphi) + b(p,\varphi) = f(\varphi), \tag{1.9}$$

$$c(u,\psi) = g(\psi), \tag{1.10}$$

where the bilinear forms $b(\cdot,\cdot)$, $c(\cdot,\cdot)$ correspond to linear first order operators. Again, we do not assume that $b(\psi,\varphi) = c(\varphi,\psi)$. The problem (1.9), (1.10) is discretized by the method analogous to (1.7), (1.8),

$$(u_h,\varphi_h) + b(p_h,\varphi_h) = f(\varphi_h) \quad \forall \varphi_h \in X_h, \tag{1.11}$$

$$c(u_h,\psi_h) + \omega(u_h + Bp_h, B\psi_h) = g(\psi_h) + \omega f(B\psi_h) \quad \forall \psi_h \in Y_h, \tag{1.12}$$

where $(Bp,\varphi) = b(p,\varphi)$. It is now obvious that the stabilization term is the first variation of

$$\|u + Bp - f\|_2$$

with respect to p. Hence, the method is a Galerkin least squares method already considered in [13] for the elasticity problem. It is the object of this paper to demonstrate that the Galerkin least squares approach allows nearly arbitrary, in particular unsymmetric, splittings of the second order equation. For the classical and other least squares methods, we refer to [8] and [3].

With a proof similar to that for Theorem 1.2 we can show the following.

Theorem 1.3. *Assume that* (1.5) *and Assumptions* B *and* C *hold. Assume further that the solution* (u,p) *of* (1.9), (1.10) *is in the space* $H^{m,2}(\Omega)^n \times H^{m+1,2}(\Omega)$ *and that the stabilization parameter* $\omega > 0$ *is chosen independent of* h*. Then there exists a* $h_0 > 0$ *such that for* $0 < h \leq h_0$ *the discrete system* (1.11), (1.12) *has a solution which satisfies the error estimate*

$$h||u-u_h||_2 + ||p-p_h||_2 \leq ch^{m+1}\{||u||_{m,2} + ||p||_{m+1,2}\}.$$

We remark that we have assumed only that $\lambda = 0$ is not an eigenvalue of the continuous system. This carries over to the discretized system only for sufficiently small h. Thus, the condition $0 < h \leq h_0$ cannot be removed in the linear case.

A further object of this paper is to present a framework for analyzing other mixed methods for first order systems of the type (1.9), (1.10). The classical mixed finite element theory (see, e.g., [4], [6]) requires a symmetry condition for the forms $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$, but what can be done if this condition is violated? Our improvement in the analysis of first order systems is Lemma 2.1, which states that any elliptic first order system can be transformed into a first order system with "nearly" adjoint forms $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$. Using this lemma, it is a simple exercise to prove convergence for the classical mixed elements introduced in [22], [5] for general first order systems. Mixed methods for unsymmetric systems are also treated in [18], but the methodology presented there is different. The idea of Lemma 2.1 can also be applied to generalized Stokes equations [11].

Mixed methods for nonlinear first order systems are also treated in [17], [20], [16]. In all of these papers, only the symmetric case is considered. Moreover, the convergence proofs are based on L^2-estimates, which restricts the results to space dimension $n = 2$ or higher order elements.

Theorem 1.2 states that for an elliptic second order equation

$$\partial_i F_i(x, p, \nabla p) = 0 \tag{1.13}$$

any splitting of the form (1.1), (1.2) is allowed. For example, the coefficient function a_{ij} in (1.4) of Poisson's equation

$$\Delta p = g$$

is the identity matrix. If we use the splitting

$$u^1 = \partial_1 p + \partial_2 p, \quad u^2 = -\partial_1 p + \partial_2 p, \quad \partial_1 u^1 + \partial_2 u^2 = g,$$

we again obtain Poisson's equation, but now with coefficient matrix

$$A = \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}.$$

In contrast to the direct finite element method for the second order equation, the corresponding mixed method is different from the mixed method of the standard splitting

$$u = \nabla p, \quad \partial_1 u^1 + \partial_2 u^2 = g.$$

More natural splittings for (1.13) are

$$u = \nabla p, \quad \partial_i F_i(x, p, u) = 0$$

and

$$u^i = F_i(x, p, \nabla p) \ \text{ for } i = 1, \dots, n, \quad \sum_{i=1}^{n} \partial_i u^i = -F_0(x, p, \nabla p).$$

Numerical examples for both splittings are given in [23]. These splittings also give an idea of how to discretize the general elliptic equation

$$F(x, q, \nabla q, \nabla^2 q) = 0,$$

which can be treated by

$$u = \nabla p, \quad F(x, p, u, \nabla u) = 0.$$

We remark that in this case our convergence proof works only for higher order elements, i.e., $m \geq 2$.

2. Linear systems of first order. We consider the system

$$v^\alpha - b_i^\alpha \partial_i q = f^\alpha \ \text{ in } \Omega, \quad \alpha = 1, \dots, n, \tag{2.1}$$

$$\partial_i (c_i^\alpha v^\alpha) + d_i \partial_i q = g \ \text{ in } \Omega, \quad q = 0 \ \text{ on } \partial\Omega, \tag{2.2}$$

where the coefficient functions $b_i^\alpha, c_i^\alpha, d_i$ are assumed to be smooth functions of $x \in \Omega$. Introducing the bilinear forms

$$\begin{aligned} b(q,\varphi) &= -\left(b_i^\alpha \partial_i q, \varphi^\alpha\right), \\ c(v,\psi) &= -\left(\sum_{i=1}^n c_i^\alpha v^\alpha, \partial_i \psi\right) + \left(c_0^\alpha v^\alpha, \psi\right), \\ d(q,\psi) &= (d_i \partial_i q, \psi), \end{aligned}$$

the weak solution $(v,q) \in L^2(\Omega)^n \times H_0^{1,2}(\Omega)$ is defined by

$$(v,\varphi) + b(q,\varphi) = (f,\varphi) \quad \forall \varphi \in L^2(\Omega)^n, \tag{2.3}$$

$$c(v,\psi) + d(q,\psi) = (g,\psi) \quad \forall \psi \in H_0^{1,2}(\Omega). \tag{2.4}$$

The characteristic problem in the finite element analysis of (2.1), (2.2) is that we have the condition of ellipticity in (1.5), but we need conditions for the forms $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$. Therefore, the following lemma is the basis of the finite element analysis of (2.1), (2.2) and the corresponding nonlinear case in (1.7), (1.8).

LEMMA 2.1. *The following statements are equivalent:*

(i) *The system* (2.1), (2.2) *is elliptic in the sense of* (1.5); *i.e., the coefficients*

$$a_{ij} = c_i^\alpha b_j^\alpha, \quad i,j = 1,\dots,n, \tag{2.5}$$

satisfy

$$m_0|\xi|^2 \le a_{ij}(x)\xi_i\xi_j \le M_0|\xi|^2 \quad \forall \xi \in \mathbb{R}^n. \tag{2.6}$$

(ii) *The following two conditions hold:*

(a) *There exists a matrix $M(x) \in \mathbb{R}^{n\times n}$ which is as smooth as the coefficient functions b_i^α, c_i^α such that with a constant $m_1 > 0$*

$$m_1||v||_2^2 \le (v, Mv) \quad \forall v \in L^2(\Omega)^n \tag{2.7}$$

and

$$|b(q,Mv) - c(v,q)| \le c||v||_2||q||_2, \quad \forall v \in L^2(\Omega)^n,\ \forall q \in H_0^{1,2}(\Omega). \tag{2.8}$$

(b) *Setting*

$$Bq = (b_i^\alpha \partial_i q)_{\alpha=1,\dots,n},$$

we have with a constant $m_2 > 0$

$$m_2||\nabla q||^2 \le ||Bq||_2^2 + c||q||_2^2 \quad \forall q \in H_0^{1,2}(\Omega). \tag{2.9}$$

Remarks. (i) Condition (2.7) is satisfied if M is uniformly real positive. (ii) The forms $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are of first order, but the right-hand side in (2.8) is not. The intuitive meaning of condition (2.8) is that (2.3) can be transformed by the matrix M such that the transformed form of $b(\cdot,\cdot)$ is adjoint to $c(\cdot,\cdot)$ in the leading terms; i.e.,

$$b(q,Mv) = c(v,q) + l.o.t.$$

Proof. Introducing the matrices

$$\tilde{C} = (c_i^\alpha)_{i,\alpha=1,\dots,n}, \quad \tilde{B} = (b_i^\alpha)_{i,\alpha=1,\dots,n}, \quad A = (a_{ij})_{i,j=1,\dots,n},$$

formula (2.5) can be written in the form

$$A = \tilde{C}\tilde{B}^T. \tag{2.10}$$

In the next step, we derive a representation of the matrix M from condition (2.8). Since it is stated that M is as smooth as the coefficient functions, we can assume that at least $M \in C(\overline{\Omega})^{n\times n}$. Let $x_0 \in \Omega$, $B_\epsilon(x_0) \subset \Omega$, and $\varphi \in C_0^\infty(B_\epsilon(x_0))$ with $||\varphi||_2 = 1$. For $r, s \in \mathbb{R}^n$ set ($\mathbf{i} = \sqrt{-1}$)

$$q(x) = e^{2\pi \mathbf{i} r\cdot x}\varphi(x), \quad v(x) = e^{-2\pi \mathbf{i} r\cdot x}\varphi(x)s.$$

Then

$$\nabla q(x) = 2\pi \mathbf{i} r q(x) + e^{2\pi \mathbf{i} r\cdot x}\nabla\varphi(x),$$

and hence

$$b(q, Mv) - c(v, q) = \int_\Omega \sum_{i=1}^n \{-2\pi \mathbf{i} b_i^\alpha r_i \varphi M_{\alpha\beta}\varphi s_\beta + 2\pi \mathbf{i} c_i^\alpha s_\alpha \varphi r_i \varphi\}\, dx + R, \tag{2.11}$$

where the lower order terms in R satisfy the estimate

$$|R| \leq c|s|\ ||\varphi||_{1,2}^2$$

with a constant c independent of r, s. We replace r in (2.11) by λr, $\lambda \in \mathbb{R}$, and apply the mean value theorem which gives, for $x_1 \in B_\epsilon(x_0)$,

$$b(q, Mv) - c(v, q) = -\lambda 2\pi \mathbf{i} S + R, \tag{2.12}$$

where

$$S = \sum_{i=1}^n \{b_i^\alpha(x_1) r_i M_{\alpha\beta}(x_1) s_\beta - c_i^\alpha(x_1) s_\alpha r_i\}.$$

In view of the fact that condition (2.8) must hold for all $\lambda \in \mathbb{R}$ and

$$||q||_2 ||v||_2 \leq c|s|,$$

we have $S = 0$ or

$$r^T \tilde{B}(x_1) M(x_1) s = r^T \tilde{C}(x_1) s \quad \forall r, s \in \mathbb{R}^n,$$

and, since $\tilde{B}, \tilde{C}, M$ are continuous and ϵ can be chosen arbitrary small,

$$\tilde{C} = \tilde{B}M \ \text{ in } \Omega.$$

Comparing this with (2.10), we obtain that $A = \tilde{B}M\tilde{B}^T$ is real positive. This implies that M is also real positive.

Finally, we show that condition (2.9) is equivalent to the uniform regularity of the matrix $\tilde{B}$. Let $\tilde{B}(x)$ be uniformly regular, i.e.,

$$|\tilde{B}(x)^{-T} r|^2 \leq c|r|^2 \quad \forall x \in \Omega,\ \forall r \in \mathbb{R}^n.$$

Inserting $r = \tilde{B}(x)^T \nabla q(x)$ and integrating gives

$$||\nabla q||_2 \leq c||\tilde{B}^T \nabla q||_2. \tag{2.13}$$

From

$$(Bq)^\alpha = b_i^\alpha \partial_i q = (\tilde{B}^T \nabla q)^\alpha + b_0^\alpha q,$$

it follows that

$$\|Bq\|_2 \geq \|\tilde{B}^T \nabla q\|_2 - \|b_0 q\|_2,$$

and condition (2.9) is proved by (2.13). Conversely, if condition (2.9) is satisfied, we proceed with similar arguments as before. Let $x_0 \in \Omega$ and $\epsilon > 0$ such that $B_\epsilon(x_0) \subset \Omega$. Let $\varphi \in C_0^\infty(B_\epsilon(x_0))$ with $||\varphi||_2 = 1$. For $r \in \mathbb{R}^n$ set

$$q(x) = e^{2\pi \mathbf{i} r\cdot x}\varphi(x).$$

Then

$$\nabla q(x) = 2\pi \mathbf{i} r q(x) + e^{2\pi \mathbf{i} r\cdot x}\nabla\varphi(x)$$

and

$$||q||_2 = ||\varphi||_2 = 1.$$

For the terms in (2.9) we obtain

$$m_2||\nabla q||^2 \geq \frac{m_2}{2}||2\pi \mathbf{i} r\varphi||^2 - m_2||\nabla\varphi||^2 = \frac{m_2}{2}(|r|^2 - 2||\nabla\varphi||^2),$$

$$||Bq|| \leq ||b_i^\alpha \partial_i q|| \leq ||\tilde{B}^T rq||_2 + c||\varphi||_{1,2}.$$

Using the mean value theorem, we obtain for $x_1 \in B_\epsilon(x_0)$

$$||\tilde{B}^T rq||_2 = |\tilde{B}^T(x_1)r|.$$

From (2.9) and the above estimates it follows that

$$\frac{m_2}{2}(|r|^2 - 2||\nabla\varphi||_2^2) \leq 2|\tilde{B}(x_1)r|^2 + c||\varphi||_{1,2}^2.$$

Replacing r by λr, we conclude that $\tilde{B}$ is uniformly regular in Ω. □

Proof of Theorem 1.1. The theorem can clearly be proved by considering the second order equation corresponding to (2.3), (2.4). Here we want to show how the methodology of Lemma 2.1 can be applied. The finite element analysis in Lemma 4.2 is more complicated but uses the same arguments.

We consider (2.3) and a compactly perturbed (2.4), namely,

$$c(v, \psi) + d(q, \psi) - \lambda(q, \psi) = (g, \psi) \quad \forall \psi \in H_0^{1,2}(\Omega), \tag{2.14}$$

with a parameter $\lambda > 0$. Assume that $(v, q) \in L^2(\Omega)^n \times H_0^{1,2}(\Omega)$ is a solution of (2.3), (2.14). Inserting $\varphi = Mv$ into (2.3) and $\psi = q$ into (2.14) and subtracting the resulting equations yields

$$(v, Mv) + b(q, Mv) - c(v, q) - d(q, q) + \lambda||q||_2^2 = (f, Mv) - (g, q),$$

and hence, by conditions (2.7) and (2.8),

$$m_1||v||_2^2 + \lambda||q||_2^2 \le c||v||_2||q||_2 + c||q||_{1,2}||q||_2 + c||f||_2||v||_2 + ||g||_{-1,2}||q||_{1,2}. \tag{2.15}$$

From the strong formulation of (2.3) we conclude that

$$||Bq||_2^2 \le 2||v||_2^2 + 2||f||_2^2$$

and, by condition (2.9),

$$m_2||\nabla q||^2 \le 2||v||_2^2 + 2||f||_2^2 + c||q||_2^2.$$

Combining this estimate with (2.15) gives the a priori bound for sufficiently large $\lambda > 0$,

$$||v||_2^2 + ||q||_{1,2}^2 \le c\{||f||_2^2 + ||g||_{-1,2}^2\}. \tag{2.16}$$

From the bound (2.16) we conclude an inf-sup condition for the bilinear form

$$B_\lambda((v,q),(\varphi,\psi)) = (v,\varphi) + b(q,\varphi) - c(v,\psi) - d(q,\psi) + \lambda(q,\psi),$$

since

$$\begin{aligned} ||f||_2 + ||g||_{-1} &= \sup_{\varphi \in L^2(\Omega)^n} \frac{(v,\varphi) + b(q,\varphi)}{||\varphi||_2} + \sup_{\psi \in H_0^{1,2}(\Omega)} \frac{-c(v,\psi) - d(q,\psi) + \lambda(q,\psi)}{||\psi||_{1,2}} \\ &= \sup_{||\varphi||_2 = 1, ||\psi||_{1,2} = 1} B_\lambda((v,q),(\varphi,\psi)), \end{aligned}$$

and hence, by (2.16),

$$||v||_2 + ||q||_{1,2} \le c \sup_{||\varphi||_2 = 1, ||\psi||_{1,2} = 1} B_\lambda((v,q),(\varphi,\psi)).$$

In order to prove existence of a solution of (2.3), (2.14), an a priori bound analogous to (2.16) for the adjoint problem is required (see [2]). This can be done by the same method by observing that the adjoint forms satisfy condition (2.8),

$$|c(M^{-1}v,q) - b(q,v)| \le c||v||_2||q||_2$$

and M^{-1} is real positive. Thus, we have proved that the system (2.3), (2.14) has a uniquely determined solution. In view of the fact that $\lambda(q,\psi)$ is a compact perturbation of the bilinear form B_0, we obtain from standard arguments (see, e.g., [1, p. 102]) that (2.3), (2.4) satisfies Fredholm's alternative.

3. Standard finite element spaces and weighted Sobolev norms.

Throughout the rest of the paper we assume that Ω is a polygonal or polyhedral domain and that $\overline{\Omega} = \cup\Lambda$. The interpolation operator $I_h : C^0(\overline{\Omega}) \to S^m$ of a standard Lagrangian finite element space S^m satisfies

$$||\nabla^k(u - I_h u)||_{p;\Lambda} \le ch^{l-k}||\nabla^l u||_{p;\Lambda} \quad \forall u \in H^{l,p}(\Lambda), \tag{3.1}$$

for $0 \le k \le l \le m+1$ and $1 \le p \le \infty$ with $lp > n$ (see, e.g., [9]). Moreover, the inverse estimates hold,

$$||\nabla^k u_h||_{p;\Lambda} \le ch^{-k}||u_h||_{p;\Lambda} \quad \forall u_h \in S^m,\ 0 \le k \le m,\ 1 \le p \le \infty, \tag{3.2}$$

$$||u_h||_{\infty;\Lambda} \le ch^{-n/2}||u_h||_{2;\Lambda} \quad \forall u_h \in S^m. \tag{3.3}$$

For nonsmooth functions we use the approximation operator $R_h : L^1(\Omega) \to S^m$ defined in [10] with the properties

$$||\nabla^k(u - R_h u)||_{p;\Lambda} \leq ch^{l-k}||\nabla^l u||_{p;U(\Lambda)}, \quad 0 \leq k \leq l \leq m+1,\ 1 \leq p \leq \infty, \tag{3.4}$$

$$||\nabla^k R_h u||_{p;\Lambda} \leq c||\nabla^k u||_{p;U(\Lambda)}, \quad 0 \leq k \leq m,\ 1 \leq p \leq \infty, \tag{3.5}$$

where $U(\Lambda)$ denotes a neighborhood of Λ of diameter ch. Since Π is assumed to be a quasi-uniform partition, the estimates (3.4), (3.5) carry over to the domain Ω.

For $x_0 \in \Omega$ we define the weight function

$$\sigma(x) = (|x - x_0|^2 + \rho^2)^{1/2},$$

where $\rho = c^* h^\beta$ and $c^* \geq 1$, $\beta \geq 1$ will be determined later. σ satisfies

$$\sigma^{-1} \leq \rho^{-1}, \quad |\nabla^k \sigma| \leq c\sigma^{-k+1} \quad \forall k \in \mathbb{N} \tag{3.6}$$

and, in view of $\operatorname{diam}(\Lambda) \leq h$,

$$\max_{x\in\Lambda} \sigma(x) \leq c \min_{x\in\Lambda} \sigma(x). \tag{3.7}$$

Using polar coordinates, one can easily show that

$$\int_\Omega \sigma^{-\alpha}\, dx \leq \begin{cases} c & \text{for} \quad \alpha < n, \\ c|\ln\rho| & \text{for} \quad \alpha = n, \\ c\rho^{n-\alpha} & \text{for} \quad \alpha > n. \end{cases} \tag{3.8}$$

For $k \in \mathbb{N}$ and $\alpha \in \mathbb{R}$ the corresponding weighted Sobolev norms are defined by

$$||v||^2_{(\alpha)} = \sum_\Lambda \int_\Lambda |v|^2 \sigma^\alpha\, dx, \quad ||v||^2_{(k,\alpha)} = \sum_{i=0}^k ||\nabla^i v||^2_{(\alpha)}.$$

Estimate (3.7) is the crucial point which allows us to prove the interpolation and inverse estimates (3.1)–(3.5) also in weighted norms; for example,

$$\begin{aligned} ||\nabla^k(u - I_h u)||^2_{(\alpha)} &= \sum_\Lambda \int_\Lambda |\nabla^k(u - I_h u)|^2 \sigma^\alpha\, dx \\ &\leq \sum_\Lambda \max_{x\in\Lambda} \sigma^\alpha(x) ||\nabla^k(u - I_h u)||^2_{2;\Lambda} \\ &\leq ch^{2(l-k)} \sum_\Lambda \min_{x\in\Lambda} \sigma^\alpha(x) ||\nabla^l u||^2_{2;\Lambda} \\ &\leq ch^{2(l-k)} ||\nabla^l u||^2_{(\alpha)}, \end{aligned}$$

and hence,

$$||\nabla^k(u - I_h u)||_{(\alpha)} \leq ch^{l-k}||\nabla^l u||_{(\alpha)}, \quad 0 \leq k \leq l \leq m+1,\ 2l > n, \tag{3.9}$$

$$||\nabla^k(u - R_h u)||_{(\alpha)} \leq ch^{l-k}||\nabla^l u||_{(\alpha)}, \quad 0 \leq k \leq l \leq m+1, \tag{3.10}$$

$$||\nabla^k R_h u||_{(\alpha)} \leq c||\nabla^k u||_{(\alpha)}, \tag{3.11}$$

$$||\nabla^k u_h||_{(\alpha)} \leq ch^{-k}||u_k||_{(\alpha)}. \tag{3.12}$$

The next lemma presents the key idea of the weighted Sobolev norms, namely the control of the L^∞-norm by a weighted L^2-norm.

LEMMA 3.1. *Let $v \in H^{l,\infty}(\Omega)$, $1 \le l \le m+1$, and $v_h \in S^m$. Then there exists a point x_0 such that for the corresponding weighted norm*

$$||v - v_h||_\infty \le ch^{-n/2}\rho^{n/2}||v - v_h||_{(-n)} + ch^{-n/2}\rho^{n/2}h^l|\ln\rho|^{1/2}||v||_{l,\infty}.$$

Proof. Using (3.4) and (3.3), we obtain for some $x_0 \in \Lambda$

$$\begin{aligned}||v - v_h||_\infty &\le ||v - R_hv||_\infty + ||R_hv - v_h||_\infty\\ &\le ch^l||v||_{l,\infty} + |R_hv - v_h|(x_0)\\ &\le ch^l||v||_{l,\infty} + ch^{-n/2}||R_hv - v_h||_{2;\Lambda}.\end{aligned}$$

The second term on the right-hand side can be bounded by a weighted norm,

$$\begin{aligned}||R_hv - v_h||_{2;\Lambda} &\le c\rho^{n/2}||R_hv - v_h||_{(-n)}\\ &\le c\rho^{n/2}\left(||v - R_hv||_{(-n)} + ||v - v_h||_{(-n)}\right).\end{aligned}$$

Now we use (3.10), (3.8),

$$||v - R_hv||_{(-n)} \le ch^l||\nabla^l v||_{(-n)} \le ch^l|\ln\rho|^{1/2}||v||_{l,\infty},$$

which completes the proof of the lemma. □

For the duality argument in finite element analysis an a priori estimate in weighted Sobolev norms is required.

LEMMA 3.2. *Assume that conditions* B *and* C *hold for the linear problem* (1.6). *Then, for $f \in H^{1,2}(\Omega)^n$, $g \in L^2(\Omega)$, the solution $(v,q) \in L^2(\Omega)^n \times H_0^{1,2}(\Omega)$ of $L(v,q) = (f,g)$ satisfies*

$$||v||_{(1,n)} + ||q||_{(2,n)} \le c\rho^{-\kappa}(||f||_{(1,n)} + ||g||_{(n)})$$

for all $\kappa > 0$ with a constant $c = c(\kappa)$. The estimate also holds for the adjoint system.

Proof. The result is well known for second order equations (see, e.g., [19], [21]). It also holds with $\rho^{-\kappa}$ replaced by $|\ln\rho|^{1/2}$. Since our system can equivalently be transformed into a second order equation, the result follows. □

4. Finite element approximation of linear systems.

We return to the linear system (2.1), (2.2), but assume that $d_i = 0$ in view of the fact that this term is definitely of lower order. Using again the notation

$$Bq(x) = (b_i^\alpha \partial_i q)_{\alpha=1,\dots,n},$$

the solution of (2.1), (2.2) satisfies

$$(u,\varphi) + b(p,\varphi) = (f,\varphi) \quad \forall\varphi \in L^2(\Omega)^n, \tag{4.1}$$

$$c(u,\psi) + \omega\{(u,B\psi) + b(p,B\psi) - (f,B\psi)\} = (g,\psi) \quad \forall\psi \in H_0^{1,2}(\Omega). \tag{4.2}$$

The stabilized finite element approximation is defined by an analogous set of equations. Subtracting the defining equations for (u,p) and (u_h,p_h) gives

$$(u - u_h,\varphi_h) + b(p - p_h,\varphi_h) = F(\varphi_h) \quad \forall\varphi_h \in X_h, \tag{4.3}$$

$$c(u - u_h,\psi_h) + \omega(u - u_h,B\psi_h) + \omega b(p - p_h,B\psi_h) = G(\psi_h) \quad \forall\psi_h \in Y_h \tag{4.4}$$

with $F = 0, G = 0$. The functionals F, G are introduced for treating the nonlinear case described in the next section. F is a continuous linear functional on $L^2(\Omega)^n$, whereas G is continuous on $H_0^{1,2}(\Omega)$. The corresponding weighted norms of the functionals are defined by

$$||F||_{(-n)} = \sup_{v \in L^2(\Omega)^n} \frac{|F(v)|}{||v||_{(n)}},$$

$$||G||_{(-1,-n)} = \sup_{q \in H_0^{1,2}(\Omega)} \frac{|G(q)|}{||q||_{(1,n)}}.$$

In the following, we assume that the system (4.1), (4.2) for $\omega = 0$ satisfies the assumptions of the first section.

In order to simplify the presentation we assume that $l = m - 1$ and $m \geq 2$. The proof for the case $l = m = 1$ is given in [23]. The latter case does not really differ in this section, but the treatment of the nonlinearities in the next section requires a special technique already described in [15], [12].

In the following we use the weighted norm technique introduced in [19]. It allows an easy treatment of the nonlinearities, but the results are not as sharp as those obtained from the methods using regularized Green's functions (see, e.g., [14]).

We start with a technical estimate which is used only in the proof of Lemma 4.2.

LEMMA 4.1. *For $u \in H^{m,\infty}(\Omega)^n$, $u_h \in X_h$, and $\varphi_h = I_h(M(u - u_h)\sigma^{-n})$ with a smooth matrix M we have*

$$||M(u - u_h)\sigma^{-n} - \varphi_h||_{(n)} \leq ch\rho^{-1}||u - u_h||_{(-n)} + ch^m||\nabla^m u||_{(-n)}. \tag{4.5}$$

Proof. We use the interpolation estimate for I_h in (3.9):

$$||M(u - u_h)\sigma^{-n} - \varphi_h||_{(n)} \leq ch^m||\nabla^m(M(u - u_h)\sigma^{-n})||_{(n)}. \tag{4.6}$$

Recall that the weighted norm is defined for piecewise smooth functions. In view of the fact that $u_h\,|_\Lambda \in \mathbb{P}_{m-1}$ we have

$$||\nabla^m(u - u_h)||_{(-n)} = ||\nabla^m u||_{(-n)},$$

and hence, by Leibniz' rule and (3.6),

$$||\nabla^m(M(u - u_h)\sigma^{-n})||_{(n)} \leq c\sum_{i=0}^{m-1}||u - u_h||_{(i,-n-2m+2i)} + c||\nabla^m u||_{(-n)}.$$

Using (3.9) and (3.12), we immediately obtain

$$\begin{aligned}||\nabla^k(u - u_h)||_{(\alpha)} &\leq ||\nabla^k(u - u_h - I_h(u - u_h))||_{(\alpha)} + ||\nabla^k I_h(u - u_h)||_{(\alpha)} \\ &\leq ch^{m-k}||\nabla^m u||_{(\alpha)} + ch^{-k}||I_h(u - u_h) - (u - u_h)||_{(\alpha)} + ch^{-k}||u - u_h||_{(\alpha)} \\ &\leq ch^{-k}||u - u_h||_{(\alpha)} + ch^{m-k}||\nabla^m u||_{(\alpha)}.\end{aligned}$$

From this estimate we have

$$\begin{aligned}||\nabla^m(M(u - u_h)\sigma^{-n}||_{(n)} \leq c\sum_{i=0}^{m-1} h^{-i}||u - u_h||_{(-n-2m+2i)} \\ + c\sum_{i=0}^{m-1} h^{m-i}||\nabla^m u||_{(-n-2m+2i)} + c||\nabla^m u||_{(-n)}.\end{aligned}$$

Now we use the estimates

$$||v||_{(-n-2\beta)} \leq c\rho^{-\beta}||v||_{(-n)}$$

and $\rho^{-1} \leq h^{-1}$,

$$||\nabla^m(M(u-u_h)\sigma^{-n}||_{(n)} \leq ch^{-m+1}\rho^{-1}||u-u_h||_{(-n)} + c||\nabla^m u||_{(-n)}.$$

This estimate and (4.6) complete the proof of the lemma. □

LEMMA 4.2. *For pairs* (u,p) *and* $(u_h,p_h) \in X_h \times Y_h$ *satisfying* (4.3), (4.4) *we have the error estimates*

$$(4.7)\quad ||u-u_h||^2_{(-n)} + ||p-p_h||^2_{(1,-n)} \leq c||p-p_h||^2_{(-n-2)} + c\{||F||^2_{(-n)} + ||G||^2_{(-1,-n)}\} + ch^{2m-\epsilon}\left\{||u||^2_{m,\infty} + ||p||^2_{m+1,\infty}\right\}$$

for all $\epsilon > 0$ *with a constant* $c = c(\epsilon)$.

Proof. The proof follows the lines of the proof of Theorem 1.1. Lemma 4.2 can be regarded as the energy estimate of the error in a weighted norm. The terms on the right-hand side of (4.7) are denoted by

$$R = ||p-p_h||^2_{(-n-2)} + ||F||^2_{(-n)} + ||G||^2_{(-1,-n)} + h^{2m-\epsilon}\{||u||^2_{m,\infty} + ||p||^2_{m+1,\infty}\}.$$

In view of the fact that the system is elliptic, we can apply condition (2.7) in Lemma 2.1, and we obtain with a smooth matrix M

$$\begin{aligned}(4.8)\qquad m_1||u-u_h||^2_{(-n)} &\leq (u-u_h, M(u-u_h)\sigma^{-n}) \\ &= (u-u_h, M(u-u_h)\sigma^{-n} - \varphi_h) + F(\varphi_h) \\ &\quad - b(p-p_h, \varphi_h - M(u-u_h)\sigma^{-n}) \\ &\quad - b(p-p_h, M(u-u_h)\sigma^{-n}) \\ &= \text{(i)} + \text{(ii)} + \text{(iii)} + \text{(iv)}.\end{aligned}$$

Here we have used (4.3) with $\varphi_h = I_h(M(u-u_h)\sigma^{-n})$.

The first term in (4.8) is estimated by Cauchy's inequality,

$$\text{(i)} \leq ||u-u_h||_{(-n)}||M(u-u_h)\sigma^{-n} - \varphi_h||_{(n)}.$$

From Lemma 4.1 and Young's inequality $ab \leq \frac{\epsilon}{2}a^2 + \frac{1}{2\epsilon}b^2$ we conclude

$$\text{(i)} \leq \left(ch\rho^{-1} + \frac{m_1}{16}\right)||u-u_h||^2_{(-n)} + ch^{2m}||\nabla^m u||^2_{(-n)}$$

and, by $\rho = c^*h$ and c^* sufficiently large,

$$\text{(i)} \leq \frac{m_1}{8}||u-u_h||^2_{(-n)} + ch^{2m}||\nabla^m u||^2_{(-n)}.$$

For the second term, we obtain from the definition of the weighted norm of F

$$\begin{aligned}\text{(ii)} = F(\varphi_h) &\leq ||F||_{(-n)}||\varphi_h||_{(n)} \\ &\leq ||F||_{(-n)}\left(||\varphi_h - M(u-u_h)\sigma^{-n}||_{(n)} + ||M(u-u_h)\sigma^{-n}||_{(n)}\right).\end{aligned}$$

Using Lemma 4.1 and

$$||M(u-u_h)\sigma^{-n}||_{(n)} \leq c||u-u_h||_{(-n)},$$

we get

$$\text{(ii)} \leq \frac{m_1}{8}||u-u_h||^2_{(-n)} + c||F||^2_{(-n)} + ch^{2m}||\nabla^m u||^2_{(-n)}.$$

For the third term we conclude similarly from (4.5)

$$\begin{aligned}\text{(iii)} &\leq ||p-p_h||_{(1,-n)}||M(u-u_h)\sigma^{-n} - \varphi_h||_{(n)} \\ &\leq ch\rho^{-1}||p-p_h||_{(1,-n)}||u-u_h||_{(-n)} + ch^m||p-p_h||_{(1,-n)}||\nabla^m u||_{(-n)} \\ &\leq \eta||p-p_h||^2_{(1,-n)} + \frac{m_1}{8}||u-u_h||^2_{(-n)} + ch^{2m}||\nabla^m u||^2_{(-n)},\end{aligned}$$

with an $\eta > 0$ which will be determined later sufficiently small.

Collecting the preceding estimates, it follows from (4.8) that

$$(4.9)\qquad \frac{m_1}{2}||u-u_h||^2_{(-n)} \leq \eta||p-p_h||^2_{(1,-n)} + cR - b(p-p_h, M(u-u_h)\sigma^{-n}).$$

In order to estimate the last term on the right-hand side we use condition (2.8) in Lemma 2.1 in the form

$$|b(q,Mv) - c(v,q)| \leq c||v||_{(\alpha)}||q||_{(-\alpha)},$$

which can be proved in the same way as (2.8). Now

$$\begin{aligned}-b(p-&p_h, M(u-u_h)\sigma^{-n}) \\ &= -b(p-p_h, M(u-u_h)\sigma^{-n}) + c((u-u_h)\sigma^{-n}, p-p_h) - c((u-u_h)\sigma^{-n}, p-p_h) \\ &\leq c||(u-u_h)\sigma^{-n}||_{(n)}||p-p_h||_{(-n)} - c((u-u_h)\sigma^{-n}, p-p_h).\end{aligned}$$

Thus we have proved that

$$(4.10)\qquad -b(p-p_h, M(u-u_h)\sigma^{-n}) \leq \frac{m_1}{8}||u-u_h||^2_{(-n)} + c||p-p_h||^2_{(-n)} + A,$$

where $A = -c((u-u_h)\sigma^{-n}, p-p_h)$. Since we want to apply the error relation (4.4) to A, we have to shift the weight function σ^{-n} to the second argument of $c(\cdot,\cdot)$,

$$\begin{aligned}A &= \int_\Omega \sum_{i=1}^n c_i^\alpha (u-u_h)^\alpha \sigma^{-n}\partial_i(p-p_h)\,dx - \int_\Omega c_0^\alpha(u-u_h)^\alpha\sigma^{-n}(p-p_h)\,dx \\ &\leq -c(u-u_h,(p-p_h)\sigma^{-n}) + c\int_\Omega |u-u_h||\nabla\sigma^{-n}||p-p_h|\,dx \\ &\leq -c(u-u_h,(p-p_h)\sigma^{-n}) + \frac{m_1}{8}||u-u_h||^2_{(-n)} + c||p-p_h||^2_{(-n-2)}.\end{aligned}$$

Combining this estimate with (4.10) and using the fact that $||p-p_h||^2_{(-n)} \leq c||p-p_h||^2_{(-n-2)}$, we obtain

$$(4.11)\qquad -b((p-p_h), M(u-u_h)\sigma^{-n}) \leq \frac{m_1}{4}||u-u_h||^2_{(-n)} + c||p-p_h||^2_{(-n-2)} + B$$

with

$$B = -c(u - u_h, (p - p_h)\sigma^{-n}),$$

and hence, by (4.9), (4.11),

$$\frac{m_1}{4}||u - u_h||^2_{(-n)} \leq \eta||p - p_h||^2_{(1,-n)} + cR + B. \tag{4.12}$$

Term B is estimated by (4.4):

$$\begin{aligned} B = {} & -c(u - u_h, (p - p_h)\sigma^{-n} - \psi_h) + \omega(u - u_h, B(\psi_h - (p - p_h)\sigma^{-n})) \\ & + \omega b(p - p_h, B(\psi_h - (p - p_h)\sigma^{-n})) - G(\psi_h - (p - p_h)\sigma^{-n}) \\ & + \omega(u - u_h, B((p - p_h)\sigma^{-n})) + \omega b(p - p_h, B((p - p_h)\sigma^{-n})) \\ & - G((p - p_h)\sigma^{-n}), \end{aligned} \tag{4.13}$$

where $\psi_h = I_h((p - p_h)\sigma^{-n})$. In view of the fact that we gain a factor $h\rho^{-1}$ in terms containing $v - I_h v$ in contrast to the terms containing v, we restrict ourselves to the estimation of the last three terms in (4.13), namely

$$\begin{aligned} \text{(i)} &= \omega(u - u_h, B((p - p_h)\sigma^{-n})), \\ \text{(ii)} &= \omega b(p - p_h, B((p - p_h)\sigma^{-n})), \\ \text{(iii)} &= -G((p - p_h)\sigma^{-n}); \end{aligned} \tag{4.14}$$

the other terms in (4.13) can be bounded by

$$\frac{m_1}{16}||u - u_h||^2_{(-n)} + \eta||p - p_h||^2_{(1,-n)} + cR, \tag{4.15}$$

since the constant $\omega > 0$ is assumed to be fixed.

For the first term in (4.14) we obtain

$$\begin{aligned} \text{(i)} = {} & -\omega\int_\Omega (u - u_h)^\alpha b_i^\alpha \partial_i(p - p_h)\sigma^{-n}\,dx - \omega\sum_{i=1}^n \int_\Omega (u - u_h)^\alpha b_i^\alpha (p - p_h)\partial_i\sigma^{-n}\,dx \\ & \leq \omega||u - u_h||_{(-n)}||B(p - p_h)||_{(-n)} + \frac{m_1}{16}\|u - u_h\|^2_{(-n)} + c\|p - p_h\|^2_{(-n-2)}. \end{aligned}$$

The second term in (4.14) gives us the second term with the "right" sign,

$$\begin{aligned} \text{(ii)} &= -\omega\int_\Omega b_i^\alpha \partial_i(p - p_h) b_j^\alpha \partial_j((p - p_h)\sigma^{-n})\,dx \\ &\leq -\omega||B(p - p_h)||^2_{(-n)} + c||p - p_h||_{(1,-n)}||p - p_h||_{(-n-2)} \\ &\leq -\omega||B(p - p_h)||^2_{(-n)} + \eta||p - p_h||^2_{(1,-n)} + cR. \end{aligned}$$

The third term in (4.14) is simply bounded by

$$\text{(iii)} \leq ||G||_{(-1,-n)}||p - p_h||_{(1,-n)} \leq \eta||p - p_h||^2_{(1,-n)} + cR.$$

These estimates together with (4.12), (4.13), (4.15) give

$$\begin{aligned}(4.16)\qquad \frac{m_1}{8}||u-u_h||^2_{(-n)} &+ \omega||B(p-p_h)||^2_{(-n)}\\ &\le \omega||u-u_h||_{(-n)}||B(p-p_h)||_{(-n)} + 3\eta||p-p_h||^2_{(1,-n)} + cR.\end{aligned}$$

It is obvious that the term

$$\omega||u-u_h||_{(-n)}||B(p-p_h)||_{(-n)}$$

cannot be put into the left-hand side of (4.16) for all $\omega > 0$. Since we will prove the lemma for all $\omega > 0$, we will estimate $||u-u_h||_{(-n)}$ by $||B(p-p_h)||_{(-n)}$. Similar to the method described at the beginning of the proof, we use (4.3) with $\varphi_h = I_h((u-u_h)\sigma^{-n})$,

$$\begin{aligned}||u-u_h||^2_{(-n)} &= (u-u_h,(u-u_h)\sigma^{-n}-\varphi_h) + b(p-p_h,(u-u_h)\sigma^{-n}-\varphi_h)\\ &\quad + F(\varphi_h) - b(p-p_h,(u-u_h)\sigma^{-n}).\end{aligned}$$

Using Lemma 4.1 with $M = E$, the first three terms of the right-hand side can be bounded by

$$\epsilon_1||u-u_h||^2_{(-n)} + \epsilon_1||B(p-p_h)||^2_{(-n)} + cR$$

with arbitrarily small $\epsilon_1 > 0$, and for the last term we have trivially

$$-b(p-p_h,(u-u_h)\sigma^{-n}) \le ||B(p-p_h)||_{(-n)}||u-u_h||_{(-n)},$$

and hence

$$(4.17)\qquad ||u-u_h||^2_{(-n)} \le (1+\epsilon)||B(p-p_h)||^2_{(-n)} + cR$$

with $c = c(\epsilon)$ and ϵ can be chosen arbitrary small. We obtain from (4.16) and (4.17) for all $\omega > 0$

$$||u-u_h||^2_{(-n)} + ||B(p-p_h)||^2_{(-n)} \le c\eta||p-p_h||^2_{(1,-n)} + cR.$$

The lemma follows by applying the analogue of (2.9) in Lemma 2.1 and choosing η sufficiently small. □

LEMMA 4.3. *For pairs (u,p) and $(u_h,p_h) \in X_h \times Y_h$ satisfying* (4.3), (4.4) *we have the error estimates*

$$\begin{aligned}(4.18)\qquad ||p-p_h||_{(-n)} &\le c\rho^{-\kappa}h\{||u-u_h||_{(-n)} + ||p-p_h||_{(1,-n)}\}\\ &\quad + c\rho^{-\kappa}\{||F||_{(-n)} + ||G||_{(-1,-n)}\}\end{aligned}$$

with an arbitrarily small $\kappa > 0$ and a constant $c = c(\kappa)$.

Proof. Let $(v,q) \in L^2(\Omega)^n \times H_0^{1,2}(\Omega)$ be the solution of the adjoint problem

$$(\varphi,v) + c(\varphi,q) + \omega(\varphi,Bq) = 0 \quad \forall\varphi \in L^2(\Omega)^n,$$

$$b(\psi,v) + \omega b(\psi,Bq) = ((p-p_h)\sigma^{-n},\psi) \quad \forall\psi \in H_0^{1,2}(\Omega).$$

Inserting $\varphi = u - u_h$, $\psi = p - p_h$ and using the error relations (4.3), (4.4) gives

$$\begin{aligned} ||p-p_h||^2_{(-n)} &= (u-u_h, v-R_hv) + c(u-u_h, q-R_hq) \\ &\quad + \omega(u-u_h, B(q-R_hq)) + b(p-p_h, v-R_hv) \\ &\quad + \omega b(p-p_h, B(q-R_hq)) + F(R_hv) + G(R_hq). \end{aligned} \tag{4.19}$$

The bilinear forms are estimated by Cauchy's inequality and (3.10); for example,

$$\begin{aligned} (u-u_h, v-R_hv) &\leq ||u-u_h||_{(-n)}||v-R_hv||_{(n)} \\ &\leq ch||u-u_h||_{(-n)}||v||_{(1,n)}, \\ \omega b(p-p_h, B(q-R_hq)) &\leq c||p-p_h||_{(1,-n)}||q-R_hq||_{(1,n)} \\ &\leq ch||p-p_h||_{(1,-n)}||q||_{(2,n)}. \end{aligned}$$

The functionals are simply bounded by their norms and (3.11):

$$\begin{aligned} F(R_hv) + G(R_hq) &\leq ||F||_{(-n)}||R_hv||_{(n)} + ||G||_{(-1,-n)}||R_hq||_{(1,n)} \\ &\leq c||F||_{(-n)}||v||_{(1,n)} + c||G||_{(-1,-n)}||q||_{(2,n)}. \end{aligned}$$

Inserting these estimates into (4.19) gives

$$\begin{aligned} ||p-p_h||^2_{(-n)} \leq c(&h||u-u_h||_{(-n)} + h||p-p_h||_{(1,-n)} + ||F||_{(-n)} + ||G||_{(-1,-n)}) \\ &\times(||v||_{(1,n)} + ||q||_{(2,n)}). \end{aligned}$$

The lemma follows by applying Lemma 3.2,

$$||v||_{(1,n)} + ||q||_{(2,n)} \leq c\rho^{-\kappa}||(p-p_h)\sigma^{-n}||_{(n)} = c\rho^{-\kappa}||p-p_h||_{(-n)}. \qquad \square$$

LEMMA 4.4. *There exists an $h_0 > 0$ such that for all $h \leq h_0$ all solutions $(u_h, p_h) \in X_h \times Y_h$ of* (4.3), (4.4) *satisfy the error estimates*

$$\begin{aligned} h||u-u_h||_\infty + h||p-p_h||_{1,\infty} + ||p-p_h||_\infty &\leq ch^{m+1-\epsilon}(||u||_{m,\infty} + ||p||_{m+1,\infty}) \\ &\quad + ch^{-\epsilon}\{||F||_{(-n)} + ||G||_{(-1,-n)}\} \end{aligned}$$

for all $\epsilon > 0$ with constants $c = c(\epsilon)$.

Proof. We add $\rho^{-2}||p-p_h||^2_{(-n)}$ to both sides of (4.7) in Lemma 4.2, use the estimate

$$||p-p_h||^2_{(-n-2)} \leq \rho^{-2}||p-p_h||^2_{(-n)}$$

on the right-hand side of (4.7), and insert the estimate (4.18) into the resulting estimate,

$$\begin{aligned} ||u-u_h||^2_{(-n)} &+ ||p-p_h||^2_{(-1,-n)} + \rho^{-2}||p-p_h||^2_{(-n)} \\ &\leq c\rho^{-2-2\kappa}h^2\{||u-u_h||^2_{(-n)} + ||p-p_h||^2_{(1,-n)}\} \\ &\quad + c\rho^{-2-2\kappa}\{||F||^2_{(-n)} + ||G||^2_{(-1,-n)}\} \\ &\quad + ch^{2m-\epsilon}\{||u||^2_{m,\infty} + ||p||^2_{m+1,\infty}\}, \end{aligned} \tag{4.20}$$

where κ and ϵ can be chosen arbitrarily small. For small κ we choose $\rho = c^* h = h^{1-2\kappa}$ such that

$$c\rho^{-2-2\kappa}h^2 = ch^{(1-2\kappa)(-2-2\kappa)}h^2 \leq \frac{1}{2}$$

for $h \leq h_0$. By this procedure, the first term on the right-hand side of (4.20) can be put into the left-hand side. We determine

$$\max(||u-u_h||_\infty, ||p-p_h||_{1,\infty}, \rho^{-1}||p-p_h||_\infty)$$

and choose x_0 such that Lemma 3.1 can be applied to the corresponding term. In view of the fact that κ can be chosen arbitrarily small, the lemma is proved. □

5. Proof of Theorem 1.2. We return to the definition of the discrete solution in (1.7), (1.8). Subtracting this system from the corresponding continuous system yields

$$(u-u_h, \varphi_h) = (b(\cdot, p, \nabla p) - b(\cdot, p_h, \nabla p_h), \varphi_h) \quad \forall \varphi_h \in X_h, \tag{5.1}$$

$$\begin{aligned} -\sum_{i=1}^n (c_i(\cdot,p,u) - c_i(\cdot,p_h,u_h), \partial_i\psi_h) + (c_0(\cdot,p,u) - c_0(\cdot,p_h,u_h), \psi_h) \\ + \omega(u-u_h-b(\cdot,p,\nabla p) + b(\cdot,p_h,\nabla p_h), B(p_h)\psi_h) \\ = (g(\cdot,p,\nabla p) - g(\cdot,p_h,\nabla p_h), \psi_h) \quad \forall \psi_h \in Y_h. \end{aligned} \tag{5.2}$$

Setting

$$\bar{b}_i^\alpha = \int_0^1 b_i^\alpha(x, tp + (1-t)p_h, t\nabla p + (1-t)\nabla p_h)\,dt,$$

we have

$$\begin{aligned} (b(\cdot,p,\nabla p) - b(\cdot,p_h,\nabla p_h), \varphi_h) &= (\bar{b}_i^\alpha \partial_i(p-p_h), \varphi_h^\alpha) \\ &= -b(p-p_h, \varphi_h) + F(p_h;\varphi_h), \end{aligned}$$

where

$$b(q,v) = -\int_\Omega b_i^\alpha \partial_i q\, v\, dx$$

and

$$F(p_h;\varphi_h) = ((\bar{b}_i^\alpha - b_i^\alpha)\partial_i(p-p_h), \varphi_h)$$

satisfies

$$|F(p_h;\varphi_h)| \leq c\int_\Omega (|p-p_h|^2 + |\nabla(p-p_h)|^2)|\varphi_h|\,dx. \tag{5.3}$$

Thus we have shown that (5.1) is of the form

$$(u-u_h,\varphi_h) + b(p-p_h,\varphi_h) = F(p_h;\varphi_h) \quad \forall \varphi_h \in X_h, \tag{5.4}$$

which is exactly (4.3). The first two terms in (5.2) are treated by the same method

$$-\sum_{i=1}^{n}(c_i(\cdot,p,u)-c_i(\cdot,p_h,u_h),\partial_i\psi_h)+(c_0(\cdot,p,u)-c_0(\cdot,p_h,u_h),\psi_h)$$

$$=c(u-u_h,\psi_h)+d_1(p-p_h,\psi_h)-G_1(u_h,p_h;\psi_h),$$

where G_1 satisfies

$$|G_1(u,p;\psi_h)|\le c\int_\Omega(|p-p_h|+|u-u_h|)^2(|\psi_h|+|\nabla\psi_h|)\,dx. \tag{5.5}$$

For the right-hand side we obtain similarly

$$(g(\cdot,p,\nabla p)-g(\cdot,p_h,\nabla p_h),\psi_h)=-d_2(p-p_h,\psi_h)+G_2(p_h;\psi_h)$$

with

$$|G_2(p_h;\psi_h)|\le c\int_\Omega(|p-p_h|+|\nabla p-\nabla p_h|)^2|\psi_h|\,dx. \tag{5.6}$$

For the remaining term in (5.2) we obtain

$$\begin{aligned}\omega(u-u_h&-b(\cdot,p,\nabla p)+b(\cdot,p_h,\nabla p_h),B(p_h)\psi_h)\\ &=\omega(u-u_h-b(\cdot,p,\nabla p)+b(\cdot,p_h,\nabla p_h),(B(p_h)-B(p))\psi_h)\\ &\quad+\omega(u-u_h-b(\cdot,p,\nabla p)+b(\cdot,p_h,\nabla p_h),B(p)\psi_h)\\ &=-G_3(u_h,p_h;\psi_h)+\omega(u-u_h,B(p)\psi_h)\\ &\quad+\omega b(p-p_h,B(p)\psi_h)-G_4(u_h,p_h;\psi_h)\end{aligned}$$

with G_3,G_4 satisfying

$$\begin{aligned}&|G_3(u_h,p_h;\psi_h)|+|G_4(u_h,p_h;\psi_h)|\\ &\qquad\le c\int_\Omega(|u-u_h|^2+|p-p_h|^2+|\nabla p-\nabla p_h|^2)(|\psi_h|+|\nabla\psi_h|)\,dx.\end{aligned} \tag{5.7}$$

Now (5.1), (5.2) are exactly of the form (4.3), (4.4) with $G=G_1+G_2+G_3+G_4$, $d=d_1+d_2$. From (5.3), (5.5), (5.6), (5.7), it follows with the aid of Cauchy's inequality that

$$\begin{aligned}||F(u_h;\cdot)||_{(-n)}&\le\left(\int_\Omega(|p-p_h|^2+|\nabla(p-p_h)|^2)^2\sigma^{-n}\,dx\right)^{1/2}\\ &\le c|\ln h|^{1/2}||p-p_h||_{1,\infty}^2,\end{aligned} \tag{5.8}$$

$$||G(u_h,p_h;\cdot)||_{(-1,-n)}\le c|\ln h|^{1/2}(||u-u_h||_\infty^2+||p-p_h||_{1,\infty}^2). \tag{5.9}$$

Define the set

$$K_R = \{(v_h, q_h) \in X_h \times Y_h : \ ||u - v_h||_\infty + ||p - q_h||_{1,\infty} \le R\}$$

and the operator $T : K_R \to X_h \times Y_h$, $T(v_h, q_h) = (u_h, p_h)$ by

$$(u - u_h, \varphi_h) + b(p - p_h, \varphi_h) = F(v_h; \varphi_h) \quad \forall \varphi_h \in X_h, \tag{5.10}$$

$$\begin{aligned} c(u - u_h, \psi_h) + \omega(u - u_h, B\psi_h) + \omega b(p - p_h, B\psi_h) \\ = G(v_h, q_h; \psi_h) \ \ \forall \psi_h \in Y_h, \end{aligned} \tag{5.11}$$

where $B = B(p)$. In order to demonstrate that T is well defined let $h \le h_0$ such that Lemma 4.4 can be applied. Set $F = 0$, $G = 0$, $u = 0$, $p = 0$. Then it follows from Lemma 4.4 that $u_h = 0$ and $p_h = 0$. This proves that (5.10), (5.11) is a regular system for h sufficiently small.

By Lemma 4.4, (u_h, p_h) satisfies

$$\begin{aligned} ||u - u_h||_\infty + ||p - p_h||_{1,\infty} &\le c(u,p)h^{m-\epsilon} \\ &\quad + ch^{-1-\epsilon}\left\{||F(v_h; \cdot)||_{(-n)} + ||G(v_h, q_h; \cdot)||_{(-1,-n)}\right\} \\ &\le c(u,p)h^{m-\epsilon} + ch^{-1-2\epsilon}R^2, \end{aligned}$$

where we have used (5.8), (5.9) and that $(v_h, q_h) \in K_R$. Choosing $R = 2c(u,p)h^{m-\epsilon}$ and recalling $m \ge 2$, the last estimate demonstrates that $T : K_R \to K_R$ for sufficiently small h. By Brouwer's fixed point theorem, T has a fixed point, which is a solution of (1.7), (1.8). The better estimate for $||p - p_h||_\infty$ follows from Lemma 4.4.

REFERENCES

[1] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, van Nostrand, Princeton, NJ, 1965.

[2] I. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundation of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, London, 1972, pp. 3–359.

[3] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-square type*, SIAM Rev., 40 (1998), pp. 789–837.

[4] F. Brezzi, *On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers*, RAIRO, 8 (R-2) (1974), pp. 129–151.

[5] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, Berlin, 1989.

[7] F. Brezzi and J. Pitkäranta, *On the stabilization of the finite element approximation of the Stokes equations*, in Efficient Solutions of Elliptic Systems (Proc. GAMM-Seminar, Kiel, 1984), Notes Numer. Fluid Mech. 10, Vieweg, Braunschweig, Germany, 1984, pp. 11–19.

[8] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least-squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[10] P. Clement, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numer., 9 (R-2) (1975), pp. 77–84.

[11] M. Dobrowolski, *Finite element methods for elliptic systems with constraints*, Numer. Linear Algebra Appl., 6 (1999), pp. 115–124.

[12] M. Dobrowolski and R. Rannacher, *Finite element methods for nonlinear elliptic systems of second order*, Math. Nachr., 94 (1980), pp. 155–172.

[13] L. P. Franca and R. Stenberg, *Error analysis of some Galerkin least squares methods for the elasticity equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1680–1697.

[14] J. Frehse and R. Rannacher, *Eine L^1–Fehlerabschätzung für diskrete Grundlösungen in der Methode der finiten Elemente*, Bonner Math. Schriften, 89 (1976), pp. 92–114.

[15] J. Frehse and R. Rannacher, *Asymptotic L^∞-error estimates for linear finite element approximation of quasilinear boundary value problems*, SIAM J. Numer. Anal., 15 (1978), pp. 418–431.

[16] M. Lee and F. A. Milner, *Mixed finite element methods for nonlinear elliptic problems: The h–p version*, J. Comput. Appl. Math., 85 (1997), pp. 239–261.

[17] F. A. Milner, *Mixed finite element methods for quasilinear second-order elliptic problems*, Math. Comp., 44 (1985), pp. 303–320.

[18] R. A. Nicolaides, *Existence, uniqueness and approximation for generalized saddle point problems*, SIAM J. Numer. Anal., 19 (1982), pp. 349–357.

[19] J. A. Nitsche, *L^∞-convergence of finite element approximations*, in Lecture Notes in Math. 606, Springer, New York, 1977, pp. 261–274.

[20] E.-J. Park, *Mixed finite element methods for nonlinear second-order elliptic problems*, SIAM J. Numer. Anal., 32 (1995), pp. 865–885.

[21] R. Rannacher, *Zur L^∞-Konvergenz linearer finiter Elemente beim Dirichlet–Problem*, Math. Z., 149 (1976), pp. 69–77.

[22] P. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.

[23] M. Villegas, *Gemischte Finite Elemente Verfahren bei elliptischen Systemen*, dissertation, Universität Würzburg, Würzburg, Germany, 1998; available online at http://ifamus.mathematik.uni-wuerzburg.de/~dobro/pub/villegas.ps.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 970–986

THE IMPLICIT UPWIND METHOD FOR 1-D SCALAR CONSERVATION LAWS WITH CONTINUOUS FLUXES[*]

MICHAEL BREUẞ[†]

Abstract. We present the first part of a theory of monotone implicit methods for scalar conservation laws. In this paper, we focus on the implicit upwind scheme. The theoretical investigation of this method is centered around a rigorously verified implicit monotonicity criterion. The relation between the upwind scheme and a discrete entropy inequality is constructed analogously to the classical approach of Crandall and Majda [M. G. Crandall and A. Majda, *Math. Comp.*, 34 (1980), pp. 1–21]. A proof of convergence is given which does not rely on a classical compactness argument. The theoretical results are complemented by a discussion of numerical aspects.

Key words. conservation laws, implicit methods, monotone methods, upwind methods

AMS subject classifications. 35L65, 65M06, 65M12

DOI. 10.1137/S0036142902410106

1. Introduction.

The aim of this paper is to describe the first part of a constructive approach to the entropy solution of hyperbolic conservation laws in the sense of Kružkov for the most general case possible in 1-D, i.e., essentially for merely continuous flux functions. Finally, we seek to develop a theory of implicit finite difference methods which is centered around a rigorously verified implicit monotonicity notion.

The validation of the monotonicity property is not trivial in the implicit case. We restrict our attention in this note on the upwind method, i.e., on continuous nonstrictly growing monotone fluxes. The extension to the mentioned more general case is the subject of a forthcoming paper. In contrast to fundamental works within this field; see, e.g., [4, 3, 16, 13], we do not rely on the Lipschitz continuity of the flux, nor on the boundedness of the solution in space, nor on boundedness with respect to the total variation of the solution. In the nonconstructive sense, the existence and uniqueness results of corresponding solutions are documented within a number of papers of Kružkov and his co-workers; see [1, 11, 12] and the references therein. Related to the subject of the present work are also problems with discontinuous fluxes; see, e.g., [6] for a useful presentation.

In order to point out the difficulties encountered when approaching the described task, let us briefly discuss the consequences of our basic assumptions.

If the flux function of a nonlinear conservation law is not Lipschitz continuous, it may happen that information is propagated with infinite speed. This is, e.g., the case in the example given by Kružkov and Panov in [12], which is concerned with the equation

$$\frac{\partial}{\partial t}u(x,t)+\frac{\partial}{\partial x}\left(\frac{|u(x,t)|^{\alpha}}{\alpha}\right)=0,\quad \alpha\in(0,1),\, t>0,\, x\in\mathbf{R}. \tag{1.1}$$

[*]Received by the editors June 24, 2002; accepted for publication (in revised form) October 20, 2004; published electronically September 20, 2005.

http://www.siam.org/journals/sinum/43-3/41010.html

[†]Technical University Brunswick, Computational Mathematics, Pockelsstraße 14, 38106 Brunswick, Germany (m.breuss@tu-bs.de).

Given the initial condition

$$u_0(x) \;=\; \begin{cases} 0 & : \quad x < -1, \\ 1 & : \quad -1 \le x \le 0, \\ 0 & : \quad x > 0, \end{cases} \tag{1.2}$$

the exact solution defined over a time interval depending on the exact choice of α reads

$$u(x,t) \;=\; \begin{cases} 0 & : \quad t > \alpha(x+1), \\ 1 & : \quad x < t \le \alpha(x+1), \\ \left(\frac{t}{x}\right)^{1/(1-\alpha)} & : \quad t \le x. \end{cases} \tag{1.3}$$

As can easily be seen, this solution incorporates a rarefaction wave extending to infinity after arbitrarily small time. Mathematically, two properties of special interest are contained within this example: the flux features a pole at $u = 0$, and the domain of the solution is infinite even for an initial condition with compact support. The first property yields that the CFL number would be effectively zero when using an explicit scheme. Also, the Kuznetsov approach to convergence [13] does not work since it explicitly uses the Lipschitz continuity of the flux (as well as the boundedness of the domain of the solution). The second property implies that the main other traditional approaches to the convergence of numerical methods are also not suitable, as can be seen as follows. There are essentially two further approaches to be noted: the first one relies on Helly's theorem, using the compactness of the space of functions of bounded variations to extract a converging sequence of numerical solutions as the grid is refined. This is, e.g., the case in the convergence proofs of TVD methods. As described in detail by LeVeque [14], this function space is only compact when employing the BV concept over a fixed compact space-time-domain, thus the compactness property of this function space is in general not applicable in the discussed case. Note that also within the paper of Crandall and Majda [4] on explicit monotone methods, the properties of this function space are used to obtain a compactness argument. The second approach relies on the concept of measure valued solutions introduced by DiPerna [5]. Also here, the compactness of the space-time-domain of the solution is assumed which is, e.g., already noted in [3].

By this discussion, it is evident that we have to employ implicit schemes in a framework which enables the use of a convergence strategy different from the mentioned classical ones. We achieve this by using the monotonicity of an implicit method as a means of nonlinear stability. It is then verified that the monotonicity property is enough to guarantee the convergence of such methods to the entropy solution in the sense of Kružkov.

The need to be sure about the monotonicity property needs to be addressed in detail within this paper. In fact, within the literature about numerical methods for conservation laws, the monotonicity of numerical schemes is generally discussed for explicit schemes; see, e.g., [7, 14] and the references therein. It makes no sense to cite here numerous more practically oriented papers on implicit methods, where the numerical flux function of a scheme which is monotone under a CFL condition in the explicit case is employed in some context, and where an implicit time-stepping procedure is applied for evaluating the fluxes, e.g., to ensure the stability of an algorithm, or, e.g., to enhance the effectiveness of a method for computing steady state solutions, etc.. Since the aim of these works is typically a totally different one from ours, it is also not at all our objective to criticize them. However, an existing mathematically rigorous investigation of the monotonicity property of implicit schemes is not known by the author. Furthermore, with respect to the continuity of the flux, it is necessary that the validation of an implicit monotonicity criterion does not involve any derivative of the flux. Consequently, we develop a monotonicity notion which is verified by comparing data sets using the induction principle.

One further point of interest is the validity of a discrete entropy condition. Since we are interested in the Kružkov entropy notion, it is adequate to use a discrete version of the corresponding entropy inequality. We found that the relation between a monotone method and a discrete entropy inequality used by Crandall and Majda [4] (which is in turn based on the previous work of Harten, Hyman, and Lax [9]) can also be used in the implicit case. This is not at all self-evident; it is based on the fact that even in the original derivation of this relation no technique is used which relies on the Lipschitz property of the flux. This was also already noted in the original paper [4]. However, the validity of this relation of course has to be verified rigorously since there are some differences with respect to the explicit case within the techniques in use. It is mandatory to stress that only the relation between the implicit upwind method and the discrete entropy condition is derived similarly to the proceeding in [4]. The remainder of the content of this paper including the convergence proof is technically completely different than the theory presented in [4].

The proof of convergence can be sketched as follows. Using pointwise comparison of data sets, the monotone implicit upwind method yields a monotonously growing approximative sequence of numerical solutions corresponding to a constructed sequence of monotonously growing discrete initial data. The sequence is L_∞-bounded by the monotonicity of the method for L_∞-bounded initial data. Thus, it is possible to use the theorem of monotone convergence by Beppo Levi to obtain strong convergence a.e. of the whole sequence to a unique limit over any arbitrarily chosen compact domain. Since the approximative sequence satisfies a discrete entropy inequality, convergence to the entropy condition of Kružkov follows by the established strong convergence a.e. of the sequence. To the knowledge of the author, this principle is not used up to now in any other work.

According to the discussion above, this paper is organized as follows. At first, implicit notions are developed which are centered around the monotonicity notion for the implicit upwind schemes. A minor stability result is established which is needed within the convergence proof. The upwind method is investigated with respect to its monotonicity. The proof of convergence to the entropy solution is illustrated. Finally, we discuss numerical approximations followed by conclusive remarks and acknowledgements.

2. The implicit upwind method.

In the following, u_m^l denotes the value of the numerical solution at $m\Delta x$ and $l\Delta t$. The presence of a uniform grid with a constant mesh ratio $\lambda = \Delta t/\Delta x$ is assumed, but this is only for a more transparent notation and is not essential.

In order to avoid some inconveniences within the notation, implicit 3-point-methods of the form

$$u_0^{n+1} = \tilde{H}(u_{-1}^{n+1}, u_0^{n+1}, u_1^{n+1}; u_0^n)$$

are sometimes addressed which includes both possible upwind schemes

$$u_j^{n+1} \;=\; u_j^n - \lambda\big[f(u_j^{n+1}) - f(u_{j-1}^{n+1})\big] \tag{2.1}$$

and

$$u_j^{n+1} \;=\; u_j^n - \lambda\big[f(u_j^{n+1}) - f(u_{j+1}^{n+1})\big].$$

However, in this paper we only discuss in detail the upwind scheme (2.1). All statements given within the paper can easily be adapted in an identical fashion to the other possible implicit upwind scheme.

Since we restrict our attention in the following to (2.1), it is convenient to use as the abbreviated form of the method

$$u_0^{n+1} = H(u_{-1}^{n+1}, u_0^{n+1}; u_0^n). \tag{2.2}$$

2.1. Implicit notions. For the completeness of the presentation, we give the definitions of the conservation form, the numerical flux function, and consistency. These notions are completely analogous to the corresponding ones for explicit schemes.

Definition 2.1 (conservation form, numerical flux function). *A* 3*-point numerical scheme is in conservation form if a continuous numerical flux function $g : \mathbf{R}^2 \to \mathbf{R}$ exists, so that the method reads*

$$\tilde{H}(u_{-1}^{n+1}, u_0^{n+1}, u_1^{n+1}; u_0^n) = u_0^n - \lambda\big[g(u_0^{n+1}, u_1^{n+1}) - g(u_{-1}^{n+1}, u_0^{n+1})\big].$$

Definition 2.2 (consistency). *A* 3*-point numerical method described by g is consistent if $g(v,v) = f(v)$ holds for all $v \in \mathbf{R}$.*
Obviously, the implicit upwind scheme (2.1) is a consistent and conservative scheme with

$$g \equiv g(u,v) \equiv g(u) = f(u). \tag{2.3}$$

The key to nonlinear stability is the notion of monotonicity.

Definition 2.3 (monotonicity). *Let two data sequences v^n and w^n be given. Let the upwind scheme* (2.1) *produce new sequences of data v^{n+1} and w^{n+1} out of the given data v^n and w^n, respectively. Then the method is monotone iff the implication*

$$v^n \geq w^n \quad \Rightarrow \quad v^{n+1} \geq w^{n+1} \tag{2.4}$$

holds in the sense of the comparison of components.
Note that Definition 2.3 captures the essence of the notion of monotonicity and is free of derivatives of H.

In the following, we use the abbreviations $w_{j-1}^{n+1} =: a$, $w_j^{n+1} =: b$ and $w_j^n =: d$ when appropriate to simplify the notation.

Theorem 2.1 (monotonicity of the upwind scheme). *The upwind method* (2.1) *is monotone if the following conditions hold:*

$$H(a + \Delta a, b; d) \geq H(a, b, c; d), \tag{2.5}$$

$$H(a, b; d + \Delta d) \geq H(a, b; d), \tag{2.6}$$

with $\Delta a, \Delta d \;\geq 0$, respectively. Thereby, the condition (2.5) *has the meaning that the scheme* (2.1) *is monotone if the flux is a nonstrictly growing monotonous function.*
Before we proceed with the proof, let us give an insight into the meaning of the properties of the mapping H which follow from (2.5) and (2.6).

In the case of explicit 3-point schemes, the monotonicity of a scheme is verified by computing the derivatives of the function k of the method,

$$u_j^{n+1} = k(u_{j-1}^n, u_j^n, u_{j+1}^n).$$

The monotonicity condition is then

$$\frac{\partial}{\partial a_i} k(a_1, a_2, a_3) \geq 0 \quad \forall\, i \in \{1, 2, 3\}.$$

Note that this means not to investigate concrete changes in u^{n+1} due to changes in a given data set u^n. Rather than that, the effects of changes in u^n are studied this way, answering the question: If we change the given data set u^n in a positive way, is the effect a nonnegative change in u^{n+1} or not?

Since we deal with continuous fluxes in this paper, we cannot generally compute a derivative. However, if we fix, for instance, our attention on the first monotonicity requirement (2.5)

$$H(a + \Delta a, b; d) \geq H(a, b; d)$$

for $\Delta a > 0$, then we may subtract the right-hand side from this inequality and divide by Δa to obtain the condition

$$\frac{H(a + \Delta a, b; d) - H(a, b; d)}{\Delta a} \geq 0$$

which takes the form of a discretized derivative. We observe the connection between the explicit and the implicit monotonicity notion.

Proof of Theorem 2.1. Let us first verify the validity of the conditions (2.5) and (2.6) before we proceed further.
To condition (2.5)*,*

$$\begin{aligned} &H(a + \Delta a, b; d) - H(a, b; d) \\ &= \big[d - \lambda\big[f(b) - f(a + \Delta a)\big]\big] - \big[d - \lambda\big[f(b) - f(a)\big]\big] \\ &= \lambda\left[f(a + \Delta a) - f(a)\right]. \end{aligned}$$

The condition (2.5) is only valid if f grows monotonously.
To condition (2.6)*,*

$$\begin{aligned} &H(a, b, c; d + \Delta d) - H(a, b, c; d) \\ &= \big[d + \Delta d - \lambda\big[f(b) - f(a)\big]\big] - \big[d - \lambda\big[f(b) - f(a)\big]\big] \\ &= \Delta d \quad (\geq 0). \end{aligned}$$

The latter condition seems to be redundant by the general form of the discussed methods. However, the condition (2.6) fixes the necessary coupling mechanism needed to obtain a comparison principle as given in Definition 2.3. Note that f does not need to be Lipschitz continuous in order to fulfill the stated monotonicity conditions. Let us now prepare the main part of the proof and let us note that (2.1) is defined for the indices $j > J$ and $n \geq 0$, whereby u_J^{n+1} must be given for some index J.

In the notation used before, the scheme (2.1) reads

$$b = d - \lambda\left(f(b) - f(a)\right),$$

or, equivalently,

$$b + \lambda f(b) = d + \lambda f(a).$$

Let us define the function

$$\mathcal{F}(b) := b + \lambda f(b).$$

For $\delta b > 0$ we have

$$\begin{aligned}&\mathcal{F}(b+\delta b) - \mathcal{F}(b)\\ &= \big[b + \delta b + \lambda f(b+\delta b)\big] - \big[b + \lambda f(b)\big]\\ &= \underbrace{\delta b}_{>0} + \lambda \underbrace{\left(f(b+\delta b) - f(b)\right)}_{\geq 0}\end{aligned}$$

since f is a nonstrictly growing monotone function. Thus, $\mathcal{F}$ is a strictly growing monotone function so that there exists a strictly increasing inverse function $\mathcal{F}^{-1}$. Now, let two sequences v^n, w^n be given with $v^n \mapsto v^{n+1}$ and $w^n \mapsto w^{n+1}$ by application of the considered scheme. Furthermore, let $(v^n = w^n) \Rightarrow \big(v^{n+1} = w^{n+1}\big)$ hold. According to the assertion of the theorem, we seek to establish the comparison principle $v^{n+1} \geq w^{n+1}$ for any two given sequences $v^n \geq w^n$ by using the properties (2.5) and (2.6) of the mapping H. The idea of the proof worked out in the following is to assume inductively $a \geq a'$ and $d \geq d'$ resulting in

$$b = \mathcal{F}^{-1}\left[d + \lambda f\left(a\right)\right] \geq \mathcal{F}^{-1}\left[d' + \lambda f\left(a'\right)\right] = b', \tag{2.7}$$

which is nothing but the pointwise realization of the sought monotone comparison principle.

Let $\mathbf{I}$ be the set of indices k with $v^n_k > w^n_k$, $v^n_k \in v^n$, $w^n_k \in w^n$. There are only a few possibilities for the composition of the set $\mathbf{I}$: It may consist of the empty set or a finite or infinite subset of $\mathbf{Z}$. The proof of the assertion of the theorem follows by diversion of cases and induction by the number of indices in $\mathbf{I}$. The aim is to show in each step of the induction the validity of the defined comparison principle.

Note that we neglect the choice of the value u^{n+1}_J which must be given within an algorithm. For our purpose, the index J can be considered as the minimum index in the index set $\mathbf{I}$ if $\mathbf{I}$ is not empty.

Case: $\mathbf{I} = \mathbf{I_0} = \emptyset$. The validity of the assertion is trivial for $\mathbf{I_0} = \emptyset$ since v^n is equal to w^n in that case.

Case: $\mathbf{I} \neq \emptyset$. As already indicated, the proof is done by induction.

Beginning of the induction (induction level 1): $\sharp\mathbf{I_1} = 1$ (*i.e.,* $\mathbf{I} = \mathbf{I_1}$).

Let $k =: J$ be the index in the arbitrarily chosen but fixed index set $\mathbf{I_1}$. By definition, this is equivalent to

$$v^n_J > w^n_J \quad \text{and} \quad v^n_i = w^n_i \;\forall\, i > J.$$

The indices i' with $i' < J$ are not important since we discuss nonstrictly growing monotone fluxes, i.e., any signal will be transported from left to right and

$$v^{n+1}_{i'} = v^n_{i'} \equiv w^n_{i'} \quad \forall\, i' < J$$

will hold for all considered time level indices n, $n+1$. We prove the comparison principle sought here by induction over the indices $\tilde{i} \geq J$.

Beginning of the induction (induction level 2*):* $\tilde{i} = J$.

By the properties of $\mathcal{F}$ worked out above, we have

$$v^{n+1}_J = \mathcal{F}^{-1}\left[v^n_J + \lambda f\left(v^{n+1}_{J-1}\right)\right]$$

$$\begin{aligned}
&\geq \mathcal{F}^{-1}\left[w_J^n + \lambda f\left(v_{J-1}^{n+1}\right)\right] \\
&= \mathcal{F}^{-1}\left[w_J^n + \lambda f\left(w_{J-1}^{n+1}\right)\right] \\
&= w_J^{n+1}.
\end{aligned}$$

Assumption of the induction (induction level 2*): For all indices* $\tilde{i} \geq J$ *but* $\tilde{i} \leq \bar{i}$ *it holds* $v_{\tilde{i}}^{n+1} \geq w_{\tilde{i}}^{n+1}$.
Induction step (induction level 2*):* $\bar{i} \mapsto \bar{i} + 1$.
Let us consider the situation at the index $\bar{i} + 1$. As before, we employ the properties of G to obtain

$$\begin{aligned}
v_{\bar{i}+1}^{n+1} &= \mathcal{F}^{-1}\left[v_{\bar{i}+1}^n + \lambda f\left(v_{\bar{i}}^{n+1}\right)\right] \\
&= \mathcal{F}^{-1}\left[w_{\bar{i}+1}^n + \lambda f\left(v_{\bar{i}}^{n+1}\right)\right] \\
&\geq \mathcal{F}^{-1}\left[w_{\bar{i}+1}^n + \lambda f\left(w_{\bar{i}}^{n+1}\right)\right] \\
&= w_{\bar{i}+1}^{n+1}.
\end{aligned}$$

This proves the assertion of the beginning of the induction over $\mathbf{I}$.
Assumption (induction level 1*):*
Consider $\sharp\mathbf{I_m} = m$ with $m > 1$, $m \in \mathbf{N}$ (i.e., $\mathbf{I} = \mathbf{I_m}$). Let the upwind scheme under consideration be monotone with respect to the proceeding for all subsets of $\mathbf{I_m}$, i.e., it holds $v^{n+1} \geq w^{n+1}$ for perturbations $v_i^n > w_i^n$, $i \in \tilde{\mathbf{I}}$, for any arbitrarily chosen but fixed subset $\tilde{\mathbf{I}}$ of $\mathbf{I_m}$.
Induction step (induction level 1*):* $m \mapsto m + 1$.
Now, $\sharp\mathbf{I_{m+1}} = m + 1$ is considered. For the following proceeding, it is useful to have in mind that the construction of the induction step has to be well defined, i.e., the proceeding does not rely on any specific choice of indices.

We first choose (in a well-defined fashion) two indices m_1, m_2 useful for the procedure.

Let $\mathbf{I_m} \subset \mathbf{I_{m+1}}$ hold without limitation of generality. Let m_1 be an arbitrary but fixed index with

$$m_1 \ \in \ \left[\mathbf{I_m} \subset \mathbf{I_{m+1}}\right].$$

Moreover, let m_2 be the index by which

$$m_2 \ \in \ \left[\mathbf{I_{m+1}} \setminus \mathbf{I_m}\right]$$

holds. In other words, m_1 is an index in both $\mathbf{I_m}$ and $\mathbf{I_{m+1}}$, while m_2 corresponds exactly to the difference between the given index set $\mathbf{I_m}$ and the new index set $\mathbf{I_{m+1}}$.

By the assumption of the induction, the implicit upwind scheme is monotone with respect to positive changes in values corresponding to the index set $\mathbf{I_m}$. This means in particular that a positive change in $v_{m_1}^n$ together with positive changes in other values v_k^n, $k \in \mathbf{I_m} \setminus \{m_1\}$, leads to nonnegative changes in the sequence v^{n+1}.

In order to undertake the induction step, we have to consider now the effect of the additional nonnegative perturbation in the index m_2.

Thus, simultaneous positive changes in $v_{m_1}^n$ and $v_{m_2}^n$ are considered while in the background there are arbitrary but fixed positive changes in the values corresponding to

$$\bar{\mathbf{I}}_{\mathbf{m}} \ := \ \mathbf{I_{m+1}} \setminus \{m_1, m_2\} \quad \subset \quad \mathbf{I_m}$$

due to the assumption of the induction.

Let us briefly consider the effects of the latter positive perturbations in the values corresponding to $\bar{\mathbf{I}}_{\mathbf{m}}$. We recall that by the assumption of the induction, the application of the upwind scheme has the effect that the comparison principle $v^{n+1} \geq w^{n+1}$ holds for positive changes in v_k^n, $k \in \bar{\mathbf{I}}_{\mathbf{m}}$. Let us fix these influences by denoting the data computed by positive perturbations in v_k^n, $k \in \bar{\mathbf{I}}_{\mathbf{m}}$, by $\bar{v}^{n+1}$. Having defined this notion, we know that $\bar{v}^{n+1} \geq w^{n+1}$ holds by the assumption of the induction step because of $\bar{\mathbf{I}}_{\mathbf{m}} \subset \mathbf{I}_{\mathbf{m}}$. Furthermore, we have achieved that $\bar{v}^{n+1}$ denotes the state before taking into account further changes caused by nonnegative perturbations in $v_{m_1}^n$ and $v_{m_2}^n$.

We now focus our attention especially on the indices m_1 and m_2 defined above. Let Δ_j^1 be a change in $\bar{v}_j^{n+1}$ induced by a positive change in $v_{m_1}^n$. Then Δ_j^1 is always nonnegative by the assumption of the induction since $m_1 \in \mathbf{I}_{\mathbf{m}}$. Analogously, let Δ_j^2 be a change in $\bar{v}_j^{n+1}$ induced by a positive change in $v_{m_2}^n$. The change Δ_j^2 is also nonnegative because of the assumed validity of condition (2.6) and application of the procedure within the level two induction.

Thus, the single effects of separately considered nonnegative perturbations at m_1 and m_2 are nonnegative. Now, for the mutual effects of such changes in data corresponding to an arbitrary but fixed index i, $i \notin \{m_1, m_2\}$, there are only two possibilities because of $m_1 \neq m_2$:

1. $m_1 > m_2$:
   We set $J := m_2$. The procedure within the level two induction especially reveals $v_{m_1-1}^{n+1} \geq w_{m_1-1}^{n+1}$, thus we obtain the estimate $v_{m_1}^{n+1} \geq w_{m_1}^{n+1}$ in the same fashion as in (2.7). Then the assertion follows by the same type of induction as in the level two induction over $\mathbf{I}$.
2. $m_2 > m_1$:
   We set $J := m_1$. The assertion follows analogously to the case $m_1 > m_2$.

Thus, the mutual effects cannot result in a different situation as investigated up to now. Note the arbitrary choice of m_1 and m_2 together with simultaneous changes in the data corresponding to the index set $\bar{\mathbf{I}}_{\mathbf{m}}$. Since there are no limitations concerning the choice of the index set $\mathbf{I}_{\mathbf{m}}$, the proceeding as a whole is well defined and so the proof is finished. □

By the monotonicity of the implicit upwind method, one can easily derive the following stability statement. This is not the nonlinear stability we seek, but it is needed as a technical assertion in the convergence proof.

Theorem 2.2 (L_∞-stability). *Let the flux f be a nonstrictly growing monotone function. Then the numerical solution obtained by the implicit upwind scheme* (2.1) *satisfies*

$$||u^n||_{L_\infty} \quad \leq \quad ||u_0||_{L_\infty}.$$

Proof. Let a sequence $u^n \in L_\infty$ be given. With $a_k := \inf_{j\in\mathbf{Z}}(u_j^k)$ and $b_k := \sup_{j\in\mathbf{Z}}(u_j^k)$ we define sequences a^n, a^{n+1} and b^n, b^{n+1} by $a_j^k := a_k$ and $b_j^k := b_k$ for all $j \in \mathbf{Z}$. Since the method is conservative,

$$a_{n+1} \;=\; u_j^n - \lambda\left\{g(a_{n+1}, a_{n+1}) - g(a_{n+1}, a_{n+1})\right\} \;\geq\; a_n \tag{2.8}$$

$$\text{and} \quad b_{n+1} \;=\; u_j^n - \lambda\left\{g(b_{n+1}, b_{n+1}) - g(b_{n+1}, b_{n+1})\right\} \;\leq\; b_n \tag{2.9}$$

holds for the sequences a^{n+1} and b^{n+1}, respectively. Consequently, by the monotonicity of the method established in Theorem 2.1 follows $(b^n \geq u^n) \Rightarrow \left(b^{n+1} \geq u^{n+1}\right)$ and

$(u^n \geq a^n) \Rightarrow \left(u^{n+1} \geq a^{n+1}\right)$ which implies the L_∞-stability of the method by (2.8) and (2.9). □

Note that the proof given above is valid for general monotone 3-point schemes.

Concerning convergence of the scheme (2.1) to the entropy solution, the following definition is useful.

Definition 2.4 (entropy consistency). *The implicit upwind scheme* (2.1) *is consistent with the entropy condition of Kružkov if there exists a continuous numerical entropy flux G which satisfies for all $l \in \mathbf{R}$ the following assertions:*

1. *Consistency with the entropy flux of Kružkov*

$$G(v;l) = sgn(v-l)\,[f(v)-f(l)]\ \ \forall v. \tag{2.10}$$

2. *Validity of a discrete entropy inequality*

$$\frac{U(u_j^{n+1};l)-U(u_j^n;l)}{\Delta t} \leq -\frac{G(u_j^{n+1};l)-G(u_{j-1}^{n+1};l)}{\Delta x}, \tag{2.11}$$

where $U(v;l) = |v-l|$ is chosen due to Kružkov.

For the sake of brevity, let

$$a \vee b := \max(a,b) \quad \text{and} \quad a \wedge b := \min(a,b) \tag{2.12}$$

hold. The important connection between the numerical entropy flux G and the numerical flux function g of the scheme (2.1) can now be established.

Lemma 2.1. *Consider the implicit upwind scheme* (2.1) *and nonstrictly growing monotone fluxes. Then the numerical entropy flux defined by*

$$G(v;l) \ := \ g(v \vee l) - g(v \wedge l) \tag{2.13}$$

is consistent with the entropy flux of Kružkov.

Proof. By use of g as in (2.3) we immediately obtain

$$G(v;l) = f(v \vee l) - f(v \wedge l) = sgn(v-l)[f(v)-f(l)]$$

by use of (2.12) for all $l \in \mathbf{R}$. □

Now follow one of the main assertions described in this paper. The proceeding within the proof is a variation of a procedure employed by Crandall and Majda [4].

Theorem 2.3. *Consider the implicit upwind scheme* (2.1) *which is consistent, conservative, and monotone for nonstrictly growing monotone fluxes. Then the scheme is also consistent with the entropy condition of Kružkov.*

Proof. By Lemma 2.1 a numerical entropy flux can be associated with the considered scheme. Thus, the consistency with the entropy flux of Kružkov is given. It remains to show the validity of a discrete entropy inequality in the sense of (2.11). Therefore, let $l \in \mathbf{R}$ be chosen arbitrarily but fixed. Using (2.12), one can derive

$$\begin{aligned} -\lambda\left\{G(u_j^{n+1};l) - G(u_{j-1}^{n+1};l)\right\} = \\ H\left(u_{j-1}^{n+1} \vee l, u_j^{n+1} \vee l; u_j^n \vee l\right) \\ -H\left(u_{j-1}^{n+1} \wedge l, u_j^{n+1} \wedge l; u_j^n \wedge l\right) - \left|u_j^n - l\right|. \end{aligned} \tag{2.14}$$

We now show the validity of

$$H\left(u_{j-1}^{n+1} \vee l, u_j^{n+1} \vee l; u_j^n \vee l\right) \ \geq \ u_j^{n+1} \vee l \tag{2.15}$$

and

$$H(u_{j-1}^{n+1} \wedge l, u_j^{n+1} \wedge l; u_j^n \wedge l) \ \leq \ u_j^{n+1} \wedge l. \tag{2.16}$$

As we will see, these relations follow straightforwardly by diversion of the cases $u_j^{n+1} \geq l$ and $u_j^{n+1} < l$ using the monotonicity conditions (2.5) and (2.6).
For the validation of (2.15) we compute the following cases.
Case $u_j^{n+1} \geq l$,

$$\begin{aligned}
&H\left(u_{j-1}^{n+1} \vee l, u_j^{n+1} \vee l; u_j^n \vee l\right) \\
&\overset{case}{=} H\left(u_{j-1}^{n+1} \vee l, u_j^{n+1}; u_j^n \vee l\right) \\
&\overset{(2.5)}{\geq} H\left(u_{j-1}^{n+1}, u_j^{n+1}; u_j^n \vee l\right) \\
&\overset{(2.6)}{\geq} H\left(u_{j-1}^{n+1}, u_j^{n+1}; u_j^n\right) \\
&= \ u_j^{n+1} \\
&\overset{case}{=} u_j^{n+1} \vee l, \text{ and}
\end{aligned}$$

Case $u_j^{n+1} < l$,

$$\begin{aligned}
&H\left(u_{j-1}^{n+1} \vee l, u_j^{n+1} \vee l; u_j^n \vee l\right) \\
&\overset{case}{=} H\left(u_{j-1}^{n+1} \vee l, l; u_j^n \vee l\right) \\
&\overset{(2.5)}{\geq} H\left(l, l; u_j^n \vee l\right) \\
&\overset{(2.6)}{\geq} H\left(l, l; l\right) \\
&= \ l \\
&\overset{case}{=} u_j^{n+1} \vee l.
\end{aligned}$$

Analogously, we compute for the validation of (2.16).
Case $u_j^{n+1} \geq l$,

$$\begin{aligned}
&H\left(u_{j-1}^{n+1} \wedge l, u_j^{n+1} \wedge l; u_j^n \wedge l\right) \\
&\overset{case}{=} H\left(u_{j-1}^{n+1} \wedge l, l; u_j^n \wedge l\right) \\
&\overset{(2.5)}{\leq} H\left(l, l; u_j^n \wedge l\right) \\
&\overset{(2.6)}{\leq} H\left(l, l; l\right) \\
&= \ l \\
&\overset{case}{=} u_j^{n+1} \wedge l, \text{ and}
\end{aligned}$$

Case $u_j^{n+1} < l$,

$$\begin{aligned}
&H\left(u_{j-1}^{n+1} \wedge l, u_j^{n+1} \wedge l; u_j^n \wedge l\right) \\
&\overset{case}{=} H\left(u_{j-1}^{n+1} \wedge l, u_j^{n+1}; u_j^n \wedge l\right) \\
&\overset{(2.5)}{\leq} H\left(u_{j-1}^{n+1}, u_j^{n+1}; u_j^n \wedge l\right)
\end{aligned}$$

$$\begin{aligned} &\overset{(2.6)}{\leq} H\left(u_{j-1}^{n+1}, u_j^{n+1}; u_j^n\right) \\ &= u_j^{n+1} \\ &\overset{case}{=} u_j^{n+1} \wedge l. \end{aligned}$$

Using $|a-b| = a \vee b - a \wedge b$, one can easily derive from (2.14) via the estimates (2.15) and (2.16) the desired discrete entropy inequality. $\square$

2.2. Convergence. We now give the convergence proof for the method (2.1) under the usual assumption of nonstrictly growing monotone fluxes.

Since the proof relies on the theorem of monotone convergence of Beppo Levi which is not common in the conservation laws literature, we first state it for convenience.

THEOREM 2.4 (Beppo Levi). *Let $h_k : \mathbf{R} \to \mathbf{R}$, $k \in \mathbf{N}$, be a sequence of integrable functions with $h_k \leq h_{k+1}$ (in the sense of pointwise comparison a.e.) for all $k \in \mathbf{N}$. If*

$$\lim_{k\to\infty} \int h_k(x)\,dx =: C < \infty$$

holds, then the function

$$h := \lim_{k\to\infty} h_k$$

is integrable and it holds

$$\int h(x)\,dx = \lim_{k\to\infty} \int h_k(x)\,dx.$$

The basic idea behind the convergence proof is the following: corresponding to a sequence $\Delta x_k \downarrow 0$, we construct a monotonously growing sequence of initial data. Then, by the monotonicity of the method, we obtain a monotonously growing sequence of numerical solutions. Since we multiply the initial function u_0 with an arbitrarily chosen but fixed test function with compact support, we only have to consider u_0 over a compact domain. Because of the assumption $u_0 \in L_\infty$ and since we have L_∞-stability, the function sequence corresponding to the numerical solutions obtained in the limit $k \to \infty$ is integrable and bounded from above. Then we can use the theorem of monotone convergence of Beppo Levi to show convergence (almost everywhere) to a limit function. More formally, we state the following theorem.

THEOREM 2.5 (Convergence of the implicit upwind method). *Let $u_0(x)$ be in $L_\infty^{loc}(\mathbf{R})$. Let the flux function of a given conservation law grow in a nonstrictly monotone fashion and be at least continuous. Consider a sequence of nested grids indexed by $k = 1, 2, \ldots$, with mesh parameters $\Delta t_k, \Delta x_k \to 0$ as $k \to \infty$, and let $u_k(x,t)$ be the numerical approximation obtained via the implicit upwind scheme. Then u_k converges to the unique entropy solution of the given conservation law as $k \to \infty$.*

Proof. For brevity of the notation, we omit the arguments (x,t) when appropriate in the following within this proof.

The most important technical detail is the special discretization of the initial condition $u_0 \in L_\infty^{loc}(\mathbf{R})$. Therefore, we use nested grids in order to compare data sets of values; i.e., refined grids always inherit cell borders.

After a suitable manipulation of u_0 on a set of Lebesgue measure zero if necessary, the initial condition is discretized on cell j by

$$u_j^0 \ := \ \inf_{x \in (\Delta x_0 (j-1), j\Delta x_0]} u_0(x). \tag{2.17}$$

Corresponding to the initial data we also define a piecewise constant function

$$u_k(x,0) := u_j^0 \quad \text{for} \quad (j-1)\Delta x_k < x \leq j\Delta x_k. \tag{2.18}$$

It is a simple matter of classical analysis not described in this paper to verify that the discretization (2.17) together with (2.18) gives on any compact interval on the x-axis a monotonously growing function sequence with

$$\lim_{k\to\infty} u_k(x,0) \ = \ u_0(x) \text{ almost everywhere} \tag{2.19}$$

by application of the theorem of monotone convergence.

Similarly, we extract discrete test elements ϕ_j^0 out of a given test function $\phi \in C_0^\infty(\mathbf{R} \times \mathbf{R}_+)$, where $\phi \geq 0$ holds in a pointwise sense, by setting

$$\begin{aligned} \phi_j^0 \ &:= \ \phi(x_j, 0) \quad \text{and} \\ \phi_k(x,0) &= \phi_j^0 \quad \text{for} \quad x_j - \Delta x_k < x \leq x_j. \end{aligned} \tag{2.20}$$

Also, we define for $n \geq 1$ the step function

$$u_k(x,t) = u_j^n \quad \text{for} \quad x_j - \Delta x_k < x \leq x_j \quad \text{and} \quad t^{n-1} < t \leq t^n.$$

Analogously, the definition of the function $\phi_k(x,0)$ can be extended to arguments $t > 0$.

Let the test function ϕ be chosen arbitrarily but fixed. Multiplication of the discrete entropy condition (2.11) corresponding to the implicit upwind scheme (2.1) with $\Delta x_k \Delta t_k$ as well as with the discrete test element ϕ_j^{n+1}, summation over $j \in \mathbf{Z}$ and $n \geq 0$ and finally summation by parts yields

$$\begin{aligned} -\Delta x \sum_{j\in\mathbf{Z}} |u_j^0 - l|\phi_j^0 \ \leq \ \Delta x \Delta t \sum_{j\in\mathbf{Z}} \sum_{n\geq 0} \Bigg[& |u_j^n - l| \frac{\phi_j^{n+1} - \phi_j^n}{\Delta t} \\ & + sgn\left(u_j^{n+1} - l\right) \left[f(u_j^{n+1}) - f(l)\right] \frac{\phi_{j+1}^{n+1} - \phi_j^{n+1}}{\Delta x} \Bigg]. \end{aligned} \tag{2.21}$$

We now prove convergence of (2.21) towards the weak form of the entropy condition due to Kružkov,

$$\int_0^\infty \int_{-\infty}^\infty \Big[|u-l|\phi_t + sgn\,(u-l)\,[f(u) - f(l)]\,\phi_x \Big]\, dxdt \geq - \int_{-\infty}^\infty |u_0 - l|\phi_0 \, dx. \tag{2.22}$$

Therefore, we have to consider an arbitrarily chosen but fixed test element which consists of a 2-tupel (ϕ, l) composed of a test function ϕ with $\phi \geq 0$, $\phi \in C_0^\infty(\mathbf{R}\times\mathbf{R}_+)$, and a test number $l \in \mathbf{R}$.

By (2.18) and (2.20) as well as by the developed notions, the inequality (2.21) is equivalent to

$$\begin{aligned} - \int_{-\infty}^\infty |u_k(x,0) - l| \, \phi_k(x,0) \, dx \leq \int_0^\infty \int_{-\infty}^\infty \Big[& |u_k(x,t) - l| \phi_t' \\ & + sgn\,(u_k(x, t+\Delta t_k) - l)\, [f(u_k(x,t+\Delta t_k)) - f(l)]\, \phi_x' \Big]\, dxdt \end{aligned} \tag{2.23}$$

with

$$\phi'_t := \frac{\phi_k(x, t + \Delta t_k) - \phi_k(x,t)}{\Delta t_k} \tag{2.24}$$

and

$$\phi'_x := \frac{\phi_k(x + \Delta x_k, t + \Delta t_k) - \phi_k(x, t + \Delta t_k)}{\Delta x_k}. \tag{2.25}$$

We first investigate the left-hand side of (2.23). Therefore, let

$$\begin{aligned} &K_\phi := \mathit{support}(\phi) \cap \{(x,t) \,|\, t = 0\} \quad \text{and} \\ &K := \{x \,|\, \exists y \in K_\phi :\ y - \Delta x_0 \le x \le y + \Delta x_0\}. \end{aligned}$$

By construction, K is compact and gives the largest possible interval on the x-axis where nonzero discrete data may occur.

Adding zero, we now cast the problem in a first step into the form

$$\begin{aligned} &\int_K |u_k(x,0) - l| \phi_k(x,0)\,dx \\ &= \int_K |u_k(x,0) - l| \big[\phi_k(x,0) - \phi(x,0) + \phi(x,0)\big]\,dx \\ &= \underbrace{\int_K |u_k(x,0) - l| \big[\phi_k(x,0) - \phi(x,0)\big]\,dx}_{=:(a)} + \underbrace{\int_K |u_k(x,0) - l| \phi(x,0)\,dx}_{=:(b)}. \end{aligned} \tag{2.26}$$

We now discuss the value of (2.26)(a) in the limit $k \to \infty$. Because of $u_0 \in L^\infty(\mathbf{R})$ and since l is chosen arbitrarily but fixed (i.e., $l < \infty$), we can estimate the term $|u_k(x,0) - l|$ in (2.26)(a) by the help of a constant $M_{u,l} < \infty$ as

$$\left| \int_K |u_k(x,0) - l| \big[\phi_k(x,0) - \phi(x,0)\big]\,dx \right| \le M_{u,l} \int_K |\phi_k(x,0) - \phi(x,0)|\;dx. \tag{2.27}$$

Since ϕ is a smooth test function, it is a simple but technical exercise of classical analysis to prove

$$\sup_{x \in K} |\phi_k(x,0) - \phi(x,0)| \overset{k\to\infty}{\longrightarrow} 0;$$

i.e., $||\phi_k(x,0) - \phi(x,0)||_\infty \to 0$ for $k \to \infty$. Using this, we receive from (2.27)

$$M_{u,l} \int_K |\phi_k(x,0) - \phi(x,0)|\;dx \le M_{u,l}\,|K| \sup_{x \in K} |\phi_k(x,0) - \phi(x,0)|\,. \tag{2.28}$$

The estimations (2.27) and (2.28) imply that the Limes for $k \to \infty$ of the investigated term goes to zero. We now turn to (2.26)(b). It is useful to add zero again in order to obtain

$$\begin{aligned} &\int_K |u_k(x,0) - l| \phi(x,0)\,dx \\ &= \int_K \big|u_k(x,0) - u(x,0) + u(x,0) - l\big| \phi(x,0)\,dx \\ &\le \underbrace{\int_K |u(x,0) - l| \phi(x,0)\,dx}_{=:(b_1)} + \underbrace{\int_K |u_k(x,0) - u(x,0)| \phi(x,0)\,dx}_{=:(b_2)}. \end{aligned} \tag{2.29}$$

Our aim is to prove that (2.29)(b_2) vanishes in the limit for $k \to \infty$. Since ϕ is continuous, we can estimate the absolute of the term (2.29)(b_2) with the help of a constant $M_\phi < \infty$ by

$$\left| \int_K |u_k(x,0) - u(x,0)| \, \phi(x,0) \, dx \right| \le M_\phi \int_K |u_k(x,0) - u_0(x)| \, dx.$$

By our construction (2.17), (2.18), $u_k(x,0)$ approaches $u_0(x)$ from below; i.e.,

$$M_\phi \int_K |u_k(x,0) - u_0(x)| \, dx = M_\phi \int_K u_0(x) - u_k(x,0) \, dx \quad \forall k.$$

Using the theorem of monotone convergence, i.e., (2.19), implies that

$$\int_K u_0(x) - u_k(x,0) \, dx$$

vanishes in the limit for $k \to \infty$. Thus, the term (2.29)(b_2) goes to zero for $k \to \infty$. To condense these results, we get by the described procedure and use of the identity $u(x,0) \equiv u_0(x)$

$$\lim_{k\to\infty} \int_{\mathbf{R}} u_k(x,0)\phi_k(x,0) \, dx \le \int_{\mathbf{R}} u_0(x)\phi(x,0) \, dx,$$

i.e.,

$$-\lim_{k\to\infty} \int_{\mathbf{R}} u_k(x,0)\phi_k(x,0) \, dx \ge -\int_{\mathbf{R}} u_0(x)\phi(x,0) \, dx$$

which displays the appropriate order with respect to the underlying relation (2.21) and the sought inequality (2.22). By analogously defining a compact domain S including the support of ϕ in space and time and using the attributes of test functions, it remains to be shown that

$$\begin{aligned} &\int_S sgn\,(u_k(x,t+\Delta t_k) - l)\,\big[f(u(x,t)) - f(l)\big]\phi_x(x,t)\,dxdt \\ \text{(2.30)} \qquad &\overset{k\to\infty}{\longrightarrow} \int_S sgn\,(u(x,t) - l)\,\big[f(u(x,t)) - f(l)\big]\phi_x(x,t)\,dxdt \end{aligned}$$

as well as

$$\text{(2.31)} \qquad \int_S \big|u(x,t) - u_k(x,t)\big| \, |\phi_t(x,t)| \, dxdt \overset{k\to\infty}{\longrightarrow} 0$$

whereby (2.31) was obtained after adding zeros in a similar fashion as within the proceeding before.

Let us first discuss the latter expression (2.31). Since ϕ_t is continuous on S, we can estimate $|\phi_t|$ in (2.31) by a constant $M_t < \infty$.

Since we obtain in the limit $k \to \infty$ a monotonously growing sequence of numerical approximations in the sense of pointwise comparison and since it is bounded from above because of $u_0 \in L_\infty(S)$ and the monotonicity of the method, the function sequence $(u_k(x,t))_{k\in\mathbf{N}}$ converges almost everywhere to an integrable limit function on S by the theorem of monotone convergence due to Levi. We set

$$u(x,t) := \lim_{k\to\infty} u_k(x,t).$$

Introducing exactly this limit function as the function $u(x,t)$ used up to now, the expression on the left-hand side of (2.31) becomes zero in the limit

$$\lim_{k\to\infty}\int_S \big|u(x,t)-u_k(x,t)\big|\,|\phi_t(x,t)|\;dxdt \le M_t\int_S u(x,t)-\lim_{k\to\infty}u_k(x,t)\,dxdt=0.$$

Note that the pointwise convergence $u_k\to u$ almost everywhere is now established and can be used in the following.

By definition of the symbols "$\vee$" and "$\wedge$" from (2.12), it holds that

$$\begin{aligned}
&\lim_{k\to\infty}\Big|\int_S sgn\,(u_k(x,t+\Delta t_k)-l)\,\big[f(u_k(x,t+\Delta t_k))-f(l)\big]\phi_x(x,t)\,dxdt\\
&-sgn\,(u(x,t)-l)\,\big[f(u(x,t))-f(l)\big]\phi_x(x,t)\,dxdt\Big|\\
\le{}&\lim_{k\to\infty}\int_S\Big|\Big[sgn\,(u_k(x,t+\Delta t_k)-l)\,\big[f(u_k(x,t+\Delta t_k))-f(l)\big]\\
&-sgn\,(u(x,t)-l)\,\big[f(u(x,t))-f(l)\big]\Big]\phi_x(x,t)\Big|\,dxdt\\
={}&\lim_{k\to\infty}\int_S\Big|\Big[\big[f(u_k(x,t+\Delta t_k)\vee l)-f(u_k(x,t+\Delta t_k)\wedge l)\big]\\
&-\big[f(u(x,t)\vee l)-f(u(x,t)\wedge l)\big]\Big]\phi_x(x,t)\Big|\,dxdt.
\end{aligned}\tag{2.32}$$

By adding zero, we obtain equivalently to (2.32) the expression

$$\begin{aligned}
\lim_{k\to\infty}\int_S\Big|\Big[&\big[f(u_k(x,t+\Delta t_k)\vee l)-f(u_k(x,t)\vee l)\big]\\
&-\big[f(u_k(x,t+\Delta t_k)\wedge l)-f(u_k(x,t)\wedge l)\big]\\
&-\big[f(u(x,t)\vee l)-f(u_k(x,t)\vee l)\big]\\
&+\big[f(u(x,t)\wedge l)-f(u_k(x,t)\wedge l)\big]\Big]\phi_x(x,t)\Big|\,dxdt.
\end{aligned}\tag{2.33}$$

The latter expression (2.33) can be estimated from above using $|\phi_x(x,t)|<M_x<\infty$ by

$$M_x\lim_{k\to\infty}\int_S\big|f(u_k(x,t+\Delta t_k)\vee l)-f(u_k(x,t)\vee l)\big|\,dxdt \tag{2.34}$$

$$+M_x\lim_{k\to\infty}\int_S\big|f(u_k(x,t+\Delta t_k)\wedge l)-f(u_k(x,t)\wedge l)\big|\,dxdt \tag{2.35}$$

$$+M_x\lim_{k\to\infty}\int_S\big|f(u(x,t)\vee l)-f(u_k(x,t)\vee l)\big|\,dxdt \tag{2.36}$$

$$+M_x\lim_{k\to\infty}\int_S\big|f(u(x,t)\wedge l)-f(u_k(x,t)\wedge l)\big|\,dxdt. \tag{2.37}$$

Let us discuss now the terms (2.34)–(2.37). The terms denoted via (2.34) and (2.35) vanish because of the continuity in the mean of the combined function $f\circ u_k$. Because of Theorem 2.2 it holds that $u_k\in L_\infty$, and since we have chosen an arbitrary but fixed test number l ($<\infty$), we can find finite estimates from above for the integrands in (2.36) and (2.37). Thus, these terms vanish when using the established pointwise convergence $u_k\to u$ almost everywhere and the theorem of Lebesgue. Thus, convergence to the entropy solution is shown. $\square$

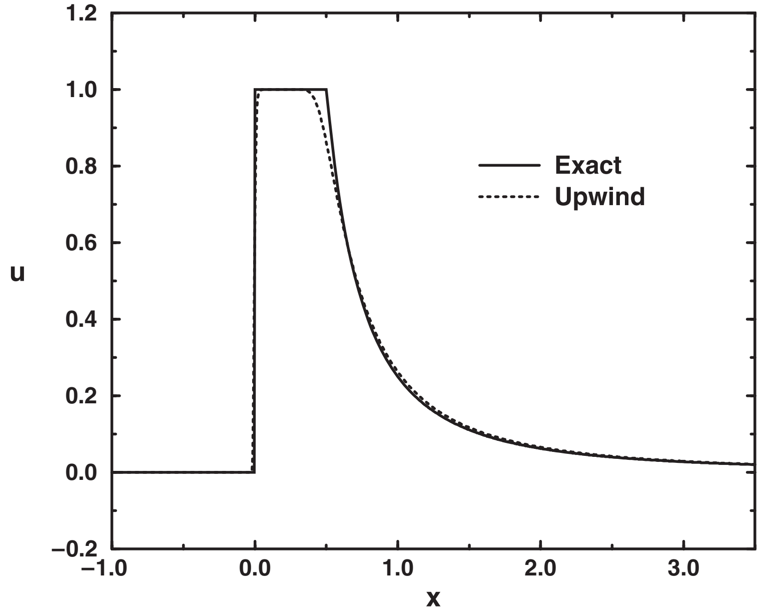


Fig. 3.1. *Numerical solution of the described example due to Kružkov, obtained by the implicit upwind method.*

3. Numerical tests.

In order to show the usefulness of the developed notions, we investigate the truly nonstandard conservation law developed by Kružkov and Panov [12] we already mentioned in the introduction, i.e., we investigate exactly the example given by (1.1), (1.2), and (1.3). For the numerical solution we use a grid of 1000 points with $\Delta x = 0.005$ and a time step size of $\Delta t = 0.00075$. We set $\alpha = 0.5$ within the flux function described above.

The solution obtained by the use of the implicit upwind scheme is displayed in Figure 3.1. The ansatz of the rarefaction wave is a bit smeared which can be explained by the monotonicity property and the employed relatively coarse grid.

4. Conclusive remarks.

We have described the fundament of a theory of implicit methods for conservation laws, extending the rigorously verified range of applicability of numerical methods. The scheme we focused our attention on is the upwind scheme, the monotonicity property of other implicit methods is the subject of a forthcoming paper.

Acknowledgements. The author is very thankful for the constructive comments of the referees which helped to improve the paper. He would also like to express his gratitude to Professor Rainer Ansorge from the University of Hamburg.

REFERENCES

[1] P. Bénilan and S. N. Kružkov, *Conservation laws with continuous flux functions*, NoDEA, 3 (1996), pp. 395–420.

[2] M. Breuss, *Numerical Methods for Conservation Laws in Non-Standard-Situations*, Ph.D. Thesis, University of Hamburg, Hamburg, Germany, 2001 (in German).

[3] F. Coquel and P. Le Floch, *Convergence of finite difference schemes for conservation laws in several space dimensions: A general theory*, SIAM J. Numer. Anal., 30 (1993), pp. 675–700.

[4] M. G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[5] R. J. DiPerna, *Measure-valued solutions to conservation laws*, Arch. Rational Mech. Anal., 88 (1985), pp. 223–270.

[6] T. Gimse, *Conservation laws with discontinuous flux functions*, SIAM. J. Math. Anal., 24 (1993), pp. 279–289.

[7] E. Godlewski and P.-A. Raviart, *Hyperbolic Systems of Conservation Laws*, Ellipses, Paris, 1991.

[8] A. Harten, *On a class of high resolution total-variation-stable finite-difference schemes*, SIAM J. Numer. Anal., 21 (1984), pp. 1–23.

[9] A. Harten, J. M. Hyman, and P. D. Lax, *On finite difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math., 29 (1976), pp. 297–322.

[10] S. N. Kružkov, *First order quasilinear equations in several independent variables*, Math. USSR Sbornik, 10 (1970), pp. 217–243.

[11] S. N. Kružkov and F. Hildebrand, *The Cauchy problem for first-order quasilinear equations when the domain of dependence on initial data is infinite*, Moscow Univ. Math. Bull., 29 (1974), pp. 75–81.

[12] S. N. Kružkov and E. Y. Panov, *Conservative quasilinear first-order laws with an infinite domain of dependence on the initial data*, Soviet Math. Dokl., 42 (1991), pp. 316–321.

[13] N. N. Kuznetsov, *Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation*, USSR Comput. Math. Math. Phys., 16 (1976), pp. 105–119.

[14] R. J. LeVeque, *Numerical Methods for Conservation Laws*, 2nd ed., Birkhäuser Verlag, 1992.

[15] S. Osher, *Riemann solvers, the entropy condition, and difference approximations*, SIAM J. Numer. Anal., 21 (1984), pp. 217–235.

[16] R. Sanders, *On convergence of monotone difference schemes with variable spatial differencing*, Math. Comput., 40 (1983), pp. 91–106.

[17] H. C. Yee, *A class of high-resolution explicit and implicit shock-capturing methods*, J. Appl. Sci. Comput., 2 (1995), pp. 1–226.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 987–1002

ON THE CONVERGENCE OF A GENERAL CLASS OF FINITE VOLUME METHODS*

HOLGER WENDLAND†

Abstract. In this paper we investigate numerical methods for solving hyperbolic conservation laws based on finite volumes and optimal recovery. These methods can, for example, be applied in certain ENO schemes. Their approximation properties depend in particular on the reconstruction from cell averages. Hence, this paper is devoted to prove convergence results for such reconstruction processes from cell averages.

Key words. optimal recovery, finite volumes, positive definite kernels, approximation orders

AMS subject classifications. 65M20 65M15 41A30

DOI. 10.1137/040612993

1. Introduction. Finite volume methods are well-established tools for solving hyperbolic conservation laws of the form

$$\frac{\partial}{\partial t}u + \sum_{\ell=1}^{d} \frac{\partial}{\partial x_\ell} f_\ell(u) = 0 \tag{1.1}$$

numerically. Here, $u : \mathbb{R}^d \times [0, \infty) \to \mathbb{R}^n$ is the vector-valued solution containing the quantity to be conserved while $f_\ell : \mathbb{R}^n \to \mathbb{R}^n$ denote the so-called flux functions.

For discretizing in space, finite volume methods use cell average information. To be more precise, for a fixed time such cell averages are employed to reconstruct the unknown function u approximately. For a good reconstruction in regions where the solution of (1.1) is known or expected to be smooth, a higher order reconstruction scheme is desirable. Hence, such high order schemes currently form a major research direction in the theory of finite volumes.

The first higher order reconstruction schemes employed were based on polynomials and suffered from the typical behavior of multivariate polynomials, such as oscillation and ill-conditioning.

Sonar and Iske [4] and Sonar [9, 10, 11] proposed to employ optimal recovery based on conditionally positive definite kernels instead. Sonar's numerical examples indicate that these recovery processes indeed lead to higher order schemes. Nonetheless, up to now there has been no mathematical proof given for this observation. In [11], he concluded with "[...] nearly nothing is known about approximation orders in the case of recovery from cell average data. [...] At the moment, however, we are faced with the fact that important theoretical results are missing in this area of research."

It is the goal of this paper to fill this theoretical gap and to show that the recovery process can lead to arbitrary high orders, provided the target function u is sufficiently smooth and the correct (conditionally) positive definite kernel is employed.

However, since our analysis is based upon approximation properties of polynomials, our proof will need slightly larger *stencils* than those proposed by Sonar. On the

*Received by the editors August 6, 2004; accepted for publication (in revised form) January 4, 2005; published electronically September 20, 2005.
http://www.siam.org/journals/sinum/43-3/61299.html

†Institut für Numerische und Angewandte Mathematik, Universität Göttingen, Lotzestr. 16-18, D-37083 Göttingen, Germany (wendland@math.uni-goettingen.de).

other hand, since the "correct" selection of stencils is still under investigation, our results might also contribute to this problem.

Moreover, the results we will achieve are not restricted to (conditionally) positive definite kernels at all. On the contrary, they will work for every interpolatory and stable reconstruction process. Finally, our results are established for an *arbitrary* space dimension.

This paper is organized as follows. In the rest of the section we will introduce some general notations we will need to state our convergence results. Section 2 is devoted to a short review on finite volume and ENO (essentially nonoscillatory) schemes. Section 3 describes how such schemes can be derived using optimal recovery. Sections 4 and 5 are the main sections where we provide our error analysis. In section 6 we take a special look at thin-plate spline approximation, which is one of the most popular reconstruction methods in this context. For numerical examples we refer the reader to the previously mentioned papers by Sonar.

We will establish our error estimates using a variety of Sobolev spaces, which we want to introduce now. Let $\Omega \subseteq \mathbb{R}^d$ be a domain. For $k \in \mathbb{N}_0$ and $1 \le p < \infty$, we define the Sobolev spaces $W_p^k(\Omega)$ to consist of all u with distributional derivatives $D^\alpha u \in L_p(\Omega)$, $|\alpha| \le k$. Associated with these spaces are the (semi-)norms

$$|u|_{W_p^k(\Omega)} = \left(\sum_{|\alpha|=k} \|D^\alpha u\|_{L_p(\Omega)}^p \right)^{1/p} \text{ and } \|u\|_{W_p^k(\Omega)} = \left(\sum_{|\alpha|\le k} \|D^\alpha u\|_{L_p(\Omega)}^p \right)^{1/p}.$$

The case $p = \infty$ is defined in the following obvious way:

$$|u|_{W_p^k(\Omega)} = \sup_{|\alpha|=k} \|D^\alpha u\|_{L_\infty(\Omega)} \text{ and } \quad \|u\|_{W_\infty^k(\Omega)} = \sup_{|\alpha|\le k} \|D^\alpha u\|_{L_\infty(\Omega)}.$$

We will also be dealing with fractional order Sobolev spaces. Let $1 \le p < \infty$, $k \in \mathbb{N}_0$, and $0 < s < 1$. We define the fractional order Sobolev spaces $W_p^{k+s}(\Omega)$ to be all u for which the following (semi-)norms are finite:

$$|u|_{W_p^{k+s}(\Omega)} := \left(\sum_{|\alpha|=k} \int_\Omega \int_\Omega \frac{|D^\alpha u(x) - D^\alpha u(y)|^p}{\|x-y\|_2^{d+ps}} dxdy \right)^{1/p},$$
$$\|u\|_{W_p^{k+s}(\Omega)} := \left(\|u\|_{W_p^k(\Omega)}^p + |u|_{W_p^{k+s}(\Omega)}^p \right)^{1/p}.$$

2. Finite volume and ENO schemes.

Finite volume schemes introduce weak solutions to (1.1) in the following sense. If $V \subseteq \mathbb{R}^d$ is an arbitrary compact, small region, called the *control volume*, then u has to satisfy the weak form of the conservation law (1.1) in the form

$$\frac{d}{dt} \int_V u(x,t)dx = -\int_{\partial V} \sum_{\ell=1}^d f_\ell(u(x,t))\eta_\ell(x)dS, \tag{2.1}$$

where $\eta(x)$ denotes the outer normal vector to the boundary ∂V. This form of (1.1) often directly results from the physical conservation law and is then in a certain sense even more natural than (1.1).

To convert (2.1) into a numerical procedure, the region $\Omega \subseteq \mathbb{R}^d$ of interest is subdivided into nonoverlapping subregions $\mathcal{T}_h = \{V_j\}$, i.e.,

$$\Omega = \bigcup_{j=1}^{N} V_j,$$

where the V_j are simplices having size $\mathcal{O}(h)$. Then, (2.1) can obviously be rewritten using the *cell averages*

$$\lambda_j(u)(t) := \overline{u}_j(t) = \frac{1}{|V_j|}\int_{V_j} u(x,t)dx, \qquad 1 \leq j \leq N.$$

Moreover, if $\mathcal{N}_j$ denotes the set of the neighboring simplices to the simplex $V_j \in \mathcal{T}_h$, we have

$$\frac{d}{dt}\lambda_j(u)(t) = -\frac{1}{|V_j|}\sum_{V\in\mathcal{N}_j}\int_{\partial V\cap\partial V_j}\sum_{\ell=1}^{d} f_\ell(u)\eta_\ell^{(V)}dS,$$

where $\eta^{(V)}$ denotes the outer unit normal vector to the boundary face $\partial V \cap \partial V_j$ of V.

If the flux is replaced by a numerical flux function or an approximate Riemann solver $H : \mathbb{R}^n \times \mathbb{R}^n \times \mathbb{R}^d \to \mathbb{R}^n$, satisfying

$$H(u,u;\eta) = \sum_{\ell=1}^{d} f_\ell(u)\eta_\ell,$$

and if the integration on the boundary hyperplane $\partial V \cap \partial V_j$ is replaced by a quadrature rule having weights w_ν and points $x_\nu(V)$, $1 \leq \nu \leq n_Q$, we derive

$$\frac{d}{dt}\lambda_j(u)(t) = -\frac{1}{|V_j|}\sum_{V\in\mathcal{N}_j}\sum_{\nu=1}^{n_Q} w_\nu H(u(x_\nu(V),t), u(x_\nu(V),t); \eta_\ell^{(V)}) + \mathcal{O}(h^{m_Q}),$$

where m_Q denotes the order of the employed quadrature rule.

Replacing the unknown values $u(x_\nu(V),t)$ simply by the cell averages spoils the approximation order and leads only to a first order scheme. However, if these values are replaced by a more accurate reconstruction $s_u(x_\nu(V),t)$, which satisfies

$$\lambda_j(s_u)(t) = \lambda_j(u)(t), \qquad 1 \leq j \leq N, \tag{2.2}$$

$$\|u - s_u\|_{L_\infty(\Omega)} = \mathcal{O}(h^p), \qquad h \to 0, \tag{2.3}$$

for all sufficiently smooth functions u, then we get for smooth f_ℓ via

$$\begin{aligned}\|H(u,u;n) - H(s_u,s_u;n)\|_2 &\leq \sum_{\ell=1}^{d}\|[f_\ell(u) - f_\ell(s_u)]\eta_\ell\|_2 \leq \sum_{\ell=1}^{d}\|[f_\ell(u) - f_\ell(s_u)]\|_2\\ &\leq C_f\|u - s_u\|_2\end{aligned}$$

finally,

$$\frac{d}{dt}\lambda_j(u)(t) = -\frac{1}{|V_j|}\sum_{V\in\mathcal{N}_j}\sum_{\nu=1}^{n_Q} w_\nu H(s_u(x_\nu(V),t), s_u(x_\nu(V),t); \eta_\ell^{(V)}) + \mathcal{O}(h^{\min\{p,m_Q\}}).$$

Hence, it is crucial to have reconstruction processes satisfying (2.2) and (2.3). In the next section, we will describe one possibility, which is based upon optimal recovery. However, the error estimate (2.3) is actually intrinsic to the reconstruction requirements (2.2), at least if the reconstruction process is stable in a sense which we will soon make precise. This is a consequence of the following result, which we will prove in this paper.

THEOREM 2.1. *Let k be a positive integer, $0 < s \le 1, 1 \le p < \infty, 1 \le q \le \infty$ and let α be a multi-index satisfying $k > |\alpha| + d/p$ or, for $p = 1, k \ge |\alpha| + d$. Let Ω be a bounded set satisfying an interior cone condition. Suppose Ω is decomposed into finitely many, nonoverlapping subdomains V_j such that every ball $B \subseteq \Omega$ with radius h contains at least one subdomain V_j. If the function $u \in W_p^{k+s}(\Omega)$ satisfies $\lambda_j(u) = 0$, $1 \le j \le N$, then*

$$|u|_{W_q^{|\alpha|}(\Omega)} \le ch^{k+s-|\alpha|-d(1/p-1/q)_+}|u|_{W_p^{k+s}(\Omega)},$$

where c is a constant independent of u and h, and $(x)_+ = \max\{x, 0\}$.

The condition that every ball $B \subseteq \Omega$ of radius h contains at least one volume V_j is automatically satisfied if the volumes form a uniform decomposition of Ω consisting of volumes of size h.

From this theorem we can conclude that any reconstruction process in the sense of (2.2) immediately satisfies

$$\|u - s_u\|_{L_\infty(\Omega)} \le Ch^{k+s-d/p}|u - s_u|_{W_p^{k+s}(\Omega)},$$

so that the stability assumption on the reconstruction process has to be something like

$$|u - s_u|_{W_p^{k+s}(\Omega)} \le C|u|_{W_p^{k+s}(\Omega)}. \tag{2.4}$$

Of course, the (semi-)norm on the right-hand side might be replaced by a norm or a stronger (semi-)norm; (2.2) and (2.4) together now yield the approximation error (2.3).

However, in most applications it is not reasonable to build a reconstruction using *all* cell averages. Instead, for each cell a local reconstruction is computed using all the cell averages of cells in a neighborhood of the current cell. Moreover, to avoid unwanted oscillations, the ENO approach chooses for each cell V_j different sets of neighboring volumes, which are usually called *stencils*, computes for all theses sets a local reconstruction and then chooses the reconstruction, where the oscillation of the solution is least.

3. Optimal recovery. In this section we shortly review the idea of using optimal recovery to solve the reconstruction problem (2.2). It has initially been introduced by Sonar in [9]. It is based upon (conditionally) positive definite functions and has the advantage of satisfying a stability condition such as (2.4) automatically. For simplicity, we will from now on suppress the time variable and restrict ourselves to functions $u : \mathbb{R}^d \to \mathbb{R}$, i.e., we assume $n = 1$.

In what follows, we will denote the space of d-variate polynomials of degree less than or equal to m by $\pi_m(\mathbb{R}^d)$.

DEFINITION 3.1. *A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is said to be conditionally positive definite of order m, if for all $N \in \mathbb{N}_0$, all sets $X = \{x_1 \ldots, x_N\}$ of pairwise distinct* *points, and all $\alpha \in \mathbb{R}^N \setminus \{0\}$ satisfying $\sum_{j=1}^N \alpha_j p(x_j) = 0$ for all $p \in \pi_{m-1}(\mathbb{R}^d)$, it is satisfied that*

$$\sum_{i,j=1}^N \alpha_i \alpha_j \Phi(x_i - x_j) > 0.$$

A function that is conditionally positive definite of order $m = 0$, where no constraints on α are imposed, is called positive definite.

A positive definite function Φ gives rise to a *reproducing kernel Hilbert space* $\mathcal{H} = \mathcal{H}_\Phi$, for which Φ is the reproducing kernel. In case of a conditionally positive definite function this remains true modulo $\pi_{m-1}(\mathbb{R}^d)$, which is the null space of the semi-inner product defined by Φ. In this context, this space is also called the *native space* associated to Φ.

Now suppose we are given functionals $\mu_1, \ldots, \mu_N$, which are continuous and linearly independent over $\mathcal{H}$ (modulo $\pi_{m-1}(\mathbb{R}^d)$.) Suppose further these functionals are $\pi_{m-1}(\mathbb{R}^d)$-unisolvent meaning that $p = 0$ is the only polynomial from $\pi_{m-1}(\mathbb{R}^d)$ with $\mu_j(p) = 0$ for all $1 \leq j \leq N$. Then, there exists exactly one function of the form

$$s_u(x) = \sum_{j=1}^N \alpha_j \mu_j^y \Phi(x - y) + p(x),$$

where $p \in \pi_{m-1}(\mathbb{R}^d)$ and μ_j^y means acting with respect to the variable y, which satisfies the interpolation conditions

$$\mu_j(s_u) = \mu_j(u), \qquad 1 \leq j \leq N,$$

together with

$$\sum_{j=1}^N \alpha_j \mu_j(q) = 0, \qquad q \in \pi_{m-1}(\mathbb{R}^d).$$

Moreover, the solution s_u satisfies the stability estimates

$$\|s_u\|_{\mathcal{H}} \leq \|u\|_{\mathcal{H}}, \qquad \|u - s_u\|_{\mathcal{H}} \leq \|u\|_{\mathcal{H}},$$

and the latter leads to (2.4), if $\mathcal{H}$ is either a Sobolev or a Beppo–Levi space. Examples for (conditionally) positive definite functions are the compactly supported functions, constructed and investigated in [12, 13] by this author, for Sobolev spaces, and the thin-plate or surface splines, investigated in [2] by Duchon, for Beppo–Levi spaces. We will come back to these examples later on.

For more details on optimal recovery and scattered data approximation, we refer the reader to the recent book [14].

4. Local estimates. Let us describe the main idea for proving Theorem 2.1. Our analysis is based upon the following approach. First of all, we consider u only on small subregions Ω_ℓ of Ω and derive estimates in terms of the diameter of Ω_ℓ. These subregions can, for example, be the stencils of the ENO schemes previously described. For our matters, it is only important that Ω_ℓ itself is the union of some of the volumes V_j. If interested in estimates on all of Ω, we have to glue these local estimates together to retrieve our final error bound.

In accordance with Theorem 2.1 we assume that the smooth function u satisfies

$$\lambda_j(u) = \frac{1}{|V_j|}\int_{V_j} u(x)dx = 0, \qquad j \in \mathcal{M}_\ell,$$

where $\mathcal{M}_\ell$ contains all indices j with $V_j \subseteq \Omega_\ell$.

On the local domain Ω_ℓ we approximate u by a polynomial $P \in \pi_k(\mathbb{R}^d)$, i.e., we write

$$u = u - P + P.$$

If u is smooth, the term $u - P$ can locally be bounded in a sound way by using an averaged Taylor polynomial to u as P. To bound this specific P we have to study the action of the functionals λ_j on the space of polynomials $\pi_k(\mathbb{R}^d)$ of degree k. Suppose our functionals allow a reconstruction of the form

$$p(x) = \sum_{j\in\mathcal{M}_\ell} a_j(x)\lambda_j(p), \qquad x \in \Omega_\ell, p \in \pi_k(\mathbb{R}^d)$$

with certain numbers $a_j(x)$, $j \in \mathcal{M}_\ell$, having a uniformly bounded ℓ_1-norm. Then, using $\lambda_j(u) = 0$, we can derive

$$\begin{aligned}|p(x)| &\le \sum_{j\in\mathcal{M}_\ell} |a_j(x)||\lambda_j(p-u)| = \sum_{j\in\mathcal{M}_\ell} |a_j(x)|\frac{1}{|V_j|}\int_{V_j} |p(x) - u(x)|dx \\ &\le \|p-u\|_{L_\infty(\Omega_\ell)} \sum_{j\in\mathcal{M}_\ell} |a_j(x)|.\end{aligned}$$

Hence, we can control the norm of P again by the norm of $u - P$. Moreover, this shows that we first have to investigate polynomial approximation on local sets.

4.1. Polynomial approximation. To simplify notation in the rest of the section, we denote our current local set Ω_ℓ by $\mathcal{D}$ and assume that $\mathcal{M}_\ell = \{1, \dots, N\}$. This short section is meant to collect all necessary results on local polynomial approximation. It is based upon Brenner and Scott's book [1, Chapter 4] and on the recent article [7]. We start by introducing star-shaped regions.

DEFINITION 4.1. *A domain $\mathcal{D}$ is* star-shaped with respect to a ball $B(x_c, r) := \{x \in \mathbb{R}^d : \|x - x_c\|_2 < r\}$ *if for every $x \in \mathcal{D}$, the closed convex hull of $\{x\} \cup B$ is contained in $\mathcal{D}$.*

It should be apparent, that a bounded region which is star-shaped with respect to a ball automatically satisfies a uniform interior cone condition. To be more precise, the following result has been established in [7].

LEMMA 4.2. *If $\mathcal{D}$ is star-shaped with respect to a ball $B(x_c, r)$ and contained in a ball $B(x_c, R)$, then it satisfies an interior cone condition with radius r and angle $\theta = 2\arcsin\left(\frac{r}{2R}\right)$.*

Brenner and Scott [1, Chapter 4] discuss approximating a function $u \in W_p^k(\mathcal{D})$ by averaged Taylor polynomials $Q_k u \in \pi_{k-1}(\mathbb{R}^d)$. These results have been extended and generalized to the case of fractional Sobolev spaces by Narcowich et al. [7]. In this section, we briefly summarize these results.

The averaged Taylor polynomials are defined as follows. Let B_r be a ball relative to which $\mathcal{D}$ is star-shaped and having radius $r \ge \frac{1}{2}r_{max}$, where r_{max} is the largest radius of a ball relative to which $\mathcal{D}$ is star-shaped. The averaged Taylor polynomials are then given by

$$Q_k u(x) := \sum_{|\alpha|<k} \frac{1}{\alpha!} \int_{B_r} D^\alpha u(y)(x-y)^\alpha \phi(y)dy,$$

where ϕ is a nonnegative C^∞ "bump" function supported on B_r, satisfying both $\int_{B_r} \phi(y)dy = 1$ and $\max \phi \leq Cr^{-d}$, where $C = C_d$. The result we need is a generalization of [1, Proposition 4.3.2] and comes from [7]. It utilizes the *chunkiness parameter* γ of a star-shaped domain $\mathcal{D}$, which is defined to be $\gamma = \rho_{\mathcal{D}}/r_{max}$, where $\rho_{\mathcal{D}}$ is the diameter of $\mathcal{D}$.

LEMMA 4.3. *Let $0 < s \leq 1$. Assume that $1 < p < \infty$ and $k > |\alpha| + d/p$, or $p = 1$ and $k \geq |\alpha| + d$. Then, we have for $u \in W_p^{k+s}(\mathcal{D})$ the estimate*

$$\|D^\alpha u - D^\alpha Q_{k+1} u\|_{L_\infty(\mathcal{D})} \leq C_{k,d,p}(1+\gamma)^{d(1+1/p)} \rho_{\mathcal{D}}^{k+s-|\alpha|-d/p} |u|_{W_p^{k+s}(\mathcal{D})},$$

where $\rho_{\mathcal{D}}$ denotes the diameter of $\mathcal{D}$.

It is important to realize that the involved constant here depends on the domain $\mathcal{D}$ only via its chunkiness parameter γ.

4.2. Norming sets. To reconstruct polynomials from cell averages in a controlled way we employ *norming sets*. Norming sets have been introduced in the context of scattered data approximation on spheres [5]. The idea behind them can be described in a rather abstract setting; see [6, 14].

Let V be a finite dimensional vector space with norm $\|\cdot\|_V$ and let $Z \subseteq V^*$ be a finite set consisting of N functionals. Here, V^* denotes the dual space of V consisting of all linear and continuous functionals defined on V.

DEFINITION 4.4. *We will say that Z is a norming set for V if the mapping $T : V \to T(V) \subseteq \mathbb{R}^N$ defined by $T(v) = (z(v))_{z \in Z}$ is injective. We will call T the sampling operator.*

If Z is a norming set for V, the mapping $T : V \to T(V) \subseteq \mathbb{R}^{|Z|}$ is bijective and the norm of its inverse is given by

$$\|T^{-1}\| = \sup_{v \neq 0} \frac{\|v\|}{\|Tv\|}.$$

THEOREM 4.5. *Suppose V is a finite dimensional normed linear space and $Z = \{z_1, \dots, z_N\}$ is a norming set for V with T being the corresponding sampling operator. For every $\psi \in V^*$ there exists a vector $a \in \mathbb{R}^N$ depending only on ψ such that for every $v \in V$,*

$$\psi(v) = \sum_{j=1}^N a_j z_j(v),$$

and

$$\|a\|_{\mathbb{R}^{N*}} \leq \|\psi\|_{V^*} \|T^{-1}\|.$$

We can apply Theorem 4.5 to our situation by choosing $V = \pi_k(\mathbb{R}^d)$, $\psi = \delta_x$, or more generally $\psi = \delta_x \circ D^\alpha$ with $x \in \mathcal{D}$ and $T : \pi_k(\mathbb{R}^d) \to \mathbb{R}^N$ defined by $T(p) := (\lambda_1(p), \dots, \lambda_N(p))^T$, where λ_j denotes once again the cell average operator to the cell V_j. Thus, we have to find a bound on $\|T^{-1}\|$, which means we have to determine a constant $C > 0$ with

$$\|p\|_{L_\infty(\mathcal{D})} \le C\|T(p)\|_\infty, \qquad p \in \pi_k(\mathbb{R}^d). \tag{4.1}$$

Before providing such a result in general, let us have a look at two examples.

Example 4.6. In the first case we consider the case of linear polynomials in the univariate setting with equidistant points. Hence, our cell averages are given by

$$\lambda_j(u) = \frac{1}{h}\int_{jh}^{(j+1)h} u(x)dx.$$

Now, let our local $\mathcal{D}$ be $\mathcal{D} = [0, 2h]$. If $p \in \pi_1(\mathbb{R})$ satisfies $\|p\|_{L_\infty(\mathcal{D})} = 1$, then it obviously attains its absolute maximum either in 0 or in $2h$. In the first case (the other one is dealt with in the same way) p satisfies (without restriction) $p(0) = 1$ and $p'(0) \le 0$. Hence, we can express p as $p(x) = 1 - ax$ with $a \ge 0$ and for this p we find

$$|\lambda_1(p)| = \left|1 - \frac{1}{h}\int_0^h axdx\right| = \left|1 - a\frac{h}{2}\right| = 1 - a\frac{h}{2} \ge 1 - \frac{1}{2} = \frac{1}{2},$$

since we can conclude from $|p(2h)| \le 1$ that $|a| < 1/h$. Hence, we can choose the constant C in (4.1) as $C = 2$.

Example 4.7. Our second example deals still with linear polynomials but this time on $\mathbb{R}^2$. Again, we assume that our cells form an equidistant grid $h\mathbb{Z}^2$. Suppose our stencil consists of the four volumes $V_1 = [0,h]^2$, $V_2 = [h,2h]\times[0,h]$, $V_3 = [0,h]\times[h,2h]$, and $V_4 = [h,2h]^2$. Again, without restriction we can assume that $p \in \pi_1(\mathbb{R}^2)$ attains its maximum in $(0,0)$. Hence, we have $p(x,y) = 1 - ax - by$ with $a, b \ge 0$. A simple computation shows that

$$\lambda_1(p) = 1 - \frac{h}{2}(a+b), \quad \lambda_2(p) = \lambda_1(p) - ha, \quad \lambda_3(p) = \lambda_1(p) - hb.$$

Now if $|\lambda_1(p)| \le 1/3$, then we find $4/3 \le h(a+b) \le 8/3$ giving

$$-\frac{10}{3} \le \lambda_2(p) + \lambda_3(p) = 2\lambda_1(p) - h(a+b) \le -\frac{2}{3}.$$

Hence, if we have in addition $|\lambda_2(p)| \le 1/3$, we must have $|\lambda_3(p)| \ge 1/3$. Thus we can choose the constant C in (4.1) to be $C = 3$.

From the second example it should already be clear that the multidimensional case is in general harder to be dealt with. Moreover, if higher degree polynomials are under investigation, we cannot expect the maximum to be attained in a vertex of our simplices. Furthermore, we do not want to restrict ourselves to uniform grids. Hence, we need a somewhat more general approach. The price we have to pay for this is to use slightly larger stencils than probably really necessary.

LEMMA 4.8. *Suppose $\mathcal{D}$ is compact and satisfies a cone condition with angle θ and radius $r > 0$. Suppose $\mathcal{D}$ is disjointly decomposed into $\mathcal{D} \subseteq \cup_{j=1}^N V_j$ with subregions V_j. Suppose every ball $B(x_0, h) \subseteq \mathcal{D}$ with radius h contains at least one V_j. Then, provided that*

$$h \le \frac{r\sin\theta}{4k^2(1+\sin\theta)}, \tag{4.2}$$

the mapping $T : \pi_k(\mathbb{R}^d) \to \mathbb{R}^N$, $T(p) = (\lambda_1(p), \dots, \lambda_N(p))^T$ *is injective with*

$$\|T(p)\|_\infty = \max_{1 \le j \le N} |\lambda_j(p)| \ge \frac{1}{2}\|p\|_{L_\infty(\mathcal{D})}, \qquad p \in \pi_k(\mathbb{R}^d).$$

Proof. Suppose $p \in \pi_k(\mathbb{R}^d)$ with $\|p\|_{L_\infty(\mathcal{D})} = 1$ is given. Then there exists a point $x_0 \in \mathcal{D}$ with $|p(x_0)| = 1$. To point x_0, we can choose a cone $C(x_0)$ with angle θ and radius r which is completely contained in $\mathcal{D}$. If $\xi \in \mathbb{R}^d$ with $\|\xi\|_2 = 1$ gives the axis of the cone, it is easy to see that $C(x_0)$ completely contains the ball $B(y, h)$ with $y = x_0 + (h/\sin\theta)\xi$, provided $h \le r\sin\theta/(1+\sin\theta)$ is satisfied. Hence, by assumption, we find a volume V_j with

$$V_j \subseteq B(y, h) \subseteq C(x_0) \subseteq \mathcal{D}.$$

Moreover, the line segment $x_0 + t(x - x_0)/\|x - x_0\|_2$, $t \in [0, r]$, which joins x_0 with any $x \in V_j$ and beyond is entirely contained in $\mathcal{D}$. Hence, if we define the univariate polynomial $q(t) = p(x_0 + t(x - x_0)/\|x - x_0\|_2)$, $t \in [0, r]$, we can apply Markov's inequality for univariate polynomials in the form

$$|q'(s)| \le \frac{2k^2}{r}\|q\|_{L_\infty[0,r]}, \qquad s \in [0, r],$$

to derive

$$|p(x_0) - p(x)| = |q(0) - q(\|x - x_0\|_2)| \le \int_0^{\|x - x_0\|_2} |q'(s)| ds \le \frac{2k^2}{r}\|x - x_0\|_2$$

for every $x \in V_j \subseteq B(y, h)$. For such an x we also have the bound

$$\|x - x_0\|_2 \le \|x - y\|_2 + \|y - x_0\|_2 \le h + \frac{h}{\sin\theta} = h\left(\frac{1 + \sin\theta}{\sin\theta}\right),$$

so that we can conclude

$$\begin{aligned} |p(x_0) - \lambda_j(p)| &\le \frac{1}{|V_j|}\int_{V_j} |p(x_0) - p(x)| dx \le \frac{2k^2}{r|V_j|}\int_{V_j} \|x - x_0\|_2 dx \\ &\le \frac{2k^2(1 + \sin\theta)}{r\sin\theta} h \le \frac{1}{2}, \end{aligned}$$

provided (4.2) is satisfied. This means $\lambda_j(p) \ge \|p\|_{L_\infty(\mathcal{D})}/2$ for this particular j. □

Note that for deriving our result only those volumes V_j were necessary with $V_j \subseteq \mathcal{D}$, as long as all other assumptions are satisfied.

Now that we know the bound on the inverse of the sampling operator we still need a bound on the norm of the evaluation functional ψ; this situation is again provided in [7].

LEMMA 4.9. *Suppose* $\mathcal{D}$ *is bounded and satisfies an interior cone condition with angle* θ *and radius* r*. Then, for every* $p \in \pi_k(\mathbb{R}^d)$ *and every* $|\alpha| \le k$ *we have*

$$\|D^\alpha p\|_{L_\infty(\mathcal{D})} \le \left(\frac{2k^2}{r\sin\theta}\right)^{|\alpha|} \|p\|_{L_\infty(\mathcal{D})}.$$

As a consequence, Theorem 4.5 guarantees for every $x \in \mathcal{D}$ and every $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le k$ the existence of numbers $a_1^\alpha(x), \dots, a_N^\alpha(x)$ such that

$$D^\alpha p(x) = \sum_{j=1}^N \lambda_j(p) a_j^\alpha(x), \qquad p \in \pi_k(\mathbb{R}^d), \tag{4.3}$$

and

$$\sum_{j=1}^{N} |a_j^\alpha(x)| \le 2 \left(\frac{2k^2}{r \sin\theta} \right)^{|\alpha|}. \tag{4.4}$$

For our next result, we use the fact that a domain $\mathcal{D} \subseteq B(x_c, R)$ which is star-shaped with respect to a ball $B(x_c, r) \subseteq \mathcal{D}$ satisfies an interior cone condition with radius r and angle $\theta = 2\arcsin(r/(2R))$.

PROPOSITION 4.10. *Let k be a positive integer, $1 \le p < \infty$, $0 < s \le 1$, and let α be a multi-index satisfying $k > |\alpha| + d/p$, or, for $p = 1$, $k \ge |\alpha| + d$. Let $1 \le q \le \infty$. Suppose $\mathcal{D} = \cup V_j \subseteq B(x_c, R)$ is star-shaped with respect to $B(x_c, r)$ and covered with volumes V_j such that every ball $B \subseteq \mathcal{D}$ with radius h contains at least one of these volumes V_j. Let $\rho_{\mathcal{D}}$ denote the diameter of $\mathcal{D}$. If h satisfies* (4.2)*, then we have for every $u \in W_p^{k+s}(\mathcal{D})$ satisfying $\lambda_j(u) = 0$, $1 \le j \le N$, the error estimate*

$$\|D^\alpha u\|_{L_q(\mathcal{D})} \le C \rho_D^{k+s-|\alpha|+d(1/q-1/p)} |u|_{W_p^{k+s}(\mathcal{D})},$$

and the constant C depends only on $k, d, p, |\alpha|$, and $\theta = 2\arcsin(r/2R)$.

Proof. We use the decomposition $D^\alpha u = D^\alpha u - D^\alpha Q_{k+1} + D^\alpha Q_{k+1}$ with $Q_{k+1} \in \pi_k(\mathbb{R}^d)$ being the averaged Taylor polynomial to u.

Since the chunkiness parameter γ can be bounded by

$$1 \le \gamma \le \rho_{\mathcal{D}}/r \le 2R/r = 1/\sin(\theta/2), \tag{4.5}$$

Lemma 4.3 yields a constant C depending only on $k, d, |\alpha|, \theta$ with

$$\|D^\alpha u - D^\alpha Q_{k+1} u\|_{L_\infty(\mathcal{D})} \le C \rho_{\mathcal{D}}^{k+s-|\alpha|-d/p} |u|_{W_p^{k+s}(\mathcal{D})}.$$

On the other hand, since (4.2) is satisfied, Lemma 4.8 and Theorem 4.5 yield stable polynomial reproduction in the sense of (4.3) and (4.4). Hence, because of $\lambda_j(u) = 0$ we have for any polynomial $p \in \pi_k(\mathbb{R}^d)$,

$$\begin{aligned} |D^\alpha p(x)| &= \left| \sum_{j=1}^{N} \lambda_j(p) a_j^\alpha(x) \right| = \left| \sum_{j=1}^{N} a_j^\alpha(x) \left[\lambda_j(p) - \lambda_j(u) \right] \right| \\ &\le \sum_{j=1}^{N} |a_j^\alpha(x)| \frac{1}{|V_j|} \int_{V_j} |p(y) - u(y)| dy \\ &\le 2 \left(\frac{2k^2}{r \sin\theta} \right)^{|\alpha|} \|p - u\|_{L_\infty(\mathcal{D})}. \end{aligned}$$

Specifying $p = Q_{k+1}$ and using Lemma 4.3 and (4.5) yields

$$\begin{aligned} |D^\alpha Q_{k+1}(x)| &\le 2 C_{k,d,p} (1+\gamma)^{d(1+1/p)} \left(\frac{2k^2 \rho_{\mathcal{D}}}{r \sin\theta} \right)^{|\alpha|} \rho_{\mathcal{D}}^{k+s-|\alpha|-d/p} |u|_{W_p^{k+s}(\mathcal{D})} \\ &\le C \rho_{\mathcal{D}}^{k+s-|\alpha|-d/p} |u|_{W_p^{k+s}(\mathcal{D})} \end{aligned}$$

with a constant depending only on $k, d, p, |\alpha|$ and θ. Hence, we have established the result

$$\|D^\alpha u\|_{L_\infty(\mathcal{D})} \le C \rho_{\mathcal{D}}^{k+s-|\alpha|-d/p} |u|_{W_p^{k+s}(\mathcal{D})}$$

for $q = \infty$. Next, integrating this inequality and using the fact that $\mathcal{D}$ has volume $\mathcal{O}(\rho_{\mathcal{D}}^d)$, leads to

$$\|u\|_{L_q(\mathcal{D})} \leq C\rho_{\mathcal{D}}^{k+s-|\alpha|+d(1/q-1/p)}|u|_{W_p^{k+s}(\mathcal{D})},$$

which finishes the proof. □

Note that this result remains true even if $\mathcal{D}$ itself is not the union of its volumes. More precisely, if $\mathcal{D} \subseteq \cup V_j$ it suffices that the collection $\{V_j : V_j \subseteq \mathcal{D}\}$ satisfies the assumptions. In particular, in this situation the recovery function should only be based upon these volumes.

Proposition 4.10 yields our first main result. It covers the situation of ENO approximation whenever a sufficiently large stencil is chosen. Note, however, that the assumption on the size of the stencil can even be improved in case of linear polynomials and regular meshes, using the ideas of Examples 4.6 and 4.7.

DEFINITION 4.11. *We say that a stencil $\mathcal{D} = \cup_{j=1}^M V_j$ with volumes V_j is* admissible *if there exist constants $C_1 > 0$ and $L \geq 1$ such that*

- *$\mathcal{D}$ is star-shaped with respect to a ball $B(x_c, C_1 h)$,*
- *$\mathcal{D}$ is contained in the ball $B(x_c, LC_1 h)$,*
- *each ball of radius h contains at least one V_j.*

If a stencil is admissible in this sense, it yields convergence of the reconstruction process using cell averages.

THEOREM 4.12. *Suppose the assumption on $k, p, s, |\alpha|$, and q from Proposition* 4.10 *hold. Suppose $\mathcal{D} = \cup_{j=1}^M V_j \subseteq \mathbb{R}^d$ is an admissible stencil with constants C_1, L. Set $\theta = 2\arcsin(1/(2L))$. If*

$$C_1 \geq \frac{4k^2(1+\sin\theta)}{\sin\theta},$$

then for every $u \in W_p^{k+s}(\mathcal{D})$, which satisfies $\lambda_j(u) = 0$ for $1 \leq j \leq N$, the error estimate

$$\|D^\alpha u\|_{L_q(\mathcal{D})} \leq Ch^{k+s-|\alpha|+d(1/q-1/p)}|u|_{W_p^{k+s}(\mathcal{D})}$$

is satisfied, where C is a constant independent of u and h.

Proof. From Lemma 4.2 we know that $\mathcal{D}$ satisfies a cone condition with radius $r = C_1 h$ and angle $\theta = 2\arcsin(1/(2L))$. Our assumption on C_1 assures that condition (4.2) is satisfied, so that Proposition 4.10 yields the result. □

It is important here to note that all involved constants depend on the region $\mathcal{D}$ only by their cone condition. This will be of importance for deriving our general result in the next section.

The assumptions $k > |\alpha| + d/p$ can be weakened in case of $p = q$ using interpolation theory to $k > d/p$.

Finally, note that for Proposition 4.10 it is not important for $\mathcal{D}$ to be the collection of all local V_j. It suffices that every ball of radius h contained in $\mathcal{D}$ contains a V_j itself.

5. Global estimates.

Now we work our way towards proving Theorem 2.1. To this end, we cover our region Ω by small regions $\mathcal{D}$. The key ingredient is that the involved constants depend on the local regions only via their cone condition. Hence, if we fix r and R and use local domains $\mathcal{D}$ for which we can find an x_c such that $\mathcal{D} \subseteq B(x_c, R)$ is star-shaped with respect to $B(x_c, r)$, the chunkiness parameter and thus the angle θ of the cone condition are all the same. This means in particular that we can use the same constant C in Proposition 4.10 for any such domain.

Interestingly, the procedure which now follows is identical to the one employed in the point evaluation case, which was studied in [7]. Hence, we will rely heavily on that paper; see also [14].

Let us suppose our global region Ω is bounded and satisfies a cone condition with radius r and angle θ. We introduce the following quantities. As usual, let h be the typical size of our finite volumes. Let

$$\begin{aligned} \vartheta &:= 2\arcsin\left(\frac{\sin\theta}{4(1+\sin\theta)}\right), \\ Q(k,\theta) &:= \frac{\sin\theta\sin\vartheta}{8k^2(1+\sin\theta)(1+\sin\vartheta)} \\ R &:= Q(k,\theta)^{-1}h, \\ r &:= \frac{\sin\theta}{2(1+\sin\theta)}R. \end{aligned}$$

With these settings we define the sets $T_r := \{t \in \frac{2r}{\sqrt{d}}\mathbb{Z}^d : B(t,r) \subseteq \Omega\}$ and

$$\mathcal{D}_t = \{x \in \Omega : \mathrm{co}(\{x\} \cup B(t,r)) \subseteq \Omega \cap B(t,R)\}, \qquad t \in T_r,$$

where co(A) denotes the closed convex hull of the set A.

LEMMA 5.1 (see [7]). *With the just introduced quantities, suppose the number $h > 0$ satisfies $h \le Q(k,\theta)r$. Then, the following holds true:*

1. *Each $\mathcal{D}_t$ is star-shaped with respect to the ball $B(t,r)$ and satisfies $B(t,r) \subseteq \mathcal{D}_t \subseteq \Omega \cap B(t,R)$.*
2. *Each $\mathcal{D}_t$ satisfies a cone condition with angle ϑ and radius r.*
3. *$\Omega = \bigcup_{t\in T_r} \mathcal{D}_t$ and $\rho_{\mathcal{D}_t} \le 2R = 2h/Q(k,\theta)$.*
4. *There is a constant $M_1 > 0$ such that $\sum_{t\in T_r} \chi_{\mathcal{D}_t} \le M_1$.*
5. *There is a constant $M_2 > 0$ such that $\#T_r \le M_2 r^{-d}$.*

Here, $\chi_{\mathcal{D}_t}(x)$ is 1 if $x \in \mathcal{D}_t$ and 0 elsewhere and $\#T_r$ denotes the cardinality of T_r.

Now that we have the local sets we can formulate and prove our main result of this section; Theorem 2.1.

THEOREM 5.2. *Suppose Ω is bounded and satisfies an interior cone condition with radius $r > 0$ and angle θ. Let k be a positive integer, $0 < s \le 1$, $1 \le p < \infty$, $1 \le q \le \infty$, and let $m \in \mathbb{N}_0$ satisfy $k > m + d/p$ for $p > 1$, or $k \ge m + d$ for $p = 1$. Also, let $\{V_1, \dots, V_N\} \subseteq \Omega$ such that each ball $B \subseteq \Omega$ of radius $h \le Q(k,\theta)r$ contains at least one volume V_j. If $u \in W_p^{k+s}(\Omega)$ satisfies $\lambda_j(u) = 0$, $1 \le j \le N$, then*

$$|u|_{W_q^m(\Omega)} \le Ch^{k+s-m-d(1/p-1/q)_+}|u|_{W_p^{k+s}(\Omega)}, \tag{5.1}$$

where $(x)_+ = \max\{x, 0\}$.

Proof. We use the notation introduced in the paragraph before Lemma 5.1. First of all note that, since $h \le Q(k,\theta)r$, Lemma 5.1 is applicable. Furthermore, our definition of r, R, and $Q(k,\theta)$ establish $h = \frac{r\sin\vartheta}{4k^2(1+\sin\vartheta)}$, which allows us to apply Proposition 4.10 to the local sets $\mathcal{D}_t$. The just mentioned lemma and proposition immediately establish the result in the case $q = \infty$. On the other hand, for $1 \le q < \infty$

the decomposition of Ω implies that we have

$$\begin{aligned}|u|^q_{W^m_q(\Omega)} &= \sum_{|\alpha|=m}\int_\Omega |D^\alpha u(x)|^q dx\\ &\le \sum_{t\in T_r}\sum_{|\alpha|=m}\int_{\mathcal{D}_t}|D^\alpha u(x)|^q dx = \sum_{t\in T_r}|u|^q_{W^m_q(\mathcal{D}_t)}\\ &\le (\#T_r)^{q\left(\frac{1}{q}-\frac{1}{p}\right)_+}\left(\sum_{t\in T_r}|u|^p_{W^m_q(\mathcal{D}_t)}\right)^{q/p},\end{aligned}$$

where the last bound follows from standard inequalities relating p and q norms on finite dimensional spaces. Proposition 4.10 together with $\rho_{\mathcal{D}_t}\le 2R = 2Q(k,\theta)^{-1}h$ gives the bound

$$\left(\sum_{t\in T_r}|u|^p_{W^m_q(\mathcal{D}_t)}\right)^{1/p} \le Ch^{k+s-m+d(\frac{1}{q}-\frac{1}{p})}\left(\sum_{t\in T_r}|u|^p_{W^{k+s}_p(\mathcal{D}_t)}\right)^{1/p}.$$

Here, it has been essential that all involved constants depend only on the cone condition, which is the same for all $\mathcal{D}_t$. Now, using Lemma 5.1 again yields

$$\begin{aligned}\sum_{t\in T_r}|u|^p_{W^{k+s}_p(\mathcal{D}_t)} &\le \sum_{|\alpha|=k}\int_\Omega\sum_{t\in T_r}\chi_{\mathcal{D}_t}(x)\int_{\mathcal{D}_t}\frac{|D^\alpha u(x)-D^\alpha u(y)|^p}{\|x-y\|_2^{d+sp}}dydx\\ &\le M_1\sum_{|\alpha|=k}\int_\Omega\int_\Omega\frac{|D^\alpha u(x)-D^\alpha u(y)|^p}{\|x-y\|_2^{d+sp}}dydx\\ &\le M_1|u|^p_{W^{k+s}_p(\Omega)}.\end{aligned}$$

A final application of Lemma 5.1 together with $r=\frac{\sin\theta}{2Q(k,\theta)(1+\sin\theta)}h$ shows $\#T_r\le Ch^{-d}$. Putting all these things together and taking

$$d\left(\frac{1}{q}-\frac{1}{p}\right)-d\left(\frac{1}{q}-\frac{1}{p}\right)_+ = -d\left(\frac{1}{p}-\frac{1}{q}\right)_+$$

into account establishes the desired result. □

Let us apply this result to the compactly supported functions of minimal degree [12, 13]. They are radial functions $\Phi_{d,k}(x)=\phi_{d,k}(\|x\|_2)$, $x\in\mathbb{R}^d$, where each univariate function $\phi_{d,k}$ consists of a univariate polynomial on its support. Moreover, the d-variate function $\Phi_{d,k}$ is in $C^{2k}(\mathbb{R}^d)$. The functions $\Phi_{d,k}$ generate Sobolev spaces $W^\sigma_2(\mathbb{R}^d)$ with $\sigma = d/2+k+1/2$ as their associated reproducing kernel Hilbert spaces since their Fourier transforms $\widehat{\Phi}=\widehat{\Phi}_{d,k}$ decay like

$$C_1(1+\|\omega\|_2^2)^{-\sigma}\le\widehat{\Phi}(\omega)\le C_2(1+\|\omega\|_2^2)^{-\sigma}. \tag{5.2}$$

Corollary 5.3. *Suppose $\Omega\subseteq\mathbb{R}^d$ is bounded, satisfies a cone condition, and has a Lipschitz boundary. Suppose $\Phi\in L_1(\mathbb{R}^d)$ is positive definite and has a Fourier transform $\widehat{\Phi}$ satisfying* (5.2) *with $\sigma>d/2$. Let $1\le q\le\infty$ and $u\in W^\sigma_2(\Omega)$. If Ω is covered by volumes $\{V_j\}$ such that every ball $B\subseteq\Omega$ of radius h contains at least* *one V_j, then the error between u and its optimal recovery s_u from cell averages can be bounded by*

$$\|u - s_u\|_{L_q(\Omega)} \le Ch^{\sigma - d(1/2-1/q)_+}\|u\|_{W_2^\sigma(\Omega)}.$$

Proof. We need to extend $u \in W_2^\sigma(\Omega)$ to a function $Eu \in W_2^\sigma(\mathbb{R}^d)$. This is under the assumptions on Ω continuously possible; see the discussion in [7]. Hence, there exists a constant $C > 0$ such that $\|Eu\|_{W_2^\sigma(\mathbb{R}^d)} \le C\|u\|_{W_2^\sigma(\Omega)}$ and optimal recovery leads to

$$\|u - s_u\|_{W_2^\sigma(\Omega)} \le \|Eu - s_{Eu}\|_{W_2^\sigma(\mathbb{R}^d)} \le C\|Eu\|_{W_2^\sigma(\mathbb{R}^d)} \le C\|u\|_{W_2^\sigma(\Omega)}. \tag{5.3}$$

Since we have for the cell averages $\lambda_j(u - s_u) = 0$, we can apply Theorem 5.2 to $u - s_u$, which yields, together with (5.3), the desired result. □

For simplicity, we have stated the result only for the nonderivative case. The reader should have no problems restating it, including estimates on the derivatives.

If the target function u is known to be analytic so that it belongs even to the reproducing Hilbert space of *Gaussians* $\phi(r) = \exp(-\alpha r^2)$, $\alpha > 0$ or *multiquadrics* $\phi(r) = \sqrt{1 + r^2}$, then *any* Sobolev norm can be used, giving arbitrary convergence order. However, this is true only under rather restrictive assumptions on the target function.

6. Improved estimates for thin-plate splines. In this section we will improve the estimates derived in Theorem 5.2 in the particular situation of thin-plate splines. The ideas employed can also be used for other basis functions. However, the additional assumptions that have to be imposed on the target function are in other cases often only implicitly given.

We start by applying Theorem 5.2 to thin-plate splines. To be more precise, for $x \in \mathbb{R}^d$, we let

$$\Phi_{d,k}(x) := c_{d,k}\begin{cases} \|x\|_2^{2k-d} & \text{for } d \text{ odd}, \\ \|x\|_2^{2k-d}\log\|x\|_2 & \text{for } d \text{ even}, \end{cases}$$

where $c_{d,k}$ is a constant chosen so that $\Phi_{d,k}$ is a fundamental solution of the iterated Laplacian. In terms of the distributional Fourier transform, this is equivalent to requiring that $\widehat{\Phi}_{d,k}(\omega) = \|\omega\|_2^{-2k}$, if $\omega \neq 0$.

The reproducing kernel semi-Hilbert space associated with $\Phi_{d,k}$ is the Beppo–Levi space,

$$\mathrm{BL}_k(\mathbb{R}^d) := \{f \in C(\mathbb{R}^d) : D^\alpha f \in L_2(\mathbb{R}^d) \text{ for all } |\alpha| = k\},$$

which is equipped with the semi-inner product

$$(f, g)_{\mathrm{BL}_k(\mathbb{R}^d)} = \sum_{|\alpha|=k} \frac{k!}{\alpha!}(D^\alpha f, D^\alpha g)_{L_2(\mathbb{R}^d)},$$

i.e., BL_k is equipped with a semi-Hilbert norm comparable to $|\cdot|_{W_2^k(\mathbb{R}^d)}$. In particular, choosing $k = 2$ in case of $d = 2$ leads to the well-known thin-plate splines $\phi(r) = r^2 \log r$.

As in the case of the compactly supported functions, we need to extend the function $f \in W_2^k(\Omega)$ to a function $E_k f \in \mathrm{BL}_k(\mathbb{R}^d)$. This is possible due to Duchon

[2]. To be more precise there exists an extension operator $E_k : W_2^k(\Omega) \to \mathrm{BL}_k(\mathbb{R}^d)$ with $E_k u|\Omega = u$ and $|E_k u|_{\mathrm{BL}_k(\mathbb{R}^d)} \le \|E_k\| |u|_{\mathrm{BL}_k(\Omega)}$.

Again, we state our result only for the nonderivative case.

COROLLARY 6.1. *Suppose $\Omega \subseteq \mathbb{R}^d$ is bounded and satisfies an interior cone condition. Suppose further $k > d/2$ and $1 \le q \le \infty$. If Ω is covered by volumes $\{V_j\}$ such that every ball $B \subseteq \Omega$ of radius h contains at least one volume V_j. Then, the error between $u \in W_2^k(\Omega)$ and its optimal recovery s_u from cell averages using the thin-plate spline $\Phi_{d,k}$ has the error estimate*

$$\|u - s_u\|_{L_q(\Omega)} \le C h^{k-d(1/2-1/q)_+} |u|_{\mathrm{BL}_k(\Omega)}.$$

Proof. We extend $u \in W_2^k(\Omega)$ to a function $u = E_k u \in \mathrm{BL}_k(\mathbb{R}^d)$. Since $\mathrm{BL}_k(\mathbb{R}^d)$ is the associated reproducing kernel semi-Hilbert space, we have

$$|u - s_u|_{\mathrm{BL}_k(\mathbb{R}^d)} \le |u|_{\mathrm{BL}_k(\mathbb{R}^d)} \le C_e |u|_{\mathrm{BL}_k(\Omega)}.$$

Since the function $u - s_u$ satisfies $\lambda_j(u) = 0$ we can apply Theorem 5.2 to it, which gives the desired result. □

For here the mainly interesting case of $q = \infty$ gives

$$\|u - s_u\|_{L_\infty(\Omega)} \le C h^{k-d/2} |u|_{\mathrm{BL}_k(\Omega)}.$$

Hence, in the bivariate case $d = 2$ using classical thin-plate splines $\phi(r) = r^2 \log r$ this leads only to a first order approximation scheme. This recovers an unpublished result [3] by Gutzmer.

However, under additional assumptions on the target function u improved error estimates can be established.

THEOREM 6.2. *Under the assumptions of Corollary* 6.1, *assume that $u \in W_2^{2k}(\Omega)$ has support in Ω. Then, the error between u and its optimal recovery s_u can be bounded by*

$$\|u - s_u\|_{L_q(\Omega)} \le C h^{2k-d(1/2-1/q)_+} \|\Delta u^k\|_{L_2(\Omega)}.$$

Proof. Since u is supported in Ω it is actually globally defined. Moreover, we have the estimate

$$\|u - s_u\|_{L_q(\Omega)} \le C h^{k-d(1/2-1/q)_+} |u - s_u|_{\mathrm{BL}_k(\Omega)} \tag{6.1}$$

by Theorem 5.2. The improved estimate follows from bounding $|u - s_u|_{\mathrm{BL}_k(\Omega)}$. To this end, we use the fact that s_u is the best approximation to u from

$$\mathrm{span}\{\lambda_j^y \Phi_{d,k}(\cdot - y) : 1 \le j \le N\}.$$

Hence, using also the compact support of u, we deduce

$$|u - s_u|^2_{\mathrm{BL}_k(\Omega)} = (u - s_u, u)_{\mathrm{BL}_k(\Omega)} = \sum_{|\alpha|=k} \frac{k!}{\alpha!} \int_\Omega D^\alpha (u - s_u)(x) D^\alpha u(x) dx,$$

so that integration by parts results in

$$\begin{aligned} |u - s_u|^2_{\mathrm{BL}_k(\Omega)} &= (-1)^k \sum_{|\alpha|=k} \frac{k!}{\alpha!} \int_\Omega (u - s_u)(x) D^{2\alpha} u(x) dx \\ &= (-1)^k \int_\Omega (u - s_u)(x) \Delta^k u(x) dx, \end{aligned}$$

which can be bounded by

$$|u - s_u|^2_{\mathrm{BL}_k(\Omega)} \le \|u - s_u\|_{L_2(\Omega)} \|\Delta^k u\|_{L_2(\Omega)}.$$

Applying Theorem 5.2 to $\|u - s_u\|_{L_2(\Omega)}$ and inserting the result into (6.1) finishes the proof. □

This gives the estimate

$$\|u - s_u\|_{L_\infty(\Omega)} \le Ch^{2k-d/2} \|\Delta^k u\|_{L_2(\Omega)},$$

so that we have for thin-plate splines in $\mathbb{R}^2$ now a third order scheme, provided the target function is smooth enough and compactly supported.

This naturally raises the question of how good the approximation order can become under the best possible conditions. As in the case of pure point evaluation functionals (see [8]) it can be shown that any smooth function u with

$$\|u - s_u\|_{L_\infty(K)} = o(h^{2k})$$

for every compact subset $K \subseteq \Omega$ must already satisfy $\Delta^k u = 0$ and is in this sense trivial.

REFERENCES

[1] S. Brenner and L. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[2] J. Duchon, *Sur l'erreur d'interpolation des fonctions de plusieurs variables par les D^m-splines*, R.A.I.R.O. Anal. Numer., 12 (1978), pp. 325–334.

[3] T. Gutzmer, *Error estimates for reconstruction using thin plate spline interpolants*, Tech. Report 97-08, ETH Zürich, Switzerland, 1997.

[4] A. Iske and T. Sonar, *On the structure of function spaces in optimal recovery of point functionals for ENO-schemes by radial basis functions*, Numer. Math., 74 (1996), pp. 177–201.

[5] K. Jetter, J. Stöckler, and J. Ward, *Error estimates for scattered data interpolation on spheres*, Math. Comp., 68 (1999), pp. 733–747.

[6] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, *Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature*, Math. Comp., 70 (2001), pp. 1113–1130.

[7] F. J. Narcowich, J. D. Ward, and H. Wendland, *Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting*, Math. Comp., 74 (2005), pp. 743–763.

[8] R. Schaback and H. Wendland, *Inverse and saturation theorems for radial basis function interpolation*, Math. Comp., 71 (2002), pp. 669–681.

[9] T. Sonar, *Optimal recovery using thin plate splines in finite volume methods for the numerical solution of hyperbolic conservation laws*, IMA J. Numer. Anal., 16 (1996), pp. 549–581.

[10] T. Sonar, *On the construction of essentially non-oscillatory finite volume approximations to hyperbolic conservation laws on general triangulations: Polynomial recovery, accuracy, and stencil selection*, Comput. Methods in Appl. Mech. Engrg., 140 (1997), pp. 157–181.

[11] T. Sonar, *On families of pointwise optimal finite volume ENO approximations*, SIAM J. Numer. Anal., 35 (1998), pp. 2350–2369.

[12] H. Wendland, *Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree*, Adv. Comput. Math., 4 (1995), pp. 389–396.

[13] H. Wendland, *Error estimates for interpolation by compactly supported radial basis functions of minimal degree*, J. Approx. Theory, 93 (1998), pp. 258–272.

[14] H. Wendland, *Scattered Data Approximation*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2005.

SIAM J. Numer. Anal.
Vol. 43, No. 3, pp. 1003–1020

A PIECEWISE CONSTANT ALGORITHM FOR WEIGHTED L_1 APPROXIMATION OVER BOUNDED OR UNBOUNDED REGIONS IN $\mathbb{R}^s$*

FRED J. HICKERNELL[†], IAN H. SLOAN[‡], and GRZEGORZ W. WASILKOWSKI[§]

Abstract. Using Smolyak's construction [S. A. Smolyak, *Dokl. Akad. Nauk SSSR,* 4 (1963), pp. 240–243], we derive a new algorithm for approximating multivariate functions over bounded or unbounded regions in $\mathbb{R}^s$ with the error measured in a weighted L_1-norm. We provide upper bounds for the algorithm's cost and error for a class of functions whose mixed first order partial derivatives are bounded in the L_1-norm. In particular, we prove that the error and the cost (measured in terms of the number of function evaluations) satisfy the relation

$$\text{error} \le \frac{s\ \exp\left(\frac{1}{12(s-1)}\right)}{(s-1)\pi} \left(\frac{e\ \ln(\text{cost})}{(s-1)\sqrt{2}\ln(2)}\right)^{2(s-1)} \frac{1}{\text{cost}}$$

whenever the cost is sufficiently large relative to the number s of variables. More specifically, the inequality holds when $q \ge 2(s-1)$, where q is a special parameter defining the refinement level in Smolyak's algorithm, and hence the number of function evaluations used by the algorithm. We also discuss extensions of the results to the spaces with the derivatives bounded in L_p-norms.

Key words. Banach spaces, mixed first order partial derivatives, multivariate functions, Smolyak's construction

AMS subject classifications. 65D05, 65D15

DOI. 10.1137/S0036142903427664

1. Introduction.

In this paper, we derive a simple-to-use, piecewise constant algorithm for approximating functions in a weighted L_1 sense. Function approximation has been studied quite extensively; see, e.g., [13, 17, 22, 24] and the papers cited therein. However, such problems were considered mainly for functions with a bounded domain D, say $D = [0,1]^s$.

The worst case complexity of weighted approximation over unbounded domains D has recently been studied in, e.g., [11, 27], assuming that the corresponding function classes $\mathcal{F}$ are isotropic. The analysis of the approximation problem for tensor product spaces $\mathcal{F}$ is quite straightforward if $\mathcal{F}$ is a Hilbert space, since then desirable properties of Smolyak's construction could be used; see, e.g., [25].

In this paper, we study a weighted approximation problem with an emphasis on unbounded domains D and tensor product function classes $\mathcal{F}$ in a non-Hilbert-space setting. More specifically, we study a ρ-weighted L_1 approximation problem with the

*Received by the editors May 5, 2003; accepted for publication (in revised form) September 24, 2004; published electronically September 20, 2005. This work was partially supported by the Hong Kong Research Grants Council under grant HKBU/2020/02P, the Hong Kong Baptist University Faculty Research under grant FRG/00-01/II-62, the Australian Research Council, and the National Science Foundation under grant CCR-0095709.

http://www.siam.org/journals/sinum/43-3/42766.html

[†] Department of Mathematics, Hong Kong Baptist University, Kowloon Tong, Hong Kong. Current address: Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616 (hickernell@iit.edu).

[‡] School of Mathematics, University of New South Wales, Sydney 2052, Australia (sloan@maths.unsw.edu.au).

[§] Department of Computer Science, University of Kentucky, 773 Anderson Hall, Lexington, KY 40506-0046 (greg@cs.uky.edu).

error between f and the approximation $\mathcal{A}(f)$ measured in the following seminorm:

$$\|(f-\mathcal{A}(f))\,\rho\|_{L_1(D)} = \int_D \rho(\mathbf{x})\,|f(\mathbf{x})-\mathcal{A}(f)(\mathbf{x})|\,d\mathbf{x}. \tag{1}$$

We have chosen the L_1-norm here (as opposed to the L_r-norm for $r > 1$) since then the approximation problem is related to weighted integration. This relation is briefly discussed in section 6; here we only mention that any algorithm $\mathcal{A}$ for L_1 approximation yields an integration algorithm with the error at least as small as the error of $\mathcal{A}$. Another reason for choosing $r = 1$ is the simplicity of analysis; for $r \neq 1, 2$ the corresponding weighted L_r approximation problem is more difficult to analyze.

Throughout this article, it is assumed that $\rho(\mathbf{x}) = \prod_{k=1}^{s} \rho_k(x_k)$ is a given integrable weight function. The domain D of functions can be an arbitrary box, even $D = \mathbb{R}^s$. The class $\mathcal{F}$ is a space of functions whose dominating mixed first order derivatives are bounded in the L_1-norm. This seems to be about the minimum amount of smoothness needed to get a convergence rate of essentially $O(n^{-1})$. For example, if only the functions have bounded L_1-norm, then function values may not even be finite, and algorithms depending on function values are not be to guaranteed to converge. The case of derivatives bounded in L_p-norms seems to be harder than the case with the L_1-norm (especially for unbounded D), and is briefly addressed in section 5.

We stress that classes $\mathcal{F}$ as defined here are commonly assumed in the context of integration problems over $D = [0, 1]^s$, especially when dealing with quasi-Monte Carlo methods and discrepancy; see, e.g., [5, 12, 19] and papers cited therein. This point is further discussed in section 2. These $\mathcal{F}$ have been used even for integration problems with unbounded domains [9]. Moreover, for bounded D, these classes contain the classes MW_p^1 of periodic functions considered for approximation in [22, Chap. 4].

Instead of a condition on the dominating mixed derivatives, one sometimes assumes that all derivatives of total order α are bounded in the L_1-norm. Error analysis using such conditions suffer from the "curse of dimensionality." For example, to have a convergence rate of $O(n^{-1})$ requires all derivatives of total order s to be bounded in the L_1-norm. The condition assumed here is weaker.

The main result of the paper is the derivation and analysis of a family $\{\mathcal{A}_{q,s}\}_{q=s}^{\infty}$ of algorithms that provide approximations that are special piecewise constant functions. They are obtained by applying Smolyak's construction (see [21]) to scalar piecewise constant interpolation methods. As we shall see, given a number s of variables and a parameter $q \geq s$, the algorithm has the worst case error bounded by

$$\mathrm{error}(\mathcal{A}_{q,s}) \leq \begin{cases} s\,2^{-2q+2s-3}\binom{2q-2s+3}{q-s+1} & \text{if } q < 2(s-1), \\ s\,2^{-q}\binom{q}{s-1} & \text{if } q \geq 2(s-1). \end{cases}$$

Here q is the refinement parameter in Smolyak's algorithm. Under an additional symmetry assumption (4) that is stated later, we get

$$\mathrm{error}(\mathcal{A}_{q,s}) \leq s\,2^{-q+s-1-a}\binom{q-s}{a} \quad \text{with} \quad a = \begin{cases} \lfloor \frac{q-s+1}{3} \rfloor & \text{if } q < 4(s-1), \\ s-1 & \text{otherwise.} \end{cases}$$

Let $n = \mathrm{card}(\mathcal{A}_{q,s})$ denote the number of function evaluations used by $\mathcal{A}_{q,s}$. From [25, Lem. 7] we have $n \leq 2^{q-s+1}\binom{q-1}{s-1}$, and we conclude that for every $s \geq 2$ and $q \geq 2(s-1)$,

$$\mathrm{error}(\mathcal{A}_{q,s}) \leq \frac{s}{(s-1)\pi}\left(\frac{e\,\ln(n)}{(s-1)\sqrt{2}\ln(2)}\right)^{2(s-1)}\frac{1}{n}.$$

This rate $n^{-1}\ln^{2(s-1)}(n)$ of convergence is as good as the best known rate for L_1 approximation of periodic functions from the class $\mathcal{F} = MW_1^1$; see [22, Thm. 5.1 of Chap. 4] and [23]. Recall that MW_1^1 is contained in $\mathcal{F}$ when D^s is bounded and the approximation is with $\rho \equiv 1$. Hence $\mathcal{A}_{q,s}$ works for more general classes of problems and spaces of functions. Moreover, the result above shows in an explicit way the dependence of the errors on the dimension s.

We do not know whether $n^{-1}\ln^{2(s-1)}(n)$ is the best possible rate. However, because the Kolmogorov width for the corresponding problem over MW_1^1 equals $\Theta(n^{-1}\ln^{3(s-1)/2}(n))$ (see [22, Thm. 4.5 of Chap. 3]), the difference could only be in the exponent of the $\ln(n)$-term.

Since $\mathcal{A}_{q,s}(f)$ is a piecewise constant function, the algorithm is easy to implement. Its only drawback is in exponential gaps between consecutive numbers $\mathrm{card}(\mathcal{A}_{q,s})$ of function evaluations. However, we believe that it is of a practical interest, especially for small to moderate values of s. Implementation of the algorithm and numerical tests will be reported later.

We stress that Smolyak's construction, also referred to as *sparse grid*, *hyperbolic cross*, and *Boolean blending*, has been applied to many problems, including differential and integral equations, integration and approximation of multivariate functions, and to wavelets construction; see, e.g., [1, 2, 3, 4, 6, 7, 8, 10, 16, 18, 14, 22, 23, 25, 26] and the papers cited therein. However, with only a few exceptions, the considered functions are defined over bounded domains (e.g., $D = [0,1]^s$), and sometimes are of a special form (e.g., $f = g * B$ for a fixed function B as in MW_p^α classes). By analyzing ρ-weighted L_1 approximation, we are able to propose a simple algorithm that works well for any probability density function ρ of a tensor product form and even for $D = \mathbb{R}^s$. The proposed algorithm uses a finite number n of function values and approximates the function with error $O(n^{-1}\ln^{2(s-1)}(n))$. It is linear, simple, and easy to implement.

We would like to contrast this to a possible wavelet approach with $D = \mathbb{R}^s$ as proposed in [4]. There, the approximating algorithm is defined as the function from a special linear subspace $\mathcal{H}_n$ that minimizes the L_p distance from f, and the space $\mathcal{H}_n$ has infinite dimension. Hence, it is very difficult to implement. This should explain our interest in ρ-weighted approximation. We have chosen the L_1-norm since for the L_r-norm ($r > 1$) the analysis would be much harder and getting sharp bounds (including explicit constants) rather impossible.

We now explain the choice of low regularity $\alpha = 1$. Recall that we assume first order mixed derivatives to be bounded. As in the papers cited in the previous paragraph that deal with classes of functions of arbitrary (but fixed) regularity and bounded domains, it is possible to extend the results even to the case of $D = \mathbb{R}^s$. Indeed, using general properties of Smolyak's construction (e.g., from [25]) and results on the complexity of scalar ($s = 1$) weighted L_r approximation over $\mathbb{R}$ (see [27]), it is possible to achieve a convergence rate $O(n^{-(\alpha+(1/r-1/p)_-)}\ln^{(\alpha+1)(s-1)}(n))$. This is when the mixed partial derivatives of order α are bounded in the (weighted) L_p-norm. However, this extension to general α would (i) hold only under special assumptions on ρ (see [27, Thm. 1]), (ii) make explicit dependence on the dimension s very difficult to obtain, and (iii) make the resulting algorithms more difficult to implement and less applicable. In particular, the gaps between consecutive values of $\mathrm{card}(\mathcal{A}_{q,s})$ would increase with increasing regularity α, making the algorithm applicable only for very small values of s (say $s \le 3$). Moreover, by assuming $\alpha = 1$ we make our algorithm applicable to a larger class of functions also because the classes with $\alpha > 1$ are contained in the class with $\alpha = 1$. The explicitness of the results, as well as the simplicity and applicability of the algorithm, were our primary reasons for choosing $\alpha = 1$.

We summarize the content of this paper. Section 2 provides some basic definitions and assumptions, as well as an error bound for an arbitrary algorithm $\mathcal{A}$. Since Smolyak's construction depends on the specific choice of scalar algorithms, section 3 considers very special scalar algorithms based on a piecewise constant interpolation. The corresponding algorithm and its properties are presented in section 4. An extension to functions with derivatives bounded in L_p-norm is provided in section 5. Section 6 briefly explains why the error bounds obtained for the ρ-weighted L_1 approximation also hold for the corresponding ρ-weighted integration problem.

2. Basic definitions.

In this section, we briefly present some definitions and basic facts concerning the worst case setting. A more detailed discussion can be found, e.g., in [13, 24].

We consider a weighted L_1 approximation of functions of s variables whose domain D is a box,

$$D = \overline{(a_1, b_1)} \times \cdots \times \overline{(a_s, b_s)}.$$

The values a_i and b_i might be infinite; this is why we write $\overline{(a_i, b_i)}$ instead of $[a_i, b_i]$.

Let $\mathcal{F}$ be a Banach space of functions $f : D \to \mathbb{R}$ that will be specified later. The approximation problem depends on a weight function ρ which is assumed to have the following properties:

$$\rho(\mathbf{x}) = \prod_{k=1}^{s} \rho_k(x_k) \quad \text{and} \quad \rho_k \geq 0. \tag{2}$$

For simplicity of presentation, we also assume that

$$\int_{a_k}^{b_k} \rho_k(t)\, dt = 1 \quad \forall\, k = 1, \ldots, s.$$

However, it is enough to assume that the integrals of ρ_k are finite; in such a case, all error bounds derived in this paper should be multiplied by the constant $c = \int_D \rho(\mathbf{x})\, d\mathbf{x}$.

Functions from $\mathcal{F}$ are approximated by an algorithm $\mathcal{A}$,

$$f \sim \mathcal{A}(f) = \sum_{i=1}^{n} f(\mathbf{x}^i)\, g_i$$

for some points $\mathbf{x}^i$ and functions g_i, with the error between f and $\mathcal{A}(f)$ measured in the ρ-weighted L_1-norm; see (1). The *worst case error* (with respect to $\mathcal{F}$) of $\mathcal{A}$ is defined by

$$\text{error}(\mathcal{A}) := \sup_{\|f\| \leq 1} \|(f - \mathcal{A}(f))\, \rho\|_{L_1(D)},$$

where $\|f\|$ denotes the norm of f in the space $\mathcal{F}$. The importance of this definition is that due to linearity of $\mathcal{A}$ we have

$$\|(f - \mathcal{A}(f))\, \rho\|_{L_1} \leq \|f\|\ \text{error}(\mathcal{A}) \quad \forall\, f \in \mathcal{F}.$$

Each algorithm uses a finite number n of function evaluations. That number is called the *cardinality* and is denoted by $\mathrm{card}(\mathcal{A})$.

With the exception of sections 5 and 6, the following space $\mathcal{F}=\mathcal{F}_{1,s}$ is considered. Let $\mathcal{H}_k$ be the space of absolutely continuous functions on $\overline{(a_k,b_k)}$ whose first derivative is in $L_1((a_k,b_k))$. Let $\mathcal{H}^s = \bigotimes_{k=1}^s \mathcal{H}_k$ be the space consisting of linear combinations of functions f of the tensor product form

$$f : D \to \mathbb{R} \qquad \text{and} \qquad f(\mathbf{x}) = \prod_{k=1}^{s} h_k(x_k) \quad \text{with} \quad h_k \in \mathcal{H}_k.$$

The space $\mathcal{F}_{1,s}$ is the completion of $\mathcal{H}^s$ with respect to the following norm:

$$\|f\|_{1,s} := |f(\mathbf{c})| + \sum_{U\neq\emptyset} \|f'_U\|_{L_1(D_U)}. \tag{3}$$

Here $\mathbf{c} = [c_1,\dots,c_s] \in D$ is a fixed point, called an *anchor*. The summation is with respect to nonempty subsets U of $\{1,\dots,s\}$, and

$$f'_U(\mathbf{x}_U) := \frac{\partial^{|U|}}{\prod_{k\in U}\partial x_k} f(\mathbf{x}_U,\mathbf{c}),$$

where $(\mathbf{x}_U,\mathbf{c})$ denotes the s-dimensional vector whose kth component is x_k if $k \in U$, and c_k otherwise. By $\mathbf{x}_U$ we mean the $|U|$-dimensional vector obtained from $\mathbf{x}$ by removing all components x_k with $k \notin U$. This means that f'_U is a function defined on $D_U := \prod_{k\in U} \overline{(a_k,b_k)}$ and $\mathbf{x}_U \in D_U$. To simplify the notation, we will also write $f'_\emptyset$ and $\|f'_\emptyset\|_{L_1}$ to denote $f(\mathbf{c})$ and $|f(\mathbf{c})|$, respectively; and we often drop D_U by writing $\|\cdot\|_{L_1}$ instead of $\|\cdot\|_{L_1(D_U)}$. This allows the more concise formula

$$\|f\|_{1,s} = \sum_U \|f'_U\|_{L_1}.$$

We illustrate this for $s = 2$:

$$\begin{aligned}\|f\|_{1,2} = |f(c_1,c_2)| &+ \int_{a_1}^{b_1} \left|\frac{\partial}{\partial x_1} f(x_1,c_2)\right| dx_1 + \int_{a_2}^{b_2} \left|\frac{\partial}{\partial x_2} f(c_1,x_2)\right| dx_2 \\ &+ \int_{a_1}^{b_1}\int_{a_2}^{b_2} \left|\frac{\partial^2}{\partial x_1 \partial x_2} f(x_1,x_2)\right| dx_2\, dx_1.\end{aligned}$$

Although in general the anchor $\mathbf{c}$ and the weight ρ are not related, we shall obtain stronger results under the following symmetry condition:

$$\int_{a_k}^{c_k} \rho_k(x)\, dx = \frac{1}{2} \quad \forall k = 1,\dots,s. \tag{4}$$

Of course, this condition is satisfied if (a_k,b_k) and ρ_k are symmetric with respect to c_k ($k = 1,\dots,s$), e.g., if $a_k = -b_k$, $c_k = 0$, and $\rho_k(x) = \rho_k(-x)$. In general, for a given ρ_k one may always choose c_k to satisfy (4).

We now comment on the norm (3) and the role of the anchor $\mathbf{c}$.

Remark 1. As already mentioned in the introduction, the norm (3) and the space $\mathcal{F}_{1,s}$ are frequently assumed/studied in the context of quasi-Monte Carlo integration and discrepancies when the domain D is bounded, $D = [0,1]^s$. Then, classically, the anchor $\mathbf{c} = [1, \dots, 1]$. Of course, when $D = \mathbb{R}^s$ such a choice of the anchor seems unjustified and this is why we prefer to deal with arbitrary $\mathbf{c}$; see [9] for additional discussion.

There are a number of important results when, instead of the norm (3), a seminorm

$$|||f|||_{\alpha,p,s} := \|f^{(\alpha,\dots,\alpha)}\|_{L_p} \quad \text{with} \quad f^{(\alpha,\dots,\alpha)} = \left(\prod_{k=1}^{s} \frac{\partial^\alpha}{\partial x_k^\alpha} \right) f$$

is assumed. Of course, for $\alpha = 1$ and $p = 1$ this seminorm is very much related to (3) since it is equivalent to $\|f'_U\|_{L_1}$ with $U = \{1, \dots, s\}$. In particular, for $s = 1$, the complexities of problems with bounded $\|f\|_{1,s}$ or $|||f|||_{1,1,s}$ are equivalent. The situation changes drastically in the the multivariate case $s \geq 2$. This is because the subspace of functions with vanishing $|||\cdot|||_{\alpha,p,s}$-seminorm has infinite dimension. Therefore, to guarantee finite errors, one has to add additional restrictions on the considered class of functions. Examples of such restrictions include periodicity as in classes MW_p^α, or boundary conditions such as the vanishing of f and its partial derivatives at the points $\mathbf{x}$ with at least one component equal to zero. The presence of $\|f'_U\|_{L_1}$-terms in the definition of (3) guarantees that it is a well-defined norm without additional restrictions on the class of functions.

The following fact will play an important role. Let

$$M_k(x,t) := \begin{cases} 1 & \text{if } c_k < t < x, \\ -1 & \text{if } x < t < c_k, \\ 0 & \text{otherwise,} \end{cases}$$

$$M(\mathbf{x},\mathbf{t}) := \prod_{k=1}^{s} M_k(x_k, t_k) \quad \text{and} \quad M_U(\mathbf{x}_U, \mathbf{t}_U) := \prod_{k \in U} M_k(x_k, t_k),$$

with the convention that $M_\emptyset \equiv 1$. Then for every $f \in \mathcal{F}_{1,s}$ and every $\mathbf{x} \in D$,

$$f(\mathbf{x}) = \sum_U \int_{D_U} f'_U(\mathbf{t}_U)\, M_U(\mathbf{x}_U, \mathbf{t}_U)\, d\mathbf{t}_U. \tag{5}$$

The representation (5) has been used, at least implicitly, in a number of papers, and its short proof can be found in [9]. From it we have

$$\|f\,\rho\|_{L_1} \leq \sum_U \int_{D_U} |f'_U(\mathbf{t}_U)| \int_D |\rho(\mathbf{x})\, M_U(\mathbf{x}_U, \mathbf{t}_U)|\, d\mathbf{x}\, d\mathbf{t}_U \leq \|f\|_{1,s}.$$

This means that the approximation problem is well defined since the corresponding embedding operator is bounded. Actually, the following theorem, when applied to the zero algorithm $\mathcal{A} \equiv 0$, implies that the norm of the embedding is equal to one, i.e.,

$$\sup_{f \in \mathcal{F}} \frac{\|f\,\rho\|_{L_1}}{\|f\|_{1,s}} = 1.$$

THEOREM 1. *The error of any $\mathcal{A}$ is bounded by*

$$\text{error}(\mathcal{A}) \leq \sup_{\mathbf{t} \in D} \max_U \int_{D_U} h_U(\mathbf{x}_U, \mathbf{t}_U)\, d\mathbf{x}_U, \tag{6}$$

where

$$h_U(\mathbf{x}_U,\mathbf{t}_U) = \rho_U(\mathbf{x}_U)\,|M_U(\mathbf{x}_U,\mathbf{t}_U) - \mathcal{A}(M_U(\cdot,\mathbf{t}_U))(\mathbf{x}_U)|.$$

If $\mathcal{A}$ is based on piecewise constant interpolation then (6) *holds with equality.*

Proof. We defer the proof to section 5, where a more general result is proven. □

3. Scalar functions.

Since Smolyak's construction depends on specific algorithms for the scalar cases, we consider now approximating univariate functions whose domain is $\overline{(a_k,b_k)}$ and whose weight function is ρ_k. To simplify the notation in this section, we write a, b, and ω instead of a_k, b_k, and ρ_k.

For $i = 1, 2, \dots$, consider the set of points $x^k_{i,j}$ $(j = 0, \dots, 2^i)$ such that

$$a = x^k_{i,0} < x^k_{i,1} < \cdots < x^k_{i,2^i} = b$$

and

$$\int_{x^k_{i,j-1}}^{x^k_{i,j}} \omega(t)\,dt = 2^{-i}. \tag{7}$$

For simplicity, we will write in this section $x_{i,j}$ instead $x^k_{i,j}$.

Of course, $x_{i-1,j} = x_{i,2j}$. We take the following family of algorithms A_i $(i = 1, 2, \dots)$ based on piecewise constant interpolation:

$$A_i(f)(x) = f(x_{i,j}) \qquad \text{if } c_k \le x_{i,j} \le x < x_{i,j+1} \quad \text{or} \quad x_{i,j-1} < x \le x_{i,j} \le c_k.$$

Moreover, when $x, c_k \in (x_{i,j}, x_{i,j+1})$ then $A_i(f)(x)$ equals $f(x_{i,j})$ or $f(x_{i,j+1})$ depending on whether or not $x \le c_k$. Note that under the symmetry condition (4), $x_{1,1} = x_{i,2^{i-1}} = c_k$ and $A_i(f)(x) = f(c_k)$ for $x \in (x_{i,2^{i-1}-1}, x_{i,2^{i-1}+1})$.

For given k, define

$$\delta_{k,1}(x,t) := A_1(M_k(\cdot,t))(x)$$

and

$$\delta_{k,i}(x,t) := A_i(M_k(\cdot,t))(x) - A_{i-1}(M_k(\cdot,t))(x), \quad i \ge 2. \tag{8}$$

The following result is needed in section 4.

PROPOSITION 1. *For every t, we have*

$$\|\omega\,\delta_{k,i}(\cdot,t)\|_{L_1} \le 2^{-i} \quad \forall\, i \ge 2.$$

Moreover, $A_1(M_k(\cdot,t))$ is a constant function equal to ± 1 or zero, and if (4) *holds, then*

$$A_1(M_k(\cdot,t)) = 0.$$

The second part of the proposition can be directly checked. The first part of the proposition is an immediate consequence of the following lemma that will be used in section 5. For simplicity of presentation, we state the lemma only for arguments x and t greater than c_k. The analogous result is true for arguments smaller than c_k with the only difference being that $\delta_{i,k}(x,t) \in \{-1, 0\}$. Moreover, $\delta_{k,i}(x,t) = 0$ when c_k is between x and t.

Lemma 1. *The following statements hold for any $i \ge 2$ and $x, t > c_k$.*

(i) $\delta_{k,i}(x,t) \in \{0,1\}$.

(ii) *If $\delta_{k,i}(x,t) = 1$, then there exists $j \le 2^{i-1} - 1$ such that*

$$t \in (x_{i,2j}, x_{i,2j+1}] \qquad \textit{and} \qquad x \in (x_{i,2j+1}, x_{i,2j+2}].$$

(iii) *There is at most one $i \ge 2$ such that $\delta_{k,i}(x,t) = 1$.*

Proof. (i) follows immediately from the facts that $M_k(x,t) \in \{0,1\}$, $M_k(\cdot,t)$ is nondecreasing, and A_i uses the points used by A_{i-1}.

(ii) Let $x \in (x_{i,\ell}, x_{i,\ell+1}]$ for some ℓ. Then $A_i(M_k(\cdot,t))(x) = 1$ only if $t < x_{i,\ell}$. However, for $A_{i-1}(M_k(\cdot,t))(x) = 0$, $x_{i,\ell}$ has to be different than any point $x_{i-1,j}$ used by A_{i-1}. This means that ℓ has to be odd.

(iii) follows from (ii). Indeed, let (x,t) be fixed. If $\delta_{k,i}(x,t) = 1$, then t and x are in two neighboring subintervals with the evaluation point $x_{i,\ell}$ between them and $\ell = 2j+1$. Consider now $\delta_{k,i+n}(x,t)$ for positive n. Then the points $x_{i,\ell} = x_{i+n,m}$ with $m = 2^n \ell$ is between t and x, yet m is even. Hence, due to part (ii), $\delta_{k,i+n}(x,t)$ cannot be equal to one. This completes the proof. □

4. The algorithm.

Let $\{A_{k,i}\}$ be the families of algorithms from the previous section, each for $\omega = \rho_k$ and $(a,b) = (a_k, b_k)$, respectively. Recall that $A_{k,i}$ uses function values at points $x^k_{i,1}, \dots, x^k_{i,2^i-1}$. Define

$$\Delta_{k,1} := A_{k,1}, \qquad \Delta_{k,i} := A_{k,i} - A_{k,i-1} \quad \text{for} \quad i \ge 2$$

and

$$\mathcal{A}_{q,s} := \sum_{|\mathbf{i}| \le q} \bigotimes_{k=1}^{s} \Delta_{k,i_k} \tag{9}$$

for $q \ge s$. Here and elsewhere, $\mathbf{i} = [i_1, \dots, i_s] \in \mathbb{N}^s_+$ is a multi-index with $i_k \ge 1$ and $|\mathbf{i}| = \sum_{k=1}^s i_k$.

Theorem 2. *Let $s \ge 2$ and $q \ge s$. Then*

$$\mathrm{error}(\mathcal{A}_{q,s}) \le \begin{cases} s\, 2^{-q} \binom{q}{\lfloor \frac{q+1}{2} \rfloor} & \textit{if } q < 2(s-1), \\ s\, 2^{-q} \binom{q}{s-1} & \textit{if } q \ge 2(s-1). \end{cases}$$

If, additionally, (4) *holds, then*

$$\mathrm{error}(\mathcal{A}_{q,s}) \le s\, 2^{-q+s-1-a} \binom{q-s}{a} \quad \textit{with} \quad a = \begin{cases} \lfloor \frac{q-s+1}{3} \rfloor & \textit{if } q < 4(s-1), \\ s-1 & \textit{if } q \ge 4(s-1). \end{cases}$$

To prove this theorem, we need the following lemma. We think it is known; however, we have not found it in the literature. For $q \ge s-1$, define

$$\mathrm{B}(q,s) := \sum_{|\mathbf{i}| \ge q+1} 2^{-|\mathbf{i}|}. \tag{10}$$

Lemma 2. *For every $q \ge s-1$,*

$$\mathrm{B}(q,s) = 2^{-q} \sum_{j=0}^{s-1} \binom{q}{j} \le \overline{\mathrm{B}}(q,s), \tag{11}$$

where

$$\overline{\mathrm{B}}(q,s) := s\,2^{-q} \begin{cases} \binom{q}{\lfloor \frac{q+1}{2} \rfloor} & \textit{if } 2s \geq q+3, \\ \binom{q}{s-1} & \textit{otherwise.} \end{cases} \tag{12}$$

Proof of Lemma 2. Indeed,

$$\mathrm{B}(q,s) = \sum_{j=q+1}^{\infty} 2^{-j} \binom{j-1}{s-1}$$

and

$$\begin{aligned} 2^{-j}\binom{j-1}{s-1} &= 2^{-s} \cdot 2^{-(j-s)} \cdot \frac{(j-1)\cdots(j-s+1)}{(s-1)!} \\ &= \frac{x^s}{(s-1)!} \cdot \left(x^{j-1}\right)^{(s-1)} \Big|_{x=1/2}. \end{aligned}$$

Taking the summation with respect to j inside the differentiation, we get

$$\begin{aligned} \mathrm{B}(q,s) &= \frac{x^s}{(s-1)!} \left(\sum_{j=q+1}^{\infty} x^{j-1} \right)^{(s-1)} \Bigg|_{x=1/2} = \frac{x^s}{(s-1)!} \left(\frac{x^q}{1-x} \right)^{(s-1)} \Bigg|_{x=1/2} \\ &= \frac{x^s}{(s-1)!} \sum_{j=0}^{s-1} \binom{s-1}{j} \left(x^q\right)^{(j)} \left((1-x)^{-1}\right)^{(s-1-j)} \Bigg|_{x=1/2} \\ &= x^s \sum_{j=0}^{s-1} \frac{1}{j!(s-1-j)!} \cdot x^{q-j} \cdot q \cdots (q-j+1) \cdot \frac{(s-1-j)!}{(1-x)^{s-j}} \Bigg|_{x=1/2}, \end{aligned}$$

which, after some elementary manipulation, can be shown to be equal to $2^{-q} \sum_{j=0}^{s-1} \binom{q}{j}$. This completes the proof of (11). The upper bound $\overline{\mathrm{B}}(q,s)$ follows from this and the well-known fact that $\binom{q}{j-1} \leq \binom{q}{j}$ iff $2j \leq q+1$. This completes the proof of Lemma 2. □

Proof of Theorem 2. Note that for any scalar function $f \in \mathcal{H}_k$, $A_{k,n}(f)$ converges pointwise to f with $n \to \infty$ and that $\sum_{i=1}^{n} \Delta_{k,i} = A_{k,n}$. Therefore,

$$f(\mathbf{x}) = \sum_{\mathbf{i} \in \mathbb{N}_+^s} \bigotimes_{k=1}^{s} \Delta_{k,i_k}(f)(\mathbf{x})$$

for any $\mathbf{x}$ and any $f \in \mathcal{F}_{1,s}$. Here and elsewhere, by $\sum_{\mathbf{i} \in \mathbb{N}^s}$ we mean a double sum $\sum_{\ell=s}^{\infty} \sum_{|\mathbf{i}|=\ell}$. Hence

$$f - \mathcal{A}_{q,s}(f) = \mathcal{E}_{q,s}(f) \quad \text{with} \quad \mathcal{E}_{q,s}(f) := \sum_{|\mathbf{i}| \geq q+1} \bigotimes_{k=1}^{s} \Delta_{k,i_k}(f).$$

Due to (5) and the fact that $\mathcal{E}_{q,s}$ vanishes on constant functions,

$$\mathcal{E}_{q,s}(f) = \sum_{U \neq \emptyset} \int_{D_U} f'_U(\mathbf{t}_U)\, \mathcal{E}_{q,s}(M_U(\cdot, \mathbf{t}_U))\, d\mathbf{t}_U,$$

which yields

$$\|(f-\mathcal{A}_{q,s}(f))\,\rho\|_{L_1} \le \|f\|_{1,s}\cdot \sup_{\mathbf{t}\in D}\max_{U\neq\emptyset}\|\rho\,\mathcal{E}_{q,s}(M_U(\cdot,\mathbf{t}_U))\|_{L_1}.$$

Hence, to complete the proof, we only need to estimate the above supremum. For that purpose, note that

$$\bigotimes_{k=1}^{s}\Delta_{k,i_k}(M_U(\cdot,\mathbf{t}_U))\equiv 0 \quad \text{if } i_k\ge 2 \text{ for some } k\notin U.$$

This follows from the fact that $\Delta_{k,i}(1)\equiv 0$ for any $i\ge 2$, which, in turn, is a consequence of the fact that $A_{k,i}(1)\equiv 1$ for any $i\ge 1$. Otherwise, i.e., when $i_k=1$ for all $k\notin U$,

$$\left\|\rho\bigotimes_{k=1}^{s}\Delta_{k,i_k}(M_U(\cdot,\mathbf{t}_U))\right\|_{L_1} = \prod_{k\in U}\|\rho_k\,\Delta_{k,i_k}(M_k(\cdot,t_k))\|_{L_1} \le \prod_{k\in U}2^{-i_k},$$

independently of $\mathbf{t}$, due to Proposition 1. Therefore

$$(13)\qquad \|\rho\,\mathcal{E}_{q,s}(M_U(\cdot,\mathbf{t}_U))\|_{L_1} \le \sum_{\mathbf{i}}2^{-|\mathbf{i}|} \le \mathrm{B}(q-s+|U|,|U|) \le \overline{\mathrm{B}}(q-s+|U|,|U|)$$

with the summation in (13) over $\mathbf{i}\in\mathbb{N}^{|U|}$ such that $|\mathbf{i}|\ge q+1-(s-|U|)$. The first result in the theorem is proved by showing that $\overline{\mathrm{B}}(q-s+|U|,|U|)$ is increasing with $|U|$, i.e., $\max_U\overline{\mathrm{B}}(q-s+|U|,|U|)=\overline{\mathrm{B}}(q,s)$.

Consider first $q\ge 2(s-1)$. To see that $\overline{\mathrm{B}}(q-s+|U|,|U|)$ is increasing with $|U|$, note that

$$\begin{aligned}\frac{\overline{\mathrm{B}}(q-s+|U|+1,|U|+1)}{\overline{\mathrm{B}}(q-s+|U|,|U|)} &= \frac{(|U|+1)(q-s+|U|+1)}{2|U|^2}\\ &\ge \frac{q-s+|U|+1}{2|U|}\ge 1\end{aligned}$$

with the last inequality due to the fact that $|U|+1\le s$ and hence $|U|\le s-1\le q-s+1$.

Suppose now $q<2(s-1)$. To show that $\overline{\mathrm{B}}(q-s+|U|,|U|)$ increases with $|U|$ also in this case, we need to consider three different cases. Consider first $|U|>q-s+3$ (i.e., $2|U|>q-s+|U|+3$). Then

$$\begin{aligned}\frac{\overline{\mathrm{B}}(q-s+|U|+1,|U|+1)}{\overline{\mathrm{B}}(q-s+|U|,|U|)} &= \frac{|U|+1}{2|U|}\binom{q-s+|U|+1}{\lfloor\frac{q-s+|U|+2}{2}\rfloor}\Big/\binom{q-s+|U|}{\lfloor\frac{q-s+|U|+1}{2}\rfloor}\\ &= \frac{|U|+1}{|U|}\,\frac{2\ell+1}{2\ell+2}>1\end{aligned}$$

with the last equality due to an extra assumption that $q-s+|U|=2\ell$ (the proof for odd $q-s+|U|$ is very similar).

Consider next the case of $|U|=q-s+2$. Then

$$\frac{\overline{\mathrm{B}}(q-s+|U|+1,|U|+1)}{\overline{\mathrm{B}}(q-s+|U|,|U|)} = \frac{(q-s+3)(2q-2s+3)}{(q-s+2)(2q-2s+4)}\ge 1.$$

Consider now $|U|+1<q-s+3$. Then

$$\begin{aligned}\frac{\overline{\mathrm{B}}(q-s+|U|+1,|U|+1)}{\overline{\mathrm{B}}(q-s+|U|,|U|)} &= \frac{|U|+1}{2|U|}\binom{q-s+|U|+1}{|U|}\Big/\binom{q-s+|U|}{|U|-1}\\ &= \frac{(|U|+1)(q-s+|U|+1)}{2|U|^2}\geq 1.\end{aligned}$$

This completes the proof of the first part of the theorem.

Assume now that (4) holds. Then

$$\bigotimes_{k=1}^{s}\Delta_{k,i_k}(M_U(\cdot,\mathbf{t}_U))\equiv 0 \quad \text{if } i_k=1 \text{ for some } k\in U$$

as follows directly from the second part of Proposition 1. This means that the sum in (13) is now over $\mathbf{i}\in\mathbb{N}^{|U|}$ such that $|\mathbf{i}|\geq\max\{q+1-s+|U|,2|U|\}$ and $\mathbf{i}\geq\mathbf{2}$. Replacing $\mathbf{i}$ by $\mathbf{j}=\mathbf{i}-\mathbf{1}$, the sum becomes

$$(14)\qquad 2^{-|U|}\sum_{|\mathbf{j}|\geq\max\{q+1-s,|U|\}}2^{-|\mathbf{j}|}=2^{-|U|}\cdot\mathrm{B}(\max\{q-s,|U|-1\},|U|).$$

Therefore, for $|U|<q-s+1$ we have

$$(15)\qquad \|\rho\,\mathcal{E}_{q,s}(M_U(\cdot,\mathbf{t}_U))\|_{L_1}\leq 2^{-q-|U|+s}\sum_{j=0}^{|U|-1}\binom{q-s}{j}$$

and for $|U|\geq q-s+1$ we have

$$(16)\qquad \|\rho\,\mathcal{E}_{q,s}(M_U(\cdot,\mathbf{t}_U))\|_{L_1}\leq 2^{-|U|}\leq 2^{-q+s-1}.$$

Since $|U|\leq s$, we have to have $q\leq 2s-1$ for the latter case to happen. To estimate the maximum of the upper bound with respect to $|U|$, we consider the following cases.

Case $q\geq 3(s-1)$. Then, because $q-s\geq 2|U|-3$, the right-hand side of (15) is bounded from above by $s\,2^{-q-|U|+s}\binom{q-s}{|U|-1}$, which can be shown to increase with $|U|$ as long as $|U|\leq 1+\lfloor(q-s+1)/3\rfloor$. Since $|U|\leq s$, this yields the upper bounds

$$\|\rho\,\mathcal{E}_{q,s}(M_U(\cdot,\mathbf{t}_U))\|_{L_1}\leq s\,2^{-q}\binom{q-s}{s-1},$$

if $q\geq 4(s-1)$, and

$$\|\rho\,\mathcal{E}_{q,s}(M_U(\cdot,\mathbf{t}_U))\|_{L_1}\leq s\,2^{-q+s-1-\lfloor\frac{q-s+1}{3}\rfloor}\binom{q-s}{\lfloor(q-s+1)/3\rfloor},$$

if $q/(s-1)\in[3,4)$.

Case $2(s-1)<q<3(s-1)$. First note that the right-hand side of (15) decreases with $|U|$ when $|U|>(q-s)/2$. This follows from the fact that the value of the right-hand side for $|U|$ minus the value for $|U|+1$ equals

$$2^{-q-|U|-1+s}\left(\sum_{j=0}^{|U|-1}\binom{q-s}{j}-\binom{q-s}{|U|}\right)\geq 0$$

as claimed. Moreover, for $|U| \le (q-s)/2$, the right-hand side of (15) is bounded by $s\,2^{-q-|U|+s}\binom{q-s}{|U|-1}$, which attains its maximum for $|U| = 1 + \lfloor (q-s+1)/3 \rfloor$. Hence again,

$$\|\rho\,\mathcal{E}_{q,s}(M_U(\cdot,\mathbf{t}_U))\|_{L_1} \le s\,2^{-q+s-1-\lfloor \frac{q-s+1}{3} \rfloor}\binom{q-s}{\lfloor (q-s+1)/3 \rfloor}.$$

Case $q \le 2(s-1)$. Due to (16), we need only to estimate the right-hand side of (15) for $|U| < q-s+1$. However, as in the previous case, it is bounded by $s\,2^{-q-|U^*|+s}\binom{q-s}{|U^*|-1}$ with $|U^*| = 1 + \lfloor (q-s+1)/3 \rfloor$. Since

$$s\,2^{-q+s-1-\lfloor \frac{q-s+1}{3} \rfloor}\binom{q-s}{\lfloor (q-s+1)/3 \rfloor} \ge 2^{-q+s-1},$$

the left-hand side of the above inequality is an upper bound on $\|\rho\,\mathcal{E}_{q,s}(M_U(\cdot_U,\mathbf{t}_U))\|_{L_1}$ also in this case. This completes the proof. □

We end this section by relating the error of $\mathcal{A}_{q,s}$ to its cardinality $\mathrm{card}(\mathcal{A}_{q,s})$.

THEOREM 3. *For every* $s \ge 2$ *and* $q \ge 2(s-1)$,

$$\text{(17)} \qquad \mathrm{error}(\mathcal{A}_{q,s}) \le \frac{s\,\exp\left(\frac{1}{12(s-1)}\right)}{(s-1)\pi}\left(\frac{e\,\ln(\mathrm{card}(\mathcal{A}_{q,s}))}{(s-1)\sqrt{2}\ln(2)}\right)^{2(s-1)}\frac{1}{\mathrm{card}(\mathcal{A}_{q,s})}.$$

Proof. Since the information used by $\mathcal{A}_{q,s}$ is nested, we know from [25, Lem. 7] that

$$\text{(18)} \qquad 2^{q-s}\binom{q-1}{s-1} \le \mathrm{card}(\mathcal{A}_{q,s}) \le 2^{q-s+1}\binom{q-1}{s-1}.$$

Let's represent q as $q = (t+1)(s-1)$ with $t \ge 1$. We apply Stirling's formula ($n! = (n/e)^n\sqrt{2\pi n}\exp(\theta/(12n))$ for $\theta \in (0,\pi)$) to the error bound from Theorem 2; one can show that

$$\begin{aligned}\mathrm{error}(\mathcal{A}_{q,s}) &\le s\,2^{-(t+1)(s-1)}\binom{(t+1)(s-1)}{s-1}\\ &\le s\,2^{-(t+1)(s-1)}\,((t+1)e)^{s-1}\sqrt{\frac{t+1}{t2\pi(s-1)}}\,\exp\left(\frac{1}{12q}\right)\\ &\le s\,2^{-(t+1)(s-1)}\,((t+1)e)^{s-1}\sqrt{\frac{\exp\left(\frac{1}{12(s-1)}\right)}{\pi(s-1)}},\end{aligned}$$

with the last inequality due to the fact that $t \ge 1$. Similarly,

$$\mathrm{card}(\mathcal{A}_{q,s}) \le 2^{t(s-1)}\,(e(t+1))^{s-1}\sqrt{\frac{\exp\left(\frac{1}{12(s-1)}\right)}{\pi(s-1)}}.$$

Since $x/(\ln(x))^{2(s-1)}$ increases for $x \ge e^{2(s-1)}$, and since $\mathrm{card}(\mathcal{A}_{q,s}) \ge e^{2(s-1)}$ due to (18), we can replace $\mathrm{card}(\mathcal{A}_{q,s})$ by the right-hand side of the above inequality in the following estimation:

$$\begin{aligned}L &:= \frac{\mathrm{card}(\mathcal{A}_{q,s})\ \mathrm{error}(\mathcal{A}_{q,s})}{(\ln(\mathrm{card}(\mathcal{A}_{q,s})))^{2(s-1)}}\\ &\le \frac{s\,\exp\left(\frac{1}{12(s-1)}\right)}{\pi(s-1)}\,2^{-(s-1)}\left(\frac{e}{s-1}\,g(t)\right)^{2(s-1)},\end{aligned}$$

with

$$g(t) := \frac{t+1}{t\ln(2)+\ln(t+1)}$$

since $s-1-\ln(\pi(s-1))/2$ is positive. It is easy to verify that $\max_{t\geq 1} g(t) = 1/\ln(2)$. This yields

$$\begin{aligned} L &\leq \frac{s\ \exp\left(\frac{1}{12(s-1)}\right)}{(s-1)\pi}\, 2^{-(s-1)} \left(\frac{e}{(s-1)\ln(2)}\right)^{2(s-1)} \\ &= \frac{s\ \exp\left(\frac{1}{12(s-1)}\right)}{(s-1)\pi} \left(\frac{e}{(s-1)\sqrt{2}\ln(2)}\right)^{2(s-1)}, \end{aligned}$$

which completes the proof. □

5. Extensions. In this section, we extend some of the previous results assuming now that the functions f are from the space $\mathcal{F} = \mathcal{F}_{p,s,\gamma}$, which is the completion of $\mathcal{H}^s$ with respect to the norm

$$\|f\|_{p,s,\gamma} := \left(|f(\mathbf{c})|^p + \sum_{U\neq\emptyset} \gamma_{s,U}^{-p}\, \|f'_U\|_{L_p}^p\right)^{1/p} = \left(\sum_U \gamma_{s,u}^{-p} \|f'_U\|_{L_p}^p\right)^{1/p},$$

where $p \in [1,\infty]$ and $\gamma = \{\gamma_{s,u}\}_{s,U}$ is a family of nonnegative numbers, called weights. By a convention $0/0 = 0$. Of course, for $p = \infty$ we have $\|f\|_{\infty,s} = \max_U \gamma_{s,U}^{-1} \|f'_U\|_{L_\infty}$. This norm differs from $\|\cdot\|_{1,s}$ by using L_p instead of L_1-norms and by adding the weights $\gamma_{s,U}$.

The role of $\gamma_{s,U}$ is that they model how important certain variables and their groups are. For instance, the condition of small enough $\gamma_{s,U}$ means that $\|f'_U\|_{L_p}$ cannot be too large, and $\gamma_{s,U} = 0$ implies that $\|f'_U\|_{L_p} = 0$. Since the introduction of weighted norms in [20], high dimensional problems with such norms have been investigated in a number of papers; see, e.g., [15] and papers cited there. It is often the case that if the weights are small enough, then the error in solving a high dimensional problem is essentially no worse than that for solving a one-dimensional problem.

To stress that now a different space from $\mathcal{F}_{1,s}$ is being considered, we will write error$(\mathcal{A}; \mathcal{F}_{p,s,\gamma})$ instead error$(\mathcal{A})$. When all weights equal 1, we will simply write $\mathcal{F}_{p,s}$. For simplicity, we assume throughout this section that (4) is satisfied and that $\mathbf{c} = \mathbf{0}$. Hence, in particular $a_k < 0 < b_k$.

Note that for $p > 1$ and unbounded D it could happen that the approximation problem is not well defined since the corresponding embedding operator could be unbounded. As follows from [27, Thm. 1], the problem is well defined iff

$$\left(\int_{a_k}^{b_k} \psi_k^{p^*}(x)\,dx\right)^{1/p^*} < \infty \quad \forall k, \tag{19}$$

with

$$\psi_k(t) = \begin{cases} \int_t^{b_k} \rho_k(x)\,dx & \text{for } t \geq 0, \\ \int_{a_k}^{t} \rho_k(x)\,dx & \text{otherwise,} \end{cases} \tag{20}$$

where here and elsewhere p^* denotes the conjugate to p, i.e.,

$$\frac{1}{p}+\frac{1}{p^*}=1.$$

This is why we assume (19) throughout the rest of this paper. Of course, (19) trivially holds when D is bounded. It also holds for $p=1$ since then $p^*=\infty$ and the left-hand side of (19) should formally be replaced by $\operatorname{ess\,sup}_{t\ge c_k}\int_t^{b_k}\rho_k(x)\,dx$, which obviously is equal to $1/2$.

THEOREM 4. *The error of any algorithm $\mathcal{A}$ is bounded by*

$$\text{(21)}\qquad \operatorname{error}(\mathcal{A};\mathcal{F}_{p,s,\gamma})\le\left(\sum_U\gamma_{s,U}^{p^*}\int_{D_U}\left(\int_{D_U}h_U(\mathbf{x}_U,\mathbf{t}_U)\,d\mathbf{x}_U\right)^{p^*}d\mathbf{t}_U\right)^{1/p^*},$$

where

$$h_U(\mathbf{x}_U,\mathbf{t}_U)=\rho_U(\mathbf{x}_U)\,|M_U(\mathbf{x}_U,\mathbf{t}_U)-\mathcal{A}(M_U(\cdot,\mathbf{t}_U))(\mathbf{x}_U)|.$$

If $\mathcal{A}$ is based on piecewise constant interpolation, then we have equality in (21).

Proof. To simplify the notation, we write $m(\mathbf{x}_U,\mathbf{t}_U)$ to denote

$$m(\mathbf{x}_U,\mathbf{t}_U):=M(\mathbf{x}_U,\mathbf{t}_U)-\mathcal{A}(M(\cdot,\mathbf{t}_U))(\mathbf{x}_U).$$

Of course, $h_U(\mathbf{x}_U,\mathbf{t}_U)=\rho_U(\mathbf{x}_U)\,|m_U(\mathbf{x}_U,\mathbf{t}_U)|$.

Suppose that $p>1$. We begin with the proof of (21). Using (5), we have by Hölder's inequality

$$\begin{aligned}
&\int_D\rho(\mathbf{x})\,|f(\mathbf{x})-\mathcal{A}(f)(\mathbf{x})|\,d\mathbf{x}\\
&\quad=\int_D\rho(\mathbf{x})\left|\sum_U\int_{D_U}f_U'(\mathbf{t}_U)\,m_U(\mathbf{x}_U,\mathbf{t}_U)\,d\mathbf{t}_U\right|\,d\mathbf{x}_U\\
&\quad\le\sum_U\int_{D_U}\rho_U(\mathbf{x}_U)\int_{D_U}|f_U'(\mathbf{t}_U)\,m_U(\mathbf{x}_U,\mathbf{t}_U)|\,d\mathbf{t}_U\,d\mathbf{x}_U\\
&\quad=\sum_U\int_{D_U}|f_U'(\mathbf{t}_U)|\int_{D_U}h_U(\mathbf{x}_U,\mathbf{t}_U)\,d\mathbf{x}_U\,d\mathbf{t}_U\\
&\quad\le\sum_U\|f_U'\|_{L_p}\,\gamma_{s,U}^{-1}\left(\gamma_{s,U}^{p^*}\int_{D_U}\left(\int_{D_U}h_U(\mathbf{x}_U,\mathbf{t}_U)\,d\mathbf{x}_U\right)^{p^*}d\mathbf{t}_U\right)^{1/p^*}\\
&\quad\le\|f\|_{p,s,\gamma}\left(\sum_U\gamma_{s,U}^{p^*}\int_{D_U}\left(\int_{D_U}h_U(\mathbf{x}_U,\mathbf{t}_U)\,d\mathbf{x}_U\right)^{p^*}d\mathbf{t}_U\right)^{1/p^*}.
\end{aligned}$$

This proves (21) for $p>1$. We now show equality when $\mathcal{A}(m_U(\cdot,\mathbf{t}_U))$ is a piecewise constant function interpolating $m_U(\cdot,t_U)$. To that end, define

$$g_U(\mathbf{t}_U):=\operatorname{sign}(\mathbf{t}_U)\left(\int_{D_U}h(\mathbf{x}_U,\mathbf{t}_U)\,d\mathbf{x}_U\right)^{p^*-1}\quad\text{with}\quad\operatorname{sign}(\mathbf{t}_U):=\prod_{k\in U}\operatorname{sign}(t_k),$$

and

$$\widetilde{f}(\mathbf{y}):=\sum_U\gamma_{s,U}^{p^*}\int_{D_U}g_U(\mathbf{t}_U)\,M_U(\mathbf{y}_U,\mathbf{t}_U)\,d\mathbf{t}_U.$$

Of course, from (5), $\widetilde{f} \in \mathcal{F}_{p,s,\gamma}$ and $\widetilde{f}'_U = \gamma_{s,U}^{p^*} g_U$. Since $p p^* - p = p^*$, it is easy to check that

$$\|\widetilde{f}\|_{p,s,\gamma}, = \left(\sum_U \gamma_{s,U}^{p^*} \int_{D_U} \left(\int_{D_U} h_U(\mathbf{x}_U, \mathbf{t}_U)\, d\mathbf{x}_U \right)^{p^*} d\mathbf{t}_U \right)^{1/p}.$$

Moreover,

$$\begin{aligned}
\|(\widetilde{f} - \mathcal{A}(\widetilde{f}))\,\rho\|_{L_1} &= \int_D \rho(\mathbf{y})\, |\widetilde{f}(\mathbf{y}) - \mathcal{A}(\widetilde{f})(\mathbf{y})|\, d\mathbf{y} \\
&= \int_D \rho(\mathbf{y}) \left| \sum_U \gamma_{s,U}^{p^*} \int_{D_U} g_U(\mathbf{t}_U)\, m(\mathbf{y}_U, \mathbf{t}_U)\, d\mathbf{t}_U \right| d\mathbf{y} \\
&= \int_D \rho(\mathbf{y}) \left| \sum_U \gamma_{s,U}^{p^*} \int_{D_U} \left(\int_{D_U} h(\mathbf{x}_U, \mathbf{t}_U)\, d\mathbf{x}_U \right)^{p^*-1} \operatorname{sign}(\mathbf{t}_U)\, m(\mathbf{y}_U, \mathbf{t}_U)\, d\mathbf{t}_U \right| d\mathbf{y} \\
&= \sum_U \gamma_{s,U}^{p^*} \int_{D_U} \left(\int_{D_U} h(\mathbf{x}_U, \mathbf{t}_U)\, d\mathbf{x}_U \right)^{p^*-1} \int_{D_U} h(\mathbf{y}_U, \mathbf{t}_U)\, d\mathbf{y}_U\, d\mathbf{t}_U \\
&= \sum_U \gamma_{s,U}^{p^*} \int_{D_U} \left(\int_{D_U} h_U(\mathbf{x}_U, \mathbf{t}_U)\, d\mathbf{x}_U \right)^{p^*} d\mathbf{t}_U = \|\widetilde{f}\|_{p,s,\gamma}^p,
\end{aligned}$$

with the third-to-last equality due to the fact that

$$\operatorname{sign}(\mathbf{t}_U)\, m(\mathbf{y}_U, \mathbf{t}_U) = |m(\mathbf{y}_U, \mathbf{t}_U)| \ge 0 \quad \forall \mathbf{y}, \mathbf{t}.$$

Dividing $\|(\widetilde{f} - \mathcal{A}(\widetilde{f}))\,\rho\|_{L_1}$ by $\|\widetilde{f}\|_{p,s,\gamma}$ to obtain $\|\widetilde{f}\|_{p,s,\gamma}^{p-1} = \|\widetilde{f}\|_{p,s,\gamma}^{p/p^*}$ on the right completes the proof of the equality for $p > 1$.

For $p = 1$ the proof technique is the same with obvious modifications. □

Let $\mathcal{A}_{q,s}$ be the algorithm from section 4 and denote

$$\delta_{U,\mathbf{i}_U}(\mathbf{x}_U, \mathbf{t}_U) := \Delta_{U,\mathbf{i}_U}(M_U(\cdot, \mathbf{t}_U))(\mathbf{x}_U) = \prod_{k \in U} \delta_{k,i_k}(x_{i_k}, t_{i_k}).$$

From Lemma 1, we have the following proposition.

PROPOSITION 2. *Let $U \neq \emptyset$ and $\mathbf{t}_U \in D_U$. Then the following hold.*

(i) *For every $\mathbf{x}_U \in D_U$, $\delta_{U,\mathbf{i}_U}(\mathbf{x}_U, \mathbf{t}_U) \in \{-1, 0, 1\}$.*

(ii) *For every $\mathbf{x} \in D_U$ there exists at most one $\mathbf{i}_U$ such that $|\delta_{U,\mathbf{i}_U}(\mathbf{x}_U, \mathbf{t}_U)| = 1$.*

(iii) *For every $\mathbf{i}$, the ρ_U-probability of the set of $\mathbf{x}_U$'s with $|\delta_{U,\mathbf{i}_U}(\mathbf{x}_U, \mathbf{t}_U)| = 1$ is at most $2^{-|\mathbf{i}_U|}$.*

(iv) *If $\delta_{U,\mathbf{i}_U}(\mathbf{x}_U, \mathbf{t}_U) \neq 0$ for some $\mathbf{x}$ and $\mathbf{i}$, then $t_k \in [x_{i_k,1}^k, x_{i_k,2^{i_k}-1}^k]$ for every $k \in U$.*

As in the proof of Theorem 2, let $\mathcal{E}_{q,s}(f) = \sum_{|\mathbf{i}| \ge q+1} \bigotimes_{k=1}^s \Delta_{k,i_k}(f)$. Recall that $\Delta_{\mathbf{i}}(M_U(\cdot, \mathbf{t}_U)) \equiv 0$ if either $i_k = 1$ for some $k \in U$, or $i_k \ge 2$ for some $k \notin U$. Hence

$$\mathcal{E}_{q,s}(M_U(\cdot, \mathbf{t}_U))(\mathbf{x}_U) = \sum_{\mathbf{i}_U \in P(q,s,U)} \delta_{U,\mathbf{i}_U}(\mathbf{x}_U, \mathbf{t}_U)$$

with

$$P(q,s,U) = \left\{ \mathbf{j} \in \mathbb{N}_+^{|U|} \,:\, \mathbf{j} \ge \mathbf{2}, |\mathbf{j}| \ge q + 1 - s + |U| \right\}.$$

From Theorem 4, this yields

$$\mathrm{error}(\mathcal{A}_{q,s}, \mathcal{F}) = \left(\sum_{U \neq \emptyset} \gamma_{s,U}^{p^*} \int_{D_U} b_U^{p^*}(\mathbf{t}_U)\, d\mathbf{t}_U \right)^{1/p^*},$$

where

$$b_U(\mathbf{t}_U) := \left| \int_{D_U} \rho_U(\mathbf{x}_U) \sum_{\mathbf{i}_U \in P(q,s,U)} \delta_{U,\mathbf{i}_U}(\mathbf{x}_U, \mathbf{t}_U)\, d\mathbf{x}_U \right|. \tag{22}$$

Due to its definition (20), $\psi_k(t)$ converges to zero with $t \to b_k$ and/or $t \to a_k$. Moreover, from (iv) of Proposition 2, we know that

$$\delta_{U,\mathbf{i}_U}(\mathbf{x}_U, \mathbf{t}_U) \neq 0 \quad \text{implies } \psi_k(t_k) \geq 2^{-i_k} \text{ for every } k \in U.$$

This means that we can replace the set $P(q, s, U)$ by the even smaller set

$$P(q, s, U, \mathbf{t}_U) = \left\{ \mathbf{j} \in \mathbb{N}_+^{|U|} \,:\, \mathbf{j} \geq \mathbf{2},\ |\mathbf{j}| \geq q + 1 - s + |U|, \text{ and } 2^{-j_k} \leq \psi_k(t_k),\ \forall\, k \in U \right\}.$$

This leads to

$$\begin{aligned} b_U(\mathbf{t}_U) &\leq \sum_{\mathbf{j} \in P(q,s,U,\mathbf{t}_U)} 2^{-\mathbf{j}} \\ &\leq \min \left\{ 2^{|U|}\, \psi_U(\mathbf{t}_U)\, ,\, 2^{-|U|}\ \mathrm{B}(\max\{q - s, |U| - 1\}, |U|) \right\}, \end{aligned}$$

where, as always, $\psi_U(\mathbf{t}_U) = \prod_{k \in U} \psi_k(t_k)$. We summarize this in the following proposition.

PROPOSITION 3. *Let* (4) *hold. Then for any $s \geq 2$ and any $q \geq s$,*

$$\mathrm{error}(\mathcal{A}_{q,s,\gamma}; \mathcal{F}_{p,s,\gamma}) = \left(\sum_{U \neq \emptyset} \gamma_{s,U}^{p^*} \int_{D_U} b_U^{p^*}(\mathbf{t}_U)\, d\mathbf{t}_U \right)^{1/p^*}.$$

Moreover,

$$b_U(\mathbf{t}_U) \leq \min \left\{ 2^{|U|}\, \psi_U(\mathbf{t}_U)\, ,\, \frac{\mathrm{B}(\max\{q - s, |U| - 1\}, |U|)}{2^{|U|}} \right\}.$$

For $p > 1$ and unbounded D, the above error bound is quite complicated, due to the presence of integrals of $b_U^{p^*}$. Suppose now that D is bounded, say

$$D = [0, 1]^s.$$

Then the integrals of b_U^* can be replaced by $\mathrm{B}(\max\{q - s, |U| - 1\}, |U|)\, 2^{-|U|}$ leading to the following upper bound:

$$\mathrm{error}(\mathcal{A}_{q,s}; \mathcal{F}_{p,s,\gamma}) \leq \left(\sum_{U \neq \emptyset} \left(2^{-|U|}\, \gamma_{s,U}\ \mathrm{B}(\max\{q - s, |U| - 1\}, |U|) \right)^{p^*} \right)^{1/p^*}. \tag{23}$$

Of course, if $p = 1$, then $p^* = \infty$ and the bound (23) takes the form

$$\mathrm{error}(\mathcal{A}_{q,s}; \mathcal{F}_{1,s}) \le \max_{U \neq \emptyset} 2^{-|U|}\, \gamma_{s,U}\ \mathrm{B}(\max\{q-s, |U|-1\}, |U|),$$

which coincides with the bound from section 4 for the case when (3) holds. For $p > 1$, (23) can be further estimated from above, leading to

$$\mathrm{error}(\mathcal{A}_{q,s}; \mathcal{F}_{p,s,\gamma}) \le 2^{s/p^*} \max_{U \neq \emptyset} 2^{-|U|}\, \gamma_{s,U}\ \mathrm{B}(\max\{q-s, |U|-1\}, |U|).$$

This means that error upper bounds from section 4 also hold for $p > 1$ modulo the multiplicative factor $2^{s/p^*}$. In particular, we have the following consequence of Theorem 3.

THEOREM 5. *Let $p > 1$, $\gamma_{s,U} \equiv 1$, and $D = [0,1]^s$. Then for every $s \ge 2$ and $q \ge 2(s-1)$,*

$$\mathrm{error}(\mathcal{A}_{q,s}; \mathcal{F}_{p,s}) \le \frac{s \exp\left(\frac{1}{12(s-1)}\right), 2^{1/p^*}}{(s-1)\pi} \left(\frac{e \ln(\mathrm{card}(\mathcal{A}_{q,s}))}{(s-1)\, 2^{1/(2p^*)} \ln(2)}\right)^{2(s-1)} \frac{1}{\mathrm{card}(\mathcal{A}_{q,s})}.$$

6. Integration problem.

In this section, we briefly discuss the problem of approximating weighted integrals

$$I_\rho(f) = \int_D f(\mathbf{x})\, \rho(\mathbf{x})\, d\mathbf{x}$$

by algorithms (often called quadratures) $\mathcal{Q}$ of the form $\mathcal{Q}(f) = \sum_{i=1}^n f(\mathbf{x}^i)\, g_i$. The *worst case error* of $\mathcal{Q}$ (with respect to $\mathcal{F}_{p,s}$) is defined by

$$\mathrm{error}(\mathcal{Q}; \mathcal{F}_{p,s}, \mathrm{Int}) := \sup_{\|f\|_{p,s} \le 1} |I_\rho(f) - \mathcal{Q}(f)|.$$

Consider now the following quadrature

$$\mathcal{Q}_{q,s}(f) := I_\rho(\mathcal{A}_{q,s}(f)). \tag{24}$$

It is easy to see that

$$\mathcal{Q}_{q,s} = \sum_{|\mathbf{i}| \le q} \bigotimes_{k=1}^{s} \left(Q_{k,i_k} - Q_{k,i_k-1}\right), \quad \text{where} \quad Q_{k,i}(f) = 2^{-i} \left(f(c_k) + \sum_{j=1}^{2^i-1} f(x_{i,j}^k)\right)$$

$x_{i,j}^k$ defined by (7). Clearly,

$$\mathrm{error}(\mathcal{Q}_{q,s}; \mathcal{F}_{p,s}, \mathrm{Int}) \le \mathrm{error}(\mathcal{A}_{q,s}; \mathcal{F}_{p,s}).$$

Hence all error bounds for $\mathcal{A}_{q,s}$ obtained in the previous sections also hold for $\mathcal{Q}_{q,s}$.

REFERENCES

[1] K.I. BABENKO, *Approximation by trigonometric polynomials in a certain class of periodic functions of several variables*, Soviet. Math. Dokl., 1 (1960), pp. 672–675.

[2] G. BASZENSKI, F.J. DELVOS, AND S. JESTER, *Blending approximations with sine functions*, in Numerical Methods in Approximation Theory, Vol. 9, D. Braess and L.L. Schumaker, eds., Internat. Ser. Numer. Math. 105, Birkhäuser, Basel, 1992, pp. 1–19.

[3] F.J. Delvos, *d-variate Boolean approximation*, J. Approx. Theory, 34 (1982), pp. 99–114.

[4] R.A. DeVore, S.V. Konyagin, and V.N. Temlyakov, *Hyperbolic wavelet approximation*, Constr. Approx., 14 (1998), pp. 1–26.

[5] M. Drmota and R.F. Tichy, *Sequences, Discrepancies and Applications*, Lecture Notes in Math. 1651, Springer-Verlag, Berlin, 1997.

[6] T. Gerstner and M. Griebel, *Dimension-adaptive tensor-product quadrature,* Computing, 71 (2003), pp. 65–87.

[7] M. Griebel and W. Knapek, *Optimized tensor-product approximation spaces*, Constr. Approx., 16 (2000), pp. 252–540.

[8] M. Griebel, P. Oswald, and T. Schiekofer, *Sparse grids for boundary integral equations*, Numer. Math., 83 (1999), pp. 279–312.

[9] F.J. Hickernell, I.H. Sloan, and G.W. Wasilkowski, *On tractability of weighted integration over bounded and unbounded regions in* $\mathbb{R}^s$, Math. Comp., 73 (2004), pp. 1885–1901.

[10] Y. Li, *Applicability of Smolyak's algorithm to certain Banach spaces of functions,* J. Complexity, 18 (2002), pp. 792–814.

[11] Y. Li and G.W. Wasilkowski, *Worst case complexity of weighted approximation and integration over* $\mathbb{R}^d$*,* J. Complexity, 18 (2002), pp. 330–345.

[12] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 63, SIAM, Philadelphia, 1992.

[13] E. Novak, *Deterministic and Stochastic Error Bounds in Numerical Analysis*, Lecture Notes in Math. 1349, Springer-Verlag, Berlin, 1988.

[14] E. Novak and K. Ritter, *High dimensional integration of smooth functions over cubes*, Numer. Math., 75 (1996), pp. 79–97.

[15] E. Novak and H. Woźniakowski, *When are integration and discrepancy tractable?*, in Foundation of Computational Mathematics, Oxford, 1999, R.A. DeVore, A. Iserles, and E. Süli, eds., Cambridge University Press, Cambridge, UK, 2001, pp. 211–266.

[16] K. Petras, *On the Smolyak cubature error for analytical functions*, Adv. Comput. Math., 12 (2000), 71–93.

[17] A. Pinkus, *n-Widths in Approximation Theory,* Springer-Verlag, Berlin, 1985.

[18] L. Plaskota and G.W. Wasilkowski, *The exact exponent of sparse grid quadratures in the weighted case*, J. Complexity, 17 (2001), pp. 840–849.

[19] I.H. Sloan and S. Joe, *Lattice Methods for Multiple Integration,* Oxford University Press, Oxford, 1994.

[20] I.H. Sloan and H. Woźniakowski, *When are quasi-Monte Carlo algorithms efficient for high dimensional integrals,* J. Complexity, 14 (1998), pp. 1–33.

[21] S.A. Smolyak, *Quadrature and interpolation formulas for tensor products of certain classes of functions,* Dokl. Akad. Nauk SSSR, 4 (1963), pp. 240–243.

[22] V.N. Temlyakov, *Approximation of Periodic Functions,* Nova Science, New York, 1993.

[23] V.N. Temlyakov, *On approximate recovery of functions with bounded mixed derivative*, J. Complexity, 9 (1993), pp. 41–59.

[24] J.F. Traub, G.W. Wasilkowski, and H. Woźniakowski, *Information-Base Complexity*, Academic Press, New York, 1988.

[25] G.W. Wasilkowski and H. Woźniakowski, *Explicit cost bounds of algorithms for multivariate tensor product problems*, J. Complexity, 11 (1995), pp. 1–56.

[26] G.W. Wasilkowski and H. Woźniakowski, *Weighted tensor-product algorithms for linear multivariate problems*, J. Complexity, 15 (1999), pp. 402–447.

[27] G.W. Wasilkowski and H. Woźniakowski, *Complexity of weighted approximation over* $\mathbb{R}^1$, J. Approx. Theory, 103 (2000), pp. 223–251.

SIAM J. Numer. Anal.
Vol. 43, No. 3, pp. 1021–1042

A POSTERIORI ERROR ESTIMATES FOR THE MORTAR MIXED FINITE ELEMENT METHOD[*]

MARY F. WHEELER[†] AND IVAN YOTOV[‡]

Abstract. Several a posteriori error estimators for mortar mixed finite element discretizations of elliptic equations are derived. A residual-based estimator provides optimal upper and lower bounds for the pressure error. An efficient and reliable estimator for the velocity and mortar pressure error is also derived, which is based on solving local (element) problems in a higher-order space. The interface flux-jump term that appears in the estimators can be used as an indicator for driving an adaptive process for the mortar grids only.

Key words. mixed finite element, mortar finite element, nonmatching grids, a posteriori error estimates, adaptive mesh refinement

AMS subject classifications. 65N12, 65N15, 65N30, 65N50

DOI. 10.1137/S0036142903431687

1. Introduction.

We consider the second order elliptic problem written as a system of two first order equations

$$\mathbf{u} = -K\nabla p \quad \text{in } \Omega, \tag{1.1}$$

$$\nabla \cdot \mathbf{u} = f \quad \text{in } \Omega, \tag{1.2}$$

$$p = g \quad \text{on } \Gamma_D, \tag{1.3}$$

$$\mathbf{u} \cdot \nu = 0 \quad \text{on } \Gamma_N, \tag{1.4}$$

where $\Omega \subset \mathbf{R}^d$, $d = 2$ or 3, is a multiblock domain with a boundary $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, measure $(\Gamma_D) > 0$, ν is the outward unit normal on $\partial\Omega$, and K is a symmetric, uniformly positive definite tensor satisfying, for some $0 < k_0 \leq k_1 < \infty$,

$$k_0\xi^T\xi \leq \xi^T K(x)\xi \leq k_1\xi^T\xi \quad \forall x \in \Omega \quad \forall \xi \in \mathbf{R}^d. \tag{1.5}$$

In flow in porous media the above system models single-phase flow where p is the pressure, $\mathbf{u}$ is the Darcy velocity, and K represents the permeability divided by the viscosity.

A number of papers in recent years have studied the numerical solution of the above and related problems on multiblock domains with nonmatching grids across the interfaces. This growing interest is driven by the flexibility provided by the multiblock paradigm. Complicated geometries can be modeled as unions of relatively simple subdomains with locally constructed grids. Local features of the solution such as

[*] Received by the editors July 17, 2003; accepted for publication (in revised form) February 25, 2005; published electronically September 20, 2005.

http://www.siam.org/journals/sinum/43-3/43168.html

[†] Center for Subsurface Modeling, Institute for Computational Engineering and Sciences (ICES), Department of Aerospace Engineering & Engineering Mechanics, and Department of Petroleum and Geosystems Engineering, The University of Texas at Austin, Austin, TX 78712 (mfw@ices.utexas.edu). The research of this author was partially supported by NSF grants EIA 0121523 and DMS 0411413 and DOE grant DE-FGO2-04ER25617.

[‡] Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (yotov@math.pitt.edu). The research of this author was partially supported by NSF grants DMS 0107389 and DMS 0411694, DOE grant DE-FG02-04ER25618, and the J. Tinsley Oden Faculty Fellowship, ICES, The University of Texas at Austin.

corner singularities or large gradients can be resolved by finer grids in the local region. Large scale features such as geological faults and layers in subsurface flow can be modeled with nonmatching grids. Moreover, the resulting algebraic problem can be efficiently solved via parallel domain decomposition algorithms.

In a multiblock formulation, the equations are imposed locally on each subdomain and appropriate interface matching conditions are enforced on the interfaces. The use of mortar finite elements to impose the interface conditions is a popular approach due to its excellent stability and accuracy. For the use of mortars, the reader is referred to [10, 8, 20] and references therein for Galerkin finite element and finite volume methods, and to [40, 3, 9] in the context of mixed finite element methods.

An integral part of any successful computational method is the development of a posteriori error estimators and adaptive mesh refinement strategies. Although there is an enormous amount of literature on a posteriori error estimation and adaptivity on conforming grids (seminal works include [5, 7, 2, 35]), few papers deal with this issue for mortar finite element methods. In the case of Galerkin finite elements, error estimators have been developed in [37, 38, 30]. Even fewer results are available for the mortar mixed finite element method. Goal-oriented estimates and adaptivity are developed in [6]. Computational results from [36, 29] indicate that a judicious choice of mortar grids can lead to an accurate solution at low computational cost, but no rigorous justification is given. The goal of this paper is to develop a posteriori error estimators and an adaptive mesh refinement strategy for the mortar mixed finite element method.

Previous works on error estimation for mixed finite element methods on conforming grids include [11, 12, 17, 24, 39, 25]. In [11], mesh-dependent norms are utilized to obtain optimal residual-based error estimators. Estimators based on superconvergence error estimates are developed in [12, 24]. In [17, 39], the Helmholtz decomposition is used to derive optimal residual-based error estimators in the natural pressure and velocity norms. Hierarchical estimates and implicit estimates based on solving local problems are also investigated in [39]. Only the three-dimensional results are given in [25], where a duality argument is employed to obtain residual-based estimates. However, the velocity bounds derived there depend on a saturation assumption that may not hold in general.

In this paper we derive a posteriori error estimates that provide lower and upper bounds for the pressure, velocity, and mortar error in two and three dimensions. According to the widely accepted terminology, an estimator is referred to as *reliable* if it provides an upper bound of the error, whereas it is called *efficient* if it gives a lower bound. We employ a duality-type argument to obtain an efficient and reliable residual-based estimator for the pressure error. In addition to the usual element residual terms, the estimator involves a flux-jump term and a mortar pressure difference term on subdomain interfaces. A closely related estimator of the velocity and mortar error is also derived, which provides an optimal upper bound, but suboptimal (yet sharp) lower bound. We then proceed to derive an optimal efficient and reliable implicit estimator for the velocity based on solving local (element) problems in a higher-order space. Throughout the paper we make several reasonable saturation assumptions which are motivated by known a priori error estimates.

It was observed in [36, 29] that varying the mortar degrees of freedom while keeping the subdomain grids fixed has a substantial effect on the convergence of the interface algorithm employed to solve the algebraic system. At the same time mortar grids that are too coarse lead to deterioration of the accuracy of the method. Therefore, finding "optimal" mortar grids for given subdomain grids is an important question. The flux-jump term that appears in all estimators provides a stand-alone indicator of the nonconformity error in the mortar discretization. It can be used to drive an adaptive mesh refinement process for the mortar grids.

The rest of the paper is organized as follows. In the next section the mortar mixed finite element method is defined along with its equivalent interface formulation. In section 3, the residual-based error estimators are derived and analyzed. The implicit estimator for the velocity is developed in section 4. Computational results are presented in section 5, followed by some remarks and conclusions in section 6.

2. Formulation of the method and preliminaries. We will make use of the following standard notation. For a subdomain $G \subset \mathbf{R}^d$, the $L^2(G)$ inner product (or duality pairing) and norm are denoted by $(\cdot,\cdot)_G$ and $\|\cdot\|_G$, respectively, for scalar and vector valued functions. The Sobolev spaces $W_p^k(G)$, $k \in \mathbf{R}$, $1 \le p \le \infty$, are defined in the usual way [1] with the usual norm $\|\cdot\|_{k,p,G}$. Let $\|\cdot\|_{k,G}$ be the norm of the Hilbert space $H^k(G) = W_2^k(G)$. We omit G in the subscript if $G = \Omega$. For a section of a subdomain boundary $S \subset \cup_{i=1}^n \partial\Omega_i$ we write $\langle\cdot,\cdot\rangle_S$ and $\|\cdot\|_S$ for the $L^2(S)$ inner product (or duality pairing) and norm, respectively.

We assume that problem (1.1)–(1.4) is H^2-regular, i.e., there exists a positive constant C depending only on K and Ω such that

$$\|p\|_2 \le C(\|f\| + \|g\|_{3/2,\Gamma_D}). \tag{2.1}$$

We refer the reader to [23, 26, 21] for sufficient conditions for H^2-regularity.

To give the weak formulation of (1.1)–(1.4) we recall the usual velocity space [16]

$$H(\mathrm{div};\Omega) = \{\mathbf{v} \in (L^2(\Omega))^d : \nabla\cdot\mathbf{v} \in L^2(\Omega)\}$$

with a norm

$$\|\mathbf{v}\|_{H(\mathrm{div})} = (\|\mathbf{v}\|^2 + \|\nabla\cdot\mathbf{v}\|^2)^{1/2}$$

and define

$$\bar{\mathbf{V}} = \{\mathbf{v} \in H(\mathrm{div};\Omega) : \mathbf{v}\cdot\nu = 0 \text{ on } \Gamma_N\}.$$

A weak solution of (1.1)–(1.4) is $\mathbf{u} \in \bar{\mathbf{V}}$, $p \in L^2(\Omega)$ such that

$$(K^{-1}\mathbf{u},\mathbf{v}) = (p,\nabla\cdot\mathbf{v}) - \langle g,\mathbf{v}\cdot\nu\rangle_{\Gamma_D}, \quad \mathbf{v} \in \bar{\mathbf{V}}, \tag{2.2}$$

$$(\nabla\cdot\mathbf{u},w) = (f,w), \quad w \in L^2(\Omega). \tag{2.3}$$

It is well known (see, e.g., [16, 32]) that (2.2) and (2.3) have a unique solution.

Let $\Omega = \cup_{i=1}^n \Omega_i$ be a union of nonoverlapping subdomains. Let

$$\Gamma_{i,j} = \partial\Omega_i \cap \partial\Omega_j, \quad \Gamma = \cup_{i,j=1}^n \Gamma_{i,j}, \quad \Gamma_i = \partial\Omega_i \cap \Gamma = \partial\Omega_i \backslash \partial\Omega.$$

Let

$$\mathbf{V}_i = \{\mathbf{v} \in H(\mathrm{div};\Omega_i) : \mathbf{v}\cdot\nu_i \in L^2(\partial\Omega_i) \text{ and } \mathbf{v}\cdot\nu_i = 0 \text{ on } \partial\Omega_i \cap \Gamma_N\}, \quad \mathbf{V} = \bigoplus_{i=1}^n \mathbf{V}_i,$$

$$W_i = L^2(\Omega_i), \quad W = \bigoplus_{i=1}^n W_i = L^2(\Omega), \quad M = L^2(\Gamma).$$

It is easy to see that if the solution of (2.2) and (2.3) satisfies $\mathbf{u}\cdot\nu|_\Gamma \in L^2(\Gamma)$ and $p \in H^1(\Omega)$, then for $1 \le i \le n$

$$(K^{-1}\mathbf{u},\mathbf{v})_{\Omega_i} = (p,\nabla\cdot\mathbf{v})_{\Omega_i} - \langle \lambda, \mathbf{v}\cdot\nu_i\rangle_{\Gamma_i} - \langle g, \mathbf{v}\cdot\nu\rangle_{\partial\Omega_i\cap\Gamma_D}, \quad \mathbf{v}\in\mathbf{V}_i, \tag{2.4}$$

$$(\nabla\cdot\mathbf{u}, w)_{\Omega_i} = (f,w)_{\Omega_i}, \quad w \in W_i, \tag{2.5}$$

$$\sum_{i=1}^n \langle \mathbf{u}\cdot\nu_i, \mu\rangle_{\Gamma_i} = 0, \quad \mu \in M, \tag{2.6}$$

where $\lambda = p|_\Gamma$. (2.4)–(2.6) imply that $(\mathbf{u}, p, \lambda) \in \mathbf{V}\times W\times M$ satisfy

$$A(\mathbf{u},p,\lambda;\mathbf{v},w,\mu) = L(\mathbf{v},w,\mu) \quad \forall\ (\mathbf{v},w,\mu)\in\mathbf{V}\times W\times M, \tag{2.7}$$

where

$$\begin{aligned} &A(\mathbf{u},p,\lambda;\mathbf{v},w,\mu) \\ &\quad = \sum_{i=1}^n \big((K^{-1}\mathbf{u},\mathbf{v})_{\Omega_i} - (p,\nabla\cdot\mathbf{v})_{\Omega_i} + \langle\lambda,\mathbf{v}\cdot\nu_i\rangle_{\Gamma_i} + \sigma(\nabla\cdot\mathbf{u},w)_{\Omega_i} - \sigma\langle\mathbf{u}\cdot\nu_i,\mu\rangle_{\Gamma_i}\big)\end{aligned}$$

and

$$L(\mathbf{v},w,\mu) = \sigma(f,w) - \langle g, \mathbf{v}\cdot\nu\rangle_{\Gamma_D}.$$

Here $\sigma = 1$ or $\sigma = -1$. If $\sigma = -1$, $A(\cdot;\cdot)$ is a symmetric bilinear form, which we denote by $A^s(\cdot;\cdot)$. If $\sigma = 1$, we denote $A(\cdot;\cdot)$ by $A^c(\cdot;\cdot)$ and note that

$$A^c(\mathbf{v},w,\mu;\mathbf{v},w,\mu) = (K^{-1}\mathbf{v},\mathbf{v});$$

thus $A^c(\cdot;\cdot)$ is nonsymmetric, but coercive. Note that the solution does not depend on the choice of σ.

Let $\{\mathcal{T}_{h,i}\}_h$ be a family of finite element partitions of Ω_i, $1\le i\le n$. Let, for any $E\in\mathcal{T}_{h,i}$, $h_E = \operatorname{diam}(E)$ and let

$$h_i = \max_{E\in\mathcal{T}_{h,i}} h_E, \quad h = \max_{1\le i\le n} h_i.$$

Define ρ_E to be the largest diameter of a ball contained in $\overline{E}$. We require that each subdomain grid satisfies the nondegeneracy condition

$$\max_{E\in\mathcal{T}_{h,i}} \frac{h_E}{\rho_E} \le c_0,$$

where the constant c_0 is independent of h_i. The partitions $\mathcal{T}_{h,i}$ and $\mathcal{T}_{h,j}$ may be nonmatching along $\Gamma_{i,j}$. Let $\mathcal{T}_h = \cup_{i=1}^n \mathcal{T}_{h,i}$ and let $\mathcal{E}_h$ be the union of all interior edges (faces) not including the interfaces and the outer boundary. Let

$$\mathbf{V}_{h,i}\times W_{h,i} \subset \mathbf{V}_i\times W_i$$

be any of the usual mixed finite element spaces (i.e., the RTN spaces [34, 31, 27], BDM spaces [15], BDFM spaces [14], BDDF spaces [13], or CD spaces [18]). The order of the spaces is assumed to be the same on every subdomain. Let

$$\mathbf{V}_h = \bigoplus_{i=1}^n \mathbf{V}_{h,i}, \quad W_h = \bigoplus_{i=1}^n W_{h,i}.$$

Note that this choice leads to a nonconforming approximation since the normal components of vectors in $\mathbf{V}_h$ do not have to be continuous across Γ. Throughout the paper we will abuse notation when using the $H(\text{div})$-norm. In particular, for $\mathbf{v} \in H(\text{div}; \Omega_i)$, $i = 1, \ldots, n$,

$$\|\mathbf{v}\|_{H(\text{div})} = \left(\|\mathbf{v}\|^2 + \sum_{i=1}^{n} \|\nabla \cdot \mathbf{v}\|_{\Omega_i}^2 \right)^{1/2},$$

and for $\mathbf{v} \in H(\text{div}; E)$, $E \in \mathcal{T}_h$,

$$\|\mathbf{v}\|_{H(\text{div})} = \left(\|\mathbf{v}\|^2 + \sum_{E \in \mathcal{T}_h} \|\nabla \cdot \mathbf{v}\|_{E}^2 \right)^{1/2}.$$

For all of the above spaces

$$\nabla \cdot \mathbf{V}_{h,i} = W_{h,i}$$

and there exists a projection operator $\Pi_{h,i}$ of $(H^1(\Omega_i))^d$ onto $\mathbf{V}_{h,i}$ satisfying for any $\mathbf{q} \in (H^1(\Omega_i))^d$

$$(\nabla \cdot (\Pi_{h,i}\mathbf{q} - \mathbf{q}), w)_{\Omega_i} = 0, \quad w \in W_{h,i}, \tag{2.8}$$

$$\langle (\mathbf{q} - \Pi_{h,i}\mathbf{q}) \cdot \nu_i, \mathbf{v} \cdot \nu_i \rangle_{\partial\Omega_i} = 0, \quad \mathbf{v} \in \mathbf{V}_{h,i}. \tag{2.9}$$

Let $\Pi_h : \bigoplus (H^1(\Omega_i))^d \to \mathbf{V}_h$ be such that $\Pi_h \mathbf{q}|_{\Omega_i} = \Pi_{h,i}\mathbf{q}$ for all $\mathbf{q} \in \bigoplus (H^1(\Omega_i))^d$.

Let the mortar interface mesh $\mathcal{T}_{h,i,j}$ be a quasi-uniform finite element partition of $\Gamma_{i,j}$ and let $\mathcal{T}^{\Gamma,h} = \cup_{1 \le i < j \le n} \mathcal{T}_{h,i,j}$. For any $\tau \in \mathcal{T}_{h,i,j}$, let

$$E_\tau = \cup (E \in \mathcal{T}_h : \partial E \cap \tau \neq \emptyset).$$

We will assume that there exist constants c_1 and c_2 such that

$$c_1 h_E \le h_\tau \le c_2 h_E \quad \forall E \in E_\tau, \tag{2.10}$$

where the notation $h_S = \text{diam}(S)$ is used. Denote by $M_{h,i,j} \subset L^2(\Gamma_{i,j})$ the mortar space on $\Gamma_{i,j}$ containing at least either the continuous or discontinuous piecewise polynomials of degree $k+1$ on $\mathcal{T}_{h,i,j}$, where k is associated with the degree of the polynomials in $\mathbf{V}_h \cdot \nu$. More precisely, if $d = 3$ and e is a triangle of the mesh, we take $M_{h,i,j}|_e = P_{k+1}(e)$, the set of polynomials of degree less than or equal to $k+1$ on e. If e is a rectangle, we take $M_{h,i,j}|_e = Q_{k+1}(e)$, the set of polynomials on e for which the degree in each variable separately is less than or equal to $k+1$. Now let

$$M_h = \bigoplus_{1 \le i < j \le n} M_{h,i,j}$$

be the mortar finite element space on Γ.

In the mortar mixed finite element method for approximating (2.4)–(2.6) we seek $\mathbf{u}_h \in \mathbf{V}_h$, $p_h \in W_h$, and $\lambda_h \in M_h$ such that, for $1 \le i \le n$,

$$(K^{-1}\mathbf{u}_h, \mathbf{v})_{\Omega_i} = (p_h, \nabla \cdot \mathbf{v})_{\Omega_i} - \langle \lambda_h, \mathbf{v} \cdot \nu_i \rangle_{\Gamma_i} - \langle g, \mathbf{v} \cdot \nu_i \rangle_{\partial\Omega_i \cap \Gamma_D}, \quad \mathbf{v} \in \mathbf{V}_{h,i}, \tag{2.11}$$

$$(\nabla \cdot \mathbf{u}_h, w)_{\Omega_i} = (f, w)_{\Omega_i}, \quad w \in W_{h,i}, \tag{2.12}$$

$$\sum_{i=1}^{n} \langle \mathbf{u}_h \cdot \nu_i, \mu \rangle_{\Gamma_i} = 0, \quad \mu \in M_h. \tag{2.13}$$

(2.13) enforces weak (with respect to the mortar space M_h) continuity of the flux across the block interfaces. Existence and uniqueness of a solution of (2.11)–(2.13) are shown in [40, 3] along with optimal convergence and superconvergence for both pressure and velocity under the assumption that for all $\mu \in M_{h,i,j}$ there exists a constant C independent of h such that

$$\|\mu\|_{\Gamma_{i,j}} \le C(\|\mathcal{Q}_{h,i}\mu\|_{\Gamma_{i,j}} + \|\mathcal{Q}_{h,j}\mu\|_{\Gamma_{i,j}}), \tag{2.14}$$

where $\mathcal{Q}_{h,i} : L^2(\partial\Omega_i) \to \mathbf{V}_{h,i}\cdot\nu_i|_{\partial\Omega_i}$ is the L^2-orthogonal projection satisfying for any $\phi \in L^2(\partial\Omega_i)$

$$\langle \phi - \mathcal{Q}_{h,i}\phi, \mathbf{v}\cdot\nu_i\rangle_{\Gamma_i} = 0 \quad \forall \mathbf{v} \in \mathbf{V}_{h,i}. \tag{2.15}$$

Remark 2.1. Condition (2.14) imposes a limit on the number of mortar degrees of freedom and is easily satisfied in practice [40, 28].

We recall some a priori error estimates from [3] which will later motivate some of the saturation assumptions needed in the a posteriori error analysis. Herein l is associated with the degree of the polynomials in W_h and $\|\cdot\|_{d_h}$ is a mortar space norm defined in the next subsection. Throughout the paper C denotes a generic constant independent of h.

THEOREM 2.1. *For the solution of* (2.11)–(2.13) *if* (2.14) *holds, then*

$$\|\nabla\cdot(\mathbf{u}-\mathbf{u}_h)\| \le C\sum_{i=1}^n \|\nabla\cdot\mathbf{u}\|_{r,\Omega_i}h^r, \quad 1 \le r \le l+1,$$

$$\|\mathbf{u}-\mathbf{u}_h\| \le C\sum_{i=1}^n (\|p\|_{r+1,\Omega_i} + \|\mathbf{u}\|_{r,\Omega_i})h^r, \quad 1 \le r \le k+1,$$

$$\|p-p_h\| \le C\sum_{i=1}^n (\|p\|_{r+1,\Omega_i} + \|\mathbf{u}\|_{r,\Omega_i} + \|\nabla\cdot\mathbf{u}\|_{r,\Omega_i})h^r, \quad 1 \le r \le \min(k+1, l+1),$$

$$\|\lambda-\lambda_h\|_{d_h} \le C\sum_{i=1}^n (\|p\|_{r+1,\Omega_i} + \|\mathbf{u}\|_{r,\Omega_i})h^r, \quad 1 \le r \le k+1.$$

2.1. Interface formulation.

Method (2.11)–(2.13) can be reduced to an equivalent interface (mortar) problem. We recall this interface formulation from [22, 40, 3], as it will be used in estimating the mortar error.

Define $d_h : L^2(\Gamma)\times L^2(\Gamma) \to \mathbf{R}$ for $\varphi, \mu \in L^2(\Gamma)$ by

$$d_h(\varphi,\mu) = \sum_{i=1}^n d_{h,i}(\varphi,\mu) = -\sum_{i=1}^n \langle \mathbf{u}_h^*(\varphi)\cdot\nu_i, \mu\rangle_{\Gamma_i}, \tag{2.16}$$

where $(\mathbf{u}_h^*(\varphi), p_h^*(\varphi)) \in \mathbf{V}_h \times W_h$ solve, for $1 \le i \le n$,

$$(K^{-1}\mathbf{u}_h^*(\varphi), \mathbf{v})_{\Omega_i} = (p_h^*(\varphi), \nabla\cdot\mathbf{v})_{\Omega_i} - \langle \varphi, \mathbf{v}\cdot\nu_i\rangle_{\Gamma_i}, \quad \mathbf{v} \in \mathbf{V}_{h,i}, \tag{2.17}$$

$$(\nabla\cdot\mathbf{u}_h^*(\varphi), w)_{\Omega_i} = 0, \quad w \in W_{h,i}. \tag{2.18}$$

Define $g_h : L^2(\Gamma) \to \mathbf{R}$ by

$$g_h(\mu) = \sum_{i=1}^n g_{h,i}(\mu) = \sum_{i=1}^n \langle \bar{\mathbf{u}}_h\cdot\nu_i, \mu\rangle_{\Gamma_i},$$

where $(\bar{\mathbf{u}}_h, \bar{p}_h) \in \mathbf{V}_h \times W_h$ solve, for $1 \le i \le n$,

$$\begin{aligned}(K^{-1}\bar{\mathbf{u}}_h, \mathbf{v})_{\Omega_i} &= (\bar{p}_h, \nabla\cdot\mathbf{v})_{\Omega_i} - \langle g, \mathbf{v}\cdot\nu_i\rangle_{\partial\Omega_i\cap\Gamma_D}, \quad \mathbf{v}\in\mathbf{V}_{h,i},\\ (\nabla\cdot\bar{\mathbf{u}}_h, w)_{\Omega_i} &= (f, w)_{\Omega_i}, \quad w \in W_{h,i}.\end{aligned}$$

Then $(\mathbf{u}_h, p_h, \lambda_h)$ satisfies

$$d_h(\lambda_h, \mu) = g_h(\mu) \quad \forall \mu \in M_h, \quad \mathbf{u}_h = \mathbf{u}_h^*(\lambda_h) + \bar{\mathbf{u}}_h, \quad p_h = p_h^*(\lambda_h) + \bar{p}_h.$$

It is easy to see from (2.16) and (2.17) that

$$d_{h,i}(\varphi,\varphi) = (K^{-1}\mathbf{u}_h^*(\varphi), \mathbf{u}_h^*(\varphi))_{\Omega_i}, \tag{2.19}$$

which implies that $d_h(\cdot,\cdot)$ is positive semidefinite in $M \times M$ and, assuming (2.14), positive definite in $M_h \times M_h$. We define the norm in M_h:

$$\|\mu\|_{d_h} := d_h(\mu,\mu)^{1/2}.$$

It is shown in [40, 28] for RT_0 rectangular elements and very general hanging interface nodes and mortar grid configurations satisfying (2.14) that

$$\sum_{\tau\in\mathcal{T}^{\Gamma,h}} \|\mu\|_{1/2,\tau}^2 \le C d_h(\mu,\mu) \quad \forall \mu \in M_h. \tag{2.20}$$

The proofs in [40, 28] can be generalized in a relatively straightforward way to the other mixed finite element spaces under consideration and to higher-order elements.

The following construction will also be useful in the analysis of the mortar error. Define, for $\varphi \in L^2(\Gamma)$,

$$\mathbf{u}_h(\varphi) = \mathbf{u}_h^*(\varphi) + \bar{\mathbf{u}}_h, \quad p_h(\varphi) = p_h^*(\varphi) + \bar{p}_h.$$

We note that $(\mathbf{u}_h(\varphi), p_h(\varphi)) \in \mathbf{V}_h \times W_h$ satisfy, for $1 \le i \le n$,

$$\begin{aligned}(K^{-1}\mathbf{u}_h(\varphi), \mathbf{v})_{\Omega_i} &= (p_h(\varphi), \nabla\cdot\mathbf{v})_{\Omega_i} - \langle \varphi, \mathbf{v}\cdot\nu\rangle_{\Gamma_i}\\ &\quad - \langle g, \mathbf{v}\cdot\nu\rangle_{\partial\Omega_i\cap\Gamma_D}, \quad \mathbf{v}\in\mathbf{V}_{h,i}, \qquad (2.21)\\ (\nabla\cdot\mathbf{u}_h(\varphi), w)_{\Omega_i} &= (f, w)_{\Omega_i}, \quad w\in W_{h,i}. \qquad (2.22)\end{aligned}$$

In particular, $\mathbf{u}_h(\lambda_h) = \mathbf{u}_h$ and $p_h(\lambda_h) = p_h$.

The a priori error bounds from Theorem 2.1 motivate the following assumption on the mortar error.

Saturation assumption. There exists a constant γ such that

$$|||\lambda - \lambda_h||| := \left(\sum_{\tau\in\mathcal{T}^{\Gamma,h}} h_\tau^{-1}\|\lambda-\lambda_h\|_\tau^2\right)^{1/2} \le \gamma\|\mathbf{u}-\mathbf{u}_h\|. \tag{2.23}$$

For further justification of (2.23), note that $|||\lambda-\lambda_h|||$ is closely related to the discrete $H^{1/2}(\Gamma)$ norm and, by (2.20), to $\|\lambda-\lambda_h\|_{d_h}$. Now, assuming that

$$\|\mathbf{u}-\mathbf{u}_h(\lambda)\| \le \gamma_1\|\mathbf{u}-\mathbf{u}_h\|,$$

which is reasonable, since $\mathbf{u}_h(\lambda)$ is the numerical solution based on the true interface data, we have, using (2.19),

$$\begin{aligned}C\|\lambda-\lambda_h\|_{d_h} &\le \|\mathbf{u}_h^*(\lambda) - \mathbf{u}_h^*(\lambda_h)\| = \|\mathbf{u}_h(\lambda) - \mathbf{u}_h(\lambda_h)\| = \|\mathbf{u}_h(\lambda) - \mathbf{u}_h\|\\ &\le \|\mathbf{u}-\mathbf{u}_h(\lambda)\| + \|\mathbf{u}-\mathbf{u}_h\| \le (1+\gamma_1)\|\mathbf{u}-\mathbf{u}_h\|.\end{aligned}$$

Remark 2.2. Condition (2.14) is necessary for the solvability and accuracy of the method and for the validity of (2.23). See [40, 28] for examples of grids that satisfy (2.14). Note that (2.14) excludes the case of matching subdomain grids and a mortar grid that coincides with them. In the case of matching subdomain grids, the mortar grid has to be at least twice as coarse as their trace on the interface. Another possibility in the case of matching grids is to use the standard Lagrange multipliers from the hybrid mixed method [4], in which case the conforming mixed method solution is recovered. This trivial case is not a special case of the mortar mixed finite element method, since the mortar spaces consist of polynomials of one degree higher than the Lagrange multipliers.

2.2. Residual representation and orthogonality of error. Using the notation from (2.7), the solution of (2.11)–(2.13) $(\mathbf{u}_h, p_h, \lambda_h) \in \mathbf{V}_h \times W_h \times M_h$ satisfies

$$A(\mathbf{u}_h, p_h, \lambda_h; \mathbf{v}, w, \mu) = L(\mathbf{v}, w, \mu) \quad \forall\, (\mathbf{v}, w, \mu) \in \mathbf{V}_h \times W_h \times M_h. \tag{2.24}$$

Our goal is to derive a posteriori estimates of the error functions

$$\xi = \mathbf{u} - \mathbf{u}_h, \quad \eta = p - p_h, \quad \text{and} \quad \delta = \lambda - \lambda_h.$$

Using (2.7), $(\xi, \eta, \delta) \in \mathbf{V} \times W \times M$ satisfies the residual equation

(2.25)
$$A(\xi, \eta, \delta; \mathbf{v}, w, \mu) = L(\mathbf{v}, w, \mu) - A(\mathbf{u}_h, p_h, \lambda_h; \mathbf{v}, w, \mu) \quad \forall\, (\mathbf{v}, w, \mu) \in \mathbf{V} \times W \times M,$$

which, together with (2.24), implies the orthogonality condition

$$A(\xi, \eta, \delta; \mathbf{v}, w, \mu) = 0 \quad \forall\, (\mathbf{v}, w, \mu) \in \mathbf{V}_h \times W_h \times M_h. \tag{2.26}$$

2.3. Approximation properties. We present below some of the approximation properties of the finite element spaces. In addition to the operators defined above, we will make use of the interpolant $\mathcal{I}_h$ in the mortar space M_h, and the L^2-projection onto W_h, defined as

$$(w - \hat{w}, w_h) = 0 \quad \forall w_h \in W_h.$$

The following approximation properties hold true. For all $E \in \mathcal{T}_h$, $\tau \in \mathcal{T}^{\Gamma,h}$, $e \in \mathcal{T}_{h,i}|_{\partial\Omega_i}$, and smooth enough functions $\mathbf{v}$, w, and μ,

$$\|\mathbf{v} - \Pi_h \mathbf{v}\|_E \le C h_E \|\mathbf{v}\|_{1,E}, \tag{2.27}$$
$$\|(\mathbf{v} - \Pi_h \mathbf{v}) \cdot \nu_E\|_{\partial E} \le C h_E^s \|\mathbf{v} \cdot \nu_E\|_{s,\partial E}, \quad s = 0, 1/2, \tag{2.28}$$
$$\|w - \hat{w}\|_E \le C h_E \|w\|_{1,E}, \tag{2.29}$$
$$\|\mu - \mathcal{I}_h \mu\|_\tau \le C h_\tau^{3/2} \|\mu\|_{3/2,\tau}, \tag{2.30}$$
$$\|\mu - \mathcal{Q}_{h,i} \mu\|_e \le C h_e \|\mu\|_{1,e}. \tag{2.31}$$

Bound (2.27) can be found in [16, 33]; bounds (2.28)–(2.31) are standard interpolation and L^2-projection approximation results [19].

2.4. Some useful inequalities. In the analysis below we will make use of the trace inequalities

$$\forall E \in \mathcal{T}_h, \quad e \in \partial E, \quad \|\phi\|_e \le C(h_E^{-1/2} \|\phi\|_E + h_E^{1/2} \|\nabla \phi\|_E) \quad \forall \phi \in H^1(E), \tag{2.32}$$

$$\forall E \in \mathcal{T}_h, \quad e \in \partial E, \quad \|\phi\|_{1/2,e} \le C\|\phi\|_{1,E} \quad \forall \phi \in H^1(E), \tag{2.33}$$

$$\forall E \in \mathcal{T}_h, \quad e \in \partial E, \quad \|\mathbf{v}\cdot\nu\|_e \le C h_E^{-1/2}\|\mathbf{v}\|_E \quad \forall \mathbf{v} \in \mathbf{V}_h, \tag{2.34}$$

and the well-known inequality

$$ab \le \epsilon a^2 + \frac{1}{4\epsilon}b^2 \quad \forall \epsilon > 0. \tag{2.35}$$

3. Residual-based error estimators. In this section we derive upper and lower bounds on the error in terms of local residuals. The resulting estimators are often called explicit estimators as they involve only residual terms that depend explicitly on the input data and the computed solution and do not require the solution of additional finite element problems.

3.1. Upper bounds. Let, for all $E \in \mathcal{T}_h$, $\tau \in \mathcal{T}^{\Gamma,h}$,

$$\omega_E^2 = \|K^{-1}\mathbf{u}_h + \nabla p_h\|_E^2 h_E^2 + \|f - \nabla\cdot\mathbf{u}_h\|_E^2 h_E^2 + \|\lambda_h - p_h\|_{\partial E\cap\Gamma}^2 h_E, \tag{3.1}$$

$$\omega_\tau^2 = \|[\mathbf{u}_h\cdot\nu]\|_\tau^2 h_\tau^3, \tag{3.2}$$

where for any $\mathbf{v} \in \mathbf{V}$, $\mathbf{v}|_{\Omega_i} = \mathbf{v}_i$,

$$[\mathbf{v}\cdot\nu]|_{\Gamma_{i,j}} = \mathbf{v}_i\cdot\nu_i + \mathbf{v}_j\cdot\nu_j$$

is the jump operator. We first derive an upper bound on the pressure error η.

THEOREM 3.1. *There exists a constant C independent of h such that*

$$\|\eta\|^2 \le C\bigg\{\sum_{E\in\mathcal{T}_h}\omega_E^2 + \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\omega_\tau^2 + \sum_{e\in\mathcal{T}_h|_{\Gamma_D}}\|g - \mathcal{Q}_h g\|_e^2 h_e\bigg\}.$$

Proof. The proof is based on a duality argument. Consider the auxiliary problem

$$\begin{aligned} -\nabla\cdot K\nabla w &= \eta \quad \text{in } \Omega,\\ w &= 0 \quad \text{on } \Gamma_D,\\ K\nabla w\cdot\nu &= 0 \quad \text{on } \Gamma_N. \end{aligned}$$

The elliptic regularity assumption (2.1) implies that

$$\|w\|_2 \le C\|\eta\|. \tag{3.3}$$

Let $\mathbf{v} = -K\nabla w$ and $\mu = w|_\Gamma$. With (2.7), $(\mathbf{v}, w, \mu)$ satisfy

$$A^s(\mathbf{v}, w, \mu; \tilde{\mathbf{v}}, \tilde{w}, \tilde{\mu}) = -(\eta, \tilde{w}) \quad \forall\ (\tilde{\mathbf{v}}, \tilde{w}, \tilde{\mu}) \in \mathbf{V}\times W\times M.$$

Then, using (2.26) and (2.26),

$$\begin{aligned} \|\eta\|^2 &= -A^s(\mathbf{v}, w, \mu; \xi, \eta, \delta) = -A^s(\xi, \eta, \delta; \mathbf{v}, w, \mu)\\ &= -A^s(\xi, \eta, \delta; \mathbf{v} - \Pi_h\mathbf{v}, w - \hat{w}, \mu - \mathcal{I}_h\mu)\\ &= A^s(\mathbf{u}_h, p_h, \lambda_h; \mathbf{v} - \Pi_h\mathbf{v}, w - \hat{w}, \mu - \hat{\mu}) + (f, w - \hat{w}) + \langle g, (\mathbf{v} - \Pi_h\mathbf{v})\cdot\nu\rangle_{\Gamma_D}\\ &= \sum_{E\in\mathcal{T}_h}\big((K^{-1}\mathbf{u}_h, \mathbf{v} - \Pi_h\mathbf{v})_E - (p_h, \nabla\cdot(\mathbf{v} - \Pi_h\mathbf{v}))_E - (\nabla\cdot\mathbf{u}_h, w - \hat{w})_E\big)\\ &\quad + \sum_{i=1}^n\big(\langle\lambda_h, (\mathbf{v} - \Pi_h\mathbf{v})\cdot\nu_i\rangle_{\Gamma_i} + \langle\mathbf{u}_h\cdot\nu_i, \mu - \mathcal{I}_h\mu\rangle_{\Gamma_i}\big)\\ &\quad + (f, w - \hat{w}) + \langle g, (\mathbf{v} - \Pi_h\mathbf{v})\cdot\nu\rangle_{\Gamma_D}. \end{aligned}$$

Applying Green's formula and using (2.9),

$$\begin{aligned}\|\eta\|^2 = &\sum_{E\in\mathcal{T}_h} \big((K^{-1}\mathbf{u}_h + \nabla p_h, \mathbf{v} - \Pi_h\mathbf{v})_E + (f - \nabla\cdot\mathbf{u}_h, w - \hat{w})_E\big)\\ &+ \sum_{i=1}^{n} \big(\langle \lambda_h - p_h, (\mathbf{v} - \Pi_h\mathbf{v})\cdot\nu_i\rangle_{\Gamma_i} + \langle \mathbf{u}_h\cdot\nu_i, \mu - \mathcal{I}_h\mu\rangle_{\Gamma_i}\big)\\ &+ \langle g - \mathcal{Q}_h g, (\mathbf{v} - \Pi_h\mathbf{v})\cdot\nu\rangle_{\Gamma_D}.\end{aligned}$$

Using the Cauchy–Schwartz inequality and the approximation properties (2.27)–(2.31),

$$\begin{aligned}\|\eta\|^2 \le C\bigg\{ &\sum_{E\in\mathcal{T}_h} \big(\|K^{-1}\mathbf{u}_h + \nabla p_h\|_E h_E \|\mathbf{v}\|_{1,E} + \|f - \nabla\cdot\mathbf{u}_h\|_E h_E \|w\|_{1,E}\\ &+ \|\lambda_h - p_h\|_{\partial E\cap\Gamma} h_E^{1/2}\|\mathbf{v}\|_{1,E}\big) + \sum_{\tau\in\mathcal{T}^{\Gamma,h}} \|[\mathbf{u}_h\cdot\nu]\|_\tau h_\tau^{3/2}\|\mu\|_{3/2,\tau}\\ &+ \sum_{e\in\mathcal{T}_h|_{\Gamma_D}} \|g - \mathcal{Q}_h g\|_e h_e^{1/2}\|\mathbf{v}\|_{1/2,e}\bigg\}.\end{aligned}$$

An application of the discrete Cauchy–Schwartz inequality, the trace inequality (2.33), and (3.3) completes the proof. □

Remark 3.1. Because of the approximation property (2.31) of $\mathcal{Q}_h$ the last term in the bound of Theorem 3.1 is of higher order than the other terms. Therefore, its effect becomes negligible for small h.

To derive a bound on $\xi = \mathbf{u} - \mathbf{u}_h$ we need a saturation assumption. Let $\mathbf{V}'_h$, W'_h, and M'_h be the finite element spaces of one order higher than $\mathbf{V}_h$, W_h, and M_h, respectively. Let $\mathbf{u}'_h \in \mathbf{V}'_h$, $p'_h \in W'_h$, and $\lambda'_h \in M'_h$ be the mortar mixed finite element solution in these higher-order spaces (see (2.11)–(2.13)). The a priori error estimates from Theorem 2.1 motivate the following.

Saturation assumption. There exist constants $\beta < 1$, $\beta_{\mathrm{div}} < 1$, and $\beta_p < \infty$ such that

$$\|\mathbf{u} - \mathbf{u}'_h\| \le \beta\|\mathbf{u} - \mathbf{u}_h\|, \tag{3.4}$$

$$\|\nabla\cdot(\mathbf{u} - \mathbf{u}'_h)\| \le \beta_{\mathrm{div}}\|\nabla\cdot(\mathbf{u} - \mathbf{u}_h)\|, \tag{3.5}$$

$$\|p - p'_h\| \le \beta_p\|p - p_h\|. \tag{3.6}$$

Let

$$\xi' = \mathbf{u}'_h - \mathbf{u}_h, \quad \eta' = p'_h - p_h, \quad \text{and} \quad \delta' = \lambda'_h - \lambda_h.$$

Similar to (2.26) and (2.26), we have that $(\xi', \eta', \delta') \in \mathbf{V}'_h \times W'_h \times M'_h$ satisfy the residual equation

$$\begin{aligned}A(\xi', \eta', \delta'; \mathbf{v}'_h, w'_h, \mu'_h) = L(\mathbf{v}'_h, w'_h, \mu'_h) - A(\mathbf{u}_h, p_h, \lambda_h; \mathbf{v}'_h, w'_h, \mu'_h)&\\ \forall\ (\mathbf{v}'_h, w'_h, \mu'_h) \in \mathbf{V}'_h \times W'_h \times M'_h&\end{aligned} \tag{3.7}$$

and the orthogonality condition

$$A(\xi', \eta', \delta'; \mathbf{v}, w, \mu) = 0 \quad \forall\ (\mathbf{v}, w, \mu) \in \mathbf{V}_h \times W_h \times M_h. \tag{3.8}$$

The bounds on ξ and δ will be expressed in terms of weighted local residuals, for all $E \in \mathcal{T}_h$, $\tau \in \mathcal{T}^{\Gamma,h}$,

$$\begin{aligned}&\tilde{\omega}_E^2 = h_E^{-2}\omega_E^2 = \|K^{-1}\mathbf{u}_h + \nabla p_h\|_E^2 + \|f - \nabla\cdot\mathbf{u}_h\|_E^2 + \|\lambda_h - p_h\|_{\partial E\cap\Gamma}^2 h_E^{-1},\\ &\tilde{\omega}_\tau^2 = h_\tau^{-2}\omega_\tau^2 = \|[\mathbf{u}_h\cdot\nu]\|_\tau^2 h_\tau.\end{aligned}$$

Theorem 3.2. *Assume that the saturation assumptions* (2.23) *and* (3.4) *hold. Then there exists a constant C independent of β such that*

$$\|\xi\|^2_{H(\mathrm{div})} \le \frac{C}{(1-\beta)^2}\left\{ \sum_{E\in\mathcal{T}_h} \tilde\omega_E^2 + \sum_{\tau\in\mathcal{T}^{\Gamma,h}} \tilde\omega_\tau^2 + \sum_{e\in\mathcal{T}_h|_{\Gamma_D}} \|g-\mathcal{Q}_h g\|_e^2 h_e^{-1}\right\}.$$

Proof. The bound on $\|\nabla\cdot\xi\|$ is trivial. Indeed, for all $E\in\mathcal{T}_h$,

$$\|\nabla\cdot\xi\|_E = \|f-\nabla\cdot\mathbf{u}_h\|_E \le \tilde\omega_E.$$

To bound $\|\xi\|$, since (3.4) implies that

$$\|\xi\| \le \frac{1}{1-\beta}\|\xi'\|, \tag{3.9}$$

it is enough to bound $\|\xi'\|$. Using (3.8) and (3.7),

$$\begin{aligned}
\|K^{-1/2}\xi'\|^2 &= A^c(\xi',\eta',\delta';\xi',\eta',\delta') = A^c(\xi',\eta',\delta';\xi'-\Pi_h\xi',\eta',\delta')\\
&= L^c(\xi'-\Pi_h\xi',\eta',\delta') - A^c(\mathbf{u}_h,p_h,\lambda_h;\xi'-\Pi_h\xi',\eta',\delta')\\
&= -\sum_{E\in\mathcal{T}_h}\big((K^{-1}\mathbf{u}_h,\xi'-\Pi_h\xi')_E - (p_h,\nabla\cdot(\xi'-\Pi_h\xi'))_E + (\nabla\cdot\mathbf{u}_h,\eta')_E\big)\\
&\quad -\sum_{i=1}^n\big(\langle\lambda_h,(\xi'-\Pi_h\xi')\cdot\nu_i\rangle_{\Gamma_i} - \langle\mathbf{u}_h\cdot\nu_i,\delta'\rangle_{\Gamma_i}\big)\\
&\quad + (f,\eta') - \langle g,(\xi'-\Pi_h\xi')\cdot\nu\rangle_{\Gamma_D}.
\end{aligned}$$

The use of Green's formula and (2.9) gives

$$\begin{aligned}
\|K^{-1/2}\xi'\|^2 = &-\sum_{E\in\mathcal{T}_h}\big((K^{-1}\mathbf{u}_h+\nabla p_h,\xi'-\Pi_h\xi')_E + (\nabla\cdot\mathbf{u}_h - f,\eta')_E\big)\\
&-\sum_{i=1}^n\big(\langle\lambda_h-p_h,(\xi'-\Pi_h\xi')\cdot\nu_i\rangle_{\Gamma_i} - \langle\mathbf{u}_h\cdot\nu_i,\delta'\rangle_{\Gamma_i}\big)\\
&-\langle g-\mathcal{Q}_h g,(\xi'-\Pi_h\xi')\cdot\nu\rangle_{\Gamma_D} = T_1+\cdots+T_5.
\end{aligned}\tag{3.10}$$

For T_1, using the Cauchy–Schwartz inequality, (2.27), the inverse inequality, and (2.35), we have

$$|(K^{-1}\mathbf{u}_h+\nabla p_h,\xi'-\Pi_h\xi')_E| \le C\left(\frac{1}{4\epsilon_1}\|K^{-1}\mathbf{u}_h+\nabla p_h\|_E^2 + \epsilon_1\|\xi'\|_E^2\right). \tag{3.11}$$

Similarly for T_2,

$$|(\nabla\cdot\mathbf{u}_h - f,\eta')_E| \le \frac{1}{2}\|\nabla\cdot\mathbf{u}_h - f\|_E^2 + \frac{1}{2}\|\eta'\|_E^2. \tag{3.12}$$

To bound T_3, the use of (2.28) with $s=0$ gives, for $e\in\Gamma_i$, $e\in\partial E$,

$$|\langle\lambda_h-p_h,(\xi'-\Pi_h\xi')\cdot\nu_i\rangle_e| \le C\left(\frac{1}{4\epsilon_3}\|\lambda_h-p_h\|_e^2 h_E^{-1} + \epsilon_3\|\xi'\|_E^2\right). \tag{3.13}$$

Similarly for T_5,

$$|\langle g-\mathcal{Q}_h g,(\xi'-\Pi_h\xi')\cdot\nu\rangle_e| \le C\left(\frac{1}{4\epsilon_5}\|g-\mathcal{Q}_h g\|_e^2 h_e^{-1} + \epsilon_5\|\xi'\|_E^2\right). \tag{3.14}$$

Finally for T_4, using (2.35),

$$\begin{aligned}\left|\sum_{i=1}^{n}\langle \mathbf{u}_h\cdot\nu_i,\delta'\rangle_{\Gamma_i}\right| &= \left|\sum_{\tau\in\mathcal{T}^{\Gamma,h}}\langle[\mathbf{u}_h\cdot\nu],\delta'\rangle_\tau\right| \le \sum_{\tau\in\mathcal{T}^{\Gamma,h}} h_\tau^{1/2}\|[\mathbf{u}_h\cdot\nu]\|_\tau h_\tau^{-1/2}\|\delta'\|_\tau \\ &\le \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\left(\frac{1}{4\epsilon_4}\|[\mathbf{u}_h\cdot\nu]\|_\tau^2 h_\tau + \epsilon_4\|\delta'\|_\tau^2 h_\tau^{-1}\right).\end{aligned} \tag{3.15}$$

Combining (1.5) with (3.10)–(3.15) for small enough ϵ_1, ϵ_3, and ϵ_5,

$$\begin{aligned}\|\xi'\|^2 \le C\Bigg\{&\sum_{E\in\mathcal{T}_h}(\|K^{-1}\mathbf{u}_h+\nabla p_h\|_E^2 + \|f-\nabla\cdot\mathbf{u}_h\|_E^2 + \|\lambda_h - p_h\|_{\partial E\cap\Gamma}^2 h_E^{-1} + \|\eta'\|_E^2) \\ &+ \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\left(\frac{1}{4\epsilon_4}\|[\mathbf{u}_h\cdot\nu]\|_\tau^2 h_\tau + \epsilon_4\|\delta'\|_\tau^2 h_\tau^{-1}\right) + \sum_{e\in\mathcal{T}_h|_{\Gamma_D}}\|g-\mathcal{Q}_h g\|_e^2 h_e^{-1}\Bigg\}.\end{aligned} \tag{3.16}$$

Because of (3.6), the bound on $\|\eta\|$ from Theorem 3.1 applies to $\|\eta'\|$ as well. It remains to estimate $|||\delta'|||^2 = \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\|\delta'\|_\tau^2 h_\tau^{-1}$. Using (2.23) (with a constant γ' in the case of the higher-order spaces) and (3.4), we have

$$\begin{aligned}|||\delta'||| &\le |||\lambda-\lambda_h||| + |||\lambda-\lambda_h'||| \le \gamma\|\mathbf{u}-\mathbf{u}_h\| + \gamma'\|\mathbf{u}-\mathbf{u}_h'\| \\ &\le (\gamma+\gamma'\beta)\|\mathbf{u}-\mathbf{u}_h\|.\end{aligned} \tag{3.17}$$

Using (3.17), (3.9) and taking ϵ_4 in (3.16) small enough completes the proof. □

3.2. Lower bounds. Next, we establish lower bounds on the error, which indicate that the residual error estimators can be used effectively in an adaptive mesh refinement algorithm.

THEOREM 3.3. *There exists a constant C independent of h such that*

$$\sum_{E\in\mathcal{T}_h}\omega_E^2 + \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\omega_\tau^2 \le C\left(\|\eta\|^2 + \sum_{E\in\mathcal{T}_h} h_E^2\|\xi\|_{H(\mathrm{div};E)}^2 + \sum_{\tau\in\mathcal{T}^{\Gamma,h}} h_\tau\|\delta\|_\tau^2\right) \tag{3.18}$$

and, assuming that the saturation assumption (2.23) *holds,*

$$\sum_{E\in\mathcal{T}_h}\tilde\omega_E^2 + \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\tilde\omega_\tau^2 \le C\left(\sum_{E\in\mathcal{T}_h} h_E^{-2}\|\eta\|_E^2 + \|\xi\|_{H(\mathrm{div})}^2\right). \tag{3.19}$$

Moreover, the following local bounds hold for any $E\in\mathcal{T}_h$, $e\in\partial E$, and $\tau\in\mathcal{T}^{\Gamma,h}$:

$$\|K^{-1}\mathbf{u}_h+\nabla p_h\|_E^2 h_E^2 + \|f-\nabla\cdot\mathbf{u}_h\|_E^2 h_E^2 \le C(\|\eta\|_E^2 + \|\xi\|_{H(\mathrm{div};E)}^2 h_E^2), \tag{3.20}$$

$$\|[\mathbf{u}_h\cdot\nu]\|_\tau^2 h_\tau^3 \le C\|\xi\|_{H(\mathrm{div};E_\tau)}^2 h_\tau^2, \tag{3.21}$$

$$\|\lambda_h - p_h\|_e^2 h_E \le C(\|\eta\|_E^2 + \|\xi\|_{H(\mathrm{div};E)}^2 h_E^2 + \|\delta\|_e^2 h_E). \tag{3.22}$$

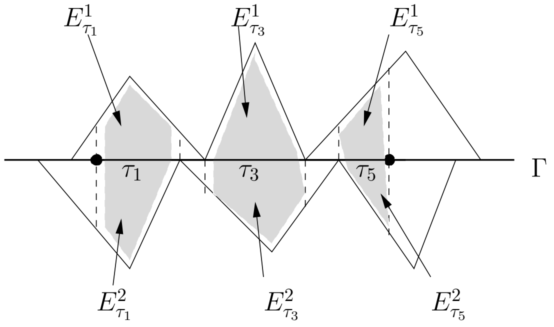


FIG. 3.1. *Construction of* E_{τ_k}.

Proof. It has been shown in [17], using a bubble function argument, that

$$\|K^{-1}\mathbf{u}_h + \nabla p_h\|_E h_E \le C(\|\eta\|_E + \|\xi\|_E h_E),$$

which, combined with

$$\|f - \nabla\cdot\mathbf{u}_h\|_E h_E = \|\nabla\cdot\xi\|_E h_E,$$

gives (3.20). To prove (3.21), consider any $\tau \in \mathcal{T}^{\Gamma,h}$. Let τ be divided by the intersection of the two traces of $\mathcal{T}_h$ on Γ into elements $\tau_1, \dots, \tau_l$. Because of (2.10) there exists $c > 0$ such that

$$h_{\tau_k} \ge c h_\tau, \quad k = 1, \dots, l. \tag{3.23}$$

Next, let us translate any point in τ_k in both directions orthogonal to Γ until an interior edge (face) of an element of $\mathcal{T}_h$ is reached. Let E_{τ_k} be the union of all such trajectories. Figure 3.1 illustrates this construction in the case of triangular grids in $\mathbf{R}^2$, where the neighboring domains are Ω_1 and Ω_2. Note that

$$E_{\tau_k} = E^1_{\tau_k} \cup E^2_{\tau_k},$$

where $E^i_{\tau_k}$, $i = 1, 2$, is a subset of an element of $\mathcal{T}_{h,i}$. Let φ_k be a continuous piecewise linear bubble function such that $0 \le \varphi_k(x) \le 1$ in E_{τ_k}, $\varphi_k = 1$ at the gravity center of τ_k and $\varphi_k = 0$ on ∂E_{τ_k}. Such a function can be easily constructed by decomposing E_{τ_k} into triangles if $d = 2$ or tetrahedra if $d = 3$. We also need an extension of $[\mathbf{u}_h\cdot\nu]_{\tau_k}$ to E_{τ_k}. Given $\psi \in H^{1/2}(\tau_k)$, define $\mathcal{R}\psi \in H^1(E_{\tau_k})$ such that $\mathcal{R}\psi$ is constant along lines perpendicular to Γ. Let

$$\phi_k = \varphi_k \mathcal{R}[\mathbf{u}_h\cdot\nu]_{\tau_k} \in H^1(E_{\tau_k}).$$

Note that $\phi_k = 0$ on ∂E_{τ_k}. Using a scaling argument similar to the one in [35] it can be shown that

$$\|\phi_k\|_{\tau_k} \le \|[\mathbf{u}_h\cdot\nu]\|_{\tau_k}, \tag{3.24}$$

$$\|\nabla\phi_k\|_{E_{\tau_k}} \le C h_{\tau_k}^{-1}\|\phi_k\|_{E_{\tau_k}}, \tag{3.25}$$

$$\|\phi_k\|_{E_{\tau_k}} \le C h_{\tau_k}^{1/2}\|\phi_k\|_{\tau_k}, \tag{3.26}$$

$$C\|[\mathbf{u}_h\cdot\nu]\|^2_{\tau_k} \le \langle\phi_k, [\mathbf{u}_h\cdot\nu]\rangle_{\tau_k}. \tag{3.27}$$

Using (3.27) and that $[\mathbf{u} \cdot \nu] = 0$,

$$
\begin{aligned}
C\|[\mathbf{u}_h \cdot \nu]\|_{\tau_k}^2 &\le \langle \mathbf{u}_{h,1} \cdot \nu_1 + \mathbf{u}_{h,2} \cdot \nu_2, \phi_k \rangle_{\tau_k} \\
&= \langle (\mathbf{u}_{h,1} - \mathbf{u}) \cdot \nu_1, \phi_k \rangle_{\tau_k} + \langle (\mathbf{u}_{h,2} - \mathbf{u}) \cdot \nu_2, \phi_k \rangle_{\tau_k}.
\end{aligned} \tag{3.28}
$$

Using Green's formula for the first term on the right-hand side, we have

$$
\begin{aligned}
\left| \langle \xi_1 \cdot \nu_1, \phi_k \rangle_{\tau_k} \right| &= \left| (\nabla \phi_k, \xi_1)_{E^1_{\tau_k}} + (\phi_k, \nabla \cdot \xi_1)_{E^1_{\tau_k}} \right| \\
&\le \|\nabla \phi_k\|_{E^1_{\tau_k}} \|\xi_1\|_{E^1_{\tau_k}} + \|\phi_k\|_{E^1_{\tau_k}} \|\nabla \cdot \xi_1\|_{E^1_{\tau_k}} \\
&\le C h_{\tau_k}^{-1} \|\phi_k\|_{E^1_{\tau_k}} \|\xi_1\|_{E^1_{\tau_k}} + \|\phi_k\|_{E^1_{\tau_k}} \|\nabla \cdot \xi_1\|_{E^1_{\tau_k}} \\
&\le C(h_{\tau_k}^{-1/2} \|\xi_1\|_{E^1_{\tau_k}} + h_{\tau_k}^{1/2} \|\nabla \cdot \xi_1\|_{E^1_{\tau_k}}) \|[\mathbf{u}_h \cdot \nu]\|_{\tau_k},
\end{aligned} \tag{3.29}
$$

where we have used (3.25) for the second inequality and (3.26), (3.24) for the third inequality. The second term on the right-hand side of (3.28) can be bounded similarly in terms of $\|\xi_2\|_{H(\mathrm{div}; E^2_{\tau_k})}$. A combination of (3.28), (3.29), and (3.23) gives (3.21).

It remains to show (3.22). By the triangle inequality,

$$
\|\lambda_h - p_h\|_e \le \|\lambda_h - \lambda\|_e + \|p - p_h\|_e. \tag{3.30}
$$

For the second term on the right-hand side we employ the trace inequality (2.32)

$$
\begin{aligned}
\|p - p_h\|_e &\le C\big(h_E^{-1/2} \|p - p_h\|_E + h_E^{1/2} \|\nabla (p - p_h)\|_E \big) \\
&\le C\big(h_E^{-1/2} \|p - p_h\|_E + h_E^{1/2} \|K^{-1} \mathbf{u}_h + \nabla p_h\|_E + h_E^{1/2} \|K^{-1}(\mathbf{u} - \mathbf{u}_h)\|_E \big) \\
&\le C\big(h_E^{-1/2} \|\eta\|_E + h_E^{1/2} \|\xi\|_{H(\mathrm{div}; E)} \big),
\end{aligned} \tag{3.31}
$$

using (3.20) for the last inequality. A combination of (3.30), (3.31), and (2.10) completes the proof of (3.22). The global bound (3.18) follows immediately from (3.20) to (3.22), using (2.10), and so does (3.19), using (2.23). □

Remark 3.2. The last two terms in (3.18) are of higher order, so $\|\eta\|$ dominates for small enough h. Therefore, this bound, combined with Theorem 3.1, implies that $\sum_{E \in \mathcal{T}_h} \omega_E^2 + \sum_{\tau \in \mathcal{T}^{\Gamma,h}} \omega_\tau^2$ is an efficient and reliable estimator for the pressure error. Because of the negative power of h in the first term on the right-hand side of (3.19), the estimator $\sum_{E \in \mathcal{T}_h} \tilde{\omega}_E^2 + \sum_{\tau \in \mathcal{T}^{\Gamma,h}} \tilde{\omega}_\tau^2$ provides only a suboptimal bound for the velocity error.

4. Error estimators based on solving local problems.

In this section we derive an implicit error estimator which requires solving local (element) boundary value problems. These problems approximate the local residual equations satisfied by the true error. The motivation for considering implicit estimators comes from the unknown generic constants that appear in the explicit estimators, as well as the suboptimality in the lower bound for the velocity error. We show that the implicit estimator provides both optimal upper and lower bounds of the error.

4.1. Global approximation to the error.

Similar to the approach in [39], we first construct a global approximation to the error based on higher-order finite element spaces. Using (2.4)–(2.6), the true error satisfies the residual equations:

$$(K^{-1}\xi, \mathbf{v})_{\Omega_i} - (\eta, \nabla\cdot\mathbf{v})_{\Omega_i} + \langle\delta, \mathbf{v}\cdot\nu_i\rangle_{\Gamma_i} = -\langle g, \mathbf{v}\cdot\nu\rangle_{\partial\Omega_i\cap\Gamma_D}$$
$$- (K^{-1}\mathbf{u}_h, \mathbf{v})_{\Omega_i} + (p_h, \nabla\cdot\mathbf{v})_{\Omega_i} - \langle\lambda_h, \mathbf{v}\cdot\nu_i\rangle_{\Gamma_i} \equiv r(\mathbf{v}), \quad \mathbf{v}\in\mathbf{V}_i, \tag{4.1}$$
$$(\nabla\cdot\xi, w)_{\Omega_i} = (f - \nabla\cdot\mathbf{u}_h, w)_{\Omega_i}, \quad w\in W_i, \tag{4.2}$$
$$\sum_{i=1}^{n}\langle\xi\cdot\nu_i, \mu\rangle_{\Gamma_i} = -\sum_{i=1}^{n}\langle\mathbf{u}_h\cdot\nu_i, \mu\rangle_{\Gamma_i}, \quad \mu\in M. \tag{4.3}$$

Recall from the previous section that $\mathbf{V}'_h \times W'_h \times M'_h$ are the mortar mixed finite element spaces of one order higher than $\mathbf{V}_h\times W_h\times M_h$ and $(\xi', \eta', \delta') \in \mathbf{V}'_h\times W'_h\times M'_h$ is the finite element approximation to (ξ, η, δ) satisfying

$$(K^{-1}\xi', \mathbf{v})_{\Omega_i} - (\eta', \nabla\cdot\mathbf{v})_{\Omega_i} + \langle\delta', \mathbf{v}\cdot\nu_i\rangle_{\Gamma_i} = r(\mathbf{v}), \quad \mathbf{v}\in\mathbf{V}'_{h,i}, \tag{4.4}$$
$$(\nabla\cdot\xi', w)_{\Omega_i} = (f - \nabla\cdot\mathbf{u}_h, w)_{\Omega_i}, \quad w\in W'_{h,i}, \tag{4.5}$$
$$\sum_{i=1}^{n}\langle\xi'\cdot\nu_i, \mu\rangle_{\Gamma_i} = -\sum_{i=1}^{n}\langle\mathbf{u}_h\cdot\nu_i, \mu\rangle_{\Gamma_i}, \quad \mu\in M'_h. \tag{4.6}$$

Note that (4.4)–(4.6) implies that $(\mathbf{u}'_h = \mathbf{u}_h + \xi', p'_h = p_h + \eta', \lambda'_h = \lambda_h + \delta')$ is the finite element approximation to $(\mathbf{u}, p, \lambda)$ in $\mathbf{V}'_h\times W'_h\times M'_h$ satisfying

$$(K^{-1}\mathbf{u}'_h, \mathbf{v})_{\Omega_i} = (p'_h, \nabla\cdot\mathbf{v})_{\Omega_i} - \langle\lambda'_h, \mathbf{v}\cdot\nu_i\rangle_{\Gamma_i} - \langle g, \mathbf{v}\cdot\nu_i\rangle_{\partial\Omega_i\cap\Gamma_D}, \quad \mathbf{v}\in\mathbf{V}'_{h,i}, \tag{4.7}$$
$$(\nabla\cdot\mathbf{u}'_h, w)_{\Omega_i} = (f, w)_{\Omega_i}, \quad w\in W'_{h,i}, \tag{4.8}$$
$$\sum_{i=1}^{n}\langle\mathbf{u}'_h\cdot\nu_i, \mu\rangle_{\Gamma_i} = 0, \quad \mu\in M'_h. \tag{4.9}$$

The saturation assumptions (3.4) and (3.5) imply

$$(1-\beta)\|\xi\| \le \|\xi'\| \le (1+\beta)\|\xi\|, \tag{4.10}$$
$$(1-\beta_{\mathrm{div}})\|\nabla\cdot\xi\| \le \|\nabla\cdot\xi'\| \le (1+\beta_{\mathrm{div}})\|\nabla\cdot\xi\|, \tag{4.11}$$

so it is enough to estimate ξ'.

4.2. Local (element) approximation to the error. For any $E\in\mathcal{T}_h$, the true error satisfies the local equations:

$$(K^{-1}\xi, \mathbf{v})_E - (\eta, \nabla\cdot\mathbf{v})_E = r_E(\mathbf{v}) - \langle p, \mathbf{v}\cdot\nu_E\rangle_{\partial E}, \quad \mathbf{v}\in\mathbf{V}(E), \tag{4.12}$$
$$(\nabla\cdot\xi, w)_E = (f - \nabla\cdot\mathbf{u}_h, w)_E, \quad w\in W(E), \tag{4.13}$$

where

$$r_E(\mathbf{v}) = -(K^{-1}\mathbf{u}_h, \mathbf{v})_E + (p_h, \nabla\cdot\mathbf{v})_E.$$

We construct a higher-order local approximation of the error by solving element subproblems: find $\psi'\in\mathbf{V}'_h(E)$ and $\theta'\in W'_h(E)$ such that

$$(K^{-1}\psi', \mathbf{v})_E - (\theta', \nabla\cdot\mathbf{v})_E = r_E(\mathbf{v}) - \langle p_A, \mathbf{v}\cdot\nu_E\rangle_{\partial E}, \quad \mathbf{v}\in\mathbf{V}'_h(E), \tag{4.14}$$
$$(\nabla\cdot\psi', w)_E = (f - \nabla\cdot\mathbf{u}_h, w)_E, \quad w\in W'_h(E), \tag{4.15}$$

where $p_A = g$ on Γ_D, $p_A = \lambda_h$ on $\partial E \cap \Gamma$, and $p_A = \tilde{p}_h$ on $\partial E \cap \mathcal{E}_h$, where $\tilde{p}_h$ is the Lagrange multiplier for $\mathbf{V}_h$ and W_h defined as

$$\langle \tilde{p}_h, \mathbf{v}\cdot\nu_E\rangle_{\partial E} = -(K^{-1}\mathbf{u}_h, \mathbf{v})_E + (p_h, \nabla\cdot\mathbf{v})_E, \quad \mathbf{v}\in\mathbf{V}_h(E). \tag{4.16}$$

Let $\tilde{p}'$ be the Lagrange multiplier for the higher-order spaces $\mathbf{V}'_h$ and W'_h satisfying

$$\langle \tilde{p}', \mathbf{v}\cdot\nu_E\rangle_{\partial E} = -(K^{-1}\mathbf{u}'_h, \mathbf{v})_E + (p'_h, \nabla\cdot\mathbf{v})_E, \quad \mathbf{v}\in V'_h(E). \tag{4.17}$$

We make the following.

Saturation assumption. There exists a constant σ such that

$$\left(\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\tilde{p}' - \tilde{p}_h\|_e^2\right)^{1/2} \le \sigma\|\mathbf{u}-\mathbf{u}_h\|. \tag{4.18}$$

Assumption (4.18) is motivated by the a priori error estimate for the Lagrange multiplier [16]

$$\left(\sum_{e\in\mathcal{E}_h} h_e^{-1}\|\bar{p} - \tilde{p}_h\|_e^2\right)^{1/2} \le Ch^{k+1},$$

where $\bar{p}$ is the L^2-projection of p onto $\mathbf{V}_h\cdot\nu|_{\mathcal{E}_h}$.

THEOREM 4.1. *Assume that the saturation assumptions* (2.23), (3.4), (3.5), *and* (4.18) *hold. Then there exist constants C_1 and C_2 independent of β and β_{div} such that*

$$\begin{aligned} C_1\left(\|\psi'\|_{H(\mathrm{div})} + \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\|[\mathbf{u}_h\cdot\nu]\|_\tau h_\tau^{1/2}\right) &\le \|\xi\|_{H(\mathrm{div})} \\ &\le \frac{C_2}{1-\beta_{max}}\left(\|\psi'\|_{H(\mathrm{div})} + \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\|[\mathbf{u}_h\cdot\nu]\|_\tau h_\tau^{1/2}\right), \end{aligned} \tag{4.19}$$

where $\beta_{max} = \max\{\beta,\beta_{\mathrm{div}}\}$.

Proof. We first note that (4.5) and (4.15) imply that on every $E\in\mathcal{T}_h$,

$$\nabla\cdot\psi' = \nabla\cdot\xi'. \tag{4.20}$$

Taking $\mathbf{v} = \psi' - \xi'$ in (4.14) and summing over all elements, we have

(4.21)
$$\begin{aligned} &\sum_{E\in\mathcal{T}_h}\big((K^{-1}(\psi'-\xi'), \psi'-\xi')_E - (\theta'-\eta', \nabla\cdot(\psi'-\xi'))_E\big) \\ &\quad= \sum_{E\in\mathcal{T}_h}\big(-(K^{-1}\xi', \psi'-\xi')_E + (\eta', \nabla\cdot(\psi'-\xi'))_E \\ &\qquad + r_E(\psi'-\xi') - \langle p_A, (\psi'-\xi')\cdot\nu_E\rangle_{\partial E}\big) \\ &\quad= \sum_{E\in\mathcal{T}_h}\big(-(K^{-1}\mathbf{u}'_h, \psi'-\xi')_E + (p'_h, \nabla\cdot(\psi'-\xi'))_E - \langle\tilde{p}_h, (\psi'-\xi')\cdot\nu_E\rangle_{\partial E\cap\mathcal{E}_h} \\ &\qquad - \langle g, (\psi'-\xi')\cdot\nu\rangle_{\partial E\cap\Gamma_D} - \langle\lambda_h, (\psi'-\xi')\cdot\nu_E\rangle_{\partial E\cap\Gamma}\big) \\ &\quad= \sum_{E\in\mathcal{T}_h}\big(\langle\tilde{p}'-\tilde{p}_h, (\psi'-\xi')\cdot\nu_E\rangle_{\partial E\cap\mathcal{E}_h} + \langle\lambda'_h - \lambda_h, (\psi'-\xi')\cdot\nu_E\rangle_{\partial E\cap\Gamma}\big), \end{aligned}$$

using (4.7) and (4.17) for the last equality. For the first term on the right-hand side, using the saturation assumption (4.18) and (2.34), we have

$$\begin{aligned}\left|\sum_{E\in\mathcal{T}_h}\langle\tilde{p}'-\tilde{p}_h,(\psi'-\xi')\cdot\nu_E\rangle_{\partial E\cap\mathcal{E}_h}\right| &\le \sum_{e\in\mathcal{E}_h} h_e^{-1/2}\|\tilde{p}'-\tilde{p}_h\|_e h_e^{1/2}\|(\psi'-\xi')\cdot\nu_e\|_e\\ &\le C\|\xi\|\|\psi'-\xi'\|.\end{aligned}\tag{4.22}$$

For the second term on the right-hand side of (4.21) we write, using (2.34), (2.23), and (3.17),

$$\begin{aligned}\left|\sum_{E\in\mathcal{T}_h}\langle\lambda_h'-\lambda_h,(\psi'-\xi')\cdot\nu_E\rangle_{\partial E\cap\Gamma}\right| &= \left|\sum_{i=1}^n\langle\delta',(\psi'-\xi')\cdot\nu_i\rangle_{\Gamma_i}\right|\\ \le \sum_{\tau\in\mathcal{T}^{\Gamma,h}} h_\tau^{-1/2}\|\delta'\|_\tau h_\tau^{1/2}\|[(\psi'-\xi')\cdot\nu]\|_\tau&\\ \le |||\delta'|||\,\|\psi'-\xi'\| \le C\|\xi\|\|\psi'-\xi'\|.&\end{aligned}\tag{4.23}$$

Combining (4.21)–(4.23) and using (4.20), we obtain

$$\|\psi'-\xi'\|\le C\|\xi\|,$$

which implies, using the triangle inequality and (4.10),

$$\|\psi'\|\le C\|\xi\|.\tag{4.24}$$

Taking $w=\nabla\cdot\psi'$ in (4.15) immediately gives

$$\|\nabla\cdot\psi'\|\le\|\nabla\cdot\xi\|,$$

which, combined with (4.24), implies

$$\|\psi'\|_{H(\mathrm{div})}\le C\|\xi\|_{H(\mathrm{div})}.$$

Combining the above bound with (3.21) completes the proof of the left inequality in (4.19). To show the right inequality in (4.19), taking $\mathbf{v}=\xi'$ in (4.4), and using (4.14), we have

$$(K^{-1}(\xi'-\psi'),\xi')=\sum_{i=1}^n\big((\eta'-\theta',\nabla\cdot\xi')_{\Omega_i}-\langle\delta',\xi'\cdot\nu_i\rangle_{\Gamma_i}\big).\tag{4.25}$$

For the first term on the right-hand side of (4.25) we use (4.20) and the argument from (4.21) to obtain

$$\begin{aligned}\sum_{i=1}^n(\eta'-\theta',\nabla\cdot\xi')_{\Omega_i} &= \sum_{E\in\mathcal{T}_h}(\eta'-\theta',\nabla\cdot\psi')_E\\ &= \big(K^{-1}(\xi'-\psi'),\psi'\big)-\sum_{E\in\mathcal{T}_h}\big(\langle\tilde{p}'-\tilde{p}_h,\psi'\cdot\nu_E\rangle_{\partial E\cap\mathcal{E}_h}\\ &\quad+\langle\lambda_h'-\lambda_h,\psi'\cdot\nu_E\rangle_{\partial E\cap\Gamma}\big),\end{aligned}\tag{4.26}$$

which, combined with (4.25), implies

$$\begin{aligned}&\big(K^{-1}(\xi'-\psi'),\xi'-\psi'\big)=-\langle\delta',\xi'\cdot\nu_i\rangle_{\Gamma_i}\\ &\quad-\sum_{E\in\mathcal{T}_h}\big(\langle\tilde{p}'-\tilde{p}_h,\psi'\cdot\nu_E\rangle_{\partial E\cap\mathcal{E}_h}+\langle\lambda_h'-\lambda_h,\psi'\cdot\nu_E\rangle_{\partial E\cap\Gamma}\big).\end{aligned}\tag{4.27}$$

For the first term on the right-hand side we have, using (4.6),

$$\left|\sum_{i=1}^{n}\langle \delta', \xi'\cdot\nu_i\rangle_{\Gamma_i}\right| = \left|\sum_{i=1}^{n}\langle \delta', \mathbf{u}_h\cdot\nu_i\rangle_{\Gamma_i}\right| \le \epsilon_1\|\xi\|^2 + \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\frac{1}{4\epsilon_1}\|[\mathbf{u}_h\cdot\nu]\|_\tau^2 h_\tau, \tag{4.28}$$

where the inequality is obtained using the argument in (3.15) and (3.17). The last two terms on the right-hand side of (4.27) can be bounded similarly to (4.22) and (4.23):

(4.29)

$$\left|\sum_{E\in\mathcal{T}_h}\left(\langle \tilde{p}' - \tilde{p}_h, \psi'\cdot\nu_E\rangle_{\partial E\cap\mathcal{E}_h} + \langle \lambda_h' - \lambda_h, \psi'\cdot\nu_E\rangle_{\partial E\cap\Gamma}\right)\right| \le C\left(\epsilon_2\|\xi\|^2 + \frac{1}{4\epsilon_2}\|\psi'\|^2\right).$$

Combining (4.27)–(4.29),

$$\|\xi' - \psi'\|^2 \le C\left((\epsilon_1+\epsilon_2)\|\xi\|^2 + \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\frac{1}{4\epsilon_1}\|[\mathbf{u}_h\cdot\nu]\|_\tau^2 h_\tau + \frac{1}{4\epsilon_2}\|\psi'\|^2\right),$$

which implies, using the triangle inequality, (4.10), and taking ϵ_1 and ϵ_2 small enough,

$$\|\xi\| \le \frac{C}{1-\beta}\left(\|\psi'\| + \sum_{\tau\in\mathcal{T}^{\Gamma,h}}\|[\mathbf{u}_h\cdot\nu]\|_\tau h_\tau^{1/2}\right).$$

An application of (4.11) and (4.20) completes the proof. ∎

5. Numerical results.

In this section we test the performance of the residual-based error estimator. The estimator is used as a local error indicator that drives an adaptive mesh refinement process. The following algorithm describes the adaptive procedure.

ALGORITHM.

1. Solve the problem on a coarse (both subdomain and mortar) grid.
2. For each subdomain Ω_i
   (a) Compute
   $$\omega_i = \left(\sum_{E\in\mathcal{T}_{h,i}}\omega_E^2 + \sum_{\tau\in\mathcal{T}^{\Gamma_i,h}}\omega_\tau^2\right)^{1/2}.$$
   (b) If $\omega_i > 0.5\max_{1\le j\le n}\omega_j$, refine $\mathcal{T}_{h,i}$.
   (c) If any neighboring subdomain grid has been refined two times more than Ω_i, refine $\mathcal{T}_{h,i}$.
3. For each interface $\Gamma_{i,j}$, if either Ω_i or Ω_j has been refined, refine $\mathcal{T}_{h,i,j}$.
4. Solve the problem on the refined grid. If either the desired error tolerance or the maximum refinement level has been reached, exit; otherwise go to step 2.

Several comments are in order. First, we employ the pressure error estimator based on ω_E and ω_τ, defined in (3.1) and (3.2), since it provides an efficient and reliable estimate of the L^2 pressure error, due to Theorems 3.1 and 3.3. Second, the refinement rule 2(c) is needed to reduce the effect of discretization error due to large ratios between grid sizes in neighboring subdomains. Third, according to rule 3, mortar grids are refined if either adjacent subdomain grid is refined.

In the examples below, the subdomains are discretized by the lowest-order Raviart–Thomas spaces. Discontinuous piecewise linear mortar spaces are used on the interfaces.

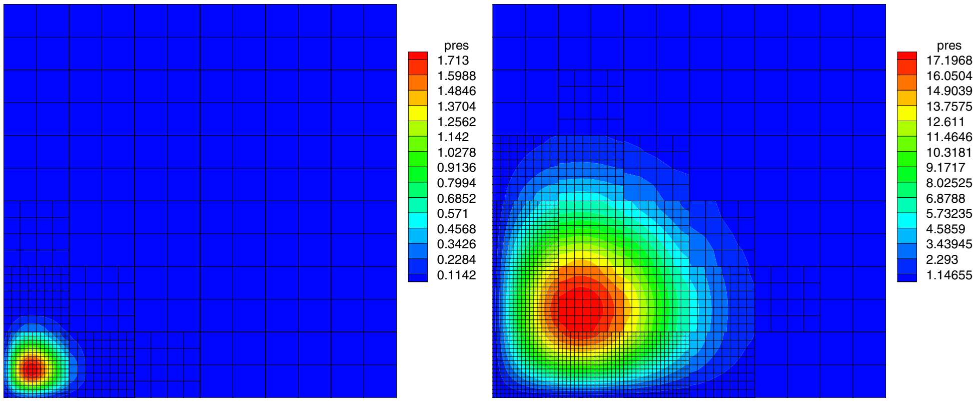


Fig. 5.1. *Computed pressure on the fourth grid level for examples* 1 *and* 2.

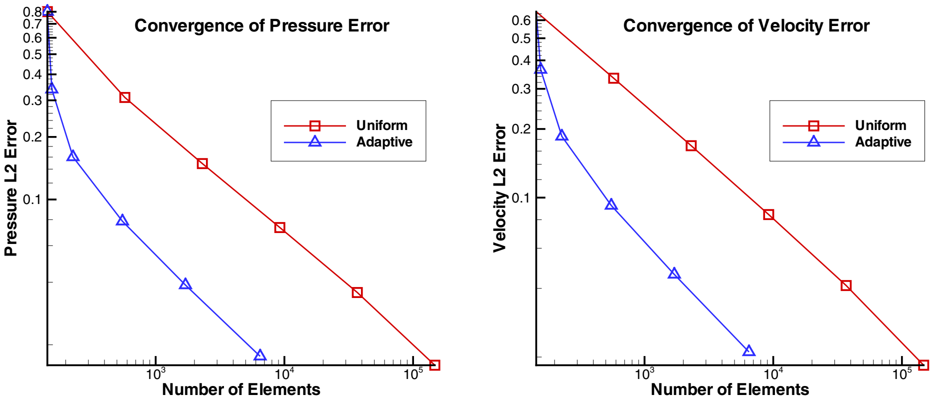


Fig. 5.2. *Convergence of pressure and velocity error for example* 1.

We first illustrate the above algorithm for several two-dimensional problems. In all examples the domain is the unit square, decomposed into 6×6 subdomains. The coarse grid in each subdomain is 2×2 with a single mortar element on each interface. In the first two examples we test problems with boundary layers. The true pressure solution is

$$p(x, y) = 1000\, x\, y\, e^{-k(x^2+y^2)},$$

where $k = 100$ in example 1 and $k = 10$ in example 2. In both cases $K = I$. The computed pressure after three refinements is shown in Figure 5.1. We note that in both cases the grids are appropriately refined along the boundary layers. In the second example the exponential drop is less steep. This causes an extended boundary layer, which is resolved by a strip of fine subdomain grids along the boundary. In Figure 5.2, the pressure and velocity errors in example 1 are plotted as functions of the total number of finite elements. The convergence of the error for all other examples is similar and is not shown. We observe that the adaptive solution needs about 20 times fewer elements to provide the same accuracy.

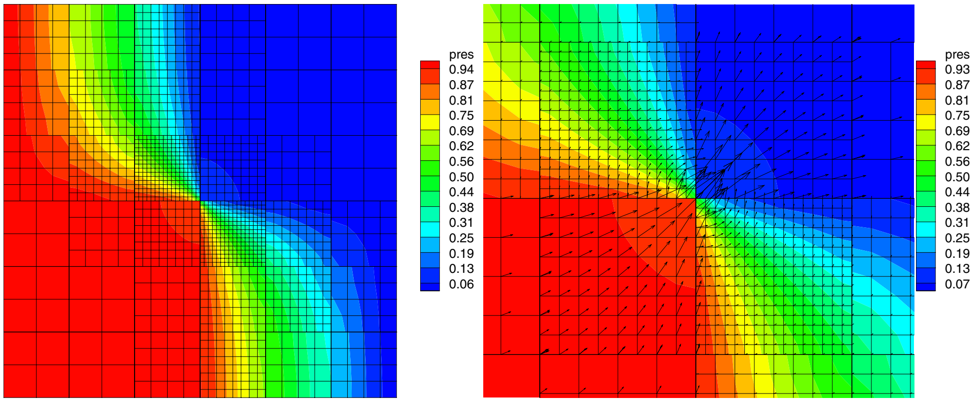


Fig. 5.3. *Computed solution on the fourth grid level for example* 3. *Left: pressure on the full grid. Right: pressure and velocity near the singularity.*

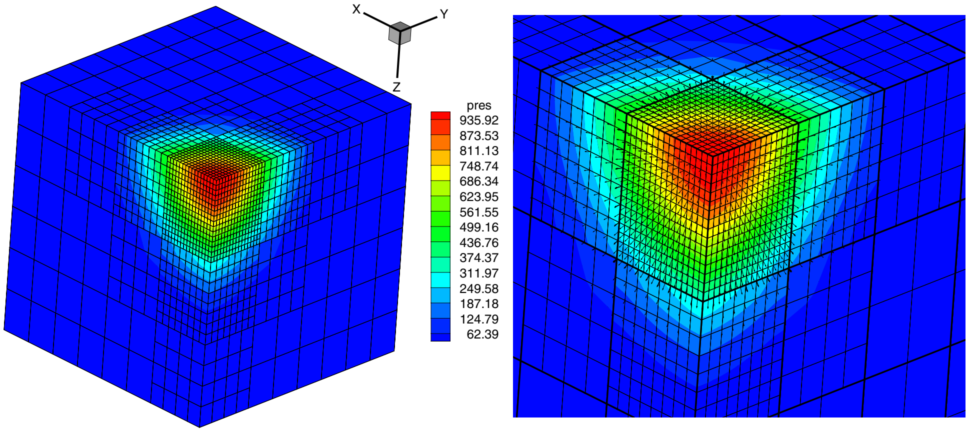


Fig. 5.4. *Computed solution on the fourth grid level for example* 4. *Left: pressure on the full grid. Right: pressure and velocity zoom.*

In the next example, motivated by the modeling flow in heterogeneous porous media, we test a problem with a discontinuous permeability tensor K. The domain is divided into four subregions by the lines $x = 0.5$ and $y = 0.5$. The permeability is $K = 100I$ in the lower-left and upper-right regions and $K = I$ in the other two regions. Dirichlet boundary conditions $p = 1$ on the left and $p = 0$ on the right and no-flow boundary conditions on the top and bottom force the flow from left to right. It is known for the true solution that $p \in H^{1+\alpha}$ for some $0 < \alpha < 1$ with singularity occurring at the cross-point. The computed solution after three refinements is shown in Figure 5.3. As expected the grids are finest near the singularity and are also refined in the low permeability region to resolve the high pressure gradient. Some of the grids in the high permeability region are refined as well, due to the refinement rule 2(c).

Finally, a three-dimensional example is presented. The unit cube is divided into $4 \times 4 \times 3$ subdomains. The true pressure

$$p = 1000\, e^{-10(x^2+y^2+z^2)}$$

exhibits a steep exponential decay near the origin. The computed pressure and velocity on the fourth grid level are given in Figure 5.4. The steep pressure gradient and large velocity are well resolved by the fine computational grids near the origin.

6. Conclusions. In this paper, several two- and three-dimensional a posteriori error estimators for mortar mixed finite element methods for elliptic equations have been derived. A residual-based error estimator provides optimal upper and lower bounds for the pressure error. A closely related error estimator for the velocity gives an optimal upper bound, but suboptimal lower bound. The negative power of h that appears is due to the different order of derivatives involved in the L^2-norm and the $H(\text{div})$-norm. An efficient and reliable implicit estimator for the velocity is also derived, which is based on solving local (element) problems. All estimators include a term that measures the jump of flux across subdomain interfaces. This term provides a measure of nonconformity in the mortar discretization. In cases where the subdomain grids are fixed and optimal mortar grids need to be obtained, this flux-jump term can be used to drive an adaptive process for the mortar grids independently of the subdomain grids.

Acknowledgment. The authors thank Wolfgang Bangerth for his useful comments.

REFERENCES

[1] R. A. Adams, *Sobolev spaces*, in Pure and Applied Mathematics, Vol. 65, Academic Press, New York, 1975.

[2] M. Ainsworth and J. T. Oden, *A unified approach to a posteriori error estimation using element residual methods*, Numer. Math., 65 (1993), pp. 23–50.

[3] T. Arbogast, L. C. Cowsar, M. F. Wheeler, and I. Yotov, *Mixed finite element methods on non-matching multiblock grids*, SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.

[4] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modèl. Math. Anal. Numèr., 19 (1985), pp. 7–32.

[5] I. Babuska and W. C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[6] W. Bangerth and M. F. Wheeler, *A Posteriori Error Estimates and Adaptivity for a Mixed Mortar Method for the Laplace Equation*, preprint.

[7] R. E. Bank and A. Weiser, *Some a posteriori error estimators for elliptic partial differential equations*, Math. Comp., 44 (1985), pp. 283–301.

[8] F. Ben Belgacem, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[9] F. Ben Belgacem, *The mixed mortar finite element method for the incompressible Stokes problem: Convergence analysis*, SIAM J. Numer. Anal., 37 (2000), pp. 1085–1100.

[10] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, H. Brezis and J. L. Lions, eds., Longman Scientific and Technical, Harlow, UK, 1994, pp. 13–51.

[11] D. Braess and R. Verfürth, *A posteriori error estimators for the Raviart–Thomas element*, SIAM J. Numer. Anal., 33 (1996), pp. 2431–2444.

[12] J. H. Brandts, *Superconvergence and a posteriori error estimation for triangular mixed finite elements*, Numer. Math., 68 (1994), pp. 311–324.

[13] F. Brezzi, J. Douglas, Jr., R. Duràn, and M. Fortin, *Mixed finite elements for second order elliptic problems in three variables*, Numer. Math., 51 (1987), pp. 237–250.

[14] F. Brezzi, J. Douglas, Jr., M. Fortin, and L. D. Marini, *Efficient rectangular mixed finite elements in two and three space variables*, RAIRO Modèl. Math. Anal. Numèr., 21 (1987), pp. 581–604.

[15] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed elements for second order elliptic problems*, Numer. Math., 88 (1985), pp. 217–235.

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

[17] C. Carstensen, *A posteriori error estimate for the mixed finite element method*, Math. Comp., 66 (1997), pp. 465–476.

[18] Z. Chen and J. Douglas, Jr., *Prismatic mixed finite elements for second order elliptic problems*, Calcolo, 26 (1989), pp. 135–148.

[19] P. G. Ciarlet, *The finite element method for elliptic problems*, in Studies in Mathematics and Its Applications, Vol. 4, North-Holland, Amsterdam, 1978.

[20] R. Ewing, R. Lazarov, T. Lin, and Y. Lin, *Mortar finite volume element approximations of second order elliptic problems*, East-West J. Numer. Math., 8 (2000), pp. 93–110.

[21] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[22] R. Glowinski and M. F. Wheeler, *Domain decomposition and mixed finite element methods for elliptic problems*, in First International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Periaux, eds., SIAM, Philadelphia, PA, 1988, pp. 144–172.

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

[24] R. H. W. Hoppe and B. I. Wohlmuth, *Adaptive multilevel techniques for mixed finite element discretizations of elliptic boundary value problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1658–1681.

[25] R. Kirby, *Residual a posteriori error estimates for the mixed finite element method*, Comput. Geosci., 7 (2003), pp. 197–214.

[26] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, 1972.

[27] J. C. Nedelec, *Mixed finite elements in* $\mathbf{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[28] G. Pencheva and I. Yotov, *Balancing domain decomposition for mortar mixed finite element methods on non-matching grids*, Numer. Linear Algebra Appl., 10 (2003), pp. 159–180.

[29] M. Peszyńska, M. F. Wheeler, and I. Yotov, *Mortar upscaling for multiphase flow in porous media*, Comput. Geosci., 6 (2002), pp. 73–100.

[30] J. Pousin and T. Sassi, *Domain decomposition with non matching grids and adaptive finite element techniques*, East-West J. Numer. Math., 8 (2000), pp. 243–256.

[31] R. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.

[32] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. Lions, eds., Elsevier Science, Amsterdam, 1991, pp. 523–639.

[33] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 523–639.

[34] J. M. Thomas, These de Doctorat d'etat, 'a l'Universite Pierre et Marie Curie, 1977.

[35] R. Verfürth, *A posteriori error estimation and adaptive mesh-refinement techniques*, J. Comput. Appl. Math., 50 (1994), pp. 67–83.

[36] M. F. Wheeler and I. Yotov, *Physical and computational domain decompositions for modeling subsurface flows*, in Tenth International Conference on Domain Decomposition Methods, Contemporary Mathematics, Vol. 218, J. Mandel et al., eds., American Mathematical Society, Providence, RI, 1998, pp. 217–228.

[37] B. I. Wohlmuth, *Hierarchical a posteriori error estimators for mortar finite element methods with Lagrange multipliers*, SIAM J. Numer. Anal., 36 (1999), pp. 1636–1658.

[38] B. I. Wohlmuth, *A residual based error estimator for mortar finite element discretizations*, Numer. Math., 84 (1999), pp. 143–171.

[39] B. I. Wohlmuth and R. H. W. Hoppe, *A comparison of a posteriori error estimators for mixed finite element discretizations by Raviart–Thomas elements*, Math. Comp., 68 (1999), pp. 1347–1378.

[40] I. Yotov, *Mixed Finite Element Methods for Flow in Porous Media*, Ph.D. thesis, Rice University, Houston, TX, 1996. TR96-09, Dept. Comp. Appl. Math., Rice University and TICAM Report 96-23, University of Texas at Austin.

SIAM J. Numer. Anal.
Vol. 43, No. 3, pp. 1043–1068

THE GAUGE–UZAWA FINITE ELEMENT METHOD. PART I: THE NAVIER–STOKES EQUATIONS*

RICARDO H. NOCHETTO† AND JAE-HONG PYO‡

Abstract. The gauge–Uzawa FEM is a new first order fully discrete projection method which combines advantages of both the gauge and Uzawa methods within a variational framework. A time step consists of a sequence of $d+1$ Poisson problems, d being the space dimension, thereby avoiding both the incompressibility constraint as well as dealing with boundary tangential derivatives as in the gauge method. This allows for a simple finite element discretization in space of any order in both two and three dimensions. This first part introduces the method for the Navier–Stokes equations of incompressible fluids and shows unconditional stability and error estimates for both velocity and pressure via a variational approach under realistic regularity assumptions. Several numerical experiments document performance of the gauge–Uzawa FEM and compare it with other projection methods.

Key words. projection method, gauge method, Uzawa method, Navier–Stokes equation

AMS subject classifications. 65M12, 65M15, 65M60

DOI. 10.1137/040609756

1. Introduction.

Given an open bounded polygon (or polyhedron) Ω in $\mathbb{R}^d$ with $d = 2$ (or 3), we consider the time-dependent Navier–Stokes equations

$$\begin{aligned} \mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p - \mu\triangle\mathbf{u} &= \mathbf{f} \quad \text{in } \Omega, \\ \operatorname{div}\mathbf{u} &= 0 \quad \text{in } \Omega, \\ \mathbf{u}(\mathbf{x},0) &= \mathbf{u}^0 \quad \text{in } \Omega \end{aligned} \tag{1.1}$$

with a vanishing Dirichlet boundary condition $\mathbf{u} = \mathbf{0}$ on $\partial\Omega$ and a pressure mean-value $\int_\Omega p = 0$. This system models the dynamics of an incompressible viscous Newtonian fluid. The viscosity $\mu = Re^{-1}$ is the reciprocal of the Reynolds number. The unknowns are a vector function $\mathbf{u}$ (velocity) and a scalar function p (pressure).

The incompressibility condition $\operatorname{div}\mathbf{u} = 0$ in (1.1) leads to a saddle point structure, which requires compatibility between the discrete spaces for $\mathbf{u}$ and p [1, 2, 10] (*inf-sup condition*). To circumvent this difficulty, projection methods have been studied since the late 1960s which exploit the time dependence in (1.1) [4, 9, 11, 18, 21, 24, 25]. However, such methods

- yield momentum equations inconsistent with the first equation in (1.1);
- impose artificial boundary conditions on pressure (or related variables), which are responsible for boundary layers and reduced accuracy [4, 9];
- require sometimes knowing a suitable initial pressure which is incompatible with the elliptic nature of the Lagrange multiplier p and equation $\operatorname{div}\mathbf{u} = 0$ [11, 18];

*Received by the editors June 8, 2004; accepted for publication (in revised form) February 28, 2005; published electronically September 20, 2005.

http://www.siam.org/journals/sinum/43-3/60975.html

†Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu, www.math.umd.edu/~rhn). The research of this author was partially supported by NSF grants DMS-9971450 and DMS-0204670.

‡Department of Mathematics, Purdue University, West Lafayette, IN 47907 (pjh@math.purdue.edu, www.math.purdue.edu/~pjh). The research of this author was partially supported by NSF grant DMS-9971450.

- are often studied without space discretization [3, 4, 18, 20, 21, 25], and the ensuing analysis may not apply to full discretizations;
- often require unrealistic regularity assumptions in their analysis, particularly so for fully discrete schemes; for instance, $\mathbf{u}_{tt} \in L^\infty(\mathbf{H}^2)$, $\mathbf{u}_{ttt} \in L^\infty(\mathbf{H}^1)$, $p_{tt} \in L^\infty(H^2)$, $p_{ttt} \in L^\infty(L^2)$ are required in [11] for a Chorin finite element method, and similar strong assumptions are made in [27] for a gauge finite difference method.

The gauge method is a projection method, due to Osedelets [17] and E and Liu [7], meant to circumvent these difficulties. It introduces new variables $\mathbf{a}$ and ϕ (gauge) such that $\mathbf{u} = \mathbf{a} + \nabla\phi$ and couple them via the boundary condition $\mathbf{u} = \mathbf{0}$. The method has been studied in [27] using asymptotic methods and in [16] employing variational techniques. The boundary coupling is responsible for accuracy degradation in problems with singular solutions (due to reentrant corners), as will be illustrated below. It also makes the use of finite element methods (FEM) problematic for space discretization. In this paper, we construct a gauge–Uzawa FEM (GU-FEM) which inherits some beneficial properties of both the gauge method and the Uzawa method and which avoids dealing with boundary derivatives. We also prove that the fully discrete method is unconditionally stable and derive error estimates for both velocity and pressure under realistic regularity requirements.

1.1. The gauge–Uzawa finite element method. To motivate the new method we start from the gauge method of Oseledets [17] and E and Liu [7]; see also [16, 19]. Let ϕ be an auxiliary scalar variable, the so-called *gauge* variable, and $\mathbf{a}$ be an unknown vector such that $\mathbf{u} = \mathbf{a} + \nabla\phi$. If ϕ and p satisfy the heat equation $\partial_t\phi - \mu\Delta\phi = -p$, then the momentum and incompressibility equations become

$$\begin{aligned} \partial_t\mathbf{a} + (\mathbf{u}\cdot\nabla)\mathbf{u} - \mu\triangle\mathbf{a} &= \mathbf{f} \quad \text{in } \Omega,\\ -\triangle\phi &= \operatorname{div}\mathbf{a} \quad \text{in } \Omega. \end{aligned}$$

This formulation is equivalent to (1.1) at the PDE level. We are now free to choose boundary conditions for the nonphysical variables $\mathbf{a}$ and ϕ for as long as $\mathbf{u} = \mathbf{0}$ is enforced. Hereafter, we employ a Neumann condition on ϕ which, according to [7, 16, 19, 27], is the most advantageous:

$$\partial_{\boldsymbol{\nu}}\phi = 0, \quad \mathbf{a}\cdot\boldsymbol{\nu} = 0, \quad \mathbf{a}\cdot\boldsymbol{\tau} = -\partial_{\boldsymbol{\tau}}\phi;$$

$\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are the unit vectors in the normal and tangential directions, respectively. Upon discretizing in time via the backward Euler method [7, 27] and a semi-implicit treatment of the convection term, we end up with the following unconditionally stable method [16, 19].

Algorithm 1 (gauge method). Start with $\phi^0 = 0$ and $\mathbf{a}^0 = \mathbf{u}^0$. Repeat the steps.
Step 1: Find $\mathbf{a}^{n+1}$ as the solution of

$$\text{(1.2)} \qquad \begin{aligned} \frac{\mathbf{a}^{n+1} - \mathbf{a}^n}{\tau} + (\mathbf{u}^n\cdot\nabla)(\mathbf{a}^{n+1} + \nabla\phi^n) - \mu\triangle\mathbf{a}^{n+1} &= \mathbf{f}(t^{n+1}) \quad \text{in } \Omega,\\ \mathbf{a}^{n+1}\cdot\boldsymbol{\nu} = 0, \qquad \mathbf{a}^{n+1}\cdot\boldsymbol{\tau} &= -\partial_{\boldsymbol{\tau}}\phi^n \quad \text{on } \partial\Omega. \end{aligned}$$

Step 2: Find ϕ^{n+1} as the solution of

$$\begin{aligned} -\triangle\phi^{n+1} &= \operatorname{div}\mathbf{a}^{n+1} \qquad \text{in } \Omega,\\ \partial_{\boldsymbol{\nu}}\phi^{n+1} &= 0 \qquad \text{on } \partial\Omega. \end{aligned}$$

Step 3: Update $\mathbf{u}^{n+1}$ according to

$$\mathbf{u}^{n+1} = \mathbf{a}^{n+1} + \nabla\phi^{n+1}. \tag{1.3}$$

We point out that the momentum equation is linear in $\mathbf{a}^{n+1}$, and that the explicit boundary condition $\mathbf{a}^{n+1}\cdot\boldsymbol{\tau} = -\partial_{\boldsymbol{\tau}}\phi^n$ is crucial to decouple the equations for $\mathbf{a}^{n+1}$ and ϕ^{n+1}. Since this formulation is consistent with (1.1), except for $\mathbf{u}^{n+1}\cdot\boldsymbol{\tau} = \partial_{\boldsymbol{\tau}}(\phi^{n+1}-\phi^n)$, normal mode analysis can be used to show full accuracy for smooth solutions [3, 20]. However, several deficiencies of this algorithm are now apparent.

- The boundary term $\partial_{\boldsymbol{\tau}}\phi^n$ is nonvariational and thus difficult to implement within a finite element context, especially in three dimensions.
- The computation of $\partial_{\boldsymbol{\tau}}\phi^n$, which involves numerical differentiation, yields loss of accuracy and is problematic at corners of $\partial\Omega$ where $\boldsymbol{\tau}$ is not well defined. This is remarkably important for reentrant corners as illustrated in the comparisons below.
- The computation of $p^{n+1} = \mu\Delta\phi^{n+1} - \tau^{-1}(\phi^{n+1}-\phi^n)$ is also unstable. This yields a reduced rate of convergence or lack of convergence altogether [16, 19, 27].
- Numerical experiments indicate that the polynomial degree for ϕ must be of higher order than that for p [19]. A suitable combination of finite element spaces for $(\mathbf{a},\mathbf{u},\phi,p)$ is continuous piecewise polynomials $(\mathcal{P}^2,\mathcal{P}^2,\mathcal{P}^3,\mathcal{P}^1)$, which is consistent with (1.3) and the previous expression for p^{n+1}. This computation is, however, rather costly since ϕ is just an auxiliary variable without intrinsic interest [19].

The purpose of this paper is to construct and study the gauge–Uzawa FEM, which overcomes these shortcomings without losing advantages of the gauge method. We start by introducing a new vector variable $\widehat{\mathbf{u}}^{n+1}$ having zero boundary values

$$\widehat{\mathbf{u}}^{n+1} = \mathbf{a}^{n+1} + \nabla\phi^n.$$

Inserting this into (1.2), we readily get

$$\frac{\widehat{\mathbf{u}}^{n+1}-\mathbf{u}^n}{\tau} + (\mathbf{u}^n\cdot\nabla)\widehat{\mathbf{u}}^{n+1} - \mu\triangle\widehat{\mathbf{u}}^{n+1} + \mu\nabla\triangle\phi^n = \mathbf{f}(t^{n+1}) \quad \text{in } \Omega. \tag{1.4}$$

To deal with the third order term $\nabla\triangle\phi^n$, which is a source of trouble due to lack of commutativity of the differential operators at the discrete level, we introduce the variable $s^{n+1} = \triangle\phi^{n+1}$ and note the connection with the *Uzawa iteration*:

$$s^{n+1} = \triangle\phi^{n+1} = -\operatorname{div}\mathbf{a}^{n+1} = \triangle\phi^n - \operatorname{div}\widehat{\mathbf{u}}^{n+1} = s^n - \operatorname{div}\widehat{\mathbf{u}}^{n+1}. \tag{1.5}$$

If we also set $\rho^{n+1} = \phi^{n+1} - \phi^n$, then

$$-\triangle\rho^{n+1} = -\triangle(\phi^{n+1}-\phi^n) = \operatorname{div}\widehat{\mathbf{u}}^{n+1}. \tag{1.6}$$

Combining (1.4), (1.5), and (1.6) we arrive at the discrete-time gauge–Uzawa method.

In order to introduce the finite element discretization we need further notation. Let $H^s(\Omega)$ be the Sobolev space with s derivatives in $L^2(\Omega)$, set $\mathbf{L}^2(\Omega) = (L^2(\Omega))^d$ and $\mathbf{H}^s(\Omega) = (H^s(\Omega))^d$, where $d = 2$ or 3, and denote by $L^2_0(\Omega)$ the subspace of $L^2(\Omega)$ of functions with vanishing mean value. We indicate with $\|\cdot\|_s$ the norm in $H^s(\Omega)$ and with $\langle\cdot\,,\cdot\rangle$ the inner product in $L^2(\Omega)$. Let $\mathfrak{T} = \{K\}$ be a shape-regular quasi-uniform partition of Ω of mesh size h into closed elements K [1, 2, 10]. The vector and scalar finite element spaces are

$$\begin{aligned}
&\mathbb{W}_h := \{\mathbf{v}_h \in \mathbf{L}^2(\Omega) : \mathbf{v}_h|_K \in \mathcal{P}(K) \quad \forall K \in \mathfrak{T}\}, \quad \mathbb{V}_h := \mathbb{W}_h \cap \mathbf{H}^1_0(\Omega),\\
&\mathbb{P}_h := \{q_h \in L^2_0(\Omega) \cap C^0(\Omega) : q_h|_K \in \mathcal{Q}(K) \quad \forall K \in \mathfrak{T}\},
\end{aligned}$$

where $\mathcal{P}(K)$ and $\mathcal{Q}(K)$ are spaces of polynomials with degree bounded uniformly with respect to $K \in \mathfrak{T}$ [2, 10]. We stress that the space $\mathbb{P}_h$ is composed of continuous functions for (1.6) to make sense. This implies the crucial equality

$$\langle \operatorname{div} \mathbf{v}_h \,,\, q_h \rangle = -\langle \mathbf{v}_h \,,\, \nabla q_h \rangle \quad \forall \mathbf{v}_h \in \mathbb{V}_h, \quad q_h \in \mathbb{P}_h.$$

Using the following discrete counterpart of the form $\mathfrak{N}(\mathbf{u}, \mathbf{v}, \mathbf{w}) = \langle (\mathbf{u} \cdot \nabla)\mathbf{v} \,,\, \mathbf{w} \rangle$:

$$\mathfrak{N}_h(\mathbf{u}_h, \mathbf{v}_h, \mathbf{w}_h) = \frac{1}{2} \langle (\mathbf{u}_h \cdot \nabla)\mathbf{v}_h \,,\, \mathbf{w}_h \rangle - \frac{1}{2} \langle (\mathbf{u}_h \cdot \nabla)\mathbf{w}_h \,,\, \mathbf{v}_h \rangle, \tag{1.7}$$

we are ready to write the gauge–Uzawa finite element method.

Algorithm 2 (gauge–Uzawa FEM). Start with $s_h^0 = 0$ and $\mathbf{u}_h^0$ as a solution of $\langle \mathbf{u}_h^0 \,,\, \mathbf{w}_h \rangle = \langle \mathbf{u}^0 \,,\, \mathbf{w}_h \rangle$ for all $\mathbf{w}_h \in \mathbb{V}_h$.

Step 1: Find $\widehat{\mathbf{u}}_h^{n+1} \in \mathbb{V}_h$ as the solution of

$$\begin{aligned} \tau^{-1} \langle \widehat{\mathbf{u}}_h^{n+1} - \mathbf{u}_h^n \,,\, \mathbf{w}_h \rangle + \mathfrak{N}_h(\mathbf{u}_h^n, \widehat{\mathbf{u}}_h^{n+1}, \mathbf{w}_h) + \mu \langle \nabla \widehat{\mathbf{u}}_h^{n+1} \,,\, \nabla \mathbf{w}_h \rangle \\ - \mu \langle s_h^n \,,\, \operatorname{div} \mathbf{w}_h \rangle = \langle \mathbf{f}(t^{n+1}) \,,\, \mathbf{w}_h \rangle \quad \forall \mathbf{w}_h \in \mathbb{V}_h. \end{aligned} \tag{1.8}$$

Step 2: Find $\rho_h^{n+1} \in \mathbb{P}_h$ as the solution of

$$\langle \nabla \rho_h^{n+1} \,,\, \nabla \psi_h \rangle = \langle \operatorname{div} \widehat{\mathbf{u}}_h^{n+1} \,,\, \psi_h \rangle \qquad \forall \psi_h \in \mathbb{P}_h. \tag{1.9}$$

Step 3: Update $s_h^{n+1} \in \mathbb{P}_h$ according to

$$\langle s_h^{n+1} \,,\, q_h \rangle = \langle s_h^n \,,\, q_h \rangle - \langle \operatorname{div} \widehat{\mathbf{u}}_h^{n+1} \,,\, q_h \rangle \qquad \forall q_h \in \mathbb{P}_h. \tag{1.10}$$

Step 4: Update $\mathbf{u}_h^{n+1} \in \mathbb{V}_h + \nabla \mathbb{P}_h$ according to

$$\mathbf{u}_h^{n+1} = \widehat{\mathbf{u}}_h^{n+1} + \nabla \rho_h^{n+1}. \tag{1.11}$$

We note that $\mathbf{u}_h^{n+1}$ is a discontinuous function across interelement boundaries and that, in light of (1.9), $\mathbf{u}_h^{n+1}$ is discrete divergence free in the sense that

$$\langle \mathbf{u}_h^{n+1} \,,\, \nabla \psi_h \rangle = 0 \qquad \forall \psi_h \in \mathbb{P}_h. \tag{1.12}$$

In addition, the discrete pressure $p_h^{n+1} \in \mathbb{P}_h$ can be computed via

$$p_h^{n+1} = \mu s_h^{n+1} - \tau^{-1} \rho_h^{n+1}. \tag{1.13}$$

Consequently, the ensuing momentum equations for either $(\widehat{\mathbf{u}}^{n+1}, p^n)$ or $(\mathbf{u}^{n+1}, p^{n+1})$ are fully consistent with (1.1), a distinctive feature of this new formulation:

$$\begin{aligned} \tau^{-1} \langle \widehat{\mathbf{u}}_h^{n+1} - \widehat{\mathbf{u}}_h^n \,,\, \mathbf{w}_h \rangle + \mathfrak{N}_h(\mathbf{u}_h^n, \widehat{\mathbf{u}}_h^{n+1}, \mathbf{w}_h) + \mu \langle \nabla \widehat{\mathbf{u}}_h^{n+1} \,,\, \nabla \mathbf{w}_h \rangle \\ - \langle p_h^n \,,\, \operatorname{div} \mathbf{w}_h \rangle = \langle \mathbf{f}(t^{n+1}) \,,\, \mathbf{w}_h \rangle \quad \forall \mathbf{w}_h \in \mathbb{V}_h. \end{aligned} \tag{1.14}$$

1.2. Comparison with other projection methods. We now compare the gauge–Uzawa FEM of Algorithm 2 with the original Chorin method [4, 25], the Chorin–Uzawa method [18], and the gauge method of Algorithm 1 [7, 16, 19] using finite elements of degree 2 for $\mathbf{u}, \widehat{\mathbf{u}}, \mathbf{a}$, of degree 1 for p, s, ρ, and of degree 3 for ϕ.

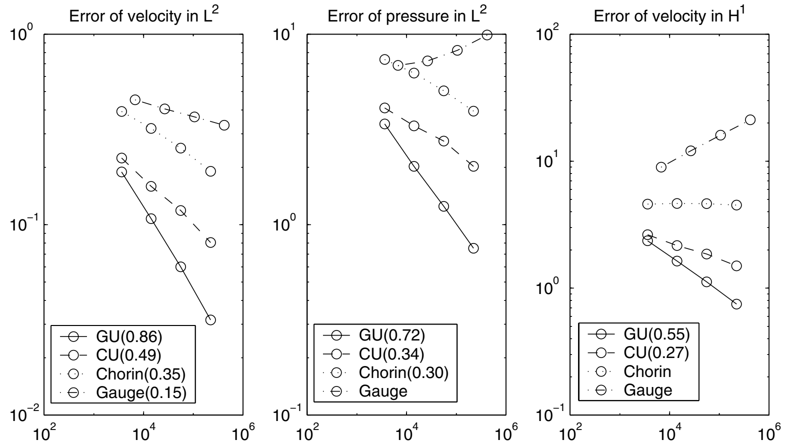


FIG. 1.1. *Error decay vs. number of degrees of freedom for four projection methods; the errors are measured in $L^2(\mathbf{L}^2)$ and $L^2(\mathbf{H}^1)$ for velocity and $L^2(L^2)$ for pressure. Velocity and pressure do not always converge for the gauge method, even though we use the best finite element combination $(\mathcal{P}^2, \mathcal{P}^1, \mathcal{P}^3)$ for $(\mathbf{u}, p, \phi)$. The gauge–Uzawa FEM exhibits a superior performance overall. Numbers in parentheses are the experimental orders of convergence.*

We consider the L-shaped domain $\Omega = ((-1,1)\times(-1,1)) - ([0,1)\times(-1,0])$ and the corresponding time-dependent singular solution of the Stokes equation ($\mathfrak{N}_h = 0$) [26]

$$\mathbf{u}(r,\theta) = \frac{3-\cos(5t)}{4} r^{\alpha} \begin{bmatrix} \cos(\theta)\psi'(\theta) + (1+\alpha)\sin(\theta)\psi(\theta) \\ \sin(\theta)\psi'(\theta) - (1+\alpha)\cos(\theta)\psi(\theta) \end{bmatrix},$$

$$p(r,\theta) = -\frac{3-\cos(5t)}{4} r^{\alpha-1} \frac{(1+\alpha)^2\psi'(\theta) + \psi'''(\theta)}{1-\alpha},$$

where $\omega = \frac{3\pi}{2}, \alpha = 0.544,$

$$\psi(\theta) = \frac{\sin((1+\alpha)\theta)\cos(\alpha\omega)}{1+\alpha} - \cos((1+\alpha)\theta) + \frac{\sin((\alpha-1)\theta)\cos(\alpha\omega)}{1-\alpha} + \cos((\alpha-1)\theta),$$

and $T = 5$. This example is not covered by theory because $(\mathbf{u}, p)(\cdot, t) \notin \mathbf{H}^2(\Omega)\times H^1(\Omega)$ due to $\alpha < 1$ (see Lemma 2.1); it provides, however, quite strong computational support to GU-FEM and hints at the need for further analysis. The initial mesh and time steps are $\tau = h = 1/8$ and are subsequently halved for every experiment.

Figure 1.1 clearly shows the superior performance of the gauge–Uzawa FEM, particularly so in regard to pressure approximation for which the gauge method fails to converge. These experiments, as well as those in section 7, were carried out within the software platform ALBERT of Schmidt and Siebert [22].

1.3. The main results. We now summarize our theoretical results of the rest of this paper for the gauge–Uzawa FEM. In section 3 we prove stability.

Theorem 1.1 (stability). *The gauge–Uzawa FEM is unconditionally stable in the sense that, for all $\tau > 0$, the following a priori bound holds:*

$$\begin{aligned}\left\|\mathbf{u}_h^{N+1}\right\|_0^2 + \sum_{n=0}^{N}\left\|\mathbf{u}_h^{n+1}-\mathbf{u}_h^{n}\right\|_0^2 + \frac{\mu\tau}{2}\sum_{n=0}^{N}\left\|\nabla\widehat{\mathbf{u}}_h^{n+1}\right\|_0^2 \\ + 2\sum_{n=0}^{N}\left\|\nabla\rho_h^{n+1}\right\|_0^2 + \mu\tau\left\|s_h^{N+1}\right\|_0^2 \le \left\|\mathbf{u}_h^0\right\|_0^2 + C\tau\sum_{n=0}^{N}\left\|\mathbf{f}(t^{n+1})\right\|_{-1}^2.\end{aligned} \tag{1.15}$$

We then study the rate of convergence of various unknowns under appropriate assumptions A1–A6 described in section 2. In section 4 we prove error estimates for velocity.

Theorem 1.2 (error estimates for velocity). *If* A1–A6 *hold and $h^2 \le C\tau$, with $C > 0$ arbitrary, then we have the error estimates*

$$\tau\sum_{n=0}^{N}\left\|\nabla\left(\mathbf{u}(t^{n+1})-\widehat{\mathbf{u}}_h^{n+1}\right)\right\|_0^2 \le C(\tau+h^2),$$

$$\tau\sum_{n=0}^{N}\left(\left\|\mathbf{u}(t^{n+1})-\mathbf{u}_h^{n+1}\right\|_0^2 + \left\|\mathbf{u}(t^{n+1})-\widehat{\mathbf{u}}_h^{n+1}\right\|_0^2\right) \le C(\tau+h^2)^2.$$

Given a sequence $\{W^n\}_{n=0}^N$, we define its discrete time derivative to be

$$\delta W^{n+1} := \frac{W^{n+1}-W^n}{\tau}.$$

We also define the discrete weight $\sigma^n := \min(t^n, 1)$ for $1 \le n \le N$. In section 5 we derive an error estimate for time derivative of velocity and utilize it in section 6 to prove and error estimate for pressure.

Theorem 1.3 (error estimates for time derivative of velocity and pressure). *Let* A1–A6 *hold and $C_1h^2 \le \tau \le C_2h^{\frac{d}{3}(1+\varepsilon)}$ be valid with arbitrary constants $C_1 > 0$ and $C_2 > 0$, where d is the space dimension. Then the following weighted estimates hold:*

$$\tau\sum_{n=0}^{N}\sigma^{n+1}\left(\left\|\delta(\mathbf{u}(t^{n+1})-\mathbf{u}_h^{n+1})\right\|_0^2 + \left\|p(t^{n+1})-p_h^{n+1}\right\|_0^2\right) \le C(\tau+h^2).$$

If NLC of section 2 *is also satisfied, then the following uniform error estimates are valid:*

$$\tau\sum_{n=0}^{N}\left(\left\|\delta(\mathbf{u}(t^{n+1})-\mathbf{u}_h^{n+1})\right\|_0^2 + \left\|p(t^{n+1})-p_h^{n+1}\right\|_0^2\right) \le C(\tau+h^2).$$

Proofs of Theorems 1.1–1.3 follow the variational approach of [16, 19]. These error estimates are similar to those known for the Chorin method [11, 18, 21], but require realistic and thus provable regularity when $t \downarrow 0$. They do not explain, though, the rates observed in practice for smooth solutions, as reported in section 7.1, which remains an open question. We also show in section 7.2 how to accommodate other boundary conditions.

2. Basic assumptions and regularity. This section is mainly devoted to stating assumptions and basic regularity results. We refer the reader to Constantin and Foias [5], Heywood and Rannacher [12], and Prohl [18] for details.

2.1. Regularity. We start with three basic assumptions about data Ω, $\mathbf{u}^0$, $\mathbf{f}$, and $\mathbf{u}$. We consider first the stationary Stokes equations which will be used in a duality argument:

$$\begin{aligned} -\triangle \mathbf{v} + \nabla q &= \mathbf{g} \quad \text{in } \Omega, \\ \operatorname{div} \mathbf{v} &= 0 \quad \text{in } \Omega, \\ \mathbf{v} &= \mathbf{0} \quad \text{on } \partial\Omega. \end{aligned} \tag{2.1}$$

Assumption A1 (regularity of $(\mathbf{v}, q)$). The unique solution $(\mathbf{v}, q) \in H_0^1(\Omega) \times L_0^2(\Omega)$ of the stationary Stokes equations (2.1) satisfies

$$\|\mathbf{v}\|_2 + \|q\|_1 \leq C\|\mathbf{g}\|_0.$$

We notice that A1 is valid provided $\partial\Omega$ is of class C^2 [5], or if Ω is a *convex* two-dimensional polygon [13] or three-dimensional polyhedron [6].

Assumption A2 (data regularity). The initial velocity $\mathbf{u}^0$ and the forcing term $\mathbf{f}$ in (1.1) satisfy

$$\mathbf{u}^0 \in \mathbf{H}^2(\Omega) \cap \mathbf{Z}(\Omega) \qquad \text{and} \qquad \mathbf{f},\, \mathbf{f}_t \in \mathbf{L}^\infty(0, T; \mathbf{L}^2(\Omega)),$$

where $\mathbf{Z}(\Omega) := \{\mathbf{z} \in \mathbf{H}_0^1(\Omega) : \operatorname{div} \mathbf{z} = 0\}$.

Assumption A3 (regularity of the solution $\mathbf{u}$). There exists $M > 0$ such that

$$\sup_{t\in[0,T]} \|\nabla \mathbf{u}(t)\|_0 \leq M.$$

We note that A3 is always satisfied in two dimensions, whereas it is valid in three dimensions provided $\left\|\mathbf{u}^0\right\|_1$ and $\|\mathbf{f}\|_{L^\infty(0,T;\mathbf{L}^2(\Omega))}$ are sufficiently small [12].

Lemma 2.1 (uniform and weighted a priori estimates [12]). *Let* $\sigma(t) = \min\{t, 1\}$ *be a weight function and* $0 < T \leq \infty$. *If* A1–A3 *holds, the solution* $(\mathbf{u}, p)$ *of* (1.1)

$$\sup_{0<t<T} \left(\|\mathbf{u}\|_2 + \|\mathbf{u}_t\|_0 + \|p\|_1\right) \leq M, \qquad \int_0^T \|\mathbf{u}_t\|_1^2\, dt \leq M, \tag{2.2}$$

$$\sup_{0<t<T} \left(\sigma(t)\|\mathbf{u}_t\|_1^2\right) \leq M, \qquad \int_0^T \sigma(t)\left(\|\mathbf{u}_t\|_2^2 + \|\mathbf{u}_{tt}\|_0^2 + \|p_t\|_1^2\right) dt \leq M. \tag{2.3}$$

Consequently, $(\mathbf{u}, p) \in L^\infty(0, T; \mathbf{H}^2(\Omega) \times H^1(\Omega))$ provided A1–A3 are valid.

The following *nonlocal* assumption is used to remove the weight $\sigma(t)$ for the error estimates for $\mathbf{u}_t$ in section 5 and pressure in section 6.

Assumption NLC (nonlocal compatibility). The data $\mathbf{u}^0$ and $\mathbf{f}^0 = \mathbf{f}(0, \cdot)$ are such that $\|\nabla \mathbf{u}_t(0)\|_0 \leq M$.

In view of [12, Corollary 2.1], we realize that NLC is equivalent to the initial data $\mathbf{u}^0, p^0 = p(0, \cdot), \mathbf{f}^0$ satisfying the overdetermined system

$$\Delta p^0 = \operatorname{div}(\mathbf{f}^0 - (\mathbf{u}^0 \cdot \nabla)\mathbf{u}^0) \quad \text{in } \Omega, \qquad \nabla p^0 = \Delta \mathbf{u}^0 + \mathbf{f}^0 - (\mathbf{u}^0 \cdot \nabla)\mathbf{u}^0 \quad \text{on } \partial\Omega.$$

This is true if $\mathbf{u}^0 = \mathbf{f}^0 = \mathbf{0}$, in which case also $p^0 = 0$ and $\|\nabla \mathbf{u}_t(0)\|_0 = 0$. However, $\|\nabla \mathbf{u}_t(t)\|_0$ blows-up in general as $t \downarrow 0$, thereby uncovering the practical limitations of results based on higher regularity than (2.2) and (2.3) uniformly for $t \downarrow 0$ [11, 27].

LEMMA 2.2 (uniform a priori estimates [12, Corollary 2.1]). *Suppose* A1–A3 *hold and let* $0 < T \leq \infty$. *Then NLC is valid if and only if*

$$\int_0^T \|\mathbf{u}_{tt}(t)\|_0^2\, dt + \sup_{0<t<T} \|\nabla \mathbf{u}_t(t)\|_0^2 \leq M. \tag{2.4}$$

Furthermore, if NLC holds, then $\int_0^T (\|p_t(t)\|_1^2 + \|\mathbf{u}_t(t)\|_2^2)\, dt \leq M$.

LEMMA 2.3 (a priori estimates on $\mathbf{Z}(\Omega)^*$ [16, 19]). *If* A1–A3 *hold, then we have*

$$\int_0^T \|\mathbf{u}_{tt}(t)\|_*^2\, dt \leq M, \tag{2.5}$$

where $\mathbf{Z}(\Omega)^*$ *is a dual space of* $\mathbf{Z}(\Omega)$. *If NLC also holds, then* $\sup_{0<t<T} \|\mathbf{u}_{tt}(t)\|_*^2 \leq M$.

LEMMA 2.4 (div-grad relation [15, 16, 19, 24]). *If* $\mathbf{v} \in \mathbf{H}_0^1(\Omega)$, *then*

$$\|\text{div } \mathbf{v}\|_0 \leq \|\nabla \mathbf{v}\|_0.$$

2.2. Properties of FEM.

We impose the following properties on $\mathbb{V}_h, \mathbb{P}_h$.

Assumption A4 (discrete inf-sup). There exists a constant $\beta > 0$ such that

$$\inf_{q_h \in \mathbb{P}_h} \sup_{\mathbf{v}_h \in \mathbb{V}_h} \frac{\langle \text{div } \mathbf{v}_h\, ,\, q_h \rangle}{\|\mathbf{v}_h\|_1 \|q_h\|_0} \geq \beta.$$

Assumption A5 (shape regularity and quasi-uniformity [1, 2, 10]). There exists a constant $C > 0$ such that the ratio between the diameter h_K of an element $K \in \mathfrak{T}$ and the diameter of the largest ball contained in K is bounded uniformly by C, and h_K is comparable with the mesh size h for all $K \in \mathfrak{T}$.

Assumption A6 (approximability [1, 2, 10]). For each $(\mathbf{v}, q) \in \mathbf{H}^2(\Omega) \times H^1(\Omega)$, there exist approximations $(\mathbf{v}_h, q_h) \in \mathbb{V}_h \times \mathbb{P}_h$ such that

$$\|\mathbf{v} - \mathbf{v}_h\|_0 + h\|\mathbf{v} - \mathbf{v}_h\|_1 \leq Ch^2 \|\mathbf{v}\|_2 \quad \text{and} \quad \|q - q_h\|_0 \leq Ch\|q\|_1.$$

The low order accuracy of A6 is consistent with the regularity setting of A1–A3. A higher order FEM could be used as well, particularly so if $(\mathbf{u}, p)$ is sufficiently smooth, and in this case the space error estimates below would accordingly be of a higher order.

Now let $(\mathbf{v}_h, q_h) \in \mathbb{V}_h \times \mathbb{P}_h$ indicate the finite element solution of (2.1), namely,

$$\begin{aligned} \langle \nabla \mathbf{v}_h\, ,\, \nabla \mathbf{w}_h \rangle - \langle q_h\, ,\, \text{div } \mathbf{w}_h \rangle = \langle \mathbf{g}\, ,\, \mathbf{w}_h \rangle \quad & \forall \mathbf{w}_h \in \mathbb{V}_h, \\ \langle r_h\, ,\, \text{div } \mathbf{v}_h \rangle = 0 \qquad\qquad & \forall r_h \in \mathbb{P}_h. \end{aligned} \tag{2.6}$$

LEMMA 2.5 (error estimates for mixed FEM [1, 2, 10]). *Let* $(\mathbf{v}, q) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ *be the solutions of* (2.1) *and* $(\mathbf{v}_h, q_h) = \mathfrak{S}_h(\mathbf{v}, q) \in \mathbb{V}_h \times \mathbb{P}_h$ *be the Stokes projections defined by* (2.6), *respectively. If* A4–A6 *hold, then*

$$\|\mathbf{v} - \mathbf{v}_h\|_0 + h\|\mathbf{v} - \mathbf{v}_h\|_1 + h\|q - q_h\|_0 \leq Ch^2 \left(\|\mathbf{v}\|_2 + \|q\|_1\right). \tag{2.7}$$

If A1 *also holds, then the right-hand side is bounded by* $Ch^2\|\mathbf{g}\|_0$ *and*

$$\|\mathbf{g}\|_* \le C\|\nabla\mathbf{v}\|_0 \le Ch\|\mathbf{g}\|_0 + C\|\nabla\mathbf{v}_h\|_0, \tag{2.8}$$

$$|\!|\!|\mathbf{v}-\mathbf{v}_h|\!|\!| := \|\mathbf{v}-\mathbf{v}_h\|_{\mathbf{L}^\infty(\Omega)} + \|\nabla(\mathbf{v}-\mathbf{v}_h)\|_{\mathbf{L}^3(\Omega)} \le C\|\mathbf{g}\|_0. \tag{2.9}$$

Proof. Inequality (2.7) is standard [1, 2, 10]. To prove (2.8) we simply test (2.1) with an arbitrary $\mathbf{z} \in \mathbf{Z}(\Omega)$ for the first inequality, and next use (2.7) for the second one. To establish (2.9) we just deal with the L^∞-norm since the other can be treated similarly. If I_h denotes the Clement interpolant, then $\|\mathbf{v} - I_h\mathbf{v}\|_{\mathbf{L}^\infty(\Omega)} \le C\|\mathbf{v}\|_2$ and

$$\|I_h\mathbf{v}-\mathbf{v}_h\|_{\mathbf{L}^\infty(\Omega)} \le Ch^{-d/2}\|I_h\mathbf{v}-\mathbf{v}_h\|_{\mathbf{L}^2(\Omega)} \le C\|\mathbf{v}\|_2$$

as a consequence of an inverse estimate and (2.7). This completes the proof. ∎

Remark 2.6 (H^1 stability of q_h). The bound $\|\nabla q_h\|_0 \le C(\|\mathbf{v}\|_2 + \|q\|_1)$ is a simple by-product of (2.7). To see this, we add and subtract I_hq, use the stability of I_h in H^1, and observe that (2.7) implies $\|\nabla(q_h - I_hq)\|_0 \le Ch^{-1}\|q_h - I_hq\| \le C$.

We finally state without proof several properties of the nonlinear form $\mathfrak{N}_h$. In view of (1.7), we have the following properties of $\mathfrak{N}_h$ for all $\mathbf{u}_h, \mathbf{v}_h\mathbf{w}_h \in \mathbb{V}_h$:

$$\mathfrak{N}_h(\mathbf{u}_h,\mathbf{v}_h,\mathbf{w}_h) = -\mathfrak{N}_h(\mathbf{u}_h,\mathbf{w}_h,\mathbf{v}_h), \qquad \mathfrak{N}_h(\mathbf{u}_h,\mathbf{v}_h,\mathbf{v}_h) = 0 \tag{2.10}$$

and

$$\operatorname{div}\mathbf{u} = 0 \Rightarrow \mathfrak{N}_h(\mathbf{u},\mathbf{v}_h,\mathbf{w}_h) = \mathfrak{N}(\mathbf{u},\mathbf{v}_h,\mathbf{w}_h) = -\mathfrak{N}(\mathbf{u},\mathbf{w}_h,\mathbf{v}_h).$$

Applying Sobolev imbedding lemma yields the following useful results.

LEMMA 2.7 (bounds on nonlinear convection [11, 12]). *Let* $\mathbf{u}, \mathbf{v} \in \mathbf{H}^2(\Omega)$ *with* $\operatorname{div}\mathbf{u} = 0$*, and let* $\mathbf{u}_h, \mathbf{v}_h, \mathbf{w}_h \in \mathbb{V}_h$*. Then*

$$\mathfrak{N}_h(\mathbf{u},\mathbf{v}_h,\mathbf{w}_h) \le C\begin{cases}\|\mathbf{u}\|_1\|\mathbf{v}_h\|_1\|\mathbf{w}_h\|_1,\\ \|\mathbf{u}\|_2\|\nabla\mathbf{v}_h\|_0\|\mathbf{w}_h\|_0,\\ \|\mathbf{u}\|_2\|\mathbf{v}_h\|_0\|\nabla\mathbf{w}_h\|_0,\end{cases} \tag{2.11}$$

$$\mathfrak{N}_h(\mathbf{u}_h,\mathbf{v},\mathbf{w}_h) \le \|\mathbf{u}_h\|_0\|\mathbf{v}\|_2\|\nabla\mathbf{w}_h\|_0. \tag{2.12}$$

In addition,

$$\mathfrak{N}_h(\mathbf{u}_h,\mathbf{v}_h,\mathbf{w}_h) \le C\begin{cases}\|\mathbf{u}_h\|_0|\!|\!|\mathbf{v}_h|\!|\!|\|\nabla\mathbf{w}_h\|_0,\\ \|\mathbf{u}_h\|_{\mathbf{L}^3(\Omega)}\|\mathbf{v}_h\|_1\|\nabla\mathbf{w}_h\|_0.\end{cases} \tag{2.13}$$

3. Theorem 1.1: Stability.

In this section, we show that the gauge–Uzawa FEM is unconditionally stable via a standard energy method. We choose $\mathbf{w}_h = 2\tau\widehat{\mathbf{u}}_h^{n+1}$ in (1.8) and observe the following relation for the first term in (1.8):

$$\left\langle \widehat{\mathbf{u}}_h^{n+1} - \mathbf{u}_h^n\,,\, \widehat{\mathbf{u}}_h^{n+1}\right\rangle = \left\langle \mathbf{u}_h^{n+1} - \mathbf{u}_h^n\,,\, \mathbf{u}_h^{n+1}\right\rangle + \left\langle \nabla\rho_h^{n+1}\,,\, \nabla\rho_h^{n+1}\right\rangle$$

because of (1.11). Since the convection term vanishes from (2.10), we then obtain

$$\begin{aligned}\left\|\mathbf{u}_h^{n+1}\right\|_0^2 - \|\mathbf{u}_h^n\|_0^2 + \left\|\mathbf{u}_h^{n+1} - \mathbf{u}_h^n\right\|_0^2 + 2\left\|\nabla\rho_h^{n+1}\right\|_0^2 + 2\mu\tau\left\|\nabla\widehat{\mathbf{u}}_h^{n+1}\right\|_0^2 \\ = 2\mu\tau\left\langle s_h^n\,,\, \operatorname{div}\widehat{\mathbf{u}}_h^{n+1}\right\rangle + 2\tau\left\langle \mathbf{f}(t^{n+1})\,,\, \widehat{\mathbf{u}}_h^{n+1}\right\rangle.\end{aligned}$$

According to (1.10), we can write

$$2\left\langle s_h^n\,,\,\mathrm{div}\;\widehat{\mathbf{u}}_h^{n+1}\right\rangle = 2\left\langle s_h^n\,,\,s_h^n - s_h^{n+1}\right\rangle = \|s_h^n\|_0^2 - \left\|s_h^{n+1}\right\|_0^2 + \left\|s_h^n - s_h^{n+1}\right\|_0^2.$$

Combining now (1.10) with Lemma 2.4, we infer that $\left\|s_h^{n+1} - s_h^n\right\|_0 \le \left\|\mathrm{div}\;\widehat{\mathbf{u}}_h^{n+1}\right\|_0 \le \left\|\nabla\widehat{\mathbf{u}}_h^{n+1}\right\|$, whence

$$\begin{aligned}\left\|\mathbf{u}_h^{n+1}\right\|_0^2 - \|\mathbf{u}_h^n\|_0^2 + \left\|\mathbf{u}_h^{n+1} - \mathbf{u}_h^n\right\|_0^2 + 2\left\|\nabla\rho_h^{n+1}\right\|_0^2 + 2\mu\tau\left\|\nabla\widehat{\mathbf{u}}_h^{n+1}\right\|_0^2 \\ + \mu\tau\left\|s_h^{n+1}\right\|_0^2 - \mu\tau\|s_h^n\|_0^2 \le \frac{\tau}{2}\left\|\mathbf{f}(t^{n+1})\right\|_{-1}^2 + \frac{3\mu\tau}{2}\left\|\nabla\widehat{\mathbf{u}}_h^{n+1}\right\|_0^2.\end{aligned}$$

Adding over n from 0 to N, we obtain (1.15) and complete the proof of Theorem 1.1. It is important to notice that it is the *semi-implicit* treatment of convection in (1.8) that is responsible for unconditional stability.

4. Theorem 1.2: Error analysis for velocity.

In this section, we prove weak and strong error estimates for velocity for the gauge–Uzawa FEM of Algorithm 2. The proof is rather intricate because of the limited regularity of section 2.1, particularly that $\mathbf{u}_{tt} \notin L^2(0,T;\mathbf{L}^2(\Omega))$, and consists of three steps as follows.

- *Time-discrete Stokes.* We first consider a sequence of Stokes equations with the exact forcing and convection, namely, $\mathbf{U}^{n+1} \in \mathbf{H}_0^1(\Omega), P^{n+1} \in L_0^2(\Omega)$ satisfy $\mathbf{U}^0 = \mathbf{u}^0$ and

  (4.1)
  $$\delta\mathbf{U}^{n+1} - \mu\Delta\mathbf{U}^{n+1} + \nabla P^{n+1} = \mathbf{f}(t^{n+1}) - \big((\mathbf{u}\cdot\nabla)\mathbf{u}\big)(t^{n+1}), \qquad \mathrm{div}\;\mathbf{U}^{n+1} = 0.$$

  In Lemma 4.1, we derive estimates for the errors

  $$\mathbf{G}^{n+1} := \mathbf{u}(t^{n+1}) - \mathbf{U}^{n+1}, \qquad g^{n+1} := p(t^{n+1}) - P^{n+1},$$

  which rely solely on the regularity $\mathbf{u}_{tt} \in L^2([0:T]:\mathbf{Z}(\Omega)^*)$ of Lemma 2.3. This is possible because the test function $\mathbf{w} = \mathbf{u}(t^{n+1}) - \mathbf{U}^{n+1}$ is divergence free and thus allows us to work on the spaces $\mathbf{Z}(\Omega)$ and $\mathbf{Z}(\Omega)^*$.
- *Stokes projection.* We define $(\mathbf{U}_h^{n+1}, P_h^{n+1}) := \mathfrak{S}_h(\mathbf{u}(t^{n+1}), p(t^{n+1})) \in \mathbb{V}_h \times \mathbb{P}_h$ to be the Stokes projection of the true solution at time t^{n+1}, and derive error estimates in Lemma 4.3 for the errors

  $$\mathbf{G}_h^{n+1} := \mathbf{u}(t^{n+1}) - \mathbf{U}_h^{n+1}, \qquad g_h^{n+1} := p(t^{n+1}) - P_h^{n+1}.$$

  We point out that this choice of space discretization is more handy than discretizing (4.1) by finite elements, and still gives estimates for the errors $\mathbf{F}^{n+1} := \mathbf{U}^{n+1} - \mathbf{U}_h^{n+1}$ and $f^{n+1} := P^{n+1} - P_h^{n+1}$ by combining the first two steps.
- *Comparing* (4.1) *with* (1.8)–(1.11). We derive strong estimates of order 1/2 and use then to prove weak estimates of order 1 for the errors

  $$\text{(4.2)} \quad \mathbf{E}^{n+1} := \mathbf{U}^{n+1} - \mathbf{u}_h^{n+1}, \quad \widehat{\mathbf{E}}^{n+1} := \mathbf{U}^{n+1} - \widehat{\mathbf{u}}_h^{n+1}, \quad e^{n+1} := P^{n+1} - p_h^{n+1}.$$

  This is the most technical step; therefore, we now must deal with the fact that $\widehat{\mathbf{u}}_h^{n+1}$ is not divergence free, whereas $\mathbf{u}_h^{n+1}$ does not vanish on $\partial\Omega$; this is carried out in section 4.3. Upon combining the estimates of these three steps, we readily obtain Theorem 1.2.

4.1. Time-discrete Stokes problem. We now show error bounds for (4.1).

LEMMA 4.1 (uniform estimates). *Let* A1–A3 *hold. Then*

$$\left\|\mathbf{G}^{N+1}\right\|_0^2+\sum_{n=0}^{N}\left\|\mathbf{G}^{n+1}-\mathbf{G}^n\right\|_0^2+\mu\tau\sum_{n=0}^{N}\left\|\nabla\mathbf{G}^{n+1}\right\|_0^2\le C\tau^2, \tag{4.3}$$

$$\tau\sum_{n=0}^{N}\left\|g^{n+1}\right\|_0^2\le C\tau. \tag{4.4}$$

Proof. We subtract (4.1) from (1.1) at $t=t^{n+1}$ and thereby write

$$\delta\mathbf{G}^{n+1}-\mu\Delta\mathbf{G}^{n+1}+\nabla g^{n+1}=\mathbf{R}^{n+1}:=\frac{1}{\tau}\int_{t^n}^{t^{n+1}}(t-t^n)\mathbf{u}_{tt}(\cdot,t)\,dt, \tag{4.5}$$

where $\mathbf{R}^{n+1}$ is the truncation error. We multiply this elliptic PDE by the admissible test function $2\tau\mathbf{G}^{n+1}\in\mathbf{Z}(\Omega)$ to arrive at

$$\left\|\mathbf{G}^{n+1}\right\|_0^2-\left\|\mathbf{G}^n\right\|_0^2+\left\|\mathbf{G}^{n+1}-\mathbf{G}^n\right\|_0^2+2\mu\tau\left\|\nabla\mathbf{G}^{n+1}\right\|_0^2\le 2\tau\left\|\mathbf{R}^{n+1}\right\|_*\left\|\nabla\mathbf{G}^{n+1}\right\|_0.$$

Adding over n and using (2.5) yield (4.3). To prove (4.4) we use the error equation (4.5) to obtain any $\mathbf{w}\in\mathbf{H}_0^1(\Omega)$

$$\left\langle g^{n+1}\,,\,\mathrm{div}\,\mathbf{w}\right\rangle\le\frac{1}{\tau}\left\|\mathbf{G}^{n+1}-\mathbf{G}^n\right\|_0\|\mathbf{w}\|_0+\mu\left\|\nabla\mathbf{G}^{n+1}\right\|_0\|\nabla\mathbf{w}\|_0+\left\|\mathbf{R}^{n+1}\right\|_0\|\mathbf{w}\|_0.$$

Since $\left\|\mathbf{R}^{n+1}\right\|_0^2\le\frac{1}{2}\int_{t^n}^{t^{n+1}}\sigma\|\mathbf{u}_{tt}\|_0^2$, (2.3) and (4.3) together with the continuous inf-sup condition imply (4.4). ∎

LEMMA 4.2 (weighted estimates). *Let* A1–A3 *hold. Then*

$$\sigma^{N+1}\left\|\delta\mathbf{G}^{N+1}\right\|_0^2+\sum_{n=1}^{N}\sigma^{n+1}\left\|\delta\mathbf{G}^{n+1}-\delta\mathbf{G}^n\right\|_0^2+\frac{\mu\tau}{2}\sum_{n=1}^{N}\sigma^{n+1}\left\|\nabla\delta\mathbf{G}^{n+1}\right\|_0^2\le C\tau, \tag{4.6}$$

$$\sup_{0\le n\le N+1}\sigma^n\|g^n\|_0^2+\sum_{n=0}^{N}\sigma^{n+1}\left(\left\|g^{n+1}\right\|_0^2+\left\|\delta g^{n+1}\right\|_0^2\right)\le C\tau. \tag{4.7}$$

If NLC is also valid, then (4.6) *and* (4.7) *become uniform, namely, without weights.*

Proof. To prove (4.6) we subtract two consecutive equations (4.5) and thus derive an equation for $\delta\mathbf{G}^{n+1}$. We next multiply this equation by $2\sigma^{n+1}\delta\mathbf{G}^{n+1}$ and proceed as in Lemma 4.1 to discover that $I^{n+1}:=2\sigma^{n+1}\left\langle\delta(\mathbf{G}^{n+1}-\mathbf{G}^n)\,,\,\delta\mathbf{G}^{n+1}\right\rangle$ and $II^{n+1}:=2\tau\sigma^{n+1}\left\langle\delta\mathbf{R}^{n+1}\,,\,\delta\mathbf{G}^{n+1}\right\rangle$ must be estimated. We see that

$$I^{n+1}=\sigma^{n+1}\left\|\delta\mathbf{G}^{n+1}\right\|_0^2-\sigma^n\left\|\delta\mathbf{G}^n\right\|_0^2+\sigma^{n+1}\left\|\delta\mathbf{G}^{n+1}-\delta\mathbf{G}^n\right\|_0^2-(\sigma^{n+1}-\sigma^n)\|\delta\mathbf{G}^n\|_0^2$$

and realize that, upon summation over n, the first two terms on the right-hand side telescope, whereas the last one leads to $\frac{1}{\tau}\sum_{n=1}^{N}\left\|\mathbf{G}^{n+1}-\mathbf{G}^n\right\|_0^2\le C\tau$ in view of (4.3). On the other hand, II^{n+1} can be written equivalently as follows:

$$\begin{aligned}II^{n+1}&=2\sigma^{n+1}\left\langle\mathbf{R}^{n+1}\,,\,\delta\mathbf{G}^{n+1}\right\rangle-2\sigma^n\left\langle\mathbf{R}^n\,,\,\delta\mathbf{G}^n\right\rangle\\&\quad+2\sigma^n\left\langle\mathbf{R}^n\,,\,\delta\mathbf{G}^n-\delta\mathbf{G}^{n+1}\right\rangle+2(\sigma^n-\sigma^{n+1})\left\langle\mathbf{R}^n\,,\,\delta\mathbf{G}^{n+1}\right\rangle.\end{aligned}$$

We now add on n and observe that the first two terms telescope. The third term can be handled via the estimate $\sum_{n=1}^{N}\sigma^n\|\mathbf{R}^n\|_0^2 \le C\tau\int_0^T\sigma\|\mathbf{u}_{tt}\|_0^2 \le C\tau$, which results from (2.3), together with the bound for $\sum_{n=1}^{N} I^{n+1}$. Using again $\sum_{n=1}^{N}\sigma^n\|\mathbf{R}^n\|_0^2 \le C\tau$, now coupled with $\sum_{n=1}^{N}\|\delta\mathbf{G}^n\|_0^2 \le C$ from (4.3), takes care of the last term in II^{n+1}.

We finally observe that the presence of weights allows us to employ regularity (2.3) for $\mathbf{u}_{tt}$. If we further assume NLC, then we could omit weights and instead resort to regularity (2.4) to establish uniform bounds. This completes the proof. □

4.2. Stokes projection. We now establish simple estimates for $(\mathbf{G}_h^{n+1}, g_h^{n+1})$.

LEMMA 4.3 (Stokes projection). *Let* A1–A6 *hold. Then*

$$\big\|\mathbf{G}_h^{n+1}\big\|_0 + h\big\|\mathbf{G}_h^{n+1}\big\|_1 + h\big\|g_h^{n+1}\big\|_0 \le Ch^2, \tag{4.8}$$

$$\tau\sum_{n=0}^{N}\sigma^{n+1}\Big(\big\|\delta\mathbf{G}_h^{n+1}\big\|_0^2 + h^2\big\|\delta\mathbf{G}_h^{n+1}\big\|_1^2 + h^2\big\|\delta g_h^{n+1}\big\|_0^2\Big) \le Ch^4. \tag{4.9}$$

If NLC also holds, then (4.9) *becomes* uniform*, namely, without weights.*

Proof. Estimate (4.8) is a direct consequence of Lemma 2.5 and (2.2). Since the Stokes operator $\mathfrak{S}_h$ is linear, we readily have $(\delta\mathbf{U}_h^n, \delta P_h^n) = \mathfrak{S}_h(\delta\mathbf{u}(t^n), \delta p(t^n))$, and Lemma 2.5 applies again. Upon multiplying by $\tau\sigma^{n+1}$, the square of the right-hand side of (2.7) can be bounded by

$$h^4\tau^{-1}\sum_{n=0}^{N}\sigma^{n+1}\Big(\big\|\mathbf{u}(t^{n+1}) - \mathbf{u}(t^n)\big\|_2^2 + \big\|p(t^{n+1}) - p(t^n)\big\|_1^2\Big).$$

We examine the velocity term only since the other one is similar. For $n = 0$ we recall (2.2), along with $\sigma^1 = \tau$, to write $\sigma^1\big\|\mathbf{u}(t^1) - \mathbf{u}(t^0)\big\|_2^2 \le C\tau$. For $n \ge 1$, instead, we use that $\sigma^{n+1} \le 2\sigma(t)$ for $t^n \le t \le t^{n+1}$, whence

$$\sum_{n=1}^{N}\sigma^{n+1}\big\|\mathbf{u}(t^{n+1}) - \mathbf{u}(t^n)\big\|_2^2 \le C\tau\int_0^T\sigma\|\mathbf{u}_t\|_2^2 \le C\tau$$

because of (2.3). This completes the proof. □

4.3. Comparing (4.1) with (1.8)–(1.11). We derive strong estimates of order 1/2 and use them to prove weak estimates of order 1 for the errors in (4.2), namely,

$$\mathbf{E}^{n+1} = \mathbf{U}^{n+1} - \mathbf{u}_h^{n+1}, \quad \widehat{\mathbf{E}}^{n+1} = \mathbf{U}^{n+1} - \widehat{\mathbf{u}}_h^{n+1}, \quad e^{n+1} = P^{n+1} - p_h^{n+1}.$$

Before embarking on this discussion, we mention several useful properties of the error functions. If $\mathbf{E}_h^{n+1} := \mathbf{U}_h^{n+1} - \mathbf{u}_h^{n+1}$, $\widehat{\mathbf{E}}_h^{n+1} := \mathbf{U}_h^{n+1} - \widehat{\mathbf{u}}_h^{n+1}$, and $\mathbf{F}^{n+1} = \mathbf{U}^{n+1} - \mathbf{U}_h^{n+1}$, then

$$\widehat{\mathbf{E}}^{n+1} = \mathbf{E}^{n+1} + \nabla\rho_h^{n+1}, \qquad \widehat{\mathbf{E}}_h^{n+1} = \mathbf{E}_h^{n+1} + \nabla\rho_h^{n+1},$$
$$\widehat{\mathbf{E}}^{n+1} = \mathbf{F}^{n+1} + \widehat{\mathbf{E}}_h^{n+1}, \qquad \mathbf{E}^{n+1} = \mathbf{F}^{n+1} + \mathbf{E}_h^{n+1},$$

as well as

$$\big\langle\mathbf{E}^{n+1}, \nabla q_h\big\rangle = \big\langle\mathbf{E}_h^{n+1}, \nabla q_h\big\rangle = \big\langle\mathbf{F}^{n+1}, \nabla q_h\big\rangle = 0 \qquad \forall q_h \in \mathbb{P}_h, \tag{4.10}$$

whence we deduce crucial orthogonality properties:

$$\|\widehat{\mathbf{E}}^{n+1}\|_0^2 = \|\mathbf{E}^{n+1}\|_0^2 + \|\nabla\rho_h^{n+1}\|_0^2, \qquad \|\widehat{\mathbf{E}}_h^{n+1}\|_0^2 = \|\mathbf{E}_h^{n+1}\|_0^2 + \|\nabla\rho_h^{n+1}\|_0^2. \tag{4.11}$$

Since $\mathbf{F}^{n+1} = \mathbf{G}_h^{n+1} - \mathbf{G}^{n+1}, f^{n+1} = g_h^{n+1} - g^{n+1}$, Lemmas 4.1 and 4.3 give rise to the following estimates provided A1–A6 hold:

$$\|\mathbf{F}^{n+1}\|_0^2 \le C(\tau^2 + h^4), \quad \mu\tau \sum_{n=1}^{N} \|\nabla \mathbf{F}^{n+1}\|_0^2 \le C(\tau^2 + h^2),$$
$$\tau \sum_{n=1}^{N} \|f^{n+1}\|_0^2 \le C(\tau + h^2). \tag{4.12}$$

We also point out that, owing to Lemma 2.4, $s_h^{n+1} \in \mathbb{P}_h$ defined in (1.10) satisfies

$$\left\|s_h^{n+1} - s_h^n\right\|_0 \le \|\nabla \widehat{\mathbf{E}}^{n+1}\|_0. \tag{4.13}$$

LEMMA 4.4 (reduced rate of convergence for velocity). *Let* A1–A6 *and* $h^2 \le C\tau$ *be valid with an arbitrary constant* $C > 0$. *Then the velocity error functions satisfy*

$$\|\mathbf{E}^{N+1}\|_0^2 + \|\widehat{\mathbf{E}}^{N+1}\|_0^2 + \mu\tau\left\|s_h^{N+1}\right\|_0^2 + \frac{1}{2}\sum_{n=0}^{N}\left\|\mathbf{E}^{n+1} - \mathbf{E}^n\right\|_0^2$$
$$+ \sum_{n=0}^{N}\left\|\nabla\rho_h^{n+1}\right\|_0^2 + \frac{\mu\tau}{2}\sum_{n=0}^{N}\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2 \le C(\tau + h^2). \tag{4.14}$$

Proof. Subtracting (1.8) from (4.1) yields, for all $\mathbf{w}_h \in \mathbb{V}_h$,

$$\tau^{-1}\left\langle \widehat{\mathbf{E}}^{n+1} - \mathbf{E}^n\, , \, \mathbf{w}_h\right\rangle + \mu\left\langle \nabla\widehat{\mathbf{E}}^{n+1}\, , \, \nabla\mathbf{w}_h\right\rangle = \left\langle P^{n+1}\, , \, \mathrm{div}\ \mathbf{w}_h\right\rangle \tag{4.15}$$
$$-\mu\left\langle s_h^n\, , \, \mathrm{div}\ \mathbf{w}_h\right\rangle - \mathfrak{N}_h(\mathbf{u}(t^{n+1}), \mathbf{u}(t^{n+1}), \mathbf{w}_h) + \mathfrak{N}_h(\mathbf{u}_h^n, \widehat{\mathbf{u}}_h^{n+1}, \mathbf{w}_h).$$

Choosing $\mathbf{w}_h = 2\tau\widehat{\mathbf{E}}_h^{n+1} = 2\tau(\widehat{\mathbf{E}}^{n+1} - \mathbf{F}^{n+1})$ in (4.15) and using (4.10), we easily get

$$\left\|\mathbf{E}^{n+1}\right\|_0^2 - \|\mathbf{E}^n\|_0^2 + \left\|\mathbf{E}^{n+1} - \mathbf{E}^n\right\|_0^2 + 2\mu\tau\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2 + 2\left\|\nabla\rho_h^{n+1}\right\|_0^2 = \sum_{i=1}^{4} A_i \tag{4.16}$$

with

$$A_1 := 2\left\langle \mathbf{E}^{n+1} - \mathbf{E}^n\, , \, \mathbf{F}^{n+1}\right\rangle + 2\mu\tau\left\langle \nabla\widehat{\mathbf{E}}^{n+1}\, , \, \nabla\mathbf{F}^{n+1}\right\rangle,$$
$$A_2 := 2\tau\left\langle P^{n+1}\, , \, \mathrm{div}\ \widehat{\mathbf{E}}_h^{n+1}\right\rangle,$$
$$A_3 := -2\tau\left(\mathfrak{N}_h(\mathbf{u}(t^{n+1}), \mathbf{u}(t^{n+1}), \widehat{\mathbf{E}}_h^{n+1}) - \mathfrak{N}_h(\mathbf{u}_h^n, \widehat{\mathbf{u}}_h^{n+1}, \widehat{\mathbf{E}}_h^{n+1})\right),$$
$$A_4 := -2\mu\tau\left\langle s_h^n\, , \, \mathrm{div}\ \widehat{\mathbf{E}}_h^{n+1}\right\rangle.$$

We now estimate each term A_i separately. Applying the Hölder inequality, we find a bound of the first term

$$A_1 \le \frac{1}{2}\left\|\mathbf{E}^{n+1} - \mathbf{E}^n\right\|_0^2 + C\left\|\mathbf{F}^{n+1}\right\|_0^2 + \frac{\mu\tau}{4}\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2 + C\mu\tau\left\|\nabla\mathbf{F}^{n+1}\right\|_0^2. \tag{4.17}$$

Since $\mathbf{U}_h^{n+1}$ is discrete divergence free, but not so $\widehat{\mathbf{u}}_h^{n+1}$, we add and subtract P_h^{n+1} and $p(t^{n+1})$, and recall (1.9) and Remark 2.6 to derive

$$A_2 = 2\tau\left\langle f^{n+1}\, , \, \mathrm{div}\ \widehat{\mathbf{E}}_h^{n+1}\right\rangle + 2\tau\left\langle \nabla g_h^{n+1}\, , \, \nabla\rho_h^{n+1}\right\rangle - 2\tau\left\langle \nabla p(t^{n+1})\, , \, \nabla\rho_h^{n+1}\right\rangle \tag{4.18}$$
$$\le C\tau^2 B^{n+1} + \frac{C\tau}{\mu}\left\|f^{n+1}\right\|_0^2 + \frac{\mu\tau}{8}\left(\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2 + \left\|\nabla\mathbf{F}^{n+1}\right\|_0^2\right) + \left\|\nabla\rho_h^{n+1}\right\|_0^2,$$

where $B^{n+1} := \left\|\mathbf{u}(t^{n+1})\right\|_2^2 + \left\|\nabla p(t^{n+1})\right\|_0^2$. To tackle A_3 we first add and subtract $\mathbf{u}(t^{n+1}), \mathbf{u}_h^n$, and realize that $\mathfrak{N}_h(\mathbf{u}_h^n, \widehat{\mathbf{E}}_h^{n+1}, \widehat{\mathbf{E}}_h^{n+1}) = 0$ according to (2.10). This yields

$$\begin{aligned}A_3 = &- 2\tau\mathfrak{N}_h(\mathbf{u}(t^{n+1}) - \mathbf{u}(t^n), \mathbf{u}(t^{n+1}), \widehat{\mathbf{E}}_h^{n+1})\\ &- 2\tau\mathfrak{N}_h(\mathbf{u}(t^n) - \mathbf{u}_h^n, \mathbf{u}(t^{n+1}), \widehat{\mathbf{E}}_h^{n+1}) - 2\tau\mathfrak{N}_h(\mathbf{u}_h^n, \mathbf{G}_h^{n+1}, \widehat{\mathbf{E}}_h^{n+1}).\end{aligned}$$

Since $\left\|\mathbf{u}(t^{n+1})\right\|_2 + \left|\!\left|\!\left|\mathbf{G}_h^{n+1}\right|\!\right|\!\right| \le C$ in view of (2.2) and (2.9), and $\widehat{\mathbf{E}}_h^{n+1} = \widehat{\mathbf{E}}^{n+1} - \mathbf{F}^{n+1}$, (2.11) and (2.13) give

$$A_3 \le \frac{C\tau^2}{\mu} D^{n+1} + \frac{C\tau}{\mu}\left(\|\mathbf{E}^n\|_0^2 + \|\mathbf{G}^n\|_0^2 + \left\|\mathbf{G}_h^{n+1}\right\|_0^2\right) + \frac{\mu\tau}{8}\|\nabla\widehat{\mathbf{F}}^{n+1}\|_0^2 + \frac{\mu\tau}{8}\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2$$

with $D^{n+1} := \int_{t^n}^{t^{n+1}} \|\mathbf{u}_t(t)\|_0^2\, dt$. Next, making use of (1.10) and (4.13), we arrive at

$$\begin{aligned}A_4 &= 2\mu\tau\left\langle s_h^n\,,\, \operatorname{div}\, \widehat{\mathbf{u}}_h^{n+1}\right\rangle = 2\mu\tau\left\langle s_h^n - s_h^{n+1}\,,\, s_h^n\right\rangle\\ &\le \mu\tau\left(\|s_h^n\|_0^2 - \left\|s_h^{n+1}\right\|_0^2\right) + \mu\tau\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2.\end{aligned}$$

Inserting the above estimates into (4.16), summing over n from 0 to N gives

$$\begin{aligned}\text{(4.19)}\qquad &\left\|\mathbf{E}^{N+1}\right\|_0^2 + \frac{1}{2}\sum_{n=0}^{N}\left\|\mathbf{E}^{n+1} - \mathbf{E}^n\right\|_0^2 + \frac{\mu\tau}{2}\sum_{n=0}^{N}\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2\\ &+ \mu\tau\left\|s_h^{N+1}\right\|_0^2 + \sum_{n=0}^{N}\left\|\nabla\rho_h^{n+1}\right\|_0^2 \le C(\tau + h^2) + \frac{C\tau}{\mu}\sum_{n=0}^{N}\|\mathbf{E}^n\|_0^2,\end{aligned}$$

where we have used (2.2) to bound B^{n+1}, D^{n+1}, together with (4.3) and (4.8) to estimate $\|\mathbf{G}^n\|_0$ and $\|\mathbf{G}_h^{n+1}\|_0$, respectively, and (4.12) as well as $h^2 \le C\tau$ to bound $\|\mathbf{F}^{n+1}\|_0, \|\mathbf{F}^{n+1}\|_0$, and $\|f^{n+1}\|_0$. The discrete Gronwall lemma finally yields (4.14) except for $\|\widehat{\mathbf{E}}^{n+1}\|_0^2$. The latter results from (4.11) and completes the proof. □

Remark 4.5 (initial errors). If $N = 0$ in (4.19), then Lemmas 4.1 and 4.3 give

$$\begin{aligned}&\left\|\mathbf{E}^1\right\|_0^2 + \frac{1}{2}\left\|\mathbf{E}^1 - \mathbf{E}^0\right\|_0^2 + \frac{\mu\tau}{2}\|\nabla\widehat{\mathbf{E}}^1\|_0^2 + \mu\tau\left\|s_h^1\right\|_0^2 + \left\|\nabla\rho_h^1\right\|_0^2\\ &\quad\le C(\tau^2 + \tau h^2 + h^4) + \frac{C\tau}{\mu}\left\|f^1\right\|_0^2 \le C(\tau + \tau h^2 + h^4)\end{aligned}$$

or $\le C(\tau^2 + \tau h^2 + h^4)$ provided NLC holds in conjunction with (4.7).

Remark 4.6 (suboptimal order). The suboptimal order $\mathcal{O}(\tau + h^2)$ of Lemma 4.4 is due to terms $\|\mathbf{F}^{n+1}\|_0^2 + \tau\|\nabla\mathbf{F}^{n+1}\|_0^2$ in (4.17) and the fact that $\widehat{\mathbf{E}}_h^{n+1}$ in (4.18) is not discrete divergence free. To improve upon this we must get rid of both terms.

Lemma 4.7 (full rate of convergence for velocity). *Let* A1–A6 *hold and* $h^2 \le C\tau$ *be valid with an arbitrary constant* $C > 0$. *Then we have*

$$\text{(4.20)}\qquad \left\|\mathbf{E}^{N+1}\right\|_*^2 + \sum_{n=0}^{N}\left\|\mathbf{E}^{n+1} - \mathbf{E}^n\right\|_*^2 + \left(\mu\tau\left\|\mathbf{E}^{n+1}\right\|_0^2 + \|\widehat{\mathbf{E}}^{n+1}\|_0^2\right) \le C(\tau^2 + h^4).$$

Proof. Let $(\mathbf{v}^n, q^n)$ and $(\mathbf{v}_h^n, q_h^n)$ be solutions of the Stokes equations (2.1) and (2.6) with $\mathbf{g} = \mathbf{E}^n$. Then Lemma 2.5 and A1 yield a crucial inequality

$$\|\mathbf{v}^n - \mathbf{v}_h^n\|_0 + h\|\mathbf{v}^n - \mathbf{v}_h^n\|_1 + h\|q^n - q_h^n\|_0 \le Ch^2\|\mathbf{E}^n\|_0. \tag{4.21}$$

Since $\mathbf{v}_h^{n+1}$ is discrete divergence free, then $\langle \nabla\rho_h^{n+1}, \mathbf{v}_h^{n+1}\rangle = 0$ and

$$\left\langle \widehat{\mathbf{E}}^{n+1} - \mathbf{E}^n , \mathbf{v}_h^{n+1} \right\rangle = \langle \mathbf{E}^{n+1} - \mathbf{E}^n , \mathbf{v}_h^{n+1}\rangle = \langle \nabla(\mathbf{v}_h^{n+1} - \mathbf{v}_h^n) , \nabla\mathbf{v}_h^{n+1}\rangle .$$

Choosing $\mathbf{w}_h = 2\tau\mathbf{v}_h^{n+1}$ in (4.15) yields

$$\left\|\nabla\mathbf{v}_h^{n+1}\right\|_0^2 - \|\nabla\mathbf{v}_h^n\|_0^2 + \left\|\nabla(\mathbf{v}_h^{n+1} - \mathbf{v}_h^n)\right\|_0^2 + 2\mu\tau\left\|\mathbf{E}^{n+1}\right\|_0^2 = \sum_{i=1}^4 A_i \tag{4.22}$$

with

$$\begin{aligned}
A_1 &:= -2\mu\tau\langle\nabla\mathbf{F}^{n+1} , \nabla\mathbf{v}_h^{n+1}\rangle,\\
A_2 &:= 2\mu\tau(\langle\mathbf{F}^{n+1} , \mathbf{E}^{n+1}\rangle + \langle\nabla\rho_h^{n+1} , \nabla q_h^{n+1}\rangle),\\
A_3 &:= 2\tau\langle P^{n+1} , \operatorname{div} \mathbf{v}_h^{n+1}\rangle,\\
A_4 &:= -2\tau(\mathfrak{N}_h(\mathbf{u}(t^{n+1}), \mathbf{u}(t^{n+1}), \mathbf{v}_h^{n+1}) - \mathfrak{N}_h(\mathbf{u}_h^n, \widehat{\mathbf{u}}_h^{n+1}, \mathbf{v}_h^{n+1})).
\end{aligned}$$

We now estimate A_1–A_4 separately. We use inequality (4.21) to get

$$\begin{aligned}
A_1 &= 2\mu\tau\left\langle\nabla\mathbf{F}^{n+1} , \nabla\left(\mathbf{v}^{n+1} - \mathbf{v}_h^{n+1}\right) - \nabla\mathbf{v}^{n+1}\right\rangle\\
&\le C\mu\tau\left(h^2\left\|\nabla\mathbf{F}^{n+1}\right\|_0^2 + \left\|\mathbf{F}^{n+1}\right\|_0^2\right) + \frac{\mu\tau}{6}\left\|\mathbf{E}^{n+1}\right\|_0^2
\end{aligned}$$

as well as

$$A_2 \le C\mu\tau\left(\left\|\mathbf{F}^{n+1}\right\|_0^2 + \left\|\nabla\rho_h^{n+1}\right\|_0^2\right) + \frac{\mu\tau}{6}\left\|\mathbf{E}^{n+1}\right\|_0^2.$$

We next use the fact that $\mathbf{v}_h^{n+1}$ is discrete divergence free and $\mathbf{v}^{n+1}$ is divergence free. Hence

$$\begin{aligned}
A_3 &= 2\tau\left\langle P^{n+1} - P_h^{n+1} , \operatorname{div}\left(\mathbf{v}_h^{n+1} - \mathbf{v}^{n+1}\right)\right\rangle\\
&\le C\tau h\left\|f^{n+1}\right\|_0\left\|\mathbf{v}^{n+1}\right\|_2 \le \frac{C\tau h^2}{\mu}\left\|f^{n+1}\right\|_0^2 + \frac{\mu\tau}{6}\left\|\mathbf{E}^{n+1}\right\|_0^2.
\end{aligned}$$

At the same time, the convection term A_4 can be rewritten as $A_4 = \sum_{i=1}^3 A_{4,i}$ with

$$\begin{aligned}
A_{4,1} &:= -2\tau\mathfrak{N}_h((\mathbf{u}(t^{n+1}) - \mathbf{u}(t^n)) + (\mathbf{u}(t^n) - \mathbf{u}_h^n), \mathbf{u}(t^{n+1}), \mathbf{v}_h^{n+1}),\\
A_{4,2} &:= 2\tau\mathfrak{N}_h(\mathbf{u}(t^n) - \mathbf{u}_h^n, \mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}, (\mathbf{v}_h^{n+1} - \mathbf{v}^{n+1}) + \mathbf{v}^{n+1}),\\
A_{4,3} &:= -2\tau\mathfrak{N}_h(\mathbf{u}(t^n), \mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}, \mathbf{v}_h^{n+1}).
\end{aligned}$$

Since $\mathbf{u}(t^n) - \mathbf{u}_h^n = \mathbf{E}^n + \mathbf{G}^n$, (2.2) in conjunction with (2.12) yields

$$A_{4,1} \le C\tau^2\mu\int_{t^n}^{t^{n+1}} \|\mathbf{u}_t(t)\|_0^2\, dt + \frac{\mu\tau}{6}\left(\|\mathbf{E}^n\|_0^2 + \|\mathbf{G}^n\|_0^2\right) + \frac{C\tau}{\mu}\left\|\nabla\mathbf{v}_h^{n+1}\right\|_0^2.$$

Before tacking $A_{4,2}$ we observe that (4.3) and (4.14) imply $\|\mathbf{u}(t^n)-\mathbf{u}_h^n\|_0 \le C(h+\tau^{1/2})$, and that (2.9) and (4.21) yield

$$\left\|\left\|\mathbf{v}_h^{n+1}-\mathbf{v}^{n+1}\right\|\right\| \le C\|\mathbf{v}^{n+1}\|_2 \le C\|\mathbf{E}^{n+1}\|_0.$$

Therefore, (2.12) and (2.13) lead to

$$A_{4,2} \le \frac{C\tau}{\mu}\left(\tau+h^2\right)\left(\left\|\nabla\mathbf{G}^{n+1}\right\|_0^2+\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2\right)+\frac{\mu\tau}{6}\left\|\mathbf{E}^{n+1}\right\|_0^2.$$

Since $\mathbf{u}(t^n)$ is divergence free, we can resort to (2.11) and (4.11) to obtain

$$A_{4,3} \le \frac{\mu\tau}{6}\left(\left\|\mathbf{E}^{n+1}\right\|_0^2+\left\|\nabla\rho_h^{n+1}\right\|_0^2+\left\|\mathbf{G}^{n+1}\right\|_0^2\right)+\frac{C\tau}{\mu}\left\|\nabla\mathbf{v}_h^{n+1}\right\|_0^2.$$

Inserting the above estimates into (4.22) and summing over n from 0 to N, we deduce

$$\begin{aligned}(4.23)\qquad &\left\|\nabla\mathbf{v}_h^{N+1}\right\|_0^2+\sum_{n=0}^{N}\left\|\nabla(\mathbf{v}_h^{n+1}-\mathbf{v}_h^n)\right\|_0^2+\mu\tau\sum_{n=0}^{N}\left\|\mathbf{E}^{n+1}\right\|_0^2\\ &\le C\left(\tau^2+h^4\right)+\frac{C\tau}{\mu}\sum_{n=0}^{N}\left\|\nabla\mathbf{v}_h^{n+1}\right\|_0^2\end{aligned}$$

because of (4.3), (4.12), and (4.14) bound the remaining terms. The discrete Gronwall lemma and (2.8) allow us to remove the rightmost term in (4.23), and thereby arrive at (4.20) upon invoking (2.8). However, this does not give a bound for $\|\widehat{\mathbf{E}}^{n+1}\|_0$, which comes from (4.11) and (4.14) instead. The proof is thus complete. □

Proof of Theorem 1.2. This is a consequence of Lemmas 4.1, 4.4, and 4.7. □

Remark 4.8 (estimates for $\left\|\nabla\mathbf{v}_h^1\right\|_0$). These estimates will be crucial in section 5 and can be extracted from (4.23) upon invoking Remark 4.5 and choosing $N=0$. Since $\mathbf{v}_h^0=0$ because $\mathbf{E}^0$ is orthogonal to $\mathbb{V}_h$, (4.23) reduces to

$$\left\|\nabla\mathbf{v}_h^1\right\|_0^2 \le C\tau(\tau^2+h^4)+\frac{C\tau h^2}{\mu}\left\|f^1\right\|_0^2 \le C\tau(\tau^2+h^2).$$

On the other hand, if NLC is also valid, then $\|f^1\|_0^2 \le C\tau$ and $\left\|\nabla\mathbf{v}_h^1\right\|_0^2 \le C\tau(\tau^2+h^4)$.

5. Theorem 1.3: Error analysis for time derivative of velocity.

In this section we embark on an error analysis for the time derivative of velocity.

LEMMA 5.1 (stability of time-derivative of velocity). *Let* A1–A6 *hold and* $h^2 \le C_1h^2 \le \tau \le C_2h^{\frac{d}{3}(1+\varepsilon)}$ *be valid with arbitrary constants* $C_1, C_2 > 0$. *Then the error functions satisfy the weighted estimates*

$$\begin{aligned}(5.1)\qquad &\sigma^{N+1}\left\|\delta\mathbf{E}^{N+1}\right\|_0^2+\sum_{n=1}^{N}\sigma^{n+1}\left\|\delta\mathbf{E}^{n+1}-\delta\mathbf{E}^n\right\|_0^2+\sum_{n=1}^{N}\sigma^{n+1}\left\|\nabla\delta\rho_h^{n+1}\right\|_0^2\\ &+\mu\tau\sigma^{N+1}\left\|\delta s_h^{N+1}\right\|_0^2+\mu\tau\sum_{n=1}^{N}\sigma^{n+1}\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\|_0^2 \le C.\end{aligned}$$

If NLC is also valid, then (5.1) *becomes uniform, namely, without weights.*

Proof. Subtracting two consecutive expressions (4.15) yields

$$
\begin{aligned}
(5.2)\qquad &\left\langle \delta\widehat{\mathbf{E}}^{n+1} - \delta\mathbf{E}^{n} \,,\, \mathbf{w}_h \right\rangle + \mu\tau \left\langle \nabla\delta\widehat{\mathbf{E}}^{n+1} \,,\, \nabla\mathbf{w}_h \right\rangle \\
&\quad = \tau \left\langle \delta P^{n+1} \,,\, \operatorname{div} \mathbf{w}_h \right\rangle - \mu\tau \left\langle \delta s_h^{n} \,,\, \operatorname{div} \mathbf{w}_h \right\rangle \\
&\qquad - \mathfrak{N}_h(\mathbf{u}(t^{n+1}), \mathbf{u}(t^{n+1}), \mathbf{w}_h) + \mathfrak{N}_h(\mathbf{u}_h^{n}, \widehat{\mathbf{u}}_h^{n+1}, \mathbf{w}_h) \\
&\qquad + \mathfrak{N}_h(\mathbf{u}(t^{n}), \mathbf{u}(t^{n}), \mathbf{w}_h) - \mathfrak{N}_h(\mathbf{u}_h^{n-1}, \widehat{\mathbf{u}}_h^{n}, \mathbf{w}_h).
\end{aligned}
$$

Choosing $\mathbf{w}_h = 2\delta\widehat{\mathbf{E}}_h^{n+1} = 2\delta(\widehat{\mathbf{E}}^{n+1} - \mathbf{F}^{n+1})$ in (5.2) and using (4.10) implies

$$
\begin{aligned}
(5.3)\qquad &\left\|\delta\mathbf{E}^{n+1}\right\|_0^2 - \left\|\delta\mathbf{E}^{n}\right\|_0^2 + \left\|\delta\mathbf{E}^{n+1} - \delta\mathbf{E}^{n}\right\|_0^2 \\
&\quad + 2\left\|\nabla\delta\rho_h^{n+1}\right\|_0^2 + 2\mu\tau\left\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\right\|_0^2 = \sum_{i=1}^{4} A_i
\end{aligned}
$$

with

$$
\begin{aligned}
A_1 &:= 2\langle \delta\mathbf{E}^{n+1} - \delta\mathbf{E}^{n} \,,\, \delta\mathbf{F}^{n+1}\rangle + 2\mu\tau\langle \nabla\delta\widehat{\mathbf{E}}^{n+1} \,,\, \nabla\delta\mathbf{F}^{n+1}\rangle, \\
A_2 &:= 2\tau\left\langle \delta P^{n+1} \,,\, \operatorname{div} \delta\widehat{\mathbf{E}}_h^{n+1}\right\rangle, \\
A_3 &:= -2\mu\tau\left\langle \delta s_h^{n} \,,\, \operatorname{div} \delta\widehat{\mathbf{E}}_h^{n+1}\right\rangle, \\
A_4 &:= -2\mathfrak{N}_h(\mathbf{u}(t^{n+1}), \mathbf{u}(t^{n+1}), \delta\widehat{\mathbf{E}}_h^{n+1}) + 2\mathfrak{N}_h(\mathbf{u}_h^{n}, \widehat{\mathbf{u}}_h^{n+1}, \delta\widehat{\mathbf{E}}_h^{n+1}), \\
&\quad + 2\mathfrak{N}_h(\mathbf{u}(t^{n}), \mathbf{u}(t^{n}), \delta\widehat{\mathbf{E}}_h^{n+1}) - 2\mathfrak{N}_h(\mathbf{u}_h^{n-1}, \widehat{\mathbf{u}}_h^{n}, \delta\widehat{\mathbf{E}}_h^{n+1}).
\end{aligned}
$$

We now estimate each term A_i separately. First, we easily find out that

$$
\begin{aligned}
A_1 &\le \frac{\mu\tau}{14}\left\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\right\|_0^2 + C\mu\tau\left\|\nabla\delta\mathbf{F}^{n+1}\right\|_0^2 + \frac{1}{2}\left\|\delta\mathbf{E}^{n+1} - \delta\mathbf{E}^{n}\right\|_0^2 + C\left\|\delta\mathbf{F}^{n+1}\right\|_0^2, \\
A_2 &= 2\tau\left\langle \delta p(t^{n+1}) - \delta g^{n+1} \,,\, \operatorname{div} \delta\widehat{\mathbf{E}}_h^{n+1}\right\rangle \\
&\le \frac{C}{\mu}\int_{t^n}^{t^{n+1}} \|p_t(t)\|_0^2 \, dt + \frac{C\tau}{\mu}\left\|\delta g^{n+1}\right\|_0^2 + \frac{\mu\tau}{14}\left\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\right\|_0^2 + \frac{\mu\tau}{14}\left\|\nabla\delta\mathbf{F}^{n+1}\right\|_0^2.
\end{aligned}
$$

Since $\mathbf{U}_h^{n+1}$ is discrete divergence free, $A_3 = 2\mu\tau\left\langle \delta s_h^{n} \,,\, \operatorname{div} \delta\widehat{\mathbf{u}}_h^{n+1}\right\rangle$. Consequently, making use of (1.10) and (4.13), we arrive at

$$
A_3 = 2\mu\tau\left\langle \delta s_h^{n} \,,\, \delta s_h^{n} - \delta s_h^{n+1}\right\rangle \le \mu\tau\left(\left\|\delta s_h^{n}\right\|_0^2 - \left\|\delta s_h^{n+1}\right\|_0^2\right) + \mu\tau\left\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\right\|_0^2.
$$

At the same time, we further split A_4 to read $A_4 = A_{4,1} + A_{4,2}$ with

$$
\begin{aligned}
A_{4,1} &:= -2\tau\Big(\mathfrak{N}_h\big(\delta\mathbf{u}(t^{n+1}), \mathbf{u}(t^{n+1}), \delta\widehat{\mathbf{E}}_h^{n+1}\big) - \mathfrak{N}_h\big(\delta\mathbf{u}(t^{n}), \mathbf{u}(t^{n}), \delta\widehat{\mathbf{E}}_h^{n+1}\big) \\
&\qquad - \mathfrak{N}_h\big(\mathbf{u}(t^{n}) - \mathbf{u}_h^{n}, \mathbf{u}(t^{n+1}), \delta\widehat{\mathbf{E}}_h^{n+1}\big) + \mathfrak{N}_h\big(\mathbf{u}(t^{n-1}) - \mathbf{u}_h^{n-1}, \mathbf{u}(t^{n}), \delta\widehat{\mathbf{E}}_h^{n+1}\big)\Big), \\
A_{4,2} &:= -2\Big(\mathfrak{N}_h(\mathbf{u}_h^{n}, \mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}, \delta\widehat{\mathbf{E}}_h^{n+1}) - \mathfrak{N}_h\big(\mathbf{u}_h^{n-1}, \mathbf{u}(t^{n}) - \widehat{\mathbf{u}}_h^{n}, \delta\widehat{\mathbf{E}}_h^{n+1}\big)\Big).
\end{aligned}
$$

In light of (2.2) and definitions of $\mathbf{G}^i$ and $\mathbf{E}^i$, (2.12) produces

$$
A_{4,1} \le \frac{C}{\mu}D^{n+1} + \frac{\mu\tau}{14}\left\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\right\|_0^2 + \frac{\mu\tau}{14}\left\|\nabla\delta\mathbf{F}^{n+1}\right\|_0^2 + \frac{C}{\mu\tau}\sum_{i=n-1}^{n}\left(\left\|\mathbf{G}^i\right\|_0^2 + \left\|\mathbf{E}^i\right\|_0^2\right)
$$

with $D^{n+1} := \int_{t^{n-1}}^{t^{n+1}} \|\mathbf{u}_t(t)\|_0^2\,dt$. To bound $A_{4,2}$ we rewrite as $A_{4,2} = \sum_{i=1}^3 B_i$ with

$$\begin{aligned}
B_1 &:= -2\mathfrak{N}_h\big(\mathbf{u}_h^n, \mathbf{G}_h^{n+1}, \delta\widehat{\mathbf{E}}_h^{n+1}\big),\\
B_2 &:= 2\mathfrak{N}_h\big(\mathbf{u}_h^{n-1}, \mathbf{G}_h^{n}, \delta\widehat{\mathbf{E}}_h^{n+1}\big),\\
B_3 &:= -2\mathfrak{N}_h\big(\mathbf{u}_h^n, \widehat{\mathbf{E}}_h^{n+1}, \delta\widehat{\mathbf{E}}_h^{n+1}\big) + 2\mathfrak{N}_h(\mathbf{u}_h^{n-1}, \widehat{\mathbf{E}}_h^{n}, \delta\widehat{\mathbf{E}}_h^{n+1}).
\end{aligned}$$

Since $\big|\!\big|\!\big|\mathbf{G}_h^{n+1}\big|\!\big|\!\big| \le C\big(\|\mathbf{u}(t^{n+1})\|_2 + \|p(t^{n+1})\|_1\big) \le C$, (2.11) and (2.13) give

$$\begin{aligned}
B_1 &= 2\mathfrak{N}_h\left((\mathbf{u}(t^n) - \mathbf{u}_h^n) - \mathbf{u}(t^n), \mathbf{G}_h^{n+1}, \delta\widehat{\mathbf{E}}_h^{n+1}\right)\\
&\le \frac{C}{\mu\tau}\left(\|\mathbf{E}^n\|_0^2 + \|\mathbf{G}^n\|_0^2 + \big\|\mathbf{G}_h^{n+1}\big\|_0^2\right) + \frac{\mu\tau}{14}\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\|_0^2 + \frac{\mu\tau}{14}\big\|\nabla\delta\mathbf{F}^{n+1}\big\|_0^2,
\end{aligned}$$

as well as

$$B_2 \le \frac{C}{\mu\tau}\left(\big\|\mathbf{E}^{n-1}\big\|_0^2 + \big\|\mathbf{G}^{n-1}\big\|_0^2 + \|\mathbf{G}_h^n\|_0^2\right) + \frac{\mu\tau}{14}\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\|_0^2 + \frac{\mu\tau}{14}\big\|\nabla\delta\mathbf{F}^{n+1}\big\|_0^2.$$

Invoking crucial properties of $\mathfrak{N}_h$, written in (2.10), we infer that

$$B_3 = \frac{2}{\tau}\big(\mathfrak{N}_h\big(\mathbf{u}_h^n, \widehat{\mathbf{E}}_h^{n+1}, \widehat{\mathbf{E}}_h^n\big) + \mathfrak{N}_h\big(\mathbf{u}_h^{n-1}, \widehat{\mathbf{E}}_h^{n}, \widehat{\mathbf{E}}_h^{n+1}\big)\big) = 2\tau\mathfrak{N}_h\big(\delta\mathbf{u}_h^n, \delta\widehat{\mathbf{E}}_h^{n+1}, \widehat{\mathbf{E}}_h^n\big).$$

Hence

$$B_3 = -2\tau\mathfrak{N}_h\big(\delta\mathbf{G}_h^n - \delta\mathbf{u}(t^n), \delta\widehat{\mathbf{E}}_h^{n+1}, \widehat{\mathbf{E}}_h^n\big) - 2\tau\mathfrak{N}_h\big(\delta\mathbf{E}_h^n, \delta\widehat{\mathbf{E}}_h^{n+1}, \widehat{\mathbf{E}}_h^n\big) = B_4 + B_5.$$

Since $\|\widehat{\mathbf{E}}_h^n\|_1 \le C$ according to (4.12) and (4.14), (2.11) yields

$$\begin{aligned}
B_4 &\le C\tau\left(\|\delta\mathbf{G}_h^n\|_1 + \|\delta\mathbf{u}(t^n)\|_1\right)\big\|\delta\widehat{\mathbf{E}}_h^{n+1}\big\|_1\big\|\widehat{\mathbf{E}}_h^n\big\|_1\\
&\le \frac{C\tau}{\mu}\|\nabla\delta\mathbf{G}_h^n\|_0^2 + \frac{C}{\mu}\int_{t^n}^{t^{n+1}} \|\mathbf{u}_t(t)\|_1^2\,dt + \frac{\mu\tau}{14}\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\|_0^2 + \frac{\mu\tau}{14}\big\|\nabla\delta\mathbf{F}^{n+1}\big\|_0^2.
\end{aligned}$$

We now deal with B_5 via (2.13), namely, $B_5 \le C\tau\|\delta\mathbf{E}_h^n\|_{\mathbf{L}^3(\Omega)}\|\delta\widehat{\mathbf{E}}_h^{n+1}\|_1\|\widehat{\mathbf{E}}_h^n\|_1$. In contrast to [16], here we no longer have $\mathbf{E}_h^{n+1} \in \mathbf{H}_0^1$ and we have to resort to the inverse inequality $\|\delta\mathbf{E}_h^n\|_{\mathbf{L}^3(\Omega)} \le Ch^{-\frac{d}{6}}\|\delta\mathbf{E}_h^n\|_0$, whence

$$B_5 \le \underbrace{\frac{C\tau h^{-\frac{d}{3}}}{\mu}\|\delta\mathbf{E}_h^n\|_0^2\big\|\nabla\widehat{\mathbf{E}}_h^n\big\|_0^2}_{=:\Lambda^n} + \frac{\mu\tau}{14}\big\|\nabla\delta\widehat{\mathbf{E}}^{n+1}\big\|_0^2 + \frac{\mu\tau}{14}\big\|\nabla\delta\mathbf{F}^{n+1}\big\|_0^2.$$

We postpone the discussion of Λ^n until the end since it is rather delicate. We now insert the above estimates into (5.3), multiply by the weight σ^{n+1}, and add over n from 1 to N. Arguing as in Lemma 4.4, we see that the first two terms in (5.3) become

$$\sigma^{N+1}\big\|\delta\mathbf{E}^{N+1}\big\|_0^2 - \sigma^1\big\|\delta\mathbf{E}^1\big\|_0^2 - \tau\sum_{n=1}^N \|\delta\mathbf{E}^n\|_0^2 \ge -C + \sigma^{N+1}\big\|\delta\mathbf{E}^{n+1}\big\|_0^2. \tag{5.4}$$

On the other hand, we resort to the property $\frac{\sigma^{n+1}}{\sigma^n} \le 2$ for $n \ge 1$ to write

$$\sum_{n=1}^{N} \sigma^{n+1} A_2 \le \frac{C\tau}{\mu} \sum_{n=1}^{N} \sigma^{n+1} \big\|\delta g^{n+1}\big\|_0^2 + \frac{C}{\mu} \int_{t^1}^{t^{N+1}} \sigma(t) \|p_t(t)\|_0^2 \, dt \\ + \frac{\mu\tau}{14} \sum_{n=1}^{N} \sigma^{n+1} \big(\|\nabla \delta \widehat{\mathbf{E}}^{n+1}\|_0^2 + \|\nabla \delta \mathbf{F}^{n+1}\|_0^2\big). \tag{5.5}$$

Collecting these estimates, and using Lemmas 4.1–4.4 and 4.7, we get, for $D_1, D_2 > 0$,

$$\sigma^{N+1} \big\|\delta \mathbf{E}^{N+1}\big\|_0^2 + \frac{1}{2} \sum_{n=1}^{N} \sigma^{n+1} \big\|\delta \mathbf{E}^{n+1} - \delta \mathbf{E}^n\big\|_0^2 + \frac{\mu\tau}{2} \sum_{n=1}^{N} \sigma^{n+1} \big\|\nabla \delta \widehat{\mathbf{E}}^{n+1}\big\|_0^2 \\ + \sum_{n=1}^{N} \sigma^{n+1} \big\|\nabla \delta \rho_h^{n+1}\big\|_0^2 + \mu\tau\sigma^{N+1} \big\|\delta s_h^{N+1}\big\|_0^2 \le D_1 + D_2 \sum_{n=1}^{N} \sigma^{n+1} \Lambda^n.$$

To complete this proof, it suffices to show $\sum_{n=1}^{N} \sigma^{n+1} B_6^n \le C$. To do so, we start with a simpler form of the above estimate, namely,

$$\sigma^{N+1} \big\|\delta \mathbf{E}^{N+1}\big\|_0^2 \le D_1 + D_2 \tau h^{-\frac{d}{3}} \sum_{n=1}^{N} \sigma^{n+1} \|\delta \mathbf{E}_h^n\|_0^2 \big\|\nabla \widehat{\mathbf{E}}_h^n\big\|_0^2. \tag{5.6}$$

Since $\tau^2 \sum_{n=1}^{N} \|\delta \mathbf{E}^n\|_0^2 = \sum_{n=1}^{N} \big\|\mathbf{E}^n - \mathbf{E}^{n-1}\big\|_0^2 \le C\tau$ and $\|\nabla \widehat{\mathbf{E}}_h^n\|_0^2 \le C$ according to Lemma 4.4, we readily obtain the rough estimate

$$\sigma^{N+1} \big\|\delta \mathbf{E}^{N+1}\big\|_0^2 \le C h^{-\frac{d}{3}}.$$

To improve upon this, we utilize $\sum_{n=1}^{N} \|\nabla \widehat{\mathbf{E}}_h^n\|_0^2 \le C$, a by-product of (4.12) and (4.14). Hence

$$\sigma^{N+1} \big\|\delta \mathbf{E}^{N+1}\big\|_0^2 \le D_1 + D_2 \tau h^{-\frac{2d}{3}} \sum_{n=1}^{N} \big\|\nabla \widehat{\mathbf{E}}_h^n\big\|_0^2 \le C\tau h^{-\frac{2d}{3}}.$$

We realize that the net effect is a an additional factor $C\tau h^{-d/3}$ in (5.6). After m iterations, we obtain

$$\sigma^{N+1} \big\|\delta \mathbf{E}^{N+1}\big\|_0^2 \le M(m) \big(\tau h^{-\frac{d}{3}}\big)^m h^{-\frac{d}{3}},$$

where $M(m) > 0$ possibly grows with m. Since $\tau h^{-\frac{d}{3}} \le C_2 h^{\frac{d\varepsilon}{3}}$ for $m > \varepsilon^{-1}$ we obtain $\sum_{n=1}^{N} \sigma^{n+1} \Lambda^n \le C$. This shows our assertion (5.1).

If NLC is valid, so is Lemma 2.2, thereby making unnecessary the use of weight σ^{n+1} in (5.4) and (5.5). This yields an inequality similar to (5.1) without weights, and implies the asserted uniform estimate. ∎

Lemma 5.2 (rate of convergence for time-derivative of velocity). *Let* A1–A6 *hold and* $C_1 h^2 \le \tau \le C_2 h^{\frac{d}{3}(1+\varepsilon)}$ *be valid with arbitrary constants* $C_1, C_2 > 0$. *Then the error function* $\mathbf{E}^n$ *satisfies the weighted estimates*

$$\sigma^{N+1} \big\|\delta \mathbf{E}^{N+1}\big\|_*^2 + \sum_{n=1}^{N} \sigma^{n+1} \left(\big\|\delta \mathbf{E}^{n+1} - \delta \mathbf{E}^n\big\|_*^2 + \mu\tau \big\|\delta \mathbf{E}^{n+1}\big\|_0^2 \right) \le C(\tau + h^2). \tag{5.7}$$

If NLC is also valid, then the following uniform error estimates hold:

$$\left\|\delta\mathbf{E}^{N+1}\right\|_*^2 + \frac{1}{2}\sum_{n=1}^{N}\left\|\delta\mathbf{E}^{n+1} - \delta\mathbf{E}^{n}\right\|_*^2 + \mu\tau\sum_{n=1}^{N}\left\|\delta\mathbf{E}^{n+1}\right\|_0^2 \le C(\tau + h^2). \tag{5.8}$$

Proof. Let $(\mathbf{v}^n, q^n)$ and $(\mathbf{v}_h^n, q_h^n)$ be solutions of the Stokes equations (2.1) and (2.6) with $\mathbf{g} = \mathbf{E}^{n+1}$. Choosing $\mathbf{w}_h = 2\delta\mathbf{v}_h^{n+1}$ in (5.2), we arrive at

$$\left\|\nabla\delta\mathbf{v}_h^{n+1}\right\|_0^2 - \|\nabla\delta\mathbf{v}_h^{n}\|_0^2 + \left\|\nabla(\delta\mathbf{v}_h^{n+1} - \delta\mathbf{v}_h^{n})\right\|_0^2 + 2\mu\tau\left\|\delta\mathbf{E}^{n+1}\right\|_0^2 = \sum_{i=1}^{4} A_i \tag{5.9}$$

with

$$\begin{aligned}
A_1 &:= -2\mu\tau\big\langle\nabla\delta\mathbf{F}^{n+1}\,,\,\nabla\delta\mathbf{v}_h^{n+1}\big\rangle,\\
A_2 &:= 2\mu\tau\big(\langle\delta\mathbf{E}^{n+1}\,,\,\delta\mathbf{F}^{n+1}\rangle + \big\langle\nabla\delta q_h^{n+1}\,,\,\nabla\delta\rho_h^{n+1}\big\rangle\big),\\
A_3 &:= 2\tau\langle\delta P^{n+1}\,,\,\mathrm{div}\ \delta\mathbf{v}_h^{n+1}\rangle,\\
A_4 &:= 2\mathfrak{N}_h\big(\mathbf{u}_h^n, \widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}_h^{n+1}\big) - 2\mathfrak{N}_h\big(\mathbf{u}_h^{n-1}, \widehat{\mathbf{u}}_h^{n}, \delta\mathbf{v}_h^{n+1}\big)\\
&\quad - 2\mathfrak{N}_h\big(\mathbf{u}(t^{n+1}), \mathbf{u}(t^{n+1}), \delta\mathbf{v}_h^{n+1}\big) + 2\mathfrak{N}_h\big(\mathbf{u}(t^{n}), \mathbf{u}(t^{n}), \delta\mathbf{v}_h^{n+1}\big).
\end{aligned}$$

Except for A_4, we can proceed as in Lemma 4.7 to estimate A_1–A_3, whence

$$\begin{aligned}
A_1 &\le C\mu\tau\left(h^2\left\|\nabla\delta\mathbf{F}^{n+1}\right\|_0^2 + \left\|\delta\mathbf{F}^{n+1}\right\|_0^2\right) + \frac{\mu\tau}{6}\left\|\delta\mathbf{E}^{n+1}\right\|_0^2,\\
A_2 &\le C\mu\tau\left(\left\|\delta\mathbf{F}^{n+1}\right\|_0^2 + \left\|\nabla\delta\rho_h^{n+1}\right\|_0^2\right) + \frac{\mu\tau}{6}\left\|\delta\mathbf{E}^{n+1}\right\|_0^2,\\
A_3 &\le \frac{C\tau h^2}{\mu}\left\|\delta f^{n+1}\right\|_0^2 + \frac{\mu\tau}{6}\left\|\delta\mathbf{E}^{n+1}\right\|_0^2.
\end{aligned}$$

The remaining term A_4 gives rise to rather technical calculations. A tedious but simple rearrangement yields $A_4 = \sum_{i=1}^{6} A_{4,i}$ with each term A_i to be examined separately:

$$\begin{aligned}
A_{4,1} &:= -2\mathfrak{N}_h(\mathbf{u}(t^{n+1}) - 2\mathbf{u}(t^n) + \mathbf{u}(t^{n-1}), \mathbf{u}(t^{n+1}), \delta\mathbf{v}_h^{n+1}),\\
A_{4,2} &:= -2\mathfrak{N}_h((\mathbf{u}(t^n) - \mathbf{u}_h^n) - (\mathbf{u}(t^{n-1}) - \mathbf{u}_h^{n-1}), \mathbf{u}(t^{n+1}), \delta\mathbf{v}_h^{n+1}),\\
A_{4,3} &:= -2\mathfrak{N}_h(\mathbf{u}_h^n - \mathbf{u}_h^{n-1}, \mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}_h^{n+1}),\\
A_{4,4} &:= -2\mathfrak{N}_h(\mathbf{u}(t^n) - \mathbf{u}(t^{n-1}), \mathbf{u}(t^{n+1}) - \mathbf{u}(t^n), \delta\mathbf{v}_h^{n+1}),\\
A_{4,5} &:= -2\mathfrak{N}_h(\mathbf{u}(t^{n-1}) - \mathbf{u}_h^{n-1}, \mathbf{u}(t^{n+1}) - \mathbf{u}(t^n), \delta\mathbf{v}_h^{n+1}),\\
A_{4,6} &:= -2\mathfrak{N}_h(\mathbf{u}_h^{n-1}, (\mathbf{u}(t^{n+1}) - \mathbf{u}(t^n)) - (\widehat{\mathbf{u}}_h^{n+1} - \widehat{\mathbf{u}}_h^{n}), \delta\mathbf{v}_h^{n+1}).
\end{aligned}$$

Since $\left\|\mathbf{u}(t^{n+1}) - 2\mathbf{u}(t^n) + \mathbf{u}(t^{n-1})\right\|_0^2 \le C\tau^2\int_{t^{n-1}}^{t^{n+1}}\sigma\|\mathbf{u}_{tt}\|_0^2\,dt$, (2.2) and (2.12) yield

$$A_{4,1} \le C\tau\int_{t^{n-1}}^{t^{n+1}}\sigma(t)\|\mathbf{u}_{tt}(t)\|_0^2\,dt + C\tau\left\|\nabla\delta\mathbf{v}_h^{n+1}\right\|_0^2,$$

as well as

$$A_{4,2} \le \frac{\mu\tau}{8}\big(\|\delta\mathbf{G}^n\|_0^2 + \|\delta\mathbf{E}^n\|\big) + \frac{C\tau}{\mu}\left\|\nabla\delta\mathbf{v}_h^{n+1}\right\|_0^2.$$

Dealing with $A_{4,3}$ entails further rearrangement as follows:

$$\begin{aligned}A_{4,3} &= 2\tau\mathfrak{N}_h\big(\delta\mathbf{u}(t^n) - \delta\mathbf{u}_h^n, \mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}_h^{n+1} - \delta\mathbf{v}^{n+1}\big)\\ &\quad + 2\tau\mathfrak{N}_h\big(\delta\mathbf{u}(t^n) - \delta\mathbf{u}_h^n, \mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}^{n+1}\big)\\ &\quad - 2\tau\mathfrak{N}_h\big(\delta\mathbf{u}(t^n), \mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}_h^{n+1} - \delta\mathbf{v}^{n+1}\big)\\ &\quad - 2\tau\mathfrak{N}_h\big(\delta\mathbf{u}(t^n), \mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}^{n+1}\big).\end{aligned}$$

In view of (2.12) and (2.13), we can thus write

$$\begin{aligned}A_{4,3} &\le C\tau\|\delta\mathbf{u}(t^n) - \delta\mathbf{u}_h^n\|_{\mathbf{L}^3(\Omega)}\big\|\mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}\big\|_1\big\|\delta\mathbf{v}^{n+1} - \delta\mathbf{v}_h^{n+1}\big\|_1\\ &\quad + C\tau\|\delta\mathbf{u}(t^n) - \delta\mathbf{u}_h^n\|_0\big\|\mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}\big\|_1\big\|\delta\mathbf{v}^{n+1}\big\|_2\\ &\quad + C\tau\|\delta\mathbf{u}(t^n)\|_1\big\|\mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}\big\|_1\big\|\delta\mathbf{v}^{n+1} - \delta\mathbf{v}_h^{n+1}\big\|_1\\ &\quad + C\tau\|\delta\mathbf{u}(t^n)\|_1\big\|\mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1}\big\|_0\big\|\delta\mathbf{v}^{n+1}\big\|_2.\end{aligned}$$

Since $\|\delta(\mathbf{v}^{n+1} - \mathbf{v}_h^{n+1})\|_1 \le Ch\|\mathbf{E}^{n+1}\|_0$, because of (2.7), we see that the problematic term with $\mathbf{L}^3$ norm can be easily handled. In fact, invoking Lemma 5.1 together with an inverse inequality from L^3 to L^2 gives

$$\sigma^n\|\delta\mathbf{u}(t^n) - \delta\mathbf{u}_h^n\|_0^2 + \sigma^n h^2\|\delta\mathbf{u}(t^n) - \delta\mathbf{u}_h^n\|_{\mathbf{L}^3(\Omega)}^2 \le C.$$

We note that this inequality also holds without weight σ^n if NLC is valid. Since, according to (2.2), we have $\|\delta\mathbf{u}(t^n)\|_0^2 \le M$ and $\|\delta\mathbf{u}(t^n)\|_1^2 \le \tau^{-1}\int_{t^n}^{t^{n+1}} \|\mathbf{u}_t(t)\|_1^2\,dt \le M\tau^{-1}$, after a simple calculation we get

$$A_{4,3} \le \frac{C}{\mu}(\tau + h^2)D^n + \frac{C\tau}{\sigma^n\mu}\left(\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0^2 + \big\|\nabla\mathbf{G}^{n+1}\big\|_0^2\right) + \frac{\mu\tau}{8}\big\|\delta\mathbf{E}^{n+1}\big\|_0^2,$$

where $D^n := \int_{t^{n-1}}^{t^n} \|\nabla\mathbf{u}_t(t)\|_0^2\,dt$. We again use the bound for $\|\delta\mathbf{u}(t^n)\|_1$ to get

$$A_{4,4} \le C\tau^2\|\delta\mathbf{u}(t^n)\|_1\big\|\delta\mathbf{u}(t^{n+1})\big\|_1\big\|\delta\mathbf{v}_h^{n+1}\big\|_1 \le C\tau D^n + C\tau\big\|\nabla\delta\mathbf{v}_h^{n+1}\big\|_0^2.$$

To estimate $A_{4,5}, A_{4,6}$ we again have to handle an L^3 norm, this time for $\mathbf{u}(t^n) - \mathbf{u}_h^n$. Combining once again Lemma 5.1 with an inverse estimate yields $h\|\mathbf{u}(t^n) - \mathbf{u}_h^n\|_{\mathbf{L}^3(\Omega)} \le C\|\mathbf{u}(t^n) - \mathbf{u}_h^n\|_0 \le C(\tau + h^2)^{\frac{1}{2}}$. Consequently,

$$\begin{aligned}A_{4,5} &\le C\tau\big\|\mathbf{u}(t^{n-1}) - \mathbf{u}_h^{n-1}\big\|_{\mathbf{L}^3(\Omega)}\big\|\delta\mathbf{u}(t^{n+1})\big\|_1\big\|\delta\mathbf{v}^{n+1} - \delta\mathbf{v}_h^{n+1}\big\|_1\\ &\quad + C\tau\big\|\mathbf{u}(t^{n-1}) - \mathbf{u}_h^{n-1}\big\|_0\big\|\delta\mathbf{u}(t^{n+1})\big\|_1\big\|\delta\mathbf{v}^{n+1}\big\|_2\\ &\le \frac{C}{\mu}(\tau + h^2)D^{n+1} + \frac{\mu\tau}{8}\big\|\delta\mathbf{E}^{n+1}\big\|_0^2.\end{aligned}$$

In addition, since

$$\begin{aligned}A_{4,6} &= 2\tau\mathfrak{N}_h(\mathbf{u}(t^{n-1}) - \mathbf{u}_h^{n-1}, \delta\mathbf{u}(t^{n+1}) - \delta\widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}_h^{n+1} - \delta\mathbf{v}^{n+1})\\ &\quad + 2\tau\mathfrak{N}_h(\mathbf{u}(t^{n-1}) - \mathbf{u}_h^{n-1}, \delta\mathbf{u}(t^{n+1}) - \delta\widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}^{n+1})\\ &\quad - 2\tau\mathfrak{N}_h(\mathbf{u}(t^{n-1}), \delta\mathbf{u}(t^{n+1}) - \delta\widehat{\mathbf{u}}_h^{n+1}, \delta\mathbf{v}_h^{n+1}),\end{aligned}$$

a similar argument leads to

$$A_{4,6} \le \frac{C\tau}{\mu}\left(\|\delta\widehat{\mathbf{E}}^{n+1} + \delta\mathbf{G}^{n+1}\|_0^2 + (\tau + h^2)\|\delta\widehat{\mathbf{E}}^{n+1} + \delta\mathbf{G}^{n+1}\|_1^2\right) + \frac{\mu\tau}{8}\|\delta\mathbf{E}^{n+1}\|_0^2.$$

We now multiply both sides of (5.9) by the weight σ^{n+1} and sum over n for $1 \leq n \leq N$. We first examine the ensuing first two terms on the left-hand side of (5.9). In light of $\sigma^1 = \tau$, $h^2 \leq C\tau$, $\sum_{n=1}^N \|\nabla\delta\mathbf{v}_h^n\|_0^2 \leq C$ (see Lemma 4.7) and $\sigma^1 \big\|\nabla\delta\mathbf{v}_h^1\big\|_0^2 \leq C\tau$ (see Remark 4.8), we deduce

$$\sum_{n=1}^N \Big(\sigma^{n+1}\big\|\nabla\delta\mathbf{v}^{n+1}\big\|_0^2 - \sigma^n\big\|\nabla\delta\mathbf{v}_h^n\big\|_0^2 - (\sigma^{n+1}-\sigma^n)\big\|\nabla\delta\mathbf{v}_h^n\big\|_0^2\Big)$$
$$\geq \sigma^{N+1}\big\|\nabla\delta\mathbf{v}_h^{N+1}\big\|_0^2 - \sigma^1\big\|\nabla\delta\mathbf{v}_h^1\big\|_0^2 - \tau\sum_{n=1}^N \|\nabla\delta\mathbf{v}_h^n\|_0^2 \geq \sigma^{N+1}\big\|\nabla\delta\mathbf{v}_h^{N+1}\big\|_0^2 - C\tau.$$

Since $\frac{\sigma^{n+1}}{\sigma^n} \leq 2$ for $n \geq 1$, we can replace σ/σ^n in $A_{4,3}$ by a constant. Therefore, we can achieve an estimate for $\sigma^{n+1}\|\nabla\delta\mathbf{v}_h^{n+1}\|_0^2$ with the aid of Lemmas 4.1, 4.7, and 5.1, as well as the discrete Gronwall lemma. The asserted *weighted* error estimate follows from (2.8).

If NLC is valid, we do not need to multiply (5.9) by σ^{n+1} to derive the *uniform* error estimate (5.8). In this case, we have, instead, $\|\delta\mathbf{G}^n\|_0 + \|\delta\mathbf{E}^n\|_0 \leq C$ (see Lemmas 4.2 and 5.1). We finally proceed as before to obtain (5.8). $\square$

6. Theorem 1.3: Error analysis for pressure. We derive here the error of pressure of Theorem 1.3 by exploiting all previous results.

Lemma 6.1 (rate of convergence for pressure). *Let* A1–A6 *hold and* $C_1h^2 \leq \tau \leq C_2h^{\frac{d}{3}(1+\varepsilon)}$ *be valid with arbitrary constants* $C_1, C_2 > 0$. *Then the pressure error function satisfies the weighted estimates*

$$\tau\sum_{n=0}^N \sigma^{n+1}\big\|e_h^{n+1}\big\|_0^2 \leq C\left(\tau + h^2\right). \tag{6.1}$$

If NLC is also valid, then the following uniform error estimate holds:

$$\tau\sum_{n=0}^N \big\|e_h^{n+1}\big\|_0^2 \leq C(\tau + h^2). \tag{6.2}$$

Proof. Since $p_h^{n+1} = \mu s_h^{n+1} - \tau^{-1}\rho_h^{n+1}$ and $\widehat{\mathbf{E}}_h^{n+1} = \mathbf{E}_h^{n+1} + \nabla\rho_h^{n+1}$ according to (1.11) and (1.13), we can rearrange (4.15) to read $\big\langle e_h^{n+1}\,,\,\mathrm{div}\,\mathbf{w}_h\big\rangle = A_1 + A_2$ with

$$\begin{aligned} A_1 &:= \langle\delta\mathbf{E}^{n+1}\,,\,\mathbf{w}_h\rangle + \mu\langle\nabla\widehat{\mathbf{E}}^{n+1}\,,\,\nabla\mathbf{w}_h\rangle - \langle\mu(s_h^{n+1} - s_h^n) + f^{n+1}\,,\,\mathrm{div}\,\mathbf{w}_h\rangle,\\ A_2 &:= \mathfrak{N}_h\left(\mathbf{u}(t^{n+1}),\mathbf{u}(t^{n+1}),\mathbf{w}_h\right) - \mathfrak{N}_h\big(\mathbf{u}_h^n,\widehat{\mathbf{u}}_h^{n+1},\mathbf{w}_h\big). \end{aligned}$$

In view of (4.13), A_1 can be bounded as follows:

$$\sup_{\mathbf{w}_h\in\mathbb{V}_h}\frac{|A_1|}{\|\nabla\mathbf{w}_h\|_0} \leq C\|\delta\mathbf{E}^{n+1}\|_0 + C\mu\|\nabla\widehat{\mathbf{E}}^{n+1}\|_0 + C\big\|f^{n+1}\big\|_0.$$

The remaining term A_2 can be further split as follows:

$$\begin{aligned} A_2 = &-\mathfrak{N}_h\big(\mathbf{u}(t^{n+1}) - \mathbf{u}(t^n),\mathbf{u}(t^{n+1}),\mathbf{z}_h^{n+1}\big)\\ &-\mathfrak{N}_h\big(\mathbf{u}(t^n) - \mathbf{u}_h^n,\mathbf{u}(t^{n+1}),\mathbf{z}_h^{n+1}\big)\\ &-\mathfrak{N}_h\big(\mathbf{u}(t^n),\mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1},\mathbf{z}_h^{n+1}\big)\\ &-\mathfrak{N}_h\big(\mathbf{u}_h^n - \mathbf{u}(t^n),\mathbf{u}(t^{n+1}) - \widehat{\mathbf{u}}_h^{n+1},\mathbf{z}_h^{n+1}\big). \end{aligned}$$

The only problematic term is the last one because it requires use of (2.13). To this end, note that $\|\mathbf{u}(t^n) - \mathbf{u}_h^n\|_{\mathbf{L}^3(\Omega)} \le C + Ch^{-\frac{d}{6}}(\tau^{\frac{1}{2}} + h) \le C$, as results from adding and subtracting $I_h\mathbf{u}(t^n)$, and employing an inverse inequality together with (4.14). Therefore, since (2.2) implies $\int_{t^n}^{t^{n+1}} \|\mathbf{u}_t(t)\|_0 \, dt \le C\tau$,

$$\sup_{\mathbf{w}_h \in \mathbb{V}_h} \frac{|A_2|}{\|\nabla \mathbf{w}_h\|_0} \le C\tau + C\big(\|\mathbf{E}^n\|_0 + \|\mathbf{G}^n\|_0 + \|\widehat{\mathbf{E}}^{n+1}\|_1 + \|\mathbf{G}^{n+1}\|_1\big).$$

Altogether, invoking the inf-sup condition $A4$ in conjunction with (4.3) and (4.14), we obtain

$$\begin{aligned} \beta\|e_h^{n+1}\|_0 &\le \sup_{\mathbf{w}_h \in \mathbb{V}_h} \frac{\langle e_h^{n+1}, \operatorname{div} \mathbf{w}_h\rangle}{\|\nabla \mathbf{w}_h\|_0} \\ &\le C\big(\tau^{\frac{1}{2}} + h\big) + C\big(\|\delta\mathbf{E}^{n+1}\|_0 + \|\widehat{\mathbf{E}}^{n+1}\|_1 + \|\mathbf{G}^{n+1}\|_1 + \|f^{n+1}\|_0\big). \end{aligned}$$

What remains now is to square, multiply by $\tau\sigma^{n+1}$ (resp., τ in case NLC is valid), and sum over n from 0 to N. Recalling (4.3), (4.12), (4.14), and (5.7), assertion (6.1) (resp., (6.2)) follows immediately. This concludes the proof. □

7. Numerical experiments.

7. Numerical experiments. In this section, we document the computational performance of the gauge–Uzawa FEM with two relevant examples. They were both computed within the finite element toolbox ALBERT of Schmidt and Siebert [22].

7.1. Example 1: Smooth solution. We first test the performance of GU-FEM with a smooth solution. Let $\Omega = [0,1] \times [0,1]$ and the solution be given by

$$\begin{cases} u(x,y,t) = \cos(t)(x^2 - 2x^3 + x^4)(2y - 6y^2 + 4y^3), \\ v(x,y,t) = -\cos(t)(y^2 - 2y^3 + y^4)(2x - 6x^2 + 4x^3), \\ p(x,y,t) = \cos(t)\left(x^2 + y^2 - \dfrac{2}{3}\right). \end{cases}$$

The forcing term $\mathbf{f}(t)$ is determined accordingly for any μ; here $\mu = 1$. Computations are carried out with the Taylor–Hood $(\mathcal{P}^2, \mathcal{P}^1)$ finite element pair on quasi-uniform meshes of size h. However, the coarsest mesh is quite distorted to avoid superconvergence effects. In view of the error estimates of Theorems 1.2 and 1.3, we adopt the parabolic relation $\tau = h^2$ to avoid dominance of either space or time error over the other. Table 7.1 shows second order accuracy for both velocity and pressure in $L^2(\mathbf{H}^1(\Omega) \times L^2(\Omega))$. This is better than predicted by the theory for solutions with minimal regularity, especially when $t \downarrow 0$, and hints at the need for further analysis beyond the present techniques, perhaps involving also other norms as suggested in Table 7.1.

7.2. Example 2: Backward step and do-nothing boundary condition. In order to explore the applicability of the gauge–Uzawa FEM beyond the theory, we consider the backward step flow problem with *do-nothing boundary condition*; this is a natural boundary condition for the stress, namely,

$$(-\nabla \mathbf{u} + \mathbf{I}p) \cdot \boldsymbol{\nu} = 0 \quad \text{on } \Gamma_{out}, \tag{7.1}$$

where $\Gamma_{out} \subset \partial\Omega$. This condition can be imposed on fluid problems with an open outlet without forcing. Conditions involving the stress and geometric quantities such

Table 7.1

Example 7.1*: The error decay of the gauge–Uzawa FEM for a smooth solution and several norms for velocity and pressure. The computations are performed with the Taylor–Hood* ($\mathcal{P}^2, \mathcal{P}^1$) *finite element pair on quasi-uniform meshes. The meshes are distorted though to prevent superconvergence effects. The table shows second order accuracy for both velocity and pressure for the relation* $\tau = h^2$.

h	$\mathbf{u}(t^n) - \mathbf{u}_h^n$				$p(t^n) - p_h^n$		
	$L^\infty(\mathbf{L}^2)$	$L^\infty(\mathbf{L}^\infty)$	$L^2(\mathbf{L}^2)$	$L^2(\mathbf{H}^1)$	$L^\infty(L^2)$	$L^\infty(L^\infty)$	$L^2(L^2)$
1/8	6.21e-4	1.61e-3	1.57e-3	2.21e-2	1.05e-2	8.95e-2	2.71e-2
1/16	1.57e-4	4.06e-4	4.31e-4	6.44e-3	2.75e-3	2.93e-2	8.37e-3
1/32	3.94e-5	9.99e-5	1.11e-4	1.72e-3	6.94e-4	8.87e-3	2.33e-3
1/64	9.84e-6	2.47e-5	2.80e-5	4.42e-4	1.74e-4	2.59e-3	6.20e-4
1/128	2.46e-6	6.14e-6	7.02e-6	1.12e-4	4.35e-5	7.37e-4	1.61e-4
Order	1.99	2.01	1.99	1.98	1.99	1.81	1.94

as mean curvature are ubiquitous in dealing with free boundary problems for fluids. The mere fact that projection methods decouple velocity and pressure computations, and that both $\mathbf{u}$ and p appear together in (7.1) makes its implementation a challenge. This is the case for several projections methods such as the Chorin method [4, 21, 18] and the gauge method [7, 8, 27].

Since the momentum equation (1.8) is consistent for the pair $(\widehat{\mathbf{u}}_h^{n+1}, p_h^n)$, as written in (1.14), to impose (7.1) on the gauge–Uzawa method, we use the modified form $(-\nabla \mathbf{u}^{n+1} + \mathbf{I}p^n) \cdot \boldsymbol{\nu} = 0$. This amounts to solving (1.14), namely,

$$\tau^{-1} \left\langle \widehat{\mathbf{u}}_h^{n+1} - \widehat{\mathbf{u}}_h^n \, , \, \mathbf{w}_h \right\rangle + \mathfrak{N}_h(\mathbf{u}_h^n, \widehat{\mathbf{u}}_h^{n+1}, \mathbf{w}_h) + \mu \left\langle \nabla \widehat{\mathbf{u}}_h^{n+1} \, , \, \nabla \mathbf{w}_h \right\rangle$$
$$- \mu \langle p_h^n \, , \, \mathrm{div} \, \mathbf{w}_h \rangle = \langle \mathbf{f}(t^{n+1}) \, , \, \mathbf{w}_h \rangle,$$

but with test function $\mathbf{w}_h$ free on Γ_{out}. This leads, however, to an incompatible Poisson problem (1.9) if we insist on a homogeneous Neumann condition; note that now it is plausible that $\int_{\partial\Omega} \widehat{\mathbf{u}}_h^{n+1} \cdot \boldsymbol{\nu} \neq 0$.

To circumvent this issue, we consider a space-continuous gauge–Uzawa formulation. In view of (1.9) and (1.12), we can write

$$\left\langle \nabla \rho^{n+1} \, , \, \nabla \psi \right\rangle = - \left\langle \widehat{\mathbf{u}}^{n+1} \, , \, \nabla \psi \right\rangle = \left\langle \mathbf{u}^{n+1} - \widehat{\mathbf{u}}^{n+1} \, , \, \nabla \psi \right\rangle \qquad \forall \psi \in \mathbb{P}.$$

This amounts to the natural boundary condition $\partial_{\boldsymbol{\nu}} \rho^{n+1} = (\mathbf{u}^{n+1} - \widehat{\mathbf{u}}^{n+1}) \cdot \boldsymbol{\nu}$, which is not computable since we do not yet know $\mathbf{u}^{n+1}$. We now decompose $\partial\Omega$ into an inflow part Γ_{in}, where we prescribe velocity, an outflow part Γ_{out}, where we impose (7.1), and the rest where $\widehat{\mathbf{u}}^{n+1} \cdot \boldsymbol{\nu} = \mathbf{u}^{n+1} \cdot \boldsymbol{\nu} = 0$. Since $\int_{\partial\Omega} \mathbf{u}^{n+1} \cdot \boldsymbol{\nu} = \int_\Omega \mathrm{div} \, \mathbf{u}^{n+1} = 0$,

$$\int_{\Gamma_{out}} \mathbf{u}^{n+1} \cdot \boldsymbol{\nu} = - \int_{\Gamma_{in}} \mathbf{u}^{n+1} \cdot \boldsymbol{\nu} = - \int_{\Gamma_{in}} \widehat{\mathbf{u}}^{n+1} \cdot \boldsymbol{\nu},$$

whence $\int_{\Gamma_{out}} (\mathbf{u}^{n+1} - \widehat{\mathbf{u}}^{n+1}) \cdot \boldsymbol{\nu} = - \int_{\partial\Omega} \widehat{\mathbf{u}}^{n+1} \cdot \boldsymbol{\nu}$. We thus solve (1.9) with a constant flux condition, namely,

$$\partial_{\boldsymbol{\nu}} \rho^{n+1} = -|\Gamma_{out}|^{-1} \int_{\partial\Omega} \widehat{\mathbf{u}}^{n+1} \cdot \boldsymbol{\nu} \quad \text{on } \Gamma_{out}.$$

We consider a simple geometry consisting of a backward step flow with *do-nothing boundary condition.* This example has been studied extensively and our results are consistent with those in the literature [14, 23]. The computational domain Ω is $[0, 6] \times$

$[0,1]$ with an obstacle $[1.2,1.6]\times[0,0.4]$ (see Figure 7.1). No slip boundary condition is imposed except on the inflow boundary Γ_{in} and on the outflow boundary Γ_{out}. We assign $\mathbf{u}=(1,0)$ on Γ_{in} and (7.1) on Γ_{out} for all time t. The viscosity is $\mu=0.005$ and the discretization parameters are $\tau=0.05$ and $h=1/32$.

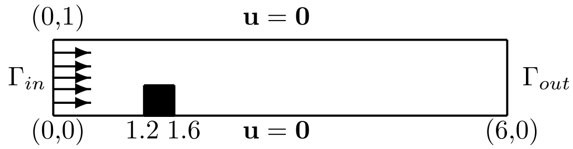


FIG. 7.1. *Example 7.2: The computational domain and boundary values. The viscosity is $\mu=0.005$ and the discretization parameters are $\tau=0.05$ and $h=1/32$.*

Figure 7.2 is a time sequence of streamlines and velocity vector fields for $t=1$, 2, 5, and 50. For $t=50$ the evolution already became stationary. Figure 7.3 displays zooms of the recirculation zone behind the step.

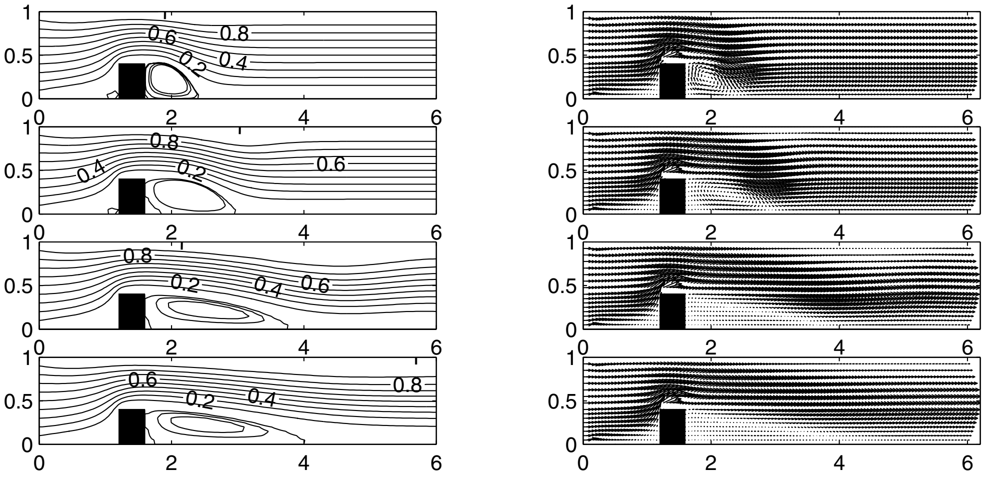


FIG. 7.2. *Example 7.2: The streamlines and velocity vector fields at times $t=1$, 2, 5, and 50.*

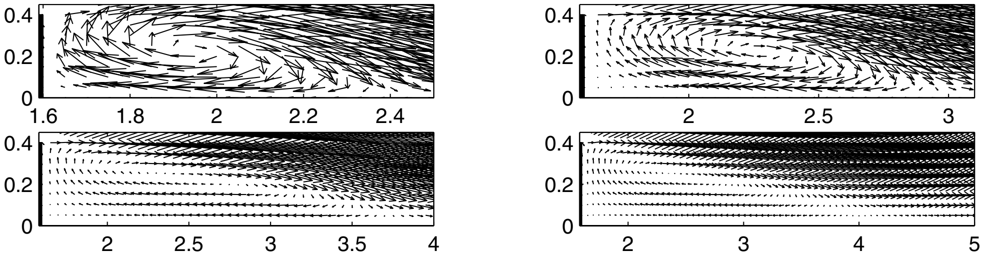


FIG. 7.3. *Example 7.2: Zooms of a velocity vector field in the recirculation zone behind the step at times $t=1$, 2, 5, and 50.*

REFERENCES

[1] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods,* Springer-Verlag, New York, 1994.

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

[3] D. Brown, R. Cortez, and M. Minion, *Accurate projection methods for the incompressible Navier–Stokes equations,* J. Comput. Phys., 168 (2001), pp. 464–499.

[4] A.J. Chorin, *Numerical solution of the Navier–Stokes equations,* Math. Comp., 22 (1968), pp. 745–762.

[5] P. Constantin and C. Foias, *Navier–Stokes Equations,* Chicago Lectures in Math., University of Chicago Press, Chicago, IL, 1988.

[6] M. Dauge, *Stationary Stokes and Navier–Stokes systems on two or three-dimensional domains with corners,* SIAM J. Math. Anal., 20 (1989), pp. 74–97.

[7] W. E and J.-G. Liu, *Gauge method for viscous incompressible flows,* Commun. Math. Sci., 1 (2003), pp. 317–332.

[8] W. E and J.-G. Liu, *Gauge finite element method for incompressible flows,* Internat. J. Numer. Methods Fluids, 34 (2000), pp. 701–710.

[9] W. E and J.-G. Liu, *Projection method* I*: Convergence and numerical boundary layers,* SIAM J. Numer. Anal., 32 (1995), pp. 1017–1057.

[10] V. Girault and P.A. Raviart, *Finite Element Methods for Navier–Stokes Equations,* Springer-Verlag, New York, 1986.

[11] J.L. Guermond and L. Quartapelle, *On the approximation of the unsteady Navier–Stokes equations by finite element projection methods,* Numer. Math., 80 (1998), pp. 207–238.

[12] J.G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem.* I*. Regularity of solutions and second-order error estimates for spatial discretization,* SIAM J. Numer. Anal., 19 (1982), pp. 57–77.

[13] R.B. Kellogg and J.E. Osborn, *A regularity result for the stokes problems in a convex polygon,* J. Funct. Anal., 21 (1976), pp. 397–431.

[14] H. Laval and L. Quartapelle, *A fractional-step Taylor–Gallerkin method for unsteady incompressible flows,* Internat. J. Numer. Methods Fluids, 11 (1990), pp. 501–513.

[15] R.H. Nochetto and J.-H. Pyo, *Optimal relaxation parameter for the Uzawa method,* Numer. Math., 98 (2004), pp. 696–702.

[16] R.H. Nochetto and J.-H. Pyo, *Error estimates for semi-discrete Gauge methods for the Navier–Stokes equations,* Math. Comp., 74 (2005), pp. 521–542.

[17] V.I. Oseledets, *A new form of writing out the Navier–Stokes equation. The Hamiltonian formalism,* Russian Math. Surveys, 44 (1989), pp. 210–211.

[18] A. Prohl, *Projection and Quasi-Compressiblity Methods for Solving the Incompressible Navier–Stokes Equations,* B.G. Teubner, Stuttgart, 1997.

[19] J.-H. Pyo, *The Gauge–Uzawa and Related Projection Finite Element,* Ph.D. dissertation, University of Maryland, 2002.

[20] J.-H. Pyo and J. Shen, *Normal mode analysis of second-order projection methods for incompressible flows*, Discrete Contin. Dyn. Syst. Ser. B, 5 (2005), pp. 817–840.

[21] J. Shen, *On error estimates of projection methods for Navier–Stokes equation: First order schemes,* SIAM J. Numer. Anal., 29 (1992), pp. 57–77.

[22] A. Schmidt and K.G. Siebert, *Design of Adaptive Finite Element Software: The Finite Element Toolbox: ALBERTA,* Lect. Notes Comput. Sci. Eng. 42, Springer-Verlag, Berlin, 2005.

[23] L.Q. Tang and T.H. Tsang, *A least-squares finite element method for time-dependent incompressible flows with thermal convection,* Internat. J. Numer. Methods Fluids, 17 (1993), pp. 271–289.

[24] R. Temam, *Navier–Stokes Equations: Theory and Numerical Analysis,* North-Holland, Amsterdam, 1977.

[25] R. Temam, *Sur l'approximation de la solution des equations de Navier–Stokes par la methode des pas fractionnaires.* II, Arch. Ration. Mech. Anal., 33 (1969), pp. 377–385.

[26] R. Verfürth, *A posteriori error estimators for the Stokes equations,* Numer. Math., 55 (1989), pp. 309–325.

[27] C. Wang and J.-G. Liu, *Convergence of gauge method for incompressible flow,* Math. Comp., 232 (2000), pp. 1385–1407.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1069–1090

EXPLICIT EXPONENTIAL RUNGE–KUTTA METHODS FOR SEMILINEAR PARABOLIC PROBLEMS*

MARLIS HOCHBRUCK† AND ALEXANDER OSTERMANN‡

Abstract. The aim of this paper is to analyze explicit exponential Runge–Kutta methods for the time integration of semilinear parabolic problems. The analysis is performed in an abstract Banach space framework of sectorial operators and locally Lipschitz continuous nonlinearities. We commence by giving a new and short derivation of the classical (nonstiff) order conditions for exponential Runge–Kutta methods, but the main interest of our paper lies in the stiff case. By expanding the errors of the numerical method in terms of the solution, we derive new order conditions that form the basis of our error bounds for parabolic problems. We show convergence for methods up to order four, and we analyze methods that were recently presented in the literature. These methods have classical order four, but they do not satisfy some of the new conditions. Therefore, an order reduction is expected. We present numerical experiments which show that this order reduction in fact arises in practical examples. Based on our new conditions, we finally construct methods that do not suffer from order reduction.

Key words. exponential integrators, Runge–Kutta methods, semilinear parabolic problems, stiff order conditions, explicit high-order methods, convergence, order reduction

AMS subject classifications. Primary, 65M12; Secondary, 65L06

DOI. 10.1137/040611434

1. Introduction.

Motivated by recent interest in exponential integrators for stiff problems [3, 9, 11, 12] and inspired by the promising numerical experiments reported in those papers, we present error bounds for a class of explicit exponential Runge–Kutta methods for semilinear parabolic problems

$$u'(t) + Au(t) = g(t, u(t)), \qquad u(t_0) = u_0. \tag{1.1}$$

The idea behind exponential integrators is an old one and dates back to the 1960s. The early literature on exponential one-step methods comprises [4, 6, 14, 23, 24, 25]. In most of those papers, the methods were constructed by making use of the variation-of-constants formula for the solution of (1.1). All of them use the exponential function of the matrix $-hA$ in order to step from time t to time $t+h$, or some rational approximations thereof. Exponential multistep methods have been considered in [13, 21, 26].

Although the first exponential integrators were proposed many years ago, such methods have not been regarded as practical for a long time. This view, however, has changed recently as new methods for computing or approximating the product of a matrix exponential function with a vector have been developed; see the review [19] and references therein. For large problems, polynomial approximations have to be applied, in general, either based on Chebyshev polynomials or using variants of the Lanczos process or the Arnoldi method. (These techniques are also reviewed in [19].) For parabolic problems it has been shown recently that one can achieve grid independent convergence of the Lanczos process by working with a shifted and inverted matrix

*Received by the editors July 12, 2004; accepted for publication (in revised form) December 27, 2004; published electronically September 20, 2005.
http://www.siam.org/journals/sinum/43-3/61143.html

†Mathematisches Institut, Heinrich-Heine Universität Düsseldorf, Universitätsstr. 1, D–40225 Düsseldorf, Germany (marlis@am.uni-duesseldorf.de).

‡Institut für Mathematik, Universität Innsbruck, Technikerstr. 13, A–6020 Innsbruck, Austria (alexander.ostermann@uibk.ac.at).

[5, 20]. The linear systems arising in each step of the Lanczos process can either be solved directly (if fast direct solvers are available) or with a preconditioned iterative method. Details can be found in [5].

The numerical comparisons presented in [11, 12, 18] reveal a number of examples where explicit exponential integrators outperform standard integrators. However, while the convergence behavior of implicit or linearly implicit Runge–Kutta or multistep methods for parabolic problems is nowadays well understood [15, 16], no analysis is known so far for explicit exponential integrators. In our paper [10] we analyzed implicit exponential Runge–Kutta methods. There, a crucial step in the convergence proofs was to show that all stages have a sufficiently small defect when the true solution is inserted in place of the numerical solution. This is no longer true for explicit schemes of order at least three. Therefore, new techniques have to be used for proving error bounds in this case. Our aim with this paper is to derive new order conditions for stiff problems and, based on these, to give error bounds for parabolic problems. The new conditions will then enable us to analyze the methods presented in the literature and, further, to construct new methods that do not suffer from order reduction.

The outline of the paper is as follows. In section 2, we define a general class of exponential Runge–Kutta methods and give conditions under which these methods preserve equilibria. We further give a simple and short derivation of the classical order conditions for arbitrary order. Our convergence analysis of (1.1) will be performed in the standard framework of analytic semigroups and locally Lipschitz continuous nonlinearities in a Banach space. This abstract framework is recalled in section 3. Our main results are contained in section 4, where we derive new order conditions for explicit exponential Runge–Kutta methods applied to parabolic problems. These conditions, which comprise the classical order conditions, are given in Table 2 up to order four. Based on them, we show convergence for explicit exponential Runge–Kutta methods up to order four, under appropriate temporal smoothness of the exact solution. The convergence results for the exponential Euler method and for second-order methods are given in Theorems 4.2 and 4.3, respectively, and our main result is Theorem 4.7. The new order conditions are further used in section 5 to analyze methods from the literature and to construct new methods. In particular, we will show that neither of the exponential classical Runge–Kutta methods from [3, 12] is of order four, in general, when applied to parabolic problems. Our order conditions enable us, however, to construct a new exponential variant of the classical Runge–Kutta method which is of full order four. In section 6 we present numerical experiments which show that the order reductions predicted by our theory may in fact arise in practical examples.

2. Order conditions and general properties.

We consider the following general class of one-step methods for solving (1.1):

$$u_{n+1} = \chi(-hA)u_n + h\sum_{i=1}^{s} b_i(-hA)G_{ni}, \tag{2.1a}$$

$$U_{ni} = \chi_i(-hA)u_n + h\sum_{j=1}^{s} a_{ij}(-hA)G_{nj}, \tag{2.1b}$$

$$G_{nj} = g(t_n + c_j h, U_{nj}). \tag{2.1c}$$

Here, the method coefficients χ, χ_i, a_{ij}, and b_i are constructed from exponential functions, or rational approximations of such functions evaluated at the matrix or operator $-hA$.

For consistency reasons, we always assume that $\chi(0) = \chi_i(0) = 1$. It seems worth mentioning that (2.1) reduces to a Runge–Kutta method with coefficients $b_i = b_i(0)$ and $a_{ij} = a_{ij}(0)$ if we consider the limit $A \to 0$. The latter method will be called the *underlying Runge–Kutta method* henceforth, while (2.1) will be referred to as an *exponential Runge–Kutta method* in the following. We suppose throughout the paper that the underlying Runge–Kutta method satisfies

$$\sum_{j=1}^{s} b_j(0) = 1, \qquad \sum_{j=1}^{s} a_{ij}(0) = c_i, \quad i = 1, \dots, s,$$

which makes it invariant under the transformation of (1.1) to autonomous form.

A desirable property of numerical methods is that they preserve equilibria $u^\star$ of the autonomous problem $u'(t) + Au(t) = g(u(t))$. Requiring $U_{ni} = u_n = u^\star$ for all i and $n \geq 0$ immediately yields the necessary and sufficient conditions. It turns out that the coefficients of the method have to satisfy

$$\sum_{j=1}^{s} b_j(z) = \frac{\chi(z) - 1}{z}, \qquad \sum_{j=1}^{s} a_{ij}(z) = \frac{\chi_i(z) - 1}{z}, \quad i = 1, \dots, s. \tag{2.2}$$

Without further mention, we will assume throughout the paper that these conditions are fulfilled.

With the help of (2.2), the functions χ and χ_i can be eliminated in (2.1). The numerical scheme then takes the form

$$u_{n+1} = u_n + h \sum_{i=1}^{s} b_i(-hA)\big(G_{ni} - Au_n\big), \tag{2.3a}$$

$$U_{ni} = u_n + h \sum_{j=1}^{s} a_{ij}(-hA)\big(G_{nj} - Au_n\big). \tag{2.3b}$$

Conditions (2.2) also imply that we can restrict ourselves to autonomous problems

$$u'(t) + Au(t) = g(u(t)), \qquad u(t_0) = u_0, \tag{2.4}$$

since all methods satisfying (2.2) are invariant under the transformation of (1.1) to autonomous form.

For $A = 0$, the methods reduce to classical Runge–Kutta methods, for which the order conditions are well known. In the case of $A \neq 0$, nonstiff order conditions of order up to 5 have been presented by Friedli [6] using Taylor series expansion of the exact and the numerical solutions; see also [25] and references therein. Our approach below is based on trees, which allows us to extract the order conditions for arbitrary orders in a systematic and simple way, very similar to classical Runge–Kutta methods. A more technical approach based on the theory of B-series is chosen in [1].

Writing (2.4) in the form

$$u' = g(u) - Au = F(u)$$

shows that the Taylor expansion of the exact solution contains the elementary differentials of F which are represented by the usual trees. Note that these differentials coincide with those of g except for $F' = g' - A$ and F itself. Splitting the trees that contain F', we obtain trees with two kinds of nodes. The nodes corresponding to A

are represented by open (white) circles in Table 1, whereas the evaluation of F and of derivatives of g is represented by filled (black) circles. This gives the subclass of bicolored trees with black branching nodes and black leaves.

For the numerical scheme written in the form (2.3), we define $\widetilde{F}(u) = g(u) - Au_n$. Then, (2.3) can be formally interpreted as an ordinary Runge–Kutta method with the usual trees corresponding to the elementary differentials of $\widetilde{F}$. The differentials of $\widetilde{F}$ coincide with those of g except for the evaluation of the functions themselves where $F(u_n) = \widetilde{F}(u_n)$. Moreover, taking the series expansion of the coefficients a_{ij} and b_i into account leads to the same bicolored trees as for the exact solution.

The derivation of the order conditions from the trees proceeds as follows. For sake of simplicity in presentation, we formally define

$$a_{ij}(z) = \sum_{k\geq 0} \alpha_{ij}^{(k)} z^k, \qquad b_i(z) = \sum_{k\geq 0} \beta_i^{(k)} z^k. \tag{2.5}$$

A black node preceded by another black node is interpreted as $\alpha_{ij}^{(0)}$, whereas a black root is interpreted as $\beta_i^{(0)}$. This is exactly the same interpretation as for the underlying Runge–Kutta method. Moreover, k subsequent white nodes followed by a black node are interpreted as $\beta_i^{(k)}$ or $\alpha_{ij}^{(k)}$, respectively, depending on whether they appear at the root or not, respectively. This is seen directly from (2.5) with $z = -hA$. The order conditions up to order four are displayed in Table 1.

3. Analytical framework.

Our analysis below will be based on an abstract formulation of (1.1) as an evolution equation in a Banach space $(X, \|\cdot\|)$. Let $\mathcal{D}(A)$ denote the domain of A in X. Our basic assumptions on the operator A are those of [8].

Assumption 1. Let $A : \mathcal{D}(A) \to X$ be sectorial; i.e., A is a densely defined and closed linear operator on X satisfying the resolvent condition

$$\|(\lambda I - A)^{-1}\|_{X\leftarrow X} \leq \frac{M}{|\lambda - a|} \tag{3.1}$$

on the sector $\{\lambda \in \mathbb{C} : \vartheta \leq |\arg(\lambda - a)| \leq \pi,\ \lambda \neq a\}$ for $M \geq 1$, $a \in \mathbb{R}$, and $0 < \vartheta < \pi/2$.

Under this assumption, the operator $-A$ is the infinitesimal generator of an analytic semigroup $\{\mathrm{e}^{-tA}\}_{t\geq 0}$. For $\omega > -a$, the fractional powers of $\widetilde{A} = A + \omega I$ are well defined. The following stability bounds are proved in [10]. They are crucial in our analysis.

Lemma 3.1. *Under Assumption* 1 *and for fixed* $\omega > -a$, *the following bounds hold uniformly on* $0 \leq t \leq T$:

$$\|\mathrm{e}^{-tA}\|_{X\leftarrow X} + \|t^\gamma \widetilde{A}^\gamma \mathrm{e}^{-tA}\|_{X\leftarrow X} \leq C, \qquad \gamma \geq 0, \tag{3.2a}$$

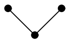
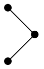
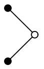
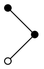
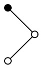
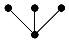
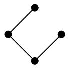
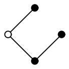
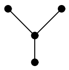
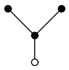
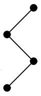
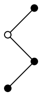
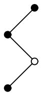
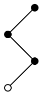
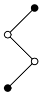
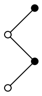
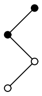
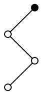
$$\left\| hA \sum_{j=1}^{n-1} \mathrm{e}^{-jhA} \right\|_{X\leftarrow X} \leq C. \tag{3.2b}$$

The next lemma (often called Abel's partial summation) is a discrete version of the integration-by-parts formula. Its proof is straightforward.

Lemma 3.2. *For* $W_k = \sum_{j=0}^{k} w_j$ *the following summation-by-parts formula holds:*

$$\sum_{j=0}^{n-1} w_j v_{n-j} = W_{n-1} v_1 - \sum_{j=0}^{n-2} W_j (v_{n-j-1} - v_{n-j}). \tag{3.3}$$

Table 1
Order trees and nonstiff order conditions for exponential Runge–Kutta methods.

No.	Tree	Order	Differential	Order condition
1		1	F	$\sum \beta_i^{(0)} = 1$
2		2	$g'F$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} = \frac{1}{2}$
3		2	AF	$\sum \beta_i^{(1)} = \frac{1}{2}$
4		3	$g''(F,F)$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} \alpha_{ik}^{(0)} = \frac{1}{3}$
5		3	$g'g'F$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} \alpha_{jk}^{(0)} = \frac{1}{6}$
6		3	$g'AF$	$\sum \beta_i^{(0)} \alpha_{ij}^{(1)} = \frac{1}{6}$
7		3	$Ag'F$	$\sum \beta_i^{(1)} \alpha_{ij}^{(0)} = \frac{1}{6}$
8		3	AAF	$\sum \beta_i^{(2)} = \frac{1}{6}$
9		4	$g'''(F,F,F)$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} \alpha_{ik}^{(0)} \alpha_{il}^{(0)} = \frac{1}{4}$
10		4	$g''(g'F,F)$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} \alpha_{ik}^{(0)} \alpha_{kl}^{(0)} = \frac{1}{8}$
11		4	$g''(AF,F)$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} \alpha_{ik}^{(1)} = \frac{1}{8}$
12		4	$g'g''(F,F)$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} \alpha_{jk}^{(0)} \alpha_{jl}^{(0)} = \frac{1}{12}$
13		4	$Ag''(F,F)$	$\sum \beta_i^{(1)} \alpha_{ij}^{(0)} \alpha_{ik}^{(0)} = \frac{1}{12}$
14		4	$g'g'g'F$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} \alpha_{jk}^{(0)} \alpha_{kl}^{(0)} = \frac{1}{24}$
15		4	$g'g'AF$	$\sum \beta_i^{(0)} \alpha_{ij}^{(0)} \alpha_{jk}^{(1)} = \frac{1}{24}$
16		4	$g'Ag'F$	$\sum \beta_i^{(0)} \alpha_{ij}^{(1)} \alpha_{jk}^{(0)} = \frac{1}{24}$
17		4	$Ag'g'F$	$\sum \beta_i^{(1)} \alpha_{ij}^{(0)} \alpha_{jk}^{(0)} = \frac{1}{24}$
18		4	$g'AAF$	$\sum \beta_i^{(0)} \alpha_{ij}^{(2)} = \frac{1}{24}$
19		4	$Ag'AF$	$\sum \beta_i^{(1)} \alpha_{ij}^{(1)} = \frac{1}{24}$
20		4	$AAg'F$	$\sum \beta_i^{(2)} \alpha_{ij}^{(0)} = \frac{1}{24}$
21		4	$AAAF$	$\sum \beta_i^{(3)} = \frac{1}{24}$

The stability estimate (3.2a) enables us to define the bounded operators

$$\varphi_j(-tA) = \frac{1}{t^j}\int_0^t \mathrm{e}^{-(t-\tau)A}\frac{\tau^{j-1}}{(j-1)!}\,\mathrm{d}\tau, \qquad j \geq 1. \tag{3.4}$$

We note for later use that $\varphi_0(z) = \mathrm{e}^z$ and

$$\varphi_{k+1}(z) = \frac{\varphi_k(z) - 1/k!}{z}, \qquad \varphi_k(0) = \frac{1}{k!}, \qquad k \geq 0. \tag{3.5}$$

Our basic assumptions on g are those of [8] and [22]. We thus choose $0 \leq \alpha < 1$ and define $V = \mathcal{D}(\widetilde{A}^\alpha) \subset X$, where $\widetilde{A}$ denotes the shifted operator $\widetilde{A} = A + \omega I$ with $\omega > -a$. The linear space V is a Banach space with norm $\|v\|_V = \|\widetilde{A}^\alpha v\|$. Note that V does not depend on ω, since different choices of ω lead to equivalent norms. Our main hypothesis on the nonlinearity g is the following.

Assumption 2. Let $g : [0,T] \times V \to X$ be locally Lipschitz-continuous in a strip along the exact solution u. Thus there exists a real number $L(R,T)$ such that

$$\|g(t,v) - g(t,w)\| \leq L\|v - w\|_V \tag{3.6}$$

for all $t \in [0,T]$ and $\max(\|v - u(t)\|_V, \|w - u(t)\|_V) \leq R$.

Example. It is well known that reaction-diffusion equations fit into this abstract framework, as well as the incompressible Navier–Stokes equations in two and three space dimensions; see, e.g., [8, Chapter 3] and [17, section 7.3].

For high-order convergence results, we have to assume more regularity.

Assumption 3. We suppose that (1.1) possesses a sufficiently smooth solution $u : [0,T] \to V$ with derivatives in V, and that $g : [0,T] \times V \to X$ is sufficiently often Fréchet differentiable in a strip along the exact solution. All occurring derivatives are supposed to be uniformly bounded.

Note that, under the above assumption, the composition

$$f : [0,T] \to X : t \mapsto f(t) = g(t,u(t))$$

is a smooth mapping, too. This will be used frequently.

4. Convergence results for exponential methods. In this section, we discuss the exponential counterparts of classical Runge–Kutta methods and study their convergence properties for the semilinear problem (1.1). More precisely, we consider methods with

$$\chi(z) = \mathrm{e}^z \qquad \text{and} \qquad \chi_i(z) = \mathrm{e}^{c_i z}, \quad 1 \leq i \leq s. \tag{4.1}$$

Our main interest lies in *explicit* methods for which, due to $c_1 = 0$,

$$\chi_1(z) = 1 \qquad \text{and} \qquad a_{ij}(z) = 0, \qquad 1 \leq i \leq j \leq s. \tag{4.2}$$

Implicit methods have been analyzed in detail in our recent paper [10]. The scheme (2.1) satisfying (4.1) and (4.2) is called an *explicit exponential Runge–Kutta method* henceforth. For the coefficients of the exponential Runge–Kutta method (2.1) we assume stability assumptions similar to (3.2a)

$$\|\phi(-tA)\|_{X \leftarrow X} + \left\|t^\gamma \widetilde{A}^\gamma \phi(-tA)\right\|_{X \leftarrow X} \leq C, \qquad 0 \leq \gamma \leq 1, \tag{4.3}$$

for $\phi = b_i$ or $\phi = a_{ij}$, $i,j = 1,\ldots,s$. This assumption is satisfied by all methods considered below, since $b_i(z)$ and $a_{ij}(z)$ are (linear) combinations of the functions $\varphi_k(z)$ and $\varphi_k(c_l z)$, respectively.

4.1. Error recursion and representation of the defects. In order to simplify the notation, we again set $f(t) = g(t, u(t))$. Our proofs are heavily based on the representation of the exact solution by the variation-of-constants formula

$$u(t_n + \theta h) = \mathrm{e}^{-\theta h A} u(t_n) + \int_0^{\theta h} \mathrm{e}^{-(\theta h - \tau)A} f(t_n + \tau)\, \mathrm{d}\tau. \tag{4.4}$$

We expand f into a Taylor series with remainder in integral form to get

$$f(t_n + \tau) = \sum_{j=1}^{q} \frac{\tau^{j-1}}{(j-1)!} f^{(j-1)}(t_n) + \int_0^{\tau} \frac{(\tau - \sigma)^{q-1}}{(q-1)!} f^{(q)}(t_n + \sigma)\, \mathrm{d}\sigma. \tag{4.5}$$

On the one hand, inserting this into the right-hand side of (4.4) yields

$$\begin{aligned} u(t_n + c_i h) = \mathrm{e}^{-c_i h A} u(t_n) &+ \sum_{j=1}^{q_i} (c_i h)^j \varphi_j(-c_i h A) f^{(j-1)}(t_n) \\ &+ \int_0^{c_i h} \mathrm{e}^{-(c_i h - \tau)A} \int_0^{\tau} \frac{(\tau - \sigma)^{q_i - 1}}{(q_i - 1)!} f^{(q_i)}(t_n + \sigma)\, \mathrm{d}\sigma\, \mathrm{d}\tau. \end{aligned} \tag{4.6}$$

On the other hand, inserting the exact solution into the numerical scheme gives

$$u(t_n + c_i h) = \mathrm{e}^{-c_i h A} u(t_n) + h \sum_{j=1}^{i-1} a_{ij}(-hA) f(t_n + c_j h) + \Delta_{ni}, \tag{4.7a}$$

$$u(t_{n+1}) = \mathrm{e}^{-hA} u(t_n) + h \sum_{i=1}^{s} b_i(-hA) f(t_n + c_i h) + \delta_{n+1} \tag{4.7b}$$

with defects Δ_{ni} and δ_{n+1}. Substituting (4.5) into (4.7a), we obtain

$$\begin{aligned} u(t_n + c_i h) = \mathrm{e}^{-c_i h A} u(t_n) &+ h \sum_{k=1}^{i-1} a_{ik}(-hA) \sum_{j=1}^{q_i} \frac{(c_k h)^{j-1}}{(j-1)!} f^{(j-1)}(t_n) \\ &+ h \sum_{k=1}^{i-1} a_{ik}(-hA) \int_0^{c_k h} \frac{(c_k h - \sigma)^{q_i - 1}}{(q_i - 1)!} f^{(q_i)}(t_n + \sigma)\, \mathrm{d}\sigma + \Delta_{ni}. \end{aligned} \tag{4.8}$$

Subtracting (4.6) from (4.8) gives the following explicit representation of the defects,

$$\Delta_{ni} = \sum_{j=1}^{q_i} h^j \psi_{j,i}(-hA) f^{(j-1)}(t_n) + \Delta_{ni}^{[q_i]}, \tag{4.9}$$

with remainders $\Delta_{ni}^{[q_i]}$ and with

$$\psi_{j,i}(-hA) = \varphi_j(-c_i h A) c_i^j - \sum_{k=1}^{i-1} a_{ik}(-hA) \frac{c_k^{j-1}}{(j-1)!}. \tag{4.10}$$

Similarly, we get for the defects at time t_{n+1}

$$\delta_{n+1} = \sum_{j=1}^{q} h^j \psi_j(-hA) f^{(j-1)}(t_n) + \delta_{n+1}^{[q]}, \tag{4.11}$$

where

$$\psi_j(-hA) = \varphi_j(-hA) - \sum_{k=1}^{s} b_k(-hA)\frac{c_k^{j-1}}{(j-1)!}. \tag{4.12}$$

For the remainders in (4.9) and (4.11), we have the following estimate.

LEMMA 4.1. *Let* $0 < \nu \le 1$ *and* $\widetilde{A}^{\nu-1} f^{(r)} \in L^\infty(0,T;V)$. *Then,*

$$h^{1-\nu}\|\Delta_{ni}^{[r]}\|_V + \|\widetilde{A}^{\nu-1}\Delta_{ni}^{[r]}\|_V \le Ch^{r+1} \sup_{0\le\tau\le 1} \|\widetilde{A}^{\nu-1} f^{(r)}(t_n+\tau h)\|_V, \tag{4.13a}$$

$$\left\| \sum_{j=0}^{n-1} \mathrm{e}^{-jhA}\delta_{n-j}^{[r]} \right\|_V \le Ch^r \sup_{0\le t\le t_n} \|\widetilde{A}^{\nu-1} f^{(r)}(t)\|_V \tag{4.13b}$$

holds with a constant C, *uniformly in* $0 \le t_n \le T$.

Proof. Both estimates follow at once from the stability bound

$$\left\|t^\gamma \widetilde{A}^\gamma \mathrm{e}^{-tA}\right\|_{V\leftarrow V} + \left\|t^\gamma \widetilde{A}^\gamma \phi(-tA)\right\|_{V\leftarrow V} \le C, \qquad 0 \le \gamma \le 1,$$

for $\phi = b_i$ or $\phi = a_{ij}$, $i,j = 1,\dots,s$. The latter is a consequence of (3.2a) and (4.3), respectively. □

Let $e_n = u_n - u(t_n)$ and $E_{ni} = U_{ni} - u(t_n + c_i h)$ denote the differences between the numerical and the exact solutions. Subtracting (4.7) from the numerical method (2.1) satisfying (4.1) and (4.2) gives the error recursion

$$E_{ni} = \mathrm{e}^{-c_i hA} e_n + h\sum_{j=1}^{i-1} a_{ij}(-hA)\Big(g(t_n+c_jh, U_{nj}) - f(t_n+c_jh)\Big) - \Delta_{ni}, \tag{4.14a}$$

$$e_{n+1} = \mathrm{e}^{-hA} e_n + h\sum_{i=1}^{s} b_i(-hA)\Big(g(t_n+c_ih, U_{ni}) - f(t_n+c_ih)\Big) - \delta_{n+1}. \tag{4.14b}$$

We will derive bounds for these errors.

4.2. Convergence of the exponential Euler method. For $s = 1$, the only reasonable choice is the exponential form of Euler's method. Applied to (1.1), it is

$$u_{n+1} = \mathrm{e}^{-hA} u_n + h\varphi_1(-hA) g(t_n, u_n). \tag{4.15}$$

For this method, we have the following convergence result.

THEOREM 4.2. *Let the initial value problem* (1.1) *satisfy Assumptions* 1–2 *and consider for its numerical solution the exponential Euler method* (4.15). *Further assume that* $f : [0,T] \to X$ *is differentiable and that* $\beta \in (0,1]$ *is such that* $\widetilde{A}^{\beta-1} f' \in L^\infty(0,T;V)$. *Then, the error bound*

$$\|u_n - u(t_n)\|_V \le C\cdot h \sup_{0\le t\le t_n} \|\widetilde{A}^{\beta-1} f'(t)\|_V$$

holds uniformly in $0 \le nh \le T$. *The constant* C *depends on* T, *but it is independent of* n *and* h.

Proof. The exponential Euler method satisfies the error recursion

$$e_{n+1} = \mathrm{e}^{-hA} e_n + h\varphi_1(-hA)\big(g(t_n,u_n) - f(t_n)\big) - \delta_{n+1} \tag{4.16}$$

with defects $\delta_{n+1} = \delta_{n+1}^{[1]}$ given by (4.7b) and (4.11). Solving recursion (4.16) gives

$$e_n = h \sum_{j=0}^{n-1} \mathrm{e}^{-(n-j-1)hA} \varphi_1(-hA)\big(g(t_j, u_j) - f(t_j)\big) - \sum_{j=0}^{n-1} \mathrm{e}^{-jhA} \delta_{n-j}.$$

Using (3.2a), Assumption 2, and Lemma 4.1, we may estimate this in V by

$$\|e_n\|_V \le Ch \sum_{j=0}^{n-2} t_{n-j-1}^{-\alpha} \|e_j\|_V + Ch^{1-\alpha}\|e_{n-1}\|_V + Ch \sup_{0\le t\le t_n} \|\widetilde{A}^{\beta-1} f'(t)\|_V.$$

The application of a discrete Gronwall lemma [10, Lemma 4] now concludes the proof. □

Remark. The above theorem can also be deduced from [10], since the exponential Euler method is a collocation method with $s = 1$ and $c_1 = 0$.

4.3. Convergence results for second-order methods. We first derive a bound for the errors of the internal stages. Due to (2.2) and (4.1), we always have $\psi_{1,j} = 0$. From (4.9) and (4.14a) we thus get, with the help of Assumption 2,

$$\|E_{ni}\|_V \le C\,\|e_n\|_V + Ch^{1-\alpha} \max_{2\le j\le i-1} \|E_{nj}\|_V + \|\Delta_{ni}^{[1]}\|_V.$$

Applying Lemma 4.1 then shows the a priori bound

$$\|E_{ni}\|_V \le C\,\|e_n\|_V + Ch^{1+\nu} \sup_{0\le\tau\le 1} \|\widetilde{A}^{\nu-1} f'(t_n+\tau h)\|_V. \tag{4.17}$$

For second-order methods, we will satisfy the order conditions

$$\psi_1(-hA) = 0, \qquad \psi_2(-hA) = 0 \tag{4.18}$$

and take $q = 2$ in (4.11). We are now in the position to state the convergence theorem.

THEOREM 4.3. *Let the initial value problem* (1.1) *satisfy Assumptions* 1 *and* 2*, and consider for its solution an exponential Runge–Kutta method* (2.1) *satisfying* (4.1)*,* (4.2)*, and* (4.18)*. Further assume that $f : [0, T] \to X$ is twice differentiable and that $\beta, \kappa \in (0, 1]$ are such that $\widetilde{A}^{\beta-1} f' \in L^\infty(0, T; V)$ and $\widetilde{A}^{\kappa-1} f'' \in L^\infty(0, T; V)$. Then, the error bound*

$$\|u_n - u(t_n)\|_V \le C \cdot h^{1+\beta} \sup_{0\le t\le t_n} \|\widetilde{A}^{\beta-1} f'(t)\|_V + C \cdot h^2 \sup_{0\le t\le t_n} \|\widetilde{A}^{\kappa-1} f''(t)\|_V$$

holds uniformly in $0 \le nh \le T$. The constant C depends on T, but it is independent of n and h.

Remark. If $f', f'' \in L^\infty(0, T; X)$, the above theorem is applicable with $\beta = \kappa = 1 - \alpha$. For $\beta = 1$, the theorem yields order 2. Under the slightly weaker regularity assumptions $f'(0) \in V$ and $f'' \in L^1(0, T; V)$, we obtain an alternative second-order error bound

$$\|u_n - u(t_n)\|_V \le C \cdot h^2 \left(\|f'(0)\|_V + \int_0^{t_n} \|f''(\tau)\|_V \,\mathrm{d}\tau \right).$$

This follows easily from the proof of Theorem 4.3.

Proof. Solving the error recursion (4.14b) in the same way as in the proof of Theorem 4.2 gives

$$\|e_n\|_V \le Ch^{1-\alpha}\|E_{n-1}\|_V + Ch\sum_{j=0}^{n-2} t_{n-j-1}^{-\alpha}\|E_j\|_V + \Big\|\sum_{j=0}^{n-1} \mathrm{e}^{-jhA}\delta_{n-j}\Big\|_V$$

with $\|E_j\|_V = \max_{2\le k\le s}\|E_{jk}\|_V$. After inserting the bounds (4.17) and (4.13b), the proof is concluded by applying a discrete Gronwall lemma. □

If $g : [0,T]\times V \to X$ is twice differentiable, its derivatives

$$J_n = \frac{\partial g}{\partial u}\big(t_n, u(t_n)\big), \qquad K_n = \frac{\partial^2 g}{\partial t\partial u}\big(t_n, u(t_n)\big)$$

are bounded operators from V to X.

Lemma 4.4. *Under Assumption* 3*, we have*

$$g(t_n + c_ih, U_{ni}) - f(t_n + c_ih) = J_nE_{ni} + c_ihK_nE_{ni} + Q_{ni}, \tag{4.19a}$$

$$g(t_n, u_n) - f(t_n) = J_ne_n + Q_n \tag{4.19b}$$

with remainders Q_{ni} *and* Q_n *satisfying the bounds*

$$\|Q_{ni}\| \le C\cdot\big(h^2 + \|E_{ni}\|_V\big)\|E_{ni}\|_V, \qquad \|Q_n\| \le C\cdot\|e_n\|_V^2,$$

as long as the errors E_{ni} *and* e_n *remain in a sufficiently small neighborhood of 0.*

Proof. Using Taylor series expansion, we get

$$\begin{aligned} g(t_n + c_ih, U_{ni}) - f(t_n + c_ih) &= \frac{\partial g}{\partial u}\big(t_n + c_ih, u(t_n + c_ih)\big)E_{ni} \\ &+ \int_0^1 (1-\tau)\frac{\partial^2 g}{\partial u^2}\big(t_n + c_ih, u(t_n + c_ih) + \tau E_{ni}\big)(E_{ni}, E_{ni})\,\mathrm{d}\tau. \end{aligned}$$

Expanding the first term on the right-hand side at t_n yields the desired result. □

Inserting (4.19a) into the recursion (4.14b) shows that the main contribution of the defects to the global error is given by the term

$$h\sum_{j=0}^{n-1}\mathrm{e}^{-jhA}\sum_{i=1}^{s} b_i(-hA)J_n\Delta_{n-j-1,i}^{[1]}. \tag{4.20}$$

Let $\gamma < 1$ and $0 \le \mu \le 1-\beta$ be such that

$$\big\|\widetilde{A}^{-\gamma}J_n\,\widetilde{A}^{\mu}\big\|_{V\leftarrow V} \le C. \tag{4.21}$$

This estimate holds trivially for $\gamma = \alpha$ and $\mu = 0$. But there are more favorable situations in which $\mu > 0$. Then, the expression (4.20) is bounded by

$$C\cdot h^{1+\beta+\mu}\sup_{0\le t\le t_n}\|\widetilde{A}^{\beta-1}f'(t)\|_V, \tag{4.22}$$

which slightly improves the error bound of the theorem. We do not elaborate this point here further.

4.4. Convergence results for higher-order methods. An explicit expansion of E_{ni} in terms of Δ_{ni} and e_n is easily obtained with the help of Lemma 4.4 by inserting the expressions of the lemma recursively into (4.14a). Depending on the size of s and α, however, this expansion will consist of a great number of terms that all give rise to order conditions for the method. For example, the arising term

$$h^{\ell+2}\sum_{j_1=1}^{i-1} a_{ij_1}(-hA)J_n\cdots\sum_{j_\ell=1}^{j_{\ell-1}} a_{j_{\ell-1}j_\ell}(-hA)J_n\cdot\psi_{2,j_\ell}(-hA)f'(t_n) \tag{4.23}$$

has classical order $\ell+2$ and can be neglected for $\ell\ge 2$ in a fourth-order method. In our context, however, it has only order $1+(\ell+1)(1-\alpha)$, and this might be a small number, even for large ℓ. In that case, the term can no longer be neglected.

To keep our presentation simple, we will consider in the remainder of this section only the case $\alpha=0$. It is therefore no longer necessary to distinguish between the spaces V and X.

LEMMA 4.5. *Under the assumptions of Lemma* 4.4 *and* $\alpha=0$, *there exist bounded operators* $\mathcal{N}_{ni}(e_n)$ *on* X *such that*

$$\begin{aligned}E_{ni} = \mathcal{N}_{ni}(e_n)e_n &- h^2\psi_{2,i}(-hA)f'(t_n) - h^3\psi_{3,i}(-hA)f''(t_n)\\ &- h^3\sum_{j=2}^{i-1} a_{ij}(-hA)J_n\psi_{2,j}(-hA)f'(t_n) + h^4\mathcal{R}_{ni},\end{aligned}$$

with uniformly bounded remainders $\|\mathcal{R}_{ni}\|\le C$.

Proof. The formula follows easily from recursion (4.14a), the representation of the defects (4.9), and from Lemma 4.4 by an induction argument. □

An important consequence of the above lemma is the following representation of the errors e_n. It is the key result for obtaining the stiff order conditions.

LEMMA 4.6. *Under the assumptions of Lemma* 4.4 *and* $\alpha=0$, *there exist uniformly bounded operators* $\mathcal{N}_n(e_n)$ *on* X *such that the global errors* e_n *satisfy the recursion*

$$\begin{aligned}e_{n+1} = \mathrm{e}^{-hA}e_n &+ h\mathcal{N}_n(e_n)e_n - h^2\psi_2(-hA)f'(t_n)\\ &- h^3\psi_3(-hA)f''(t_n) - h^3\sum_{i=1}^{s} b_i(-hA)J_n\psi_{2,i}(-hA)f'(t_n)\\ &- h^4\psi_4(-hA)f'''(t_n) - h^4\sum_{i=1}^{s} b_i(-hA)J_n\psi_{3,i}(-hA)f''(t_n)\\ &- h^4\sum_{i=1}^{s} b_i(-hA)J_n\sum_{j=2}^{i-1} a_{ij}(-hA)J_n\psi_{2,j}(-hA)f'(t_n)\\ &- h^4\sum_{i=1}^{s} b_i(-hA)c_iK_n\psi_{2,j}(-hA)f'(t_n) + h^5\mathcal{R}_n,\end{aligned} \tag{4.24}$$

with uniformly bounded remainders $\|\mathcal{R}_n\|\le C$.

Proof. The formula follows easily from recursion (4.14b), the representation of the defects (4.11), and Lemma 4.5. □

The stiff order conditions can easily be identified in (4.24). For clarity, we have collected them in Table 2. With these preparations, we are now in the position to formulate a more general convergence result.

TABLE 2

Stiff order conditions for explicit exponential Runge–Kutta methods for $\alpha = 0$. Here J and K denote arbitrary bounded operators on X. The functions ψ_i and $\psi_{k,\ell}$ are defined in (4.12) *and* (4.10)*, respectively.*

No.	Order	Order condition
1	1	$\psi_1(-hA) = 0$
2	2	$\psi_2(-hA) = 0$
3	2	$\psi_{1,i}(-hA) = 0$
4	3	$\psi_3(-hA) = 0$
5	3	$\sum_{i=1}^{s} b_i(-hA) J \psi_{2,i}(-hA) = 0$
6	4	$\psi_4(-hA) = 0$
7	4	$\sum_{i=1}^{s} b_i(-hA) J \psi_{3,i}(-hA) = 0$
8	4	$\sum_{i=1}^{s} b_i(-hA) J \sum_{j=2}^{i-1} a_{ij}(-hA) J \psi_{2,j}(-hA) = 0$
9	4	$\sum_{i=1}^{s} b_i(-hA) c_i K \psi_{2,i}(-hA) = 0$

THEOREM 4.7. *Let the initial value problem* (1.1) *satisfy Assumptions* 1–3 *with* $\alpha = 0$*, and consider for its numerical solution an explicit exponential Runge–Kutta method* (2.1) *satisfying* (4.1) *and* (4.2)*. For* $2 \le p \le 4$*, assume that the order conditions of Table* 2 *hold up to order* $p-1$ *and that* $\psi_p(0) = 0$*. Further assume that the remaining conditions of order* p *hold in a weaker form with* $b_i(0)$ *instead of* $b_i(-hA)$ *for* $2 \le i \le s$*. Then, the numerical solution* u_n *satisfies the error bound*

$$\|u_n - u(t_n)\| \le C \cdot h^p,$$

uniformly in $0 \le nh \le T$*. The constant* C *depends on* T*, but it is independent of* n *and* h*.*

Proof. Inserting the order conditions into the recursion of Lemma 4.6 yields

$$e_{n+1} = \mathrm{e}^{-hA} e_n + h\mathcal{N}_n(e_n)e_n + h^p \mathcal{T}_n + h^{p+1}\mathcal{R}_n$$

with bounded remainders $\mathcal{R}_n$ depending on p. Here, $\mathcal{T}_n$ denotes the terms multiplying h^p in Lemma 4.6. Solving the above error recursion gives

$$e_n = h \sum_{j=0}^{n-1} \mathrm{e}^{-(n-j)hA} \mathcal{N}_j(e_j)e_j + h^p \sum_{j=0}^{n-1} \mathrm{e}^{-jhA} \mathcal{T}_{n-j-1} + h^{p+1} \sum_{j=0}^{n-1} \mathrm{e}^{-jhA} \mathcal{R}_{n-j-1}.$$

In order to bound the second term on the right-hand side, we use the assumptions on the conditions of order p and apply Lemma 4.8 below. Using further the stability estimate (3.2a) and the bounds for $\mathcal{R}_j$ and for $\mathcal{N}_j(e_j)$ finally yields

$$\|e_n\| \le Ch \sum_{j=0}^{n-1} \|e_j\| + Ch^p.$$

Thus, an application of the classical Gronwall lemma concludes the proof. ∎

The following lemma was used in the above proof.

Lemma 4.8. *Under the above assumptions, let $\varrho_i : [0,T] \to X$ for $1 \le i \le s$ denote differentiable functions with bounded derivatives. Then*

$$\Big\| \sum_{j=0}^{n-1} \mathrm{e}^{-jhA} \sum_{i=1}^{s} \big(b_i(-hA) - b_i(0)\big) \varrho_i(t_{n-j-1}) \Big\| \le C. \tag{4.25}$$

Proof. We first note that there exist bounded operators $\widetilde{b}_i(-hA)$ with

$$b_i(-hA) - b_i(0) = hA \cdot \widetilde{b}_i(-hA).$$

The bound (4.25) now follows at once from Lemma 3.2 with $w_j = \mathrm{e}^{-jhA} hA \cdot \widetilde{b}_i(-hA)$ and $v_j = \varrho_i(t_{j-1})$ by using the stability bound (3.2b). □

Remark. The proof of Theorem 4.7 does not extend immediately to variable step sizes. The reason for this lies in the use of the summation-by-parts formula (3.3). Assuming, however, that the method satisfies the conditions of order p in the strong sense of Table 2, then Lemma 4.8 is no longer needed. In that case, the theorem obviously holds for variable step sizes, too.

4.5. Existence of explicit methods of arbitrarily high order. In our recent paper [10] we have shown that, under the above assumptions, an s-stage exponential Runge–Kutta method of collocation type with coefficients $\widehat{a}_{ij}(-hA)$ and $\widehat{b}_i(-hA)$ satisfies the error bound

$$\|\widehat{u}_n - u(t_n)\|_V \le C\, h^s \sup_{0 \le t \le T} \|f^{(s)}(t)\|,$$

uniformly on $0 \le t_n \le T$. It is further shown there that the equations for the internal stages can be solved by fixed-point iteration,

$$\widehat{U}_{ni}^{(k)} = \mathrm{e}^{-c_i hA} \widehat{u}_n + h \sum_{i=1}^{s} \widehat{a}_{ij}(-hA)\, g\big(t_n + c_j h, \widehat{U}_{nj}^{(k-1)}\big), \qquad \widehat{U}_{nj}^{(0)} = \widehat{u}_n.$$

For $\alpha = 0$, we obviously gain one power of h in each iteration. Performing s iterations and setting $U_{ni} = \widehat{U}_{ni}^{(s)}$, $1 \le i \le s$, and

$$u_{n+1} = \mathrm{e}^{-hA} u_n + h \sum_{i=1}^{s} \widehat{b}_i(-hA)\, g\big(t_n + c_j h, U_{nj}\big)$$

thus defines an *explicit* exponential Runge–Kutta method of order s with s^2 stages. The construction shows that there exist explicit exponential Runge–Kutta methods of arbitrarily high order. For general α, however, we gain only $h^{1-\alpha}$ in each iteration. Therefore, a lot of explicit stages might be necessary for obtaining order s.

5. A discussion of explicit methods of orders up to four. In this section, we consider some examples of methods up to order four. Since the first-order exponential Euler method has already been analyzed in section 4.2, we commence here with second-order methods. In the Butcher tableaus below, we use the abbreviations

$$\varphi_{i,j} = \varphi_{i,j}(-hA) = \varphi_i(-c_j hA), \qquad 2 \le j \le s. \tag{5.1}$$

5.1. Second-order methods. Second-order methods require two internal stages at least. For two stages, the order conditions are

$$b_1(-hA) + b_2(-hA) = \varphi_1(-hA), \tag{5.2a}$$

$$b_2(-hA)c_2 = \varphi_2(-hA), \tag{5.2b}$$

$$a_{21}(-hA) = c_2\varphi_1(-c_2hA). \tag{5.2c}$$

A straightforward elimination leads to the following one-parameter family of exponential Runge–Kutta methods

$$\begin{array}{c|cc} 0 & & \\ c_2 & c_2\,\varphi_{1,2} & \\ \hline & \varphi_1 - \frac{1}{c_2}\varphi_2 & \frac{1}{c_2}\varphi_2 \end{array}. \tag{5.3}$$

The coefficients are displayed as usual in a Butcher tableau. It is also possible to omit the function φ_2 by weakening condition (5.2b) to $b_2(0)c_2 = \varphi_2(0) = \frac{1}{2}$. This yields another one-parameter family of methods

$$\begin{array}{c|cc} 0 & & \\ c_2 & c_2\,\varphi_{1,2} & \\ \hline & (1-\frac{1}{2c_2})\varphi_1 & \frac{1}{2c_2}\varphi_1 \end{array}. \tag{5.4}$$

Note that the choice $c_2 = \frac{1}{2}$ yields $b_1 = 0$.

Methods (5.3) and (5.4) have been proposed already by Strehmel and Weiner [25, Example 4.2.2] in the context of adaptive Runge–Kutta methods, where the functions φ_j are usually approximated by certain rational functions. It is shown in [25, Section 4.5.3] that both methods are B-consistent of order one.

Method (5.3) satisfies the assumptions of Theorem 4.3 and thus converges with order $1+\beta$, in general. Method (5.4) can be analyzed in a similar way, as it differs from (5.3) in the choice of the quadrature rule for u_{n+1} only. The defects at the grid points now have the form

$$\begin{aligned} \delta_{n+1} &= \frac{h^2}{2}\big(2\varphi_2(-hA) - \varphi_1(-hA)\big)f'\left(t_n + \frac{h}{2}\right) \\ &\quad + \int_0^h \mathrm{e}^{-(h-\tau)A}\int_{h/2}^{\tau}(\tau-\xi)f''\left(t_n+\xi\right)\,\mathrm{d}\xi. \end{aligned} \tag{5.5}$$

Since $2\varphi_2(0) - \varphi_1(0) = 0$, the first term of (5.5) contributes with

$$C\cdot h^{1+\beta}\left(\|\widetilde{A}^{\beta-1}f'(0)\|_V + \int_0^{t_n}\|\widetilde{A}^{\beta-1}f''(t)\|_V\,\mathrm{d}t\right) \tag{5.6}$$

to the global error. This is seen with the help of Lemma 3.2. Thus, Theorem 4.3 holds with the additional error term (5.6) for the modified method (5.4), too. In particular, we get a second-order error bound for $\beta = 1$. Note, however, that (5.6) cannot be improved under a condition of the type (4.21). For situations in which $f'(t), f''(t) \notin V$, method (5.3) is therefore preferable to method (5.4), which can be affected by an order reduction; see also Figure 6.3 below.

5.2. Third-order methods. We will continue with 3-stage methods. In the following, we assume $\alpha = 0$. The order conditions for 3-stage methods are

$$b_1(-hA) + b_2(-hA) + b_3(-hA) = \varphi_1(-hA), \tag{5.7a}$$

$$b_2(-hA)c_2 + b_3(-hA)c_3 = \varphi_2(-hA), \tag{5.7b}$$

$$a_{21}(-hA) = c_2\varphi_1(-c_2hA), \tag{5.7c}$$

$$a_{31}(-hA) + a_{32}(-hA) = c_3\varphi_1(-c_3hA), \tag{5.7d}$$

$$b_2(-hA)c_2^2 + b_3(-hA)c_3^2 = 2\varphi_3(-hA), \tag{5.7e}$$

$$b_2(-hA)Jc_2^2\varphi_2(-c_2hA) + b_3(-hA)J\psi_{2,3} = 0, \tag{5.7f}$$

where

$$\psi_{2,3} = c_3^2\varphi_2(-c_3hA) - c_2a_{32}(-hA).$$

Condition (5.7f) can be satisfied by $b_2 = 0$ and $\psi_{2,3} = 0$ or $b_2 = \gamma b_3$ and $c_2^2\varphi_{2,2} + \gamma\psi_{2,3} = 0$. However, both choices contradict conditions (5.7b) and (5.7e). We thus weaken (5.7e) to

$$b_2(0)c_2^2 + b_3(0)c_3^2 = 2\varphi_3(0) = \frac{1}{3}. \tag{5.7g}$$

The choice $b_2 = 0$ leads to the following one-parameter family of third-order methods:

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2\varphi_{1,2} & & \\ \frac{2}{3} & \frac{2}{3}\varphi_{1,3} - \frac{4}{9c_2}\varphi_{2,3} & \frac{4}{9c_2}\varphi_{2,3} & \\ \hline & \varphi_1 - \frac{3}{2}\varphi_2 & 0 & \frac{3}{2}\varphi_2 \end{array} . \tag{5.8}$$

The other choice $b_2 = \gamma b_3$ leads to the two-parameter family of methods of order three

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2\varphi_{1,2} & & \\ c_3 & c_3\varphi_{1,3} - a_{32} & \gamma c_2\varphi_{2,2} + \frac{c_3^2}{c_2}\varphi_{23} & \\ \hline & \varphi_1 - b_2 - b_3 & \frac{\gamma}{\gamma c_2 + c_3}\varphi_2 & \frac{1}{\gamma c_2 + c_3}\varphi_2 \end{array} , \tag{5.9}$$

where γ, c_2, and c_3 have to satisfy the restriction

$$2(\gamma c_2 + c_3) = 3(\gamma c_2^2 + c_3^2). \tag{5.10}$$

Another possibility is to weaken (5.7f) to

$$b_2(0)c_2^2\varphi_2(-c_2hA) + b_3(0)\big(c_3^2\varphi_2(-c_3hA) - c_2a_{32}(-hA)\big) = 0. \tag{5.11}$$

The order conditions of the underlying Runge–Kutta method of order three show that

$$b_2(0) = \frac{2 - 3c_3}{6c_2(c_2 - c_3)}, \qquad b_3(0) = \frac{2 - 3c_2}{6c_3(c_3 - c_2)};$$

see [7, Exercise II.1.4]. Here, we have to choose $c_2 \neq 2/3$ and $c_2 \neq c_3$ except for $c_2 = c_3 = 2/3$ where $b_2(0) + b_3(0) = 3/4$.

A three-parameter family of methods that involve φ_1 and φ_2 only is given by

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2\varphi_{1,2} & & \\ c_3 & c_3\varphi_{1,3} - a_{32} & \frac{c_3^2}{c_2}\varphi_{2,3} - \frac{2-3c_3}{2-3c_2}c_3\varphi_{2,2} & \\ \hline & \varphi_1 - b_2 - b_3 & \alpha_2\varphi_1 + \beta_2\varphi_2 & \alpha_3\varphi_1 + \beta_3\varphi_2 \end{array} . \tag{5.12}$$

In addition to c_2 and c_3, one of the coefficients α_i, β_i can be chosen arbitrarily, and the remaining ones are determined by the linear system

$$c_2\alpha_2 + c_3\alpha_3 = 0, \qquad \alpha_3 + \frac{1}{2}\beta_3 = b_3(0), \qquad c_2\beta_2 + c_3\beta_3 = 1.$$

Note that for $\alpha_2 = 0$ and $\beta_2 = \gamma\beta_3$ we recover condition (5.10) and obtain the scheme (5.9). Setting $\alpha_2 = \beta_2 = 0$ leads to $b_2(-hA) = 0$ and the condition $c_3 = 2/3$, which results in the scheme (5.8).

We next discuss some related methods which can be found in the literature.

Strehmel and Weiner [25, Example 4.5.4] proved that for a 3-stage adaptive Runge–Kutta method with second-order B-consistency, the condition $b_2 = 0$ is necessary. They proposed the following family of such methods:

$$\begin{array}{c|ccc} 0 & & & \\ c_2 & c_2\varphi_{1,2} & & \\ 1 & \varphi_{1,3} - \frac{1}{c_2}\varphi_{2,3} & \frac{1}{c_2}\varphi_{2,3} & \\ \hline & \varphi_1 - \varphi_2 & 0 & \varphi_2 \end{array} .$$

These methods satisfy conditions (5.7a), (5.7b), (5.7c), and (5.7d), i.e., all conditions of Table 2 up to order two. However, the conditions (5.7f) and (5.7g) of order three are not satisfied, not even in a weak form. Therefore, these methods are of second order only. Since they are more expensive than the 2-stage methods proposed in the previous section, the latter should be preferred for semilinear parabolic problems.

Cox and Matthews [3] constructed a method called ETD3RK, which reads

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & \\ 1 & -\varphi_{1,3} & 2\varphi_{1,3} & \\ \hline & 4\varphi_3 - 3\varphi_2 + \varphi_1 & -8\varphi_3 + 4\varphi_2 & 4\varphi_3 - \varphi_2 \end{array} . \tag{5.13}$$

This method uses φ_3 and satisfies conditions (5.7a), (5.7b), (5.7c), (5.7d), and (5.7g). However, condition (5.7f) is satisfied only in a very weak form (where all arguments are evaluated for $A = 0$). This leads to an order reduction to order two in the worst case. The same conditions are satisfied by ETD2RK3 from [3].

The method ETD2CF3 is a variant of the commutator-free Lie group method CF3 due to Celledoni, Marthinsen, and Owren [2]. It is given by

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{3} & \frac{1}{3}\varphi_{1,2} & & \\ \frac{2}{3} & \frac{2}{3}\varphi_{1,3} - \frac{4}{3}\varphi_{2,3} & \frac{4}{3}\varphi_{2,3} & \\ \hline & \varphi_1 - \frac{9}{2}\varphi_2 + 9\varphi_3 & 6\varphi_2 - 18\varphi_3 & -\frac{3}{2}\varphi_2 + 9\varphi_3 \end{array} .$$

This method satisfies (5.7a)–(5.7c) and weak forms of (5.7d) and (5.7f), where $b_i(-hA)$ is evaluated for $A = 0$. Therefore this method is of order three for $\alpha = 0$.

5.3. Fourth-order methods. Again we assume $\alpha = 0$. The order conditions for s-stage methods up to order four are

$$\sum_{i=1}^{s} b_i(-hA) = \varphi_1(-hA), \tag{5.14a}$$

$$\sum_{i=2}^{s} b_i(-hA)c_i = \varphi_2(-hA), \tag{5.14b}$$

$$\sum_{j=1}^{i-1} a_{ij}(-hA) = c_i\varphi_1(-c_ihA), \tag{5.14c}$$

$$\sum_{i=2}^{s} b_i(-hA)c_i^2 = 2\varphi_3(-hA), \tag{5.14d}$$

$$\sum_{i=2}^{s} b_i(-hA)J\Bigg(\varphi_2(-c_ihA)c_i^2 - \sum_{j=2}^{i-1} a_{ij}(-hA)c_j\Bigg) = 0, \tag{5.14e}$$

$$\sum_{i=2}^{s} b_i(-hA)c_i^3 = 6\varphi_4(-hA), \tag{5.14f}$$

$$\sum_{i=2}^{s} b_i(-hA)J\Bigg(\varphi_3(-c_ihA)c_i^3 - \frac{1}{2}\sum_{j=2}^{i-1} a_{ij}(-hA)c_j^2\Bigg) = 0, \tag{5.14g}$$

$$\sum_{i=2}^{s} b_i(-hA)J\sum_{j=2}^{i-1} a_{ij}(-hA)J\Bigg(\varphi_2(-c_jhA)c_j^2 - \sum_{k=2}^{j-1} a_{jk}(-hA)c_k\Bigg) = 0, \tag{5.14h}$$

$$\sum_{i=2}^{s} b_i(-hA)c_iK\Bigg(\varphi_2(-c_ihA)c_i^2 - \sum_{j=2}^{i-1} a_{ij}(-hA)c_j\Bigg) = 0. \tag{5.14i}$$

From (5.14a), (5.14b), and (5.14d) we deduce that any fourth-order method has to involve φ_1, φ_2, and φ_3. By Theorem 4.7 it is sufficient to satisfy condition 6 of Table 2 in the weakened form $\psi_4(0) = 0$, i.e., to replace (5.14f) by

$$\sum_{i=2}^{s} b_i(0)c_i^3 = \frac{1}{4}. \tag{5.14j}$$

Cox and Matthews [3] proposed the following exponential variant of the classical Runge–Kutta method:

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & & \\
\frac{1}{2} & 0 & \frac{1}{2}\varphi_{1,3} & & \\
1 & \frac{1}{2}\varphi_{1,3}\big(\varphi_{0,3}-1\big) & 0 & \varphi_{1,3} & \\
\hline
 & \varphi_1 - 3\varphi_2 + 4\varphi_3 & 2\varphi_2 - 4\varphi_3 & 2\varphi_2 - 4\varphi_3 & 4\varphi_3 - \varphi_2
\end{array}\ . \tag{5.15}$$

This method satisfies conditions 1–4 of Table 2, the weakened but sufficient condition 6 ($\psi_4(0) = 0$), but not conditions 5, 7, 8, and 9. However, it satisfies a weakened form of conditions 5 and 9 (because $\psi_{2,2}(0) + \psi_{2,3}(0) = 0$ and $\psi_{2,4}(0) = 0$), and a very weak form of conditions 7 and 8 (where all arguments are evaluated for $A = 0$). In the worst case, this leads to an order reduction to order two only.

Krogstad's method [12] for (1.1) is given by

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,3} - \varphi_{2,3} & \varphi_{2,3} & & \\
1 & \varphi_{1,4} - 2\varphi_{2,4} & 0 & 2\varphi_{2,4} & \\
\hline
 & \varphi_1 - 3\varphi_2 + 4\varphi_3 & 2\varphi_2 - 4\varphi_3 & 2\varphi_2 - 4\varphi_3 & -\varphi_2 + 4\varphi_3
\end{array}\,. \tag{5.16}$$

This method satisfies conditions 1–5 and 9 of Table 2, the weakened but sufficient condition 6 ($\psi_4(0) = 0$), but not conditions 7 and 8, which are only satisfied in a very weak form (where all arguments are evaluated for $A = 0$). In the worst case, this leads to an order reduction to order three.

Strehmel and Weiner's method [25, Example 4.5.5] can be written as

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,3} - \frac{1}{2}\varphi_{2,3} & \frac{1}{2}\varphi_{2,3} & & \\
1 & \varphi_{1,4} - 2\varphi_{2,4} & -2\varphi_{2,4} & 4\varphi_{2,4} & \\
\hline
 & \varphi_1 - 3\varphi_2 + 4\varphi_3 & 0 & 4\varphi_2 - 8\varphi_3 & -\varphi_2 + 4\varphi_3
\end{array}\,. \tag{5.17}$$

This method satisfies the conditions of Table 2 in exactly the same way as Krogstad's method. It thus converges in our situation with order three in the worst case. Strehmel and Weiner proved that the method is B-consistent of order two.

Remark. Under favorable circumstances, each of the above methods can show a higher order of convergence (generically up to order four). We will shortly discuss a typical situation when this happens. For instance, the method of Cox and Matthews satisfies condition 5 of Table 2 only in the very weak form

$$\sum_{i=1}^{s} b_i(0) J_n \psi_{2,i}(0) = 0.$$

According to Theorem 4.7, this may result in an order reduction down to order two. The term corresponding to condition 5 contributes to the global error via

$$h^3 \sum_{j=0}^{n-1} \mathrm{e}^{-(n-j-1)hA} \sum_{i=1}^{s} b_i(-hA) J_j \psi_{2,i}(-hA) f''(t_j)$$

$$= h^3 \sum_{j=0}^{n-1} \mathrm{e}^{-(n-j-1)hA} \sum_{i=1}^{s} \Big(b_i(-hA) - b_i(0) \Big) J_j \psi_{2,i}(0) f''(t_j) \tag{5.18a}$$

$$+\, h^3 \sum_{j=0}^{n-1} \mathrm{e}^{-(n-j-1)hA} \sum_{i=1}^{s} b_i(-hA) J_j \Big(\psi_{2,i}(-hA) - \psi_{2,i}(0) \Big) f''(t_j). \tag{5.18b}$$

The first part (5.18a) has order three by Lemma 4.8. In order to improve the second term, we assume that the operators A and J_j are such that $A^{-1}J_j\widetilde{A}^{\mu}$ is bounded on X for some $0 \le \mu \le 1$. Then, due to

$$\begin{aligned}&\big\|A^{-1}J_j\big(\psi_{2,i}(-hA)-\psi_{2,i}(0)\big)\big\|_{X\leftarrow X}\\ &\quad= \big\|A^{-1}J_j(h\widetilde{A})^{\mu}\cdot(h\widetilde{A})^{-\mu}\big(\psi_{2,i}(-hA)-\psi_{2,i}(0)\big)\big\|_{X\leftarrow X} \le Ch^{\mu},\end{aligned}$$

one gains (by applying Lemma 4.8 once more) an additional order μ in the term (5.18b). Exactly this happens in the numerical examples of section 6.

One might ask whether it is possible to modify the above methods in such a way that they have order four for semilinear parabolic problems. In fact this cannot be done without adding further stages, which is seen as follows. Assume $s = 4$, $c_2 = c_3 = 1/2$, and $c_4 = 1$. Due to the order conditions of the underlying method, we have $b_4(0) \neq 0$ and $|b_2(0)| + |b_3(0)| > 0$. Condition 5 of Table 2 immediately yields $\psi_{2,4} = 0$. Moreover, $b_3 \neq 0$ since $\psi_{2,2} \neq 0$. Hence, condition 5 can be satisfied only if $b_2 = \gamma b_3$. This leads to $\gamma\psi_{2,2} + \psi_{2,3} = 0$, which gives $a_{32} = (1+\gamma)\frac{1}{2}\varphi_{2,3}$. This choice of a_{32} contradicts condition 8 even in the weakened form, where $b_i(-hA)$ is replaced by $b_i(0)$.

Thus we consider the case $s = 5$ and add the node $c_5 = 1/2$. In order to avoid the difficulty encountered above, we have to choose $b_2 = b_3 = 0$. This requires $b_4b_5 \neq 0$, and therefore we have to enforce $\psi_{2,4} = \psi_{2,5} = 0$ by condition 5 for a method of order three. Thus condition 9 is satisfied automatically. Condition 8 shows that $a_{42} = \gamma a_{43}$ and $a_{52} = \gamma a_{53}$. For simplicity, we choose $\gamma = 1$. This gives $\psi_{2,2} + \psi_{2,3} = 0$, which leads to $a_{32} = \varphi_{2,3}$. Unfortunately, condition 7 cannot be satisfied in a strong form, because $\psi_{3,4} = \psi_{3,5} = 0$ contradicts $\psi_{2,4} = \psi_{2,5} = 0$. Hence we require only the weak form with $b_i(0)$ instead of $b_i(-hA)$. This yields the following fourth-order scheme:

$$\begin{array}{c|ccccc}
0 & & & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,2} & & & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,3}-\varphi_{2,3} & \varphi_{2,3} & & & \\
1 & \varphi_{1,4}-2\varphi_{2,4} & \varphi_{2,4} & \varphi_{2,4} & & \\
\frac{1}{2} & \frac{1}{2}\varphi_{1,5}-2a_{5,2}-a_{5,4} & a_{5,2} & a_{5,2} & \frac{1}{4}\varphi_{2,5}-a_{5,2} & \\
\hline
 & \varphi_1-3\varphi_2+4\varphi_3 & 0 & 0 & -\varphi_2+4\varphi_3 & 4\varphi_2-8\varphi_3
\end{array} \tag{5.19}$$

with

$$a_{5,2} = \frac{1}{2}\varphi_{2,5} - \varphi_{3,4} + \frac{1}{4}\varphi_{2,4} - \frac{1}{2}\varphi_{3,5}.$$

6. Numerical experiments.

In this section we present some numerical experiments in order to verify the sharpness of our error bounds.

As a first example we consider the semilinear parabolic problem

$$\frac{\partial U}{\partial t}(x,t) - \frac{\partial^2 U}{\partial x^2}(x,t) = \frac{1}{1+U(x,t)^2} + \Phi(x,t) \tag{6.1}$$

for $x \in [0,1]$ and $t \in [0,1]$, subject to homogeneous Dirichlet boundary conditions. The source function Φ is chosen in such a way that the exact solution of the problem is $U(x,t) = x(1-x)\,\mathrm{e}^t$.

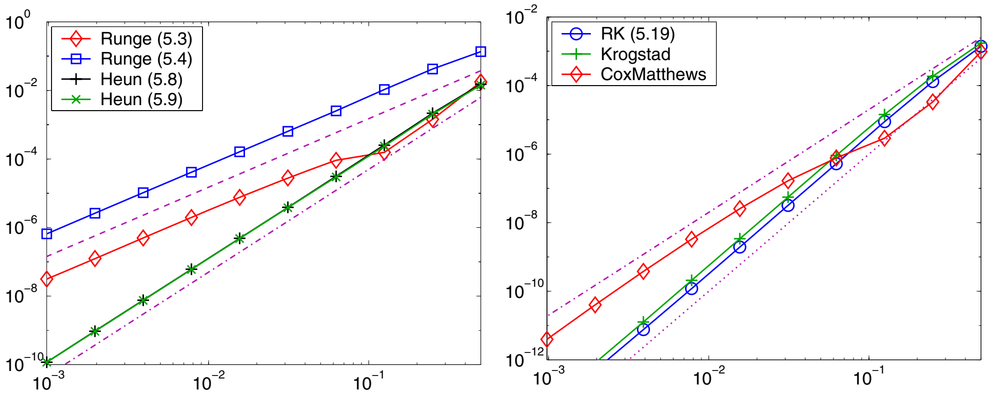


FIG. 6.1. *The errors of various explicit exponential Runge–Kutta methods of orders two to four when applied to* (6.1)*. The errors are measured in the maximum norm at* $t = 1$ *and plotted as functions of the step size. For comparison, we added lines with slope two* (*dashed*)*, three* (*dash-dotted*)*, and four* (*dotted*)*.*

We discretize this problem in space by standard finite differences with 200 grid points. Due to our theory, we expect to see order two for the two variants (5.3) and (5.4) with $c_2 = \frac{1}{2}$ of the exponential Runge method, and order three for the two variants (5.8) and (5.9) of the exponential Heun method with $c_2 = 1/3$ and $\gamma = 1.52$. Note that for this example $\|A^{-1}JA\|$ is bounded. This gives us order four for Krogstad's method and order three for the Cox and Matthews method. All these orders are confirmed by the results illustrated in Figure 6.1, where the errors are measured in the maximum norm.

Next we consider the semilinear parabolic problem

$$\frac{\partial U}{\partial t}(x,t) - \frac{\partial^2 U}{\partial x^2}(x,t) = \int_0^1 U(x,t)\,\mathrm{d}x + \Phi(x,t) \tag{6.2}$$

for $x \in [0,1]$ and $t \in [0,1]$, subject to homogeneous Dirichlet boundary conditions. The source function Φ is again chosen such that $U(x,t) = x(1-x)\,\mathrm{e}^t$ is the exact solution.

We discretize this problem in space as in the first example, with the trapezoidal rule for the approximation of the integral. The numerical results are displayed in Figure 6.2. Again we can see order two for the two variants of the exponential Runge method and order three for the two variants of the exponential Heun method. However, since in this example only $\|A^{-1}JA^{1/2}\|$ is bounded, Krogstad's method and the Cox and Matthews method suffer from an order reduction. For Krogstad's method, we thus obtain order 3.5, while for the Cox–Matthews method we have order 2.5 only. The new exponential variant of the classical Runge–Kutta method is of full order four in this example.

In Figure 6.3 we present the errors of the 2- and 3-stage methods in the V-norm for the choice $X = L^2$ and $\alpha = 1/2$. It can be seen that method (5.3) is of order two, while method (5.4) only is of order 1.75. The order reduction for (5.4) is perfectly explained by Theorem 4.3 with $\beta = 3/4 - \varepsilon$ for arbitrary $\varepsilon > 0$, since the derivatives of f are smooth functions that do not satisfy the boundary conditions. Method (5.3), however, has full order two, since condition (4.21) holds with $\gamma = 1$ and $\mu = 1/2$. See also our detailed discussion on this topic at the end of section 4.3. By similar

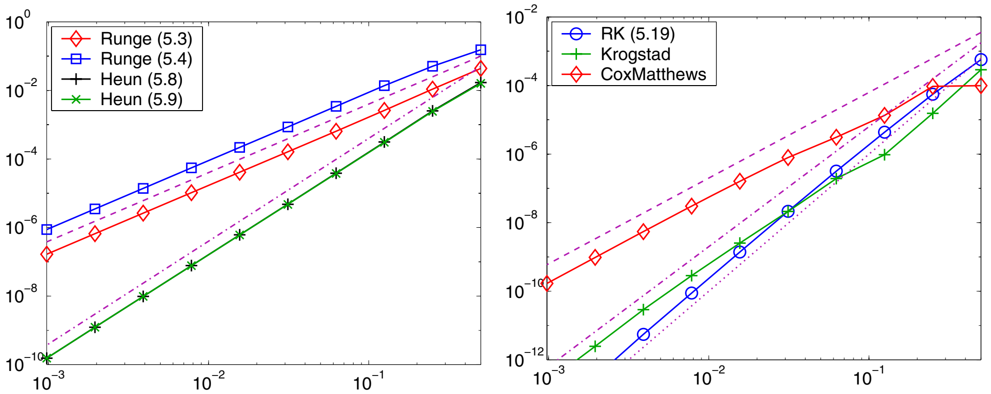


FIG. 6.2. *The errors of various explicit exponential Runge–Kutta methods of orders two to four when applied to* (6.2). *The errors are measured in the maximum norm at* $t = 1$ *and plotted as functions of the step size. For comparison, we added lines with different slopes. In the left panel, slope two is represented by a dashed line and slope three by a dash-dotted line. In the right panel, slope* 2.5 *is represented by a dashed line, slope* 3.5 *by a dash-dotted line, and slope* 4 *by a dotted line.*

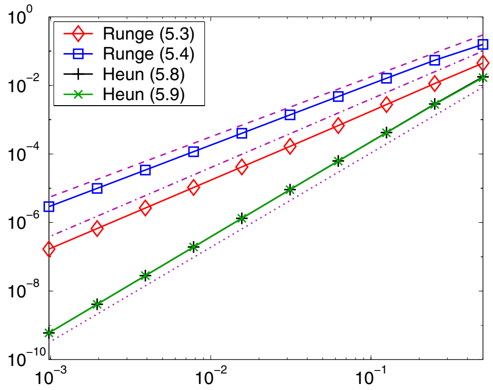


FIG. 6.3. *The errors of various explicit exponential Runge–Kutta methods up to order three when applied to* (6.2). *The errors are measured at* $t = 1$ *in a discrete* V*-norm for* $X = L^2$ *and* $\alpha = 1/2$. *They are plotted as a function of the step size. For comparison, we added lines with different slopes: slope* 1.75 *is represented by a dashed line, slope* 2 *by a dash-dotted line, and slope* 2.75 *by a dotted line.*

considerations it can be explained why the 3-stage methods show order 2.75 instead of order three in this norm.

It is beyond the scope of this paper to discuss implementation details and to compare the efficiency of exponential methods with standard implicit schemes. Such comparisons have been presented in [3, 11, 12]. Our aim here was to understand the convergence behavior of explicit methods and to present new order conditions which allow us to construct methods up to order four in a systematic way.

Acknowledgments. The authors thank Ch. Lubich and J. Schweitzer for their continuous interest in our work and for inspiring discussions.

REFERENCES

[1] H. Berland, B. Owren, and B. Skaflestad, *B-series and order conditions for exponential integrators*, SIAM J. Numer. Anal., to appear.

[2] E. Celledoni, A. Marthinsen, and B. Owren, *Commutator-free Lie group methods*, Future Generation Computer Systems, 19 (2003), pp. 341–352.

[3] S. Cox and P. Matthews, *Exponential time differencing for stiff systems*, J. Comput. Phys., 176 (2002), pp. 430–455.

[4] B. Ehle and J. Lawson, *Generalized Runge–Kutta processes for stiff initial value problems*, J. Inst. Math. Appl., 16 (1975), pp. 11–21.

[5] J. V. D. Eshof and M. Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential*, SIAM J. Sci. Comput., to appear.

[6] A. Friedli, *Verallgemeinerte Runge–Kutta Verfahren zur Lösung steifer Differentialgleichungssysteme*, in Numerical Treatment of Differential Equations, Lecture Notes in Math. 631, R. Burlirsch, R. Grigorieff, and J. Schröder, eds., Springer, Berlin, 1978, pp. 35–50.

[7] E. Hairer, S. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I—*Nonstiff Problems*, Springer Series Comput. Math. 8, 2nd ed., Springer-Verlag, Berlin, Heidelberg, 1991.

[8] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Lecture Notes in Math. 840, Springer, Berlin, Heidelberg, 1981.

[9] M. Hochbruck, C. Lubich, and H. Selhofer, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.

[10] M. Hochbruck and A. Ostermann, *Exponential Runge–Kutta methods for parabolic problems*, Appl. Numer. Math., 53 (2005), pp. 323–339.

[11] A.-K. Kassam and L. N. Trefethen, *Fourth-order time stepping for stiff PDEs*, SIAM J. Sci. Comput., 26 (2005), pp. 1214–1233.

[12] S. Krogstad, *Generalized integrating factor methods for stiff PDEs*, J. Comput. Phys., 203 (2005), pp. 72–88.

[13] J. D. Lambert and S. T. Sigurdsson, *Multistep methods with variable matrix coefficients*, SIAM J. Numer. Anal., 9 (1972), pp. 715–733.

[14] J. D. Lawson, *Generalized Runge–Kutta processes for stable systems with large Lipschitz constants*, SIAM J. Numer. Anal., 4 (1967), pp. 372–380.

[15] C. Lubich and A. Ostermann, *Linearly implicit time discretization of non-linear parabolic equations*, IMA J. Numer. Anal., 15 (1995), pp. 555–583.

[16] C. Lubich and A. Ostermann, *Runge–Kutta approximation of quasi-linear parabolic equations*, Math. Comp., 64 (1995), pp. 601–627.

[17] A. Lunardi, *Analytic Semigroups and Optimal Regularity in Parabolic Problems*, Birkhäuser, Basel, Switzerland, 1995.

[18] B. Minchev, *Lie Group Integrators with Nonautonomous Frozen Vector Fields*, Tech. Report 276, Department of Informatics, University of Bergen, Bergen, Norway, 2004.

[19] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3–49.

[20] I. Moret and P. Novati, *RD Rational approximations of the matrix exponential*, BIT, 44 (2004), pp. 595–615.

[21] S. Nørsett, *An A-stable modification of the Adams–Bashforth methods*, in Conference on the Numerical Solution of Differential Equations, J. Morris, ed., Lecture Notes in Math. 109, Springer, Berlin, 1969, pp. 214–219.

[22] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[23] D. Pope, *An exponential method of numerical integration of ordinary differential equations*, Comm. ACM, 6 (1963), pp. 491–493.

[24] K. Strehmel and R. Weiner, *B-convergence results for linearly implicit one step methods*, BIT, 27 (1987), pp. 264–281.

[25] K. Strehmel and R. Weiner, *Linear-implizite Runge–Kutta Methoden und ihre Anwendungen*, Teubner, Leipzig, 1992.

[26] J. Verwer, *On generalized linear multistep methods with zero-parasitic roots and an adaptive principal root*, Numer. Math., 27 (1977), pp. 143–155.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1091–1111

THE POSTPROCESSED MIXED FINITE-ELEMENT METHOD FOR THE NAVIER–STOKES EQUATIONS[*]

BLANCA AYUSO[†], BOSCO GARCÍA-ARCHILLA[‡], AND JULIA NOVO[†]

Abstract. A postprocessing technique for mixed finite-element methods for the incompressible Navier–Stokes equations is studied. The technique was earlier developed for spectral and standard finite-element methods for dissipative partial differential equations. The postprocessing amounts to solving a Stokes problem on a finer grid (or higher-order space) once the time integration on the coarser mesh is completed. The analysis presented here shows that this technique increases the convergence rate of both the velocity and the pressure approximations. Numerical experiments are presented that confirm both this increase in the convergence rate and the corresponding improvement in computational efficiency.

Key words. Navier–Stokes equations, mixed finite-element methods

AMS subject classifications. 65M60, 65M20, 65M15, 65M12

DOI. 10.1137/040602821

1. Introduction.

This paper in a sense culminates the development of a postprocessing technique to increase the accuracy and computational efficiency of Galerkin methods for dissipative partial differential equations introduced in [18]. We turn to the equations which gave rise to this postprocessing technique, the incompressible Navier–Stokes equations, and we address those Galerkin methods for these equations which, when complex-shaped bodies are present, are acknowledged to be of wider applicability, mixed finite-element (MFE) methods.

The postprocessing technique we study here was originally developed for spectral methods [18], [19]. At that moment, either its analysis and understanding or its development seemed to depend heavily on the properties of the Fourier modes, although this was not a shortcoming to prove its usefulness in the study of nonlinear shell vibrations [27]. In later works [13], [14], the dependence on the Fourier modes was overcome. Of particular importance to the present work, besides [14], has been the development of the postprocessing technique for finite-element methods in [20], [15]. In [20], it was devised how to carry out the postprocessing without the help of an approximate inertial manifold [11], [12], a concept more suited to spectral methods and eigenfunction expansion. In [15], it is shown what gains can be expected when postprocessing low-order elements.

As is usually the case with MFE methods, it is the experience and understanding gained in previous works (see [14], [15], [16], [17], [18], [19], [20], and the references cited therein) with simpler equations and methods which has allowed the present one to be written. Furthermore, although for simplicity we focus on Hood–Taylor [26] elements, the postprocessing technique can be easily adapted to other kinds of mixed

[*] Received by the editors January 2, 2004; accepted for publication (in revised form) March 18, 2005; published electronically September 23, 2005.
http://www.siam.org/journals/sinum/43-3/60282.html

[†] Departamento de Matemáticas, Universidad Autónoma de Madrid, Madrid, Spain (blanca.ayuso@uam.es, julia.novo@uam.es). The research of the first author was supported by project HPRN-CT-2002-00284. The research of the third author was supported by DGI-MCYT under project MTM2004-02847 (cofinanced by FEDER funds) and by JCYL under project VA044/03.

[‡] Departamento de Matemática Aplicada II, Universidad de Sevilla, Sevilla, Spain (bosco@matina.us.es). This author's research was supported by DGICYT project BFM2003-00336.

elements. In fact, in [3] (see also [5]) the so-called mini-element is shown to render similar gains as Hood–Taylor elements when postprocessed if the provisions in [15] are taken into account.

Let us describe what this postprocessing technique is. We consider the incompressible Navier–Stokes equations, which, in appropriate dimensionless variables, can be written as

$$\begin{aligned} u_t - \nu\Delta u + (u\cdot\nabla)u + \nabla p &= f, \\ \operatorname{div}(u) &= 0 \end{aligned} \tag{1.1}$$

in a bounded domain $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ with smooth boundary subject to homogeneous Dirichlet boundary conditions $u = 0$ on $\partial\Omega$. In (1.1), u is the velocity field, p the pressure, and f a given force field. Suppose that for the solution u and p corresponding to a given initial condition

$$u(\cdot, 0) = u_0; \tag{1.2}$$

we are interested in its value at a certain time $T > 0$. We first compute MFE approximations u_h and p_h to the velocity and pressure, respectively, by integrating in time the corresponding discretization of (1.1)–(1.2) from $t = 0$ to $t = T$. Then, in the postprocessing step, we obtain an approximation to the solution $\tilde{u}$, $\tilde{p}$ of the Stokes problem

$$\begin{aligned} \left.\begin{aligned} -\nu\Delta\tilde{u} + \nabla\tilde{p} = f - \tfrac{d}{dt}u_h(T) - (u_h(T)\cdot\nabla)u_h(T) \\ \operatorname{div}(\tilde{u}) = 0 \end{aligned}\right\} \quad & \text{in } \Omega, \\ \tilde{u} = 0 \quad & \text{on } \partial\Omega. \end{aligned} \tag{1.3}$$

The MFE of this last step is either the same-order Hood–Taylor element over a finer grid or a higher-order Hood–Taylor element over the same grid. The rate of convergence of the discrete velocity and pressure in the resulting method is proved to be the same as the rate of convergence of the MFE used in the postprocessed step. The overcost of the postprocessed procedure is nearly negligible since the Stokes problem using the enhanced MFE is solved only once, when the time integration has been completed. In this respect, it radically differs from some other research [2], [32], with low-order MFEs for the Navier–Stokes equations that also developed from the ideas in [11] and [12], since in [2] and [32] computations with the enhanced element or on the finer grid are carried out all the way through the interval $(0, T]$.

Some superconvergence results are obtained in the paper and are used as a tool to get the rate of convergence of the postprocessed method. In particular, we derive a superconvergence result for the error between the MFE approximation to the velocity and the discrete Stokes projection introduced in [24]. For simplicity of analysis, we derive these results under the strong regularity hypotheses in (2.2), which, as pointed out in [24], are unrealistic in practical situations. In a more practical setting, assumptions (2.2) should be assumed from some positive time $t_0 > 0$ onwards, and, as we comment in section 2, computations (and their analysis) up to this time should take into account the lower regularity at $t = 0$.

Finally, we remark that recent research [16], [17] has shown the usefulness of the postprocessing technique in obtaining efficient a posteriori error estimators in partial differential equations of evolution, a field much less developed than in the case of steady problems. The application of the postprocessing technique to get a posteriori error estimates for Navier–Stokes equations using the results obtained in this paper will be the subject of future work.

The rest of the paper is as follows. In section 2 we recall some properties of MFE methods and collect some inequalities to be used later. In section 3 we first specify the postprocessing technique and then carry out the convergence analysis. Finally, in section 4 numerical experiments are presented to assess the capabilities of the new technique.

2. Preliminaries and notations. Let Ω be a bounded domain in $\mathbb{R}^d$, $d = 2, 3$, not necessarily convex, but of class $\mathcal{C}^m$, $m \geq 3$, and let H and V be the Hilbert spaces $H = \{u \in \big(L^2(\Omega)\big)^d, \,|\, \mathrm{div}(u) = 0,\ u \cdot n_{|\partial\Omega} = 0\}$, $V = \{u \in \big(H_0^1(\Omega)\big)^d, \,|\, \mathrm{div}(u) = 0\}$, endowed with the inner product of $L^2(\Omega)^d$ and $H_0^1(\Omega)^d$, respectively. For $1 \leq q \leq \infty$ and $l \geq 0$, we consider the standard Sobolev spaces, $W^{l,q}(\Omega)^d$, of functions with derivatives up to order l in $L^q(\Omega)$, and $H^l(\Omega)^d = W^{l,2}(\Omega)^d$. The norm in $H^l(\Omega)^d$ will be denoted by $\|\cdot\|_l$ while $\|\cdot\|_{-l}$ will represent the norm of its dual space. We consider also the quotient spaces $H^l(\Omega)/\mathbb{R}$ with norm $\|p\|_{H^l/\mathbb{R}} = \inf\{\|p + c\|_l \mid c \in \mathbb{R}\}$.

We shall frequently use the following Sobolev's imbeddings [1]. There exists a constant $C = C(\Omega, q)$ such that for $q \in [1, \infty)$, $q' < \infty$, it holds that

$$\|v\|_{L^{q'}(\Omega)^d} \leq C\|v\|_{W^{s,q}(\Omega)^d}, \qquad \frac{1}{q} \geq \frac{1}{q'} \geq \frac{1}{q} - \frac{s}{d} > 0, \quad v \in W^{s,q}(\Omega)^d. \tag{2.1}$$

For $q' = \infty$, (2.1) holds with $\frac{1}{q} < \frac{s}{d}$.

Let $\Pi : L^2(\Omega)^d \longrightarrow H$ be the Leray projector that maps each function in $L^2(\Omega)^d$ onto its divergence-free part. We denote by $\mathcal{A}$ the Stokes operator in Ω:

$$\mathcal{A} : \mathcal{D}(\mathcal{A}) \subset H \longrightarrow H, \quad \mathcal{A} = -\Pi\Delta, \quad \mathcal{D}(\mathcal{A}) = H^2(\Omega)^d \cap V.$$

Applying Leray's projector to (1.1), the equations can be written in the form

$$u_t + \nu\mathcal{A}u + B(u, u) = \Pi f \qquad \text{in } \Omega,$$

where $B(u, u) = \Pi((u \cdot \nabla)u)$.

In what follows we will assume that the solution (u, p) of (1.1)–(1.2) satisfies

$$\max_{0\leq t\leq T} \big(\|u(t)\|_r + \|p(t)\|_{H^{r-1}/\mathbb{R}}\big) < \infty, \quad \max_{0\leq t\leq T} \big(\|u_t(t)\|_r + \|p_t(t)\|_{H^{r-1}/\mathbb{R}}\big) < \infty. \tag{2.2}$$

We refer the reader to [30] for a study about the regularity of the solutions of the Navier–Stokes equations. Notice, however, that, as pointed out in [24], it is unrealistic to assume such a strong regularity up to time $t = 0$. The assumption in (2.2) is for simplicity in the analysis. In a more realistic setting, $t = 0$ should be replaced by some positive time t_0, and error bounds requiring less regularity such as those in [24] and [25] should be considered from $t = 0$ to $t = t_0$. In order to maintain the accuracy levels that a higher regularity would allow from t_0 onwards, computations up to $t = t_0$ should be carried out on an adequate finer grid. Notice also that among the conditions to ensure (2.2) (see, e.g., Theorem 4 in [23]) is that Ω is of class $\mathcal{C}^r$.

Let $\mathcal{T}_h = (\tau_i^h, \phi_i^h)_{i\in I_h}$, $h > 0$, be a family of partitions of suitable domains Ω_h, where the parameter h is the maximum diameter of the elements $\tau_i^h \in \mathcal{T}_h$ and ϕ_i^h are the mappings of the reference simplex τ_0 onto τ_i^h. We restrict ourselves to quasi-uniform and regular meshes $\mathcal{T}_h$.

Let $r \geq 2$, we consider the finite-element spaces

$$\widehat{S}_{h,r} = \big\{\chi_h \in \mathcal{C}^0(\overline{\Omega_h}) \,|\, \chi_{h|_{\tau_i^h}} \circ \phi_i^h \in P^{r-1}(\tau_0)\big\} \subset H^1(\Omega_h),$$

$$\overset{\circ}{S}_{h,r} = \big\{\chi_h \in \mathcal{C}^0(\overline{\Omega_h}) \,|\, \chi_{h|_{\tau_i^h}} \circ \phi_i^h \in P^{r-1}(\tau_0),\ \chi_h(x) = 0\ \forall x \in \partial\Omega_h\big\} \subset H_0^1(\Omega_h),$$

where $P^{r-1}(\tau_0)$ denotes the space of polynomials of degree at most $r-1$ on τ_0. As a consequence of restricting our study to quasi-uniform partitions, the following inverse inequality holds (see, e.g., [9, Theorem 3.2.6]) $\forall \tau = \tau_i^h \in \mathcal{T}_h$, with $\operatorname{diam}(\tau) = h_\tau \le h$, $v_h \in (\overset{\circ}{S}_{h,r})^d$:

(2.3)
$$\|v_h\|_{W^{m,q}(\tau)^d} \le Ch^{l-m-d(\frac{1}{q'}-\frac{1}{q})}\|v_h\|_{W^{l,q'}(\tau)^d}, \quad 0 \le l \le m \le 2, \quad 1 \le q' \le q \le \infty.$$

In order to guarantee convergence of the MFE approximation, we choose a stable combination of two finite-element spaces (see [7]). We introduce the finite-element spaces in which our MFE approximation to (u, p) will be carried out. We shall denote by $(X_{h,r}, Q_{h,r-1})$ the so-called Hood–Taylor element, where

$$X_{h,r} = \left(\overset{\circ}{S}_{h,r}\right)^d, \quad Q_{h,r-1} = \widehat{S}_{h,r-1} \cap L^2(\Omega_h)/\mathbb{R}, \qquad r \ge 3.$$

For this mixed element a uniform inf-sup condition is satisfied (see [26], [6]), that is, there exists a constant $\beta > 0$ independent of the mesh grid size h such that

$$\inf_{q_h \in Q_{h,r-1}} \sup_{v_h \in X_{h,r}} \frac{(q_h, \nabla \cdot v_h)}{\|v_h\|_1 \|q_h\|_{L^2/\mathbb{R}}} \ge \beta. \tag{2.4}$$

The approximate velocity solution belongs to the discretely divergence-free space

$$V_{h,r} = X_{h,r} \cap \left\{ \chi_h \in H_0^1(\Omega_h) \ : \ \int_{\Omega_h} q_h \operatorname{div}(\chi_h) = 0 \ \ \forall q_h \in Q_{h,r-1} \right\}.$$

We observe that for the Hood–Taylor element, $V_{h,r}$ is not a subspace of V.

For any $v \in C_0(\Omega)^d$, we consider the standard interpolant operator $I_h : C_0(\Omega)^d \longrightarrow X_{h,r}$. Let $v \in H^r(\Omega)^d \cap H_0^1(\Omega)^d$; it is well known that I_h satisfies

$$\|v - I_h(v)\|_{L^2(\Omega\cap\Omega_h)^d} + h\|v - I_h(v)\|_{H^1(\Omega\cap\Omega_h)^d} \le Ch^r\|v\|_{H^r(\Omega)^d}. \tag{2.5}$$

We briefly discuss next under what circumstances (2.5) can be extended to a global estimate (i.e., to an estimate in Ω and not just in $\Omega \cap \Omega_h$). The interpolation operator $I_h(v)$ is extended by zero in $\Omega \setminus \Omega_h$, and defining $\delta(h) = \max_{x\in\partial\Omega_h} \operatorname{dist}(x, \partial\Omega)$, one obtains

$$\|v - I_h(v)\|_{L^2(\Omega)^d} + h\|v - I_h(v)\|_{H^1(\Omega\cap\Omega_h)^d} \le C(h^r + \delta(h))\|v\|_{H^r(\Omega)^d}. \tag{2.6}$$

For $x \in \Omega \cap \Omega_h$, (2.5) (and so (2.6)) follows from standard theory of interpolation and the Bramble–Hilbert lemma (see, e.g., [9, p. 192]). For $x \in \Omega \setminus \Omega_h$, $v(x)$ can be bounded by means of the mean-value theorem,

$$\|v - I_h(v)\|_{L^2(\Omega\setminus\Omega_h)^d} = \|v\|_{L^2(\Omega\setminus\Omega_h)^d} \le \delta(h)\|\nabla v\|_{L^2(\Omega)^d}.$$

We observe that using isoparametric elements $\delta(h) \le Ch^r$, and so in (2.6) the right-hand side is further bounded by $Ch^r\|v\|_{W^{r,q}(\Omega)^d}$ (see [9, section 4.4]). As regards the global estimate for the gradient, isoparametric modification is not enough to preserve the optimal approximability properties of the finite-element space. Following [3], we shall assume in what follows the use of superparametric elements at the boundary. By this type of approximation we mean that $\delta(h) \le Ch^{2r-2}$ so that the outside effects will not pollute the optimal estimate. Under these assumptions [3], [4], the interpolant I_h satisfies

$$\|v - I_h(v)\|_{L^2(\Omega)^d} + h\|v - I_h(v)\|_{H^1(\Omega)^d} \leq Ch^r \|v\|_{H^r(\Omega)^d}. \tag{2.7}$$

Notice then that the condition $\delta(h) \leq Ch^{2r-2}$ allows us to forget about the discrepancies between Ω and Ω_h in most of the arguments that follow. Observe, however, that one must then assume that Ω is piecewise of class $\mathcal{C}^{2r-2}$.

For each fixed time $t \in [0, T]$ the solution (u, p) of (1.1)–(1.2) is also the solution of a Stokes problem with right-hand side $f - u_t - (u \cdot \nabla)u$. We will denote by $(s_h, q_h) \in (X_{h,r}, Q_{h,r-1})$, its MFE approximation satisfying

$$\begin{aligned} \nu(\nabla s_h, \nabla \phi_h) - (q_h, \nabla \cdot \phi_h) &= \nu(\nabla u, \nabla \phi_h) - (p, \nabla \cdot \phi_h) \\ &= (f - u_t - (u \cdot \nabla u), \phi_h) \quad \forall \phi_h \in X_{h,r}, \\ (\nabla \cdot s_h, \psi_h) &= 0 \quad \forall \psi_h \in Q_{h,r-1}. \end{aligned} \tag{2.8}$$

We observe that $s_h = S_h(u) : V \longrightarrow V_{h,r}$ is the so-called discrete Stokes projection of the solution (u, p) of (1.1)–(1.2) (see [24]) and satisfies

$$(\nabla S_h(u), \nabla \chi_h) = (\nabla u, \nabla \chi_h) - (p, \nabla \cdot \chi_h) = (f - u_t - (u \cdot \nabla)u, \chi_h) \quad \forall\ \chi_h \in V_{h,r}.$$

The following bound holds for $2 \leq l \leq r$:

$$\|u - s_h\|_0 + h\|u - s_h\|_1 \leq Ch^l \left(\|u\|_l + \|p\|_{H^{l-1}/\mathbb{R}}\right). \tag{2.9}$$

The proof of (2.9) for $\Omega = \Omega_h$ can be found in [25]. For the general case superparametric approximation at the boundary is assumed; see [3], [4]. Under the same conditions, the bound for the pressure is [21]

$$\|p - q_h\|_{L^2/\mathbb{R}} \leq C_\beta h^{l-1} \left(\|u\|_l + \|p\|_{H^{l-1}/\mathbb{R}}\right), \tag{2.10}$$

where the constant C_β depends on the constant β in the inf-sup condition (2.4).

Since we are assuming that Ω is of class $\mathcal{C}^m$ with $m \geq 3$ (and that $\delta(h) \leq Ch^{2r-2}$) using standard duality arguments and (2.9), one obtains [3], [4]

$$\|u - s_h\|_{-s} \leq Ch^{r+s}(\|u\|_r + \|p\|_{H^{r-1}/\mathbb{R}}), \qquad 0 \leq s \leq \min(r-2, 1). \tag{2.11}$$

Let $\Pi_{h,r} : L^2(\Omega)^d \longrightarrow V_{h,r}$ be the discrete Leray's projection defined by demanding that $(\Pi_{h,r}(u), \chi_h) = (u, \chi_h)\ \forall \chi_h \in V_{h,r}$. By definition, the projection is stable in the L^2 norm. For divergence-free functions, by writing $\Pi_{h,r}u = (\Pi_{h,r}u - S_h(u)) + S_h(u)$ and using the quasi-uniformity of the meshes, one easily shows that

$$\|\Pi_{h,r}u\|_1 \leq C\|u\|_1 \qquad \forall\, u \in V. \tag{2.12}$$

We will denote by $\mathcal{A}_h$ the discrete Stokes operator defined by

$$(\nabla v_h, \nabla \phi_h) = (\mathcal{A}_h v_h, \phi_h) = \left(\mathcal{A}_h^{1/2} v_h, \mathcal{A}_h^{1/2} \phi_h\right) \qquad \forall v_h, \phi_h \in V_{h,r}.$$

Since $\mathcal{A}_h$ is a discrete self-adjoint operator, it is easy to show that, for each $0 \leq \alpha < 1$, there exists a positive constant C_α, which is independent of h, such that

$$\|\mathcal{A}_h^\alpha e^{-t\mathcal{A}_h}\|_0 \leq C_\alpha t^{-\alpha} \qquad \forall\ 0 \leq \alpha < 1. \tag{2.13}$$

In our analysis we shall frequently use the following relations for $f \in L^2(\Omega)^d$:

$$\|\mathcal{A}_h^{-s/2}\Pi_{h,r}f\|_0 \le Ch^s\|f\|_0 + \|\mathcal{A}^{-s/2}\Pi f\|_0, \quad s = 1,2, \tag{2.14}$$

$$\|\mathcal{A}^{-s/2}\Pi f\|_0 \le Ch^s\|f\|_0 + \|\mathcal{A}_h^{-s/2}\Pi_{h,r}f\|_0, \quad s = 1,2. \tag{2.15}$$

These inequalities are readily deduced from the estimates $\|\mathcal{A}^{-s/2} - \mathcal{A}_h^{-s/2}\Pi_{h,r}\|_0 \le Ch^s$ for $s = 1,2$ [29]. Similarly, since $\forall v_h \in V_{h,r}$, $(\mathcal{A}_h^{-1/2}\Pi_{h,r}f, v_h) = (f, \mathcal{A}_h^{-1/2}v_h)$, it follows that

$$\|\mathcal{A}_h^{-1/2}\Pi_{h,r}f\|_0 \le C\|f\|_{-1}, \tag{2.16}$$

and since $\forall v \in V$, we have $(\mathcal{A}^{-1/2}\Pi(\Pi_{h,r}f), v) = (\Pi_{h,r}f, \mathcal{A}^{-1/2}v) = (f, \Pi_{h,r}\mathcal{A}^{-1/2}v)$, from (2.12) it follows that

$$\|\mathcal{A}^{-1/2}\Pi(\Pi_{h,r}f)\|_0 \le C\|f\|_{-1}, \qquad f \in L^2(\Omega)^2. \tag{2.17}$$

2.1. The suggested method. Let us suppose that we want to approximate the solution of (1.1)–(1.2) at time T. For $d = 3$, the final time T is assumed to satisfy $0 < T < T^*$, where T^* is the critical time until which the existence and uniqueness of a strong solution of (1.1)–(1.2) has been proven. The postprocessing technique can be seen as a two-level method. We first compute the MFE approximation to (1.1)–(1.2) at time T. Given $u^h(0)$ an initial approximation to $u(0)$, we find that $u_h : [0,T] \longrightarrow X_{h,r}$ and $p_h : [0,T] \longrightarrow Q_{h,r-1}$ satisfy

$$(\dot{u}_h, \phi_h) + \nu(\nabla u_h, \nabla\phi_h) + b_h(u_h, u_h, \phi_h) + (\nabla p_h, \phi_h) = (f, \phi_h) \quad \forall\, \phi_h \in X_{h,r}, \tag{2.18}$$

$$(\nabla \cdot u_h, \psi_h) = 0 \quad \forall\, \psi_h \in Q_{h,r-1}, \tag{2.19}$$

where $b_h(\cdot,\cdot,\cdot)$ is a suitable discrete approximation to its continuous counterpart. As an initial condition we will take $u_h(0) = S_h(u_0)$, although other choices are possible.

In the second step, the discrete velocity and pressure $(u_h(T), p_h(T))$ are postprocessed. Basically, we enhance this approximation by solving a single discrete Stokes problem, via MFE. The MFE in this step, denoted by $(\widetilde{X}, \widetilde{Q})$, is either

- the same-order Hood–Taylor element over a finer grid $(\widetilde{X}, \widetilde{Q}) = (X_{\tilde{h},r}, Q_{\tilde{h},r-1})$, $r \ge 3$, $\tilde{h} < h$, or
- a higher-order Hood–Taylor element over the same grid $(\widetilde{X}, \widetilde{Q}) = (X_{\tilde{h},r+1}, Q_{\tilde{h},r})$, $r \ge 3$, $\tilde{h} = h$.

That is, we shall search for $(\tilde{u}_h, \tilde{p}_h) \in (\widetilde{X}, \widetilde{Q})$ satisfying

$$\nu\big(\nabla\tilde{u}_{\tilde{h}}, \nabla\tilde{\phi}\big) + \big(\nabla\tilde{p}_{\tilde{h}}, \tilde{\phi}\big) = (f, \tilde{\phi}) - b_{\tilde{h}}(u_h(T), u_h(T), \tilde{\phi}) - (\dot{u}_h(T), \tilde{\phi}) \quad \forall\, \tilde{\phi} \in \widetilde{X}, \tag{2.20}$$

$$\big(\nabla \cdot \tilde{u}_{\tilde{h}}, \tilde{\psi}\big) = 0 \quad \forall\, \tilde{\psi} \in \widetilde{Q}. \tag{2.21}$$

We will denote by $\widetilde{V}$ the corresponding discretely divergence-free space that can be either $\widetilde{V} = V_{\tilde{h},r}$ or $\widetilde{V} = V_{h,r+1}$ depending on the selection of the postprocessed space. The discrete Leray's projection into $\widetilde{V}$ will be denoted by $\widetilde{\Pi}_{\tilde{h}}$, and we will represent by $\widetilde{\mathcal{A}}_{\tilde{h}}$ the discrete Stokes operator acting on functions in $\widetilde{V}$.

The postprocessed Hood–Taylor approximation to the velocity, $\tilde{u}_{\tilde{h}}$, is the solution of the pressure-free formulation

$$\nu\big(\nabla\tilde{u}_{\tilde{h}}, \nabla\tilde{\chi}_h\big) = \big(f, \tilde{\chi}_h\big) - b_{\tilde{h}}\big(u_h(T), u_h(T), \tilde{\chi}_h\big) - \big(\dot{u}_h(T), \tilde{\chi}_h\big) \quad \forall\, \tilde{\chi}_h \in \widetilde{V}. \tag{2.22}$$

In the next section, we show that the solution $(\tilde{u}_h, \tilde{p}_h)$ of (2.20)–(2.21) is a more accurate approximation to the solution of (1.1)–(1.2) than the Galerkin MFE approximation (u_h, p_h) that solves (2.18)–(2.19).

For the discrete approximation to the nonlinear term, following [24], we define b_h in the following way:

$$b_h(u_h, v_h, \phi_h) = ((u_h \cdot \nabla)v_h, \phi_h) + \frac{1}{2}(\operatorname{div}(u_h)v_h, \phi_h) \quad \forall\, u_h, v_h, \phi_h \in X_{h,r} \subset H_0^1(\Omega)^d.$$

For all $u, v \in H_0^1(\Omega)^d$, the corresponding continuous operator will be denoted by $F(u, v) = (u \cdot \nabla)v + (1/2)\operatorname{div}(u)v$. Extending the definition of b_h to functions in $H_0^1(\Omega)^d$ (not necessarily in $X_{h,r}$), we observe that $\forall\, u, v, w \in H_0^1(\Omega)^d$, $b_h(u, v, w) = (F(u, v), w)$. It is straightforward to verify that b_h enjoys the skew-symmetry property

$$b_h(u, v, w) = -b_h(u, w, v) \qquad \forall\, u, v, w \in H_0^1(\Omega)^d. \tag{2.23}$$

Let us observe that $B(u, v) = \Pi F(u, v)$ if $u \in V$. Finally, we shall denote by

$$B_h(u, v) = \Pi_{h,r} F(u, v) \qquad \forall\, u,\, v\, \in\, H_0^1(\Omega)^d.$$

3. Analysis of the postprocessed method.

This section is devoted to the analysis of convergence of the postprocessed MFE method. Our first aim will be to show a superconvergence result for the error between the MFE approximation to the velocity u_h and the Stokes projection of the velocity field u, s_h. This superconvergence behavior occurs for both the L^2 and H^1 norms, as will be shown in Theorem 3.7 and Corollary 3.8, respectively. In the first part of the section, we shall concentrate our efforts in Theorem 3.7. It will be achieved by a stability plus consistency argument (Propositions 3.2 and 3.6, respectively). For the purpose of analysis, we shall mainly be concerned with the pressure-free formulation associated with (2.18)–(2.19). If (u_h, p_h) is the MFE approximation to the solution (u, p) of (1.1)–(1.2), then $u_h \in V_{h,r}$ is the solution of

$$(\dot{u}_h, \chi_h) + \nu(\nabla u_h, \nabla \chi_h) + b_h(u_h, u_h, \chi_h) = (f, \chi_h) \qquad \forall \chi_h \in V_{h,r}, \tag{3.1}$$

which can also be expressed in abstract operator form as

$$\dot{u}_h + \nu \mathcal{A}_h u_h + B_h(u_h, u_h) = \Pi_{h,r} f. \tag{3.2}$$

The Stokes projection s_h satisfies the abstract equation

$$\dot{s}_h + \nu \mathcal{A}_h s_h + B_h(s_h, s_h) = \Pi_{h,r} f + T_h, \tag{3.3}$$

where $T_h(t)$ is the truncation error, defined as

$$T_h(t) = \dot{s_h} - \Pi_{h,r}(u_t) + B_h(s_h, s_h) - B_h(u, u). \tag{3.4}$$

Let us now consider mappings $v_h : [0, T] \longrightarrow V_{h,r}$ satisfying the following threshold condition:

$$\|s_h(t) - v_h(t)\|_0 \le c_\tau h^2 \qquad \forall\, t \in [0, t_1], \quad 0 < t_1 \le T. \tag{3.5}$$

We define their truncation error as

$$\widehat{T}_h = \dot{v}_h + \nu \mathcal{A}_h v_h + B_h(v_h, v_h) - \Pi_{h,r} f. \tag{3.6}$$

Prior to establishing the stability restricted to the threshold (3.5) (Proposition 3.2), we prove a lemma which provides some estimates for the convective term.

LEMMA 3.1. *Let (u,p) be the solution of the Navier–Stokes problem* (1.1)–(1.2). *Let $s_h = S_h(u)$ be the discrete Stokes projection of the velocity field u and let $v_h : [0,T] \longrightarrow V_{h,r}$ satisfy the threshold condition* (3.5). *Then, there exists a constant $K>0$, independent of t_1 in* (3.5), *such that $\forall\, t \in [0,t_1]$,*

$$\|F(s_h(t),s_h(t)) - F(v_h(t),v_h(t))\|_0 \le K\|s_h(t)-v_h(t)\|_1, \tag{3.7}$$
$$\|F(s_h(t),s_h(t)) - F(v_h(t),v_h(t))]\|_{-1} \le K\|s_h(t)-v_h(t)\|_0, \tag{3.8}$$

where the constant $K = K\big(c_\tau, \max_{0\le t\le T}(\|u(t)\|_2 + \|p(t)\|_{H^1/\mathbb{R}})\big)$.

Proof. In order to simplify the notation, we shall omit the dependence on t in the proof. Denote by $e_h = v_h - s_h$. We proceed by standard duality arguments, using the splitting

$$F(v_h,v_h) - F(s_h,s_h) = F(v_h,e_h) + F(e_h,s_h). \tag{3.9}$$

We start by showing (3.7). We first observe that

$$\begin{aligned}\|F(e_h,s_h)\|_0 &= \sup_{\|\phi\|_0=1}\left|(e_h\cdot\nabla s_h,\phi) + \frac{1}{2}((\nabla\cdot e_h)s_h,\phi)\right|\\ &\le C\|e_h\|_{L^{2d/(d-1)}(\Omega)^d}\|\nabla s_h\|_{L^{2d}(\Omega)^d} + C\|e_h\|_1\|s_h\|_\infty.\end{aligned}$$

Let us show that both $\|s_h\|_\infty$, $\|\nabla s_h\|_{L^{2d}(\Omega)^d}$ are bounded. Since, by virtue of Sobolev's imbeddings (2.1), we have $\|s_h\|_\infty \le C\|\nabla s_h\|_{L^{2d}(\Omega)^d}$, we only need to bound the second term. Application of the inverse inequality (2.3) and the error estimates (2.9) and (2.7) together with (2.1) give

$$\begin{aligned}\|\nabla s_h\|_{L^{2d}(\Omega)^d} &\le Ch^{\frac{-(1+d)}{2}}(\|s_h-u\|_0 + \|u-I_hu\|_0) + \|\nabla I_h u\|_{L^{2d}(\Omega)^d}\\ &\le Ch^{(3-d)/2}(\|u\|_2 + \|p\|_{H^1/\mathbb{R}}) + C\|u\|_{W^{1,2d}(\Omega)^d} \le K.\end{aligned} \tag{3.10}$$

Using again (2.1) we obtain

$$\|e_h\|_{L^{2d/(d-1)}(\Omega)^d} \le C\|e_h\|_{1/2} \le C\|e_h\|_1,$$

and so $\|F(e_h,s_h)\|_0 \le K\|e_h\|_1$. As regards the other term in (3.9), the same arguments lead to

$$\begin{aligned}\|F(v_h,e_h)\|_0 &= \sup_{\|\phi\|_0=1}\left|(v_h\cdot\nabla e_h,\phi) + \frac{1}{2}((\nabla\cdot v_h)e_h,\phi)\right|\\ &\le C\|v_h\|_\infty\|e_h\|_1 + C\|\nabla v_h\|_{L^{2d}(\Omega)^d}\|e_h\|_{L^{2d/(d-1)}(\Omega)^d}.\end{aligned}$$

As before, to conclude we must show that the above norms of v_h are bounded. We only need to handle $\|\nabla v_h\|_{L^{2d}(\Omega)^d}$. Using the inverse inequality (2.3) and the threshold conditions (3.5) and (3.10), we find

$$\|\nabla v_h\|_{L^{2d}(\Omega)^d} \le h^{\frac{-(1+d)}{2}}\|v_h - s_h\|_0 + \|\nabla s_h\|_{L^{2d}(\Omega)^d} \le c_\tau h^{(3-d)/2} + K \le K.$$

Therefore, (3.7) follows. We now show (3.8). Applying (3.9), we find

$$\|F(v_h,v_h) - F(s_h,s_h)\|_{-1} \le \|F(v_h,e_h)\|_{-1} + \|F(e_h,s_h)\|_{-1}, \tag{3.11}$$

so that the proof is reduced to estimate each of the above negative norms on the right-hand side. Using the skew-symmetry property (2.23), one gets for the first term:

$$\begin{aligned}\|F(v_h,e_h)\|_{-1} &= \sup_{\|\phi\|_1=1}\left|-((v_h\cdot\nabla)\phi,e_h)-\frac{1}{2}((\nabla\cdot v_h)\phi,e_h)\right|\\ &\le \sup_{\|\phi\|_1=1}\left(\|e_h\|_0\|v_h\|_\infty\|\phi\|_1+\|e_h\|_0\|\nabla\cdot v_h\|_{L^{2d/(d-1)}}\|\phi\|_{L^{2d}(\Omega)^d}\right)\le K\|e_h\|_0.\end{aligned}$$

Regarding the other term in (3.11), integrating by parts, we obtain

$$\begin{aligned}\|F(e_h,s_h)\|_{-1} &= \sup_{\|\phi\|_1=1}\left|\frac{1}{2}\big((e_h\cdot\nabla)s_h,\phi\big)-\frac{1}{2}\big((e_h\cdot\nabla)\phi,s_h\big)\right|\\ &\le \sup_{\|\phi\|_1=1}(\|e_h\|_0\|\nabla s_h\|_{L^{2d/(d-1)}(\Omega)^d}\|\phi\|_{L^{2d}(\Omega)^d}+\|e_h\|_0\|\phi\|_1\|s_h\|_\infty)\le K\|e_h\|_0.\end{aligned}$$

This finishes the proof of (3.8). □

PROPOSITION 3.2 (stability). *Let $T>0$ be fixed; let $s_h=S_h(u)$ be the discrete Stokes projection of the velocity field u solution of* (1.1)–(1.2) *and let $v_h:[0,T]\longrightarrow V_{h,r}$ satisfy the threshold condition* (3.5). *Then, there exists a positive constant $K_s>0$ such that $\forall\, t_1\le T$, the following estimate holds:*

$$\max_{0\le t\le t_1}\|s_h(t)-v_h(t)\|_0\le e^{K_st_1}\Bigg(\|s_h(0)-v_h(0)\|_0 + \max_{0\le t\le t_1}\left\|\int_0^t e^{-\nu(t-s)\mathcal{A}_h}[T_h(s)-\widehat{T}_h(s)]ds\right\|_0\Bigg), \tag{3.12}$$

where $T_h(s)$ and $\widehat{T}_h(s)$ are the truncation errors given in (3.4) *and* (3.6), *respectively.*

Proof. We denote by $e_h=s_h-v_h$. Subtracting (3.6) from (3.3), it follows that e_h satisfies the error equation

$$\dot{e}_h(t)+\nu\mathcal{A}_he_h(t)=B_h(v_h(t),v_h(t))-B_h(s_h(t),s_h(t))+T_h(t)-\widehat{T}_h(t).$$

Then, by integrating the above error equation from time 0 up to time t, we find that

$$\begin{aligned}e_h(t)&=e^{-\nu t\mathcal{A}_h}\Pi_{h,r}e_h(0)+\int_0^t e^{-\nu(t-s)\mathcal{A}_h}\Pi_{h,r}[B_h(v_h,v_h)-B_h(s_h,s_h)]ds\\ &\quad+\int_0^t e^{-\nu(t-s)\mathcal{A}_h}\Pi_{h,r}[T_h(s)-\widehat{T}_h(s)]ds.\end{aligned}$$

Since $\{e^{-\nu t\mathcal{A}_h}\Pi_{h,r}\}_{t>0}$ is a contraction $\|e^{-\nu t\mathcal{A}_h}\Pi_{h,r}e_h(0)\|_0\le\|e_h(0)\|_0$. As regards the second term, estimates (2.13), (2.16), and (3.8) from Lemma 3.1 lead to

$$\begin{aligned}&\left\|\int_0^t e^{-\nu(t-s)\mathcal{A}_h}[B_h(s_h,s_h)-B_h(v_h,v_h)]ds\right\|_0\\ &\quad\le\frac{C_{1/2}}{\sqrt{\nu}}\int_0^t\frac{\left\|\mathcal{A}_h^{-1/2}\big(\Pi_{h,r}F(s_h,s_h)-\Pi_{h,r}F(v_h,v_h)\big)\right\|_0}{\sqrt{t-s}}ds\le\frac{KC_{1/2}}{\sqrt{\nu}}\int_0^t\frac{\|e_h(s)\|_0}{\sqrt{t-s}}ds.\end{aligned}$$

Then,

$$\|e_h(t)\|_0\le\|e_h(0)\|_0+\frac{KC_{1/2}}{\sqrt{\nu}}\int_0^t\frac{\|e_h(s)\|_0}{\sqrt{t-s}}ds+\int_0^t e^{-\nu(t-s)\mathcal{A}_h}[T_h(s)-\widehat{T}_h(s)]ds.$$

And now, a standard application of the generalized Gronwall lemma (see [22, pp. 188–189]) allows us to conclude the proof. □

Proposition 3.2 is an example of stability restricted to h-dependent thresholds. This kind of stability is an alternative to establishing the a priori bounds for the approximate solution u_h required in order to handle the nonlinear term [28].

The following lemmas will be required in the proof of Proposition 3.6.

LEMMA 3.3. *For any $f \in C([0,T];L^2(\Omega)^d)$, the following estimate holds $\forall\, t \in [0,T]$:*

$$\int_0^t \left\|\mathcal{A}_h e^{-\nu(t-s)\mathcal{A}_h}\Pi_{h,r} f(s)\right\|_0 ds \le \frac{C}{\nu}|\log(h)| \max_{0\le t\le T}\|f(t)\|_0.$$

Proof. The proof follows essentially the same steps as in [14] and [20]. □

LEMMA 3.4. *Let $v \in (H^2(\Omega))^d \cap V$. Then, there exists a constant $K = K(\|v\|_2)$ such that $\forall\, w \in H_0^1(\Omega)^d$, we have that*

$$\|\mathcal{A}^{-1}\Pi[F(v,v) - F(w,w)]\|_0 \le K\big(\|v-w\|_{-1} + \|v-w\|_1\|v-w\|_0\big). \tag{3.13}$$

Proof. Throughout the proof, we shall designate $e = v - w$. We rewrite the difference of the nonlinear terms as

$$\mathcal{A}^{-1}\Pi\left(F(v,v) - F(w,w)\right) = \mathcal{A}^{-1}\Pi F(v,e) + \mathcal{A}^{-1}\Pi F(e,v) - \mathcal{A}^{-1}\Pi F(e,e). \tag{3.14}$$

Let us first estimate the last term in (3.14). Using (2.23) and (2.1), we have

$$\begin{aligned}\|\mathcal{A}^{-1}\Pi F(e,e)\|_0 &\le \sup_{\|\phi\|_0=1}\left| -\big((e\cdot\nabla)\mathcal{A}^{-1}\Pi\phi, e\big) - \frac{1}{2}\big((\nabla\cdot e)(\mathcal{A}^{-1}\Pi\phi), e\big)\right| \\ &\le \sup_{\|\phi\|_0=1}\Big(\|e\|_{L^{2d/(d-1)}(\Omega)^d}\|\nabla\mathcal{A}^{-1}\Pi\phi\|_{L^{2d}(\Omega)^d} + \|e\|_1\|\mathcal{A}^{-1}\Pi\phi\|_\infty\Big)\|e\|_0 \\ &\le \sup_{\|\phi\|_0=1}\Big(C\|e\|_{1/2}\|\mathcal{A}^{-1}\Pi\phi\|_2 + C\|e\|_1\|\mathcal{A}^{-1}\Pi\phi\|_2\Big)\|e\|_0 \le C\|e\|_1\|e\|_0.\end{aligned}$$

For the first term in the splitting (3.14), taking into account that $\mathrm{div}(v) = 0$, we find

$$\|\mathcal{A}^{-1}\Pi F(v,e)\|_0 = \sup_{\|\phi\|_0=1}\left|\big((v\cdot\nabla)\mathcal{A}^{-1}\Pi\phi, e\big)\right| \le \|e\|_{-1}\sup_{\|\phi\|_0=1}\|\nabla((v\cdot\nabla)\mathcal{A}^{-1}\Pi\phi)\|_0.$$

Therefore, we must show that the last supremum above is bounded. Using again Sobolev's imbeddings (2.1) for $\phi \in L^2(\Omega)$ with $\|\phi\|_0 = 1$, we obtain

$$\begin{aligned}\|\nabla\big((v\cdot\nabla)\mathcal{A}^{-1}\Pi\phi\big)\|_0^2 &\le \sum_{k=1}^d \|\big(\partial_k v\cdot\nabla\big)(\mathcal{A}^{-1}\Pi\phi) + (v\cdot\nabla)\big(\partial_k(\mathcal{A}^{-1}\Pi\phi)\big)\|_0^2 \\ &\le \sum_{k=1}^d \|\partial_k v\|^2_{L^{2d/(d-1)}(\Omega)^d}\|\nabla\mathcal{A}^{-1}\Pi\phi\|^2_{L^{2d}(\Omega)^d} + \|v\|_\infty^2\|\partial_k(\nabla\mathcal{A}^{-1}\Pi\phi)\|_0^2 \\ &\le C\big[\|\nabla v\|_{1/2}^2\|\mathcal{A}^{-1}\Pi\phi\|_2^2 + \|v\|_2^2\|\mathcal{A}^{-1}\Pi\phi\|_2^2\big] \le C\|v\|_2^2,\end{aligned}$$

so that $\|\mathcal{A}^{-1}\Pi F(v,e)\|_0 \le K\|e\|_{-1}$. Finally, we deal with the second term in (3.14). Integrating by parts, we get

$$\|\mathcal{A}^{-1}\Pi F(e,v)\|_0 = \sup_{\|\phi\|_0=1}\left|\frac{1}{2}((e\cdot\nabla)v, \mathcal{A}^{-1}\Pi\phi) - \frac{1}{2}((e\cdot\nabla)\mathcal{A}^{-1}\Pi\phi, v)\right|.$$

We shall estimate each supremum in the above equation separately. For the first term, we have

$$\begin{aligned}\big|(e\cdot\nabla)v,\mathcal{A}^{-1}\Pi\phi)\big| &\le \|e\|_{-1}\|\big((\mathcal{A}^{-1}\Pi\phi)\cdot\nabla\big)v\|_1\\ &\le \|e\|_{-1}\Big(\|\nabla v\|_1\|\mathcal{A}^{-1}\Pi\phi\|_\infty+\|\nabla v\|_{L^{2d/(d-1)}(\Omega)^d}\|\nabla\mathcal{A}^{-1}\Pi\phi\|_{L^{2d}(\Omega)^d}\Big)\\ &\le C\|e\|_{-1}\Big(\|v\|_2\|\mathcal{A}^{-1}\Pi\phi\|_2+\|v\|_2\|\mathcal{A}^{-1}\Pi\phi\|_2\Big)\le C\|v\|_2\|e\|_{-1}\le K\|e\|_{-1}.\end{aligned}$$

As regards the other supremum, we note that

$$\begin{aligned}\big|(e\cdot\nabla)\mathcal{A}^{-1}\Pi\phi,v)\big| &\le \sum_{k=1}^{d}\|e^k\|_{-1}\Big(\|\nabla(\partial_k\mathcal{A}^{-1}\Pi\phi)\cdot v\|_0+\|\partial_k(\mathcal{A}^{-1}\Pi\phi)\cdot(\nabla v)\|_0\Big)\\ &\le \|e\|_{-1}\Big(\|\mathcal{A}^{-1}\Pi\phi\|_2\|v\|_\infty+\|\partial_k(\mathcal{A}^{-1}\Pi\phi)\|_{L^{2d}(\Omega)^d}\|\nabla v\|_{L^{2d/(d-1)}(\Omega)^d}\Big)\\ &\le \|e\|_{-1}\Big(\|\mathcal{A}^{-1}\Pi\phi\|_2\|v\|_2+\|\mathcal{A}^{-1}\Pi\phi)\|_2\|v\|_2\Big)\le C\|v\|_2\|e\|_{-1}\le K\|e\|_{-1},\end{aligned}$$

which concludes the proof. □

LEMMA 3.5. *Let* (u,p) *be the solution of* (1.1)–(1.2). *Then, there exists a positive constant* $K=K(u,p)$ *such that,* $\forall\, t\in[0,T]$, *the truncation error defined in* (3.4) *satisfies the following bound:*

$$\|\mathcal{A}_h^{-1}T_h(t)\|_0\le Kh^{r+1}. \tag{3.15}$$

Proof. In view of definition (3.4), we observe that

$$\|\mathcal{A}_h^{-1}T_h(t)\|_0\le\|\mathcal{A}_h^{-1}\Pi_{h,r}(\dot s_h-u_t)\|_0+\|\mathcal{A}_h^{-1}\Pi_{h,r}(F(s_h,s_h)-F(u,u))\|_0.$$

We will use (2.14) with $s=2$ to bound both terms on the right-hand side. For the first, we obtain

$$\begin{aligned}\|\mathcal{A}_h^{-1}\Pi_{h,r}(\dot s_h-u_t)\|_0 &\le Ch^2\|\dot s_h-u_t\|_0+\|\mathcal{A}^{-1}\Pi(\dot s_h-u_t)\|_0\\ &\le Ch^2\|\dot s_h-u_t\|_0+\|\dot s_h-u_t\|_{-2}\le Ch^{r+1}(\|u_t\|_r+\|p_t\|_{H^{r-1}/\mathbb{R}}),\end{aligned}$$

where in the last inequality we have used that $\|\cdot\|_{-2}\le\|\cdot\|_{-1}$ and applied (2.11). As regards the second term, applying (3.7) from Lemma 3.1 and (3.13) from Lemma 3.4, we get

$$\begin{aligned}&\|\mathcal{A}_h^{-1}\Pi_{h,r}\left(F(s_h,s_h)-F(u,u)\right)\|_0\\ &\quad\le Ch^2\|F(s_h,s_h)-F(u,u)\|_0+\|\mathcal{A}^{-1}\Pi(F(s_h,s_h)-F(u,u))\|_0\\ &\quad\le Kh^2\|s_h-u\|_1+K(\|s_h-u\|_{-1}+\|s_h-u\|_1\|s_h-u\|_0).\end{aligned}$$

We observe that although Lemma 3.1 has been stated for functions $v_h\in V_{h,r}$ satisfying (3.5) can equally be applied for $v_h=u$. To conclude, we apply estimates (2.9) and (2.11) to get

$$\|\mathcal{A}_h^{-1}\Pi_{h,r}\left(F(s_h,s_h)-F(u,u)\right)\|_0\le Kh^{r+1}\left(\|u\|_r+\|p\|_{H^{r-1}/\mathbb{R}}\right). \qquad \square$$

Proposition 3.6 (consistency). *Let (u,p) be the solution of* (1.1)–(1.2). *Then, there exists a positive constant $K=K(u,p,\nu)$ such that*

$$\max_{0\le t\le T}\left\| \int_0^t e^{-\nu(t-s)\mathcal{A}_h} T_h(s)ds\right\|_0 \le Kh^{r+1}|\log(h)|. \tag{3.16}$$

Proof. Let us start by noticing that

$$\left\| \int_0^t e^{-\nu(t-s)\mathcal{A}_h} T_h(s)ds\right\|_0 \le \int_0^t \left\| \mathcal{A}_h e^{-\nu(t-s)\mathcal{A}_h} \mathcal{A}_h^{-1} T_h(s)\right\|_0 ds.$$

By virtue of Lemma 3.3, the last integral reduces to

$$\int_0^t \left\| \mathcal{A}_h e^{-\nu(t-s)\mathcal{A}_h} \Pi_{h,r}\mathcal{A}_h^{-1} T_h(s)\right\|_0 ds \le \frac{C}{\nu}|\log(h)| \max_{0\le t\le T}\left\|\mathcal{A}_h^{-1}T_h(t)\right\|_0,$$

and then, since Lemma 3.5 provides the required estimate for the truncation error, we reach (3.16). □

Theorem 3.7 (superconvergence for the velocity). *Let (u,p) be the solution of* (1.1)–(1.2)*, let s_h be the Stokes projection of u, and let u_h be the Hood–Taylor element approximation to u. Then, there exist positive constants $K(u,p,\nu)$ and h_0 such that, for every $h\in(0,h_0]$,*

$$\max_{0\le t\le T}\|s_h(t)-u_h(t)\|_0 \le K(u,p,\nu)h^{r+1}|\log(h)|. \tag{3.17}$$

Proof. Since $u_h(0)=s_h(0)$, the proof follows from Proposition 3.2 (applied to $v_h=u_h$) and Proposition 3.6. The threshold condition (3.5) needed for Proposition 3.2 to be valid is easily proved by a standard bootstrap argument (see, e.g., [20] and [3]). □

Next, we derive the superconvergence result for the error between the MFE approximation u_h to the velocity and the Stokes projection s_h in the H^1 norm.

Corollary 3.8. *Let (u,p) be the solution of the Navier–Stokes problem* (1.1)–(1.2)*, let s_h be the discrete Stokes projection of u, and let u_h be the Hood–Taylor element approximation to u. Then, there exist positive constants $K(u,p,\nu)$ and h_0 such that, for every $h\in(0,h_0]$, the following bound holds:*

$$\max_{0\le t\le T}\|s_h(t)-u_h(t)\|_1 \le K(u,p,\nu)h^{r}|\log(h)|. \tag{3.18}$$

Proof. The result follows from Theorem 3.7 and the inverse inequality (2.3). □

As a consequence of Theorem 3.7 and Corollary 3.8, the optimal rate of convergence for u_h is obtained.

Corollary 3.9. *Let (u,p) be the solution of the Navier–Stokes problem* (1.1)–(1.2)*, and let the conditions of Theorem* 3.7 *be satisfied. Then, for $s=0,1$,*

$$\max_{0\le t\le T}\|u(t)-u_h(t)\|_s \le Ch^{r-s}\Big(\max_{0\le t\le T}(\|u\|_r+\|p\|_{H^{r-1}/\mathbb{R}})+K(u,p,\nu)h|\log(h)|\Big). \tag{3.19}$$

Proof. By rewriting $u-u_h=(u-s_h)+(s_h-u_h)$, and appealing to Theorem 3.7 and Corollary 3.8 together with estimate (2.9), we reach (3.19). □

The following lemma provides several estimations (in different norms) for the time derivative of the error in the MFE approximation to the velocity.

LEMMA 3.10. *Let (u,p) be the solution of* (1.1)–(1.2), *and let $u_h : [0,T] \to V_{h,r}$ be the Hood–Taylor approximation to the velocity. Then, the following estimates hold:*

$$\max_{0<t\le T} \left\| u_t(t) - \dot{u}_h(t)\right\|_0 \le K(u,p,\nu) h^{r-1} |\log(h)|, \tag{3.20}$$

$$\max_{0<t\le T} \left\| \mathcal{A}^{-1}\Pi(u_t(t) - \dot{u}_h(t))\right\|_0 \le K(u,p,\nu) h^{r+1} |\log(h)|, \tag{3.21}$$

$$\max_{0<t\le T} \left\| u_t(t) - \dot{u}_h(t)\right\|_{-1} \le K(u,p,\nu) h^{r} |\log(h)|. \tag{3.22}$$

Proof. For simplicity, we shall drop the explicit dependence on the time t in the proof. We consider the splitting

$$u_t - \dot{u}_h = (u_t - \dot{s}_h) + (\dot{s}_h - \dot{u}_h).$$

Since the first term can be readily estimated in the different norms by means of (2.9) and (2.11), we will concentrate only on the second one in the rest of the proof. Let us denote $e_h = s_h - u_h$. The time derivative of e_h satisfies the equation

$$\dot{e}_h = -\nu \mathcal{A}_h e_h + B_h(u_h,u_h) - B_h(u,u) + \Pi_{h,r}(\dot{s}_h - u_t).$$

We shall start by proving (3.20). Applying the inverse inequality (2.3), the stability of $\Pi_{h,r}$ in the L^2 norm, (3.7) from Lemma 3.1 and (2.9), we get

$$\begin{aligned}
\|\dot{e}_h\|_0 &\le \nu \left\|\mathcal{A}_h^{1/2}\mathcal{A}_h^{1/2} e_h\right\|_0 + \|B_h(u_h,u_h) - B_h(u,u)\|_0 + \|\Pi_{h,r}(\dot{s}_h - u_t)\|_0 \\
&\le C\nu h^{-1}\left\|\mathcal{A}_h^{1/2} e_h\right\|_0 + K\|e_h\|_1 + Ch^r(\|u_t\|_r + \|p_t\|_{H^{r-1}/\mathbb{R}}) \\
&\le (C\nu h^{-1} + K)\|e_h\|_1 + O(h^r) \le (C\nu h^{-1} + K)K(u,p,\nu)h^r|\log(h)| + O(h^r)
\end{aligned}$$

after applying Corollary 3.8 in the last inequality, and so (3.20) is shown. Notice that Lemma 3.1 has been applied for $v_h = u_h$ and taking u instead of s_h. It is immediate to check that the proof of the lemma remains valid in this case.

We deal next with (3.21). We first observe that

$$\begin{aligned}
\|\mathcal{A}^{-1}\Pi\dot{e}_h\|_0 \le{}& \nu\|\mathcal{A}^{-1}\Pi A_h e_h\|_0 + \|\mathcal{A}^{-1}\Pi(B_h(u_h,u_h) - B_h(u,u))\|_0 \\
&+\|\mathcal{A}^{-1}\Pi(\dot{s}_h - \Pi_{h,r}u_t)\|_0.
\end{aligned} \tag{3.23}$$

Let us now bound each term on the right-hand side of (3.23). For the first, taking into account the relation (2.15), and applying the inverse inequality (2.3) and Theorem 3.7, we obtain

$$\|\mathcal{A}^{-1}\Pi\mathcal{A}_h e_h\|_0 \le Ch^2\|\mathcal{A}_h e_h\|_0 + \left\|\mathcal{A}_h^{-1}\mathcal{A}_h e_h\right\|_0 \le C\|e_h\|_0 \le K(u,p,\nu)h^{r+1}|\log(h)|.$$

As regards the second term, by writing $\mathcal{A}^{-1}\Pi\Pi_{h,r} = (\mathcal{A}^{-1}\Pi - \mathcal{A}_h^{-1}\Pi_{h,r})\Pi_{h,r} + (\mathcal{A}_h^{-1}\Pi_{h,r} - \mathcal{A}^{-1}\Pi) + \mathcal{A}^{-1}\Pi$, then (3.7) and Lemma 3.4 give

$$\begin{aligned}
\left\|\mathcal{A}^{-1}\Pi(B_h(u_h,u_h) - B_h(u,u))\right\|_0 &= \left\|\mathcal{A}^{-1}\Pi[\Pi_{h,r}(F(u_h,u_h) - F(u,u))]\right\|_0 \\
&\le Ch^2\left\|F(u_h,u_h) - F(u,u)\right\|_0 + K(\|u_h - u\|_{-1} + \|u_h - u\|_1\|u_h - u\|_0) \\
&\le Kh^2\|u_h - u\|_1 + K(\|e_h\|_0 + \|s_h - u\|_{-1} + \|u_h - u\|_1\|u_h - u\|_0),
\end{aligned}$$

so that applying Theorem 3.7, (2.11), and (3.19) the desired bound for this term is reached. Finally, for the last term on the right-hand side of (3.23), we use (2.17) and (2.11) to get

$$\|\mathcal{A}^{-1}\Pi(\dot{s}_h - \Pi_{h,r}u_t)\|_0 \leq \left\|\mathcal{A}^{-1/2}\Pi(\dot{s}_h - \Pi_{h,r}u_t)\right\|_0 \leq C\|s_h - u_t\|_{-1}$$
$$\leq Ch^{r+1}(\|u_t\|_r + \|p_t\|_{H^{r-1}/\mathbb{R}}).$$

To conclude, we now show (3.22). As we show in Lemma 3.11

$$\|\dot{e}_h\|_{-1} \leq Ch\|\dot{e}_h\|_0 + C\left\|\mathcal{A}^{-1/2}\Pi\dot{e}_h\right\|_0.$$

We have already proved that $\|\dot{e}_h\|_0 \leq Kh^{r-1}|\log(h)|$. Reasoning exactly as we did with $\|\mathcal{A}^{-1}\Pi\dot{e}_h\|_0$, we also get $\|A^{-1/2}\Pi\dot{e}_h\|_0 \leq Kh^r|\log(h)|$, and then the proof is complete. □

Lemma 3.11. *There exists a positive constant independent of h such that*

$$\|f_h\|_{-1} \leq Ch\|f_h\|_0 + C\left\|\mathcal{A}^{-1/2}\Pi f_h\right\|_0 \qquad \forall f \in V_{h,r}.$$

Proof. For $\phi \in H_0^1(\Omega)$ we have the (L^2-orthogonal) decomposition $\phi = \Pi\phi + (I - \Pi)\phi$, for which we have that $(I - \Pi)\phi = \nabla\chi$ for some $\chi \in H^2(\Omega)$ and, for some constant $C > 0$,

$$\|\Pi\phi\|_1 \leq C\,\|\phi\|_1\,, \qquad \|\nabla\chi\|_1 \leq C\,\|\phi\|_1 \tag{3.24}$$

(see, e.g., [10]). Thus, $(f_h, \phi) = (f_h, \Pi\phi) + (f_h, \nabla\chi)$. But, on the one hand, $(f_h, \Pi\phi) = (\Pi f_h, \Pi\phi) = (\mathcal{A}^{-1/2}\Pi f_h, \mathcal{A}^{1/2}\Pi\phi)$; on the other hand, since $f_h \in V_{h,r}$, we may write $(f_h, \nabla(\chi - I_h(\chi))$, where $I_h(\chi)$ is the standard interpolant of χ in $\hat{S}_{h,r-1}$. Now, standard interpolation bounds and (3.24) finish the proof. □

Theorem 3.12 (superconvergence for the pressure). *Let (u, p) be the solution of the Navier–Stokes equations* (1.1)–(1.2)*; let p_h be the Hood–Taylor approximation to the pressure p, and let q_h be the MFE approximation to p in the Stokes problem* (2.8). *Then, there exist positive constants $K(u, p, \nu)$ and h_0 such that, for every $h \in (0, h_0]$,*

$$\max_{0\leq t\leq T}\|p_h(t) - q_h(t)\|_{L^2/\mathbb{R}} \leq \frac{1}{\beta}K(u, p, \nu)h^r|\log(h)|, \tag{3.25}$$

where β is the constant in the inf-sup *condition* (2.4).

Proof. Subtracting (2.8) from (2.18), we obtain for the difference $p_h - q_h$

$$(p_h - q_h, \nabla\cdot\phi_h) = \nu(\nabla(u_h - s_h), \nabla\phi_h) + (F(u_h, u_h) - F(u, u), \phi_h) + (\dot{u}_h - u_t, \phi_h)$$

$\forall\phi_h \in X_{h,r}$. Using the inf-sup condition (2.4),

$$\beta\|p_h - q_h\|_{L^2/\mathbb{R}} \leq \nu\|u_h - s_h\|_1 + \|F(u_h, u_h) - F(u, u)\|_{-1} + \|\dot{u}_h - u_t\|_{-1}.$$

Applying Corollary 3.8, (3.8) from Lemma 3.1, and (3.20) from Lemma 3.10, we get

$$\beta\|p_h - q_h\|_{L^2/\mathbb{R}} \leq \nu Kh^r|\log(h)| + \|u - u_h\|_0 + Kh^r|\log(h)|.$$

Finally, thanks to Corollary 3.9, (3.25) is reached. □

As a consequence of Theorem 3.12 and (2.10), and by writing $p - p_h = (p - q_h) + (q_h - p_h)$, we also obtain the optimal rate of convergence for of the pressure.

Corollary 3.13. *Let (u, p) be the solution of the Navier–Stokes equations* (1.1)–(1.2)*, and let the conditions of Theorem* 3.12 *be satisfied. Then,*

$$\max_{0\leq t\leq T}\|p(t) - p_h(t)\|_{L^2/\mathbb{R}} \leq Ch^{r-1}\max_{0\leq t\leq T}\left(\|u\|_r + \|p\|_{H^{r-1}/\mathbb{R}}\right) + K(u, p, \nu)h^r. \tag{3.26}$$

Next, we state the rate of convergence of the postprocessed MFE approximation $(\tilde u_{\tilde h}, \tilde p_{\tilde h}) \in (\widetilde X, \widetilde Q)$ that solves (2.20)–(2.21).

THEOREM 3.14. *Let $T > 0$ be fixed. Let (u_h, p_h) be the MFE approximation to the solution (u, p) of* (1.1)–(1.2)*, and let $(\tilde u_{\tilde h}, \tilde p_{\tilde h})$ be the postprocessed MFE approximation at time T. Then, there exist constants $K_1(u, p, \nu)$, $K_0(u, p, \nu)$ such that*

(i) *if the postprocessing element is $(\widetilde X, \widetilde Q) = (X_{\tilde h,r}, Q_{\tilde h,r-1})$, then*

$$\|u(T) - \tilde u_{\tilde h}\|_1 \le C(\tilde h)^{r-1}\big(\|u(T)\|_r + \|p(T)\|_{H^{r-1}/\mathbb{R}}\big) + K_1(u, p, \nu) h^r |\log(h)|, \tag{3.27}$$

$$\|u(T) - \tilde u_{\tilde h}\|_0 \le C(\tilde h)^{r}\big(\|u(T)\|_r + \|p(T)\|_{H^{r-1}/\mathbb{R}}\big) + K_0(u, p, \nu) h^{r+1} |\log(h)|; \tag{3.28}$$

(ii) *if at time T the solution $(u(T), p(T))$ belongs to $(H^{r+1}(\Omega)^d \cap V) \times H^r(\Omega)/\mathbb{R}$, and the postprocessing element is $(\widetilde X, \widetilde Q) = (X_{h,r+1}, Q_{h,r})$, then*

$$\|u(T) - \tilde u_{\tilde h}\|_1 \le C h^{r}\big(\|u(T)\|_{r+1} + \|p(T)\|_{H^{r}/\mathbb{R}}\big) + K_1(u, p, \nu) h^r |\log(h)|, \tag{3.29}$$

$$\|u(T) - \tilde u_{\tilde h}\|_0 \le C h^{r+1}\big(\|u(T)\|_{r+1} + \|p(T)\|_{H^{r}/\mathbb{R}}\big) + K_0(u, p, \nu) h^{r+1} |\log(h)|. \tag{3.30}$$

Proof. Let $\widetilde S_{\tilde h}(u) \in \widetilde V$ be the Stokes projection of the solution of (1.1)–(1.2) at time T that satisfies

$$\begin{aligned} \Big(\nabla \widetilde S_{\tilde h}(u), \nabla \tilde\chi_h\Big) &= \big(\nabla u(T), \nabla \tilde\chi_h\big) - \big(p(T), \nabla\cdot \tilde\chi_h\big) \\ &= \big(f(T) - u_t(T) - F(u(T), u(T)), \tilde\chi_h\big) \quad \forall \tilde\chi_h \in \widetilde V. \end{aligned} \tag{3.31}$$

Then, we consider the splitting $\|u(T) - \tilde u_{\tilde h}\|_l \le \|u(T) - \widetilde S_{\tilde h}(u)\|_l + \|\widetilde S_{\tilde h}(u) - \tilde u_{\tilde h}\|_l$, $l = 0, 1$. The first term can be readily estimated by using (2.9), so that, for $l = 0, 1$,

$$\|u(T) - \widetilde S_{\tilde h}(u)\|_l \le \begin{cases} C(\tilde h)^{r-l}(\|u(T)\|_r + \|p(T)\|_{H^{r-1}/\mathbb{R}}), & \widetilde V = \widetilde V_{\tilde h,r}, \\ C h^{r+1-l}(\|u(T)\|_{r+1} + \|p(T)\|_{H^{r}/\mathbb{R}}), & \widetilde V = \widetilde V_{h,r+1}. \end{cases}$$

We will concentrate now on the second term. Subtracting (3.31) from (2.22), one finds

$$\begin{aligned} \nu(\nabla(\tilde u_{\tilde h} - \widetilde S_{\tilde h}(u)), \nabla \tilde\chi_h) &= b_{\tilde h}(u(T), u(T), \tilde\chi_h) - b_{\tilde h}(u_h(T), u_h(T), \tilde\chi_h) \\ &\quad + (u_t(T) - \dot u_h(T), \tilde\chi_h) \quad \forall \tilde\chi_h \in \widetilde V. \end{aligned} \tag{3.32}$$

Then, by setting $\tilde\chi_h = \tilde u_{\tilde h} - \widetilde S_{\tilde h}(u) \in \widetilde V$, we find

$$\nu\|\nabla(\tilde u_{\tilde h} - \widetilde S_{\tilde h}(u))\|_0 \le \big\|F(u(T), u(T)) - F(u_h(T), u_h(T))\big\|_{-1} + \|u_t(T) - \dot u_h(T)\|_{-1}.$$

For the first term above, applying (3.8) from Lemma 3.1 and Corollary 3.9, we get

$$\begin{aligned} \big\|F(u(T), u(T)) - F(u_h(T), u_h(T))\big\|_{-1} &\le K\|u(T) - u_h(T)\|_0 \\ &\le K h^r(\|u(T)\|_r + \|p(T)\|_{H^{r-1}/\mathbb{R}}). \end{aligned}$$

For the second term, (3.22) from Lemma 3.10 gives $\|u_t - \dot u_h\|_{-1} \le K h^r |\log(h)|$. Then

$$\|\tilde u_{\tilde h} - \widetilde S_{\tilde h}(u)\|_1 \le K_1(u, p, \nu) h^r |\log(h)|, \tag{3.33}$$

and the proof for the H^1 norm is complete. We next deal with the estimate in the L^2 norm. Writing (3.32) in abstract operator form, we find that

$$\nu \widetilde{\mathcal A}_{\tilde h}(\tilde u_{\tilde h} - \widetilde S_{\tilde h}(u)) = \widetilde\Pi_{\tilde h}[F(u(T), u(T)) - F(u_h(T), u_h(T))] + \widetilde\Pi_{\tilde h}[u_t(T) - \dot u_h(T)].$$

Then, applying $\widetilde{\mathcal{A}}_{\tilde h}^{-1}$ to both sides of the above equation, we obtain

$$\begin{aligned}\|\tilde u_{\tilde h} - \widetilde S_{\tilde h}(u)\|_0 \le \frac{1}{\nu}\Big(&\big\|\widetilde{\mathcal{A}}_{\tilde h}^{-1}\widetilde\Pi_{\tilde h}[F(u(T),u(T)) - F(u_h(T),u_h(T))]\big\|_0\\ &+\big\|\widetilde{\mathcal{A}}_{\tilde h}^{-1}\widetilde\Pi_{\tilde h}[u_t(T)-\dot u_h(T)]\big\|_0\Big).\end{aligned}$$

Thus, our aim is reduced to estimate each of the above norms. As regards the nonlinear term, taking into account (2.14), with $s=2$, we find

$$\begin{aligned}\big\|\widetilde{\mathcal{A}}_{\tilde h}^{-1}\widetilde\Pi_{\tilde h}[F(u,u)-F(u_h,u_h)]\big\|_0 \le{}& C\tilde h^2\big\|F(u,u)-F(u_h,u_h)\big\|_0\\ &+\big\|\mathcal{A}^{-1}\Pi[F(u,u)-F(u_h,u_h)]\big\|_0.\end{aligned}$$

Now, using estimates (3.7) from Lemma 3.1 and (3.13) from Lemma 3.4, we get

$$\|\widetilde{\mathcal{A}}_{\tilde h}^{-1}\widetilde\Pi_{\tilde h}[F(u,u)-F(u_h,u_h)]\big\|_0 \le C\tilde h^2\|u-u_h\|_1 + C(\|u-u_h\|_{-1}+\|u-u_h\|_0\|u-u_h\|_1).$$

To conclude, we shall estimate each term in both sums. The required estimates in the L^2 and H^1 norms are granted by Corollary 3.9. As regards the estimate in the H^{-1} norm, note that by means of (2.11) and (3.17), one readily finds

$$\begin{aligned}\|u-u_h\|_{-1} &\le \|u-s_h\|_{-1} + \|s_h-u_h\|_{-1} \le \|u-s_h\|_{-1}+\|s_h-u_h\|_0\\ &\le Ch^{r+1}(\|u\|_r + \|p\|_{H^{r-1}/\mathbb{R}}) + Kh^{r+1}|\log(h)|.\end{aligned}$$

Then, we finally get $\big\|\widetilde{\mathcal{A}}_{\tilde h}^{-1}\widetilde\Pi_{\tilde h}[F(u,u)-F(u_h,u_h)]\big\|_0 \le Kh^{r+1}|\log(h)|$. We next deal with the estimate for the time derivative. Applying again (2.14) with $s=2$ together with estimates (3.20) and (3.21) from Lemma 3.10, we reach

$$\begin{aligned}\big\|\widetilde{\mathcal{A}}_{\tilde h}^{-1}\widetilde\Pi_{\tilde h}\,[u_t(T)-\dot u_h(T)]\big\|_0 &\le \tilde h^2\,\|u_t(T)-\dot u_h(T)\|_0 + \|\mathcal{A}^{-1}\Pi[u_t(T)-\dot u_h(T)]\|_0\\ &\le K\tilde h^2 h^{r-1}|\log(h)| + Kh^{r+1}|\log(h)| \le Kh^{r+1}|\log(h)|.\end{aligned}$$

Hence the proof for the L^2 norm is also finished. □

THEOREM 3.15. *Let $T>0$ be fixed. Let (u_h,p_h) be the MFE approximation to the solution (u,p) of* (1.1)–(1.2). *Let $(\tilde u_{\tilde h},\tilde p_{\tilde h})$ be the postprocessed MFE approximation at time T. Then, there exists a constant $K(u,p,\nu)$ such that*

(i) *if the postprocessing element is $(\widetilde X,\widetilde Q) = (X_{\tilde h,r},Q_{\tilde h,r-1})$, then*

$$\begin{aligned}\|p(T)-\tilde p_{\tilde h}\|_{L^2/\mathbb{R}} \le{}& C_\beta(\tilde h)^{r-1}\big(\|u(T)\|_r + \|p(T)\|_{H^{r-1}/\mathbb{R}}\big)\\ &+K(u,p,\nu,\beta)h^r|\log(h)|;\end{aligned}\tag{3.34}$$

(ii) *if at time T the solution $(u(T),p(T))$ belongs to $(H^{r+1}(\Omega)^d\cap V)\times H^r(\Omega)/\mathbb{R}$, and the postprocessing element is $(\widetilde X,\widetilde Q)=(X_{h,r+1},Q_{h,r})$, then*

$$\|p(T)-\tilde p_{\tilde h}\|_{L^2/\mathbb{R}} \le C_\beta h^r\big(\|u(T)\|_{r+1}+\|p(T)\|_{H^r/\mathbb{R}}\big) + K(u,p,\nu,\beta)h^r|\log(h)|.\tag{3.35}$$

Proof. Let us denote by $\tilde q_{\tilde h}$ the MFE approximation to the pressure $p(T)$ obtained by solving the Stokes problem (2.8) at time T in the postprocessed space $(\widetilde X,\widetilde Q)$. Adding and subtracting $\tilde q_{\tilde h}$, we get

$$\|p(T)-\tilde p_{\tilde h}\|_{L^2/\mathbb{R}} \le \|p(T)-\tilde q_{\tilde h}\|_{L^2/\mathbb{R}} + \|\tilde q_{\tilde h}-\tilde p_{\tilde h}\|_{L^2/\mathbb{R}}.$$

The first term can easily be estimated applying (2.10):

$$\|p(T)-\tilde q_{\tilde h}\|_{L^2/\mathbb{R}} \le \begin{cases} C_\beta(\tilde h)^{r-1}\big(\|u(T)\|_r \\ +\|p(T)\|_{H^{r-1}/\mathbb{R}}\big), & \big(\widetilde X,\widetilde Q\big)=(X_{\tilde h,r},Q_{\tilde h,r-1}),\\ C_\beta h^r\big(\|u(T)\|_{r+1} \\ +\|p(T)\|_{H^r/\mathbb{R}}\big), & \big(\widetilde X,\widetilde Q\big)=(X_{h,r+1},Q_{\tilde h,r}).\end{cases}$$

Let us now bound the second term. Using the equations that satisfy $\tilde p_{\tilde h}$ and $\tilde q_{\tilde h}$ ((2.20), (2.8), respectively), we deduce

$$\Big(\tilde p_{\tilde h}-\tilde q_{\tilde h},\nabla\cdot\tilde\phi\Big)=\nu\Big(\nabla\big(\tilde u_{\tilde h}-\widetilde S_{\tilde h}(u)\big),\nabla\tilde\phi\Big)+\big(F(u_h,u_h)\\ -F(u,u),\tilde\phi\big)+\big(\dot u_h-u_t,\tilde\phi\big)\qquad\forall\tilde\phi\in\widetilde X.$$

Using the inf-sup condition (2.4), we obtain

$$\beta\|\tilde p_{\tilde h}-\tilde q_{\tilde h}\|_{L^2/\mathbb{R}}\le\nu\|\tilde u_{\tilde h}-\widetilde S_{\tilde h}\|_1+\|F(u_h,u_h)-F(u,u)\|_{-1}+\|u_h-u_t\|_{-1}.$$

Taking into account (3.33), (3.8) from Lemma 3.1, and (3.22) from Lemma 3.10, we reach

$$\|\tilde p_{\tilde h}-\tilde q_{\tilde h}\|_{L^2/\mathbb{R}}\le\frac{1}{\beta}\left(Kh^r|\log(h)|+\|u-u_h\|_0+Kh^r|\log(h)|\right),$$

so that, applying Corollary 3.8, we have completed the proof. □

Remark 3.1. Observe that for the velocity we used piecewise polynomials of degree at least 2. In general, the postprocessed method does not increase the rate of convergence in the L^2 norm in the linear case although an improvement in the energy norm is obtained. The application of the postprocessing technique to the mini-element approximation to Navier–Stokes equations is studied in [3], [5].

4. Numerical experiments. In this section, we present some numerical experiments in order to support the analysis developed in the paper and to assess the merit of the postprocessed method when compared with the standard MFE method. We consider the Navier–Stokes equations (1.1) over the domain $\Omega=[0,1]\times[0,1]$ subject to homogeneous Dirichlet boundary conditions. The value of the viscosity in the experiments is $\nu=1$, and the final time is $T=1.2$. We set to zero the initial velocity field u_0 (1.2) and choose the external force f so that the exact solution is

$$\begin{aligned} u^1(x,y,t)&=-6\cdot[1-\cos(\pi t)]\big(\sin^3(\pi x)\sin^2(\pi y)\cos(\pi y)\big),\quad (x,y,t)\in\Omega\times[0,T],\\ u^2(x,y,t)&=6\cdot\big(1-\cos(\pi t)\big)\big(\sin^2(\pi x)\sin^3(\pi y)\cos(\pi x)\big),\quad (x,y,t)\in\Omega\times[0,T],\\ p(x,y,t)&=(\sin(2\pi t)/2)\big(\sin^4(\pi x)+\sin^3(\pi y)\big)-\mathfrak{p}_0,\quad (x,y,t)\in\Omega\times[0,T],\end{aligned}$$

where $\mathfrak{p}_0$ denotes the mean of the pressure. In spite of the simplicity of this solution and its lack of physical meaning, we remark that our main interest has been to check the improvement in the rate of convergence achieved with the postprocessing technique and whether this also increases the efficiency of the standard MFE approximation.

In our calculations we take the so-called regular pattern triangulations of Ω, which are induced by the set of nodes $(i/N,j/N)$, $0\le i,\ j\le N$, where $N=|\Omega|/h$ is an integer. The MFE approximation to (1.1)–(1.2) is carried out using the Hood–Taylor element $(X_{h,3},Q_{h,2})$ that we will denote by $P2P1$. That is, we use Lagrange quadratic

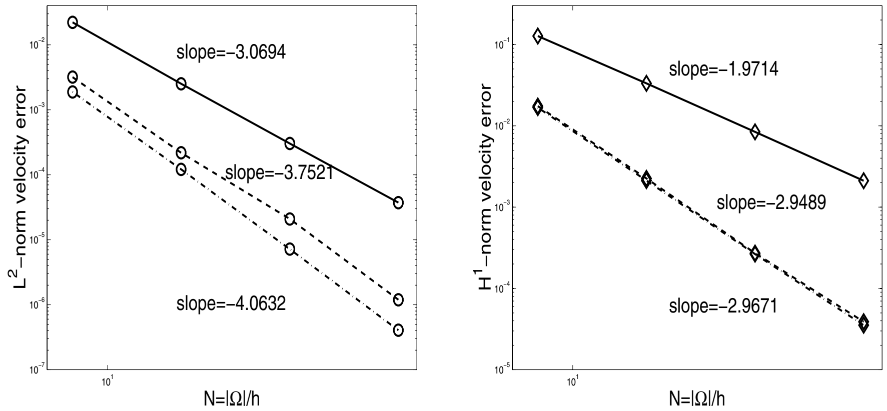


FIG. 4.1. *Convergence diagrams for the first component of the velocity with $P2P1$ (continuous line), $P3P2$ (dashed-dotted line), and the postprocessed method with $P3P2$ (dashed line). On the left the errors are measured in the L^2 norm (circles $\circ$) and on the right in the H^1 norm (diamonds $\diamondsuit$).*

elements for the approximation to the velocity and linear elements to approximate the pressure. For the postprocessing step, due to the smoothness of the solution (u, p), we perform the experiments not only with the same MFE over a finer grid, $(X_{h',3}, Q_{h',2})$, $h' < h$, but also with the higher-order Hood–Taylor element over the same grid, $(X_{h,4}, Q_{h,3})$; i.e., Lagrange cubic for the velocity and Lagrange quadratic for the pressure. This element will be denoted by $P3P2$.

For the time integration we use the well-known semi-implicit method where linear terms are approximated by the implicit midpoint rule (i.e., the Crank–Nicolson method) and nonlinear terms by the two-step explicit Adams formula (see, e.g., [8, p. 105]). The modified Stokes problems that arise at each step are solved by means of a standard projection method [31, pp. 27–28] (see also [3, section 4.6]).

For each h used in the triangulations of Ω, every experiment was carried out with different values of the time step dt. There is always a point, depending on h, at which further reduction of the time step dt does not reduce the errors anymore. This means that the error arising from the time discretization is smaller than the error arising from the MFE discretization. To avoid wrong conclusions from our numerical experiments, we have been careful to ensure that the dominant error in all the computations presented here is the spatial discretization error. For the computational cost in the efficiency diagrams shown here, we use the largest time step among those in which the spatial discretization error is dominant.

In what follows, we use the same symbols in all the plots to represent the relative errors. For the velocity we plot the errors in the first component. Similar errors are obtained for the second. The different methods are distinguished by the line used to join the symbols. For the MFE-$P2P1$ approximation, we use continuous line, and for the MFE-$P3P2$ dashed-dotted line. The MFE-$P2P1$ has been postprocessed in two different ways: using $P3P2$ (dashed line) and refining the mesh (dotted line).

In Figure 4.1 we present two convergence diagrams showing the errors committed by the methods when used with $h = |\Omega|/N$, $N = 8, 16, 32, 64$, both in the L^2 norm (left) and the H^1 norm (right). We have plotted the errors of the MFE-$P2P1$ and $P3P2$ methods and the postprocessed errors with $P3P2$. One can observe that the postprocessing technique with $P3P2$ provides an approximate velocity with about the

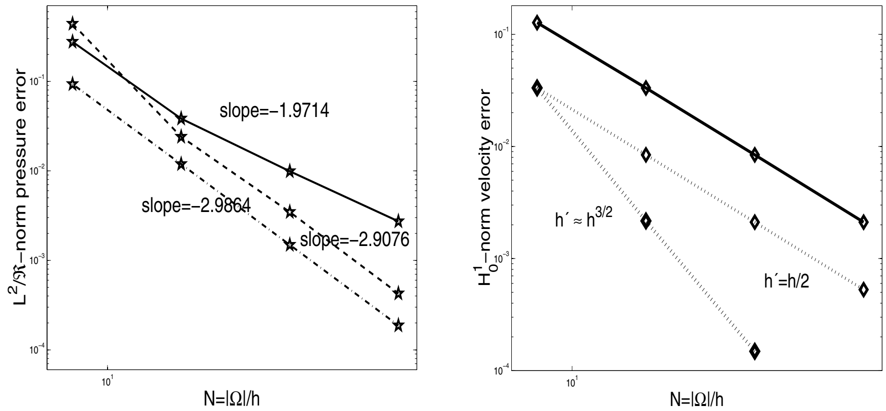


Fig. 4.2. *Left: convergence diagram for the pressure approximation with $P2P1$ (continuous line), $P3P2$ (dashed-dotted lines), and the postprocessed method with $P3P2$ (dashed lines). Right: convergence diagram for the first component of the velocity approximation with $P2P1$ (continuous line) and the postprocessed $P2P1$ over finer grids.*

same accuracy as that corresponding to the MFE-$P3P2$ method. This is especially true for the H^1 norm, in which the two methods produce virtually the same errors. Measures of the slopes of the plots confirm the rates predicted by the theory (i.e., the errors in the plots decrease like $N^{\text{slope}} = \text{const}.h^{-\text{slope}}$).

Similar conclusions can be reached from the errors of the approximations to the pressure in Figure 4.2 (left). Except for the first point, which correspond to $h = 1/8$, the postprocessed errors lies on a line (almost) parallel to the one joining the MFE-$P3P2$ errors. The rate of convergence of these two methods is one unit larger than that of the MFE-$P2P1$ in agreement with what the theory predicts.

In Figure 4.2 (right), we plot the errors obtained postprocessing the MFE-$P2P1$ refining the grid. We have represented the errors measured in the H^1 norm; similar results have been obtained for the L^2 norm. In view of Theorem 3.14, in order to get a gain of one order of convergence in the H^1 norm, we should use a mesh of size $h' \approx h^{3/2}$. The improvement in the rate of convergence of the postprocessed method can be observed in the figure. We can also observe in the plot that using a refined mesh of size $h' = h/2$ (only one regular refinement), the errors are considerably reduced. In fact, observe that the postprocessed error with $h' = h/2$ is almost the same as that of the standard MFE-$P2P1$ carried out using a mesh of size $h/2$ over the full interval $[0, T]$. This fact can be of interest when the cost of the postprocessing step with a refined mesh of size a power of h is not affordable for computational reasons.

The relevant question now is whether the improvement in the rate of convergence also implies improved efficiency. In Figure 4.3, we have represented the same errors as in Figure 4.1 (right) and Figure 4.2 (left) against the smallest amount of time needed to achieve them. We have also plotted the errors of the postprocessed method refining the mesh (Figure 4.2 (right)). In the plot we observe that the efficiency of the two postprocessing procedures is very similar. We can conclude that the postprocessed method really improves the efficiency of the standard MFE method for both approximations to the velocity and to the pressure. For any error that we may demand, the postprocessed method achieves that error in less computing time than the standard $P2P1$ and $P3P2$-MFE methods. The reason for this improvement is that the error of

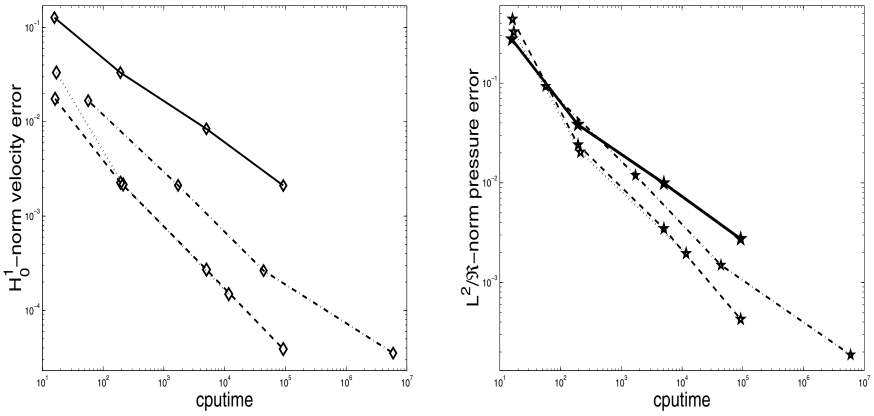


FIG. 4.3. *Efficiency diagrams for the first component of the velocity in the H^1 norm (left) and the pressure (right) with $P2P1$ (continuous line), $P3P2$ (dashed-dotted line) and the postprocessed method with $P3P2$ (dashed line) and refining the grid (dotted line).*

the MFE-$P2P1$ method is reduced when the postprocessing is done, but this is done at very little cost: that of solving a single discrete Stokes problem at the final time.

All numerical experiments were carried out on a Pentium IV, with 1 GB of Rimm memory, under the Solaris8 (Intel) operating system, with SUN Workshop 5 compilers. The programs were written in Fortran 77.

REFERENCES

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

[2] A. Ait Ou Ammi and M. Marion, *Nonlinear Galerkin methods and mixed finite elements: Two-grid algorithms for the Navier–Stokes equations*, Numer. Math., 68 (1994), pp. 189–213.

[3] B. Ayuso, *The Postprocessed Mixed Finite Element Method for the Navier–Stokes Equations*, Ph.D. thesis, Universidad Autónoma de Madrid, Madrid, Spain, 2003.

[4] B. Ayuso and B. García-Archilla, *Regularity constants of the Stokes problem. Application to finite-element methods on curved domains*, Math. Models Methods Appl. Sci., 15 (2005), pp. 437–470.

[5] B. Ayuso, J. de Frutos, and J. Novo, *Improving the accuracy of the mini-element approximation to Navier–Stokes equations*, submitted for publication.

[6] F. Brezzi and R. S. Falk, *Stability of higher-order Hood–Taylor methods*, SIAM J. Numer. Anal., 28 (1991), pp. 581–590.

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

[8] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer Ser. Comput. Math., Springer-Verlag, Berlin, 1988.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[10] P. Constantin and C. Foias, *Navier–Stokes Equations*, Chicago Lectures in Math., The University of Chicago, Chicago, 1988.

[11] C. Foias, O. Manley, and R. Temam, *Modelling of the interaction of small and large eddies in two dimensional turbulent flows*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 93–118.

[12] C. Foias, G. R. Sell, and E. S. Titi, *Exponential tracking and approximation of inertial manifolds for dissipative nonlinear equations*, J. Dynam. Differential Equations, 1 (1989), pp. 199–243.

[13] J. de Frutos, B. García-Archilla, and J. Novo, *A postprocessed Galerkin method with Chebyshev or Legendre polynomials*, Numer. Math., 86 (2000), pp. 419–442.

[14] J. de Frutos and J. Novo, *A spectral element method for the Navier–Stokes equations with improved accuracy*, SIAM J. Numer. Anal., 38 (2000), pp. 799–819.

[15] J. de Frutos and J. Novo, *Postprocessing the linear-finite-element method*, SIAM J. Numer. Anal., 40 (2002), pp. 805–819.

[16] J. de Frutos and J. Novo, *A posteriori error estimation with the p-version of the finite element method for nonlinear parabolic differential equations*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4893–4904.

[17] J. de Frutos and J. Novo, *Element-wise a posteriori estimates based on hierarchical bases for non-linear parabolic problems*, Internat. J. Numer. Methods Engrg., 63 (2005), pp. 1146–1173.

[18] B. García-Archilla, J. Novo, and E. S. Titi, *Postprocessing the Galerkin method: A novel approach to approximate inertial manifolds*, SIAM J. Numer. Anal., 35 (1998), pp. 941–972.

[19] B. García-Archilla, J. Novo, and E. S. Titi, *An approximate inertial manifold approach to postprocessing Galerkin methods for the Navier–Stokes equations*, Math. Comp., 68 (1999), pp. 893–911.

[20] B. García-Archilla and E. S. Titi, *Postprocessing the Galerkin method: The finite-element case*, SIAM J. Numer. Anal., 37 (2000), pp. 470-499.

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

[22] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, 1991.

[23] J. G. Heywood, *The Navier–Stokes equations: On the existence, regularity and decay of solutions*, Indiana Univ. Math. J., 29 (1980), pp. 639–681.

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

[25] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem. Part* III: *Smoothing property and higher order error estimates for spatial discretization*, SIAM J. Numer. Anal., 25 (1988), pp. 489–512.

[26] P. Hood and C. Taylor, *A numerical solution of the Navier–Stokes equations using the finite element technique*, Comput. Fluids, 1 (1973), pp. 73–100.

[27] C. R. Laing, A. McRobie, and J. M. T. Thompson, *The post-processed Galerkin method applied to non-linear shell vibrations*, Dynam. Stability Systems, 14 (1999), pp. 163–181.

[28] J. C. López Marcos and J. M. Sanz-Serna, *Stability and convergence in numerical analysis,* III. *Linear investigation of nonlinear stability*, IMA J. Numer. Anal., 8 (1988), pp. 71–84.

[29] H. Okamoto, *On the semidiscrete finite element approximation for the nonstationary Stokes equation*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 29 (1982), pp. 241–260.

[30] R. Temam, *Navier Stokes Equations and Nonlinear Functional Analysis*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 41, SIAM, Philadelphia, PA, 1983.

[31] S. Turek, *Efficient Solvers for Incompressible Flow Problems. An Algorithmic and Computational Approach*, Springer-Verlag, Berlin, 1999.

[32] O. Walsh, *On Approximate Inertial Manifolds for the Navier–Stokes Equations Using Finite Elements*, Ph.D. thesis, University of British Columbia, Vancouver, BC, Canada, 1994.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1112–1138

FULLY DISCRETE FINITE ELEMENT APPROXIMATION FOR ANISOTROPIC SURFACE DIFFUSION OF GRAPHS*

KLAUS DECKELNICK†, GERHARD DZIUK‡, AND CHARLES M. ELLIOTT§

Abstract. We analyze a fully discrete numerical scheme for approximating the evolution of graphs for surfaces evolving by anisotropic surface diffusion. The scheme is based on the idea of second order operator splitting for the nonlinear geometric fourth order equation. This yields two coupled spatially second order problems, which are approximated by linear finite elements. The time discretization is semi-implicit. We prove error bounds for the resulting scheme and present numerical test calculations that confirm our analysis and illustrate surface diffusion.

Key words. surface diffusion, anisotropic, geometric motion, second order operator splitting, nonlinear partial differential equation, finite element, fully discrete, error estimates, fourth order parabolic equation

AMS subject classifications. 65N30, 35K55

DOI. 10.1137/S0036142903434874

1. Introduction.

This article is concerned with the geometric problem of determining an evolving surface $\Gamma(t)$ whose motion is governed by the highly nonlinear fourth order geometric anisotropic surface diffusion equation

$$V = \Delta_\Gamma \mathcal{H}_\gamma \quad \text{on } \Gamma(t), \tag{1.1}$$

where V and Δ_Γ denote, respectively, the normal velocity and the Laplace–Beltrami (surface Laplacian) operator for $\Gamma(t)$. Furthermore, $\mathcal{H}_\gamma$ denotes the anisotropic mean curvature of the surface with respect to the positive, convex, and 1-homogeneous surface energy density $\gamma : \mathbb{R}^{n+1}\backslash\{0\} \to \mathbb{R}$. We can introduce $\mathcal{H}_\gamma$ formally as the first variation of the surface energy

$$A_\gamma(\Gamma) = \int_\Gamma \gamma(\nu), \tag{1.2}$$

where ν denotes the unit normal to Γ.

Modelling morphological surface evolution and growth is fundamental in materials science and the study of microstructure. The surface evolution law (1.1) is referred to as surface diffusion because it models the diffusion of mass within the bounding surface of a solid body. At the atomistic level atoms on the surface move along the

*Received by the editors September 17, 2003; accepted for publication (in revised form) November 11, 2004; published electronically September 23, 2005. This work was carried out whilst the authors participated in the 2003 programme "Computational Challenges in Partial Differential Equations" at the Isaac Newton Institute, Cambridge, UK. The work was supported by the Deutsche Forschungsgemeinschaft via DFG-Forschergruppe Nonlinear partial differential equations: Theoretical and numerical analysis and via DFG-Graduiertenkolleg: "Nichtlineare Differentialgleichungen: Modellierung, Theorie, Numerik, Visualisierung." The graphical presentations were performed with the packages GRAPE and Xgraph.

http://www.siam.org/journals/sinum/43-3/43487.html

†Institut für Analysis und Numerik, Otto-von-Guericke-Universität Magdeburg, Universitätsplatz 2, 39106 Magdeburg, Germany (Klaus.Deckelnick@mathematik.uni-magdeburg.de).

‡Institut für Angewandte Mathematik, Albert-Ludwigs-Universität Freiburg, Hermann–Herder–Straße 10, 79104 Freiburg i. Br., Germany (gerd@mathematik.uni-freiburg.de).

§Department of Mathematics, University of Sussex, Falmer, Brighton, BN1 9RF, UK (c.m.elliott@sussex.ac.uk).

surface due to a driving force consisting of a chemical potential difference. For a surface with surface energy density $\gamma(\nu)$ the appropriate chemical potential in this setting is the anisotropic curvature $\mathcal{H}_\gamma$. This leads to the flux law

$$\rho V = -\operatorname{div}_\Gamma \mathbf{j},$$

where ρ is the mass density and $\mathbf{j}$ is the mass flux in the surface, with the constitutive flux law [19], [21]

$$\mathbf{j} = -D\nabla_\Gamma \mathcal{H}_\gamma.$$

Here, D is the diffusion constant. From these equations we obtain the law (1.1) after an appropriate nondimensionalization. The notion of surface diffusion is due to Mullins [21] and for a review we refer the reader to [5].

Our sign convention is that $\mathcal{H}_\gamma$ with respect to the outer normal is positive for the Wulff shape $\mathcal{W} := \{p \in \mathbb{R}^{n+1} \mid \langle p, q\rangle \le \gamma(q)\ \forall q \in \mathbb{R}^{n+1}\}$.

This evolution has interesting geometrical properties: if $\Gamma(t)$ is a closed surface bounding a domain $\Omega(t)$, then the volume of $\Omega(t)$ is preserved and the surface energy (or weighted surface area) of $\Gamma(t)$ decreases. The corresponding result in the graph case is given in Lemma 2.2. At present, the existence and uniqueness theory for surface diffusion is limited to the isotropic case $\gamma(q) := |q|$, $q \in \mathbb{R}^{n+1}$. For example, it is known that for closed curves in the plane or closed surfaces in $\mathbb{R}^3$ balls are asymptotically stable subject to small perturbations; see [15], [17]. However, topological changes such as pinch-off are possible [18], [20], and a one-dimensional graph may lose its graph property in finite time whilst the surface evolves smoothly [16].

In what follows we shall study evolving surfaces $\Gamma(t)$ which can be described, for each $t \ge 0$, as the graph of a height function $u(\cdot, t)$ over some base domain $\Omega \subset \mathbb{R}^n$, i.e., $\Gamma(t) = \{(x, u(x,t)) \in \mathbb{R}^{n+1} \mid x \in \Omega\}$. The area element and a unit normal, denoted by $Q(u)$ and $\nu(u)$, are then given by

$$Q(u) = \sqrt{1 + |\nabla u|^2}, \qquad \nu(u) = \frac{(\nabla u, -1)}{\sqrt{1+|\nabla u|^2}} = \frac{(\nabla u, -1)}{Q(u)}$$

so that we can calculate the surface energy or weighted area for a graph Γ given by the height function u as

$$A_\gamma(\Gamma) = \mathcal{I}_\gamma(u) := \int_\Omega \gamma(\nu(u))Q(u) = \int_\Omega \gamma(\nabla u, -1)$$

in view of the homogeneity of γ. Thus the first variation of A_γ in the direction of a function $\phi \in C_0^\infty(\Omega)$ is

$$\begin{aligned}\frac{d}{d\epsilon}\mathcal{I}_\gamma(u + \epsilon\phi)_{|\epsilon=0} &= \sum_{i=1}^n \int_\Omega \gamma_{p_i}(\nabla u, -1)\phi_{x_i} = -\sum_{i,j=1}^n \int_\Omega \gamma_{p_ip_j}(\nabla u, -1)u_{x_ix_j}\phi \\ &= -\int_\Omega \mathcal{H}_\gamma \phi = \int_\Omega w\phi,\end{aligned}$$

where we use $-w$ to denote the anisotropic or weighted mean curvature of the surface in the graph case so that

$$w := -\sum_{i,j=1}^n \gamma_{p_ip_j}(\nabla u, -1)u_{x_ix_j}. \tag{1.3}$$

In order to translate (1.1) into a differential equation for $u = u(x,t)$, we observe that the normal velocity V of $\Gamma(t)$ is given by $V = -\frac{u_t}{Q(u)}$. Furthermore, if $v : \Omega \to \mathbb{R}$, then the Laplace–Beltrami operator on $\Gamma(t)$ is given by (see (2.5) below)

$$\Delta_\Gamma v = \frac{1}{Q(u)} \nabla \cdot \left(\left(Q(u) I - \frac{\nabla u \otimes \nabla u}{Q(u)} \right) \nabla v \right),$$

where $\otimes$ denotes the usual tensor product of two vectors in $\mathbb{R}^n$. Thus, anisotropic surface diffusion for graphs is defined by the following highly nonlinear fourth order evolutionary equation:

$$u_t = -\nabla \cdot \left(\left(Q(u) I - \frac{\nabla u \otimes \nabla u}{Q(u)} \right) \nabla \left(\sum_{i,j=1}^{n} \gamma_{p_i p_j}(\nabla u, -1) u_{x_i x_j} \right) \right). \tag{1.4}$$

The aim of this paper is to analyze a fully discrete finite element approximation of the initial-boundary value problem in the case of *graphs.* We use the second order splitting method for fourth order problems proposed by Elliott, French, and Milner [14] for the fourth order Cahn–Hilliard equation and subsequently employed for surface diffusion by Deckelnick, Dziuk, and Elliott [12]. Thus the space discretization is accomplished using H^1 conforming finite element spaces. For example, continuous piecewise linear elements on triangulations are sufficient. On the other hand, in time we use a novel semi-implicit discretization which requires only the solution of linear algebraic equations but which preserves the Liapunov structure. This ensures the natural stability properties of the scheme with a time step independent of the spatial mesh size. The scheme involves stabilizing the explicit Euler scheme by adding a semi-implicit linear form which involves the discrete time derivative. This stabilizing form has two terms. One involves the anisotropy and is designed to yield a stable linearization. The second term is of higher order with respect to the time step and is based on the Laplace–Beltrami form. It is designed to yield the L^2 stability bound, (3.11), on the discrete solution similar to that enjoyed by the solution of the partial differential equation. A similar idea was previously used in [11] for the anisotropic mean curvature flow of graphs and in [23] for surface diffusion. The main achievement of the paper is the derivation of a priori geometric error bounds. We prove optimal order bounds for the difference of the normals measured in the L^2 norm over either the continuous surface $\Gamma(t)$ or the discrete surface $\Gamma_h(t)$ and the L^2 norm on the discrete surface of the difference of the tangential gradients of the anisotropic mean curvature. This latter bound is equivalent to an H^{-1} bound on the difference in normal velocities. Some numerical computations are presented which confirm the analysis and which illustrate the effect of anisotropy.

A second order splitting finite element scheme for axially symmetric surfaces was presented by Coleman, Falk, and Moakher [7], [8] together with some stability results and interesting numerical computations illustrating pinch-off and the formation of beads. A first finite element error analysis for the second order splitting method for surface diffusion in the axially symmetric case was presented by Deckelnick, Dziuk, and Elliott [12]. Subsequently, Bänsch, Morin, and Nochetto [1] developed an optimal order continuous in time finite element error analysis for the second order splitting method in the case of multidimensional graphs. Our work has the distinctive feature of analyzing a fully discrete second order splitting finite element method for nonlinear surface anisotropy using a stable semi-implicit time stepping scheme.

Remark 1.1. The analysis is easily extended to the more general evolution law

$$V = \Delta_\Gamma(\mathcal{H}_\gamma - f) + g, \tag{1.5}$$

where f is a force arising from an extra term in the energy and g is a surface growth term. For example, including mechanical energy leads to the appearance of f and in epitaxial growth g models the deposition of atoms.

Remark 1.2. Our results are presented for zero Neumann boundary conditions with exact quadrature. The results and arguments also hold without change for the case of Ω being a box and periodic boundary conditions. Minor modifications are required for homogeneous Dirichlet boundary conditions. These three sets of conditions have the property of being variationally separated and allow the second order splitting method to work.

Remark 1.3. The approach to surface diffusion in this paper is entirely analogous to the work of Elliott, French, and Milner [14] for the Cahn–Hilliard equation where u is an order or phase field variable and w is the chemical potential. The variational gradient flow structure is identical in each setting. Indeed the degenerate Cahn–Hilliard equation yields a diffuse interface approximation to surface diffusion [4].

The paper is organized as follows. In section 2 we introduce some notation and assumptions. We set up the numerical scheme and derive some preliminary estimates in section 3, whilst section 4 contains the proof of the error bounds. Finally, section 5 contains some numerical results.

2. Notation and assumptions.

2.1. Differential geometry. Let Γ be a C^2 hypersurface in $\mathbb{R}^{n+1}$ with unit normal ν. For any function $\bar{\eta} = \bar{\eta}(x_1, \dots, x_{n+1})$ defined in a neighborhood $\mathcal{N} \subset \mathbb{R}^{n+1}$ of Γ we define its tangential gradient on Γ by

$$\nabla_\Gamma \bar{\eta} := D\bar{\eta} - \langle D\bar{\eta}, \nu\rangle \nu,$$

where on $\mathbb{R}^{n+1}$ $\langle \cdot, \cdot \rangle$ denotes the usual scalar product and $D\bar{\eta}$ denotes the usual gradient. The tangential gradient $\nabla_\Gamma \bar{\eta}$ depends only on the values of $\bar{\eta}$ on Γ and $\langle \nabla_\Gamma \bar{\eta}, \nu \rangle = 0$. The Laplace–Beltrami operator on Γ is defined as the tangential divergence of the tangential gradient, i.e.,

$$\Delta_\Gamma \bar{\eta} = \langle \nabla_\Gamma, \nabla_\Gamma \bar{\eta} \rangle.$$

Let Γ have a boundary $\partial\Gamma$ whose intrinsic unit outer normal, tangential to Γ, is denoted by μ. Then the surface Green's formula is

$$\int_\Gamma \langle \nabla_\Gamma \bar{\xi}, \nabla_\Gamma \bar{\eta} \rangle = \int_{\partial\Gamma} \bar{\xi} \langle \nabla_\Gamma \bar{\eta}, \mu \rangle - \int_\Gamma \bar{\xi} \Delta_\Gamma \bar{\eta}. \tag{2.1}$$

We now turn to the situation in hand where $\Gamma(t) = \{(x, u(x,t)) \in \mathbb{R}^{n+1} \mid x \in \Omega\}$. For functions $v = v(x)$, $x \in \Omega$, we use the extension $\bar{v}(x, x_{n+1}) = v(x)$ and define

$$\nabla_\Gamma v := \nabla_\Gamma \bar{v} = D\bar{v} - \langle D\bar{v}, \nu(u) \rangle \nu(u) = P(\nu(u)) D\bar{v},$$

where we observe that $D\bar{v} = (\nabla v, 0)$, $\nu(u) = (\nabla u, -1)/Q(u)$ and $P(\nu(u))$ is given by

$$P(\nu(u)) := I - \nu(u) \otimes \nu(u).$$

Here, we have used the tensor product notation $y \otimes y := yy^T$. It follows that

$$\langle \nabla_\Gamma v, \nabla_\Gamma \eta \rangle = \nabla v \cdot \nabla \eta - \frac{1}{Q(u)^2} \nabla v \cdot \nabla u \nabla \eta \cdot \nabla u = \frac{1}{Q(u)} (\nabla v)^t E(\nabla u) \nabla \eta, \tag{2.2}$$

where

$$E(\nabla u) := Q(u) I - \frac{\nabla u \otimes \nabla u}{Q(u)}.$$

For later use we note that

$$\langle P(\nu(u)) D\bar{v}, D\bar{w} \rangle Q(u) = (\nabla v)^t E(\nabla u) \nabla w, \tag{2.3}$$

$$(\nabla v)^t E(\nabla u) \nabla v \geq \frac{|\nabla v|^2}{Q(u)}. \tag{2.4}$$

Integrating (2.2) over Γ we derive

$$\int_\Gamma \langle \nabla_\Gamma \bar{v}, \nabla_\Gamma \bar{\eta} \rangle = \int_\Omega \langle \nabla_\Gamma v, \nabla_\Gamma \eta \rangle Q(u) = \int_\Omega (\nabla v)^t E(\nabla u) \nabla \eta.$$

If we combine this relation with (2.1) we obtain for test functions η, which vanish on $\partial\Omega$

$$\int_\Gamma \bar{\eta}\, \Delta_\Gamma \bar{v} = \int_\Omega \eta \nabla \cdot (E(\nabla u) \nabla v) = \int_\Omega \eta \frac{1}{Q(u)} \nabla \cdot (E(\nabla u) \nabla v) Q(u),$$

so that

$$\Delta_\Gamma v := \Delta_\Gamma \bar{v} = \frac{1}{Q(u)} \nabla \cdot (E(\nabla u) \nabla v). \tag{2.5}$$

2.2. The anisotropy. We suppose that $\gamma : \mathbb{R}^{n+1} \setminus \{0\} \to \mathbb{R}$ is smooth with $\gamma(p) > 0$ for $p \in \mathbb{R}^{n+1} \setminus \{0\}$ and that γ is positively homogeneous of degree one, i.e.,

$$\gamma(\lambda p) = |\lambda| \gamma(p) \quad \forall\, \lambda \neq 0,\ p \neq 0. \tag{2.6}$$

Here, $|\cdot|$ denotes the Euclidean norm. It is not difficult to verify that (2.6) implies

$$\langle \gamma'(p), p \rangle = \gamma(p), \qquad \langle \gamma''(p) p, q \rangle = 0, \tag{2.7}$$

$$\gamma_{p_i}(\lambda p) = \frac{\lambda}{|\lambda|} \gamma_{p_i}(p), \qquad \gamma_{p_i p_j}(\lambda p) = \frac{1}{|\lambda|} \gamma_{p_i p_j}(p) \tag{2.8}$$

for all $p \in \mathbb{R}^{n+1} \setminus \{0\}$, $q \in \mathbb{R}^{n+1}$, $\lambda \neq 0$, and $i, j \in \{1, \dots, n+1\}$. Finally, we assume that there exists $\gamma_0 > 0$ such that

$$\langle D^2 \gamma(p) q, q \rangle \geq \gamma_0 |q|^2 \quad \forall\, p, q \in \mathbb{R}^{n+1},\ |p| = 1,\ \langle p, q \rangle = 0. \tag{2.9}$$

Further information about the geometric properties and physical relevance of anisotropic energy functionals can be found, respectively, in [2] and [24].

2.3. Function spaces. By $(\cdot,\cdot)$ we denote the $L^2(\Omega)$ inner product $(v, \eta) := \int_\Omega v(x)\eta(x)dx$ for $v, \eta \in L^2(\Omega)$ with norm $\|v\| := (v,v)^{\frac{1}{2}}$. Also $H^{m,p}(\Omega)$ denotes the usual Sobolev space with the corresponding norm being given by $\|u\|_{H^{m,p}(\Omega)} = (\sum_{k=0}^m \|D^k u\|_{L^p(\Omega)}^p)^{\frac{1}{p}}$ with the usual modification for $p = \infty$. For $p = 2$ we simply write $H^m(\Omega) = H^{m,2}(\Omega)$ with norm $\|\cdot\|_{H^m(\Omega)}$.

2.4. The variational formulation and initial-boundary value problem. Rather than discretizing the fourth order equation (1.4) we use the height u of the graph and the anisotropic curvature of the graph w as variables and consider the two second order equations (1.1), (1.3),

$$u_t = \nabla \cdot (E(\nabla u)\nabla w), \tag{2.10}$$

$$w = -\sum_{i,j=1}^{n} \gamma_{p_i p_j}(\nabla u, -1) u_{x_i x_j}. \tag{2.11}$$

The system is closed using Neumann boundary conditions and an initial condition for u.

$$E(\nabla u)\nabla w \cdot \nu_{\partial\Omega} = 0, \tag{2.12}$$

$$\langle \gamma'(\nu(u)), (\nu_{\partial\Omega}, 0)\rangle = 0, \tag{2.13}$$

$$u(\cdot, 0) = u_0. \tag{2.14}$$

The first equation, (2.12), is the zero mass flux condition whereas the second equation, (2.13), is the natural variational boundary condition which defines w as the variational derivative or chemical potential for the surface energy functional. Note that an initial condition on w is not required.

In order to write down the variational formulation it is convenient to introduce the following forms:

Laplace–Beltrami (LB) form,

$$\mathcal{E}(u; w, \eta) := \int_\Omega (\nabla w)^t E(\nabla u)\nabla \eta dx$$

Anisotropic mean curvature (AMC) form,

$$\mathcal{A}(u, \eta) := \sum_{i=1}^{n} \int_\Omega \gamma_{p_i}(\nu(u))\eta_{x_i} dx.$$

Then it is straightforward to show the following equivalence between the classical form of the initial-boundary value problem and the variational formulation.

LEMMA 2.1. *Let $u \in C^1([0, T]; C^4(\bar{\Omega}))$, $u(\cdot, 0) = u_0$, and $w \in C^0([0, T]; C^2(\bar{\Omega}))$. Then (u, w) is a solution of* (2.10)–(2.13) *iff* $u(\cdot, 0) = u_0$ *and the following variational equations are satisfied:*

$$(\partial_t u, \eta) + \mathcal{E}(u; w, \eta) = 0 \quad \forall\, \eta \in H^1(\Omega), \tag{2.15}$$

$$(w, \eta) - \mathcal{A}(u, \eta) = 0 \quad \forall\, \eta \in H^1(\Omega). \tag{2.16}$$

LEMMA 2.2. *The solution (u, w) satisfies for each $t \in [0, T]$ the surface energy equation*

$$\mathcal{I}_\gamma(u) + \int_0^t \mathcal{E}(u; w, w)\, ds = \mathcal{I}_\gamma(u_0) \tag{2.17}$$

and the conservation laws

$$(u, 1) = (u_0, 1), \qquad (w, 1) = 0. \tag{2.18}$$

Furthermore, for each $t \in [0,T]$ we have the bound

$$\|u(t)\|^2 + \int_0^t \|w\|^2 \, ds \le C(\gamma, u_0, T). \tag{2.19}$$

Proof. Taking $\eta = w$ in (2.15) and $\eta = \partial_t u$ in (2.16) and subtracting the resulting equations yields (2.17). Taking $\eta = 1$ in (2.15) and (2.16) yields (2.18).

In order to prove the first part of (2.19), we use $\eta = u$ in (2.15) and apply (2.29) which gives

$$\frac{1}{2}\frac{d}{dt}\|u\|^2 = -\mathcal{E}(u;w,u) \le \mathcal{E}(u;w,w)^{\frac{1}{2}}\mathcal{E}(u;u,u)^{\frac{1}{2}} \le \frac{1}{2}\mathcal{E}(u;w,w) + \frac{1}{2}\int_\Omega Q(u).$$

Integrating this inequality with respect to time we obtain with the help of (2.17),

$$\|u(t)\|^2 \le \|u_0\|^2 + \int_0^t \mathcal{E}(u;w,w)ds + \frac{1}{\inf_{|p|=1}\gamma(p)}\int_0^t I_\gamma(u)\, ds \le C(\gamma, u_0, T).$$

Using $\eta = w$ in (2.16) we deduce

$$\|w\|^2 = \mathcal{A}(u,w) \le \sup_{|p|=1}|\gamma'(p)| \int_\Omega |\nabla w| \le C\left(\int_\Omega \frac{|\nabla w|^2}{Q(u)}\right)^{\frac{1}{2}}\left(\int_\Omega Q(u)\right)^{\frac{1}{2}},$$

so that (2.4) and similar arguments as above yield

$$\int_0^t \|w\|^2 ds \le C\int_0^t \mathcal{E}(u;w,w)ds + C\int_0^t Q(u)\, ds \le C(\gamma, u_0, T). \qquad \square$$

Remark 2.3. The surface energy equation (2.17) can be written as

$$\int_{\Gamma(t)} \gamma(\nu) + \int_0^T \int_{\Gamma(t)} |\nabla_\Gamma \mathcal{H}_\gamma|^2 = \int_{\Gamma(0)} \gamma(\nu). \tag{2.20}$$

The conservation of u is equivalent to the conservation of the volume lying below the graph of the surface. That the integral over Ω of the anisotropic mean curvature is zero is a consequence of the fact that constant vertical variations in the height of the graph do not change the anisotropic surface area.

2.5. Geometric lemmas. The following algebraic relations are elementary.

LEMMA 2.4.

$$|\nabla(u-v)|^2 = (Q(u)-Q(v))^2 + |\nu(u)-\nu(v)|^2 Q(u)Q(v), \tag{2.21}$$

$$\left|\frac{1}{Q(u)} - \frac{1}{Q(v)}\right| \le |\nu(u)-\nu(v)|, \tag{2.22}$$

$$|Q(u)-Q(v)| \le Q(u)Q(v)|\nu(u)-\nu(v)|. \tag{2.23}$$

LEMMA 2.5 (properties of the anisotropy and the AMC form $\mathcal{A}$). *Let $u, v \in H^{1,\infty}(\Omega)$. Then*

$$\mathcal{A}(v, u-v) \ge \mathcal{I}_\gamma(u) - \mathcal{I}_\gamma(v) - \bar{\gamma}\int_\Omega |\nu(u)-\nu(v)|^2 Q(u), \tag{2.24}$$

where

$$\bar{\gamma} := \frac{1}{\sqrt{5}-1} \max\left\{ \sup_{|p|=1} |\gamma'(p)|, \sup_{|p|=1} |\gamma''(p)| \right\}. \tag{2.25}$$

If in addition $|\nabla u| \leq K$ *a.e. in* Ω*, then*

$$|\mathcal{A}(u,\eta) - \mathcal{A}(v,\eta)| \leq C(\gamma, K) \int_\Omega |\nu(u) - \nu(v)||\nabla\eta|. \tag{2.26}$$

Proof. The first inequality follows from the estimate

$$\sum_{i=1}^{n} \gamma_{p_i}(\nu(v))(u-v)_{x_i} \geq \gamma(\nu(u))Q(u) - \gamma(\nu(v))Q(v) - \bar{\gamma}|\nu(u) - \nu(v)|^2 Q(u) \tag{2.27}$$

which is contained in the proof of Theorem 3.1 in [11, p. 430]. Let us next turn to (2.26). Lemma 6.1 in [11] implies that there exists $c_0 = c_0(K) > 0$ such that

$$|s\nu(u) + (1-s)\nu(v)| \geq c_0 \quad \text{a.e. in } \Omega \quad \forall\, s \in [0,1]. \tag{2.28}$$

Note that c_0 is independent of v. As a consequence,

$$\begin{aligned} |\gamma_{p_i}(\nu(u)) - \gamma_{p_i}(\nu(v))| &= \left| \sum_{j=1}^{n+1} \int_0^1 \gamma_{p_i p_j}(s\nu(u) + (1-s)\nu(v)) ds (\nu_j(u) - \nu_j(v)) \right| \\ &\leq \frac{1}{c_0} \max_{|p|=1} |D^2\gamma(p)||\nu(u) - \nu(v)| \leq C(\gamma, K)|\nu(u) - \nu(v)|, \end{aligned}$$

since $D^2\gamma$ is positively homogeneous of degree -1. This yields (2.26). □

LEMMA 2.6 (properties of the LB form $\mathcal{E}$). *Let* $u, v \in H^{1,\infty}(\Omega)$*. Then*

$$|\mathcal{E}(u; w, \eta)| \leq \mathcal{E}(u; w, w)^{\frac{1}{2}} \mathcal{E}(u; \eta, \eta)^{\frac{1}{2}}. \tag{2.29}$$

If in addition $|\nabla u| \leq K$ *a.e. in* Ω*, then*

$$\mathcal{E}(v; u-v, u-v) \leq C(K) \int_\Omega |\nu(u) - \nu(v)|^2 Q(v), \tag{2.30}$$

$$|\mathcal{E}(u; \eta_1, \eta_2) - \mathcal{E}(v; \eta_1, \eta_2)| \leq C(K)\|\nabla\eta_1\|_\infty \int_\Omega |\nu(u) - \nu(v)||\nabla\eta_2|Q(v), \tag{2.31}$$

$$\begin{aligned} |\mathcal{E}(u; \eta_1, \eta_2) - \mathcal{E}(v; \eta_1, \eta_2)| &\leq \epsilon\mathcal{E}(v; \eta_1, \eta_1) \\ &\quad + \frac{C(K)}{\epsilon} \|\nabla\eta_2\|_\infty^2 \int_\Omega |\nu(u) - \nu(v)|^2 Q(v). \end{aligned} \tag{2.32}$$

Proof. Using (2.3) together with Young's inequality we have

$$\begin{aligned} |\mathcal{E}(u; w, \eta)| &= \left| \int_\Omega \langle P(\nu(u))D\bar{w}, D\bar{\eta}\rangle Q(u) \right| \\ &\leq \int_\Omega \langle P(\nu(u))D\bar{w}, D\bar{w}\rangle^{\frac{1}{2}} \langle P(\nu(u))D\bar{\eta}, D\bar{\eta}\rangle^{\frac{1}{2}} Q(u) \\ &\leq \mathcal{E}(u; w, w)^{\frac{1}{2}} \mathcal{E}(u; \eta, \eta)^{\frac{1}{2}}. \end{aligned}$$

Next, observing that $(\nabla(u-v),0) = Q(u)\nu(u) - Q(v)\nu(v)$ we obtain

$$\begin{aligned}
&\langle P(\nu(v))(\nabla(u-v),0),(\nabla(u-v),0)\rangle \\
&\quad = \langle (I-(\nu(v)\otimes\nu(v)))(Q(u)\nu(u)-Q(v)\nu(v)),(Q(u)\nu(u)-Q(v)\nu(v))\rangle \\
&\quad = Q(u)^2(1-\langle\nu(u),\nu(v)\rangle^2) = Q(u)^2(1-\langle\nu(u),\nu(v)\rangle)(1+\langle\nu(u),\nu(v)\rangle) \\
&\quad \le Q(u)^2|\nu(u)-\nu(v)|^2,
\end{aligned}$$

since $1-\langle\nu(u),\nu(v)\rangle = \frac{1}{2}|\nu(u)-\nu(v)|^2$. Multiplication of the above inequality by $Q(v)$ followed by integration over Ω yields (2.30). From the definition of $P(\nu(u))$ and (2.23) we infer

$$|P(\nu(u))Q(u) - P(\nu(v))Q(v)| \le C(K)|\nu(u)-\nu(v)|Q(v),$$

which implies (2.31). Finally, writing $D\bar\eta = (\nabla\eta,0)$ and using (2.23) as well as (2.4) we have

$$\begin{aligned}
|\mathcal{E}(u;\eta_1,\eta_2) - \mathcal{E}(v;\eta_1,\eta_2)| &\le \int_\Omega |\langle P(\nu(v))D\bar\eta_1, D\bar\eta_2\rangle||Q(v)-Q(u)| \\
&\quad + \int_\Omega |\langle (P(\nu(v)) - P(\nu(u)))D\bar\eta_1, D\bar\eta_2\rangle|Q(u) \\
&\le \int_\Omega \langle P(\nu(v))D\bar\eta_1, D\bar\eta_1\rangle^{\frac{1}{2}}\langle P(\nu(v))D\bar\eta_2, D\bar\eta_2\rangle^{\frac{1}{2}}|\nu(u)-\nu(v)|Q(u)Q(v) \\
&\quad + C(K)\int_\Omega |\nu(u)-\nu(v)|\sqrt{Q(v)}\frac{|\nabla\eta_1|}{\sqrt{Q(v)}}|\nabla\eta_2| \\
&\le \epsilon\mathcal{E}(v;\eta_1,\eta_1) + \frac{C(K)}{\epsilon}\|\nabla\eta_2\|_\infty^2\int_\Omega |\nu(u)-\nu(v)|^2Q(v).
\end{aligned}$$

This concludes the proof of (2.32). □

Remark 2.7. We note that inequalities (2.30) and (2.32) were proved in [1] as Lemmas 4.7 and 4.5, respectively. The argument used above, employing the projection P, is more direct and slightly simpler than the one used in [1] in that it avoids the splitting of Ω into subsets.

LEMMA 2.8. *Let $u,v \in H^{1,\infty}(\Omega)$ with $|\nabla u| \le K$ a.e. in Ω. There exists a constant $c_1 > 0$ which depends only on K and γ_0 from* (2.9) *such that for*

$$D := \int_\Omega (\gamma(\nu(v)) - \langle\gamma'(\nu(u)),\nu(v)\rangle)Q(v)$$

we have

$$D \ge c_1\int_\Omega |\nu(u)-\nu(v)|^2Q(v).$$

Proof. This is just a reformulation of Lemma 3.2 in [9]. □

3. Discretization.

3.1. The finite element approximation.

We now turn to the discretization of (2.15), (2.16). Let $\mathcal{T}_h$ be a family of triangulations of Ω with maximum mesh size $h := \max_{\tau\in\mathcal{T}_h}\operatorname{diam}(\tau)$. We suppose that $\bar\Omega$ is the union of the elements of $\mathcal{T}_h$ so that element edges lying on the boundary are curved. Furthermore, we suppose that the triangulation is nondegenerate in the sense that $\max_{\tau\in\mathcal{T}_h}\frac{\operatorname{diam}(\tau)}{\rho_\tau} \le \kappa$, where the constant $\kappa > 0$ is independent of h and ρ_τ denotes the radius of the largest ball which is contained in $\bar{\tau}$. The discrete space is defined by

$$\mathcal{S}^h := \{v_h \in C^0(\bar{\Omega}) \mid v_h \text{ is a linear polynomial on each } \tau \in \mathcal{T}_h\}.$$

There exists an interpolation operator $\Pi^h : H^2(\Omega) \to \mathcal{S}^h$ such that

$$\|v - \Pi^h v\| + h\|\nabla(v - \Pi^h v)\| \le ch^2\|v\|_{H^2(\Omega)} \quad \forall\, v \in H^2(\Omega). \tag{3.1}$$

We are now in position to give a precise formulation of our numerical scheme. Let $\Delta t := \frac{T}{N}$ for an integer N and $t_m := m\Delta t$, $m = 0, \dots, N$. We denote by U^m, W^m the approximations to $u(\cdot, t_m)$ and $w(\cdot, t_m)$, respectively. Furthermore, let

$$\delta_t v^m := \frac{v^{m+1} - v^m}{\Delta t}.$$

In order to formulate a semi-implicit scheme requiring just the solution of linear equations we introduce the following form.

Stabilizing Anisotropic (SA) form,

$$\mathcal{B}(u; v, \eta) := \lambda \mathcal{B}_0(u; v, \eta) + \Delta t \mathcal{E}(u; v, \eta), \tag{3.2}$$

where

$$\mathcal{B}_0(u; v, \eta) := \int_\Omega \frac{\gamma(\nu(u))}{Q(u)} \nabla v \cdot \nabla w dx. \tag{3.3}$$

Remark 3.1. The purpose of the form $\mathcal{B}_0$ is to stabilize $\mathcal{A}$, which will be evaluated at the old time step. The second part in $\mathcal{B}$ is introduced in order to gain control on $\|U^m\|$ (see the proof of Lemma 3.4 below, in particular (3.14) and (3.15)) and the corresponding error in the convergence analysis.

Scheme 3.2. *We seek for each $m \in [1, N]$ a pair $\{U^m, W^m\} \in \mathcal{S}^h \times \mathcal{S}^h$ satisfying for $m \ge 0$*

$$(\delta_t U^m, \eta) + \mathcal{E}(U^m; W^{m+1}, \eta) = 0 \quad \forall\, \eta \in \mathcal{S}^h, \tag{3.4}$$

$$(W^{m+1}, \eta) - \mathcal{A}(U^m, \eta) - \Delta t \mathcal{B}(U^m; \delta_t U^m, \eta) = 0 \quad \forall\, \eta \in \mathcal{S}^h. \tag{3.5}$$

For simplicity we impose the initial condition,

$$U^0 := \Pi^h u_0. \tag{3.6}$$

The scheme does not require W^0. The constant λ is chosen to satisfy

$$\lambda \gamma_{\min} > \bar{\gamma}, \quad \text{where } \gamma_{\min} = \inf_{|p|=1} \gamma(p) > 0 \tag{3.7}$$

in order to ensure stability (see Lemma 3.4 below).

Lemma 3.3 (properties of the SA form $\mathcal{B}$). *Suppose that $u, v \in H^{1,\infty}(\Omega)$. Then*

$$\mathcal{B}(u; v, v) \le \left(\lambda \sup_{|p|=1} \gamma(p) + \Delta t\right) \mathcal{E}(u; v, v). \tag{3.8}$$

If in addition $|\nabla u| \le K$ a.e. in Ω, then

$$\begin{aligned} &|\mathcal{B}(u; \eta_1, \eta_2) - \mathcal{B}(v; \eta_1, \eta_2)| \\ &\quad \le C\|\nabla \eta_1\|_{L^\infty} \left(\int_\Omega |\nu(u) - \nu(v)||\nabla \eta_2| + \Delta t \int_\Omega |\nu(u) - \nu(v)||\nabla \eta_2| Q(v) \right). \end{aligned} \tag{3.9}$$

Proof. The inequality (3.8) follows immediately from (2.4). Next, if $|\nabla u| \le K$ a.e. in Ω, we deduce from (2.28) that

$$\begin{aligned}&\left|\frac{\gamma(\nu(u))}{Q(u)} - \frac{\gamma(\nu(v))}{Q(v)}\right| \\ &\quad \le \frac{1}{Q(u)}\left|\int_0^1 \langle\gamma'(s\nu(u)+(1-s)\nu(v))ds, \nu(u)-\nu(v)\rangle\right| + C\left|\frac{1}{Q(u)} - \frac{1}{Q(v)}\right| \\ &\quad \le C|\nu(u)-\nu(v)|.\end{aligned}$$

Combining this inequality with (2.31) implies (3.9). □

3.2. Stability.

Lemma 3.4. *Suppose that* (3.7) *holds. Then the unique discrete solution satisfies*

$$\max_{m\in[0,N]} \mathcal{I}_\gamma(U^m) + \Delta t\sum_{k=1}^N \mathcal{E}(U^{k-1}; W^k, W^k) \le C(\gamma, U^0), \tag{3.10}$$

$$\max_{m\in[0,N]} \|U^m\|^2 + \Delta t\sum_{k=1}^N \|W^k\|^2 \le C(\lambda,\gamma,U^0,T). \tag{3.11}$$

Proof. Taking $\eta = \Delta t W^{m+1}$ in (3.4), $\eta = \Delta t\delta_t U^m$ in (3.5) and adding yields

$$\begin{aligned}&\Delta t\mathcal{E}(U^m; W^{m+1}, W^{m+1}) + \mathcal{A}(U^m, U^{m+1}-U^m) \\ &\qquad + (\Delta t)^2\mathcal{B}(U^m; \delta_t U^m, \delta_t U^m) = 0.\end{aligned} \tag{3.12}$$

Lemma 2.5 implies

$$\begin{aligned}\mathcal{A}(U^m, U^{m+1}-U^m) &\ge I_\gamma(U^{m+1}) - I_\gamma(U^m) - \bar\gamma\int_\Omega |\nu(U^{m+1})-\nu(U^m)|^2 Q(U^{m+1}) \\ &\ge I_\gamma(U^{m+1}) - I_\gamma(U^m) - (\Delta t)^2\frac{\bar\gamma}{\gamma_{\min}}\mathcal{B}_0(U^m; \delta_t U^m, \delta_t U^m),\end{aligned}$$

where we have used (2.21). Inserting the above inequality into (3.12) and recalling the definition of $\mathcal{B}$ we infer

$$\begin{aligned}&\mathcal{I}_\gamma(U^{m+1}) - \mathcal{I}_\gamma(U^m) + \Delta t\mathcal{E}(U^m; W^{m+1}, W^{m+1}) \\ &\quad + \left(\lambda - \frac{\bar\gamma}{\gamma_{\min}}\right)(\Delta t)^2\mathcal{B}_0(U^m; \delta_t U^m, \delta_t U^m) + (\Delta t)^3\mathcal{E}(U^m, \delta_t U^m, \delta_t U^m) \le 0.\end{aligned} \tag{3.13}$$

Summation over m yields (3.10) as well as

$$\begin{aligned}&(\Delta t)^2\sum_{m=0}^{N-1}\mathcal{B}_0(U^m; \delta_t U^m, \delta_t U^m) \\ &\qquad + (\Delta t)^3\sum_{m=0}^{N-1}\mathcal{E}(U^m; \delta_t U^m, \delta_t U^m) \le C(\lambda,\gamma,U^0).\end{aligned} \tag{3.14}$$

Next, using $\eta = \Delta t U^{m+1}$ in (3.4) we deduce

$$\begin{aligned}
&\frac{1}{2}\|U^{m+1}\|^2 - \frac{1}{2}\|U^m\|^2 + \frac{1}{2}\|U^{m+1} - U^m\|^2 = \Delta t \mathcal{E}(U^m; W^{m+1}, U^{m+1})\\
&\quad \le \Delta t \mathcal{E}(U^m; W^{m+1}, W^{m+1})^{\frac{1}{2}} \mathcal{E}(U^m; U^{m+1}, U^{m+1})^{\frac{1}{2}}\\
&\quad \le \Delta t \mathcal{E}(U^m; W^{m+1}, W^{m+1})^{\frac{1}{2}} \left(\mathcal{E}(U^m; U^m, U^m)^{\frac{1}{2}} + \Delta t \mathcal{E}(U^m; \delta_t U^m, \delta_t U^m)^{\frac{1}{2}}\right)\\
&\quad \le \Delta t \mathcal{E}(U^m; W^{m+1}, W^{m+1}) + \Delta t \int_\Omega Q(U^m) + (\Delta t)^3 \mathcal{E}(U^m; \delta_t U^m, \delta_t U^m).
\end{aligned} \tag{3.15}$$

Finally, using $\eta = \Delta t W^{m+1}$ in (3.5) we obtain with the help of (2.4) and (3.8) that

$$\begin{aligned}
\Delta t \|W^{m+1}\|^2 &= \Delta t \mathcal{A}(U^m, W^{m+1}) + (\Delta t)^2 \mathcal{B}(U^m; \delta_t U^m, W^{m+1})\\
&\le \Delta t \sup_{|p|=1} |\gamma'(p)| \left(\int_\Omega \frac{|\nabla W^{m+1}|^2}{Q(U^m)}\right)^{\frac{1}{2}} \left(\int_\Omega Q(U^m)\right)^{\frac{1}{2}} \qquad (3.16)\\
&\quad + (\Delta t)^2 \mathcal{B}(U^m; \delta_t U^m, \delta_t U^m)^{\frac{1}{2}} \mathcal{B}(U^m; W^{m+1}, W^{m+1})^{\frac{1}{2}}\\
&\le \Delta t \mathcal{E}(U^m; W^{m+1}, W^{m+1}) + C(\gamma) \Delta t \int_\Omega Q(U^m) \qquad (3.17)\\
&\quad + C(\Delta t)^2 \mathcal{B}(U^m; \delta_t U^m, \delta_t U^m).
\end{aligned}$$

Now (3.11) follows from summing (3.15), (3.16) over m, the inequality $\int_\Omega Q(U^m) \le C(\gamma) I_\gamma(U^m)$, and (3.10), (3.14). □

Remark 3.5. It follows in particular that

$$\max_{m \in [0,N]} \int_\Omega Q(U^m) \le C(\gamma, U^0). \tag{3.18}$$

3.3. Boundary conditions, domain perturbation, and quadrature.

For Neumann boundary conditions it is sufficient for the union of the elements to contain Ω, provided exact quadrature is used. The above analysis can be easily extended to higher order elements. On the other hand, when using piecewise linear elements it is convenient to use a quadrature rule based on mass lumping for the L^2 inner products. The other integrals require just the measure of the regions of integration. In the case of Dirichlet boundary conditions it is necessary either to analyze the effect of domain perturbation in the case of linear finite elements with a polygonal interpolation of Ω or to analyze isoparametric approximations for higher order elements.

4. Error bounds.

We set

$$u^m := u(\cdot, t_m), \qquad w^m := w(\cdot, t_m), \qquad S^m := \delta_t u^m - \partial_t u(\cdot, t_{m+1}).$$

Then we have for the continuous problem the analogue of the discrete scheme,

$$(\delta_t u^m, \eta) + \mathcal{E}(u^{m+1}; w^{m+1}, \eta) = (S^m, \eta) \quad \forall \eta \in H^1(\Omega), \tag{4.1}$$

$$(w^m, \eta) - \mathcal{A}(u^m, \eta) = 0 \quad \forall \eta \in H^1(\Omega). \tag{4.2}$$

It is convenient to introduce the errors

$$e_u^m := u^m - U^m =: \rho_u^m + \theta_u^m, \qquad e_w^m := w^m - W^m := \rho_w^m + \theta_w^m,$$

where

$$\rho_u^m := u^m - \Pi^h u^m, \qquad \rho_w^m := w^m - \Pi^h w^m$$

are the interpolation errors. It is our goal to prove the following error bounds.

Theorem 4.1. *Let (u,w) solve (2.10)–(2.14) and satisfy the regularity $u \in H^{1,\infty}(0,T;H^{2,\infty}(\Omega))$, $u_{tt} \in L^\infty(0,T;H^{1,\infty}(\Omega))$, $w \in H^{1,\infty}(0,T;H^{2,\infty}(\Omega))$, $w_{tt} \in L^2(0,T;L^2(\Omega))$. Suppose also that (3.7) holds. Then there exists $\delta > 0$ such that for $0 < \Delta t \le \delta$*

$$\max_{m\in[0,N]} \left(\|e_u^m\|^2 + \int_\Omega |\nu(u^m) - \nu(U^m)|^2 Q(U^m) \right) \\ + \Delta t \sum_{k=1}^{N} (\|e_w^k\|^2 + \mathcal{E}(U^{k-1}; e_w^k, e_w^k)) \le C(h^2 + (\Delta t)^2),$$

where C and δ depend on γ, Ω, T, λ and the solution u.

The rest of this section will be devoted to the proof of Theorem 4.1. Subtracting (3.4), (3.5) and (4.1), (4.2) yields, for all $\eta \in \mathcal{S}^h$, the error equations

$$(\delta_t e_u^m, \eta) + \mathcal{E}(u^{m+1}; w^{m+1}, \eta) - \mathcal{E}(U^m; W^{m+1}, \eta) = (S^m, \eta), \tag{4.3}$$

$$(e_w^{m+1}, \eta) - \mathcal{A}(u^m, \eta) + \mathcal{A}(U^m, \eta) + \Delta t \mathcal{B}(U^m; \delta_t U^m, \eta) = (w^{m+1} - w^m, \eta). \tag{4.4}$$

4.1. An a priori estimate in the energy norm. The first step is to emulate the energy bounds obtained for the continuous and discrete solutions by testing (4.3) and (4.4) with $e_w^{m+1} - \rho_w^{m+1} \in \mathcal{S}^h$ and $\delta_t e_u^m - \delta_t \rho_u^m \in \mathcal{S}^h$ yielding

$$\begin{aligned}(\delta_t e_u^m, e_w^{m+1}) &+ \mathcal{E}(u^{m+1}; w^{m+1}, e_w^{m+1}) - \mathcal{E}(U^m; W^{m+1}, e_w^{m+1}) \\ &= (\delta_t e_u^m, \rho_w^{m+1}) + \mathcal{E}(u^{m+1}; w^{m+1}, \rho_w^{m+1}) \\ &\quad - \mathcal{E}(U^m; W^{m+1}, \rho_w^{m+1}) + (S^m, e_w^{m+1} - \rho_w^{m+1}),\end{aligned} \tag{4.5}$$

$$\begin{aligned}(e_w^{m+1}, \delta_t e_u^m) &- \mathcal{A}(u^m, \delta_t e_u^m) + \mathcal{A}(U^m, \delta_t e_u^m) + \Delta t \mathcal{B}(U^m; \delta_t U^m, \delta_t e_u^m) \\ &= (e_w^{m+1}, \delta_t \rho_u^m) - \mathcal{A}(u^m, \delta_t \rho_u^m) + \mathcal{A}(U^m, \delta_t \rho_u^m) + \Delta t \mathcal{B}(U^m; \delta_t U^m, \delta_t \rho_u^m) \\ &\quad + \Delta t(\delta_t w^m, \delta_t e_u^m - \delta_t \rho_u^m).\end{aligned} \tag{4.6}$$

Combining these equations and multiplying by Δt yields

$$\begin{aligned}\Delta t(\mathcal{A}(u^m, \delta_t e_u^m) &- \mathcal{A}(U^m, \delta_t e_u^m)) \\ &+ \Delta t(\mathcal{E}(u^{m+1}; w^{m+1}, e_w^{m+1}) - \mathcal{E}(U^m; W^{m+1}, e_w^{m+1})) \\ &+ (\Delta t)^2 \mathcal{B}(U^m; \delta_t e_u^m, \delta_t e_u^m) = \Delta t(\mathcal{A}(u^m, \delta_t \rho_u^m) - \mathcal{A}(U^m, \delta_t \rho_u^m)) \\ &+ \Delta t(\mathcal{E}(u^{m+1}; w^{m+1}, \rho_w^{m+1}) - \mathcal{E}(U^m; W^{m+1}, \rho_w^{m+1})) \\ &+ \Delta t(S^m, e_w^{m+1} - \rho_w^{m+1}) - \Delta t(e_w^{m+1}, \delta_t \rho_u^m) \\ &+ \Delta t(\delta_t e_u^m, \rho_w^{m+1}) - (\Delta t)^2(\delta_t w^m, \delta_t e_u^m - \delta_t \rho_u^m) \\ &+ (\Delta t)^2(\mathcal{B}(U^m; \delta_t u^m, \delta_t e_u^m) - \mathcal{B}(U^m; \delta_t U^m, \delta_t \rho_u^m)) := \sum_{j=1}^{7} R_j^m.\end{aligned} \tag{4.7}$$

The proof of the error bounds is based on estimating the terms on both sides of the above equation. We begin with the left-hand side of (4.7) which we denote by L^m. First we recall the following lemma.

Lemma 4.2. *Let*

$$\mathcal{D}^m := \int_\Omega (\gamma(\nu(U^m)) - \langle \gamma'(\nu(u^m)), \nu(U^m) \rangle) Q(U^m).$$

Then we have for $m \in [0, N-1]$ *and small* Δt

$$\begin{aligned}\Delta t(\mathcal{A}(u^m, \delta_t e_u^m) - \mathcal{A}(U^m, \delta_t e_u^m)) &\geq \mathcal{D}^{m+1} - \mathcal{D}^m \\ &- (\bar{\gamma} + C\Delta t)\int_\Omega \frac{|\nabla(e_u^{m+1} - e_u^m)|^2}{Q(U^m)} \\ &- C\Delta t\left((\Delta t)^2 + \int_\Omega |\nu(u^{m+1}) - \nu(U^{m+1})|^2 Q(U^{m+1})\right).\end{aligned}$$

Proof. See [11, Lemma 4.2]. □

Lemma 4.2 and the definition of $\mathcal{B}_0$ now imply

$$\begin{aligned}&\Delta t(\mathcal{A}(u^m, \delta_t e_u^m) - \mathcal{A}(U^m, \delta_t e_u^m)) \\ &\quad\geq \mathcal{D}^{m+1} - \mathcal{D}^m - (\Delta t)^2\left(\frac{\bar{\gamma}}{\gamma_{\min}} + C\Delta t\right)\mathcal{B}_0(U^m; \delta_t e_u^m, \delta_t e_u^m) \\ &\qquad - C\Delta t\left((\Delta t)^2 + \int_\Omega |\nu(u^{m+1}) - \nu(U^{m+1})|^2 Q(U^{m+1})\right).\end{aligned} \tag{4.8}$$

Next we examine

$$\begin{aligned}&\Delta t(\mathcal{E}(u^{m+1}; w^{m+1}, e_w^{m+1}) - \mathcal{E}(U^m; W^{m+1}, e_w^{m+1})) \\ &\quad= \Delta t\mathcal{E}(U^m; e_w^{m+1}, e_w^{m+1}) + \Delta t(\mathcal{E}(u^{m+1}; w^{m+1}, e_w^{m+1}) - \mathcal{E}(u^m; w^{m+1}, e_w^{m+1})) \\ &\qquad + \Delta t(\mathcal{E}(u^m; w^{m+1}, e_w^{m+1}) - \mathcal{E}(U^m; w^{m+1}, e_w^{m+1})) \\ &\quad=: \alpha_1^m + \alpha_2^m + \alpha_3^m.\end{aligned}$$

We infer from (3.18) and (2.4) that

$$\begin{aligned}|\alpha_2^m| &\leq C(\Delta t)^2\|\nabla w^{m+1}\|_{L^\infty}\int_\Omega |\nabla e_w^{m+1}| \leq C(\Delta t)^2\left(\int_\Omega Q(U^m)\right)^{\frac{1}{2}}\left(\int_\Omega \frac{|\nabla e_w^{m+1}|^2}{Q(U^m)}\right)^{\frac{1}{2}} \\ &\leq \epsilon\Delta t\mathcal{E}(U^m, e_w^{m+1}, e_w^{m+1}) + \frac{C}{\epsilon}(\Delta t)^3.\end{aligned}$$

Furthermore, (2.32) yields

$$|\alpha_3^m| \leq \epsilon\Delta t\mathcal{E}(U^m, e_w^{m+1}, e_w^{m+1}) + \frac{C}{\epsilon}\|\nabla w^{m+1}\|_{L^\infty}^2 \int_\Omega |\nu(u^m) - \nu(U^m)|^2 Q(U^m).$$

Combining (4.8) and the estimates for α_2^m, α_3^m we derive

$$\begin{aligned}L^m &\geq \mathcal{D}^{m+1} - \mathcal{D}^m + (1 - 2\epsilon)\Delta t\mathcal{E}(U^m; e_w^{m+1}, e_w^{m+1}) \\ &\quad + (\Delta t)^2\left(\lambda - \frac{\bar{\gamma}}{\gamma_{\min}} - C\Delta t\right)\mathcal{B}_0(U^m; \delta_t e_u^m, \delta_t e_u^m) + (\Delta t)^3\mathcal{E}(U^m; \delta_t e_u^m, \delta_t e_u^m) \\ &\quad - \frac{C}{\epsilon}\Delta t\Bigg((\Delta t)^2 + \int_\Omega |\nu(u^m) - \nu(U^m)|^2 Q(U^m) \\ &\qquad\qquad + \int_\Omega |\nu(u^{m+1}) - \nu(U^{m+1})|^2 Q(U^{m+1})\Bigg).\end{aligned} \tag{4.9}$$

4.2. L^2-estimates.

In order to proceed and estimate the terms R_j^m on the right-hand side of (4.7), we need to derive bounds on the L^2-norms of e_w^{k+1} and e_u^{k+1}.

LEMMA 4.3. *We have for* $m \in [0, N-1]$

$$\|e_w^{m+1}\|^2 \le \mathcal{E}(U^m; e_w^{m+1}, e_w^{m+1}) + C(\Delta t)^2 \mathcal{B}(U^m; \delta_t e_u^m, \delta_t e_u^m) \\ + C \int_\Omega |\nu(u^m) - \nu(U^m)|^2 Q(U^m) + C(h^2 + (\Delta t)^2).$$

Proof. Inserting $\eta = e_w^{m+1} - \rho_w^{m+1}$ into (4.4) and using (2.26) we infer

$$\begin{aligned} &\|e_w^{m+1}\|^2 \\ &\quad = (e_w^{m+1}, e_w^{m+1} - \rho_w^{m+1}) + (e_w^{m+1}, \rho_w^{m+1}) \\ &\quad = \mathcal{A}(u^m, e_w^{m+1} - \rho_w^{m+1}) - \mathcal{A}(U^m, e_w^{m+1} - \rho_w^{m+1}) \\ &\qquad -\Delta t \mathcal{B}(U^m; \delta_t U^m, e_w^{m+1} - \rho_w^{m+1}) + \Delta t(\delta_t w^m, e_w^{m+1} - \rho_w^{m+1}) \\ &\qquad + (e_w^{m+1}, \rho_w^{m+1}) \le C \int_\Omega |\nu(u^m) - \nu(U^m)|(|\nabla e_w^{m+1}| + |\nabla \rho_w^{m+1}|) \\ &\qquad + \Delta t |\mathcal{B}(U^m; \delta_t U^m, e_w^{m+1} - \rho_w^{m+1})| + C\Delta t(\|e_w^{m+1}\| + \|\rho_w^{m+1}\|) \\ &\qquad + \|e_w^{m+1}\| \|\rho_w^{m+1}\| \le \frac{1}{2}\|e_w^{m+1}\|^2 + \Delta t |\mathcal{B}(U^m; \delta_t U^m, e_w^{m+1} - \rho_w^{m+1})| \\ &\qquad + C((\Delta t)^2 + h^2) + \frac{1}{4}\int_\Omega \frac{|\nabla e_w^{m+1}|^2}{Q(U^m)} + C\int_\Omega |\nu(u^m) - \nu(U^m)|^2 Q(U^m). \end{aligned}$$

It remains to bound the term involving $\mathcal{B}$. Clearly,

$$\begin{aligned} &|\mathcal{B}(U^m; \delta_t U^m, e_w^{m+1})| \\ &\quad \le \mathcal{B}(U^m; \delta_t U^m, \delta_t U^m)^{\frac{1}{2}} \mathcal{B}(U^m; e_w^{m+1}, e_w^{m+1})^{\frac{1}{2}} \\ &\quad \le \left(\mathcal{B}(U^m; \delta_t u^m, \delta_t u^m)^{\frac{1}{2}} + \mathcal{B}(U^m; \delta_t e_u^m, \delta_t e_u^m)^{\frac{1}{2}}\right) \mathcal{B}(U^m; e_w^{m+1}, e_w^{m+1})^{\frac{1}{2}} \\ &\quad \le C\left(\left(\int_\Omega Q(U^m)\right)^{\frac{1}{2}} + \mathcal{B}(U^m; \delta_t e_u^m, \delta_t e_u^m)^{\frac{1}{2}}\right) \mathcal{E}(U^m; e_w^{m+1}, e_w^{m+1})^{\frac{1}{2}}, \end{aligned}$$

by (3.8). Recalling (3.18) we deduce

$$\begin{aligned} &\Delta t |\mathcal{B}(U^m; \delta_t U^m, e_w^{m+1})| \\ &\qquad \le \frac{1}{4}\mathcal{E}(U^m; e_w^{m+1}, e_w^{m+1}) + C((\Delta t)^2 + (\Delta t)^2 \mathcal{B}(U^m; \delta_t e_u^m, \delta_t e_u^m)). \end{aligned}$$

Similarly,

$$\Delta t |\mathcal{B}(U^m; \delta_t U^m, \rho_w^{m+1})| \le C((\Delta t)^2 + h^2) + C(\Delta t)^2 \mathcal{B}(U^m; \delta_t e_u^m, \delta_t e_u^m).$$

If we insert these inequalities into the estimate for $\|e_w^{m+1}\|$ and use (2.4) we arrive at the desired bound. □

LEMMA 4.4. *We have for* $0 \le m \le N$

$$\max_{k \in [0,m]} \|e_u^k\|^2 \le C\left(\Delta t \sum_{k=0}^{m-1} \mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1}) + (\Delta t)^3 \sum_{k=0}^{m-1} \mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k)\right) \\ + C((\Delta t)^2 + h^2) + C\Delta t \sum_{k=0}^{m-1} \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k).$$

Proof. Clearly,

$$\begin{aligned}
&\frac{1}{2}\|e_u^{k+1}\|^2 - \frac{1}{2}\|e_u^k\|^2 + \frac{1}{2}\|e_u^{k+1} - e_u^k\|^2 \\
&\qquad = \Delta t(\delta_t e_u^k, e_u^{k+1}) = \Delta t(\delta_t e_u^k, \theta_u^{k+1}) + \Delta t(\delta_t e_u^k, \rho_u^{k+1}) \\
&\qquad = \Delta t(\mathcal{E}(U^k; W^{k+1}, \theta_u^{k+1}) - \mathcal{E}(u^{k+1}; w^{k+1}, \theta_u^{k+1})) \\
&\qquad\quad + \Delta t(S^k, \theta_u^{k+1}) + \Delta t(\delta_t e_u^k, \rho_u^{k+1}),
\end{aligned} \tag{4.10}$$

where the last inequality follows from (4.3) with the choice $\eta = \Delta t \theta_u^{k+1}$. To begin,

$$\begin{aligned}
&|\mathcal{E}(U^k; W^{k+1}, \theta_u^{k+1}) - \mathcal{E}(u^{k+1}; w^{k+1}, \theta_u^{k+1})| \\
&\quad \le |\mathcal{E}(U^k; e_w^{k+1}, \theta_u^{k+1})| + |\mathcal{E}(U^k; w^{k+1}, \theta_u^{k+1}) - \mathcal{E}(u^k; w^{k+1}, \theta_u^{k+1})| \\
&\qquad + |\mathcal{E}(u^k; w^{k+1}, \theta_u^{k+1}) - \mathcal{E}(u^{k+1}; w^{k+1}, \theta_u^{k+1})| \\
&\quad = I + II + III.
\end{aligned}$$

Before we estimate these terms we first note that (2.30) and (3.18) imply

$$\begin{aligned}
&\mathcal{E}(U^k; \theta_u^{k+1}, \theta_u^{k+1}) \\
&\quad \le 2\mathcal{E}(U^k; e_u^{k+1}, e_u^{k+1}) + 2\mathcal{E}(U^k; \rho_u^{k+1}, \rho_u^{k+1}) \\
&\quad \le 4\mathcal{E}(U^k; e_u^k, e_u^k) + 4(\Delta t)^2\mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k) + C\|\nabla \rho_u^{k+1}\|_{L^\infty}^2 \int_\Omega Q(U^k) \\
&\quad \le C\int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k) + 4(\Delta t)^2\mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k) + Ch^2.
\end{aligned} \tag{4.11}$$

We then infer from (2.29) and (4.11)

$$\begin{aligned}
I &\le \mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1})^{\frac{1}{2}} \mathcal{E}(U^k; \theta_u^{k+1}, \theta_u^{k+1})^{\frac{1}{2}} \le \mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1}) \\
&\quad + C\biggl((\Delta t)^2\mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k) + h^2 + \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k)\biggr).
\end{aligned}$$

Next, (2.32) together with (4.11) implies

$$\begin{aligned}
II &\le \mathcal{E}(U^k; \theta_u^{k+1}, \theta_u^{k+1}) + C\|\nabla w^{k+1}\|_{L^\infty}^2 \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k) \\
&\le C\left((\Delta t)^2\mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k) + h^2 + \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k)\right),
\end{aligned}$$

as well as

$$\begin{aligned}
III &\le \mathcal{E}(U^k; \theta_u^{k+1}, \theta_u^{k+1}) + C\|\nabla w^{k+1}\|_{L^\infty}^2 \int_\Omega |\nu(u^{k+1}) - \nu(u^k)|^2 Q(u^k) \\
&\le C\left((\Delta t)^2\mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k) + (\Delta t)^2 + h^2 + \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k)\right).
\end{aligned}$$

Collecting the above estimates we derive

$$\begin{aligned}
&\Delta t|\mathcal{E}(U^k; W^{k+1}, \theta_u^{k+1}) - \mathcal{E}(u^{k+1}; w^{k+1}, \theta_u^{k+1})| \le \Delta t\mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1}) \\
&\quad + C\Delta t\left((\Delta t)^2\mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k) + (\Delta t)^2 + h^2 + \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k)\right).
\end{aligned}$$

Next,

$$\Delta t|(S^k,\theta_u^{k+1})| \le C(\Delta t)^2(\|e_u^k\| + \|e_u^{k+1}-e_u^k\| + \|\rho_u^{k+1}\|)$$
$$\le \frac{1}{4}\|e_u^{k+1}-e_u^k\|^2 + C\Delta t\|e_u^k\|^2 + C\Delta t((\Delta t)^2+h^4).$$

If we insert the above estimates into (4.10), sum from $k=0$ to $m-1$, and rearrange terms, we obtain

$$\begin{aligned}\frac{1}{2}\|e_u^m\|^2 \le{}& \frac{1}{2}\|e_u^0\|^2 + \Delta t\sum_{k=0}^{m-1}\mathcal{E}(U^k;e_w^{k+1},e_w^{k+1}) + C(\Delta t)^3\sum_{k=0}^{m-1}\mathcal{E}(U^k;\delta_t e_u^k,\delta_t e_u^k)\\ &+ C((\Delta t)^2+h^2) + C\Delta t\sum_{k=0}^{m-1}\int_\Omega |\nu(u^k)-\nu(U^k)|^2 Q(U^k)\\ &+ C\Delta t\sum_{k=0}^{m-1}\|e_u^k\|^2 + \Delta t\sum_{k=0}^{m-1}(\delta_t e_u^k,\rho_u^{k+1}).\end{aligned}$$

Integrating by parts discretely in time we infer

$$\left|\Delta t\sum_{k=0}^{m-1}(\delta_t e_u^k,\rho_u^{k+1})\right| = \left|-\Delta t\sum_{k=0}^{m-1}(e_u^k,\delta_t\rho_u^k) + (e_u^m,\rho_u^m) - (e_u^0,\rho_u^0)\right| \le Ch^2,$$

since $\max_{k\in[0,N]}\|e_u^k\|^2 \le C$ by Lemma 3.4. Thus,

$$\begin{aligned}\|e_u^m\|^2 \le{}& 2\Delta t\sum_{k=0}^{m-1}\mathcal{E}(U^k;e_w^{k+1},e_w^{k+1}) + C(\Delta t)^3\sum_{k=0}^{m-1}\mathcal{E}(U^k;\delta_t e_u^k,\delta_t e_u^k) + C((\Delta t)^2+h^2)\\ &+ C\Delta t\sum_{k=0}^{m-1}\int_\Omega|\nu(u^k)-\nu(U^k)|^2Q(U^k) + C\Delta t\sum_{k=0}^{m-1}\|e_u^k\|^2.\end{aligned}$$

The result now follows with the help of a discrete Gronwall argument. □

4.3. Estimating the right-hand side of (4.7). Invoking (2.26) we obtain

$$\begin{aligned}|R_1^k| &= \Delta t|\mathcal{A}(u^k,\delta_t\rho_u^k) - \mathcal{A}(U^k,\delta_t\rho_u^k)|\\ &\le C\Delta t\int_\Omega|\nu(u^k)-\nu(U^k)||\nabla\delta_t\rho_u^k| \le C\Delta t h\Big(\int_\Omega|\nu(u^k)-\nu(U^k)|^2\Big)^{\frac{1}{2}}\\ &\le C\Delta t h^2 + C\Delta t\int_\Omega|\nu(u^k)-\nu(U^k)|^2Q(U^k).\end{aligned}\tag{4.12}$$

Lemma 2.6 and (3.18) imply

$$
\begin{aligned}
|R_2^k| &\le \Delta t|\mathcal{E}(u^{k+1}; w^{k+1}, \rho_w^{k+1}) - \mathcal{E}(u^k; w^{k+1}, \rho_w^{k+1})| \\
&\quad + \Delta t|\mathcal{E}(u^k; w^{k+1}, \rho_w^{k+1}) - \mathcal{E}(U^k; w^{k+1}, \rho_w^{k+1})| \\
&\quad + \Delta t|\mathcal{E}(U^k; e_w^{k+1}, \rho_w^{k+1})| \le C\Delta t\|\nabla w^{k+1}\|_{L^\infty} \\
&\quad \times\bigg(\int_\Omega |\nu(u^{k+1}) - \nu(u^k)||\nabla \rho_w^{k+1}|Q(u^k) \\
&\qquad + \int_\Omega |\nu(u^k) - \nu(U^k)||\nabla \rho_w^{k+1}|Q(U^k)\bigg) \\
&\quad + \Delta t\mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1})^{\frac{1}{2}}\mathcal{E}(U^k; \rho_w^{k+1}, \rho_w^{k+1})^{\frac{1}{2}} \\
&\le \epsilon\Delta t\mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1}) + \frac{C}{\epsilon}\Delta t((\Delta t)^2 + h^2) \\
&\quad + C\Delta t\int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k).
\end{aligned}
\tag{4.13}
$$

Next, Lemma 4.3 gives

$$
\begin{aligned}
|R_3^k + R_4^k| &\le C(\Delta t)^2(\|e_w^{k+1}\| + \|\rho_w^{k+1}\|) + \Delta t\|e_w^{k+1}\|\|\delta_t \rho_u^k\| \\
&\le \epsilon\Delta t\|e_w^{k+1}\|^2 + \frac{C}{\epsilon}\Delta t((\Delta t)^2 + h^4) \\
&\le \epsilon\Delta t\mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1}) + \epsilon(\Delta t)^3\mathcal{B}(U^k; \delta_t e_u^k, \delta_t e_u^k) \\
&\quad + \epsilon\Delta t\int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k) + \frac{C}{\epsilon}\Delta t((\Delta t)^2 + h^2).
\end{aligned}
\tag{4.14}
$$

Integrating by parts discretely in time yields

$$
\begin{aligned}
\left|\sum_{k=0}^{m-1} R_5^k\right| &= \left|-\Delta t\sum_{k=0}^{m-1}(e_u^k, \delta_t \rho_w^k) + (e_u^m, \rho_w^m) - (e_u^0, \rho_w^0)\right| \\
&\le \epsilon \max_{k\in[0,m]}\|e_u^k\|^2 + \frac{C}{\epsilon}h^4.
\end{aligned}
\tag{4.15}
$$

Similarly,

$$
\begin{aligned}
\left|\sum_{k=0}^{m-1} R_6^k\right| &\le \left|-(\Delta t)^2\sum_{k=0}^{m-1}(\delta_t w^k, \delta_t e_u^k)\right| + (\Delta t)^2\left|\sum_{k=0}^{m-1}(\delta_t w^k, \delta_t \rho_u^k)\right| \\
&\le (\Delta t)^2\Bigg|\sum_{k=1}^{m-1}\left(\frac{w^{k+1} - 2w^k + w^{k-1}}{(\Delta t)^2}, e_u^k\right) \\
&\qquad - \Delta t(\delta_t w^{m-1}, e_u^m) + \Delta t(\delta_t w^0, e_u^0)\Bigg| + Ch^2\Delta t \\
&\le C\Delta t\max_{k\in[0,m]}\|e_u^k\| + Ch^2\Delta t \le \epsilon\max_{k\in[0,m]}\|e_u^k\|^2 + \frac{C}{\epsilon}((\Delta t)^2 + h^4).
\end{aligned}
\tag{4.16}
$$

Finally, let us write

$$
\begin{aligned}
R_7^k &= (\Delta t)^2(\mathcal{B}(U^k; \delta_t u^k, \delta_t e_u^k) - \mathcal{B}(U^k; \delta_t u^k, \delta_t \rho_u^k) + \mathcal{B}(U^k; \delta_t e_u^k, \delta_t \rho_u^k)) \\
&= I + II + III.
\end{aligned}
$$

In view of the definition of $\mathcal{B}$ we have

$$\begin{aligned} I &= (\Delta t)^2 \lambda \int_\Omega \frac{\gamma(\nu(u^k))}{Q(u^k)} \nabla \delta_t u^k \cdot \nabla \delta_t e_u^k \\ &\quad + (\Delta t)^2 \lambda \int_\Omega \left(\frac{\gamma(\nu(U^k))}{Q(U^k)} - \frac{\gamma(\nu(u^k))}{Q(u^k)} \right) \nabla \delta_t u^k \cdot \nabla \delta_t e_u^k \\ &\quad + (\Delta t)^3 \mathcal{E}(U^k; \delta_t u^k, \delta_t e_u^k) = (\Delta t)^2 \lambda (G^k, \nabla \delta_t e_u^k) + I_2 + I_3, \end{aligned}$$

where we have written $G^k := \frac{\gamma(\nu(u^k))}{Q(u^k)} \nabla \delta_t u^k$. We infer from (2.28) and (2.4) that

$$\begin{aligned} |I_2| &\le C(\Delta t)^2 \int_\Omega |\nu(u^k) - \nu(U^k)| |\nabla \delta_t e_u^k| \\ &\le \epsilon (\Delta t)^3 \int_\Omega \frac{|\nabla \delta_t e_u^k|^2}{Q(U^k)} + \frac{C}{\epsilon} \Delta t \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k) \\ &\le \epsilon (\Delta t)^3 \mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k) + \frac{C}{\epsilon} \Delta t \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k). \end{aligned}$$

Furthermore, (2.29) and (3.18) yield

$$|I_3| \le \epsilon (\Delta t)^3 \mathcal{E}(U^k; \delta_t e_u^k, \delta_t e_u^k) + \frac{C}{\epsilon} (\Delta t)^3.$$

Observing that $\mathcal{B}(U^k; \delta_t \rho_u^k, \delta_t \rho_u^k) \le Ch^2$ we finally have

$$\begin{aligned} |II| &\le (\Delta t)^2 \mathcal{B}(U^k; \delta_t u^k, \delta_t u^k)^{\frac{1}{2}} \mathcal{B}(U^k; \delta_t \rho_u^k, \delta_t \rho_u^k)^{\frac{1}{2}} \le C \Delta t ((\Delta t)^2 + h^2), \\ |III| &\le (\Delta t)^2 \mathcal{B}(U^k; \delta_t e_u^k, \delta_t e_u^k)^{\frac{1}{2}} \mathcal{B}(U^k; \delta_t \rho_u^k, \delta_t \rho_u^k)^{\frac{1}{2}} \\ &\le \epsilon (\Delta t)^2 \mathcal{B}(U^k; \delta_t e_u^k, \delta_t e_u^k) + \frac{C}{\epsilon} \Delta t ((\Delta t)^2 + h^2). \end{aligned}$$

Summing the above estimates, integrating the first term in I by parts in time, and taking into account the estimate (which follows from (2.21) and (2.23))

$$\|\nabla e_u^k\|_{L^1} \le \left(\int_\Omega \frac{|\nabla e_u^k|^2}{Q(U^k)} \right)^{\frac{1}{2}} \left(\int_\Omega Q(U^k) \right)^{\frac{1}{2}} \le C \left(\int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k) \right)^{\frac{1}{2}},$$

we derive

$$\begin{aligned} \left| \sum_{k=0}^{m-1} R_7^k \right| &\le \lambda \left| -(\Delta t)^2 \sum_{k=0}^{m-1} (\delta_t G^k, \nabla e_u^k) + \Delta t (G^m, \nabla e_u^m) - \Delta t (G^0, \nabla e_u^0) \right| \\ &\quad + \frac{C}{\epsilon} ((\Delta t)^2 + h^2) + \epsilon (\Delta t)^2 \sum_{k=0}^{m-1} \mathcal{B}(U^k, \delta_t e_u^k, \delta_t e_u^k) \\ &\quad + \frac{C}{\epsilon} \Delta t \sum_{k=0}^{m-1} \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k) \\ &\le \epsilon \int_\Omega |\nu(u^m) - \nu(U^m)|^2 Q(U^m) + \epsilon (\Delta t)^2 \sum_{k=0}^{m-1} \mathcal{B}(U^k, \delta_t e_u^k, \delta_t e_u^k) \\ &\quad + \frac{C}{\epsilon} ((\Delta t)^2 + h^2) + \frac{C}{\epsilon} \Delta t \sum_{k=0}^{m-1} \int_\Omega |\nu(u^k) - \nu(U^k)|^2 Q(U^k). \end{aligned} \tag{4.17}$$

Collecting (4.12)–(4.17) and recalling Lemma 2.8 finally yields

$$\begin{aligned}\left|\sum_{k=0}^{m-1}\sum_{j=1}^{7} R_j^k\right| &\le \epsilon\Delta t\sum_{k=0}^{m-1}\mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1})\\ &\quad + \epsilon(\Delta t)^2\sum_{k=0}^{m-1}\mathcal{B}_0(U^k;\delta_t e_u^k, \delta_t e_u^k) + \epsilon\max_{k\in[0,N]}\|e_u^k\|^2\\ &\quad + \epsilon(\Delta t)^3\sum_{k=0}^{m-1}\mathcal{E}(U^k;\delta_t e_u^k,\delta_t e_u^k) + \frac{C}{\epsilon}((\Delta t)^2+h^2)\\ &\quad + C\Delta t\sum_{k=0}^{m}\mathcal{D}^k + \epsilon\mathcal{D}^m.\end{aligned}\tag{4.18}$$

4.4. Completion of the proof of the error bound. We are now in position to complete the proof of the error estimate. Starting from the relation $\sum_{k=0}^{m-1} L^k = \sum_{k=0}^{m-1}\sum_{j=1}^{7} R_j^k$ and using (4.9) together with (4.18) and Lemma 4.4 we deduce

$$\begin{aligned}&(1-\epsilon)\mathcal{D}^m + (\Delta t)^2\left(\lambda - \frac{\bar\gamma}{\gamma_{\min}} - \epsilon - C\Delta t\right)\sum_{k=0}^{m-1}\mathcal{B}_0(U^k;\delta_t e_u^k,\delta_t e_u^k)\\ &\qquad + (1-C\epsilon)\Delta t\sum_{k=0}^{m-1}\mathcal{E}(U^k; e_w^{k+1}, e_w^{k+1}) + (1-C\epsilon)(\Delta t)^3\sum_{k=0}^{m-1}\mathcal{E}(U^k;\delta_t e_u^k,\delta_t e_u^k)\\ &\quad \le \mathcal{D}^0 + \frac{C}{\epsilon}((\Delta t)^2+h^2) + \frac{C}{\epsilon}\Delta t\sum_{k=0}^{m}\mathcal{D}^k.\end{aligned}$$

It follows from (2.7) that $\mathcal{D}^0 = \int_\Omega(\gamma(\nu(U^0)) - \gamma(\nu(u^0)) - \langle\gamma'(\nu(u^0)), (\nu(U^0)-\nu(u^0))\rangle)Q(U^0)$ so that by Taylor expansion and (2.28) $\mathcal{D}^0 \le Ch^2$. After choosing ϵ and Δt sufficiently small we obtain

$$\begin{aligned}\mathcal{D}^m + \frac{\Delta t}{2}\sum_{k=1}^{m}\mathcal{E}(U^k, e_w^{k+1}, e_w^{k+1}) + c_0(\Delta t)^2\sum_{k=0}^{m-1}\mathcal{B}(U^k;\delta_t e_u^k,\delta_t e_u^k)\\ \le C((\Delta t)^2+h^2) + C\Delta t\sum_{k=0}^{m-1}\mathcal{D}^k.\end{aligned}$$

Gronwall's lemma together with Lemma 2.8 implies that

$$\begin{aligned}\max_{m\in[0,N]}\int_\Omega|\nu(u^m)-\nu(U^m)|^2 Q(U^m) + \Delta t\sum_{k=1}^{N}\mathcal{E}(U^{k-1}, e_w^k, e_w^k)\\ + (\Delta t)^2\sum_{k=0}^{N-1}\mathcal{B}(U^k;\delta_t e_u^k,\delta_t e_u^k) \le C((\Delta t)^2+h^2)\end{aligned}$$

and the remainder of the proof of Theorem 4.1 now follows from Lemmas 4.3 and 4.4.

5. Numerical results.

5.1. The algebraic problem. Let $\{\chi_j\}$ denote the usual nodal basis functions for $\mathcal{S}^h$. Set

$$M_{i,j} = (\chi_i, \chi_j), \qquad E^m_{i,j} = \mathcal{E}(U^m; \chi_i, \chi_j), \qquad B^m_{i,j} = \mathcal{B}(U^m; \chi_i, \chi_j)$$

and

$$F^m_j = -\mathcal{A}(U^m, \chi_j) + \mathcal{B}(U^m; U^m, \chi_j).$$

It follows that the nodal values $\mathbf{U}^{m+1}$, $\mathbf{W}^{m+1}$ solve the linear algebraic system

$$\begin{aligned} \frac{1}{\Delta t} M\mathbf{U}^{m+1} + E^m \mathbf{W}^{m+1} &= \frac{1}{\Delta t} M\mathbf{U}^m, \\ B^m \mathbf{U}^{m+1} - M\mathbf{W}^{m+1} &= \mathbf{F}^m. \end{aligned}$$

Note that the structure of this system is of the same form as that arising in discretizations of the Cahn–Hilliard equation. Eliminating $\mathbf{W}^{m+1}$ by inverting the mass matrix in the second equation leads to the "fourth order" system

$$\frac{1}{\Delta t} M\mathbf{U}^{m+1} + E^m M^{-1} B^m \mathbf{U}^{m+1} = \frac{1}{\Delta t} M\mathbf{U}^m + E^m M^{-1}\mathbf{F}^m. \tag{5.1}$$

In our practical computations we have used mass lumping, so that M becomes a diagonal matrix. Although the system is unsymmetric, both the biconjugate gradient (BICG) and conjugate gradient (CG) methods were used to solve the linear equations. Remarkably, it was discovered that CG converged.

5.2. Convergence tests. We measured the actual error in different norms for several quantities for test problems, for which we know the continuous solutions. For this we have to extend our method to include right-hand sides f and g as indicated in (1.5). The tables contain the errors for the graph $u = u(x,t)$,

$$\begin{aligned} E_{\infty,2,u} &= \max_{m\in[0,N]} \|u^m - U^m\|, \\ E_{\infty,2,\nu} &= \max_{m\in[0,N]} \left(\int_\Omega |\nu(u^m) - \nu(U^m)|^2 Q(U^m) \right)^{\frac{1}{2}}, \end{aligned}$$

and for the curvature $w = w(x,t)$,

$$\begin{aligned} E_{2,2,w} &= \left(\Delta t \sum_{m=0}^{N} \|w^m - W^m\|^2 \right)^{\frac{1}{2}}, \\ E_{2,\mathcal{E},\nabla w} &= \left(\Delta t \sum_{m=0}^{M-1} \mathcal{E}(U^m; w^m - W^m, w^m - W^m) \right)^{\frac{1}{2}}. \end{aligned}$$

These are the errors which were estimated in Theorem 4.1. Additionally we provide the errors

$$E_{\infty,\infty,u} = \max_{m\in[0,N]} \|u^m - U^m\|_{L^\infty}, \qquad E_{\infty,2,\nabla u} = \max_{m\in[0,N]} \|\nabla u^m - \nabla U^m\|.$$

Table 5.1

Errors for the isotropic test problem with $\Delta t = 0.1h$.

h	$E_{\infty,2,u}$	*eoc*	$E_{\infty,2,\nu}$	*eoc*	$E_{2,2,w}$	*eoc*	$E_{2,\mathcal{E},\nabla w}$	*eoc*
1.0	8.495	-	0.4538	-	0.2264	-	5.534	-
0.7368	3.299	3.10	0.1702	3.21	0.6294	-3.35	2.965	2.04
0.4203	0.6255	2.96	0.06580	1.69	0.2343	1.76	1.097	1.77
0.2219	0.1564	2.17	0.03241	1.11	0.06291	2.06	0.4664	1.34
0.1137	0.04360	1.91	0.01622	1.04	0.01597	2.05	0.2234	1.10
0.05754	0.01306	1.77	0.008113	1.02	0.003942	2.05	0.1109	1.03

We also measure the error

$$E_{2,2,\nabla w} = \left(\Delta t \sum_{m=0}^{M-1} \int_\Omega \frac{|\nabla w^m - \nabla W^m|^2}{Q(U^m)} \right)^{\frac{1}{2}},$$

which is bounded from above by $E_{2,\mathcal{E},\nabla w}$. The error in the normal velocity is given by

$$E_{2,2,V} = \left(\Delta t \sum_{m=1}^{N} \int_\Omega (V(u^m) - V(U^m))^2 Q(U^m) \right)^{\frac{1}{2}},$$

where

$$V(u^m) = -\frac{u_t(\cdot, t_m)}{Q(u^m)}, \qquad V(U^m) = -\frac{U^m - U^{m-1}}{\Delta t\, Q(U^m)}.$$

Between two spatial discretization levels with grid sizes h_1 and h_2 we compute the experimental order of convergence

$$eoc(h_1, h_2) = \log \frac{E(h_1)}{E(h_2)} \left(\log \frac{h_1}{h_2} \right)^{-1}$$

for the errors $E(h_1)$ and $E(h_2)$ for each of the error norms.

For isotropic surface diffusion we used the function

$$u(x,t) = \frac{1}{2} \cos(t) \left(1 + |x|^2 - \frac{3}{4}|x|^4 + \frac{1}{6}|x|^6 \right)$$

as continuous solution on the domain $\Omega = \{x \in \mathbb{R}^2 \mid |x| < 1\}$ and on the time interval $[0,T] = [0,1]$. We calculated the right-hand side g from the equation

$$g = V - \Delta_\Gamma \mathcal{H}_\gamma,$$

and used this function as a right-hand side in our algorithm to compute U^m and W^m. We have chosen $\lambda = 1$. In Tables 5.1 and 5.2 we show the results for the time step size $\Delta t = 0.1\,h$ and Tables 5.3 and 5.4 contain the results for $\Delta t = h^2$. The results confirm the theoretical estimates from Theorem 4.1. Obviously the errors $E_{\infty,2,\nu}$ and $E_{\infty,2,\nabla u}$ as well as the errors $E_{2,\mathcal{E},\nabla w}$ and $E_{2,2,\nabla w}$ exhibit the same orders of convergence.

The anisotropic case was tested, see Tables 5.5 and 5.6, with the exact solution

$$u(x,t) = \sqrt{1 - 4t - 4x_1^2 - x_2^2}$$

Table 5.2
Errors for the isotropic test problem with $\Delta t = 0.1h$.

h	$E_{\infty,\infty,u}$	*eoc*	$E_{2,2,V}$	*eoc*	$E_{\infty,2,\nabla u}$	*eoc*	$E_{2,2,\nabla w}$	*eoc*
1.0	5.027	-	9.113	-	0.4676	-	5.529	-
0.7368	1.848	3.28	3.548	3.09	0.1767	3.19	2.952	2.06
0.4203	0.3365	3.03	0.6565	3.01	0.06754	1.71	1.090	1.78
0.2219	0.07905	2.27	0.2053	1.82	0.03305	1.12	0.4636	1.34
0.1137	0.01990	2.06	0.1093	0.94	0.01654	1.04	0.2221	1.10
0.05754	0.004986	2.03	0.07361	0.58	0.008272	1.02	0.1102	1.03

Table 5.3
Absolute errors for the isotropic test problem with $\Delta t = h^2$.

h	$E_{\infty,2,u}$	*eoc*	$E_{\infty,2,\nu}$	*eoc*	$E_{2,2,w}$	*eoc*	$E_{2,\mathcal{E},\nabla w}$	*eoc*
1.	1.523	-	0.5929	-	0.1119	-	3.135	-
0.7368	0.5954	3.08	0.1827	3.85	0.4998	-4.90	2.203	1.16
0.4203	0.5108	0.27	0.06818	1.76	0.1906	1.72	0.9358	1.53
0.2219	0.1661	1.76	0.03228	1.17	0.06028	1.80	0.4549	1.13
0.1137	0.04476	1.96	0.01622	1.03	0.01591	1.99	0.2234	1.06
0.05754	0.01146	2.00	0.008113	1.02	0.004031	2.02	0.1112	1.02

Table 5.4
Absolute errors for the isotropic test problem with $\Delta t = h^2$.

h	$E_{\infty,\infty,u}$	*eoc*	$E_{2,2,V}$	*eoc*	$E_{\infty,2,\nabla u}$	*eoc*	$E_{2,2,\nabla w}$	*eoc*
1.	1.003	-	0.9711	-	0.5960	-	3.116	-
0.7368	0.3781	3.19	0.7349	0.91	0.1854	3.82	2.189	1.16
0.4203	0.2202	0.96	0.6887	0.12	0.06911	1.76	0.9296	1.53
0.2219	0.07354	1.72	0.2628	1.51	0.03292	1.16	0.4522	1.13
0.1137	0.01989	1.96	0.1163	1.22	0.01654	1.03	0.2221	1.06
0.05754	0.005087	2.00	0.05621	1.07	0.008271	1.02	0.1105	1.03

Table 5.5
Absolute errors for the anisotropic test problem with $\Delta t = h^2$.

h	$E_{\infty,2,u}$	*eoc*	$E_{\infty,2,\nu}$	*eoc*	$E_{2,2,w}$	*eoc*	$E_{2,\mathcal{E},\nabla w}$	*eoc*
0.1250	0.1475e-1	-	0.1354e-1	-	0.1409e-1	-	0.1207e-3	-
0.7138e-1	0.4999e-2	1.93	0.8346e-2	0.86	0.3483e-2	2.50	0.5734e-4	1.33
0.3807e-1	0.1458e-2	1.96	0.4399e-2	1.02	0.8862e-3	2.18	0.1997e-4	1.68
0.1964e-1	0.3937e-3	1.98	0.2216e-2	1.04	0.2221e-3	2.09	0.6971e-5	1.59
0.9969e-2	0.1032e-3	1.98	0.1110e-2	1.02	0.5553e-4	2.05	0.3079e-5	1.21

Table 5.6
Absolute errors for the anisotropic test problem with $\Delta t = h^2$.

h	$E_{\infty,\infty,u}$	*eoc*	$E_{2,2,V}$	*eoc*	$E_{\infty,2,\nabla u}$	*eoc*	$E_{2,2,\nabla w}$	*eoc*
0.1250	0.8037e-1	-	0.7644e-1	-	0.1547e-1	-	0.1200e-3	-
0.7138e-1	0.2658e-1	1.98	0.4285e-1	1.03	0.9390e-2	0.89	0.5710e-4	1.33
0.3807e-1	0.7753e-2	1.96	0.2293e-1	0.99	0.4894e-2	1.04	0.1988e-4	1.68
0.1964e-1	0.2093e-2	1.98	0.1182e-1	1.00	0.2453e-2	1.04	0.6921e-5	1.59
0.9969e-2	0.5481e-3	1.98	0.5997e-2	1.00	0.1227e-2	1.02	0.3080e-5	1.22

on the domain $\Omega = \{x \in \mathbb{R}^2 \mid |x| < 0.125\}$ and for $t \in [0, 0.125]$. Domain and time interval have to be relatively small in order to remain in the setting of a graph. As in the isotropic case we have used a right-hand side g, and since u does not satisfy the natural boundary condition, we have extended the concept to include the inhomogeneous Neumann boundary condition $\langle \gamma'(\nabla u, -1), (\nu_{\partial\Omega}, 0)\rangle = c$ for a given

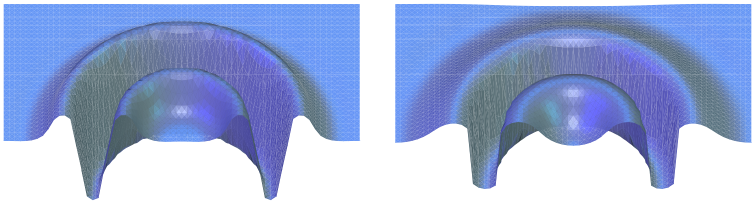


FIG. 5.1. *Initial function which leads to loss of the graph property after short time and solution becoming vertical (cut along the x_1-x_3 plane of symmetry).*

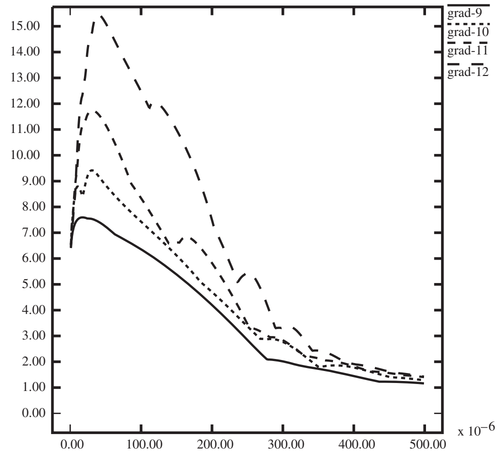


FIG. 5.2. *Lipschitz-norm of the discrete solution (vertical axis) plotted as a function of time* $t \in [0, 0.0005]$ *for the initial function from* (5.1) *for different spatial discretization levels.*

function c on $\partial\Omega$. As anisotropy we have used

$$\gamma(p) = \sqrt{0.25p_1^2 + p_2^2 + p_3^2}$$

and the stabilizing parameter was $\lambda = 1$.

We add an example of a surface which moves under isotropic surface diffusion and which loses its graph property in finite time. Nevertheless the discrete solution exists for all times. In Figure 5.1 two steps of the evolution are shown. In Figure 5.2 the maxima of the moduli of the gradients of the discrete solution is plotted as a function of time. The computational domain is $\Omega = (-1, 1)^2$ and the time interval is $[0, 0.0005]$. The graph of the solution becomes vertical after a short time, but the discrete solution continues to exist. We show the maximal gradient for the discretization levels 9, 10, 11, and 12. Observe that the number $1/h$ is 8.0, 11.32, 16.0, and 22.63 for these levels and by comparison with peaks in the graph of Figure 5.2 we see the suggestion of "infinite" gradients.

5.3. Numerical experiments. We end this section with two illustrative computations. First, we demonstrate the smoothing property of isotropic surface diffusion by choosing a highly oscillatory initial function u_0,

$$u_0(x) = 1 + 0.1\big(\sin(2(m+1)\pi x_1) + \sin(2m\pi x_1)(\sin(2(m+1)\pi x_2) + \sin(2m\pi x_2))\big) \tag{5.2}$$

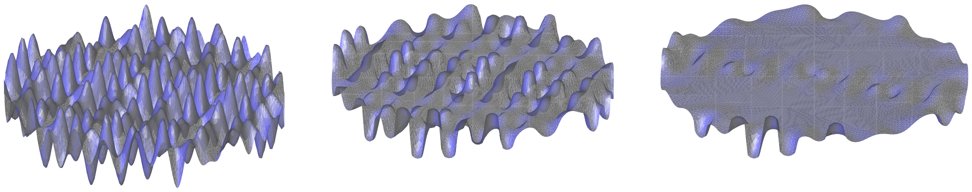


FIG. 5.3. *Solution u for the initial data* (5.2) *at times* 0.0, 3.5×10^{-6}, *and* 6.3×10^{-6}.

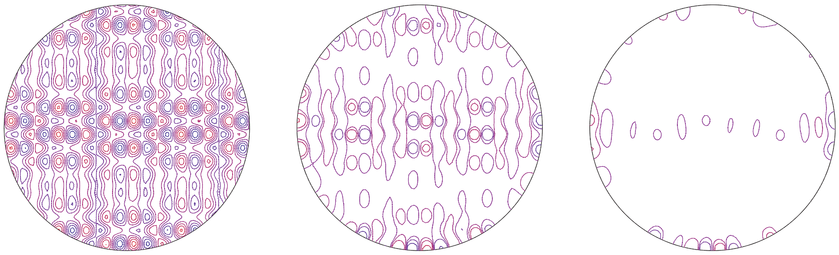


FIG. 5.4. *Level lines of the solution from Figure* 5.3.

with $m = 4$. The computational domain is the unit disk $\Omega = \{x \in \mathbb{R}^2 \mid |x| < 1\}$, and we have used natural boundary conditions. The grid has to be fine in order to capture the frequency of the initial function. In order to show the rapid smoothing of u_0 we have chosen an extremely small time step proportional to h^4. In Figure 5.3 we show the solution at the times 0.0, 7.0×10^{-6} and 1.4×10^{-5}. Figure 5.4 shows level lines of the solution for these time steps. The level lines are equally distributed between the values 0.65 and 1.35 and are the same in all three cases.

Second, we computed an example for anisotropic surface diffusion with an extremely strong anisotropy. The anisotropy is chosen to be a regularized l^1 norm,

$$\gamma(p) = \sum_{j=1}^{3} \sqrt{p_j^2 + \varepsilon^2 |p|^2}, \tag{5.3}$$

where we have chosen $\varepsilon = 10^{-3}$. Thus the Frank diagram is a smoothed octahedron and the Wulff shape is a smoothed cube. The initial data were taken to depend on three random numbers $r_1, r_2, r_3 \in (0, 1)$,

$$\begin{aligned} u_0(x) = {} & \frac{1}{4}\left(\sin(2\pi r_1 x_1) + \frac{1}{4}\sin(3\pi r_2 x_2)\right)(0.1\sin(2\pi r_3 x_1) + \sin(5\pi r_1 x_2)) \\ & \times \sin(2\pi r_2 x_1 x_2). \end{aligned} \tag{5.4}$$

We used Neumann boundary conditions and the right-hand side (for the curvature equation) $f = 1 - x_1^2 - x_2^2$. The domain is given as $\Omega = (-1, 1) \times (-1, 1)$, and the triangulation contains 16641 vertices and 32768 triangles. We chose $\lambda = 4$. In Figure 5.5 we show the graph of the solution u in the direction of the x_1-axis. Figure 5.6 shows the graph for the time steps 0, 50, and 200. The Wulff shape (a smooth cube) appears in the solution as a consequence of the right-hand side f.

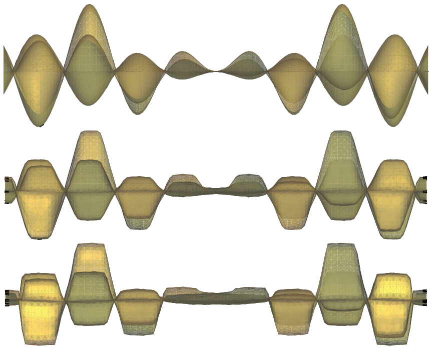


Fig. 5.5. *Anisotropic surface diffusion for the initial function* (5.4) *with anisotropy* (5.3), *viewed from the x_1-axis. Time steps* 0, 50, 200.

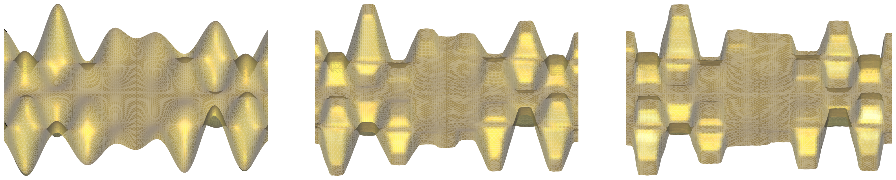


Fig. 5.6. *The solution from Figure* 5.5 *shown as graph.*

REFERENCES

[1] E. Bänsch, P. Morin, and R.H. Nochetto, *Surface diffusion of graphs: Variational formulation, error analysis, and simulation*, SIAM J. Numer. Anal., 42 (2004), pp. 773–799.

[2] G. Bellettini and M. Paolini, *Anisotropic motion by mean curvature in the context of Finsler geometry*, Hokkaido Math. J., 25 (1996), pp. 537–566.

[3] A.J. Bernoff, A.L. Bertozzi, and T.P. Witelski, *Axisymmetric surface diffusion: Dynamics and stability of self-similar pinchoff*, J. Statist. Phys., 93 (1998), pp. 725–776.

[4] J.W. Cahn, C.M. Elliott, and A. Novick-Cohen, *The Cahn-Hilliard equation with a concentration dependent mobility: Motion by minus the Laplacian of the mean curvature*, European J. Appl. Math., 7 (1996), pp. 287–301.

[5] J.W. Cahn and J.E. Taylor, *Surface motion by surface diffusion*, Acta metall. mater., 42 (1994), pp. 1045–1063.

[6] W.C. Carter, A.R. Roosen, J.W. Cahn, and J.E. Taylor, *Shape evolution by surface diffusion and surface attachment limited kinetics on completely faceted surfaces*, Acta metall. mater., 43 (1995), pp. 4309–4323.

[7] B.D. Coleman, R.S. Falk, and M. Moakher, *Stability of cylindrical bodies in the theory of surface diffusion*, Phys. D, 89 (1995), pp. 123–135.

[8] B.D. Coleman, R.S. Falk, and M. Moakher, *Space-time finite element methods for surface diffusion with applications to the theory of the stability of cylinders*, SIAM J. Sci. Comput., 17 (1996), pp. 1434–1448.

[9] K. Deckelnick and G. Dziuk, *Discrete anisotropic curvature flow of graphs*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1203–1222.

[10] K. Deckelnick and G. Dziuk, *Error estimates for a semi-implicit fully discrete finite element scheme for the mean curvature flow of graphs*, Interfaces Free Bound., 2 (2000), pp. 341–359.

[11] K. Deckelnick and G. Dziuk, *A fully discrete numerical scheme for weighted mean curvature flow*, Numer. Math., 91 (2002), pp. 423–452.

[12] K. Deckelnick, G. Dziuk, and C.M. Elliott, *Error analysis of a semidiscrete numerical scheme for diffusion in axially symmetric surfaces*, SIAM J. Numer. Anal., 41 (2003), pp. 2161–2179.

[13] G. Dziuk, *Numerical schemes for the mean curvature flow of graphs*, in IUTAM Symposium on Variations of Domains and Free-Boundary Problems in Solid Mechanics, P. Argoul, M.

Frémond, and Q.S. Nguyen, eds., Kluwer Academic Publishers, Dordrecht-Boston-London, 1999, pp. 63–70.

[14] C.M. Elliott, D.A. French, and F.A. Milner, *A second order splitting method for the Cahn-Hilliard equation*, Numer. Math., 54 (1989), pp. 575–590.

[15] C.M. Elliott and H. Garcke, *Existence results for diffusive surface motion laws*, Adv. Math. Sci. Appl., 7 (1997), pp. 467–490.

[16] C.M. Elliott and S. Maier-Paape, *Losing a graph with surface diffusion*, Hokkaido Math. J., 30 (2001), pp. 297–305.

[17] J. Escher, U.F. Mayer, and G. Simonett, *The surface diffusion flow for immersed hypersurfaces*, SIAM J. Math. Anal., 29 (1998), pp. 1419–1433.

[18] Y. Giga and K. Ito, *On pinching of curves moved by surface diffusion*, Commun. Appl. Anal., 2 (1998), pp. 393–405.

[19] C. Herring, *Surface diffusion as a motivation for sintering*, in The Physics of Powder Metallurgy, W.E. Kingston, ed., McGraw Hill, New York, 1951, pp. 143–179.

[20] U.F. Mayer and G. Simonett, *Self-intersections for the surface diffusion and the volume-preserving mean curvature flow*, Differential Integral Equations, 13 (2000), pp. 1189–1199.

[21] W.W. Mullins, *Theory of thermal grooving*, J. Appl. Phys., 28 (1957), pp. 333–339.

[22] F.A. Nichols and W.W. Mullins, *Surface–(interface–) and volume–diffusion contributions to morphological changes driven by capillarity*, Trans. Metall. Soc. AIME, 233 (1965), pp. 1840–1847.

[23] P. Smereka, *Semi-implicit level set methods for curvature and surface diffusion motion*, J. Sci. Comput., 19 (2003), pp. 439–456.

[24] J.E. Taylor, J.W. Cahn, and C.A. Handwerker, *Geometric models of crystal growth*, Acta metall. mater., 40 (1992), pp. 1443–1474.

[25] H. Wong, M.J. Miksis, P.W. Voorhees, and S.H. Davis, *Universal pinch off of rods by capillarity-driven surface diffusion*, Scripta Mater., 39 (1998), pp. 55–60.

SIAM J. Numer. Anal.
Vol. 43, No. 3, pp. 1139–1154

NUMERICAL INTEGRATION OF STOCHASTIC DIFFERENTIAL EQUATIONS WITH NONGLOBALLY LIPSCHITZ COEFFICIENTS*

G. N. MILSTEIN†‡ AND M. V. TRETYAKOV‡

Abstract. We propose a new concept which allows us to apply any numerical method of weak approximation to a very broad class of stochastic differential equations (SDEs) with nonglobally Lipschitz coefficients. Following this concept, we discard the approximate trajectories which leave a sufficiently large sphere. We prove that accuracy of any method of weak order p is estimated by $\varepsilon + O(h^p)$, where ε can be made arbitrarily small with increasing radius of the sphere. The results obtained are supported by numerical experiments.

Key words. SDEs with nonglobally Lipschitz coefficients, numerical integration of SDEs in the weak sense

AMS subject classifications. 60H35, 65C30, 60H10

DOI. 10.1137/040612026

1. Introduction.

Stochastic differential equations (SDEs) with nonglobally Lipschitz coefficients possessing unique solutions make up a very important class in applications. For instance, Langevin-type equations and gradient systems with noise belong to this class [10, 9, 1, 5, 11]. At the same time, most numerical methods for SDEs are derived under the global Lipschitz condition [3, 7]. If this condition is violated, the behavior of many standard numerical methods in the whole space can lead to incorrect conclusions (see, for instance, [9, 1, 5, 11, 4]). This situation is very alarming since we are forced to refuse many effective methods and/or to resort to some comparatively complicated and inefficient numerical procedures. In [6] (see also Example 3.3 here), applying an explicit quasi-symplectic method of weak approximation to a Langevin equation with nonglobally Lipschitz coefficients for calculating an ergodic limit, the authors found an explosive behavior of some approximate trajectories. The explosions are observed outside of a comparatively large sphere after a relatively large time and very rarely. Clearly, the exploding approximate trajectories badly reproduce the actual behavior of the considered system. We have also found that if these rare trajectories are discarded, then the explicit quasi-symplectic method gives much better results than the implicit Euler method, which does not have any exploding trajectories. From the heuristic point of view, this is rather natural. Roughly speaking, the value of an ergodic limit depends, on the whole, on the behavior of trajectories in a bounded (though large) domain on a finite (though large) time interval. Consequently, any method that is effective for systems with globally Lipschitz coefficients has to work well for systems with nonglobally Lipschitz coefficients as well if one rejects a small number of "bad" trajectories.

In this paper, we propose a new concept which allows us to apply any method of weak approximation to a very broad class of SDEs with nonglobally Lipschitz coef-

*Received by the editors July 21, 2004; accepted for publication (in revised form) January 13, 2005; published electronically September 23, 2005. This research was supported by EPSRC Research Grant GR/R90017/01.

http://www.siam.org/journals/sinum/43-3/61202.html

†Department of Mathematics, Ural State University, Lenin Str. 51, 620083 Ekaterinburg, Russia (Grigori.Milstein@usu.ru).

‡Department of Mathematics, University of Leicester, Leicester LE1 7RH, UK (M.Tretiakov@le.ac.uk).

ficients. Roughly speaking, we require for SDEs from this class just to have regular solutions on a time interval $[t_0, T]$ and to have sufficiently smooth coefficients; i.e., the assumptions made (which are given in terms of Lyapunov functions in section 2) usually hold for SDEs of applicable interest. (We note that convergence of the explicit Euler method is proved under more restrictive assumptions in [1].) Following the concept proposed here, we discard the approximate trajectories which leave a sufficiently large sphere $S_R := \{x : |x| < R\}$. The theoretical justification of the concept is given in section 2. We prove that accuracy of any method of weak order p is estimated by $\varepsilon + O(h^p)$, where ε can be made arbitrarily small with growing R, and $|O(h^p)| \leq Kh^p$ can be made arbitrarily small with decreasing h (of course, K depends on R). Thus, we obtain that the violation of the global Lipschitz condition is not fatal for applying any method of numerical integration. Since the Monte Carlo technique is used for simulation of a mean $Ef(X(T))$, the error estimation $\varepsilon + O(h^p)$ should be increased by the Monte Carlo error. It turns out that in practice the error given by ε is much smaller than the joint numerical integration and Monte Carlo error. Furthermore, in principle, due to the concept of rejecting "bad" trajectories, we can choose a suitable method for solving a system of SDEs with nonglobally Lipschitz coefficients, taking into account all the known methods of numerical integration [3, 7]. The application of the concept is discussed in section 3, where some numerical experiments are presented.

An implication of the concept proposed here for the calculation of ergodic limits will be considered in a separate publication.

2. Integration via paths in a bounded domain. Consider the system of Ito SDEs

$$dX = a(t, X)dt + \sum_{l=1}^{q} \sigma_l(t, X)dw_l(t), \qquad X(t_0) = x, \tag{2.1}$$

where X, a, σ_l are d-dimensional column-vectors and $w_l(t)$, $l = 1, \ldots, q$, are independent standard Wiener processes.

We suppose the coefficients of (2.1) to be sufficiently smooth functions in $[t_0, T] \times \mathbf{R}^d$, and any solution $X(t; t_0, x)$ of (2.1) to be regular on $[t_0, T]$. We recall that a process is called regular if it is defined for all $t_0 \leq t \leq T$. Denote by $\mathbf{C}^2$ the class of functions defined on $[t_0, T] \times \mathbf{R}^d$ and twice continuously differentiable with respect to x and once with respect to t. A sufficient condition of regularity (see [2]) consists of the existence of a Lyapunov function $V \in \mathbf{C}^2$, $V \geq 0$, which satisfies the inequality

$$LV(t, x) \leq c_0 V(t, x) + c_1, \qquad (t, x) \in [t_0, T] \times \mathbf{R}^d, \tag{2.2}$$

and

$$V_R := \min_{t_0 \leq t \leq T,\ |x| \geq R} V(t, x), \qquad \lim_{R \to \infty} V_R = \infty, \tag{2.3}$$

where c_0 and c_1 are some constants and L is the following generating operator:

$$LV(t, x) = \frac{\partial V}{\partial t}(t, x) + \sum_{i=1}^{d} a^i(t, x) \frac{\partial V}{\partial x^i}(t, x) + \frac{1}{2} \sum_{i,j=1}^{d} a^{ij}(t, x) \frac{\partial^2 V}{\partial x^i \partial x^j}(t, x), \tag{2.4}$$

$$a^{ij} := \sum_{l=1}^{q} \sigma_l^i \sigma_l^j.$$

Moreover, if (2.2) and (2.3) are fulfilled and if an initial distribution for x (x can be random) is such that $EV(t_0, x)$ exists, then $EV(t, X(t; t_0, x))$ exists for all $t_0 \le t \le T$. For instance, if V has an m-polynomial growth at infinity, then there exist the moments of order m for X. We note that it is not required that c_0 be negative. For definiteness, we consider $c_0 > 0$ and $c_1 > 0$.

Let $S_R := \{x : |x| < R\}$ be an open sphere in $\mathbf{R}^d$, $Q_R = [t_0, T) \times S_R$ be a cylinder in $\mathbf{R}^{d+1}$, $\Gamma_R = \bar{Q}_R \backslash Q_R$, where $\bar{Q}_R$ is the closure of Q_R. The set Γ_R is a part of the boundary of cylinder Q_R consisting of the upper base and the lateral surface. Let τ_R be the first-passage time of the process $(t, X(t; t_0, x))$, $t_0 \le t \le T$, to Γ_R. Clearly, $t_0 \le \tau_R \le T$. We introduce the following events:

$$\begin{aligned} \Omega_R &: = \{\omega : |X(s; t_0, x)| < R,\ t_0 \le s < T\} = \{\omega : \tau_R = T\}, \\ \Lambda_R &: = \{\omega : \exists\ s \in [t_0, T) \ \text{ such that } \ |X(s; t_0, x)| \ge R\} = \{\omega : \tau_R < T\}. \end{aligned} \tag{2.5}$$

Let us obtain an upper bound for the probability

$$p_R := P(\tau_R < T) = P(\Lambda_R), \tag{2.6}$$

assuming (2.2) and (2.3) (see [2]). Introduce the nonnegative function

$$U(t, x) = (c_1 + 1)(T - t) + \exp(c_0(t_0 - t))V(t, x), \tag{2.7}$$

where $V(t, x)$ is a function satisfying (2.2)–(2.3). We get

$$LU(t, x) = -(c_1 + 1) - c_0 \exp(c_0(t_0 - t))V(t, x) + \exp(c_0(t_0 - t))LV(t, x) \le -1. \tag{2.8}$$

Due to the Ito formula, we have

$$\begin{aligned} dU(t, X(t; t_0, x)) &= LU(t, X(t; t_0, x))dt \\ &+ \exp(c_0(t_0 - t)) \sum_{i=1}^{d} \frac{\partial V}{\partial x^i}(t, X(t; t_0, x)) \sum_{l=1}^{q} \sigma_l^i(X(t; t_0, x))dw_l(t). \end{aligned} \tag{2.9}$$

Hence

$$U(\tau_R, X(\tau_R; t_0, x)) - U(t_0, x) = \int_0^{\tau_R} LU dt + \int_0^{\tau_R} \exp(c_0(t_0 - t)) \sum_{i=1}^{d} \frac{\partial V}{\partial x^i} \sum_{l=1}^{q} \sigma_l^i \ dw_l(t). \tag{2.10}$$

The expectation of the second integral on the right-hand side of (2.10) is equal to zero according to the martingale property. Therefore, due to (2.8), we get

$$EU(\tau_R, X(\tau_R; t_0, x)) \le U(t_0, x) = (c_1 + 1)(T - t_0) + V(t_0, x). \tag{2.11}$$

By Chebyshev's inequality, we obtain from (2.11)

$$\begin{aligned} p_R \exp(c_0(t_0 - T)) \min_{t_0 \le t \le T,\ |x| \ge R} V(t, x) &\le p_R \min_{t_0 \le t \le T,\ |x| \ge R} U(t, x) \\ &\le (c_1 + 1)(T - t_0) + V(t_0, x), \end{aligned} \tag{2.12}$$

whence

$$p_R \le \exp(c_0(T-t_0))\frac{(c_1+1)(T-t_0)+V(t_0,x)}{\min_{t_0\le t\le T,\ |x|\ge R} V(t,x)}, \tag{2.13}$$

and therefore

$$\lim_{R\to\infty} p_R = 0.$$

PROPOSITION 2.1. *Let* (2.2) *and* (2.3) *be fulfilled. Let $f(x)$ be a function such that*

$$|f(x)| \le V(t,x), \qquad t_0 \le t \le T,\ x\in \mathbf{R}^d. \tag{2.14}$$

Then for any $x\in R^d$ and $\varepsilon>0$ there exists $R(x,\varepsilon)>0$ such that for any $R>R(x,\varepsilon)$

$$|Ef(X(T;t_0,x)) - E[f(X(T;t_0,x))\chi_{\Omega_R}(\omega)]| < \varepsilon. \tag{2.15}$$

Proof. Clearly,

$$\lim_{R\to\infty} f(X(T;t_0,x))\chi_{\Omega_R}(\omega) = f(X(T;t_0,x)),\ \text{a.s.}$$

Now the conclusion of this proposition follows from the existence of $EV(T,X(T;t_0,x))$ and the Lebesgue theorem on majorized convergence. □

The significance of this proposition consists of the capability to disregard the trajectories running off too far. We are about to show that when systems under consideration are numerically integrated, the approximating trajectories running off too far can also be discarded. Due to this possibility, we are able, in principle, to use any known method of numerical integration for calculating means. In this respect we shall rest on the developed theory of weak approximation for SDEs with globally Lipschitz coefficients [3, 7]. To this aim we introduce an auxiliary system with globally Lipschitz coefficients, which coincides with the original system in a sphere $S_{R'}$ somewhat wider than S_R: $S_{R'}\supset S_R$, where $R'=R+r$ and $r>0$ is a constant.

Let the coefficients a^i, σ^i_l and the function f have continuous derivatives up to some order. The requirement on smoothness depends on a particular numerical method used; in general, the higher the order of the method, the more derivatives are needed. We construct coefficients $a^i_R, (\sigma^i_l)_R$ and function f_R so that in $[t_0,T]\times S_{R'}$ they coincide with a^i, σ^i_l and f, respectively, and, in addition, they are bounded in $[t_0,T]\times\mathbf{R}^d$ together with their derivatives up to the same order. This can be done in the following way. Introduce the function $\varphi(z)$ of one variable z:

$$\varphi(z) = \begin{cases} z, & -R'\le z\le R', \\ R' + \displaystyle\int_{R'}^{z}\frac{dz'}{1+(z'-R')^k}, & z>R', \\ -R' - \displaystyle\int_{z}^{-R'}\frac{dz'}{1+(-R'-z')^k}, & z<-R', \end{cases} \tag{2.16}$$

where $k\ge 2$ is a natural number. Clearly, $\varphi(z)$ is bounded on $\mathbf{R}^1$ together with its derivatives up to order k. Let $g(t,x^1,\dots,x^d)$ be a function with some continuous derivatives defined in $[t_0,T]\times\mathbf{R}^d$. It is easily seen that

$$g_R(t,x^1,\dots,x^d) := g(t,\varphi(x^1),\dots,\varphi(x^d))$$

satisfies the above-mentioned conditions. Moreover, there exists a constant $\rho > r$ (which does not depend on R) such that for any $x = (x^1, \dots, x^d) \in \mathbf{R}^d$ the point $(\varphi(x^1), \dots, \varphi(x^d)) \in S_{R+\rho}$. Therefore

$$\sup_{x\in\mathbf{R}^d} |f_R(x)| \le \max_{|x|\le R+\rho} |f(x)| \,. \tag{2.17}$$

Introduce the auxiliary system of SDEs

$$dX_R = a_R(X_R)dt + \sum_{l=1}^{q} (\sigma_l^i)_R(X_R)dw_l(t) \,. \tag{2.18}$$

We emphasize that this system is used in our theoretical proofs only; it is not used in simulation.

Proposition 2.2. *Assume that* $V(t,x)$ *satisfies* (2.2), (2.3), *and*

$$\frac{\min_{t_0\le t\le T,\ |x|\ge R+\rho} V(t,x)}{\min_{t_0\le t\le T,\ |x|\ge R} V(t,x)} \le c, \tag{2.19}$$

where c *is a constant which is independent of* R. *Let* $f(x)$ *be a function such that*

$$\lim_{R\to\infty} \frac{\max_{|x|\le R} |f(x)|}{\min_{t_0\le t\le T,\ |x|\ge R} V(t,x)} = 0. \tag{2.20}$$

Then for any $x \in R^d$ *and* $\varepsilon > 0$ *there exists* $R(x,\varepsilon) > 0$ *such that for any* $R > R(x,\varepsilon)$

$$|Ef_R(X_R(T;t_0,x)) - Ef(X(T;t_0,x))| < \varepsilon. \tag{2.21}$$

Proof. Since the solutions $X(t;t_0,x)$ and $X_R(t;t_0,x)$ and also the functions $f(X(t;t_0,x))$ and $f_R(X_R(t;t_0,x))$ coincide on the interval $t \in [t_0, \tau_R]$, we have

$$E[f_R(X_R(T;t_0,x))\chi_{\Omega_R}(\omega)] = E[f(X(T;t_0,x))\chi_{\Omega_R}(\omega)]. \tag{2.22}$$

Hence

$$\begin{aligned} &|Ef_R(X_R(T;t_0,x)) - Ef(X(T;t_0,x))| \\ &\le |E[f_R(X_R(T;t_0,x))\chi_{\Lambda_R}(\omega)]| + |E[f(X(T;t_0,x))\chi_{\Lambda_R}(\omega)]|. \end{aligned} \tag{2.23}$$

Proposition 2.1 implies

$$\lim_{R\to\infty} |E[f(X(T;t_0,x))\chi_{\Lambda_R}(\omega)]| = 0. \tag{2.24}$$

Further, due to (2.17) and (2.13), we obtain

$$\begin{aligned} E|f_R(X_R(T;t_0,x))\chi_{\tau_R<T}(\omega)| &\le \sup_{x\in\mathbf{R}^d} |f_R(x)| p_R \le \max_{|x|\le R+\rho} |f(x)| p_R \\ &\le \frac{\max_{|x|\le R+\rho} |f(x)|}{\min_{t_0\le t\le T,\ |x|\ge R+\rho} V(t,x)} \times \frac{\min_{t_0\le t\le T,\ |x|\ge R+\rho} V(t,x)}{\min_{t_0\le t\le T,\ |x|\ge R} V(t,x)} \\ &\quad \times \exp(c_0(T-t_0))[(c_1+1)(T-t_0) + V(t_0,x)]. \end{aligned} \tag{2.25}$$

Now, by using the conditions (2.19) and (2.20), we complete the proof. □

Our next step is to show that approximating paths obtained by a numerical method applied to the system (2.1) belong to the bounded domain with a large probability and that averaging via these paths gives a good approximation for the mean $Ef(X(T;t_0,x))$.

Let us start with some necessary auxiliary knowledge of the Markov chains generated by numerical methods. Consider the system of SDEs in the sense of Ito:

$$dY = b(t,Y)dt + \sum_{l=1}^{q} \gamma_l(t,Y)dw_l(t). \tag{2.26}$$

We assume that the functions $b(t,y)$ and $\gamma_l(t,y),\ (t,y) \in [t_0,T] \times \mathbf{R}^d$, have bounded derivatives with respect to t, y up to some order. In particular, the system (2.18) satisfies this assumption. In most cases a method (both mean-square and weak) can be defined by a one-step approximation of the form

$$\bar{Y}(t+h;t,y) = y + A(t,y,h;\xi), \qquad t_0 \le t < t+h \le T, \tag{2.27}$$

where ξ is a random vector having moments of a sufficiently high order and A is a vector function of dimension d.

Partition the interval $[t_0,T]$ into N equal parts with the step $h = (T-t_0)/N$: $t_0 < t_1 < \cdots < t_N = T$, $t_{k+1} - t_k = h$. According to (2.27), we construct the sequence

$$\bar{Y}_0 = Y(t_0) = y, \quad \bar{Y}_{k+1} = \bar{Y}_k + A(t_k,\bar{Y}_k,h;\xi_k), \qquad k = 0,\ldots,N-1, \tag{2.28}$$

where ξ_0 is independent of $\bar{Y}_0$, while ξ_k for $k > 0$ are independent of $\bar{Y}_0,\ldots,\bar{Y}_k$, $\xi_0,\ldots,\xi_{k-1}$. The sequence $\bar{Y}_k$ is a Markov chain. Its transition probability function is defined by

$$P(t,y,s,D) = P(\bar{Y}(s;t,y) \in D), \qquad s \ge t,\ t,s = t_0, t_0+h,\ldots,T, \tag{2.29}$$

where $\bar{Y}(s;t,y)$ is the process with discrete time starting at the moment t from y and defined by (2.28). The generating operator of the Markov chain is defined by

$$\begin{aligned} L_h U(t,y) &= \frac{1}{h}\int P(t,y,t+h,dz)[U(t+h,z) - U(t,y)] \\ &= \frac{1}{h}[EU(t+h,\bar{Y}(t+h;t,y)) - U(t,y)]. \end{aligned} \tag{2.30}$$

Denote by $\bar{\tau}_R$ the first exit time of the process $(t_k, \bar{Y}(t_k;t_0,y))$, $k = 0,\ldots,N$, from $[t_0,T) \times S_R$. Due to (2.30), we have (see [8])

$$EU(\bar{\tau}_R, \bar{Y}(\bar{\tau}_R;t_0,y)) - U(t_0,y) = E\sum_{k=0}^{\kappa-1} L_h U(t_k,\bar{Y}(t_k;t_0,y))h, \tag{2.31}$$

where κ is defined by $t_\kappa = \bar{\tau}_R$.

In what follows we use the following assumption.

(A1) *The one-step approximation* (2.27) *is at least of order two in the weak sense, and the method defined by this approximation converges at least with order one.*

By the definition of weak approximation

$$|EU(t+h,\bar{Y}(t+h;t,y)) - EU(t+h,Y(t+h;t,y))| \le Kh^2, \tag{2.32}$$

where K is a positive constant, provided that y belongs to a compact set. At the same time (see [7])

$$EU(t+h,Y(t+h;t,y)) = U(t,y) + hLU(t,y) + O(h^2), \tag{2.33}$$

where $|O(h^2)| \le Kh^2$ and K is again independent of y belonging to a compact set. Using (2.32) and (2.33), we get

$$|EU(t+h,\bar{Y}(t+h;t,y)) - U(t,y) - hLU(t,y)| \le Kh^2,$$

and then we obtain from (2.30)

$$|L_hU(t,y) - LU(t,y)| \le Kh. \tag{2.34}$$

Let us proceed to the system (2.18). Apply a numerical method of weak order $p \ge 1$ to the systems (2.1) and (2.18). As a result, we obtain two Markov chains $\bar{X}_k$ and $(\bar{X}_R)_k$. For the Markov chain $(\bar{X}_R)_k$ (but not for $\bar{X}_k$) we have for $(t,x) \in [t_0,T] \times S_R$ (see [7])

$$|Ef(X_R(T;t_0,x)) - Ef(\bar{X}_R(T;t_0,x))| \le Kh^p. \tag{2.35}$$

Let L_R be the generating operator for (2.18), and $(L_R)_h$ for $\bar{X}_R$. According to (2.34), we get

$$|(L_R)_hU(t,x) - L_RU(t,x)| \le Kh, \qquad (t,x) \in [t_0,T] \times S_R. \tag{2.36}$$

If $(t,x) \in [t_0,T] \times S_R$, then

$$L_RU(t,x) = LU(t,x). \tag{2.37}$$

Due to (2.8), we obtain that $LU \le -1$. It follows from this inequality together with (2.36) and (2.37) that for all h small enough

$$(L_R)_hU(t,x) \le 0, \qquad (t,x) \in [t_0,T] \times S_R. \tag{2.38}$$

In future we need the following assumption.

(A2) *If $(\bar{X}_R)_i \in S_R$, $i = 0,\dots,k$, then $\bar{X}_k = (\bar{X}_R)_k$, $k \le N$. (Of course, the approximating trajectories are starting from the same point: $\bar{X}_0 = (\bar{X}_R)_0 = x$.)*

The assumption is evidently true, for instance, for the explicit Euler method. Moreover, for this method even $\bar{X}_{k+1} = (\bar{X}_R)_{k+1}$ if only $\bar{X}_k = (\bar{X}_R)_k \in S_R$, though $\bar{X}_{k+1}$ may not belong to S_R. The definition of φ (see (2.16)) ensures coincidence of the coefficients of (2.1) and (2.18) in the wider domain S_{R+r}. Due to this fact, (A2) is fulfilled for h small enough if ξ_k are bounded. This is the most typical case for weak methods, while, applying mean-square methods, we can use random variables such that $\xi_k h^{1/2}$ are small if h is small (see [7]). Thus, the condition (A2) is fulfilled for typical mean-square methods as well. Nevertheless, we should pay attention that a method of the type (2.27) with initial data from $[t_0,T] \times S_R$ may depend on behavior of a system's coefficients not only in $[t_0,T] \times S_{R+r}$ but, generally speaking, in $[t_0,T] \times \mathbf{R}^d$.

Let us take h ensuring (2.38). Denote by $\bar{\tau}_R$ the first exit time of the chain $(t_k, \bar{X}_R(t_k;t_0,x))$ from $[t_0,T) \times S_R$, i.e., $\bar{X}_R(t_k;t_0,x) \in S_R$, $k = 0,1,\ldots,\kappa-1$, and either $\bar{X}_R(t_\kappa;t_0,x) = \bar{X}_R(\bar{\tau}_R;t_0,x) \notin S_R$, where $\bar{\tau}_R = t_\kappa$, or $t_\kappa = T$. Applying (2.31) to $\bar{X}_R$ and using (2.38), we obtain

$$EU(\bar{\tau}_R, \bar{X}_R(\bar{\tau}_R;t_0,x)) \le U(t_0,x) = (c_1+1)(T-t_0) + V(t_0,x). \tag{2.39}$$

Introduce the events

$$\begin{aligned} \tilde{\Omega}_R \; :&= \{\omega : |\bar{X}_R(t_k;t_0,x)| < R,\ k = 0,\ldots,N-1, \ \text{ and } \ |\bar{X}_R(T;t_0,x)| \le R\} \\ &= \{\omega : (\bar{\tau}_R = T) \setminus \big((\bar{\tau}_R = T) \cap \big(|\bar{X}_R(T;t_0,x)| > R\big)\big)\}, \\ \tilde{\Lambda}_R \; :&= \{\omega : (\exists\ t_k,\ k = 0,\ldots,N-1, \ \text{ such that} \\ &\qquad |\bar{X}_R(t_k;t_0,x)| \ge R) \cup (|\bar{X}_R(T;t_0,x)| > R)\} \\ &= \{\omega : (\bar{\tau}_R < T) \cup \big((\bar{\tau}_R = T) \cap \big(|\bar{X}_R(T;t_0,x)| > R\big)\big)\}. \end{aligned} \tag{2.40}$$

The event $\tilde{\Lambda}_R$ consists of leaving S_R by $\bar{X}_R$ at one of the moments $t_0,\ldots,t_{N-1}$ or leaving $\bar{S}_R$ at $t_N = T$. We note that in the continuous case (see (2.5)) the set $(\tau_R = T) \cap (|X_R(T;t_0,x)| > R)$ is empty.

Let

$$\bar{p}_R = P(\tilde{\Lambda}_R).$$

Analogously to (2.12) and (2.13), we apply Chebyshev's inequality and obtain from (2.39)

$$\bar{p}_R \le \exp(c_0(T-t_0)) \frac{(c_1+1)(T-t_0) + V(t_0,x)}{\min_{t_0 \le t \le T,\ |x| \ge R} V(t,x)}. \tag{2.41}$$

Further, analogously to (2.25), we obtain

$$E|f_R(\bar{X}_R(T;t_0,x))\chi_{\tilde{\Lambda}_R}(\omega)| \le \max_{|x| \le R+\rho} |f(x)|\bar{p}_R. \tag{2.42}$$

We see from the two last inequalities that the expectation $E|f_R(\bar{X}_R(T;t_0,x))\chi_{\tilde{\Lambda}_R}(\omega)|$ is as small as $E|f_R(X_R(T;t_0,x))\chi_{\tilde{\Lambda}_R}(\omega)|$ (cf. (2.25)) if only h ensures (2.38).

THEOREM 2.3. *Consider any method satisfying* (A1) *which is weakly convergent with order* p *for systems with sufficiently smooth and bounded derivatives up to some order. Let the conditions of Propositions* 2.1 *and* 2.2 *and the assumption* (A2) *be fulfilled. Then for any* $x \in R^d$ *and* $\varepsilon > 0$ *there exists* $R(x,\varepsilon) > 0$ *such that for all* $R \ge R(x,\varepsilon)$ *and sufficiently small* h

$$|Ef(X(T;t_0,x)) - E[f(\bar{X}(T;t_0,x))\chi_{\tilde{\Omega}_R}(\omega)]| \le Kh^p + \varepsilon, \tag{2.43}$$

where $K > 0$ *depends on* x *and* R.

Proof. It has been proved (see Proposition 2.2) that for any $\varepsilon > 0$

$$|Ef(X(T;t_0,x)) - Ef_R(X_R(T;t_0,x))| \le \frac{\varepsilon}{2} \tag{2.44}$$

if R is sufficiently large.

Since the coefficients of system (2.18) and the function f_R can be taken so that they have bounded derivatives up to a sufficiently high order, the mentioned method gives for sufficiently small h (see [7])

$$|Ef_R(X_R(T;t_0,x)) - Ef_R(\bar{X}_R(T;t_0,x))| \le Kh^p. \tag{2.45}$$

Let us choose $R(x,\varepsilon)$ so that for $R \geq R(x,\varepsilon)$ and sufficiently small h both inequality (2.44) and inequality

$$E|f_R(\bar{X}_R(T;t_0,x))\chi_{\tilde{\Lambda}_R}(\omega)| \leq \frac{\varepsilon}{2} \tag{2.46}$$

(see (2.41) and (2.42)) are fulfilled.

Since the assumption (A2) holds, $f_R(x) = f(x)$ for $x \in S_R$, and $\bar{X}_R = \bar{X}$ for $\omega \in \tilde{\Omega}_R$, we get

$$\begin{aligned} Ef_R(\bar{X}_R(T;t_0,x)) &= E[f_R(\bar{X}_R(T;t_0,x))\chi_{\tilde{\Omega}_R}(\omega)] + E[f_R(\bar{X}_R(T;t_0,x))\chi_{\tilde{\Lambda}_R}(\omega)] \\ &= E[f(\bar{X}(T;t_0,x))\chi_{\tilde{\Omega}_R}(\omega)] + E[f_R(\bar{X}_R(T;t_0,x))\chi_{\tilde{\Lambda}_R}(\omega)]. \end{aligned} \tag{2.47}$$

Inequality (2.43) follows from (2.44)–(2.47). Theorem 2.3 is proved. □

Remark 2.1. If a method for a particular stochastic system converges, then K in (2.43) is bounded for all R (and ε). However, as was discussed in the Introduction, a method applied to SDEs with nonglobally Lipschitz coefficients can be divergent. It is obvious that in this case K goes to infinity as $R \to \infty$ ($\varepsilon \to 0$). In practice (see, e.g., our experiments and also a comment on the choice of R in section 3), for a not too big R (and, consequently, not large K) the ε is negligibly small since the divergence is usually due to rare exploding approximate trajectories which have to be discarded. This concept of rejecting exploding trajectories is very practical; it allows us, in particular, to guarantee the accuracy of numerical results obtained even by "divergent" methods. We emphasize that the value of K depends on the choice of a numerical method, as is usual in the global Lipschitz case. Thanks to the above concept, we can exploit the whole arsenal of methods [3, 7] and choose an appropriate scheme depending on the system we are solving.

Remark 2.2. It is possible to prove that the proposed concept is also applicable in the case of the Talay–Tubaro extrapolation [12, 7]; i.e., for a sufficiently large R and all sufficiently small h the error can be expanded in powers of h:

$$\begin{aligned} &Ef(X(T;t_0,x)) - E[f(\bar{X}(T;t_0,x))\chi_{\tilde{\Omega}_R}(\omega)] \\ &\quad = \rho(R,h) + C_0h^p + \cdots + C_nh^{p+n} + O(h^{p+n+1})\,, \end{aligned} \tag{2.48}$$

where the constants $C_0,\dots,C_n$ are independent of h and $\rho(R,h) \to 0$ as $R \to \infty$ uniformly with respect to h.

Due to Remark 2.1, ρ is negligibly small for a fixed R in comparison with the term $O(h^{p+n+1})$ (for realistic, not too small h, of course). Therefore it can be supposed that $|\rho| \leq Ch^{p+n+1}$, where C is a positive constant. Then we can use (2.48) in practice to estimate the global error as well as to improve the accuracy of the method [12, 7]. For example, simulating $u = Ef(X(T;t_0,x))$ twice by a first-order scheme (i.e., $p = 1$) with two different time steps $h_1 = h$, $h_2 = \alpha h$, $\alpha > 0$, $\alpha \neq 1$, we obtain $\bar{u}^{h_1} = E[f(\bar{X}^{h_1}(T;t_0,x))\chi_{\tilde{\Omega}_R}(\omega)]$ and $\bar{u}^{h_2} = E[f(\bar{X}^{h_2}(T;t_0,x))\chi_{\tilde{\Omega}_R}(\omega)]$, respectively. We can expand (see (2.48) with $p = 1$, $n = 0$):

$$u = \bar{u}^{h_1} + C_0h_1 + \delta_1\,, \quad u = \bar{u}^{h_2} + C_0h_2 + \delta_2\,,$$

where $|\delta_i| \leq Ch^2$, $i = 1,2$. Hence C_0 can be estimated as $C_0 \simeq -\frac{\bar{u}^{h_2}-\bar{u}^{h_1}}{h_2-h_1}$, and we get the improved value

$$\bar{u}_{imp} = \bar{u}^{h_1}\frac{h_2}{h_2-h_1} - \bar{u}^{h_2}\frac{h_1}{h_2-h_1}, \qquad u = \bar{u}_{imp} + \delta,$$

where $|\delta| \leq Ch^2$.

Example 2.1. Consider the following system:

$$\begin{aligned} dP &= -\nabla F(Q)dt - \nu P dt + \sum_{l=1}^{n} \sigma_l dw_l(t), \\ dQ &= P dt, \end{aligned} \tag{2.49}$$

where ν is a positive constant and σ_l, $l = 1, \dots, n$, are n-dimensional constant linearly independent vectors. The authors of [5] prove exponential ergodicity of (2.49) assuming that $F \in \mathbf{C}^{\infty}(\mathbf{R}^n, \mathbf{R})$, $F(q) \geq 0$ for all $q \in \mathbf{R}^n$ and that there exist an $\alpha > 0$ and $0 < \beta < 1$ such that

$$\frac{1}{2}(\nabla F(q), q) \geq \beta F(q) + \nu^2 \frac{\beta(2-\beta)}{8(1-\beta)} |q|^2 - \alpha. \tag{2.50}$$

The Lyapunov function

$$\begin{aligned} V(x) = V(p,q) &= \frac{1}{2}|p|^2 + F(q) + \frac{\nu}{2}(p,q) + \frac{\nu^2}{4}|q|^2 + 1 \\ &\geq 1 + \frac{1}{8}|p|^2 + \frac{\nu^2}{12}|q|^2 \end{aligned} \tag{2.51}$$

is used to prove that for any $m \geq 1$ there exist positive $c_m,\ d_m$ such that

$$L[V(x)]^m \leq -c_m [V(x)]^m + d_m. \tag{2.52}$$

The exponential ergodicity means that for any function f with a polynomial growth the following inequality holds:

$$\left| Ef(X(t;0,x)) - \int f(z) d\mu(z) \right| \leq Ce^{-\lambda t}, \tag{2.53}$$

where $C > 0$ and $\lambda > 0$ are some constants. In (2.51)–(2.53), $x := (p,q),\ X := (P,Q)$, and μ is an invariant measure for the Markov process defined by (2.49).

Resting on (2.52), it is not difficult to verify all the assumptions of Theorem 2.3 that concern the system under consideration. Due to (2.53), application of this theorem to calculation of $Ef(X(t;0,x))$ gives an approximate value of the ergodic limit. In [6] (see also Example 3.3 below) a numerical example connected with calculation of an ergodic limit is given.

3. Numerical experiments.

3. Numerical experiments. Theorem 2.3 has the following practical implication for evaluating expectations of functionals of solutions to SDEs. We pick up a numerical method suitable for a stochastic system under consideration. We choose $R > 0$ such that the solution of the stochastic system equipped with some initial data leaves the sphere S_R of the radius R during a fixed time interval with a relatively small probability. In a lot of cases interesting from the applicable point of view (e.g., Langevin-type equations and gradient systems with noise) it is usually not difficult to guess this value of R by physical reasoning. Anyway, we can test the choice of R in practice as explained below. For the chosen R, we select a time step h for the numerical method, which ensures an accuracy appropriate for our purposes. As usual, the choice of time step h_* is appropriate if, by further decrease of the time step, we obtain a result which is close enough to the one obtained with h_*. The expectation of a functional which we are aiming to find is evaluated according to the Monte Carlo technique by running M independent realizations of the numerical solution to the considered system. According to the concept proposed in this paper, the value of the functional corresponding to sample trajectories that left the sphere S_R is set to be zero when counted to the expectation. Finally, we say that the choice of R is appropriate if its increase does not essentially affect the result. We also note that there is Monte Carlo error in this procedure, which is controlled in the standard way by choosing an appropriate M. In practice, the procedure can be modified by assigning a certain value (not zero as we do here) for the trajectories which leave the sphere S_R. This value can be chosen/adjusted in response to experimental results or by physical reasoning.

As we will see in the numerical experiments presented below, the numerical integration and Monte Carlo errors affect accuracy of simulation much more than error due to canceling "bad" trajectories (ε in (2.43)), which is usually negligibly small.

Example 3.1. Consider the stochastic differential equation

$$dX = -X^3 dt + \sigma dw(t), \quad X(0) = X_0. \tag{3.1}$$

It is demonstrated in [5] (see also [9, 11]) that the explicit Euler method for (3.1),

$$X_{k+1} = X_k - X_k^3 h + \sigma \Delta_k w, \qquad \Delta_k w := w(t_{k+1}) - w(t_k), \tag{3.2}$$

can explode.

For test purposes, we evaluate the functional

$$F = \frac{1}{2} E X^2(T) + E \int_0^T X^4(t) dt. \tag{3.3}$$

It can be shown that

$$F = \frac{1}{2}\sigma^2 T.$$

To simulate this functional, we introduce the additional equation

$$dZ = X^4(t) dt, \qquad Z(0) = 0. \tag{3.4}$$

Then

$$F = E\left(\frac{1}{2}X^2(T) + Z(T)\right). \tag{3.5}$$

The solution of (3.4) is approximated as

$$Z_{k+1} = Z_k + X_k^4 h. \tag{3.6}$$

By taking $V(x,z) = x^6 + z^2$, it is not difficult to check that the conditions of Propositions 2.1 and 2.2 are satisfied for the system (3.1), (3.4). Also, the condition (A2) holds for the explicit Euler method (3.2). Then Theorem 2.3 is applicable here; i.e., we can evaluate F from (3.3) by using approximate trajectories $(\bar{X}(t), \bar{Z}(t))$, $0 \le t \le T$, which belong to the ball $\{(x,z) : x^2 + z^2 < R^2\}$. In fact, in the case of functionals like that in (3.3) it is enough to control the paths $\bar{X}(t)$ only; i.e., the following estimate takes place (cf. (2.43)):

$$\left| E\left(\frac{1}{2}X^2(T) + Z(T)\right) - E\left(\frac{1}{2}\bar{X}^2(T) + \bar{Z}(T)\right)\chi_{\tilde{\Omega}_R}(\omega)\right| \le Kh + \varepsilon, \tag{3.7}$$

where $\tilde{\Omega}_R$ is defined by $\bar{\tau}_R$ being the first exit time of $(t, \bar{X}(t))$ from the rectangle $[0,T) \times (-R,R)$. This result is valid thanks to the fact that the right-hand sides of (3.1), (3.4) do not depend on Z. The proof is almost a word-by-word repetition of the proof of Theorem 2.3.

We also consider the weak Euler method:

$$X_{k+1} = X_k - X_k^3 h + \sigma \xi_k \sqrt{h}, \tag{3.8}$$

where ξ_k are i.i.d. (independently and identically distributed) random variables with the law $P(\xi = \pm 1) = 1/2$.

Let us choose a time step $h > 0$ for (3.8) such that

$$|X_0| \le \frac{1}{\sqrt{h}} \quad \text{and} \quad h < \frac{1}{\sigma}\left(1 - \frac{2}{3\sqrt{3}}\right).$$

Then one can directly show that $|X_1| \le \frac{1}{\sqrt{h}}$ and therefore $|X_k| \le \frac{1}{\sqrt{h}}$ for all k. Thus, trajectories of (3.8) do not explode, provided that the above conditions on the step h hold. The authors do not exclude a possibility that methods using bounded random variables (like the weak Euler method (3.8)) can weakly converge in some nonglobally Lipschitz cases. In general this question concerning convergence is rather complicated (see, e.g., the third example below and also [6]) and requires further investigation. We should stress that convergence of methods does not undermine the concept proposed in this paper. Indeed, suppose that a weak method converges but for a not very small time step it may have exploding trajectories; then results obtained with this time step should be disregarded unless this concept is applied. This is well illustrated in our examples. Further, the concept is universal. It allows us to use any numerical method in the nonglobally Lipschitz case straightaway for a very broad class of SDEs, without any additional analysis at all.

In our experiments we simulate F from (3.5) as follows:

$$\bar{F} = \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{2} \left[\bar{X}^{(m)}(T) \right]^2 + \bar{Z}^{(m)}(T) \right) \chi_{\tilde{\Omega}_R}(\omega) + \rho_{mc}, \tag{3.9}$$

where M is the number of independent realizations $\bar{X}^{(m)}(T)$, $\bar{Z}^{(m)}(T)$ of $\bar{X}(T)$, $\bar{Z}(T)$ that are found due to a numerical method of our choice. The Monte Carlo error ρ_{mc} has zero bias, and its variance equals

$$Var(\rho_{mc}) = \frac{Var\ \left(\left(\frac{1}{2}\bar{X}^2(T) + \bar{Z}(T) \right) \chi_{\tilde{\Omega}_R}(\omega) \right)}{M}; \tag{3.10}$$

i.e., the simulated

$$\hat{F} := \frac{1}{M} \sum_{m=1}^{M} \left(\frac{1}{2} \left[\bar{X}^{(m)}(T) \right]^2 + \bar{Z}^{(m)}(T) \right) \chi_{\tilde{\Omega}_R}(\omega)$$

belongs to the confidence interval

$$\hat{F} \in (E\bar{F} - c\sqrt{Var(\rho_{mc})}, E\bar{F} + c\sqrt{Var(\rho_{mc})}) \tag{3.11}$$

with the fiducial probability, for example, 0.997 for $c = 3$ and 0.95 for $c = 2$. For definiteness, we set $c = 2$ here.

Table 1

Simulation of (3.1), (3.4) *by the Euler methods* (3.2), (3.6) *and* (3.8), (3.6) *with various time steps h. See the other parameters in the text. The exact value $F = 5$. The "±" reflects the Monte Carlo error only; it does not reflect the error of the methods.*

h	(3.2), (3.6) $\bar{F}$	(3.2), (3.6) Trajectories left $(-R, R)$	(3.8), (3.6) $\bar{F}$
0.25	6.640 ± 0.010	0.03%	5.962 ± 0.006
0.1	5.409 ± 0.007	0%	5.371 ± 0.007
0.02	5.069 ± 0.007	0%	5.073 ± 0.007
0.01	5.032 ± 0.007	0%	5.037 ± 0.007

In Table 1 some results of our numerical experiments are presented. We take $\sigma = 1$, $X_0 = 0$, $M = 400000$, $T = 10$, and $R = 50$. The "±" reflects the Monte Carlo error only; it gives the confidence interval with $c = 2$ (see (3.11)). If our concept were not applied in the case of the Euler method (3.2), (3.6) with $h = 0.25$, then there would be an overflow in computer calculations. We also note in passing that both Euler methods produce quite similar results. Of course, the Euler method (3.2) is computationally more expensive than (3.8) due to the need to simulate Gaussian random variables instead of very simple random variables for (3.8).

Example 3.2. As the second test model, we choose the equation

$$dX = -X \exp(X^2) dt + dw(t), \quad X(0) = X_0, \tag{3.12}$$

and evaluate $EX^2(t)$. See also some experiments and discussion concerning (3.12) in [5].

By taking $V(x) = x^4$, it is easy to check that the conditions of Propositions 2.1 and 2.2 are satisfied for this equation. Therefore Theorem 2.3 is applicable again for methods satisfying the condition (A2). Further, (3.12) is exponentially ergodic, and the second moment evaluated with respect to the invariant measure is equal to 0.2539 up to 4 decimal points (d.p.).

We simulate (3.12) by the explicit Euler method

$$X_{k+1} = X_k - X_k \exp(X_k^2)\, h + \Delta_k w \tag{3.13}$$

on the time interval $[0, 100]$ with $X_0 = 0$. In our numerical experiments we choose $R = 5$ (we note that the force $-x \exp(x^2)$ produces a very sharp "barrier"); the number of independent realizations $M = 400000$, and as the result the Monte Carlo error $2\sqrt{Var(\rho_{mc})} \leq 0.001$. Figure 1 gives the averaged trajectories of $EX^2(t)$ for various time steps h. For $h = 0.2$, there are 356390 trajectories (i.e., 89%) that leave the interval $(-5, 5)$. Obviously, the obtained result cannot be considered reliable, and it is not presented in Figure 1. For $h = 0.1$, we have 36676 trajectories (i.e., 9%) that leave the interval $(-5, 5)$; for $h = 0.05$, 130 trajectories (i.e., 0.03%); and for $h = 0.02$, there are no trajectories out of 400000 that leave the interval $(-5, 5)$. An increase of R has almost no effect on the results. This indicates that our choice of R is appropriate. We see from Figure 1 that for $h = 0.05$ and $h = 0.02$ we obtain a quite good approximation of the ergodic limit. At the same time, we note that if our concept were not applied in this experiment, then the average trajectories for $h = 0.1$ and $h = 0.05$ would blow up. In the case of $h = 0.02$ we have not observed exploding trajectories, but this does not mean that if we continued the experiment further, we would not observe any exploding trajectories that lead to blow-up of the average trajectories. Further, we should note that the weak Euler method (cf. (3.8)) applied

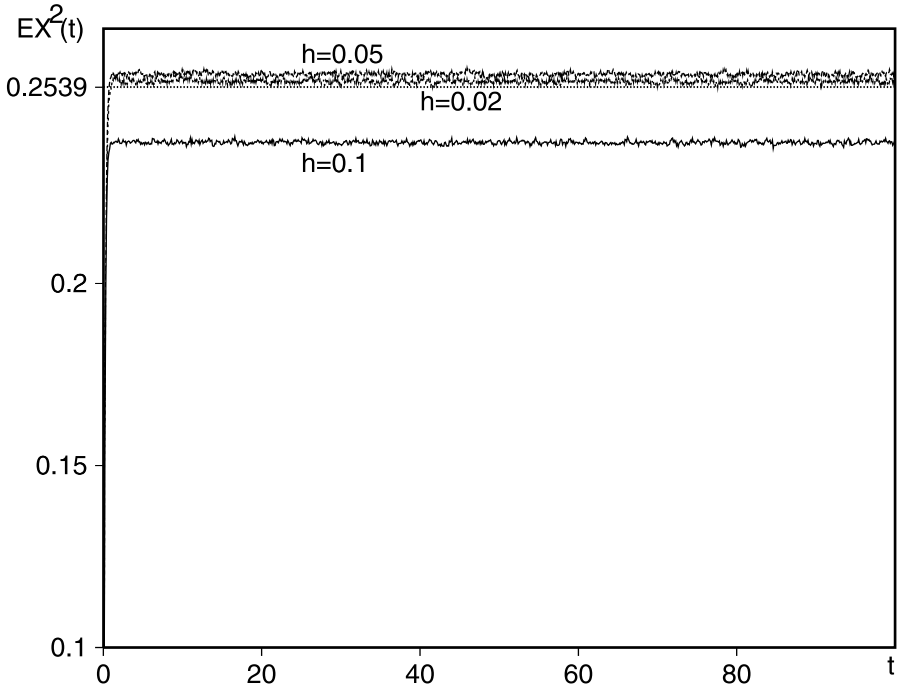


FIG. 1. *Result of simulations of* (3.12) *by the Euler method* (3.13) *with various time steps* h. *We take* $R = 5$; *see the other parameters in the text.*

to (3.12) does not explode even for $h = 0.2$. This can be explained by arguments similar to those used in the first example.

Example 3.3. Consider the oscillator with cubic restoring force and additive noise

$$dQ = Pdt,\qquad dP = \left(Q - Q^3\right)dt - \nu Pdt + \sigma dw(t), \tag{3.14}$$

where $w(t)$ is a standard Wiener process and ν and σ are positive constants. This system is exponentially ergodic, and the second moment EQ^2 evaluated with respect to the invariant measure is equal to 2.435 up to 3 d.p.

The system (3.14) belongs to the class of Langevin equations, for which quasi-symplectic methods are the most effective [6] (see also [7]). We apply an explicit quasi-symplectic method of weak order one to (3.14):

$$P_{k+1} = (1 - \nu h)\left(P_k + h\left(Q_k - Q_k^3\right) + h^{1/2}\sigma\xi_k\right),\qquad Q_{k+1} = Q_k + h\left(P_k + h\left(Q_k - Q_k^3\right)\right), \tag{3.15}$$

where ξ_k are i.i.d. random variables with the law $P(\xi = \pm 1) = 1/2$.

The results of simulating $EQ^2(t)$ by this method are presented in Figure 2. We take the parameters of (3.14) as follows: $Q(0) = P(0) = 0$, $\nu = 0.05$, and $\sigma = 1$. For realization of the proposed concept, we choose $R = 50$. The number of independent realizations M used to produce the picture is equal to 400000, which ensures the Monte Carlo error $2\sqrt{Var(\rho_{mc})} \leq 0.008$. For $h = 0.2$, there are 28 trajectories (i.e., 0.007%) that leave the ball of radius 50. We see that, applying the proposed concept

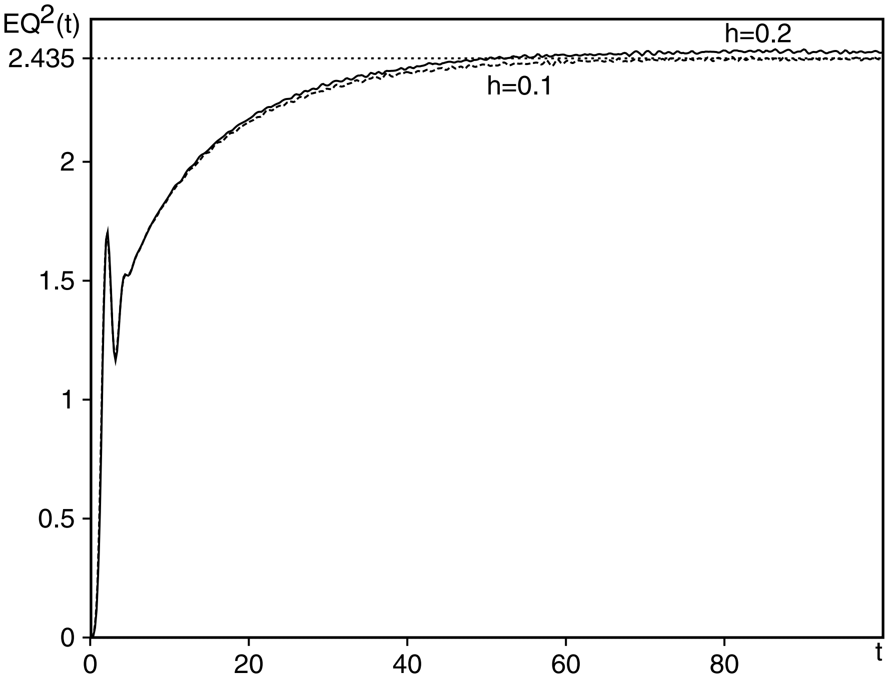


FIG. 2. *Result of simulations of* (3.14) *by the quasi-symplectic method* (3.15) *with various time steps h. We take $R = 50$; see the other parameters in the text.*

(i.e., not taking into account these 28 “bad” trajectories), we obtain a quite accurate approximation of the ergodic limit. If one did not exploit the concept here, then the results obtained with $h = 0.2$ could not be used, since the exploding trajectories lead to numerical overflow in computing the average. For $h = 0.1$, there are no trajectories out of 400000 that leave the ball of radius 50, but this does not exclude the possibility of having exploding trajectories in another series of Monte Carlo runs. Such uncertainty made it uncomfortable to use the results of such experiments before the concept developed in this paper. This concept gives us a rigorous basis for making use of any numerical method to solve nonlinear SDEs and for interpreting experiments in which occurrence of exploding trajectories is not excluded. Some other experiments with the model (3.14) are available in [6] and [5]. We also note that a further development of the concept of rejecting exploding trajectories specifically for calculation of ergodic limits will be addressed in a separate publication.

REFERENCES

[1] D. J. HIGHAM, X. MAO, AND A. M. STUART, *Strong convergence of Euler-type methods for nonlinear stochastic differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1041–1063.

[2] R. Z. HASMINSKII, *Stochastic Stability of Differential Equations*, Sijthoff & Noordhoff, Groningen, The Netherlands, 1980.

[3] P. E. KLOEDEN AND E. PLATEN, *Numerical Solution of Stochastic Differential Equations*, Springer, Berlin, 1992.

[4] H. LAMBA, J. C. MATTINGLY, AND A. M. STUART, *An Adaptive Euler–Maruyama Scheme for SDEs: Convergence and Stability*, working paper, Warwick University (England), 2005.

[5] J. C. MATTINGLY, A. M. STUART, AND D. J. HIGHAM, *Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise*, Stochastic Process. Appl., 101 (2002), pp. 185–232.

[6] G. N. Milstein and M. V. Tretyakov, *Quasi-symplectic methods for Langevin-type equations*, IMA J. Numer. Anal., 23 (2003), pp. 593–626.

[7] G. N. Milstein and M. V. Tretyakov, *Stochastic Numerics for Mathematical Physics*, Springer, New York, 2004.

[8] M. B. Nevelson and R. Z. Hasminskii, *Stochastic Approximation and Recursive Estimation*, Transl. Math. Monogr. 47, American Mathematical Society, Providence, RI, 1973.

[9] G. O. Roberts and R. L. Tweedie, *Exponential convergence of Langevin distributions and their discrete approximations*, Bernoulli, 2 (1996), pp. 341–363.

[10] C. Soize, *The Fokker–Planck Equation for Stochastic Dynamical Systems and Its Explicit Steady State Solutions*, World Scientific, London, 1994.

[11] D. Talay, *Stochastic Hamiltonian systems: Exponential convergence to the invariant measure, and discretization by the implicit Euler scheme*, Markov Process. Related Fields, 8 (2002), pp. 163–198.

[12] D. Talay and L. Tubaro, *Expansion of the global error for numerical schemes solving stochastic differential equations*, Stochastic Anal. Appl., 8 (1990), pp. 483–509.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1155–1174

VARIABLE ACCURACY OF MATRIX-VECTOR PRODUCTS IN PROJECTION METHODS FOR EIGENCOMPUTATION*

V. SIMONCINI†

Abstract. We analyze the behavior of projection-type schemes, such as the Arnoldi and Lanczos methods, for the approximation of a few eigenvalues and eigenvectors of a matrix A, when A cannot be applied exactly but only with a possibly large perturbation. This occurs, for instance, in shift-and-invert procedures or when dealing with large generalized eigenvalue problems. We theoretically show that the accuracy with which A is applied at each iteration can be relaxed, as convergence to specific eigenpairs takes place. We show that the size of the perturbation is allowed to be inversely proportional to the current residual norm, in a way that also depends on the sensitivity of the matrix A. This result provides a complete understanding of reported experimental evidence in the recent literature. Moreover, we adapt our theoretical results to devise a practical relaxation criterion to achieve convergence of the inexact procedure. Numerical experiments validate our analysis.

Key words. eigenvalues, projection methods, Krylov subspaces, inexact methods, Ritz values

AMS subject classifications. 65F15, 65F50, 65N25

DOI. 10.1137/040605333

1. Introduction.

We are interested in the behavior of projection-type procedures, such as the Arnoldi or Lanczos methods (see, e.g., [18, 2]), for the approximation of a few of the eigenvalues and corresponding eigenvectors in the eigenvalue problem

$$Ax = \lambda x, \quad \|x\| = 1, \tag{1.1}$$

where A is an $n \times n$ non-Hermitian matrix and $\|\cdot\|$ is the Euclidean norm.We focus on the case in which A cannot be applied exactly, but only with a perturbation; that is, at each iteration the operation $y = Av$ is replaced by

$$y = Av + f, \tag{1.2}$$

where f can change at each iteration and $\|f\|$ can be monitored. In general, we expect $\|f\|$ to be much larger than machine precision, so that the standard techniques of round-off error analysis are not appropriate. This is indeed the case when, for instance, shift-and-invert procedures are used to find interior eigenvalues of the given matrix, or when a generalized eigenvalue problem is considered. On large problems, both these procedures require the (approximate) solution of a linear system before applying the matrix A. Finally, inaccurate products occur when the matrix itself is an operator that needs to be estimated each time it is applied. As an alternative to a fixed perturbation tolerance, methods based on shift-and-invert power iterations have for a long time focused on increasing the accuracy as convergence was taking place; see, e.g., [9, 27] and [2, section 11.2] and references therein; see also [29] for a recent analysis of perturbed power iterations. On the other hand, in the context of projection-type methods, more recently the case of *decreasing* accuracy has been considered

*Received by the editors March 17, 2004; accepted for publication (in revised form) January 22, 2005; published electronically September 23, 2005.
http://www.siam.org/journals/sinum/43-3/60533.html

†Dipartimento di Matematica, Università di Bologna, Piazza di Porta S. Donato, 5, I-40127 Bologna, Italy; IMATI-CNR, Pavia and CIRSA, Ravenna, Italy (valeria@dm.unibo.it).

[5, 21, 7, 15]. We notice that an analogous problem has received considerable attention in the linear system setting; see, e.g., [4, 6, 32, 22, 25].

An approximate matrix-vector multiplication significantly perturbs the method, but in a way that is apparently far less dramatic than the norm of the perturbation would suggest. A large number of experiments in [5] showed that in many cases convergence towards eigenpairs of A can be achieved despite the fact that $\|f\|$ is allowed to grow as the iteration progresses. In other words, it was shown in [5] that the accuracy with which A is applied at each iteration can be relaxed as convergence takes place. It was argued that the size of the perturbation can be related to the inverse of the residual norm of the current eigenvalue approximation. However, these arguments were not supported by theoretical justifications in [5]; the not always consistent behavior of the methods under the analyzed perturbations did not allow the authors to make conclusive statements as to the reliability of using variable accuracy. We aim to fill this gap in this paper.

The aim of this paper is twofold. First, we provide a theoretical understanding of the experimental evidence reported in [5], together with an analysis of the spectral properties that may influence the inexact process. Then, we adapt our theoretical results to devise a practical relaxation criterion to achieve convergence of the inexact procedure. Assuming that the unperturbed process converges, given an approximate eigenpair (θ, z) obtained by the perturbed process, we will show that the deviation of the computable residual from the true (unobservable) residual $Az - \theta z$ can be kept below a fixed small tolerance. This result will imply that the final attainable true residual will also fall below the required tolerance.

Our analysis of variable accuracy in the matrix-vector products includes both Ritz and harmonic Ritz approximations obtained with the Arnoldi method, as well as Ritz spectral information computed by the Lanczos procedure. Although these methods usually provide different approximation quantities and satisfy different optimality properties (see, e.g., [2]), the analysis of their performance under matrix-vector perturbations can be unified within the presented framework.

In this paper we restrict our analysis to the case when the considered method is not restarted. In practice, projection-type schemes require possibly sophisticated restarting procedures such as *implicit restarting* (see [28] and also [10]), to limit memory requirements while improving the current available approximation. The problem of handling inexactness when some form of restarting is included will be the topic of future research.

The key idea beyond the success of inexact processes is related to an intrinsic property of Krylov subspace methods. Approximations are generated as $V_m u$, where the m columns of the matrix V_m span the Krylov subspace of dimension m, and u is the approximate solution of the projected problem. The perturbations f's affecting the matrix-vector operation in (1.2) at each iteration may be collected as subsequent columns of a "perturbation matrix" F_m. It turns out that the fundamental relation associated with the inexact process formally differs from the unperturbed one by an additional term $F_m u$. Ideally, if one is able to show that the components of u decrease with m—that is, for instance, that the ith component of u goes towards zero as ρ_m^i for some $0 < \rho_m < 1$—then the norm of the corresponding columns of F_m is allowed to grow, that is, larger perturbations are allowed in later iterations, while still yielding a small perturbation term $F_m u$ in norm. As a result, the perturbed fundamental relation remains significantly close to the unperturbed one, and the approximation process does not appear to be influenced. Therefore, given a problem to be solved by means of projection onto a Krylov subspace, the inexact matrix-vector product in (1.2) can be conveniently exploited if one can show that the approximation vector u has a decreasing pattern. Intuitively, a decreasing pattern is associated with a "marginal approximation" property, as the Krylov subspace grows. However, a rigorous treatment of this step is far from obvious. As one may suspect, both the difficulty in proving the existence of this pattern, as well as the constraints under which this pattern does settle down, depend on the possible complexity and nonlinearity of the problem. In the linear system setting, the decreasing pattern of u (in this case the projected system approximate solution) was proved in [22, 32]. In this paper, we solve this problem within eigenvalue computation. We will show that the matrix sensitivity and the nonlinearity of the problem play a crucial role: the conditions under which the pattern arises are different and significantly more stringent than in the linear system setting. Moreover, the analysis becomes even more complex when more than one eigenpair, or more generally an invariant subspace, is sought. Finally, deep results from matrix perturbation theory need to be employed, making the approach and the conclusions of this paper significantly different from what has been done in [22, 32] for linear systems.

In section 2 we show that some of the eigenvectors of the representation matrix of A in the approximation subspace have a decreasing pattern. This key result will be used in section 3 to show that, in the inexact Arnoldi method, the matrix-vector multiplication can be perturbed in a way that is inversely proportional to this pattern. A practical relaxation strategy for the Arnoldi method is proposed in section 3.1 and numerically tested in section 4. Our theoretical results are subsequently applied to related methods that are currently used as alternatives to general Arnoldi and that are based on the same key relations [2]. In particular, we will discuss the inexact harmonic Ritz approximations in section 5, and the inexact Lanczos method in section 6. Finally, section 7 summarizes our results and discusses some related issues.

The following notation will be used throughout. For a vector u, $\bar{u}$ denotes its conjugate, u^* its conjugate transpose, and $\|u\|$ its 2-norm. For a given $k \times k$ matrix T, an eigenpair (λ, u) consists of $\lambda \in \mathbb{C}$ and $u \in \mathbb{C}^k$ such that $Tu = \lambda u$, $\|u\| = 1$. An eigentriple (λ, u, v) of T is such that $Tu = \lambda u$ and $v^*T = \lambda v^*$ such that $\|u\| = 1$ and $v^*u = 1$. Here u indicates a right eigenvector and v a left eigenvector. Moreover, I_k denotes the identity matrix of size k (the subscript is omitted if clear from the context), while e_j is the jth column of the identity matrix of given dimension not smaller than j. Finally, Range(X) is the space generated by the columns of X, while $\Lambda(H)$ is the set of eigenvalues of a square matrix H. Exact precision arithmetic is assumed throughout the paper, and the term inexact refers to an inaccurate computation, whose error is significantly larger than the finite precision arithmetic unit. All experiments were run using MATLAB [11].

2. Bounds for the eigenvector components of the Arnoldi matrix.

2.1. Notation. Starting with a unit norm vector v_1, the Arnoldi method builds a basis V_m for the Krylov subspace $K_m(A, v_1) = \text{span}\{v_1, Av_1, \ldots, A^{m-1}v_1\}$ satisfying the following *Arnoldi relation*:

$$AV_m = V_m H_m + v_{m+1} h_{m+1,m} e_m^* = V_{m+1} \begin{bmatrix} H_m \\ h_{m+1,m} e_m^* \end{bmatrix}, \quad V_{m+1}^* V_{m+1} = I. \tag{2.1}$$

Matrix H_m is an $m \times m$ upper Hessenberg matrix, and it is the orthogonal projection and restriction of the matrix A onto the Krylov subspace. An eigenpair (θ, u) of H_m

defines the Ritz value, θ, and Ritz vector, $V_m u$, which may be used to approximate some of the eigenpairs of A [2, 18]. The accuracy in the approximation is usually monitored by means of some relative quantity involving the residual $r_m = AV_m u - \theta V_m u$; we refer to [2] and its references for a detailed discussion on stopping criteria for eigenvalue solvers. It is also important to recall that for ill-conditioned problems, small residuals do not necessarily imply small errors in the approximate eigenpair [8]; additional care should be taken in this case.

Given the eigenpair (θ, u) and using (2.1), the residual r_m and its norm can be cheaply computed as

$$r_m = v_{m+1} h_{m+1,m}\, e_m^* u, \qquad \|r_m\| = |h_{m+1,m}|\, |e_m^* u|. \tag{2.2}$$

This relation emphasizes that for the residual to be small, at least one of $h_{m+1,m}$ or $|e_m^* u|$ have to be small. In the former case, the Krylov subspace is close to an invariant subspace. On the other hand, a small $|e_m^* u|$ indicates that the mth component of the eigenvector u of H_m is small. No other knowledge of the eigenvector components is commonly employed in the convergence test, although it can be experimentally observed that the absolute values of the components of u tend to exhibit a decreasing pattern if $(\theta, V_m u)$ is a good approximation to an eigenpair of A; see, e.g., [17, 19].

In Proposition 2.2 we will show that there exists a strong relation between the magnitude of the $(k+1)$st component of u and the residual of some Ritz pair after k Arnoldi iterations, with $k < m$. This relation can be derived as a consequence of a general approximation theorem for nonnormal matrices; cf., e.g., [30]. Below we report the result with our notation. To this end, we introduce some definitions. Given an orthogonal basis U for an invariant subspace of a matrix H, and given Y so that $[U, Y]$ is unitary, Range(U) is called a simple invariant subspace of H if the spectra of $U^* H U$ and of $Y^* H Y$ do not intersect; cf. [30, Definition V.1.2]. Moreover, for two square matrices L_1, L_2 with disjoint spectra, the function $\mathrm{sep}(L_1, L_2)$ is defined as (see, e.g., [30, p. 231])

$$\mathrm{sep}(L_1, L_2) := \inf_{\|P\|=1} \|PL_1 - L_2 P\|.$$

The definition holds for $\|\cdot\|$ being any consistent family of norms; in the following we shall use the 2-norm for vectors and the induced 2-norm for matrices. If L_1 is a scalar, then $\mathrm{sep}(L_1, L_2) = \sigma_{\min}(L_2 - L_1 I)$. An analogous relation holds whenever L_2 is a scalar.

Theorem 2.1 (see [30, Theorem V.2.1, p. 230]). *Let $\mathcal{X} = [U, Y]$ be a unitary matrix and let*

$$\mathcal{X}^* H_m \mathcal{X} = \begin{bmatrix} L_1 & K \\ G & L_2 \end{bmatrix}.$$

Set $\gamma = \|G\|$, $\eta = \|K\|$. Assume that L_1 and L_2 have distinct spectra, so that $\delta := \mathrm{sep}(L_1, L_2) > 0$. Then if $\gamma\eta/\delta^2 < \frac{1}{4}$, there is a unique matrix P satisfying $\|P\| < 2\frac{\gamma}{\delta}$ such that the columns of $\widetilde{U} = (U + YP)(I + P^*P)^{-\frac{1}{2}}$ form an orthonormal basis of a simple right invariant subspace of H_m. The representation of the matrix H_m with respect to $\widetilde{U}$ is given by $\widetilde{L}_1 = (I + P^*P)^{\frac{1}{2}}(L_1 + KP)(I + P^*P)^{-\frac{1}{2}}$.

2.2. Spectral properties of the Arnoldi matrix.

We first use the result of Theorem 2.1 in the case of the approximation of one eigenpair, and then generalize it to the approximation of an invariant subspace.

Consider the principal submatrix of H_m of size k, H_k, i.e.,

$$H_m = \left[\begin{array}{cc} H_k & H_\star \\ h_{k+1,k}e_1e_k^* & * \end{array} \right], \quad H_k \in \mathbb{C}^{k\times k}, \tag{2.3}$$

and let $u^{(k)}$ be an eigenvector of H_k. Let Y be a matrix such that the matrix

$$\mathcal{X} = \left[\left[\begin{array}{c} u^{(k)} \\ \underline{0} \end{array} \right], Y \right] \in \mathbb{C}^{m\times m}$$

is unitary, where here and in the following, $\underline{0}$ pads with zeros the bottom part of a vector with a total of m components. Then

$$\underline{H_m} := Y^* H_m Y \in \mathbb{C}^{(m-1)\times(m-1)} \tag{2.4}$$

is the orthogonal projection and restriction of H_m onto the range of Y, the space orthogonal to the space spanned by $[\begin{smallmatrix} u^{(k)} \\ \underline{0} \end{smallmatrix}]$.

Proposition 2.2. *Let $(\theta^{(k)}, u^{(k)})$ be an eigenpair of H_k, $r_k = v_{k+1}h_{k+1,k}e_k^*u^{(k)}$, $\delta_{m,k} = \sigma_{\min}(\underline{H_m} - \theta^{(k)}I) > 0$ with $\underline{H_m}$ defined in* (2.4)*, and $s_m^* = [(u^{(k)})^*, \underline{0}^*]H_m - \theta^{(k)}[(u^{(k)})^*, \underline{0}^*]$. If*

$$\|r_k\| < \frac{\delta_{m,k}^2}{4\|s_m\|}, \tag{2.5}$$

then there exists a unit norm eigenvector $u = [\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}]$ of H_m with $u_1 \in \mathbb{C}^k$ such that

$$\|u_2\| \le \frac{\tau}{\sqrt{1+\tau^2}}, \quad \textit{with} \quad \tau \in \mathbb{R}, \quad 0 \le \tau < 2\frac{\|r_k\|}{\delta_{m,k}}. \tag{2.6}$$

Moreover, if θ is the eigenvalue associated with u, we have

$$|\theta - \theta^{(k)}| \le \|s_m\|\tau. \tag{2.7}$$

Proof. Let $Y = [\begin{smallmatrix} Y_1 \\ Y_2 \end{smallmatrix}]$ be such that the matrix $\mathcal{X} = [[\begin{smallmatrix} u^{(k)} \\ \underline{0} \end{smallmatrix}], Y] \in \mathbb{C}^{m\times m}$ is unitary. Note that this implies $Y_1^*u^{(k)} = 0$ and $Y_2Y_2^* = I$. Using (2.3) gives

$$\mathcal{X}^*H_m\mathcal{X} = \left[\begin{array}{cc} \theta^{(k)} & K \\ G & \underline{H_m} \end{array} \right], \qquad \begin{array}{l} G = Y_2^*h_{k+1,k}e_1e_k^*u^{(k)}, \\ K = (u^{(k)})^*\,[H_k, H_\star]\,Y. \end{array}$$

Since Y_2 has orthonormal rows,

$$\gamma := \|G\| = \|Y_2^*h_{k+1,k}e_1e_k^*u^{(k)}\| = |h_{k+1,k}e_k^*u^{(k)}| = \|r_k\|. \tag{2.8}$$

Moreover, since $[(u^{(k)})^*, \underline{0}^*]Y = 0$, $\|K\| = \|[(u^{(k)})^*, \underline{0}^*]\,H_mY\| = \|s_m^*Y\| \le \|s_m\|$. Using Theorem 2.1, if $\frac{\gamma\|s_m\|}{\delta_{m,k}^2} < \frac{1}{4}$, i.e., if (2.5) holds, then there exists a vector $p \in \mathbb{C}^{m-1}$ satisfying

$$\tau := \|p\| < 2\frac{\gamma}{\delta_{m,k}}$$

such that the unit norm vector

$$u = \left[\begin{array}{c} u_1 \\ u_2 \end{array} \right] = \frac{1}{\sqrt{1+\|p\|^2}} \left(\left[\begin{array}{c} u^{(k)} \\ \underline{0} \end{array} \right] + \left[\begin{array}{c} Y_1 \\ Y_2 \end{array} \right] p \right)$$

is an eigenvector of H_m. To complete the proof, we notice that

$$\|u_2\| = \frac{1}{\sqrt{1+\|p\|^2}}\|Y_2 p\| \le \frac{1}{\sqrt{1+\|p\|^2}}\|p\|.$$

The bound for $|\theta - \theta^{(k)}|$ follows from the representation of H_m with respect to u in Theorem 2.1, that is, $|\theta - \theta^{(k)}| = \|Kp\| \le \|K\|\,\|p\| \le \|s_m\|\,\tau$. □

Proposition 2.2 shows that if the residual r_k is sufficiently small, then there exists an eigenvector u of H_m whose last $m-k$ components can be bounded by a quantity involving $\|r_k\|$, and thus they are also small. Each of the last $m-k$ components of u is bounded by $\tau/\sqrt{1+\tau^2}$, and hence

$$|e_j^* u| \le \frac{\tau}{\sqrt{1+\tau^2}} \le \tau \le 2\frac{\|r_k\|}{\delta_{m,k}}, \quad j = k+1,\dots,m. \tag{2.9}$$

The bound (2.9) is most interesting for $j = k+1$. Indeed, Proposition 2.2 can be applied to other principal submatrices of H_m, of size k_1 larger than k. In this case, and if a more accurate Ritz pair is computed with H_{k_1}, the last $m-k_1$ components of the corresponding eigenvector are smaller than those for H_k. In practice, this implies that for k sufficiently large, so that $\|r_k\|$ satisfies (2.5), the last $m-k$ components of u have an almost monotonically decreasing pattern.

It is remarkable that, in our setting, $\|r_k\|$ has a double role. On the one hand, $\|r_k\| = \|G\|$ (cf. (2.8)), which measures the accuracy of

$$\left(\theta^{(k)}, \begin{bmatrix} u^{(k)} \\ \underline{0} \end{bmatrix}\right) \tag{2.10}$$

as an approximate eigenpair of H_m in connection with the general approximation theorem. On the other hand,

$$r_k = v_{k+1}h_{k+1,k}e_k^* u^{(k)} = AV_k u^{(k)} - \theta^{(k)} V_k u^{(k)}$$

is the residual of $(\theta^{(k)}, V_k u^{(k)})$ as approximate eigenpair of A. This double role is what makes our eigenvector component analysis possible.

In Proposition 2.2, s_m is the *left* residual vector of the approximate pair (2.10) of H_m. For nonnormal problems, $\|s_m\|$ is bound not to be small, and in general the estimate $\|s_m\| \approx \|A\|$ can be used.

Remark 2.3. The function $\delta_{m,k}$ provides a condition number of the eigenvector problem [30, p. 241], and it reflects the proximity of $\theta^{(k)}$ to $\underline{H_m}$, although it may be much smaller than the distance between the spectrum of $\underline{H_m}$ and $\theta^{(k)}$ for nonnormal matrices [30, Example 2.4, p. 234]. We notice that if (2.10) is a good approximation to the eigenpair (θ, u) of H_m, then $\delta_{m,k}$ is close to the norm of the reduced resolvent of H_m. These comments will be used when we derive a practical relaxation criterion for the matrix-vector product with A. □

Remark 2.4. Proposition 2.2 states the *existence* of an eigenvector u of H_m satisfying (2.6). On the other hand, we are interested in the characterization of the components of a specific eigenvector u of H_m. To be able to correctly identify the eigenvector determined by Proposition 2.2 as the analyzed vector u, we will need some further hypotheses. In particular, we will require that after k iterations, $\theta^{(k)}$ and $u^{(k)}$ provide sufficiently good approximations to an eigenpair of H_m and of A. Unfortunately this apparently restrictive condition is observed to be required in practice.

In particular, eigenvector components do not exhibit a decreasing pattern until the Ritz pair reaches its final, possibly superlinear, asymptotic convergence rate; see [3] for a discussion on the occurrence of superlinear convergence rates. □

We next generalize our result to the case of an invariant subspace.

PROPOSITION 2.5. *Let the columns of $U^{(k)}$ be an orthonormal basis for a simple invariant subspace of H_k of dimension ℓ, with representation matrix $L^{(k)} = (U^{(k)})^* H_k U^{(k)}$. Let $\delta_{m,k} = \mathrm{sep}(L^{(k)}, L_2) > 0$, $R_k = v_{k+1} h_{k+1,k} e_k^* U^{(k)}$, and $S_m = [(U^{(k)})^*, O^*] H_m - L^{(k)} [(U^{(k)})^*, O^*]$, where O is the $(m-k) \times \ell$ zero matrix. If $\|R_k\| < \frac{\delta_{k,m}^2}{4\|S_m\|}$, then there exists a matrix $U = [\begin{smallmatrix} U_1 \\ U_2 \end{smallmatrix}]$, $U^*U = I$, with $U_1 \in \mathbb{C}^{k \times \ell}$ and whose columns span a simple invariant subspace of H_m, such that*

$$\|U_2\| \leq \frac{\tau}{\sqrt{1+\tau^2}}, \quad \textit{with} \quad \tau \in \mathbb{R}, \quad 0 \leq \tau < 2\frac{\|R_k\|}{\delta_{k,m}}. \tag{2.11}$$

Proof. The proof follows the lines of that of Proposition 2.2. □

The setting of Proposition 2.5 is most appropriate whenever the sought-after eigenvalues form a cluster separated from the rest of the spectrum. Indeed, $\|R_k\|$ is required to be less than $\delta_{m,k}^2 / \|S_m\|$, and $\delta_{m,k}$ may be very small if the whole cluster is not sufficiently well captured by Range$(U^{(k)})$. The result of Proposition 2.5 is particularly helpful when dealing with close to defective eigenvalues, whose eigenvector bases can be very ill-conditioned. The use of invariant subspaces considerably simplifies the analysis [30].

We would also like to comment on the case when H_m, and thus H_k, is tridiagonal. Whenever H_m is tridiagonal, e.g., for Hermitian problems and in the non-Hermitian Lanczos process, the relation between the $(k+1)$st component of an eigenvector of H_m and the residual at step k can be derived directly [20]. However, in the context of inexact matrix-vector multiplication, the resulting matrix H_m is not tridiagonal but upper Hessenberg, even in the inexact Lanczos recurrence (cf. section 6); therefore the more general result is required.

As an alternative to Proposition 2.2, van den Eshof [31] noticed that a related result could be obtained as follows. Let $(\theta^{(k)}, u^{(k))})$ be as before, with associated residual r_k, and let $\hat{u}^{(k)} = [u^{(k)}; \underline{0}]$. If $H_m = QTQ^*$ is the Schur decomposition of H_m, with $Q = [u, Q_2]$ unitary and

$$T = \begin{bmatrix} \theta & a^* \\ \underline{0} & T_{22} \end{bmatrix},$$

then we have (cf. [2, formula (7.107), p. 230])

$$\sin\theta(\hat{u}^{(k)}, u) \leq \frac{\|r_k\|}{\sigma_{\min}(T_{22} - \theta^{(k)} I)}.$$

Therefore, using $u = [u_1; u_2]$, we can write $u = \hat{u}^{(k)}(\hat{u}^{(k)})^* u + (I - \hat{u}^{(k)}(\hat{u}^{(k)})^*)u$ so that $u_2 = [0, I](I - \hat{u}^{(k)}\hat{(}u^{(k)})^*)u$, from which we obtain

$$\|u_2\| \leq \|(I - \hat{u}^{(k)}(\hat{u}^{(k)})^*)u\| = \sin\theta(\hat{u}^{(k)}, u) \leq \frac{\|r_k\|}{\sigma_{\min}(T_{22} - \theta^{(k)} I)}.$$

This relation is very similar to that in (2.6); however, $\sigma_{\min}(T_{22} - \theta^{(k)} I)$ is not the same quantity as $\delta_{m,k}$, as in the former the matrix T_{22} is completely specified by the spectral properties of the target matrix H_m and does not take into account the approximate eigenvector. Nonetheless, it is not clear whether the bound in (2.6), which also includes a condition on the residual, is sharper than the one above. Other characterizations using perturbation theorems can be found, e.g., in [30, section V.2.2]. We would also like to stress that, as shown in Proposition 2.5, the result we have used can be naturally generalized to the case of a simple invariant subspace.

3. Inexact Arnoldi method.

When A is not applied exactly, at each iteration the operation $y = Av$ is replaced by (1.2). Letting $F_m = [f_1, \dots, f_m]$ be the matrix whose columns collect all perturbations, we obtain the following inexact (perturbed) Arnoldi relation:

$$AV_m + F_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^*. \tag{3.1}$$

As in the exact case, matrix V_m has orthonormal columns; however, the space generated by its columns is no longer a Krylov subspace associated with A. Moreover, both V_m and the upper Hessenberg matrix H_m are different from those one would obtain with the exact process, and F_m measures the inexactness of the perturbed Arnoldi relation. When looking for specific eigenpairs, large perturbations may be allowed and still maintain the convergence of the eigenvalue residual to a low final accuracy. Let (θ, u) be an eigenpair of H_m. We have

$$\begin{aligned} AV_m u - \theta V_m u &= V_m H_m u - \theta V_m u + h_{m+1,m} v_{m+1} e_m^* u - F_m u \\ &= h_{m+1,m} v_{m+1} e_m^* u - F_m u. \end{aligned}$$

We call the quantity $AV_m u - \theta V_m u$ the *true residual*, whereas we call the vector $r_m = h_{m+1,m} v_{m+1} e_m^* u$ the *computed residual*, which can be monitored during the recurrence. Note that the true residual cannot be computed when A is not applied exactly. In particular, in the inexact case, the true and computed residuals differ by the vector $F_m u$. We have

$$\begin{aligned} \|(AV_m u - \theta V_m u) - r_m\| &= \|F_m u\| = \|[f_1, \dots, f_m]u\| \\ &= \left\| \sum_{k=1}^m f_k (e_k^* u) \right\| \le \sum_{k=1}^m \|f_k\| \, |e_k^* u|. \end{aligned} \tag{3.2}$$

The distance between the true and the computed residuals is small when each addend $\|f_k\| \, |e_k^* u|$, $k = 1, \dots, m$, in the last sum is small. This occurs when either of the two terms $\|f_k\|$ and $|e_k^* u|$ is small, and not necessarily both, as long as the other term remains bounded by some $O(1)$, say, constant. Therefore, if $|e_k^* u|$ is small, $\|f_k\|$ is allowed to be large and still provide a small gap between true and computed residuals. We next make this statement more precise.

Assume that a maximum of m iterations of inexact Arnoldi can be carried out, and let $(\theta, V_m u)$ be the best Ritz approximation with H_m to the sought-after eigenpair of A. Relaxation in the matrix-vector product at step $k < m$ is possible if there exists an eigenpair of H_{k-1}, denoted by $(\theta^{(k-1)}, u^{(k-1)})$, that is sufficiently close to the eigenpair (θ, u) of H_m or to an eigenpair of A. More precisely, using the notation introduced in Proposition 2.2, for relaxation to take place at step k, it must hold that

$$\|r_{k-1}\| < \frac{\delta_{m,k-1}^2}{4\|s_m\|}, \tag{3.3}$$

$$\forall \theta_j \in \Lambda(H_m), \quad \theta_j \neq \theta, \quad |\theta^{(k-1)} - \theta_j| > 2 \frac{\|s_m\| \, \|r_{k-1}\|}{\delta_{m,k-1}}. \tag{3.4}$$

Condition (3.4) ensures that the eigenpair $(\theta^{(k-1)}, u^{(k-1)})$ in Proposition 2.2 is indeed a perturbation of the analyzed eigenpair (θ, u) of H_m; cf. Remark 2.4. This requires that $\theta^{(k-1)}$ be closer to θ than to other eigenvalues of H_m. This closeness is measured in terms of the quantities $\|r_{k-1}\|$ and $\delta_{m,k-1}$, associated with $\theta^{(k-1)}$. If both (3.3) and (3.4) hold, then (θ, u) is the only eigenpair of H_m such that (2.6) and (2.7) hold at iteration $k-1$. When

$$\theta^{(k-1)} \approx \theta, \qquad \begin{bmatrix} u^{(k-1)} \\ \underline{0} \end{bmatrix} \approx u,$$

we can write

$$\delta_{m,k-1} \leq \min_{\theta_i \in \Lambda(\underline{H_m})} |\theta_i - \theta^{(k-1)}| \approx \min_{\theta_j \in \Lambda(H_m) \setminus \{\theta\}} |\theta_j - \theta^{(k-1)}|,$$

in which case (3.4) follows from (3.3). For simplicity, here and below we assume that θ is a simple eigenvalue of H_m. A generalization of the result of Theorem 3.1 can be obtained by using Proposition 2.5.

Theorem 3.1. *Assume that m inexact Arnoldi iterations have been carried out, and let (θ, u) be an eigenpair of H_m with θ simple Ritz value and $\|u\| = 1$. Given any $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$, assume that for $k = 1, \ldots, m$,*

$$\|f_k\| \leq \begin{cases} \frac{\delta_{m,k-1}}{2m\|r_{k-1}\|}\varepsilon & \text{if } k > 1 \text{ and there exists } (\theta^{(k-1)}, u^{(k-1)}) \text{ of } H_{k-1} \\ & \quad \text{satisfying (3.3) and (3.4),} \\ \frac{1}{m}\varepsilon & \text{otherwise.} \end{cases} \tag{3.5}$$

Then

$$\|(AV_m u - \theta V_m u) - r_m\| \leq \varepsilon.$$

Proof. If at step $k-1$ there exists an eigenpair $(\theta^{(k-1)}, u^{(k-1)})$ of H_{k-1} satisfying (3.3), (3.4), then θ is the only eigenvalue of H_m such that

$$|\theta - \theta^{(k-1)}| \leq 2\frac{\|s_m\|\,\|r_{k-1}\|}{\delta_{m,k-1}}.$$

Hence, Proposition 2.2 ensures that $\theta^{(k-1)}$ is a perturbation of the considered eigenvalue θ of H_m.

Let $\mathcal{K}$ be the subset of $\{2, \ldots, m\}$ such that for each $k \in \mathcal{K}$ there exists an eigenpair $(\theta^{(k-1)}, u^{(k-1)})$ of H_{k-1} satisfying (3.3) and (3.4). Then, using (3.2),

$$\begin{aligned} \|(AV_m u - \theta V_m u) - r_m\| &= \|F_m u\| \leq \sum_{k=1}^{m} \|f_k\|\,|e_k^* u| \\ &= \sum_{k \in \mathcal{K}} \|f_k\|\,|e_k^* u| + \sum_{k \notin \mathcal{K}, k \leq m} \|f_k\|\,|e_k^* u| \\ &\leq \sum_{k \in \mathcal{K}} \frac{\delta_{m,k-1}}{2m\|r_{k-1}\|}\varepsilon\,|e_k^* u| + \sum_{k \notin \mathcal{K}, k \leq m} \frac{1}{m}\varepsilon\,|e_k^* u| \\ &\leq \sum_{k \in \mathcal{K}} \frac{\delta_{m,k-1}}{2m\|r_{k-1}\|}\varepsilon\, 2\frac{\|r_{k-1}\|}{\delta_{m,k-1}} + \sum_{k \notin \mathcal{K}, k \leq m} \frac{1}{m}\varepsilon \qquad \text{(see (2.9))} \\ &= \frac{|\mathcal{K}|}{m}\varepsilon + \frac{m - |\mathcal{K}|}{m}\varepsilon = \varepsilon. \qquad \square \end{aligned}$$

If the conditions of Theorem 3.1 hold, then the difference between the two residuals is less than some fixed value ε. For a sufficiently good starting vector, the norm of the computed residual tends to zero as m goes to infinity; therefore, ε also provides a bound for the final attainable accuracy of the true residual. For this reason, it is natural to relate the value of ε to the threshold in the eigenvalue problem stopping criterion.

Note that $\delta_{m,k-1}$ may dramatically influence the size of the perturbation. For a sensitive matrix H_m, $\delta_{m,k-1}$ may be very small and thus force high accuracy in the matrix-vector product to maintain convergence. On the other hand, it is also important to realize that $\delta_{m,k-1}$ is related to the sensitivity of H_m and not of the original matrix A. Since H_m is a projection of A onto a possibly much smaller space, in general we expect H_m to be less sensitive to perturbations than A.

3.1. A practical relaxation strategy. The result of Theorem 3.1 suggests that we could derive a practical criterion for relaxing the accuracy with which A is applied at each iteration. Unfortunately, the criterion in (3.5) requires crucial information that is not available at iteration k, namely $\delta_{m,k-1} = \sigma_{\min}(\underline{H_m} - \theta^{(k-1)} I)$. This quantity emphasizes the sensitivity of the eigenproblem with H_m and cannot cheaply be replaced. We therefore suggest a relaxation strategy that mimics (3.5), while sacrificing some accuracy in $\delta_{m,k-1}$.

Assume that a maximum of m inexact iterations are to be carried out. At the first iteration, we require that $\|f_1\|$ be less than or equal to $\frac{1}{m}\varepsilon$. At iteration $k > 1$ we require that the perturbation satisfy

$$\|f_k\| \leq \frac{\min\{\alpha, \delta^{(k-1)}\}}{2m\|r_{k-1}\|}\varepsilon, \qquad \delta^{(k-1)} := \min_{\theta_j \in \Lambda(H_{k-1})\setminus\{\theta^{(k-1)}\}} |\theta^{(k-1)} - \theta_j|, \tag{3.6}$$

where α is an estimate of $\|A\|$ and is included to make the condition invariant[1] with respect to a scaling of A. The quantity $\delta^{(k-1)}$ is related to the distance of the Ritz value $\theta^{(k-1)}$ from the rest of the spectrum of H_{k-1}. Clearly, $\delta^{(k-1)}$ may in general be very different from $\delta_{m,k-1}$ (cf. also Remark 2.3). On the other hand, $\delta_{m,k-1}$ will not be too overestimated when $\theta^{(k-1)}$ and its nearby eigenvalues have stabilized to the corresponding eigenvalues of a matrix H_m that is not very sensitive to perturbations.

We should add that, in our numerical experiments, we assume that conditions (3.3) and (3.4) are always satisfied. We require only that the residual be less than one; otherwise the unit value is used instead of the residual in the test.

In (3.6), ε is some fixed constant, naturally related to the final stopping tolerance for the eigenvalue computation. In the numerical experiments of section 4, we require that the final residual be less than ε in norm. In general, ε may include some information about the eigenproblem; one could, for instance, let ε depend on the current approximation, that is, $\varepsilon_{k-1} = |\theta^{(k-1)}|\epsilon$, with ϵ fixed. This choice of ε_{k-1} is associated with the following stopping criterion for the eigenvalue solver,

$$\frac{\|r_{k-1}\|}{|\theta^{(k-1)}|} \leq \epsilon,$$

which is commonly employed in practical implementations; see, e.g., Example 4.2.

[1] We thank Julien Langou, University of Tennessee, for pointing this out.

Remark 3.2. The proposed dynamic perturbation criterion is tailored to the sought-after invariant subspace of A. Assuming that convergence is achieved in the unperturbed case, our analysis shows that specific Ritz pairs obtained by the inexact procedure can still converge to the wanted eigenpairs of A, up to a certain tolerance. Other Ritz pairs may be significantly perturbed, if in the exact scheme they approximate eigenpairs of A with a slower convergence rate. In Example 4.1 we report an experiment where eigenvalues do converge at different rates and the inexact method delivers significantly different results from the exact process (cf. Table 4.1). Our theory formalizes a similar consideration stated in [22, section 11]. □

4. Numerical experiments.

4. Numerical experiments. In this section we report on some numerical experiments that support our theoretical results for the inexact Arnoldi method using Ritz pairs.

Extensive computational experiments were carried out in [5], where the family of relaxation criteria

$$\|f_k\| \leq \frac{10^{-\alpha_0}}{\|r_{k-1}\|}\varepsilon \tag{4.1}$$

was introduced, with $\alpha_0 = 0, 1, 2$, while $\varepsilon = \|A\|\epsilon$, with ϵ equal to the required final residual accuracy. The authors reported the number of times the inexact procedure successfully achieved the required accuracy in approximating the selected eigenpairs of a wide range of matrices. They showed that in 42% of the tests the criterion (4.1) was fulfilled for $\alpha_0 = 0$, up to 81% for $\alpha_0 = 2$. In the following we shall report experiments for $\alpha_0 = 0$. We remark that our theory provides an understanding of the role of the numerator $10^{-\alpha_0}$, and it explains the reason why for $\alpha_0 = 0$ the effect of the perturbation may be severely underestimated.

Example 4.1. We consider the 900×900 matrix stemming from the centered finite difference discretization of the operator

$$Lu = -\Delta u + 100((x+y)u)_x + 100((x+y)u)_y$$

on the unit square, and we seek the eigenvalue with largest real part, $\lambda \approx 7.5127$; cf. Figure 4.1(left). In Figure 4.1(right) we report the convergence history for the exact Arnoldi method for $m = 130$ (dash-dotted line), the inexact method with the flexible accuracy criterion in (3.6) with $\alpha = 10$ (solid line), and the flexible accuracy criterion adopted by Bouras and Frayssé (dashed line); cf. (4.1). The two increasing curves report the values of $\|f_k\|$ for the two inexact methods, with $\varepsilon = 10^{-8}$. The starting vector for the iterative process was taken to be the normalized vector of all ones. Inexactness of A was simulated by adding a random perturbation vector f_k, whose norm was equal to the right-hand side of (3.5) and (4.1) (with $\alpha_0 = 0$), for our criterion and for that of Bouras and Frayssé, respectively.

Both inexact approaches replicate the exact convergence curve until they reach their final attainable accuracy. Note that the original Bouras and Frayssé criterion does not allow the method to fall below the required final residual accuracy, while this is achieved by the criterion in (3.6).

In Table 4.1 we report the last seven computed Ritz values after $m = 100$ iterations, with the exact Arnoldi method and with the inexact method, when either the new relaxation strategy or strategy (4.1) is employed. The first column reports the corresponding eigenvalues of A. We notice that for the two inexact procedures most Ritz values differ from those computed by the exact Arnoldi method, and only the

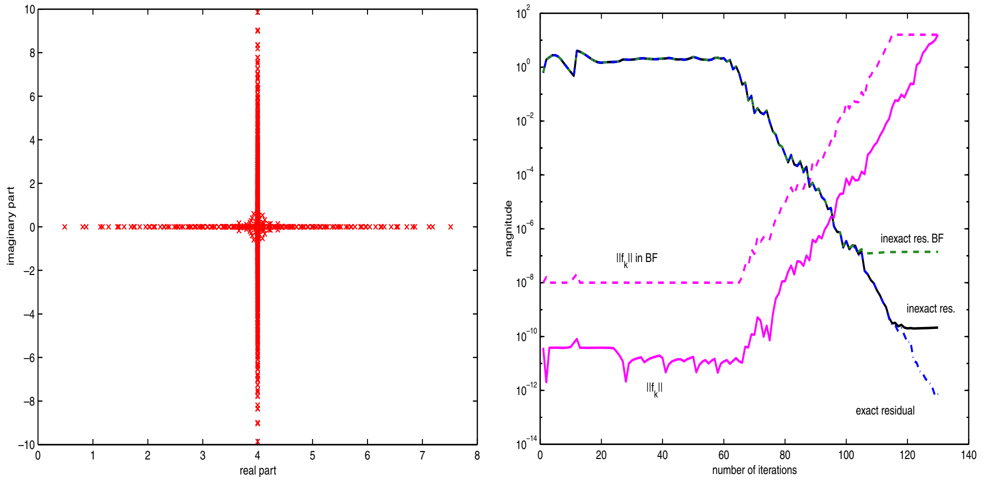


FIG. 4.1. *Example* 4.1. *Left: spectrum of A. Right: convergence curves of the exact method (dash-dotted), inexact Arnoldi with* (3.6) *(solid), and inexact Arnoldi with* (4.1) *(dashed). The increasing curves report the values of the perturbation norms, $\|f_k\|$, in* (3.6) *and* (4.1) *(labeled BF).*

TABLE 4.1

Example 4.1. *Ritz values of exact and inexact Arnoldi methods at iteration m =* 100. *Both flexible strategies* (3.6) *and* (4.1) *are considered. The underlined digits are as accurate as those obtained with the exact Arnoldi method.*

Exact eigenvalues	Arnoldi Ritz values	Flexible accuracy $\|\|f_{100}\|\| = 9.38\,10^{-5}$	Flexible accuracy BF $\|\|f_{100}\|\| = 4.91\,10^{-2}$
6.528963528	$6.5012 + 0.7235i$	$6.5010 + 0.7208i$	$6.5012 + 0.7238i$
6.553808631	$6.5012 - 0.7235i$	$6.5010 - 0.7208i$	$6.5012 - 0.7238i$
6.714551208	$6.6933 + 0.3818i$	$6.6949 + 0.3793i$	$6.6929 + 0.3821i$
6.825884813	$6.6933 - 0.3818i$	$6.6949 - 0.3793i$	$6.6929 - 0.3821i$
6.863220504	$6.8832 + 0.1068i$	$6.8846 + 0.1070i$	$6.8828 + 0.1066i$
7.122198478	$6.8832 - 0.1068i$	$6.8846 - 0.1070i$	$6.8828 - 0.1066i$
7.185702959	7.123532521	$\underline{7.123}544655$	$\underline{7.1235}21753$
7.512696262	7.512695900	$\underline{7.51269590}4$	$\underline{7.5126959}12$

first 3–4 digits remain unaltered. This is not the case for the sought-after eigenvalue $\lambda \approx 7.5126$, for which the exact and inexact Ritz values coincide with several digits of accuracy. We can thus confirm that the perturbation does affect the convergence of Ritz values that do not converge with the same rate as those that guided the perturbation magnitude; see Remark 3.2.

We next report the results obtained when looking for a group of eigenvalues, namely the three largest eigenvalues of A, with both the exact and inexact methods. We considered a starting vector with random entries normally distributed (MATLAB function `randn`, with initial state random number generator), since the previously chosen constant vector had small components onto the second largest eigenvalue. In the table below we display the *four* largest Ritz values obtained after m = 150 iterations of the exact Arnoldi and inexact Arnoldi methods, with $\varepsilon = 10^{-8}$.

Method	θ_4	θ_3	θ_2	θ_1
Exact Arnoldi	6.856543090	7.12220153908	7.18570250215	7.51269626278988
Inexact Arnoldi	6.856516751	7.12220154236	7.18570250124	7.51269626278483

In the inexact process, the perturbation was monitored by using the Frobenius norm of the residual matrix of the three largest Ritz values. The final perturbation norm was equal to $\|f_{150}\| = 2 \cdot 10^{-4}$. We observe in the table that these three Ritz values deviate from those of the exact process of at most $O(10^{-9})$. On the other hand, the fourth Ritz value matches only that of the exact Arnoldi process, to five decimal digits. Therefore, the accuracy in the inexact recurrence is again lost for the approximate eigenpairs that are not involved in the perturbation tolerance. We also note that the largest eigenvalue has more accurate digits with respect to the exact process than the other eigenvalues. In particular, it has almost full accuracy, in spite of a $2 \cdot 10^{-4}$ perturbation in norm. This phenomenon confirms the fact that for groups of eigenvalues, it is the slowest one converging that drives the allowed perturbation.

Example 4.2. We next present a typical setting where the flexible accuracy in the matrix-vector product can be fully appreciated. We consider the matrix SHERMAN5 from the Matrix Market repository [12]. This is a non-Hermitian indefinite real matrix of size $n = 3312$, and was also employed in [14] to analyze the performance of Arnoldi-type methods. The spectrum of the matrix is reported in the left plot of Figure 4.2. We approximate the (real) eigenvalue closest to zero, $\lambda \approx 4.6925 \cdot 10^{-2}$, by means of an inverted Arnoldi process. The generating vector v_1 was the normalized vector of all ones. At each inverted Arnoldi iteration, the operation $y = A^{-1}v$ should be carried out. At each (outer) inexact iteration, a system with A, namely $Ay = v$, is approximately solved with preconditioned GMRES with zero starting guess. The MATLAB incomplete LU factorization with `tol` $= 10^{-3}$ was used as right preconditioner. The GMRES iteration terminated as soon as the system residual norm reached a certain tolerance `tol`. The inner stopping criterion that we used to approximately solve $Ay = v_k$ at step k is

$$\|v_k - Ay\| \leq \frac{\min\{\alpha, \delta^{(k-1)}\}}{2m\|r_{k-1}\|/|\theta^{(k-1)}|}\varepsilon, \qquad \varepsilon = 10^{-10}, \qquad \alpha = 400.$$

In the *exact* case, we assume that we cannot afford to solve with A exactly; therefore we approximately solve the inner system with a fixed tolerance, `tol` $= 10^{-10}$. A total of $m = 12$ outer iterations was carried out. The results of our experiment are reported

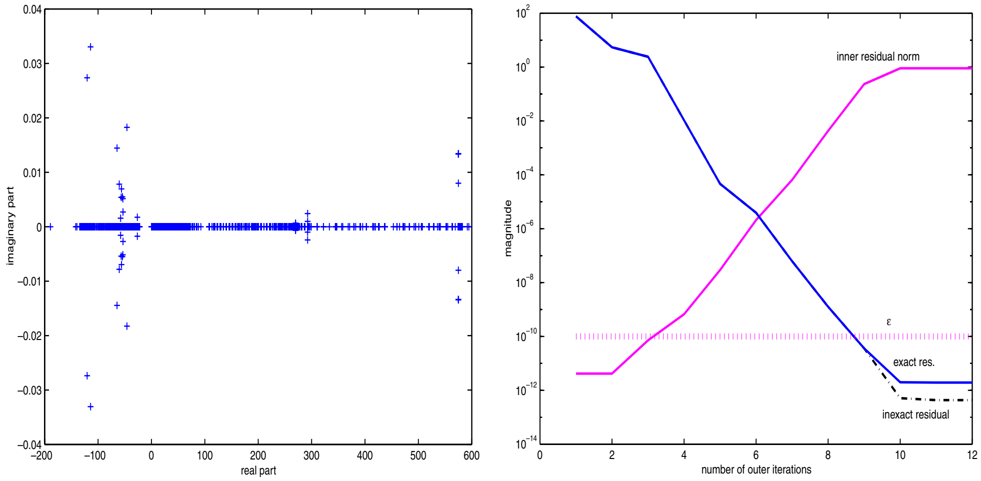


FIG. 4.2. *Example* 4.2. *Left: spectrum of A. Right: convergence curves for inverted Arnoldi (dash-dotted curve) and inexact inverted Arnoldi (solid curve). The right panel also reports the fixed tolerance $\varepsilon = 10^{-10}$ (dotted line).*

in the right plot of Figure 4.2, where the magnitude of the final inner residual at each iteration is also plotted. The eigenvalue convergence curves cannot be distinguished until the final accuracy is reached. Note that in the case of variable inner tolerance, preconditioned GMRES took 22 iterations to solve the first inner system at the required accuracy, whereas only 5 iterations were needed at the 8th inverted Arnoldi iteration, to reach a residual of the order of 10^{-2}.

5. Harmonic Ritz approximation. It has been shown [13, 16, 24] that when looking for interior eigenvalues of Hermitian as well as of non-Hermitian matrices, harmonic Ritz values may be preferred to Ritz values. Harmonic Ritz pairs are pairs $(\theta, V_m u)$, where θ and u are the eigenvalues and corresponding eigenvectors of the generalized eigenvalue problem

$$(H_m^* H_m + |h_{m+1,m}|^2 e_m e_m^*)u = \theta H_m^* u, \quad \|u\| = 1,$$

or, equivalently, of the standard eigenvalue problem

$$(H_m + |h_{m+1,m}|^2 (H_m^*)^{-1} e_m e_m^*)u = \theta u, \quad \|u\| = 1. \tag{5.1}$$

In the following, we let

$$\widetilde{H}_m = H_m + |h_{m+1,m}|^2 (H_m^*)^{-1} e_m e_m^*. \tag{5.2}$$

It was observed in [14] that a better choice as approximation to eigenpairs of A is the pair $(\rho, V_m u)$, where ρ is the Rayleigh quotient of u, that is, $\rho = u^* H_m u$. Using (5.1), for the associated residual we have

$$\begin{aligned} AV_m u - \rho V_m u &= V_m H_m u + h_{m+1,m} v_{m+1} e_m^* u - \rho V_m u \\ &= V_{m+1} \begin{pmatrix} H_m u - \rho u \\ h_{m+1,m} e_m^* u \end{pmatrix}. \end{aligned} \tag{5.3}$$

In the following we will use the *computed* residual in (5.3); namely, we define

$$r_m := V_{m+1} \begin{pmatrix} H_m u - \rho u \\ h_{m+1,m} e_m^* u \end{pmatrix}, \tag{5.4}$$

which differs from the *true* residual $AV_m u - \rho V_m u$ in the inexact case.

A result similar to that of Proposition 2.2 can be derived in terms of the matrix $\widetilde{H}_m$ in (5.2).

PROPOSITION 5.1. *For $k < m$, let $V_k u^{(k)}$ be a harmonic Ritz vector associated with H_k, with $\|u^{(k)}\| = 1$, and let $\rho^{(k)} = (u^{(k)})^* H_k u^{(k)}$. Moreover, let $\mathcal{X} = [[\begin{smallmatrix} u^{(k)} \\ \underline{0} \end{smallmatrix}], Y]$ be a unitary matrix and $\underline{\widetilde{H}_m} = Y^* \widetilde{H}_m Y$. Let $\delta_{m,k} = \sigma_{\min}(\underline{\widetilde{H}_m} - \rho^{(k)} I)$,*

$$r_k = V_{k+1} \begin{pmatrix} H_k u^{(k)} - \rho^{(k)} u^{(k)} \\ h_{k+1,k} e_k^* u^{(k)} \end{pmatrix}, \quad s_m^* = [(u^{(k)})^*, \underline{0}^*] \widetilde{H}_m - \rho^{(k)} [(u^{(k)})^*, \underline{0}^*].$$

If $\delta_{m,k} > 0$ *and*

$$\|r_k\| < \frac{\delta_{m,k}^2}{4\|s_m\|}, \tag{5.5}$$

then there exists a unit norm eigenvector $u = \left[\begin{smallmatrix} u_1 \\ u_2 \end{smallmatrix}\right]$ *of* $\widetilde{H}_m$ *with* $u_1 \in \mathbb{C}^k$ *such that*

$$\|u_2\| \le \frac{\tau}{\sqrt{1+\tau^2}} \le \tau, \quad \textit{where} \quad \tau \in \mathbb{R}, \quad 0 \le \tau < 2\frac{\|r_k\|}{\delta_{m,k}}.$$

Proof. Let Y be such that $\mathcal{X} = [\left[\begin{smallmatrix} u^{(k)} \\ \underline{0} \end{smallmatrix}\right], Y] \in \mathbb{C}^{m\times m}$ is unitary. Recalling that $e_m^*[(u^{(k)}); \underline{0}] = 0$ so that $[(u^{(k)})^*, \underline{0}^*]\widetilde{H}_m[(u^{(k)}); \underline{0}] = \rho^{(k)}$, and following the proof of Proposition 2.2, we can write

$$\mathcal{X}^*\widetilde{H}_m\mathcal{X} = \left[\begin{array}{cc} \rho^{(k)} & K \\ G & \underline{\widetilde{H}_m} \end{array}\right], \qquad \begin{array}{l} G = Y^*\widetilde{H}_m \left[\begin{array}{c} u^{(k)} \\ \underline{0} \end{array}\right], \\ K = [(u^{(k)})^*, \underline{0}^*]\widetilde{H}_m Y. \end{array}$$

Once again, we use $[(u^{(k)})^*, \underline{0}^*]e_m = 0$ and, in addition, $[(u^{(k)})^*, \underline{0}^*]Y = 0$, to obtain

$$\begin{aligned} \gamma := \|G\| &= \left\| Y^*\widetilde{H}_m \left[\begin{array}{c} u^{(k)} \\ \underline{0} \end{array}\right] \right\| = \left\| Y^* \left[\begin{array}{c} H_k u^{(k)} \\ h_{k+1,k}e_1 e_k^* u^{(k)} \end{array}\right] \right\| \\ &= \left\| Y^* \left[\begin{array}{c} H_k u^{(k)} - \rho^{(k)}u^{(k)} \\ h_{k+1,k}e_1 e_k^* u^{(k)} \end{array}\right] \right\| \le \|r_k\|. \end{aligned}$$

Moreover, $\|K\| = \|s_m^* Y\| \le \|s_m\|$. Using Theorem 2.1, if (5.5) holds, so that $\frac{\gamma\|s_m\|}{\delta_{m,k}^2} < \frac{1}{4}$, then there exists $p \in \mathbb{C}^{m-1}$ satisfying $\tau := \|p\| < 2\frac{\gamma}{\delta_{m,k}}$ such that the vector

$$u = \left[\begin{array}{c} u_1 \\ u_2 \end{array}\right] = \frac{1}{\sqrt{1+\|p\|^2}} \left(\left[\begin{array}{c} u^{(k)} \\ \underline{0} \end{array}\right] + \left[\begin{array}{c} Y_1 \\ Y_2 \end{array}\right] p \right)$$

is an eigenvector of $\widetilde{H}_m$ with unit norm, so that $\|u_2\| \le \frac{\|p\|}{\sqrt{1+\|p\|^2}}$. □

Using Proposition 5.1, we can provide a flexible criterion for the accuracy with which the matrix A is applied in the case of harmonic Ritz approximation. We first notice that, in the inexact case, (5.3) becomes

$$AV_m u - \rho V_m u = V_{m+1} \left(\begin{array}{c} H_m u - \rho u \\ h_{m+1,m}e_m^* u \end{array} \right) - F_m u = r_m - F_m u.$$

With the notation of Proposition 5.1, to bound the distance between the true and computed residuals, we require that the best harmonic Ritz approximation $(\rho^{(k-1)}, u^{(k-1)})$ after $k-1$ iterations satisfies

$$\|r_{k-1}\| < \frac{\delta_{m,k-1}^2}{4\|s_m\|}, \tag{5.6}$$

$$\left\{ \begin{array}{l} \forall\rho_j \text{ Rayleigh quotient of eigenvectors of } \widetilde{H}_m, \ \ \rho_j \ne \rho, \\ \quad |\rho^{(k-1)} - \rho_j| > 2\frac{\|s_m\|\,\|r_{k-1}\|}{\delta_{m,k-1}}. \end{array} \right. \tag{5.7}$$

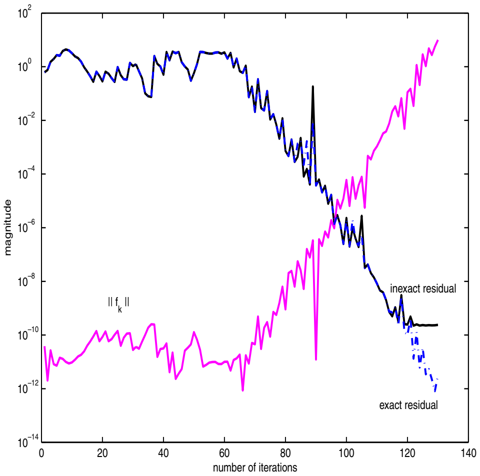


FIG. 5.1. *Matrix in Example* 4.1 *and harmonic Ritz approximation of the largest eigenvalue in modulo. Convergence curves of the exact method (dash-dotted) and inexact Harmonic approximation with* (5.9) *(solid). The increasing curve reports the values of the perturbation norms,* $\|f_k\|$*, in* (5.9).

The following result is similar to Theorem 3.1, and its proof is therefore omitted.

THEOREM 5.2. *Assume that m inexact Arnoldi iterations have been carried out, and let $V_m u$ be a harmonic Ritz vector associated with H_m, with $\|u\| = 1$, and let ρ be the Rayleigh quotient of u, $\rho = u^* H_m u$. Given any $\varepsilon \in \mathbb{R}$, $\varepsilon > 0$, assume that for $k = 1, \ldots, m$,*

$$\|f_k\| \leq \begin{cases} \dfrac{\delta_{m,k-1}}{2m\|r_{k-1}\|}\varepsilon & \begin{array}{l}\textit{if } k > 1 \textit{ and there exists } V_{k-1}u^{(k-1)} \textit{ harmonic Ritz}\\ \textit{vector of } H_{k-1} \textit{ satisfying } (5.6) \textit{ and } (5.7) \textit{ with}\\ \rho^{(k-1)} = (u^{(k-1)})^* H_{k-1} u^{(k-1)},\end{array} \\[2ex] \dfrac{1}{m}\varepsilon & \textit{otherwise.} \end{cases} \tag{5.8}$$

Then $\|(AV_m u - \rho V_m u) - r_m\| \leq \varepsilon$.

The deviation of the inexact harmonic process from its unperturbed counterpart does not significantly differ from that of inexact and exact Arnoldi. As an example, we report the convergence behavior of the exact and inexact harmonic Ritz approximations for the matrix in Example 4.1. In Figure 5.1 the convergence to the largest eigenvalue in modulo is depicted. The matrix-vector product was perturbed by using the following variant of (3.6):

$$\|f_k\| \leq \frac{\min\{\alpha, \delta^{(k-1)}\}}{2m\|r_{k-1}\|}\varepsilon, \qquad \delta^{(k-1)} := \min_{\text{all } \rho_j \neq \rho^{(k-1)}} |\rho^{(k-1)} - \rho_j|, \tag{5.9}$$

where each ρ_j is the Rayleigh quotient associated with the jth unit norm harmonic Ritz vector $u^{(j)}$ of H_{k-1}.

We also report the results after $m = 150$ iterations with the same matrix, when looking for the first three eigenvalues. Similar strategies as in Example 4.1 were used.

Method	ρ_4	ρ_3	ρ_2	ρ_1
Exact harmonic	6.8572788458	7.1222009509	7.18570255184	7.5126962627829
Inexact harmonic	6.8571791230	7.1222009690	7.18570254871	7.5126962627823

The same considerations as for the Arnoldi process apply here. The magnitude of the perturbation at the last iteration was $9 \cdot 10^{-5}$.

6. Exact and inexact Lanczos methods. In this section we show that our analysis carries over to the case of the inexact Lanczos method. In the standard nonsymmetric Lanczos method the following relations hold:

$$AQ_k = Q_k T_k + t_{k+1,k} q_{k+1} e_k^*, \quad A^* P_k = P_k T_k^* + \bar{t}_{k,k+1} p_{k+1} e_k^*, \quad P_k^* Q_k = I_k. \tag{6.1}$$

Writing $Q_{k+1} = [q_1, q_2, \dots, q_{k+1}]$, we assume $\|q_i\| = 1$, $i = 1, \dots, k+1$. This condition, together with the biorthogonality of P_k, Q_k, completely defines the column vectors in P_k, Q_k. Different normalizations would be possible, such as $\|q_i\| = 1 = \|p_i\|$, $i = 1, \dots, k$. In (6.1), T_k is a $k \times k$ nonsymmetric tridiagonal matrix; see, e.g., [2].

When the matrix-vector products with A and A^* are performed inexactly, that is,

$$q = Aq_i + f_i, \qquad p = A^* p_i + g_i, \quad i = 1, \dots, k,$$

the original Lanczos relations (6.1) transform as follows:

$$AQ_k = Q_k H_k + h_{k+1,k} q_{k+1} e_k^* + F_k, \tag{6.2}$$

$$A^* P_k = P_k K_k + \bar{h}_{k,k+1} p_{k+1} e_k^* + G_k, \quad P_k^* Q_k = I_k, \tag{6.3}$$

where we used $F_k = [f_1, \dots, f_k]$ and $G_k = [g_1, \dots, g_k]$. Here matrices H_k and K_k^* are upper Hessenberg and no longer tridiagonal. Their diagonal and near-diagonal elements are the same, whereas the remaining upper parts of the two matrices differ. Clearly, the special properties of the Lanczos iteration are lost. Indeed, the inexact Lanczos iteration is a paired long-term recurrence, as opposed to the paired three-term recurrence of the Lanczos iteration (6.1). Moreover, while in the exact recurrence the matrix T_k provides approximations to both right and left eigenvectors of A, this is no longer the case in the inexact method. Since H_k and K_k^* differ, both eigenvalue problems with H_k and K_k^* need to be solved to obtain right and left Ritz vectors. Another property not inherited by the inexact process concerns convergence. Under certain conditions, the exact Lanczos recurrence determines quadratically converging Ritz values [1, 20].[2] Since neither H_k nor K_k alone provides right and left eigenvector approximations, convergence is only linear in the inexact case.

Proposition 2.2 can be applied to each of the two matrices H_k and K_k^*, to show that the components of converging Ritz vectors have a decreasing pattern. Owing to the similarity between the inexact relations of (6.2), (6.3) with (3.1), we can thus apply the result on dynamic accuracy stated in Theorem 3.1 to each of the two inexact Lanczos recurrences. This provides us with a way to monitor the gap $\|(AQ_k u - \theta Q_k u) - r_k\|$, where (θ, u) is a right eigenpair of H_k in (6.2) and r_k is the associated computed residual. An analogous result holds for left Ritz pairs.

Example 6.1. We consider the 900×900 matrix arising from the centered finite difference approximation of the operator $Lu = -u_{xx} - u_{yy} + (x+y)u_x$ on the unit square. The starting vector is the normalized vector of all ones, for both the Q_k and P_k sequences. We are interested in the approximation of the smallest (real) eigenvalue, $\lambda \approx 2.0276 \cdot 10^{-2}$, with a final residual tolerance of $\varepsilon = 10^{-8}$. The whole spectrum is depicted in the left plot of Figure 6.1. The convergence of the exact and

[2] We thank David Day, Sandia National Laboratories, for pointing us to [1].

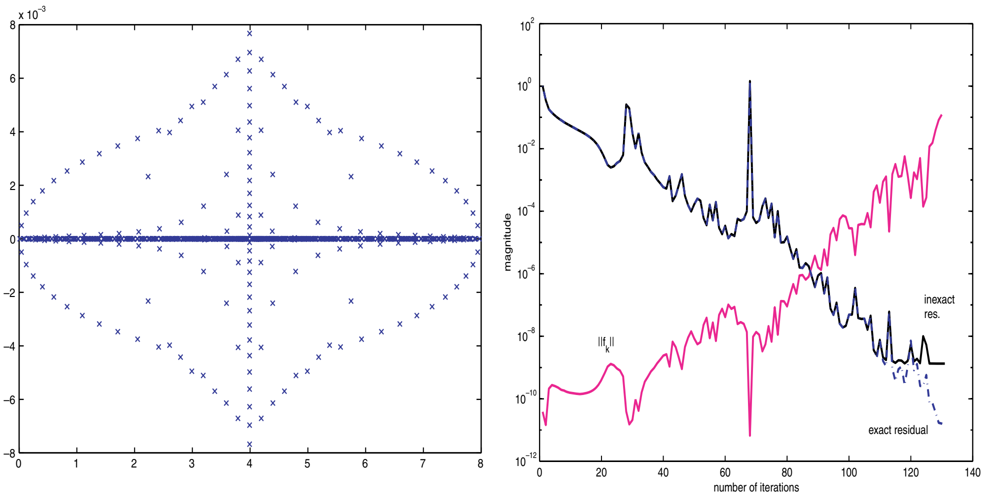


Fig. 6.1. *Example* 6.1. *Left: spectrum of A. Right: convergence curves of right residual norms of exact and inexact Lanczos. The increasing (solid) curve represents the norm of the right matrix-vector perturbation at each iteration.*

inexact Lanczos right residual norms is reported in the right plot of Figure 6.1. The perturbations in the matrix-vector products were enforced by adding random vectors f_i, g_i, $i = 1, \ldots, k$. Their norms were monitored by means of (3.6), applied distinctly to the eigenpairs of the matrices H_{i-1}, K_{i-1}. The convergence curve in the inexact case agrees with that of the exact procedure until final accuracy is reached.

7. Further comments. It was empirically observed in the literature that when approximating the eigenpairs of a matrix by means of the Arnoldi method, the accuracy in the application of the operator may in some cases be relaxed while maintaining the convergence to the sought-after eigenpairs. In this paper we have presented the theoretical foundation for the justification of this phenomenon, and provided a more robust relaxation criterion. Our results indicate that flexible accuracy can be safely employed when the approximate eigenpair is sufficiently close to the target eigenpair, depending on the sensitivity of the given matrix.

Our analysis highlights that reasonably accurate results can be obtained in spite of large perturbations. The inexact Arnoldi relation (3.1) could be written as follows:

$$(A + \mathcal{E}_m)V_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^*, \qquad \mathcal{E}_m = \sum_{k=1}^{m} f_k v_k^*.$$

A backward error analysis would suggest the rather pessimistic picture that a Ritz pair $(\theta, V_m u)$ would be an approximation to an eigenpair of $A + \mathcal{E}_m$ but not of A. We have shown that the perturbation is performed in such a way that the approximation to the target eigenpairs of A is not affected. However, other inexact Ritz pairs may be perturbed by a quantity fully influenced by $\|\mathcal{E}_m\|$.

A related question that we have not answered is whether the rate of convergence could be affected by the perturbation. More precisely, even though our theory ensures that the norm of the true residual of some selected Ritz pairs still converges to a small quantity, it is not clear whether it does so with the same convergence rate as in the unperturbed process. A similar problem is encountered in the inexact linear system setting [22, 32, 25]. In [23] it was shown that no convergence delay is observed in the inexact linear system case, unless the coefficient matrix and the right-hand side are very sensitive to perturbations. Although we expect these conclusions to carry over to the eigenvalue setting, an ad hoc analysis remains to be done.

Practical implementations require (implicit) restarting and locking of converged eigenpairs [10]. As already mentioned in the introduction, we have not addressed these important issues, which need special attention, since our theory predicts that flexible accuracy should take into account the occurrence of Ritz pairs converging to target eigenpairs of A at different rates. In the linear system setting, the problem of restarting has been recently addressed in [26].

Acknowledgments. We thank Jasper van den Eshof for his insightful remarks, and the two anonymous referees for their comments.

REFERENCES

[1] Z. Bai, D. Day, and Q. Ye, *ABLE: An adaptive block Lanczos method for non-Hermitian eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20 (1999), pp. 1060–1082.

[2] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds., *Templates for the solution of algebraic eigenvalue problems: A practical guide*, Software Environ. Tools 11, SIAM, Philadelphia, 2000.

[3] C. Beattie, M. Embree, and J. Rossi, *Convergence of restarted Krylov subspaces to invariant subspaces*, SIAM J. Matrix Anal. Appl., 25 (2004), pp. 1074–1109.

[4] A. Bouras and V. Frayssé, *Inexact matrix-vector products in Krylov methods for solving linear systems: A relaxation strategy*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 660–678.

[5] A. Bouras and V. Frayssé, *A relaxation strategy for the Arnoldi method in eigenproblems*, Technical Report 16, CERFACS, Toulouse, France, 2000.

[6] A. Bouras, V. Frayssé, and L. Giraud, *A Relaxation Strategy for Inner-Outer Linear Solvers in Domain Decomposition Methods*, Technical Report 17, CERFACS, Toulouse, France, 2000.

[7] G. H. Golub and Q. Ye, *Inexact inverse iteration for generalized eigenvalue problems*, BIT, 40 (2000), pp. 671–684.

[8] W. Kahan, B. N. Parlett, and E. Jiang, *Residual bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Numer. Anal., 19 (1982), pp. 470–484.

[9] Y. Lai, K. Lin, and W. Lin, *An inexact inverse iteration for large sparse eigenvalue problems*, Numer. Linear Algebra Appl., 4 (1997), pp. 425–437.

[10] R. B. Lehoucq, D. C. Sorensen, and C. Yang, *ARPACK Users Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*, SIAM, Philadelphia, 1998.

[11] The MathWorks, Inc., *MATLAB User's Guide*, Natick, MA, 2001.

[12] Matrix Market, *A visual repository of test data for use in comparative studies of algorithms for numerical linear algebra*, Mathematical and Computational Sciences Division, National Institute of Standards and Technology, available online at http://math.nist.gov/MatrixMarket.

[13] R. B. Morgan, *Computing interior eigenvalues of large matrices*, Linear Algebra Appl., 154-156 (1991), pp. 289–309.

[14] R. B. Morgan and M. Zeng, *Harmonic projection methods for large non-symmetric eigenvalue problems*, Numer. Linear Algebra Appl., 5 (1998), pp. 33–55.

[15] Y. Notay, *Combination of Jacobi–Davidson and conjugate gradients for the partial symmetric eigenproblem*, Numer. Linear Algebra Appl., 9 (2002), pp. 21–44.

[16] C. Paige, B. Parlett, and H. van der Vorst, *Approximate solutions and eigenvalue bounds from Krylov subspaces*, Numer. Linear Algebra Appl., 2 (1995), pp. 115–134.

[17] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Classics Appl. Math. 20, SIAM, Philadelphia, PA, 1998.

[18] Y. Saad, *Numerical Methods for Large Eigenvalue Problems*, Halstead Press, New York, 1992.

[19] V. Simoncini, *A matrix analysis of Arnoldi and Lanczos methods*, Numer. Math., 81 (1998), pp. 125–141.

[20] V. Simoncini, *Error Bounds for Ritz Values of Non-Hermitian Matrices*, in preparation.

[21] V. Simoncini and L. Eldèn, *Inexact Rayleigh quotient-type methods for eigenvalue computations*, BIT, 42 (2002), pp. 159–182.

[22] V. Simoncini and D. B. Szyld, *Theory of inexact Krylov subspace methods and applications to scientific computing*, SIAM J. Sci. Comput., 25 (2003), pp. 454–477.

[23] V. Simoncini and D. B. Szyld, *On the occurrence of superlinear convergence of exact and inexact Krylov subspace methods*, SIAM Rev., 47 (2005), pp. 247–272.

[24] G. L. G. Sleijpen and J. van den Eshof, *On the use of harmonic Ritz pairs in approximating internal eigenpairs*, Linear Algebra Appl., 358 (2003), pp. 115–137.

[25] G. L. G. Sleijpen, J. van den Eshof, and M. B. van Gijzen, *Iterative linear system solvers with approximate matrix-vector products*, Technical Report 1293, Department of Mathematics, Utrecht University, Utrecht, The Netherlands, 2003.

[26] G. L. G. Sleijpen, J. van den Eshof, and M. B. van Gijzen, *Restarted GMRES with inexact matrix-vector products*, in Proceedings of the 3rd International Conference, NAA 2004, Rousse, Bulgaria, 2004, Z. Li, L. Vulkov, and J. Walźniewski, eds., Lecture Notes in Comput Sci. 3401, Springer-Verlag, Heidelberg, 2005, pp. 494–502.

[27] P. Smit and M. Paardekooper, *The effects of inexact solvers in algorithms for symmetric eigenvalue problems*, Linear Algebra Appl., 287 (1999), pp. 337–357.

[28] D. C. Sorensen, *Implicit application of polynomial filters in a k-step Arnoldi method*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 357–385.

[29] G. W. Stewart, *On the powers of a matrix with perturbations*, Numer. Math., 26 (2003), pp. 363–376.

[30] G. W. Stewart and J.-G. Sun, *Matrix Perturbation Theory*, Academic Press, New York, 1990.

[31] J. van den Eshof, *Personal communication*, 2004.

[32] J. van den Eshof and G. Sleijpen, *Inexact Krylov subspace methods for linear systems*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 125–153.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1175–1201

CONVERGENCE OF THE SUPERCELL METHOD FOR DEFECT MODES CALCULATIONS IN PHOTONIC CRYSTALS[*]

SOFIANE SOUSSI[†]

Abstract. We present a rigorous study of the convergence of the *supercell* method used for determing defect modes in photonic crystals with compactly supported perturbations. Transverse electric and transverse magnetic polarized waves are investigated in 2-D structures. We prove an exponential convergence of the defect frequencies with the *supercell* size and give a justification of the quasi-independence of the corresponding eigenfunctions on the wave vector. We also give a characterization of the *supercell* eigenvalues corresponding to the background photonic crystal.

Key words. photonic crystal, supercell, Helmholtz equation, defect modes

AMS subject classifications. 45C05, 45L05

DOI. 10.1137/040616875

1. Introduction.

Photonic crystals are periodic structures composed of dielectric materials and designed to exhibit interesting properties, such as spectral band gaps, in the propagation of classical electromagnetic waves. In other words, monochromatic electromagnetic waves of certain frequencies do not exist in these structures. Media with band gaps have many potential applications, for example, in optical communications, filters, lasers, and microwaves; see [18], [19], [23], [29] for an introduction to photonic crystals. While necessary conditions under which band gaps exist in general are not known, Figotin and Kuchment have produced an example of high-contrast periodic medium where band gaps exist and can be characterized [16], [17]. Other band gap structures have been found through computational and physical experiments; see [2], [8], [9], [10], [12].

In order to achieve lasers, filters, fibers, or waveguides, allowed modes are required in the band gaps. These modes are obtained by creating localized defects in the periodicity and correspond to isolated eigenvalues with finite multiplicity inside the gaps. The defect mode frequency strongly depends on the defect nature. Figotin and Klein rigorously proved that when a defect is introduced into the periodic structure, i.e., a perturbation with compact support, it is possible to create a defect mode, which is an exponentially confined standing wave whose frequency lies in the band gap [13], [14], [15]; see also Ammari and Santosa [1] and Kuchment and Ong [24] for the issue of existence of exponentially confined modes guided by line defects in photonic crystals.

The defect modes as well as the guided modes associated with compact and line defects, respectively, are computed via the supercell technique. This technique consists of restricting the computation on a domain surrounding the defect with sufficient bulk crystal, called the *supercell*, with periodic conditions on its boundary. The boundary conditions on the supercell are, in principle, irrelevant if the mode is sufficiently confined. Since one would like to compute only the defect or the guided modes in the band gap, without the waste of computation and memory of finding all the eigenvalues associated with the supercell belonging to the continuous spectrum,

[*] Received by the editors October 13, 2004; accepted for publication (in revised form) January 4, 2005; published electronically September 23, 2005.

http://www.siam.org/journals/sinum/43-3/61687.html

[†] Centre de Mathématiques Appliquées, CNRS UMR 7641 & Ecole Polytechnique, 91128 Palaiseau Cedex, France (soussi@cmapx.polytechnique.fr).

one states the problem as one of finding the eigenvalues and eigenvectors closest to the mid-gap frequency.

The supercell method demonstrates very good concordance with experimental results and seems to be very accurate. However, analytic studies and rigorous proofs of convergence of this technique are essentially absent.

In this paper we address some of the basic issues of the supercell method and prove the convergence of this technique. Although one can obtain analogous results for the case of full Maxwell equations, we only address the cases of transverse electric (TE) and transverse magnetic (TM) polarized electromagnetic waves in two-dimensional photonic structures.

The outline of this paper is as follows. In the next section we review some basic facts on the spectra of periodic elliptic operators, emphasizing the Floquet–Bloch theory. We then describe in section 3 the supercell method and investigate its mathematical foundations in the TM case. Section 4 is devoted to the TE case. Finally, in section 5 the results of numerical experiments are presented to illustrate our main findings.

2. Notation and preliminary results.

Consider a photonic crystal characterized by its dielectric permittivity ε_{p}, which is a real valued, piecewise constant and periodic function belonging to the set $\{\varepsilon_{\mathrm{p}} \in L^\infty(\mathbb{R}^2/\mathbb{Z}^2) :\ 0 < \varepsilon_1 \le \varepsilon_{\mathrm{p}} \le \varepsilon_2 \text{ a.e.}\}$, where ε_1 and ε_2 are constants. The magnetic permeability is supposed constant and equal to unity throughout this paper.

We assume that the crystal is periodic with period $[0,1]^2$, i.e., that $\varepsilon_{\mathrm{p}}(x+n) = \varepsilon(x)$ for almost all $x \in \mathbb{R}^2$ and all $n \in \mathbb{Z}^2$.

The propagation of electromagnetic waves is governed by the Maxwell's equations. It is common to reduce these equations in a 2-D medium to two sets of scalar equations in the TM and TE cases. Each one can be solved by solving one scalar partial differential equation and the other scalar functions follow immediately from that solution.

These equations are the Helmholtz equation

$$\Delta u + \omega^2 \varepsilon_{\mathrm{p}} u = 0, \tag{2.1}$$

for the TM polarization, and the acoustic equation

$$\nabla \cdot \frac{1}{\varepsilon_{\mathrm{p}}} \nabla u + \omega^2 u = 0, \tag{2.2}$$

for the TE polarization.

We now recall some well-known results on the spectrum of the TM and TE operators in the periodic medium. Since we deal with a partial differential equation with periodic coefficients, it is natural to make a Floquet transform and apply the Floquet–Bloch theory.

First we briefly present the Floquet–Bloch theory applied to the TM and TE operators in periodic media.

Let $A(x,D)$ denote the TM or TE operator on $L^2(\mathbb{R}^2)$ in a periodic medium characterized by ε_{p}, where $D = -i\nabla$. This operator is invariant with respect to the discrete group of translations $\mathbb{Z}^2$ acting on $\mathbb{R}^2$. It is then natural to apply the Fourier transform on $\mathbb{Z}^2$, that is the transform assigning to a sufficiently decaying function $h(n)$ on $\mathbb{Z}^2$, the Fourier series

$$\widehat{h}(\xi) = \sum_{j \in \mathbb{Z}^2} h(j) e^{i\xi\cdot j},$$

where $\xi \in \mathbb{R}^2$. However, since we deal with functions defined on $\mathbb{R}^2$, we use the Floquet transform which is the appropriate transform in this case.

Consider a function v defined on $\mathbb{R}^2$, sufficiently decaying at infinity. We can then define its Floquet transform by

$$\mathcal{F}v(x,\xi) = \sum_{j\in\mathbb{Z}^2} v(x-j)e^{i\xi\cdot j} = \widehat{v(x-\cdot)}. \tag{2.3}$$

It is easy to check that $\mathcal{F}v(\cdot,\xi)$ is ξ-quasi-periodic with respect to the first variable, that is

$$(\mathcal{F}v)(x+n,\xi) = (\mathcal{F}v)(x,\xi)e^{i\xi\cdot n} \quad \forall x \in \mathbb{R}^2, n \in \mathbb{Z}^2.$$

Moreover, it is periodic with respect to the variable ξ, called quasi-momentum, with period lattice $[0,2\pi]^2$. It is then sufficient to know the function $\mathcal{F}v$ for $(x,\xi) \in Y \times \mathcal{B}$, where $Y = [0,1[^2$ and $\mathcal{B} = [-\pi,\pi[^2$ (called in the literature the first Brillouin zone), to recover it on $\mathbb{R}^2 \times \mathbb{R}^2$.

It turns out that the Floquet transform commutes with partial differential operators with periodic coefficients. In particular, we notice that

$$\mathcal{F}(A(x,D)u) = A(x,D)(\mathcal{F}u).$$

The Floquet transform allows us to represent a function on $L^2(\mathbb{R}^2)$ as a continuous sum of quasi-periodic functions. In fact, the Floquet theory defines an isometric mapping between $L^2(\mathbb{R}^2)$ and $L^2(\mathcal{B}, L^2_\xi(\mathbb{R}^2))$, $L^2_\xi(\mathbb{R}^2)$ being the space of ξ-quasi-periodic L^2- functions. The inverse of the Floquet transform is given by the following formula:

$$(\mathcal{F}^{-1}v)(x) = \frac{1}{|\mathcal{B}|}\int_B v(x,\xi)d\xi, \tag{2.4}$$

for any v in $L^2(\mathcal{B}, L^2_\xi(\mathbb{R}^2))$.

The isometric character of the Floquet transform, together with its commutation properties on partial differential operators with periodic coefficients, make it very useful to study spectral problems. Indeed, the spectral problem for the operator $A(x,D)$ becomes a family of spectral problems for operators $A_\xi(x,D)$ (having formally the same expression but with domains depending on ξ), acting on functions defined on a bounded set (the period lattice of the photonic crystal), with ξ-quasi-periodicity.

An alternative version to the Floquet transform is the transform Φ defined as

$$\Phi v(x,\xi) = \sum_{j\in\mathbb{Z}^2} v(x-j)e^{-i\xi\cdot(x-j)} = e^{-i\xi\cdot x}\mathcal{F}v(x,\xi).$$

The function Φv is periodic with respect to x and $(-x)$-quasi-periodic with respect to ξ with 2π-quasi-period,

$$\begin{cases} \Phi v(x+n,\xi) = \Phi v(x,\xi), & n \in \mathbb{Z}^2, \\ \Phi v(x,\xi+\zeta) = e^{-i\zeta\cdot x}\Phi v(x,\xi), & \zeta \in 2\pi\mathbb{Z}^2. \end{cases} \tag{2.5}$$

With this transform, we now deal with functions defined on a fixed space $L^2(\mathcal{B}, L^2(\mathbb{R}^2/\mathbb{Z}^2))$, while the operator $A(x,D)$ is split into a sum of operators $A(x,D-\xi)$, depending on ξ,

$$\Phi(A(x.D)u)(x,\xi) = A(x,D-\xi)(\Phi u)(x,\xi).$$

The transform Φ is still an isometric mapping between $L^2(\mathbb{R}^2)$ and $L^2(\mathcal{B}, L^2(\mathbb{R}^2/\mathbb{Z}^2))$, and its inverse transform is

$$(\Phi^{-1}v)(x) = \frac{1}{|\mathcal{B}|}\int_{\mathcal{B}} e^{ix\cdot\xi} v(x,\xi)d\xi.$$

Let Σ be the spectrum of $A(x,D)$ on $L^2(\mathbb{R}^2)$ and Σ^ξ the spectrum of $A(x, D-\xi)$ on $L^2(\mathbb{R}^2/\mathbb{Z}^2)$, then we can deduce immediately the following identity:

$$\Sigma = \cup_{\xi\in\mathcal{B}}\Sigma^\xi. \tag{2.6}$$

Now, with these tools, we are in the position to explore the spectrum of the TM and TE operators in periodic media.

In the case of the TE polarization, the operator we are studying is

$$A(x,D) = -\nabla\cdot\frac{1}{\varepsilon_{\mathrm{p}}}\nabla.$$

After the transform Φ, we get the following spectral problem:

$$-(\nabla_x - i\xi)\cdot\frac{1}{\varepsilon_{\mathrm{p}}}(\nabla_x - i\xi)v(x,\xi) = \omega^2 v(x,\xi), \quad v(\cdot,\xi)\in L^2(\mathbb{R}^2/\mathbb{Z}^2). \tag{2.7}$$

We remark that $A(x, D-\xi)$ is an elliptic self-adjoint operator on $L^2(\mathbb{R}^2/\mathbb{Z}^2)$ with compact resolvent. It follows that its spectrum is discrete with countably many positive eigenvalues denoted $\lambda_n(\xi)$ and ordered increasingly. It is easy to prove the continuity of $\lambda_n(\xi)$ on $\xi\in\mathcal{B}$. Finally, defining the intervals I_n by

$$I_n = \left[\min_{\xi\in\mathcal{B}}\lambda_n(\xi), \max_{\xi\in\mathcal{B}}\lambda_n(\xi)\right],$$

we deduce the spectrum of the TE operator

$$\Sigma_{\mathrm{TE}} = \cup_{n\in\mathbb{N}} I_n.$$

We then see clearly the band structure of the spectrum since it is a union of the intervals formed by the values of each eigenvalue when the quasi-momentum varies in the Brillouin zone. In fact, if two successive intervals are disjoint, which means that the maximal value of an eigenvalue is smaller than the minimal value of the following one, then there is a gap in the spectrum Σ_{TE} and no propagation is possible for TE waves at the corresponding frequencies. This makes all the interest of photonic crystals.

In what follows, we make the following assumption: *the spectrum of the TM operator and the TE operator in the periodic medium are absolutely continuous.* This is always assumed in photonic crystal studies. There exists a proof of absolute continuity of Maxwell operator spectrum in periodic media with smooth coefficients in [25].

Another important property of photonic crystals is a consequence of the characterization of the decay of functions in $L^2(\mathbb{R}^2)$ in terms of the smoothness of their Floquet transform in the same spirit as the Paley–Wiener theorem. Suppose that the spectrum contains some gaps, that is $\Sigma_{TE} \neq \mathbb{R}^2$ and let ω be a frequency lying in a band gap. Let G_{p} be the Green's function of the TE operator defined by

$$\nabla\cdot\frac{1}{\varepsilon_{\mathrm{p}}}\nabla G_{\mathrm{p}}(\omega;x,y) + \omega^2 G_{\mathrm{p}}(\omega;x,y) = \delta(x-y), \quad x\in\mathbb{R}^2. \tag{2.8}$$

It has been established in [3], [7] that the Floquet transform of G_{p} is analytic with respect to ω in a complex neighborhood of the real axis. In view of Paley–Wiener-type theorems, the analyticity of $\mathcal{F}G_{\mathrm{p}}$ is the key ingredient of the proof of the following result [13], [14], [15].

Lemma 2.1. *For any $\omega_0^2 > 0$, there exist two positive constants C_1 and C_2 depending only on ω_0^2 such that for any $\omega^2 \notin \Sigma_{\mathrm{TE}} \cap (0, \omega_0^2)$,*

$$|G_{\mathrm{p}}(\omega; x, y)| \leq C_1 e^{-C_2\, dist(\omega^2, \Sigma_{\mathrm{TE}})|x-y|}, \quad \textit{for } |x-y| \to +\infty. \tag{2.9}$$

Remark 2.1. The behavior of the Green's function at infinity is the essential feature of PBG materials: it explains why localized defects in photonic crystals may act as perfect cavities, when the frequency lies in a band gap. Electromagnetic waves can be represented in terms of G_{p} and thus inherit the exponential decay property.

In the case of the TM polarization, the operator we are studying is

$$A(x, D) = -\frac{1}{\varepsilon_{\mathrm{p}}}\Delta.$$

Taking the transform Φ, we get the following spectral problem:

$$-\frac{1}{\varepsilon_{\mathrm{p}}}(\nabla_x - i\xi)\cdot(\nabla_x - i\xi)v(x,\xi) = \omega^2 v(x,\xi), \quad v(\cdot,\xi) \in L^2(\mathbb{R}^2/\mathbb{Z}^2). \tag{2.10}$$

The difference with the TE case is that this operator is elliptic, self-adjoint with compact resolvent on the space of square measurable functions provided with the measure $\varepsilon_{\mathrm{p}}(x)\, dx$ instead of dx.

The results are therefore the same as for the TE case, and we get a spectrum with band structure

$$\Sigma_{\mathrm{TM}} = \cup_{n\in\mathbb{N}} I_n,$$

where $(I_n)_{n\in\mathbb{N}}$ are defined in the same way as for the TE case.

Analogous properties to the TE case hold. In particular, Lemma 2.1 holds with the Green's function associated with the TM polarization.

From now on and until otherwise mentioned, we deal with TM-polarized electromagnetic waves. We consider a background medium characterized by its dielectric permittivity ε_{p}.

First, we introduce some simplified notations. For any domain $D \subset \mathbb{R}^2$ and any measurable positive function $\rho \in L^\infty(D)$ bounded and away from 0 (i.e., $\exists \rho_-, \rho_+ \in \mathbb{R}_+$ s.t. $0 < \rho_- \leq \rho(x) \leq \rho_+ < \infty$, a.e. $x \in D$) we define the weighted L^2 space denoted by $L^2_\rho(D)$ and corresponding to square mesurable functions on D provided with the scalar product $(f,g)_{L^2_\rho(D)} = \int_D f(x)\bar{g}(x)\rho(x)\mathrm{d}x$ and the norm $\|f\|_{L^2_\rho(D)} = \int_D |f(x)|^2\rho(x)\, dx$.

Definition 2.1. *We define the self-adjoint operator A_{p} by*

$$A_{\mathrm{p}} = -\frac{1}{\varepsilon_{\mathrm{p}}}\Delta, \quad \textit{on } L^2_{\varepsilon_{\mathrm{p}}}(\mathbb{R}^2),$$

and denote by Σ_{p} its spectrum.

For $\xi \in [0, 2\pi[^2$ we define A^ξ_{p} on $L^2_{\varepsilon_{\mathrm{p}}}(\mathbb{R}^2/\mathbb{Z}^2)$ by

$$A^\xi_{\mathrm{p}} = -\frac{1}{\varepsilon_{\mathrm{p}}}(\nabla_x - i\xi)\cdot(\nabla_x - i\xi),$$

and denote by $\Sigma_{\mathrm{p}}^{\xi}$ *its spectrum.*

We create a perturbation of the background medium by modifying its dielectric permittivity into ε as follows:

$$\varepsilon(x) = \varepsilon_{\mathrm{p}}(x) - (\delta\varepsilon)\chi_{\Omega}(x), \tag{2.11}$$

where $(\delta\varepsilon)$ is a real constant and Ω is a bounded domain in $\mathbb{R}^2$.

The perturbation of the dielectric permittivity induces a modification of the TM operator into

$$A = -\frac{1}{\varepsilon}\Delta, \tag{2.12}$$

and, consequently, the spectrum Σ of A is different from the spectrum Σ_{p} of A_{p}. However, it has been proved that the perturbation of the TM operator is relatively compact and therefore it keeps the essential spectrum of A_{p} unchanged; see [14]. Since the spectrum Σ_{p} is purely continuous, the perturbation will result in the addition of eigenvalues of finite multiplicity to Σ_{p}.

The following theorem from [14] is of importance to us.

THEOREM 2.1. *Suppose that, for some* $\omega_0^2 > 0$*, the spectrum* $\Sigma_{\mathrm{p}} \cap (0, \omega_0^2)$ *of the operator* A_{p} *has a gap and suppose that the defect* $(\Omega, (\delta\varepsilon))$ *has created an isolated eigenvalue* ω^2 *in the gap. Let* u *be an associated eigenvector. Then, there exists two constants* C_1 *and* C_2*, depending only on* ω_0^2*, such that*

$$\|u\|_{L^2(B_x)} \leq C_1 e^{-C_2\, dist(\omega^2, \Sigma_{\mathrm{p}})\, dist(x,\Omega)} \|u\|_{L^2(\Omega)},$$

where B_x *is the ball of center* x *and radius one.*

Proof. The eigenmode u is solution of the following equation:

$$\Delta u + \omega^2 \varepsilon(x) u = 0. \tag{2.13}$$

It is easy then to see that u is solution of the following integral equation:

$$u(x) = (\delta\varepsilon)\omega^2 \int_{\Omega} G_{\mathrm{p}}(\omega; x, y) u(y)\, dy. \tag{2.14}$$

The proof of the theorem is then a direct consequence of the exponential decay of the Green's function in Lemma 2.1. □

Remark 2.2. This theorem has very important consequences. It explains why we can confine electromagnetic waves in defects or guide them along a defect. The use of dielectric material that has very low loss and the exponential decrease of the electromagnetic field away from the defect ensures a very efficient confinement with a cladding of few periods of the photonic crystal.

3. The supercell method. We start this section by giving a mathematical description of the supercell method.

3.1. Definitions and preliminary results. We consider the background and perturbed media introduced in the previous section with their corresponding TM operators and spectra. Since the perturbed medium is not periodic, the Floquet's theory does not apply.

To recover a periodic medium, we define an artificial medium in the following way. Without loss of generalization, we can suppose that the defect support Ω is centered at 0. For $N \in \mathbb{N}$ large enough to have $\Omega \in]-N, N[^2$, we define the $(2N)$-periodic L^∞-function ε_N by

$$\begin{cases} \varepsilon_N(x) = \varepsilon(x) \quad \forall x \in]-N, N[^2, \\ \varepsilon_N(x + 2Nj) = \varepsilon_N(x) \quad \forall x \in \mathbb{R}^2 \, \forall j \in \mathbb{N}^2. \end{cases} \tag{3.1}$$

Definition 3.1. *We define the self-adjoint operator A_N on $L^2_{\varepsilon_N}(\mathbb{R}^2)$ by*

$$A_N = -\frac{1}{\varepsilon_N}\Delta, \tag{3.2}$$

and let Σ_N be its spectrum.

For $\xi \in \mathcal{B_N} = [-\frac{\pi}{2N}, \frac{\pi}{2N}[^2$, we define the self-adjoint operator A_N^ξ on $L^2_{\varepsilon_N}\mathbb{R}^2/2N\mathbb{Z}^2)$ by:

$$A_N^\xi = -\frac{1}{\varepsilon_N}(\nabla - i\xi) \cdot (\nabla - i\xi),$$

and denote by Σ_N^ξ its spectrum.

The function ε_N defines a photonic crystal formed by the defect repeated with a $2N$-period inside the original photonic crystal. It is therefore obvious that the spectrum Σ_N is an absolutely continuous spectrum. The question is: what happens when N goes to infinity?

A natural answer is that since the repeated defects will be away from each other, they will not interact and, in the neighborhood of one defect, the operator will see almost an infinite crystal. We expect then a kind of convergence of Σ_N to the spectrum Σ corresponding to one defect in the infinite photonic crystal. So for N large enough, after taking the Floquet transform in the supercell and computing the spectrum, we will find a spectrum divided into wide bands very close to those corresponding to the background medium and very narrow bands (almost a horizontal line when plotted against the quasi-momentum) that should correspond to the defect modes of the perturbed crystal. This is what will be proved in the following subsections.

To give a characterization of the convergence of the spectrum of the supercell, we will use the Hausdorff distance denoted $\text{dist}_\mathcal{H}$, that is a measure of the resemblance of two (fixed) sets.

Definition 3.2. *Let E and F be two nonempty subsets of a metric set. We define the Hausdorff distance denoted $dist_\mathcal{H}$ between E and F as*

$$dist_\mathcal{H}(E, F) = \inf\{d \geq 0 \; \forall (x, y) \in E \times F, \; dist(x, F) < d \text{ and } dist(y, E) < d\}.$$

This means that if $\text{dist}_\mathcal{H}(E, F) = d$, then any point of one of the two sets is within distance d from some point of the other set.

Finally, we give in the following proposition an important result from the spectral theory, see [28], that will be useful for the convergence results.

Proposition 3.1. *Let A be a self-adjoint operator with a domain $\mathcal{D}(A)$ and a spectrum $\sigma(A)$, then, for $\mu \in \mathbb{R}$,*

$$dist(\mu, \sigma(A)) = \min_{\phi \in \mathcal{D}(A)} \frac{\|(A - \mu I)\phi\|}{\|\phi\|}. \tag{3.3}$$

3.2. Convergence of the "continuous spectrum." Here we give a characterization of the convergence of the part corresponding to the spectrum of the unperturbed crystal.

THEOREM 3.1. *For any $\omega_0 > 0$ and $N_0 \in \mathbb{N}$, there exists $C > 0$, depending only on ω_0, N_0 and Ω, such that*

$$\max_{\omega^2 \in \cup_{k\in[-N+1,N-1[^2\cap\mathbb{N}^2} \Sigma_{\mathrm{p}}^{\xi+k\pi/N} \cap [0,\omega_0^2]} dist(\omega^2, \Sigma_N^{\xi}) \leq \frac{C}{N^2}, \tag{3.4}$$

for any $N \geq N_0$ and any $\xi \in \mathcal{B}_{\mathcal{N}}$.

Proof. Let $k \in [-N+1, N-1[^2\cap\mathbb{N}^2$ and $\xi \in \mathcal{B}_{\mathcal{N}}$. Let ω^2 be in $\Sigma_{\mathrm{p}}^{\xi+k\pi/N} \cap [0, \omega_0^2]$. Since $\xi + k\pi/N \in \mathcal{B}$, there exists $\phi \in L^2_{\varepsilon_{\mathrm{p}}}(\mathbb{R}^2/\mathbb{Z}^2)$ with unit norm such that

$$\left(\nabla - i\left(\xi + \frac{k\pi}{N}\right)\right) \cdot \left(\nabla - i\left(\xi + \frac{k\pi}{N}\right)\right)\phi + \omega^2\varepsilon_{\mathrm{p}}\phi = 0. \tag{3.5}$$

Let $\tilde{\phi}$ be defined in $L^2_{\varepsilon_N}(\mathbb{R}^2/2N\mathbb{Z}^2)$ as

$$\tilde{\phi}(x) = \phi(x)e^{-i\frac{\pi}{N}k\cdot x}. \tag{3.6}$$

There exists a constant $0 < C_0 < 4$ such that for some integer $N_0 > 0$,

$$\|\tilde{\phi}\|_{L^2_{\varepsilon_N}(\mathbb{R}^2/2N\mathbb{Z}^2)} > C_0 N^2$$

for any $N \geq N_0$, where C_0 and N_0 are independent of $\tilde{\phi}$. The function $\tilde{\phi}$ satisfies the following equation:

$$(\nabla - i\xi) \cdot (\nabla - i\xi)\tilde{\phi} + \omega^2\varepsilon_{\mathrm{p}}\tilde{\phi} = 0, \tag{3.7}$$

which can be rewritten as follows:

$$(\nabla - i\xi) \cdot (\nabla - i\xi)\tilde{\phi} + \omega^2\varepsilon_N\tilde{\phi} = -\chi_\Omega(\delta\varepsilon)\omega^2\tilde{\phi}. \tag{3.8}$$

Let C_1 be the minimal number of unit squares in which Ω can be strictly included. Since the $L^2_{\varepsilon_{\mathrm{p}}}$-norm of $\tilde{\phi}$ in a unit square is 1, we have

$$\|\tilde{\phi}\|_{L^2(\Omega)} \leq \frac{C_1}{\sqrt{\min(\varepsilon_{\mathrm{p}})}}.$$

Thus

$$\begin{aligned} \frac{\|\frac{1}{\varepsilon_N}(\nabla - i\xi) \cdot (\nabla - i\xi)\tilde{\phi} + \omega^2\tilde{\phi}\|_{L^2_{\varepsilon_N}(\mathbb{R}^2/(2N)\mathbb{Z}^2)}}{\|\tilde{\phi}\|_{L^2_{\varepsilon_N}(\mathbb{R}^2/2N\mathbb{Z}^2)}} &= (\delta\varepsilon)\omega^2 \frac{(\int_\Omega |\tilde{\phi}|^2 \frac{1}{\varepsilon_N}\, dx)^{1/2}}{\|\tilde{\phi}\|_{L^2_{\varepsilon_N}(\mathbb{R}^2/2N\mathbb{Z}^2)}} \\ &\leq \frac{C_2}{N^2}, \end{aligned}$$

where $C_2 = |(\delta\varepsilon)|\omega_0^2 \frac{C_1}{C_0\sqrt{\min(\varepsilon_{\mathrm{p}})\min(\varepsilon_N)}}$.

It follows from Proposition 3.1 that there exists an eigenvalue ω_ξ^2 belonging to the spectrum Σ_N^ξ of the operator A_N^ξ such that

$$|\omega^2 - \omega_\xi^2| \leq \frac{C_2}{N^2},$$

which ends the proof. □

Remark 3.1. This theorem tells us that $\mathrm{card}\big(\Sigma_N^\xi \cap [0, \omega_0^2]\big)$ for $\xi \in \mathcal{B}_\mathcal{N}$ will grow at least as fast as $N^2 \mathrm{card}\big(\Sigma_p^{\xi'} \cap [0, \omega_0^2]\big)$ for any $\xi' \in \mathcal{B}$. So when we use the supercell method to determine the defect modes, we are in front of a dilemma; larger is the size of the supercell, better is the approximation of the defect eigenvalues. But this will take much more time and need much more memory size because of the size of the computational domain and the growing number of useless (in the sense that they do not correspond to the defect) eigenvalues. It is important then to determine the convergence rate of the eigenvalues corresponding to the defect.

Since we know that the spectrum $\Sigma_N = \cup_{\xi \in \mathcal{B}_\mathcal{N}} \Sigma_N^\xi$ is absolutely continuous, we deduce that each connected component of $(\mathbb{R}^2 \setminus \Sigma_N) \cap \Sigma_\mathrm{p} \cap [0, \omega_0]$ has a width smaller than $\frac{2C}{N^2}$.

In practice, because of the growth of degeneracy of the eigenvalues located in Σ_p with N, there will be almost no visible gap inside the bands of Σ_N but the remark remains useful for the perturbation brought to the edges of the bands. In particular, it is useful to check if a perturbation of the edges of a band in Σ_p is due to the presence of a defect eigenvalue in Σ close to the band or not.

3.3. Convergence of the defect eigenvalues.

Here we are concerned with the behavior of the part of the spectrum Σ_N that will give us an approximation of the defect modes (eigenvalues with finite multiplicity in Σ). Let us first try to give a characterization of this part.

DEFINITION 3.3. *For $\eta > 0$, we define $\Sigma_{\mathrm{d},N}^\eta$ as the union of the connected components of Σ_N that are at least η-distant from Σ_p.*

We also define Σ_d as the set of the defect eigenvalues of the perturbed photonic crystal

$$\Sigma_\mathrm{d} = \Sigma \setminus \Sigma_\mathrm{p}.$$

Finally, we introduce $\Sigma_{\mathrm{d},N}^{\xi,\eta}$ and Σ_d^η as

$$\Sigma_{\mathrm{d},N}^{\xi,\eta} = \{\omega^2 \in \Sigma_N^\xi : dist(\omega^2, \Sigma_\mathrm{p}) \geq \eta\}.$$
$$\Sigma_\mathrm{d}^\eta = \{\omega_\mathrm{d}^2 \in \Sigma_\mathrm{d} : dist(\omega_\mathrm{d}^2, \Sigma_\mathrm{p}) \geq \eta\}.$$

The following proposition holds.

PROPOSITION 3.2. *For every gap $]a, b[$ in Σ_p $(0 < a < b)$ satisfying $]a, b[\cap \Sigma = \emptyset$, there exists $N_1 \in \mathbb{N}$ such that, for $N \geq N_1$, $\Sigma_N \cap]a, b[= \emptyset$.*

Proof. Suppose that the proposition is false. Then for any $N_0 \in \mathbb{N}$ there exists $N \geq N_0$ and $\omega_N^2 \in]a, b[\cap \Sigma_N$. This means that there exist $\xi_N \in \mathcal{B}_\mathcal{N}$ and $\phi_N \in L_{\varepsilon_N}^2(\mathbb{R}^2/2N\mathbb{Z}^2)$ with unit norm such that

$$(\nabla - i\xi_N) \cdot (\nabla - i\xi_N)\phi_N + \omega_N^2 \varepsilon_N \phi_N = 0 \quad \text{in } L_{\varepsilon_N}^2(\mathbb{R}^2/2N\mathbb{Z}^2). \tag{3.9}$$

Now, we define $\tilde{\phi}_N$ in $L^2_\varepsilon(\mathbb{R}^2)$ by

$$\tilde{\phi}_N(x) = \int_\Omega G(\omega_N^2; x, y) e^{-i\xi_N \cdot y} \phi_N(y)\, dy. \tag{3.10}$$

The following lemma is needed.

Lemma 3.1. *There exist $N_0 > 0$ depending only on a, b and $\Sigma_{\rm p}$, such that for $N \geq N_0$, we have*

$$\|\tilde{\phi}_N\|_{L^2_\varepsilon(\mathbb{R}^2)} \geq \frac{1}{2}.$$

Proof. From the expression of $\tilde{\phi}_N$ we deduce that

$$\begin{aligned}
(\delta\varepsilon)\omega_N^2 \tilde{\phi}_N(x) &= (\delta\varepsilon)\omega_N^2 \int_\Omega G(\omega_N^2; x, y) e^{-i\xi_N \cdot y} \phi_N(y)\, dy \\
&= \int_{\mathbb{R}^2} G(\omega_N^2; x, y)(\Delta + \omega_N^2 \varepsilon_{\rm p})\big(e^{-i\xi_N \cdot y}\phi_N(y)\big)\, dy \\
&\quad - \int_{\mathbb{R}^2} G(\omega_N^2; x, y)(\Delta + \omega_N^2 \varepsilon)\big(e^{-i\xi_N \cdot y}\phi_N(y)\big)\, dy \\
&= \int_{\mathbb{R}^2} (\Delta + \omega_N^2 \varepsilon_{\rm p}) G(\omega_N^2; x, y) e^{-i\xi_N \cdot y}\phi_N(y)\, dy \\
&\quad - \int_{\mathbb{R}^2} G(\omega_N^2; x, y) e^{-i\xi_N \cdot y}((\nabla - i\xi_N)\cdot(\nabla - i\xi_N) + \omega_N^2\varepsilon)\phi_N(y)\, dy \\
&= e^{-i\xi_N \cdot x}\phi_N(x) \\
&\quad - \int_\Omega \sum_{j\in\mathbb{Z}^2, j\neq 0} (G(\omega_N^2; x, y + Nj) e^{-iN\xi_N \cdot j}) e^{-i\xi_N \cdot y}\phi_N(y) dy.
\end{aligned}$$

Let us now prove that the L^2-norm of the last term in $]-N, N[^2$ converges to 0. From the exponential decay of the Green's function, we deduce that there exist positive constants C_1 and C_2 depending only on the distance of a and b to $\Sigma_{\rm p}$ such that, for any $\omega^2 \in]a, b[$, we have [1]

$$\sum_{j\in\mathbb{Z}^2, j\neq 0} \big|G(\omega^2; x, y + Nj)\big| \leq C_1 e^{-C_2 N} \qquad \forall x \in]-N, N[^2\ \forall y \in \Omega. \tag{3.11}$$

It follows then, since $\|\phi_N\|_{L^2_{\varepsilon_N}(]-N,N[^2)} = 1$, that for any $x \in]-N, N[^2$, we have

$$\begin{aligned}
&\Big| \int_\Omega \sum_{j\in\mathbb{Z}^2, j\neq 0} (G(\omega_N^2; x, y + Nj) e^{-iN\xi_N \cdot j}) e^{-i\xi_N \cdot y}\phi_N(y) dy \Big| \\
&\quad \leq C_1 e^{-C_2 N} \int_\Omega |\phi_N(y)|\, dy \\
&\quad \leq C_1 e^{-C_2 N} \frac{|\Omega|^{\frac{1}{2}}}{\sqrt{\min(\varepsilon_N)}} ||\phi_N||_{L^2_{\varepsilon_N}(\Omega)} \\
&\quad \leq C_1 e^{-C_2 N} \frac{|\Omega|^{\frac{1}{2}}}{\sqrt{\min(\varepsilon_N)}}.
\end{aligned}$$

We then deduce that:

$$\left\| \int_\Omega \sum_{j\in\mathbb{Z}^2, j\neq 0} (G(\omega_N^2; x, y+Nj)e^{-iN\xi_N\cdot j})e^{-i\xi_N\cdot y}\phi_N(y)dy \right\|_{L^2_\varepsilon(]-N,N[^2)} \\ \leq \frac{\sqrt{\max(\varepsilon)}|\Omega|^{\frac{1}{2}}}{\sqrt{\min(\varepsilon_N)}} 2NC_1e^{-C_2N}.$$

Hence, recalling that $\|e^{-i\xi_N\cdot x}\phi_N(x)\|_{L^2_{\varepsilon_N}(]-N,N[^2)} = 1$, there exists $N_0 > 0$ such that for any $N \geq N_0$, we have

$$\|\tilde{\phi}_N\|_{L^2_\varepsilon(\mathbb{R}^2)} \geq \|\tilde{\phi}_N\|_{L^2_\varepsilon(]-N,N[^2)} \geq \frac{1}{2}. \tag{3.12}$$

Lemma 3.1 is then proved. □

We now turn to the proof of Proposition 3.2. We have

$$\begin{aligned}
\Delta\tilde{\phi}_N + \omega_N^2\varepsilon\tilde{\phi}_N &= \int_\Omega (\Delta_x + \omega_N^2\varepsilon)G(\omega_N^2; x, y)e^{-i\xi_N\cdot y}\phi_N(y)\, dy \\
&= \int_\Omega (\Delta_x + \omega_N^2\varepsilon_{\mathrm{p}})G(\omega_N^2; x, y)e^{-i\xi_N\cdot y}\phi_N(y)\, dy \\
&\quad -(\delta\varepsilon)\chi_\Omega(x)\omega_N^2 \int_\Omega G(\omega_N^2; x, y)e^{-i\xi_N\cdot y}\phi_N(y)\, dy \\
&= \chi_\Omega(x)e^{-i\xi_N\cdot x}\phi_N(x) \\
&\quad -\chi_\Omega(x)\int_{\mathbb{R}^2} G(\omega_N^2; x, y)(\Delta_y + \omega_N^2\varepsilon_{\mathrm{p}})(e^{-i\xi_N\cdot y}\phi_N(y))\, dy \\
&\quad +\chi_\Omega(x)\int_{\mathbb{R}^2} G(\omega_N^2; x, y)(\Delta_y + \omega_N^2\varepsilon)(e^{-i\xi_N\cdot y}\phi_N(y))\, dy \\
&= \chi_\Omega(x)e^{-i\xi_N\cdot x}\phi_N(x) \\
&\quad -\chi_\Omega(x)\int_{\mathbb{R}^2} (\Delta_y + \omega_N^2\varepsilon_{\mathrm{p}})G(\omega_N^2; x, y)e^{-i\xi_N\cdot y}\phi_N(y)\, dy \\
&\quad +\chi_\Omega(x)\int_{\mathbb{R}^2} G(\omega_N^2; x, y)e^{-i\xi_N\cdot y}((\nabla - i\xi_N)\cdot(\nabla - i\xi_N) + \omega_N^2\varepsilon)\phi_N(y)dy \\
&= (\delta\varepsilon)\omega_N^2\chi_\Omega(x) \\
&\qquad \int_\Omega \Bigg(\sum_{j\in\mathbb{Z}^2, j\neq 0} G(\omega_N^2; x, y+Nj)e^{-i\xi_N\cdot(y+Nj)} \Bigg)\phi_N(y)\, dy.
\end{aligned}$$

Using estimate (3.11), we deduce the existence of positive constants C_1 and C_2 depending only on the distance of a and b to Σ_{p} such that

$$\left| \sum_{j\in\mathbb{Z}^2, j\neq 0} G(\omega_N^2; x, y+Nj)e^{-i\xi_N\cdot(y+Nj)} \right| \leq C_1e^{-C_2N}, \tag{3.13}$$

for any $x, y \in \Omega$. We then obtain that

$$\begin{aligned}&\left| \int_\Omega \left(\sum_{j\in\mathbb{Z}^2, j\neq 0} G(\omega_N^2; x, y+Nj) e^{-i\xi_N\cdot(y+Nj)} \right) \phi_N(y)\, dy \right| \\ &\quad \le C_1 e^{-C_2 N} |\Omega|^{\frac{1}{2}} \|\phi_N\|_{L^2(\Omega)} \\ &\quad \le C_1 e^{-C_2 N} |\Omega|^{\frac{1}{2}}.\end{aligned}$$

This yields the following result:

$$\left\| \frac{1}{\varepsilon} \Delta \tilde{\phi}_N + \omega_N^2 \tilde{\phi}_N \right\|_{L^2_\varepsilon(\mathbb{R}^2)} \le \frac{|(\delta\varepsilon)|\omega_N^2|\Omega|}{\sqrt{\min(\varepsilon)}} C_1 e^{-C_2 N}. \tag{3.14}$$

Lemma 3.1 yields the estimate

$$\operatorname{dist}(\omega_N^2, \Sigma) \le \frac{2|(\delta\varepsilon)|\ b\ |\Omega|}{\sqrt{\min(\varepsilon)}} C_1 e^{-C_2 N},$$

from which we conclude that $\operatorname{dist}(]a,b[,\Sigma) = 0$. This is a contradiction with the assumption. The proof of the proposition is complete. ∎

Now we can prove the following result concerning the convergence to the defect modes.

Theorem 3.2. *Suppose that the perturbation has created defect eigenvalues. Then, there exist $\eta_0 > 0$ and $N_0 \in \mathbb{N}$ such that for any $\eta \le \eta_0$ and $N \ge N_0$,*

$$\Sigma_{\mathrm{d},N}^{\xi,\eta} \neq \emptyset \quad \forall \xi \in \mathcal{B}_{\mathcal{N}}.$$

Moreover, for any $\omega_0^2 > 0$ and $\eta \le \eta_0$, there exist two positive constants C_1 and C_2, depending only on ω_0^2, such that for any $\xi \in \mathcal{B}_{\mathcal{N}}$,

$$dist_{\mathcal{H}}(\Sigma_{\mathrm{d},N}^{\xi,\eta} \cap [0, \omega_0^2], \Sigma_{\mathrm{d}}^{\eta} \cap [0, \omega_0^2]) \le C_1 e^{-C_2 \eta N}. \tag{3.15}$$

Proof. Let ω_{d}^2 be a defect eigenvalue in Σ_{d}. It follows that there exists a function u in $L^2_\varepsilon(\mathbb{R}^2)$ with unit norm such that

$$\Delta u + \omega_{\mathrm{d}}^2 \varepsilon u = 0 \quad \text{in } \mathbb{R}^2. \tag{3.16}$$

Let ξ be in $\mathcal{B}_{\mathcal{N}}$. We define u^ξ in $L^2_{\varepsilon_N}(\mathbb{R}^2/2N\mathbb{Z}^2)$ by

$$u^\xi(x) = \sum_{j\in\mathbb{Z}^2} u(x+Nj) e^{i\xi\cdot(x+Nj)}.$$

Then for $x \in]-N, N[^2$, we have

$$\begin{aligned}&((\nabla - i\xi)\cdot(\nabla - i\xi) + \omega_{\mathrm{d}}^2 \varepsilon_N) u^\xi(x) \\ &\qquad = \sum_{j\in\mathbb{Z}^2} e^{i\xi\cdot(x+Nj)} (\Delta + \omega_{\mathrm{d}}^2 \varepsilon_N) u(x+Nj) \\ &\qquad = \sum_{j\in\mathbb{Z}^2} e^{i\xi\cdot(x+Nj)} (\Delta + \omega_{\mathrm{d}}^2 \varepsilon(x+Nj)) u(x+Nj) \\ &\qquad\quad + (\delta\varepsilon)\omega_{\mathrm{d}}^2 \sum_{j\in\mathbb{Z}^2} e^{i\xi\cdot(x+Nj)} (\varepsilon_N(x) - \varepsilon(x+Nj)) u(x+Nj) \\ &\qquad = -(\delta\varepsilon)\omega_{\mathrm{d}}^2 \chi_\Omega(x) \sum_{j\in\mathbb{Z}^2, j\neq 0} e^{i\xi\cdot(x+Nj)} u(x+Nj).\end{aligned}$$

On the other hand, for $x \in \mathbb{R}^2$,

$$\begin{aligned} u(x) &= \int_{\mathbb{R}^2} \delta(x-y)u(y)\,dy \\ &= \int_{\mathbb{R}^2} (\Delta + \varepsilon_{\mathrm{p}}\omega_{\mathrm{d}}^2) G(\omega_{\mathrm{d}}^2; x, y) u(y)\,dy \\ &= \int_{\mathbb{R}^2} G(\omega_{\mathrm{d}}^2; x, y)(\Delta + \varepsilon_{\mathrm{p}}\omega_{\mathrm{d}}^2) u(y)\,dy \\ &= (\delta\varepsilon)\omega_{\mathrm{d}}^2 \int_{\Omega} G(\omega_{\mathrm{d}}^2; x, y) u(y)\,dy. \end{aligned}$$

Therefore

$$\begin{aligned} &((\nabla - i\xi)\cdot(\nabla - i\xi) + \omega_{\mathrm{d}}^2\varepsilon_N)u^\xi(x) \\ &\quad -(\delta\varepsilon)^2\omega_{\mathrm{d}}^4\chi_\Omega(x)\int_\Omega \left(\sum_{j\in\mathbb{Z}^2, j\neq 0} G(\omega_{\mathrm{d}}^2; x+Nj, y) e^{i\xi\cdot(x+Nj)}\right) u(y)\,dy. \end{aligned}$$

From (3.11), it follows that there exist two positive constants C_1 and C_2, depending only on ω_0^2, such that

$$\begin{aligned} \left| \int_\Omega \left(\sum_{j\in\mathbb{Z}^2, j\neq 0} G(\omega_{\mathrm{d}}^2; x+Nj, y) e^{i\xi\cdot(x+Nj)}\right) u(y)\,dy \right| &\leq C_1 e^{-C_2 \mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_{\mathrm{p}})N} \int_\Omega |u(y)|\,dy \\ &\leq C_1 e^{-C_2 \mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_{\mathrm{p}})N} |\Omega|^{\frac{1}{2}} \|u\|_{L^2(\Omega)} \\ &\leq |\Omega|^{\frac{1}{2}} C_1 e^{-C_2 \mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_{\mathrm{p}})N}. \end{aligned}$$

Therefore

$$\left\| \frac{1}{\varepsilon_N}(\nabla - i\xi)\cdot(\nabla - i\xi)u^\xi(x) + \omega_{\mathrm{d}}^2 u^\xi(x) \right\|_{L^2_{\varepsilon_N}(]-N,N[^2)} \leq \frac{(\delta\varepsilon)^2\omega_{\mathrm{d}}^4|\Omega|}{\sqrt{\min(\varepsilon_N)}} C_1 e^{-C_2 \mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_{\mathrm{p}})N}, \tag{3.17}$$

since

$$\begin{aligned} u^\xi(x) &= u(x)e^{i\xi\cdot x} + \sum_{j\in\mathbb{Z}^2, j\neq 0} u(x+Nj)e^{i\xi\cdot(x+Nj)}, x \in]-N, N[^2, \\ \lim_{N\to+\infty} &\|u(x)e^{i\xi\cdot x}\|_{L^2_{\varepsilon_N}(]-N,N[^2)} = 1, \end{aligned}$$

and

$$\left\| \sum_{j\in\mathbb{Z}^2, j\neq 0} u(x+Nj)e^{i\xi\cdot(x+Nj)} \right\|_{L^2_{\varepsilon_N}(]-N,N[^2)} \leq |\Omega|^{\frac{1}{2}} N C_1 e^{-C_2 \mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_{\mathrm{p}})N},$$

we deduce that for N large enough,

$$\|u^\xi\|_{L^2_{\varepsilon_N}(]-N,N[^2)} \geq \frac{1}{2}.$$

Thus, we conclude that

$$\mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_N^\xi) \leq C_1 e^{-C_2 \mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_{\mathrm{p}}) N},$$

for two positive constants C_1 and C_2, depending only on ω_0^2.

It follows that

$$\max_{\omega_{\mathrm{d}} \in \Sigma_{\mathrm{d}}^\eta \cap [0, \omega_0^2]} \mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_N^\xi) \leq C_1 e^{-C_2 \eta N}, \tag{3.18}$$

uniformly for $\xi \in \mathcal{B}_{\mathcal{N}}$. Hence, any defect eigenvalue $\omega_{\mathrm{d}}^2 \in \Sigma_{\mathrm{d}}$ is a limit point of $(\Sigma_N^\xi)_{N \in \mathbb{N}}$.

Let $\eta > 0$ be small enough to get $\Sigma_{\mathrm{d}}^\eta \neq \emptyset$. Applying Proposition 3.2, we may see that there exists $N_0 \in \mathbb{N}$, depending only on ω_0^2 and η, such that $\Sigma_{\mathrm{d},N}^{\xi,\eta} \cap [0, \omega_0^2]$ has at least as many connected components as $\mathrm{card}(\Sigma_{\mathrm{d}}^\eta \cap [0, \omega_0^2])$ for $N \geq N_0$. To prove this, we take a neighborhood of $\Sigma_{\mathrm{d}}^\eta \cap [0, \omega_0^2]$ formed by disjoint intervals and away from Σ_{p}, each one of them containing exactly one defect eigenvalue. Then from Proposition 3.2, we deduce that for N large enough, the edges of these intervals will be strictly distant from Σ_N. On the other hand, we have proved here that for N large enough, the intersection of every interval with Σ_N^ξ is not empty. This means that $\Sigma_{\mathrm{d},N}^{\xi,\eta}$ is not empty if we take η small enough and then let N be large enough. By the same manner, (3.18) can be written as

$$\max_{\omega_{\mathrm{d}} \in \Sigma_{\mathrm{d}}^\eta \cap [0, \omega_0^2]} \mathrm{dist}(\omega_{\mathrm{d}}^2, \Sigma_{\mathrm{d},N}^{\xi,\eta}) \leq C_1 e^{-C_2 \eta N}, \tag{3.19}$$

uniformly for $\xi \in \mathcal{B}_{\mathcal{N}}$. The proof of the first part of the theorem is then done.

Now, let $\xi \in \mathcal{B}_{\mathcal{N}}$ and let $\omega^2 \in \Sigma_{\mathrm{d},N}^{\xi,\eta}$. There exists $\phi \in L^2_{\varepsilon_N}(\mathbb{R}^2/2N\mathbb{Z}^2)$ with unit norm such that

$$(\nabla - i\xi) \cdot (\nabla - i\xi)\phi + \omega^2 \varepsilon_N \phi = 0.$$

Then, we define u in $L^2_\varepsilon(\mathbb{R}^2)$ by

$$u(x) = \int_\Omega G(\omega^2; x, y)\phi(y) e^{-i\xi \cdot y}\, dy.$$

Let us now find a lower bound for $\|u\|_{L^2_\varepsilon(\mathbb{R}^2)}$. We compute

$$\begin{aligned}
(\delta\varepsilon)\omega^2 u(x) &= \int_{\mathbb{R}^2} G(\omega^2; x, y)\big(\Delta + \omega^2 \varepsilon_{\mathrm{p}}\big)\big(\phi(y) e^{-i\xi \cdot y}\big)\, dy \\
&\quad - \int_{\mathbb{R}^2} G(\omega^2; x, y)\big(\Delta + \omega^2 \varepsilon\big)\big(\phi(y) e^{-i\xi \cdot y}\big)\, dy \\
&= \phi(x) e^{-i\xi \cdot x} \\
&\quad - \int_{\mathbb{R}^2} G(\omega^2; x, y) e^{-i\xi \cdot y}\big((\nabla - i\xi) \cdot (\nabla - i\xi) + \omega^2 \varepsilon\big)\phi(y)\, dy \\
&= \phi(x) e^{-i\xi \cdot x} \\
&\quad -(\delta\varepsilon)\omega^2 \int_\Omega \sum_{j \in \mathbb{Z}^2, j \neq 0} (G(\omega^2; x, y + Nj) e^{-i\xi \cdot (y + Nj)})\phi(y)\, dy.
\end{aligned}$$

Since there exist positive constants C_1 and C_2, depending only on ω_0^2, such that

$$\left| \sum_{j\in\mathbb{Z}^2, j\neq 0} (G(\omega^2; x, y+Nj)e^{-i\xi\cdot(y+Nj)}) \right| \leq C_1 e^{-C_2\eta N} \ \forall x \in]-N,N[^2 \ \forall y \in \Omega, \tag{3.20}$$

for any $\omega^2 \in [0, \omega_0^2]$ such that $\operatorname{dist}(\omega^2, \Sigma_{\mathrm{p}}) \geq \eta$, we deduce that

$$\left\| \int_\Omega \sum_{j\in\mathbb{Z}^2, j\neq 0} (G(\omega^2; x, y+Nj)e^{-i\xi\cdot(y+Nj)})\phi(y)\, dy \right\|_{L^2_\varepsilon(]-N,N[^2)} \leq N C_1 e^{-C_2\eta N}, \tag{3.21}$$

where the constants C_1 and C_2 are different from the previous ones but have the same dependence. Recalling that $\|\phi\|_{L^2_{\varepsilon_N}(]-N,N[^2)} = 1$, we deduce the existence of $N_0 > 0$ such that

$$\|\phi\|_{L^2_\varepsilon(\mathbb{R}^2)} \geq \|\phi\|_{L^2_\varepsilon(]-N,N[^2)} \geq \frac{1}{2}. \tag{3.22}$$

On the other hand,

$$\begin{aligned}
(\Delta + \omega^2\varepsilon)u(x) &= \int_\Omega (\Delta_x + \omega^2\varepsilon) G(\omega^2; x, y)\phi(y)e^{-i\xi\cdot y}\, dy \\
&= \chi_\Omega(x)\phi(x)e^{-i\xi\cdot x} - (\varepsilon_{\mathrm{p}}(x) - \varepsilon(x))\,\omega^2 \int_\Omega G(\omega^2; x, y)\phi(y)e^{-i\xi\cdot y}\, dy \\
&= \chi_\Omega(x)\phi(x)e^{-i\xi\cdot x} - \chi_\Omega(x)(\delta\varepsilon)\omega^2 \int_\Omega G(\omega^2; x, y)\phi(y)e^{-i\xi\cdot y}\, dy \\
&= \chi_\Omega(x)\phi(x)e^{-i\xi\cdot x} \\
&\quad -\chi_\Omega(x) \int_{\mathbb{R}^2} G(\omega^2; x, y) \left(\Delta + \omega^2\varepsilon_{\mathrm{p}}\right) \left(\phi(y)e^{-i\xi\cdot y}\right)\, dy \\
&\quad +\chi_\Omega(x) \int_{\mathbb{R}^2} G(\omega^2; x, y) \left(\Delta + \omega^2\varepsilon\right) \left(\phi(y)e^{-i\xi\cdot y}\right)\, dy \\
&= \chi_\Omega(x)\phi(x)e^{-i\xi\cdot x} \\
&\quad -\chi_\Omega(x) \int_{\mathbb{R}^2} \left(\Delta_y + \omega^2\varepsilon_{\mathrm{p}}\right) G(\omega^2; x, y)\phi(y)e^{-i\xi\cdot y}\, dy \\
&\quad +\chi_\Omega(x) \int_{\mathbb{R}^2} G(\omega^2; x, y)e^{-i\xi\cdot y} \left((\nabla - i\xi)\cdot(\nabla - i\xi) + \omega^2\varepsilon\right)\phi(y)\, dy \\
&= \chi_\Omega(x)(\delta\varepsilon)\omega^2 \int_{\mathbb{R}^2} G(\omega^2; x, y)e^{-i\xi\cdot y}\phi(y) \left(\sum_{j\in\mathbb{Z}^2, j\neq 0} \chi_\Omega(y - Nj) \right) dy \\
&= \chi_\Omega(x)(\delta\varepsilon)\omega^2 \int_\Omega \sum_{j\in\mathbb{Z}^2, j\neq 0} (G(\omega^2; x, y+Nj)e^{-i\xi\cdot(y+Nj)})\phi(y)\, dy.
\end{aligned}$$

Therefore, it follows from (3.20) that

$$\left|\Delta u(x) + \omega^2\varepsilon u(x)\right| \leq |(\delta\varepsilon)|\omega_0^2|\Omega|^{\frac{1}{2}} C_1 e^{-C_2\eta N},$$

for any $x \in \Omega$. Consequently,

$$\left\| \frac{1}{\varepsilon}\Delta u + \omega^2 u \right\|_{L^2_\varepsilon(\mathbb{R}^2)} \leq \frac{|(\delta\varepsilon)|\omega_0^2|\Omega|}{\sqrt{\min(\varepsilon)}} C_1 e^{-C_2\eta N}. \tag{3.23}$$

From (3.22), we readily get

$$\operatorname{dist}(\omega^2, \Sigma) \leq C_1 e^{-C_2 \eta N},$$

where C_1 and C_2 are different from the previous ones but have the same dependence.

Since $\operatorname{dist}(\omega^2, \Sigma_{\mathrm{p}}) \geq \eta$, we easily arrive at

$$\operatorname{dist}(\omega^2, \Sigma_{\mathrm{d}}^{\eta}) \leq C_1 e^{-C_2 \eta N},$$

which ends the proof of the theorem. □

An immediate consequence of this theorem is the following corollary.

Corollary 3.1. *Suppose that the perturbation has created defect eigenvalues. Then, there exists $\eta_0 > 0$ and $N_0 \in \mathbb{N}$ such that $\Sigma_{\mathrm{d},N}^{\eta} \neq \emptyset$ for $\eta \leq \eta_0$ and $N \geq N_0$.*

Moreover, there exists $N_1 \in \mathbb{N}$ depending only on η such that the number of connected components of $\Sigma_{\mathrm{d},N}^{\eta} \cap [0, \omega_0^2]$ is at least equal to card$\left(\Sigma_{\mathrm{d}}^{\eta} \cap [0, \omega_0^2]\right)$ and the width of each component decays exponentially with N.

Proof. The proof follows immediately from the fact that each eigenvalue in $\Sigma_{\mathrm{d},N}^{\xi,\eta}$ is continuous with respect to ξ, and

$$\Sigma_N = \cup_{\xi \in \mathcal{B}_{\mathcal{N}}} \Sigma_N^{\xi}. \qquad \square$$

Remark 3.2. These results are very important and practical for determining the defect modes of 2-D photonic crystals. Indeed, after identifying the background continuous spectrum by computing numerically $\Sigma_{\mathrm{p}}^{\xi}$ for $\xi \in \mathcal{B}$, we have the gaps and we can have constants C_1 and C_2 such that

$$\left|G(\omega^2; x, y)\right| \leq C_1 e^{-C_2 \operatorname{dist}(\omega^2, \Sigma_{\mathrm{p}}) N}.$$

Then we compute Σ_N^{ξ} for some $\xi \in \mathcal{B}_{\mathcal{N}}$, and from the eigenvalues that are not located in Σ_{p} we deduce an approximation of the defect eigenvalues.

4. The TE polarization.

In this section we deal with the TE polarization. The same results hold, but the proofs are slightly different. This is a consequence of the dependence of the domain of the acoustic operator on the inverse of the dielectric function. So when we perturb ε_{p} into ε, the operator $-\nabla.\frac{1}{\varepsilon_{\mathrm{p}}}\nabla$ is transformed into $-\nabla.\frac{1}{\varepsilon}\nabla$ and we see clearly that, in general, these operators do not have the same domain. So the proofs have to be adjusted.

4.1. Definition and preliminary results.

First, we introduce some analogous notations to those in Definition 2.1.

Definition 4.1. *Let A_{p} be the self-adjoint operator defined by*

$$A_{\mathrm{p}} = -\nabla.\frac{1}{\varepsilon_{\mathrm{p}}}\nabla, \quad \text{on } L^2(\mathbb{R}^2),$$

and let Σ_p denote its spectrum.

For $\xi \in [0, 2\pi[^2$, we define A_{p}^{ξ} on $L^2(\mathbb{R}^2/\mathbb{Z}^2)$ by

$$A_{\mathrm{p}}^{\xi} = -(\nabla_x - i\xi) \cdot \frac{1}{\varepsilon_{\mathrm{p}}}(\nabla_x - i\xi),$$

and denote by $\Sigma_{\mathrm{p}}^{\xi}$ its spectrum.

We perturb the background periodic medium on a bounded domain as done in (2.11).

It has been proved that the spectrum of A_{p} is absolutely continuous and that the perturbation is relatively compact and so does not affect the essential spectrum of A_{p}. The perturbation will then result in the addition of eigenvalues of finite multiplicity to Σ_{p}.

We define ε_N, A_N, A_N^ξ, Σ_N, and Σ_N^ξ in the same manner as in section 3.1. To avoid the problem of the dependence of the domain on ε, we introduce a new operator that will have the same spectral properties as those of A_{p}.

DEFINITION 4.2. *Let B_{p} be the self-adjoint operator defined on $L^2_{\varepsilon_{\mathrm{p}}}(\mathbb{R}^2)^2$ by*

$$B_{\mathrm{p}} = -\frac{1}{\varepsilon_{\mathrm{p}}}\nabla\nabla\cdot$$

For $\xi \in [0, 2\pi[^2$, we define B_{p}^ξ on $L^2_{\varepsilon_{\mathrm{p}}}(\mathbb{R}^2/\mathbb{Z}^2)^2$ by

$$B_{\mathrm{p}}^\xi = -\frac{1}{\varepsilon_{\mathrm{p}}}(\nabla - i\xi)(\nabla - i\xi)\cdot$$

We also define B_N and B_N^ξ analogously as done for A_{p}.

The operator B_{p} is a self-adjoint periodic differential operator on $L^2_{\varepsilon_{\mathrm{p}}}(\mathbb{R}^2/\mathbb{Z}^2)^2$ but is not elliptic since its kernel has infinite dimension. Actually, the kernel is the subspace of divergence free vectors. We cannot apply the same technique as for A_{p} to prove that the spectrum of B_{p}^ξ is a set of positive eigenvalues that accumulate at infinity and that the spectrum of B_{p} is an absolutely continuous spectrum with band structure located in $\mathbb{R}^+$. It is, however, possible to extend this operator into a larger elliptic self-adjoint operator that will coincide with B_{p} on a subspace that is complementary with the kernel of B_{p} (see [23]). We can deduce then that the spectrum of B_{p} in $\mathbb{R}^+ \setminus \{0\}$ is absolutely continuous and that 0 is an eigenvalue with infinite multiplicity. This technique is used to prove the band structure of the Maxwell operator. Another way to characterize the structure of the spectrum of B_{p} is to relate it to the spectrum of A_{p}. This is given by the following theorem.

THEOREM 4.1. *For any $\xi \in [0, 2\pi[^2$, the spectra of B_{p}^ξ, B_{p}, B_N^ξ, B_N, and B are $\Sigma_{\mathrm{p}}^\xi \cup \{0\}$, Σ_{p}, $\Sigma_N^\xi \cup \{0\}$, Σ_N, and Σ, respectively. Moreover,*

(i) *The operators B_{p}^ξ and B_N^ξ have exactly the same eigenvalues as A_{p}^ξ and A_N^ξ, respectively, except for* 0 *which is an eigenvalue of A_{p}^0 and A_N^0 of multiplicity* 1 *and is not an eigenvalue of A_{p}^ξ and A_N^ξ when $\xi \neq 0$ while it is an eigenvalue of B_{p}^ξ and B_N^ξ for any ξ with infinite multiplicity.*

(ii) *The spectra of B_{p} and B_N are absolutely continuous spectra in $\mathbb{R}^+ \setminus \{0\}$ and* 0 *is an eigenvalue of infinite multiplicity.*

(iii) *The operators A and B have the same absolutely continuous spectrum and the eigenvalues have exactly the same multiplicity for A and B except for* 0 *that is an eigenvalue of B with infinite multiplicity.*

Proof. Let $\xi \in [0, 2\pi[^2$ and $\omega^2 \geq 0$. Suppose that either $\xi \neq 0$ or $\omega^2 \neq 0$ and that ω^2 is in the spectrum of A_{p}^ξ. Then there exists $\phi \in L^2(\mathbb{R}^2/\mathbb{Z}^2)$ such that $\phi \neq 0$ and

$$(\nabla - i\xi)\cdot\frac{1}{\varepsilon_{\mathrm{p}}}(\nabla - i\xi)\phi + \omega^2\phi = 0.$$

We can easily see that since ξ and ω^2 are not simultaneously equal to 0, $(\nabla - i\xi)\phi \neq 0$.

Let $\psi = \dfrac{1}{\varepsilon_{\rm p}}(\nabla - i\xi)\phi \in L^2_{\varepsilon_{\rm p}}(\mathbb{R}^2/\mathbb{Z}^2)^2$. Then

$$(\nabla - i\xi)(\nabla - i\xi)\cdot\psi + \omega^2\varepsilon_{\rm p}\psi = 0,$$

which means that ω^2 is an eigenvalue of $B^\xi_{\rm p}$. Moreover, if ϕ_1 and ϕ_2 are two linearly independent eigenvectors related to the same eigenvalue $\omega^2 \neq 0$, then $\psi_1 = \frac{1}{\varepsilon_{\rm p}}(\nabla - i\xi)\phi_1$ and $\psi_2 = \frac{1}{\varepsilon_{\rm p}}(\nabla - i\xi)\phi_2$ are linearly independent.

We conclude that all the eigenvalues of $A^\xi_{\rm p}$ except for the eigenvalue 0 of $A^0_{\rm p}$ are eigenvalues of $B^\xi_{\rm p}$. We will see that 0 is an infinite multiplicity eigenvalue of $A^0_{\rm p}$.

Conversely, let ω^2 be an eigenvalue of $B^\xi_{\rm p}$ and let $\psi \in L^2_{\varepsilon_{\rm p}}(\mathbb{R}^2/\mathbb{Z}^2)^2$ be such that $\psi \neq 0$ and satisfies

$$(\nabla - i\xi)(\nabla - i\xi)\cdot\psi + \omega^2\varepsilon_{\rm p}\psi = 0.$$

Suppose that $(\nabla - i\xi)\cdot\psi = 0$. Then, since $\psi \neq 0$, we have $\omega^2 = 0$. We also obtain that $\nabla\cdot(e^{-i\xi\cdot x}\psi) = 0$, or, equivalently, that there exists $\alpha \in L^2(\mathbb{R}^2/\mathbb{Z}^2)$ such that

$$e^{-i\xi\cdot x}\psi = \nabla\times(\alpha e^{-i\xi\cdot x}),$$

where $\nabla\times\alpha = (\partial_2\alpha, -\partial_1\alpha)$. It follows that

$$\psi = \nabla\times\alpha - i\begin{pmatrix}\xi_2\\ -\xi_1\end{pmatrix}\alpha.$$

Hence, 0 is an eigenvalue of $B^\xi_{\rm p}$ with infinite multiplicity.

In the case where $(\nabla - i\xi)\cdot\psi \neq 0$, let $\phi = (\nabla - i\xi)\cdot\psi \in L^2(\mathbb{R}^2/\mathbb{Z}^2)$. Then,

$$(\nabla - i\xi)\cdot\frac{1}{\varepsilon_{\rm p}}(\nabla - i\xi)\phi + \omega^2\phi = 0,$$

which means that ω^2 is an eigenvalue of $A^\xi_{\rm p}$. We can also show that if ψ_1 and ψ_2 are two linearly independent eigenvectors of $B^\xi_{\rm p}$ related to the same eigenvalue $\omega^2 \neq 0$, then $\phi_1 = (\nabla - i\xi)\cdot\psi_1$ and $\phi_2 = (\nabla - i\xi)\cdot\psi_2$ are linearly independent.

The same proof holds for A^ξ_N and B^ξ_N and for the eigenvalues of A and B. □

As a consequence of the above theorem, we can recover the properties of the spectra of A^ξ_N and A_N by studying those of B^ξ_N and B_N to which we can apply mainly the same technique as in the TM case since their domain does not depend on ε.

To this end we need to give an analogous result to Lemma 2.1 for the operator $B_{\rm p}$.

LEMMA 4.1. *For any $z \notin \Sigma_{\rm p}$ and $l > 0$ we have*

$$\|\chi_{x,l}R(z)\chi_{y,l}\| \leq \left(\frac{9}{\eta}\right)e^{(\sqrt{2}l/4)}e^{-m_z|x-y|} \quad \forall x, y \in \mathbb{R}^2, \tag{4.1}$$

with

$$m_z = \frac{\eta}{4(2\varepsilon_-^{-1} + |z| + \eta)}, \tag{4.2}$$

where $\eta = dist(z, \Sigma_{\rm p})$, $\varepsilon_- = \min_{x\in\mathbb{R}^2}\varepsilon_{\rm p}(x)$, and $\chi_{x,l}$ is the characteristic function of the cube $\{y = (y_1, y_2) \in \mathbb{R}^2 : |y_1 - x_1| < \frac{l}{2}$ and $|y_2 - x_2| < \frac{l}{2}\}$.

Proof. The proof is exactly the same as the one for the Helmholtz operator which uses a Combe–Thomas argument and can be found in [13], [14], [15].

Let B_a denote the operators formally given by

$$B_a = e^{a\cdot x} B_p e^{-a\cdot x}, \quad a \in \mathbb{R}^2, \tag{4.3}$$

as the closed densely defined operators (uniquely) introduced by the corresponding quadratic forms defined on $C_0^1(\mathbb{R}^2)$ by

$$\mathcal{B}_a[\psi] = \left\langle \nabla\cdot(e^{a\cdot x}\psi), \frac{1}{\varepsilon_{\mathrm{p}}(x)}\nabla\cdot(e^{-a\cdot x}\psi)\right\rangle = \left\langle (\nabla + a)\cdot\psi, \frac{1}{\varepsilon_{\mathrm{p}}(x)}(\nabla - a)\cdot\psi\right\rangle. \tag{4.4}$$

We also introduce the quadratic form $\mathcal{Q}_a$ as

$$\begin{aligned}\mathcal{Q}_a[\psi] &= \mathcal{B}_a[\psi] - \mathcal{B}_0[\psi]\\ &= \left\langle a\cdot\psi, \frac{1}{\varepsilon_{\mathrm{p}}(x)}\nabla\cdot\psi\right\rangle - \left\langle \nabla\cdot\psi, \frac{1}{\varepsilon_{\mathrm{p}}(x)}a\cdot\psi\right\rangle\\ &\quad - \left\langle a\cdot\psi, \frac{1}{\varepsilon_{\mathrm{p}}(x)}a\cdot\psi\right\rangle.\end{aligned}$$

Since

$$\left|\left\langle a\cdot\psi, \frac{1}{\varepsilon_{\mathrm{p}}(x)}\nabla\cdot\psi\right\rangle\right| \le \frac{1}{2}|a|\left(\left\langle \psi, \frac{1}{\varepsilon_{\mathrm{p}}(x)}\psi\right\rangle + \left\langle \nabla\cdot\psi, \frac{1}{\varepsilon_{\mathrm{p}}(x)}\nabla\cdot\psi\right\rangle\right), \tag{4.5}$$

we have

$$|\mathcal{Q}_a[\psi]| \le |a|\mathcal{B}_0[\psi] + |a|(1+|a|)\varepsilon_-^{-1}\|\psi\|^2 \quad \forall\psi \in C_0^1(\mathbb{R}^2). \tag{4.6}$$

Then we require $|a| < 1$ and use Theorem VI.3.9 in [21] to conclude that $\mathcal{B}_a$ is a closable sectorial form and define B_a as the unique m-sectorial operator associated with it. If, in addition, $z \notin \Sigma_p$ and

$$\Lambda \equiv 2\left\|(|a|(1+|a|)\varepsilon_-^{-1} + |a|B_{\mathrm{p}})(B_{\mathrm{p}} - zI)^{-1}\right\| < 1, \tag{4.7}$$

we can conclude that $z \notin \Sigma_a$ (the spectrum of B_a) and

$$\|R_a(z) - R_0(z)\| \le \frac{4\Lambda}{(1-\Lambda)^2}\|R_0(z)\|, \tag{4.8}$$

where $R_a(z) = (B_a - zI)^{-1}$.

Since

$$\begin{aligned}\Lambda &= 2\left\|(|a|(1+|a|)\varepsilon_-^{-1} + |a|z)(B_{\mathrm{p}} - zI)^{-1} + |a|\right\|\\ &\le 2|a|(((1+|a|)\varepsilon_-^{-1} + |z|)\eta^{-1} + 1)\\ &\le 2|a|((2\varepsilon_-^{-1} + |z|)\eta^{-1} + 1),\end{aligned}$$

it is sufficient to take

$$|a| < \frac{\eta}{2(2\varepsilon_-^{-1} + |z| + \eta)}, \tag{4.9}$$

to ensure $\Lambda < 1$. In fact, we take

$$|a| < m_z = \frac{\eta}{4(2\varepsilon_-^{-1} + |z| + \eta)}, \tag{4.10}$$

so that we get $\Lambda < \frac{1}{2}$. It follows that

$$\|R_a(z)\| \leq \left(1 + \frac{4\Lambda}{(1-\Lambda)^2}\right) \|R_0(z)\| \leq \frac{9}{\eta}. \tag{4.11}$$

Now, let $x_0, y_0 \in \mathbb{R}^2$, $l > 0$, and take

$$a = \frac{m_z}{|x_0 - y_0|}(x_0 - y_0).$$

We have

$$\begin{aligned}\|\chi_{x_0,l} R_0(z) \chi_{y_0,l}\| &= \|\chi_{x_0,l} e^{-a\cdot x} R_a(z) e^{a\cdot x} \chi_{y_0,l}\| \\ &= e^{-m_z\cdot|x_0-y_0|} \|\chi_{x_0,l} e^{-a\cdot(x-x_0)} R_a(z) e^{a\cdot(x-y_0)} \chi_{y_0,l}\| \\ &\leq \frac{9}{\eta} e^{-m_z\cdot|x_0-y_0|} \|\chi_{x_0,l} e^{-a\cdot(x-x_0)}\|_\infty \|\chi_{y_0,l} e^{-a\cdot(x-y_0)}\|_\infty.\end{aligned}$$

We also notice that

$$\|\chi_{x_0,l} e^{\pm a\cdot(x-x_0)}\|_\infty \leq e^{\frac{l}{\sqrt{2}} m_z},$$

and since $m_z \leq \frac{1}{4}$, the theorem is proved. □

As a consequence, the matricial Green's kernel of B_p has a similar exponential decay as the Green's kernel of A_p. Let $\omega^2 \notin \Sigma_p$, we define the matricial Green's kernel $K(\omega^2; x, y)$ as the solution to

$$\nabla\nabla \cdot K(\omega^2; x, y) + \omega^2 \varepsilon_\mathrm{p} K(\omega^2; x, y) = \delta(x-y) \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{4.12}$$

Here we shall impose an outgoing radiation condition on $\nabla \cdot K$ in order to ensure uniqueness. As a direct consequence of the previous lemma the following result holds.

Corollary 4.1. *For any $\omega_0^2 > 0$, there exist two positive constants C_1 and C_2, depending only on $\omega_0^2 > 0$, such that for any $\omega^2 \notin \Sigma_\mathrm{p} \cap (0, \omega_0^2)$,*

$$|K(\omega^2; x, y)| \leq C_1 e^{-C_2 \, dist(\omega^2, \Sigma_\mathrm{p})|x-y|}, \quad \text{for } |x-y| \to +\infty. \tag{4.13}$$

Now we are ready to prove the analogous results to those concerning the TM polarization.

4.2. Convergence of the "continuous spectrum." As done for the TM polarization, we give an estimate of the perturbation brought to the continuous spectrum of A_p by the supercell method.

Theorem 4.2. *For any $\omega_0 > 0$ and $N_0 \in \mathbb{N}$, there exists $C > 0$, depending only on ω_0, N_0 and Ω, such that*

$$\max_{\omega^2 \in \cup_{k\in]-N+1,N-1[^2\cap\mathbb{N}^2} \Sigma_\mathrm{p}^{\xi+k\pi/N} \cap [0,\omega_0^2]} dist(\omega^2, \Sigma_N^\xi) \leq \frac{C}{N^2}, \tag{4.14}$$

for any $N \geq N_0$ *and any* $\xi \in \mathcal{B}_{\mathcal{N}}$.

Proof. Let $k \in]-N+1, N-1[^2 \cap \mathbb{N}^2$ and $\xi \in \mathcal{B}_{\mathcal{N}}$. Let ω^2 be in $\Sigma_{\mathrm{p}}^{\xi+k\pi/N} \cap [0, \omega_0^2]$. If $\omega^2 = 0$, then necessarily $\xi = 0$ and $k = 0$ and in that case we know that $0 \in \Sigma_N^{\xi}$.

Let us consider now $\omega^2 \neq 0$. From Theorem 4.1, we deduce that ω^2 is in the spectrum of $B_{\mathrm{p}}^{\xi+k\pi/N}$.

Since $\xi + k\pi/N \in \mathcal{B}$, there exists $\phi \in L^2_{\varepsilon_{\mathrm{p}}}(\mathbb{R}^2/\mathbb{Z}^2)^2$ with unit norm such that

$$\left(\nabla - i\left(\xi + \frac{k\pi}{N}\right)\right)\left(\nabla - i\left(\xi + \frac{k\pi}{N}\right)\right) \cdot \phi + \omega^2 \varepsilon_{\mathrm{p}} \phi = 0. \tag{4.15}$$

Let $\tilde{\phi}$ be defined in $L^2_{\varepsilon}(\mathbb{R}^2/2N\mathbb{Z}^2)^2$ as

$$\tilde{\phi}(x) = \phi(x) e^{-i\frac{\pi}{N}k\cdot x}. \tag{4.16}$$

For any $0 < C_0 < 4$, there exits an integer N_0 independent of $\tilde{\phi}$, such that for any $N > N_0$, we have $\|\tilde{\phi}\|_{L^2_{\varepsilon}(\mathbb{R}^2/2N\mathbb{Z}^2)^2} > C_0 N^2$, and it satisfies the following equation:

$$(\nabla - i\xi)(\nabla - i\xi) \cdot \tilde{\phi} + \omega^2 \varepsilon_{\mathrm{p}} \tilde{\phi} = 0, \tag{4.17}$$

which can be written as

$$(\nabla - i\xi)(\nabla - i\xi) \cdot \tilde{\phi} + \omega^2 \varepsilon \tilde{\phi} = -\chi_{\Omega}(\delta\varepsilon)\omega^2 \tilde{\phi}. \tag{4.18}$$

We prove then in the same way as done for the TM case that there exists an eigenvalue ω_ξ^2 belonging to the spectrum of B_N^{ξ}, that is $\Sigma_N^{\xi} \cup \{0\}$, satisfying

$$|\omega^2 - \omega_\xi^2| \leq \frac{C}{N^2},$$

since we considered $\omega^2 \neq 0$, for N large enough $\omega_\xi^2 \neq 0$ and then $\omega_\xi^2 \in \Sigma_N^{\xi}$. This means that

$$\mathrm{dist}(\omega^2, \Sigma_N^{\xi}) \leq \frac{C}{N^2}.$$

The theorem is then proved. □

4.3. Convergence of the defect eigenvalues. Analogously to the TM polarization, we give a characterization of the part of the spectrum Σ_N corresponding to the defect eigenvalues of Σ. We use the notations introduced in Definition 3.3. The following proposition holds.

PROPOSITION 4.1. *For every gap* $]a, b[$ *in* Σ_{p} *(*$0 < a < b$*) satisfying* $]a, b[\cap \Sigma = \emptyset$*, there exists* $N_1 \in \mathbb{N}$ *such that, for* $N \geq N_1$, $\Sigma_N \cap]a, b[= \emptyset$.

Proof. Suppose that the proposition is false. Then for any $N_0 \in \mathbb{N}$ there exists $N \geq N_0$ and $\omega_N^2 \in]a, b[\cap \Sigma_N$. This means that ω_N^2 is in the spectrum of B_N. Then there exist $\xi_N \in \mathcal{B}_{\mathcal{N}}$ and $\phi_N \in L^2_{\varepsilon_N}(\mathbb{R}^2/2N\mathbb{Z}^2)^2$ with unit norm such that

$$(\nabla - i\xi_N)(\nabla - i\xi_N) \cdot \phi_N + \omega_N^2 \varepsilon_N \phi_N = 0, \quad \text{in } L^2_{\varepsilon_N}(\mathbb{R}^2/2N\mathbb{Z}^2)^2. \tag{4.19}$$

Now, define $\tilde{\phi}_N$ in $L^2_{\varepsilon}(\mathbb{R}^2)$ as

$$\tilde{\phi}_N(x) = \int_{\Omega} K(\omega_N^2; x, y) e^{-i\xi_N \cdot y} \phi_N(y)\, dy. \tag{4.20}$$

Using $\tilde{\phi}_N$, we prove in a similar way as for Proposition 3.2 that

$$\frac{\|\frac{1}{\varepsilon}\nabla\nabla\cdot\tilde{\phi}_N+\omega_N^2\tilde{\phi}_N\|_{L^2_\varepsilon(\mathbb{R}^2)^2}}{\|\tilde{\phi}_N\|_{L^2_\varepsilon(\mathbb{R}^2)^2}}\leq C_1e^{-C_2N}, \tag{4.21}$$

for some positive constants C_1 and C_2. Since ω_N^2 is away from 0, then

$$\operatorname{dist}(\omega_N^2,\Sigma)\leq C_1e^{-C_2N}, \tag{4.22}$$

which leads to a contradiction. □

Now we give the main result for the TE case about the convergence of the eigenvalues of the supercell corresponding to the defect.

THEOREM 4.3. *Suppose that the perturbation has created defect eigenvalues. Then, there exists $\eta_0>0$ and $N_0\in\mathbb{N}$ such that for any $\eta\leq\eta_0$ and $N\geq N_0$,*

$$\Sigma_{\mathrm{d},N}^{\xi,\eta}\neq\emptyset\quad\forall\xi\in\mathcal{B}_{\mathcal{N}}.$$

Moreover, for any $\omega_0^2>0$ and $\eta\leq\eta_0$, there exists two positive constants C_1 and C_2 depending only on ω_0^2 such that for any $\xi\in\mathcal{B}_{\mathcal{N}}$,

$$dist_{\mathcal{H}}(\Sigma_{\mathrm{d},N}^{\xi,\eta}\cap[0,\omega_0^2],\Sigma_{\mathrm{d}}^{\eta}\cap[0,\omega_0^2])\leq C_1e^{-C_2\eta N}. \tag{4.23}$$

Proof. Since we deal with a part of the spectrum that is away from 0, the statements are exactly the same when considering the spectra related to B_{p} instead of A_{p}. The proof then becomes similar to the one of Theorem 3.2. □

Note that the Corollary 3.1 holds for the TE polarization.

5. Numerical experiments.

The numerical simulations presented in this section are computed with the *MIT Photonic-Bands* (MPB) package [20]. We consider a 2-D photonic crystal in which the dielectric permittivity takes the values of 1 and 12. The structure of the crystal is shown in Figure 5.1 where the dark area corresponds to dielectric permittivity 12.

We investigate only the TE polarization. We compute the TE-spectrum of this structure for the first 8 bands. This is shown in Figure 5.2 where we notice the presence of two gaps between the first and the second bands and between the second and the third bands. The singularities of the last band come from the fact that it crosses the following band, which is not represented on the diagram.

Then we introduce a defect to this periodic structure by changing the dielectric permittivity in one disc from 1 into 12. The corresponding 7×7 supercell is represented in Figure 5.3. We compute the TE-spectrum in the supercell for a fixed wave number

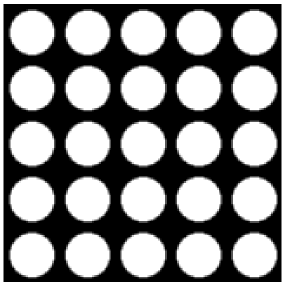


FIG. 5.1. *The periodic structure.*

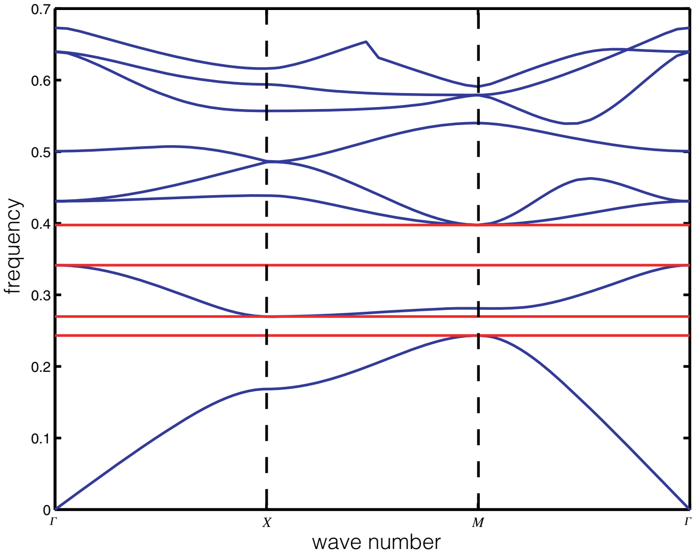


Fig. 5.2. *TE-spectrum of the periodic structure.*

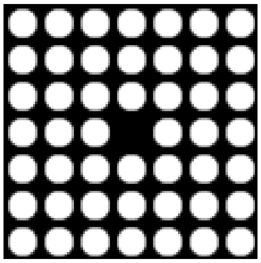


Fig. 5.3. *The 7×7 supercell.*

Table 5.1
Defect frequencies and relative difference with the 7×7 supercell.

Supercell size	3×3	5×5	7×7
Defect frequency 1	0.3706 0.6%	0.3687 0.05%	0.3685
Defect frequency 2	0.3574 0.3%	0.3563 ≤00.3%	0.3563

and for different sizes of the supercell (3,5,7). The results are shown in Figure 5.4. The horizontal dashed lines delimit the gaps of the periodic medium.

We notice clearly the presence of two defect eigenvalues in the second gap. The values of the defect frequencies and the relative difference with the 7×7 supercell results are shown in Table 5.1.

The convergence of the continuous spectrum is in $1/N$ but the multiplicative constant depends on the dispersion of the band considered (the differential of the frequency with respect to the wave vector). This explains why the convergence in the first band (the most dispersive) is the lowest.

In Figure 5.5 we plotted the defect frequencies against the wave number. In the 3×3 supercell, the defect frequencies oscillate with an amplitude about 1% while the oscillation is about 0.1% in the 5×5 supercell and about 0.05% in the 7×7 supercell.

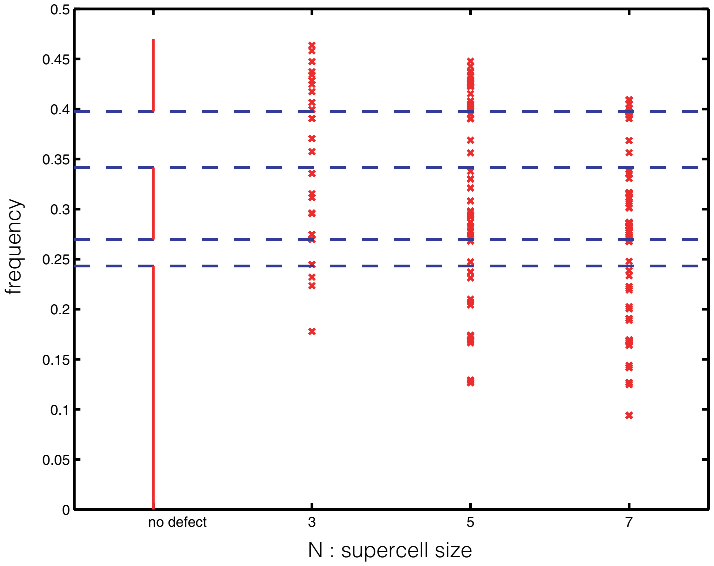
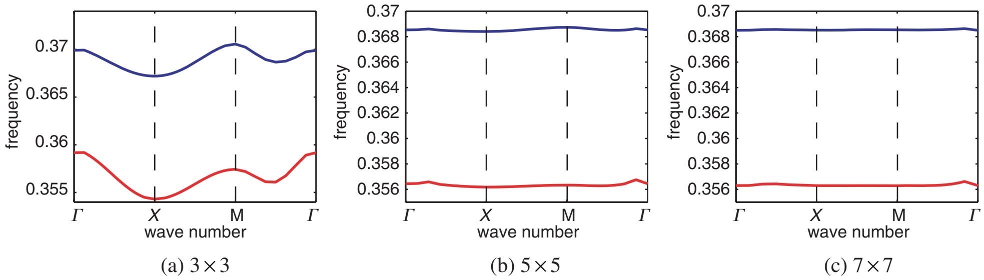


Fig. 5.4. *TE-spectrum of the supercell.*



(a) 3×3

(b) 5×5

(c) 7×7

Fig. 5.5. *Dependence of the defect frequencies on the wave number.*

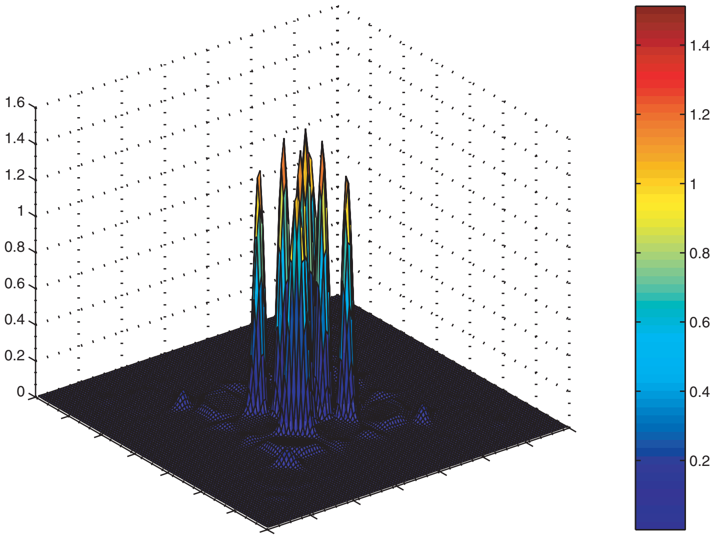


Fig. 5.6. *Energy distribution in the first defect mode.*

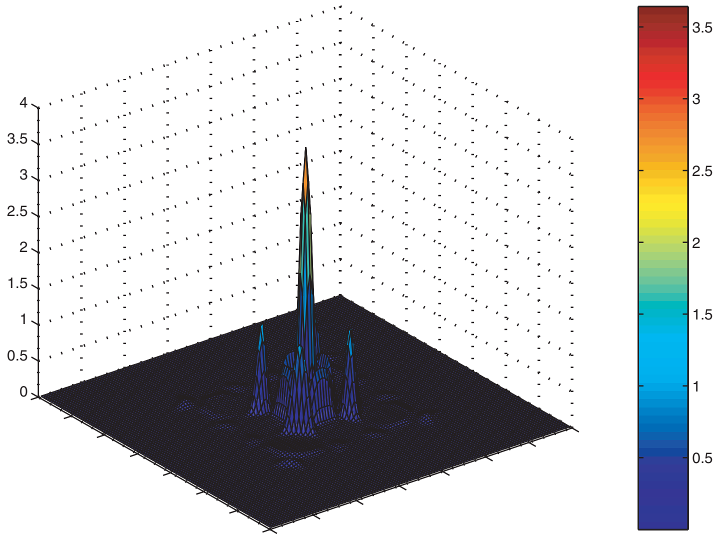


FIG. 5.7. *Energy distribution in the second defect mode.*

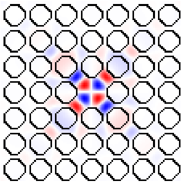


FIG. 5.8. *Magnetic field distribution in the first defect mode.*

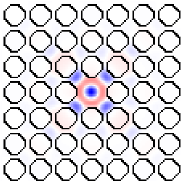


FIG. 5.9. *Magnetic field distribution in the second defect mode.*

Finally, in Figures 5.6–5.9 we represent the energy distribution and the magnetic field for the defect modes in the case of the 7 × 7 supercell.

6. Conclusion. In this paper we presented a rigorous proof of the convergence of the supercell method. The convergence speed is related to the exponential decay of the Green's function. If (ω_a^2, ω_b^2) is a gap of the photonic crystal (ω_a^2, ω_b^2 belong to the spectrum), then it was proved that for $\omega^2 \in (\omega_a^2, \omega_b^2)$, the exponential decay of the Green's function is of the form

$$O\Big(\exp\Big(-C\sqrt{|\omega^2-\omega_a^2||\omega^2-\omega_b^2|}\;|x|\Big)\Big). \tag{6.1}$$

It follows that the convergence of the defect eigenvalues will be slower when they are closer to the edges of the gap. This is not an important problem since these modes are useless. Actually, we are interested in the localization property of the defect modes which is weak for such eigenvalues.

Finally, we remark that this method becomes very costly when looking for defects lying over few bands. For example, if we look for a defect eigenvalue lying in a gap between the fourth and the fifth band, when computing the spectrum of the 5×5 supercell, every band will contribute with 5^2 eigenvalues and the defect eigenvalue will be the 101st eigenvalue which requires many calculations. We believe that it should be possible to determine such eigenvalues in a faster way with integral operator methods.

REFERENCES

[1] H. Ammari and F. Santosa, *Guided waves in a photonic bandgap structure with a line defect*, SIAM J. Appl. Math., 64 (2004), pp. 2018–2033.

[2] W. Axmann and P. Kuchment, *An efficient finite element method for computing spectra of photonic and acoustic band-gap materials.* I. *Scalar case,* J. Comput. Phys., 150 (1999), pp. 468–481.

[3] J. M. Barbaroux, J. M. Combes, and P. D. Hislop, *Localization near bad edges for random Schrödinger operators,* Helv. Phys. Acta, 70 (1997), pp. 16–43.

[4] A. Bjarklev, *Optical Fiber Amplifiers: Design and System Application*, Artech House, Boston, 1993.

[5] J. Broeng, D. Mogilevstev, S. E. Barkou, and A. Bjarklev, *Photonic crystals fibers: a new class of optical waveguides*, Optical Fiber Technol., 5 (1999), pp. 305–330.

[6] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[7] J. M. Combes and L. Thomas, *Asymptotic behavior of eigenfunctions for multiparticle Schrödinger operators*, Comm. Math. Phys., 34 (1973), pp. 251–270.

[8] S. J. Cox and D. C. Dobson, *Band structure optimization of two-dimensional photonic crystals in H-polarization*, J. Comput. Phys., 158 (2000), pp. 214–224.

[9] D. C. Dobson, *An efficient method for band structure calculations in* 2*D photonic crystals*, J. Comput. Phys., 149 (1999), pp. 363–376.

[10] D. C. Dobson, J. Gopalakrishnan, and J. E. Pasciak, *An efficient method for band structure calculations in* 3*D photonic crystals*, J. Comput. Phys., 161 (2000), pp. 668–679.

[11] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, Springer-Verlag, Berlin, 1983.

[12] A. Figotin and Y. A. Godin, *The computation of spectra of some* 2*D photonic crystals*, J. Comput. Phys., 136 (1997), pp. 585–598.

[13] A. Figotin and A. Klein, *Localization of light in lossless inhomogeneous dielectrics*, J. Opt. Soc. Amer. A, 15 (1998), pp. 1423–1435.

[14] A. Figotin and A. Klein, *Localized classical waves created by defects*, J. Statist. Phys., 86 (1997), pp. 165–177.

[15] A. Figotin and A. Klein, *Midgap defect modes in dielectric and acoustic media*, SIAM J. Appl. Math., 58 (1998), pp. 1748–1773.

[16] A. Figotin and P. Kuchment, *Band-gap structure of spectra of periodic dielectric and acoustic media.* I: *Scalar model*, SIAM J. Appl. Math., 56 (1996), pp. 68–88.

[17] A. Figotin and P. Kuchment, *Band-gap structure of spectra of periodic dielectric and acoustic media.* II: 2*D photonic crystals*, SIAM J. Appl. Math., 56 (1996), pp. 1561–1620.

[18] J. D. Joannopoulos, R. D. Meade, and J. N. Winn, *Photonic Crystals. Molding the Flow of Light*, Princeton University Press, Princeton, NJ, 1995.

[19] S. G. Johnson and J. D. Joannopoulos, *Photonic Crystals. The Road from Theory to Practice*, Kluwer Acad. Publ., Dordrecht, The Netherlands, 2002.

[20] S. G. Johnson and J. D. Joannopoulos, *Block-iterative frequency-domain methods for Maxwell's equations in a planewave basis*, Optics Express, 8 (2001), pp. 173–190.

[21] T. Kato, *Perturbation Theory for Linear Operators*, Die Gundlehren der Math. Wissenschoften, Band 132, Springer-Verlag, New York, 1966.

[22] J. C. Knight, J. Broeng, T. A. Birks, and P. St. J. Russel, *Photonic band gap guidance in optical fibers*, Science, 282 (1998), pp. 1476–1478.

[23] P. Kuchment, *The mathematics of photonic crystals*, in Mathematical Modelling in Optical Science, Bao, Cowsar, and Masters, eds., Frontiers Appl. Math. 22, SIAM, Philadelphia, 2001, pp. 207–272.

[24] P. Kuchment and B. S. Ong, *On guided waves in photonic crystal waveguides*, Contemp. Math., 3391 (2003), pp. 105–115.

[25] A. Morame, *The absolute continuity of the spectrum of Maxwell operator in a periodic media*, J. Math. Phys., 41 (2000), pp. 7099–7108.

[26] N. A. Mortensen, *Effective area of photonic crystal fibers*, Optics Express, 10 (2002), pp. 341–348.

[27] J. C. Nédélec, *Acoustic and Electromagnetic Equations. Integral Representations for Harmonic Problems*, Springer-Verlag, New York, 2001.

[28] R. Reed and B. Simon, *Methods of Modern Mathematical Physics* I*: Functional Analysis*, Academic Press, New York, 1972.

[29] K. Sakoda, *Optical Properties of Photonic Crystals*, Springer-Verlag, Berlin, 2001.

[30] S. Soussi, *Modeling photonic crystal fibers*, preprint 2004.

[31] E. Yablonovitch, *Inhibited spontaneous emission in solid-state physic and electronics*, Phys. Rev. Lett., 58 (1987), pp. 2059–2062.

SIAM J. Numer. Anal.
Vol. 43, No. 3, pp. 1202–1230

THE COMPUTATION OF CONICAL DIFFRACTION COEFFICIENTS IN HIGH-FREQUENCY ACOUSTIC WAVE SCATTERING*

B. D. BONNER†, I. G. GRAHAM†, AND V. P. SMYSHLYAEV†

Abstract. When a high-frequency acoustic or electromagnetic wave is scattered by a surface with a conical point, the component of the asymptotics of the scattered wave corresponding to diffraction by the conical point can be represented as an asymptotic expansion, valid as the wave number $k \to \infty$. The *diffraction coefficient* is the coefficient of the principal term in this expansion and is of fundamental interest in high-frequency scattering. It can be computed by solving a family of homogeneous boundary value problems for the Laplace–Beltrami–Helmholtz equation (parametrized by a complex wave number–like parameter ν) on a portion of the unit sphere bounded by a simple closed contour ℓ, and then integrating the resulting solutions with respect to ν. In this paper we give the numerical analysis of a method for carrying out this computation (in the case of acoustic waves) via the boundary integral method applied on ℓ, emphasizing the practically important case when the conical scatterer has lateral edges. The theory depends on an analysis of the integral equation on ℓ, which shows its relation to the corresponding integral equation for the planar Helmholtz equation. This allows us to prove optimal convergence for piecewise polynomial collocation methods of arbitrary order. We also discuss efficient quadrature techniques for assembling the boundary element matrices. We illustrate the theory with computations on the classical canonical open problem of a trihedral cone.

Key words. acoustic wave scattering, high-frequency asymptotics, diffraction coefficients, conical points, lateral edges, boundary integral method, collocation, mesh grading, convergence

AMS subject classifications. 65N38, 65R20, 35P25, 78A45

DOI. 10.1137/040603358

1. Introduction.

When an incident plane acoustic or electromagnetic wave is scattered by a bounded impenetrable (three-dimensional) obstacle, the asymptotic behavior of the scattered wave when the frequency is large is described by the classical geometric theory of diffraction (GTD) [28]. The asymptotics of the scattered field when the wave number $k \to \infty$ is known from the GTD to be composed of a number of components corresponding to "reflections" or "diffractions" by particular parts of the boundary. Along with the component corresponding to simple reflection of nongrazing incident waves at smooth parts of the obstacle, or a more complicated grazing incidence which leads to asymptotics in the shadow [27] and special boundary-layer asymptotics in the "penumbra" (see, e.g., [7] and the references therein), the scattered wave's asymptotics may also contain components arising from diffraction by nonsmooth "singular" points of the scattering surface, such as edges or conical points. From the GTD [28] (and its further developments), the principal parts of those components are known to be described by the (diffracted component of the) far field of waves scattered by the *tangent cone* at the singular point(s). This is due to the so-called principle of localization (which is the essence of the GTD). Many authors

*Received by the editors January 19, 2004; accepted for publication (in revised form) August 17, 2004; published electronically September 23, 2005. This work was completed while the second author was visiting the program on Computational Challenges in PDEs at the Newton Institute for Mathematical Sciences, Cambridge, UK.

http://www.siam.org/journals/sinum/43-3/60335.html

†Department of Mathematical Sciences, University of Bath, Bath BA2 7AY, UK (bradley_bonner@yahoo.co.uk, I.G.Graham@bath.ac.uk, vps@maths.bath.ac.uk).

have considered the problem of describing the asymptotics of the diffracted wave for various "canonical" cones (see, e.g., [12, 15, 16, 11, 6] and the references therein).

This problem has been studied in detail when the obstacle is a cone with a smooth lateral surface, and ideal boundary conditions are applied; see, e.g., [12, 6] and the references therein, where explicit formulae for the principal asymptotics of the diffracted wave were derived. (By "ideal" we mean pure Dirichlet or Neumann boundary conditions in the acoustic case and perfectly conducting boundary conditions in the electromagnetic case. See [9, 2, 10] for some results on nonideal boundary conditions.)

For example, consider the scalar (acoustic) case, with an incident plane wave $U^{inc}(\mathbf{x}) = \exp(-ik\boldsymbol{\omega}_0 \cdot \mathbf{x})$, with the point $\boldsymbol{\omega}_0 \in S^2$ (the unit sphere in $\mathbb{R}^3$), describing the direction of incidence. Then both the scattered wave U^{sc} and the total wave $U := U^{inc} + U^{sc}$ satisfy the three-dimensional Helmholtz equation, $(\Delta + k^2)U = 0$, in the domain of propagation, and U^{sc} satisfies an appropriate version of the radiation conditions. The theory in [33, 4, 6] describes the behavior of the diffracted component $U^{diff}(\mathbf{x})$ of $U^{sc}(\mathbf{x})$ at any point $\mathbf{x}$ in the domain of propagation. Using spherical coordinates centered at the conical point—$\mathbf{x} = r\boldsymbol{\omega}$ with $\boldsymbol{\omega} \in S^2$ and $r > 0$ denoting the distance of $\mathbf{x}$ from the conical point—it follows from the general recipes of the GTD that (with either Dirichlet or Neumann conditions imposed on the surface of the scatterer) U^{diff} has the asymptotic representation

$$U^{diff}(\mathbf{x}, k, \boldsymbol{\omega}_0) = 2\pi \frac{\exp(ikr)}{kr} f(\boldsymbol{\omega}, \boldsymbol{\omega}_0) + O((kr)^{-2}), \qquad k \to \infty. \tag{1.1}$$

Here the distribution $f(\boldsymbol{\omega}, \boldsymbol{\omega}_0)$, which is infinitely smooth everywhere except at the so-called *singular directions* (where it is typically infinite), is the important *diffraction coefficient* (also known as the kernel of the *scattering matrix*) and describes the intensity of the diffracted wave in the particular direction $\boldsymbol{\omega}$. (See, e.g., [6, 13] and the references therein for precise descriptions of the distributional spaces.)

This paper deals with the numerical analysis and implementation of methods for computing $f(\boldsymbol{\omega}, \boldsymbol{\omega}_0)$. Following [4] and [6], to obtain a formula for f, we take O to be the vertex of the conical obstacle, Ξ (which is indicated by dotted lines in Figure 1), and let M denote the portion of the unit sphere S^2 which is exterior to Ξ. M is a submanifold of S^2 with boundary, which we denote by ℓ (see Figure 1 again). Let Δ^* denote the Laplace–Beltrami operator on S^2 and introduce the *spherical* Green's function $g(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu)$ on M (also known as a "spectral function"), satisfying

$$(\Delta^* + \nu^2 - 1/4) g(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) = \delta(\boldsymbol{\omega} - \boldsymbol{\omega}_0), \quad \boldsymbol{\omega}, \boldsymbol{\omega}_0 \in M \text{ and } \nu \in \mathbb{C}\ , \tag{1.2}$$

where δ denotes the Dirac delta function and the differentiation on the left-hand side is with respect to $\boldsymbol{\omega}$. As a function of $\boldsymbol{\omega}$, g is also required to satisfy a Dirichlet or Neumann boundary condition on ℓ (whichever is given in the original scattering problem). Once g is known, the diffraction coefficient in (1.1) is then given by the formula (see [33, 6])

$$f(\boldsymbol{\omega}, \boldsymbol{\omega}_0) = \lim_{s \to 0+} \frac{i}{\pi} \int_\gamma \exp(-i\nu\pi - s\nu) g(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) \nu d\nu. \tag{1.3}$$

The integrals in (1.3) are known to converge uniformly as $s \to 0+$ away from the singular directions; cf. [8]. The infinite integration contour γ in (1.3) has to be chosen in the complex plane, so that the (positive) numbers $\sqrt{\lambda_j}$ (where λ_j ranges over all eigenvalues of the self-adjoint operator $-\Delta^* + 1/4$ on M, subject to the appropriate

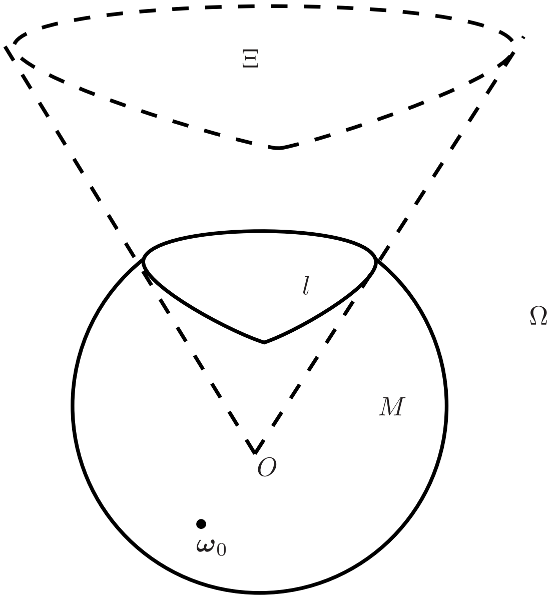


FIG. 1. *Geometry of obstacle.*

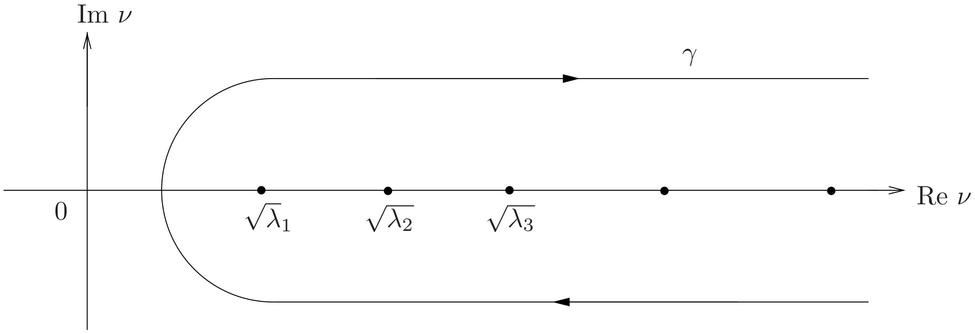


FIG. 2. *Contour of integration.*

boundary condition on ℓ) lie on its right and also so that when $\text{Re}(\nu) \to \infty$, along γ, $\text{Im}(\nu) \to \pm a$ for some constant $a > 0$ (see [6]). The function $g(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu)$ is known from spectral theory to be analytic in ν, except for poles at $\nu = \sqrt{\lambda_j}$, provided $\boldsymbol{\omega} \neq \boldsymbol{\omega}_0$—see Figure 2.

Thus the computational procedure for realizing the asymptotic formula (1.1) requires the following: (i) the computation of the Green's function $g(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu)$ for all required incidence directions $\boldsymbol{\omega}_0$ and observation directions $\boldsymbol{\omega} \in M$ and (ii) the computation of the integral in (1.3) for sufficiently small positive s, by quadrature. Note that (ii) in turn implies that $g(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu)$ must be evaluated for sufficiently many $\nu \in \gamma$ to ensure an accurate answer.

The Green's function g in (1.3) can be replaced (see [33, 4]) by its regular part $g^r := g - g_0$, where g_0 is the (known) fundamental solution for the operator $(\Delta^* + \nu^2 - 1/4)$ on all of S^2 (see section 2). Moreover, for certain configurations of $\boldsymbol{\omega}, \boldsymbol{\omega}_0$ (which, say, in the case of a smooth, fully illuminated, and convex cone corresponds to the direction of observation $\boldsymbol{\omega}$ with no reflected wave [33, 4, 25]—see (2.1) for a precise statement), the right-hand side of (1.3) can be transformed by deforming the contour of integration γ onto the imaginary axis and then interchanging the limit with the integral. These modifications yield the simpler formula

$$f(\boldsymbol{\omega},\boldsymbol{\omega}_0) = -\frac{i}{\pi}\int_{-\infty}^{\infty} \exp(\tau\pi) g^r(\boldsymbol{\omega},\boldsymbol{\omega}_0, i\tau)\tau d\tau, \tag{1.4}$$

with the integral convergent absolutely. In fact it can be shown (see [13, section 6.4] and the references therein) that

$$\exp(\tau\pi) g^r(\boldsymbol{\omega},\boldsymbol{\omega}_0, i\tau) \sim \left\{ \begin{array}{ll} \exp(\alpha_1\tau) \ , & \tau \to -\infty \ , \\ \exp(-\alpha_2\tau) \ , & \tau \to \infty \ , \end{array} \right. \tag{1.5}$$

where α_1, α_2 are positive numbers depending on the location of $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_0$, provided $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_0$ satisfy the technical condition (2.1) below.

The configurations of $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_0$ for which the formulation (1.4) is possible are described by a geometrical condition (see [6, section 2.3]). All our computations in this paper are for cases in which (1.4) is valid. In other cases one must compute the limit (1.3) as it stands, leading to a more complicated approximation problem directly employing (1.3) with sufficiently small s [6].

In [4] and [6] a numerical method was proposed for the computation of (1.4) and (1.3). The boundary integral method was used to compute g^r. (g^r satisfies the homogeneous PDE $(\Delta^* + \nu^2 - 1/4)g^r(\boldsymbol{\omega},\boldsymbol{\omega}_0,\nu) = 0$, on the manifold M, subject to an inhomogeneous boundary condition on its boundary ℓ.) This was implemented in [4] and [6] in the case when Ξ is a smooth cone (i.e., ℓ is a smooth contour) using, in effect, a simple trapezoidal-Nyström-type integral equation solver combined with the trapezoidal rule for computing (1.3) or (1.4). The approach of [33, 4, 6] was also extended to the electromagnetic case [34], which was implemented numerically in [5].

The papers [4] and [6] contained no convergence analysis of the method and, moreover, dealt only with the case of a smooth cone Ξ. The case of a cone with lateral edges is of fundamental importance in both the high-frequency theory of diffraction (where it is one of the unsolved canonical problems [28]) and in practice, where high-frequency scattering by antennas or corners of buildings is a key problem in microwave engineering. In such cases ℓ contains corners.

Although the integral equation method reduces the computation of $g(\boldsymbol{\omega},\boldsymbol{\omega}_0,\nu)$ to a computation on the (one-dimensional) contour ℓ on the surface of the unit sphere S^2, this equation has to be solved many times for different values of ν (and also more times if different $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_0$ are to be considered). Moreover, as we shall see, the evaluation of the kernel in the integral equation arising from the spherical PDE (1.2) is much more costly than for typical boundary integral equations in planar scattering theory. Thus there is strong practical demand for the development of an efficient algorithm, in particular one which solves the integral equation with the highest accuracy and the minimal number of kernel evaluations. Thus the purposes of this paper are as follows:

(i) To propose an efficient method for computing diffraction coefficients which is robust even when the cone Ξ has lateral edges and to analyze its convergence.

(ii) To minimize the number of kernel evaluations required in the implementation.

(iii) To demonstrate its use in the computation of diffraction coefficients in several sample cases.

The plan of the paper is as follows. In section 2 we describe briefly the boundary integral method for computing g^r. This leads to nonstandard integral equations posed on the spherical contour ℓ, which possibly contains corners. In section 3 we obtain the important properties of the integral operators which arise, including the case when the cone Ξ has lateral edges. In section 4 we describe a flexible numerical method based on collocation with piecewise polynomials and prove its convergence as a means of approximating $g^r(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu)$. Finally in section 5 we provide computations of diffraction coefficients for several sample problems. We also give in section 5 outline descriptions of various technical issues such as the computation of the contour integral in (1.3) and the evaluation of the kernel which appears in the integral operator. In particular, we note that because of the exponential decay (1.5), the domain of integration in (1.4) can be replaced by $[-N_1, N_2]$ with $N_i = O(r \log(n))$ at the cost of an error of $O(1/n^r)$. Therefore very large values of N_i (equivalently very large values of $|\tau|$) are not required in our computations.

Although this paper considers only diffraction coefficients for acoustic scattering, the related and more difficult electromagnetic case is described in [13] and the references therein.

2. Formulae for the conical diffraction coefficients. Throughout the paper we shall assume that the cone Ξ has a finite number of smooth (analytic) faces, joined at lateral edges, and that the angle between pairs of adjacent faces lies in $(0, 2\pi)$ (i.e., cuspoid edges are excluded). As in [13], we also assume that M and $S^2\backslash\overline{M}$ are simply connected subsets of S^2 and that the contour ℓ is a simple closed curve, consisting of a finite number of analytic arcs, also joined at noncuspoid corners. (For much of what we are going to do below, weaker smoothness assumptions away from edges would suffice, but we suppress this extra generality in the interest of readability.)

For $\boldsymbol{\omega}, \boldsymbol{\omega}' \in S^2$ we define $\theta(\boldsymbol{\omega}, \boldsymbol{\omega}')$ to be the geodesic distance between two points $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ on the sphere S^2 (i.e., $\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}') = \boldsymbol{\omega} \cdot \boldsymbol{\omega}'$, $0 \leq \theta(\boldsymbol{\omega}, \boldsymbol{\omega}') \leq \pi$). The configurations of $\boldsymbol{\omega}$ and $\boldsymbol{\omega}_0$ which ensure that (1.3) can be rewritten as (1.4) can now be described (for a convex fully illuminated cone) by the following condition (see also [33]):

$$\theta_1(\boldsymbol{\omega}, \boldsymbol{\omega}_0) := \min_{\boldsymbol{\omega}' \in \ell} \{\theta(\boldsymbol{\omega}, \boldsymbol{\omega}') + \theta(\boldsymbol{\omega}', \boldsymbol{\omega}_0)\} > \pi. \tag{2.1}$$

When $\theta_1(\boldsymbol{\omega}, \boldsymbol{\omega}_0) \leq \pi$ the formula (1.3) may either be undefined on the so-called singular directions or have to be interpreted in an appropriate distributional sense; for more details see [6, 25, 8]. We will not discuss this here, but the reader may refer to [6] and [13] for more detail, including the case when the cone is not fully illuminated.

As mentioned in section 1, the regular part g^r of the Green's function g in (1.2) is defined by

$$g^r(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) = g(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) - g_0(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu), \tag{2.2}$$

where g_0 is given by

$$g_0(\boldsymbol{\omega}, \boldsymbol{\omega}', \nu) = -\frac{1}{4\cos(\pi\nu)} P_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}')), \tag{2.3}$$

with P_k denoting the Legendre special function of the first kind of index k (see, e.g., [1, p. 332]). It is well known (see, e.g., [33, 4, 6]) that g_0 satisfies

$$(\Delta^* + \nu^2 - 1/4) g_0(\boldsymbol{\omega}, \boldsymbol{\omega}', \nu) = \delta(\boldsymbol{\omega} - \boldsymbol{\omega}')\ ,\ \ \boldsymbol{\omega}, \boldsymbol{\omega}' \in S^2 \tag{2.4}$$

(where the differentiation is with respect to $\boldsymbol{\omega}$); i.e., it is the fundamental solution for the operator $\Delta^* + \nu^2 - 1/4$ on all of the sphere S^2. Comparing (2.4) and (1.2), we see that for each $\boldsymbol{\omega}_0 \in M$ and $\nu \in \mathbb{C}$, the function g^r, as a function of $\boldsymbol{\omega}$, satisfies the homogeneous PDE (see [4, 6])

$$(\Delta^* + \nu^2 - 1/4)g^r(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) = 0, \quad \boldsymbol{\omega} \in M, \tag{2.5}$$

subject to the boundary condition on ℓ,

$$\left.\begin{array}{lll} \text{either} & g^r(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) = -g_0(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu), & \text{the Dirichlet case} \\ \text{or} & (\partial g^r/\partial \mathbf{m})(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) = -(\partial g_0/\partial \mathbf{m})(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu), & \text{the Neumann case} \end{array}\right\} \text{for all } \boldsymbol{\omega} \in \ell . \tag{2.6}$$

The boundary condition to be imposed on g^r is inherited from the boundary condition imposed on the original scattering problem. In (2.6) and throughout the paper, we make use of the following notational convention.

NOTATION 2.1. *With each $\boldsymbol{\omega} \in \ell$, not a corner point, we associate a unit normal $\mathbf{m} = \mathbf{m}(\boldsymbol{\omega})$ to ℓ at $\boldsymbol{\omega}$ which lies in the plane tangent to the unit sphere S^2 at $\boldsymbol{\omega}$ and is oriented outward from M. We also associate with $\boldsymbol{\omega}$ the unit tangent to ℓ at $\boldsymbol{\omega}$ denoted by $\mathbf{t} = \mathbf{t}(\boldsymbol{\omega})$, oriented so that $\mathbf{t}(\boldsymbol{\omega})$, $\mathbf{m}(\boldsymbol{\omega})$, $\boldsymbol{\omega}$ form an orthogonal right-handed triple (see Figure 3). (We usually suppress the dependence on $\boldsymbol{\omega}$ from the notation for simplicity.) Then $\partial/\partial\mathbf{m}$ denotes the (outward) normal derivative with respect to $\boldsymbol{\omega} \in \ell$. For any other point $\boldsymbol{\omega}'$ in ℓ, we analogously define the unit normal and tangent vectors $\mathbf{m}'$ and $\mathbf{t}'$ and normal derivative $\partial/\partial\mathbf{m}'$.*

The problem (2.5), (2.6) can now be solved by an integral equation method on ℓ. Here we follow the classical indirect approach, e.g., [3], adapted to the present problem in [4] and [6], although we note that a direct approach based on Green's formula would also be possible.

In the Dirichlet case, we seek the solution in the form of a double layer potential,

$$g^r(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) = \int_\ell \frac{\partial g_0}{\partial \mathbf{m}'}(\boldsymbol{\omega}, \boldsymbol{\omega}', \nu) u(\boldsymbol{\omega}', \nu) d\boldsymbol{\omega}', \qquad \boldsymbol{\omega} \in M \ . \tag{2.7}$$

Taking limits as $\boldsymbol{\omega}$ tends to the contour ℓ in (2.7) and using the jump conditions of the double layer potential and the Dirichlet boundary condition from (2.6), we obtain the second-kind integral equation:

$$\frac{1}{2}u(\boldsymbol{\omega}, \nu) + \int_\ell \frac{\partial g_0}{\partial \mathbf{m}'}(\boldsymbol{\omega}, \boldsymbol{\omega}', \nu) u(\boldsymbol{\omega}', \nu) d\boldsymbol{\omega}' = -g_0(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) \tag{2.8}$$

for all smooth points $\boldsymbol{\omega} \in l$. This equation is given in [4]. A rigorous justification for potential theory on manifolds with smooth boundaries is given in a very general context in [19]. For corner points the factor $1/2$ has to be replaced by a factor related to the corner angle; cf. [14]. However, since we will estimate errors for our boundary integral equations in L^2-type spaces, these points are unimportant. Notice that since $\boldsymbol{\omega} \in \ell$ and $\boldsymbol{\omega}_0 \in M$, the right-hand side (2.8) is never singular.

Analogously, the Neumann problem is solved with the single layer potential:

$$g^r(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) = \int_\ell g_0(\boldsymbol{\omega}, \boldsymbol{\omega}', \nu) u(\boldsymbol{\omega}', \nu) d\boldsymbol{\omega}', \qquad \boldsymbol{\omega} \in M \ . \tag{2.9}$$

Taking the normal derivative and fitting the boundary condition leads to

$$-\frac{1}{2}u(\boldsymbol{\omega}, \nu) + \int_\ell \frac{\partial g_0}{\partial \mathbf{m}}(\boldsymbol{\omega}, \boldsymbol{\omega}', \nu) u(\boldsymbol{\omega}', \nu) d\boldsymbol{\omega}' = -\frac{\partial g_0}{\partial \mathbf{m}}(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) \ . \tag{2.10}$$

We can write (2.8), (2.10) (almost everywhere) in operator form as

$$\Big(I + \mathcal{L}_B\Big)u = b_B \ , \quad \text{with} \ \ (\mathcal{L}_B u)(\boldsymbol{\omega}) = \int_\ell L_B(\boldsymbol{\omega}, \boldsymbol{\omega}')u(\boldsymbol{\omega}')d\boldsymbol{\omega}' \ , \quad B = D, N, \tag{2.11}$$

with solution $u(\boldsymbol{\omega}, \nu)$ abbreviated by $u(\boldsymbol{\omega})$. In the Dirichlet case the data are

$$b_D(\boldsymbol{\omega}) := -2g_0(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) \ , \quad L_D(\boldsymbol{\omega}, \boldsymbol{\omega}') := 2\frac{\partial g_0}{\partial \mathbf{m}'}(\boldsymbol{\omega}, \boldsymbol{\omega}', \nu) \ , \tag{2.12}$$

and in the Neumann case,

$$b_N(\boldsymbol{\omega}) := 2\frac{\partial g_0}{\partial \mathbf{m}}(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu), \quad L_N(\boldsymbol{\omega}, \boldsymbol{\omega}') := -2\frac{\partial g_0}{\partial \mathbf{m}}(\boldsymbol{\omega}, \boldsymbol{\omega}', \nu) \ . \tag{2.13}$$

Although the operators in (2.11), with the kernels from (2.12) or (2.13), are not classical, we will show that their properties are analogous to those of the standard layer potentials for the Helmholtz equation on the boundary of a planar domain.

3. Integral operators.

3.1. Preliminary results.

3.1. Preliminary results. The aim of this subsection is to identify the principal parts of the kernels L_D and L_N. This is done in Theorem 3.3. To prove this we need two technical lemmas.

LEMMA 3.1. *Using Notation* 2.1, *we have*

$$L_D(\boldsymbol{\omega}, \boldsymbol{\omega}') = \frac{1}{2\cos(\pi\nu)} \ P'_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}')) \ \mathbf{t}'.(\boldsymbol{\omega} \wedge \boldsymbol{\omega}') \ , \tag{3.1}$$

$$L_N(\boldsymbol{\omega}, \boldsymbol{\omega}') = -\frac{1}{2\cos(\pi\nu)} \ P'_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}')) \ \mathbf{t}.(\boldsymbol{\omega}' \wedge \boldsymbol{\omega}) \ . \tag{3.2}$$

Proof. By employing spherical polar coordinates $\boldsymbol{\omega}' = (\sin\theta'\cos\phi', \sin\theta'\sin\phi', \cos\theta')^T$, for any $v : S^2 \to \mathbb{R}$, we have the representation

$$\frac{\partial v}{\partial \mathbf{m}'}(\boldsymbol{\omega}') = \nabla_{\boldsymbol{\omega}'} \{v \circ \boldsymbol{\omega}'\} \cdot \mathbf{m}' \ ,$$

where $\nabla_{\boldsymbol{\omega}'}$ is the spherical gradient

$$\nabla_{\boldsymbol{\omega}'} = \frac{1}{\sin\theta'}\mathbf{e}_{\phi'}\frac{\partial}{\partial\phi'} + \mathbf{e}_{\theta'}\frac{\partial}{\partial\theta'} \ ,$$

with

$$\mathbf{e}_{\phi'} = (-\sin\phi', \cos\phi', 0)^T \quad \text{and} \quad \mathbf{e}_{\theta'} = (\cos\theta'\cos\phi', \cos\theta'\sin\phi', -\sin\theta')^T \ .$$

Since $\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}') = \boldsymbol{\omega} \cdot \boldsymbol{\omega}'$, we have

$$\frac{\partial}{\partial \mathbf{m}'}P_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}')) = -P'_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}')) \ \nabla_{\boldsymbol{\omega}'} \{\boldsymbol{\omega} \cdot \boldsymbol{\omega}'\} \cdot \mathbf{m}'. \tag{3.3}$$

Now an easy calculation shows that

$$\nabla_{\boldsymbol{\omega}'}\Big\{\boldsymbol{\omega} \cdot \boldsymbol{\omega}'\Big\} \cdot \mathbf{m}' = \Big\{(\boldsymbol{\omega} \cdot \mathbf{e}_{\phi'}) \ \mathbf{e}_{\phi'} + (\boldsymbol{\omega} \cdot \mathbf{e}_{\theta'}) \ \mathbf{e}_{\theta'}\Big\} \cdot \mathbf{m}' \ = \boldsymbol{\omega} \cdot \mathbf{m}'.$$

Thus from (3.3), (2.3), and (2.12), we have

$$L_D(\boldsymbol{\omega},\boldsymbol{\omega}') = \frac{1}{2\cos(\pi\nu)}\ P'_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\omega},\boldsymbol{\omega}'))\ \boldsymbol{\omega}\cdot\mathbf{m}'\ . \tag{3.4}$$

Since $\mathbf{t}'$, $\mathbf{m}'$, and $\boldsymbol{\omega}'$ form a right-handed triple, we have $\mathbf{m}' = \boldsymbol{\omega}' \wedge \mathbf{t}'$, and so

$$L_D(\boldsymbol{\omega},\boldsymbol{\omega}') = \frac{1}{2\cos(\pi\nu)}\ P'_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\omega},\boldsymbol{\omega}'))\ \boldsymbol{\omega}\cdot(\boldsymbol{\omega}'\wedge\mathbf{t}')\ ,$$

which is equivalent to (3.1) by cyclic permutation. Since $L_N(\boldsymbol{\omega},\boldsymbol{\omega}') = -L_D(\boldsymbol{\omega}',\boldsymbol{\omega})$, (3.2) follows easily. □

The next lemma identifies the asymptotic behavior of $P'_{\nu+\frac{1}{2}}(x)$ for x close to -1. In Theorem 3.3, we will combine this with (3.1), (3.2) to identify the behavior of L_D and L_N near $\boldsymbol{\omega} = \boldsymbol{\omega}'$.

Lemma 3.2. *For all $k \in \mathbb{C}$, $P_k(x)$ is an analytic function of $x \in (-1,3)$. Moreover, for $x \in (-3,1)$,*

$$P_k(x) = a_k(x)\log\left(\frac{1+x}{2}\right) + b_k(x),$$

where $a_k(x)$ and $b_k(x)$ are both analytic on $(-3,1)$, with

$$a_k(-1) = \frac{\sin(\pi k)}{\pi} \qquad \text{and} \qquad b_k(-1) = \frac{\sin(\pi k)}{\pi}\ \{\psi(k)+\psi(-k-1)+2\gamma\},$$

where $\psi(k) = -\gamma - \sum_{r=1}^{\infty}(1/(k+r) - 1/r)$ and γ is the Euler constant [1, p. 255].

Proof. From [1, equation (8.1.2)] we get the following representation of P_k:

$$P_k(x) = F\left(-k, k+1; 1; \frac{1-x}{2}\right), \tag{3.5}$$

where F is the hypergeometric function. It follows from [1, p. 556] that $F(-k,k+1;1;z)$ is a convergent power series for $-1 \le z < 1$. Therefore, by (3.5), $P_k(x)$ is analytic for $x \in (-1,3)$ and in particular for $x \in (-1,1)$. This proves the first statement in the theorem.

Furthermore, from [24, Chapter V, equation (53)] we have that

$$P_k(x) = a_k(x)\log\left(\frac{1+x}{2}\right) + b_k(x),$$

where

$$a_k(x) = \frac{\sin(\pi k)}{\pi}F(-k,k+1;1;(1+x)/2) \tag{3.6}$$

and

$$\begin{aligned} b_k(x) = \frac{\sin(\pi k)}{\pi}\Bigg\{&[\psi(k)+\psi(-k-1)+2\gamma]F(-k,k+1;1;(1+x)/2) \\ &+\sum_{r=1}^{\infty}B(k,r)\phi(k,r)\left(\frac{1+x}{2}\right)^r\Bigg\}\ . \end{aligned} \tag{3.7}$$

Here

$$B(k,r) = \frac{(-k)\dots(-k+r-1)(k+1)\dots(k+r)}{(r!)^2}$$

and

$$\phi(k,r) = \sum_{j=1}^{r} \left\{ \frac{2k(k+1)+j}{(j^2-k^2-k-j)j} \right\}.$$

As remarked above, $F(-k,k+1;1;(1+x)/2)$ is a convergent power series for $-1 \le (1+x)/2 < 1$, so $a_k(x)$ is analytic for $x \in (-3,1)$. Moreover $a_k(-1) = \sin(\pi k)/\pi$ follows from [1, p. 556]).

Turning to b_k, it is clear that the first term on the right-hand side of (3.7) is also analytic for $x \in (-3,1)$ and that the assertions about b_k will then follow, provided the domain of convergence of the power series

$$\sum_{r=1}^{\infty} B(k,r)\phi(k,r)\left(\frac{1+x}{2}\right)^r \tag{3.8}$$

can be shown to be $(-3,1)$. To obtain this result, note that $\lim_{r\to\infty}\phi(k,r)$ is clearly finite. If $\lim_{r\to\infty}\phi(k,r) \neq 0$, then it follows that $|\phi(k,r+1)/\phi(k,r)| \to 1$ as $r \to \infty$. Then

$$\begin{aligned}&\lim_{r\to\infty} \frac{|B(k,r+1)\ \phi(k,r+1)((1+x)/2)^{r+1}|}{|B(k,r)\ \phi(k,r)((1+x)/2)^r|}\\ &= \left|\frac{1+x}{2}\right| \lim_{r\to\infty}\left|\frac{(-k+r)(k+r+1)}{(r+1)^2}\,\frac{\phi(k,r+1)}{\phi(k,r)}\right| = \left|\frac{1+x}{2}\right|,\end{aligned} \tag{3.9}$$

and (3.8) is convergent for $x \in (-3,1)$ by the ratio test. However, if $\lim_{r\to\infty}\phi(k,r) = 0$, then, for large enough r, $|\phi(k,r)| < 1$. Since (3.9) also shows that the power series $\sum_{r=1}^{\infty} B(k,r)((1+x)/2)^r$ converges for $x \in (-3,1)$, the comparison test then ensures that (3.8) also converges for $x \in (-3,1)$. □

We now combine Lemmas 3.1 and 3.2 to obtain the following theorem.

THEOREM 3.3. *Recall Notation* 2.1.

(i) *For* $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \ell$,

$$L_D(\boldsymbol{\omega},\boldsymbol{\omega}') = -\frac{\mathbf{t}'\cdot(\boldsymbol{\omega}\wedge\boldsymbol{\omega}')}{\pi|\boldsymbol{\omega}-\boldsymbol{\omega}'|^2} + F_D(\boldsymbol{\omega},\boldsymbol{\omega}'), \tag{3.10}$$

$$L_N(\boldsymbol{\omega},\boldsymbol{\omega}') = \frac{\mathbf{t}\cdot(\boldsymbol{\omega}'\wedge\boldsymbol{\omega})}{\pi|\boldsymbol{\omega}-\boldsymbol{\omega}'|^2} + F_N(\boldsymbol{\omega},\boldsymbol{\omega}'), \tag{3.11}$$

where F_D *and* F_N *are bounded functions on* $\ell \times \ell$.

(ii) *When* $\boldsymbol{\omega}$ *is* not *a corner point of* ℓ,

$$\frac{\mathbf{t}'\cdot(\boldsymbol{\omega}\wedge\boldsymbol{\omega}')}{\pi|\boldsymbol{\omega}-\boldsymbol{\omega}'|^2} \quad \text{and} \quad \frac{\mathbf{t}\cdot(\boldsymbol{\omega}'\wedge\boldsymbol{\omega})}{\pi|\boldsymbol{\omega}-\boldsymbol{\omega}'|^2} \tag{3.12}$$

are both $\mathcal{C}^\infty$ *functions of* $\boldsymbol{\omega}'$ *in a neighborhood of* $\boldsymbol{\omega}$ *and, for* $B = D$ *or* N,

$$F_B(\boldsymbol{\omega},\boldsymbol{\omega}') = O(|\boldsymbol{\omega}-\boldsymbol{\omega}'|^2 \log|\boldsymbol{\omega}-\boldsymbol{\omega}'|) \quad \text{as} \quad \boldsymbol{\omega}' \to \boldsymbol{\omega}\ . \tag{3.13}$$

Proof. We give the proof for L_D; the argument for L_N is analogous.

(i) From Lemma 3.2 with $k = \nu - 1/2$, we have, for $x \in (-1, 1)$,

$$P'_{\nu-\frac{1}{2}}(x) = \frac{-\cos(\pi\nu)}{\pi}\left\{\frac{1}{x+1}\right\} + r(x), \tag{3.14}$$

where

$$r(x) = \left[\frac{a_{\nu-\frac{1}{2}}(x) - a_{\nu-\frac{1}{2}}(-1)}{x-(-1)} + a'_{\nu-\frac{1}{2}}(x)\log\left(\frac{x+1}{2}\right) + b'_{\nu-\frac{1}{2}}(x)\right] . \tag{3.15}$$

Also note that since $\boldsymbol{\omega}, \boldsymbol{\omega}' \in S^2$, we have

$$-\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}') + 1 = -\boldsymbol{\omega}.\boldsymbol{\omega}' + 1 = \frac{1}{2}|\boldsymbol{\omega} - \boldsymbol{\omega}'|^2 . \tag{3.16}$$

Hence

$$P'_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}')) = -\frac{2\cos(\pi\nu)}{\pi|\boldsymbol{\omega} - \boldsymbol{\omega}'|^2} + r(-1 + |\boldsymbol{\omega} - \boldsymbol{\omega}'|^2/2). \tag{3.17}$$

Therefore combining (3.1) with (3.15) and (3.17), we obtain the formula (3.10), where

$$F_D(\boldsymbol{\omega}, \boldsymbol{\omega}') = \frac{1}{2\cos(\pi\nu)}\ r(-1 + |\boldsymbol{\omega} - \boldsymbol{\omega}'|^2/2)\ \mathbf{t}' \cdot (\boldsymbol{\omega} \wedge \boldsymbol{\omega}') . \tag{3.18}$$

To complete the proof of (i) we now show that F_D is bounded on $\ell \times \ell$. To do this, choose a fixed δ satisfying $0 < \delta < \pi/2$ and first consider $\boldsymbol{\omega}, \boldsymbol{\omega}'$ in the range

$$0 \le \theta(\boldsymbol{\omega}, \boldsymbol{\omega}') \le \pi - \delta . \tag{3.19}$$

Then there exists $\epsilon > 0$ such that $-1 \le -\cos\theta(\boldsymbol{\omega}, \boldsymbol{\omega}') \le 1 - \epsilon$, and hence it follows from (3.16) that

$$-1 \le -1 + |\boldsymbol{\omega} - \boldsymbol{\omega}|^2/2 \le 1 - \epsilon. \tag{3.20}$$

Substituting (3.15) into (3.18) we obtain

$$\begin{aligned} &2\cos(\pi\nu)F_D(\boldsymbol{\omega}, \boldsymbol{\omega}') \\ &= \mathbf{t}' \cdot (\boldsymbol{\omega} \wedge \boldsymbol{\omega}')\left\{\frac{a_{\nu-\frac{1}{2}}(-1+|\boldsymbol{\omega} - \boldsymbol{\omega}'|^2/2) - a_{\nu-\frac{1}{2}}(-1)}{|\boldsymbol{\omega} - \boldsymbol{\omega}'|^2/2} + b'_{\nu-\frac{1}{2}}(-1+|\boldsymbol{\omega} - \boldsymbol{\omega}'|^2/2)\right\} &(3.21)\\ &+ a'_{\nu-\frac{1}{2}}(-1 + |\boldsymbol{\omega} - \boldsymbol{\omega}'|^2/2)\ \left\{\mathbf{t}' \cdot (\boldsymbol{\omega} \wedge \boldsymbol{\omega}')\log(|\boldsymbol{\omega} - \boldsymbol{\omega}'|^2/4)\right\} . &(3.22)\end{aligned}$$

Recall from Lemma 3.2 that $a_{\nu-\frac{1}{2}}$ and $b'_{\nu-\frac{1}{2}}$ are both analytic on $(-3, 1)$. Since $|\boldsymbol{\omega} - \boldsymbol{\omega}'|^2$ is a smooth function of $\boldsymbol{\omega}, \boldsymbol{\omega}'$, it follows that the terms inside the braces in (3.21) are smooth functions of $\boldsymbol{\omega}, \boldsymbol{\omega}' \in \ell$. Moreover

$$\begin{aligned} |\mathbf{t}' \cdot (\boldsymbol{\omega} \wedge \boldsymbol{\omega}')| \le |\mathbf{t}'||\boldsymbol{\omega} \wedge \boldsymbol{\omega}'| &= \sin\theta(\boldsymbol{\omega}, \boldsymbol{\omega}') = \{1 - \cos^2\theta(\boldsymbol{\omega}, \boldsymbol{\omega}')\}^{1/2} \\ &= \{1 - (\boldsymbol{\omega} \cdot \boldsymbol{\omega}')^2\}^{1/2} = \{(1 - \boldsymbol{\omega} \cdot \boldsymbol{\omega}')(1 + \boldsymbol{\omega} \cdot \boldsymbol{\omega}')\}^{1/2} \\ &= \frac{1}{2}|\boldsymbol{\omega} - \boldsymbol{\omega}'||\boldsymbol{\omega} + \boldsymbol{\omega}'|, \end{aligned} \tag{3.23}$$

which ensures the boundedness of (3.21). The boundedness of (3.22) follows in a similar way, using (3.23) and the analyticity of $a'_{\nu-\frac{1}{2}}$ on $(-3,1)$.

To complete the proof, consider the case when (3.19) fails, i.e., $\pi-\delta \le \theta(\boldsymbol{\omega},\boldsymbol{\omega}') \le \pi$. For this case recall from (3.5) that $P_{\nu-1/2}(x)$ is analytic for $x \in (-1,3)$. Therefore (3.1) implies that $L_D(\boldsymbol{\omega},\boldsymbol{\omega}')$ is bounded for $\pi-\delta \le \theta(\boldsymbol{\omega},\boldsymbol{\omega}') \le \pi$. Thus setting

$$F_D(\boldsymbol{\omega},\boldsymbol{\omega}) = L_D(\boldsymbol{\omega},\boldsymbol{\omega}') + \frac{\mathbf{t}'\cdot(\boldsymbol{\omega}\wedge\boldsymbol{\omega}')}{\pi|\boldsymbol{\omega}-\boldsymbol{\omega}'|^2}$$

ensures that (3.10) holds, with $F_D(\boldsymbol{\omega},\boldsymbol{\omega})$ bounded for $\pi-\delta \le \theta(\boldsymbol{\omega},\boldsymbol{\omega}') \le \pi$.

(ii) Now suppose that $\boldsymbol{\omega}$ is not a corner point and that $\boldsymbol{\omega}'$ is sufficiently close to $\boldsymbol{\omega}$ so as to ensure that there is no corner point between $\boldsymbol{\omega}$ and $\boldsymbol{\omega}'$ on ℓ. Let $\boldsymbol{\rho}$ denote an arclength parameterization of ℓ from any fixed reference point; then setting $\boldsymbol{\omega}=\boldsymbol{\rho}(s)$, the unit tangent $\mathbf{t}$ at $\boldsymbol{\omega}$ is given by $\mathbf{t}=\boldsymbol{\rho}_s(s)$, the derivative of $\boldsymbol{\rho}(s)$. Then for $\boldsymbol{\omega}'$ near $\boldsymbol{\omega}$ with $\boldsymbol{\omega}'=\boldsymbol{\rho}(\sigma)$, we have

$$(3.24)\quad |\boldsymbol{\omega}-\boldsymbol{\omega}'|/|s-\sigma| = O(1) \quad\text{and}\quad |s-\sigma|/|\boldsymbol{\omega}-\boldsymbol{\omega}'| = O(1) \qquad \text{as } \sigma\to s.$$

Also,

$$\boldsymbol{\omega}\wedge\boldsymbol{\omega}' = \boldsymbol{\rho}(s)\wedge\boldsymbol{\rho}(\sigma) = (\boldsymbol{\rho}(s)-\boldsymbol{\rho}(\sigma))\wedge\boldsymbol{\rho}(\sigma).$$

Hence

$$(3.25)\qquad \mathbf{t}'\cdot(\boldsymbol{\omega}\wedge\boldsymbol{\omega}') = \boldsymbol{\rho}_s(\sigma)\cdot((\boldsymbol{\rho}(s)-\boldsymbol{\rho}(\sigma)-(s-\sigma)\boldsymbol{\rho}_s(\sigma))\wedge\boldsymbol{\rho}(\sigma)).$$

Since $|\boldsymbol{\omega}-\boldsymbol{\omega}'|^2 = (\boldsymbol{\rho}(s)-\boldsymbol{\rho}(\sigma))\cdot(\boldsymbol{\rho}(s)-\boldsymbol{\rho}(\sigma))$, it follows that (3.12) are smooth functions as $\sigma\to s$ (i.e., $\boldsymbol{\omega}'\to\boldsymbol{\omega}$). Moreover (3.24) and (3.25) imply that $|\mathbf{t}'\cdot(\boldsymbol{\omega}\wedge\boldsymbol{\omega}')| = O(|\boldsymbol{\omega}-\boldsymbol{\omega}'|^2)$ and so (3.13) follows from (3.21) and (3.22). □

We see from Theorem 3.3 that if there are no corner points on ℓ, then L_D and L_N are bounded (in fact continuous), so in both the Dirichlet and Neumann cases the integral operator $\mathcal{L}_B$ will be compact on most standard spaces, e.g., $C(\ell), L^2(\ell)$. Then standard stability proofs for the numerical method will follow. However, if ℓ does contain a corner, compactness is lost and so another approach is needed to show stability of a numerical method. The approach we will use is to compare the integral operator $\mathcal{L}_B$ with a corresponding plane Laplace integral operator $\mathcal{K}_B$ and then use stability results which are known for the planar Laplace problem. This is done in the following subsection.

3.2. Relation to planar Laplace case. To simplify the presentation, we assume that the contour ℓ has one corner which we will denote by the point $\boldsymbol{\omega}_c \in S^2$. The case of several corners is obtained analogously. Without loss of generality, we assume $\boldsymbol{\omega}_c = (0,0,1)^T$. Suppose that $\boldsymbol{\rho}(s)$ travels around ℓ with M on the right-hand side (as indicated by the arrow in Figure 3), as s travels from $-\Lambda$ to Λ, where 2Λ is the length of ℓ. Then we can introduce the *wedge* w in the tangent plane to S^2 at $\boldsymbol{\omega}_c$ as follows.

DEFINITION 3.4. *The wedge w is defined to be the union of two straight line segments: $w = w^- \cup w^+$, where*

$$w^- = \{(0,0,1)^T + s\mathbf{t}_c^- : s\in[-\Lambda,0]\}, \qquad w^+ = \{(0,0,1)^T + s\mathbf{t}_c^+ : s\in[0,\Lambda]\},$$

and $\mathbf{t}_c^\pm = \lim_{s\to 0\pm}\boldsymbol{\rho}_s(s)$ (see Figure 3). The angle between the tangents $\mathbf{t}_c^+$ and $-\mathbf{t}_c^-$ is measured "counterclockwise" about the z axis (when viewed from outside the sphere)

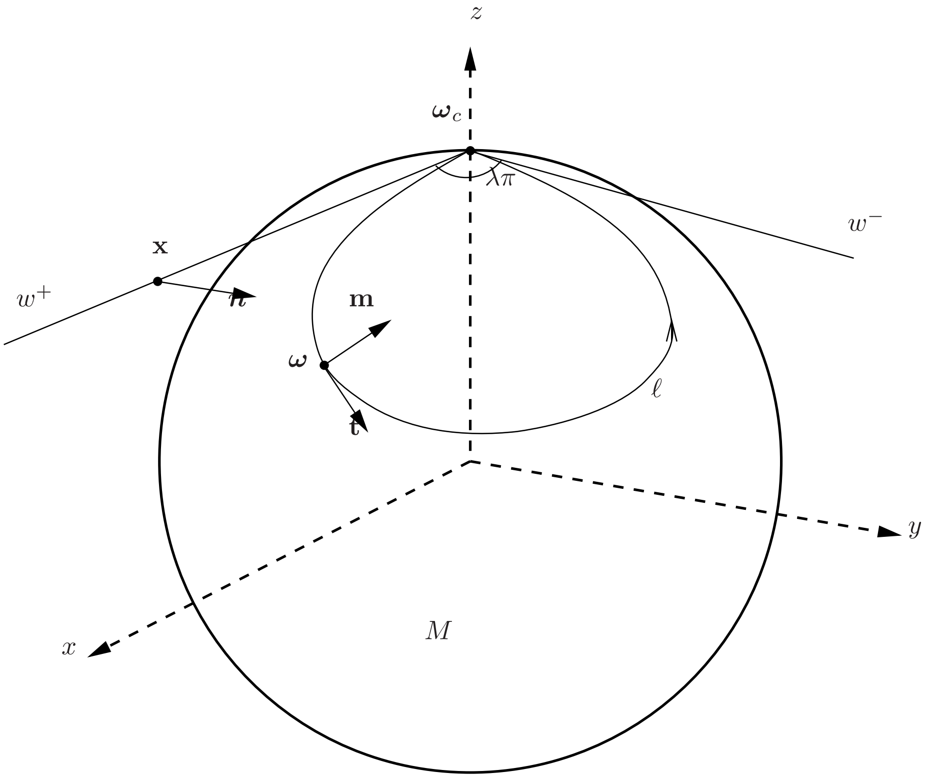


FIG. 3. *Wedge w and contour ℓ.*

from w^+ to w^- and is denoted $\lambda\pi$, where $\lambda \in (0,2)\backslash\{1\}$. Without loss of generality we choose our coordinate system so that $\mathbf{t}_c^+$ is in the direction of the x axis. Each $\mathbf{x} = s\mathbf{t}_c^\pm \in w^\pm$ can be associated with a unique $\boldsymbol{\omega} = \boldsymbol{\rho}(s) \in \ell$, and with a unit normal $\mathbf{m}$ at $\boldsymbol{\omega} \in \ell$ orientated outward from M. To $\mathbf{x}$ we associate a unit normal $\boldsymbol{n}$ to w in the plane tangent to S^2 at $\boldsymbol{\omega}_c$, orientated so that $\boldsymbol{n}\cdot\mathbf{m} \to 1$ as $s \to 0$. (See Figure 3.)

The fundamental solution of Laplace's equation on the plane is $(1/2\pi)\log|\mathbf{x}-\mathbf{x}'|$. Using this we introduce the operators

$$(\mathcal{K}_B u)(\mathbf{x}) = \int_w K_B(\mathbf{x},\mathbf{x}')u(\mathbf{x}')d\mathbf{x}' \ , \qquad B = D, N \ .$$

Analogously to (2.12), (2.13), the Dirichlet and Neumann kernels are

$$K_D(\mathbf{x},\mathbf{x}') := \frac{1}{\pi}\frac{\partial}{\partial \boldsymbol{n}'}\left\{\log|\mathbf{x}-\mathbf{x}'|\right\} = -\frac{(\mathbf{x}-\mathbf{x}')\cdot\boldsymbol{n}'}{\pi|\mathbf{x}-\mathbf{x}'|^2}, \tag{3.26}$$

$$K_N(\mathbf{x},\mathbf{x}') := -\frac{1}{\pi}\frac{\partial}{\partial \boldsymbol{n}}\left\{\log|\mathbf{x}-\mathbf{x}'|\right\} = -\frac{(\mathbf{x}-\mathbf{x}')\cdot\boldsymbol{n}}{\pi|\mathbf{x}-\mathbf{x}'|^2}. \tag{3.27}$$

Here $\boldsymbol{n}, \boldsymbol{n}'$ are unit normals to $\boldsymbol{\omega}$ at $\mathbf{x}, \mathbf{x}' \in w$, as described in Definition 3.4.

Theorem 3.5 will show that the principal singularity of L_B near $\boldsymbol{\omega} = \boldsymbol{\omega}' = \boldsymbol{\omega}_c$ is the same as K_B near $\mathbf{x} = \mathbf{x}' = \boldsymbol{\omega}_c$. This is useful because the properties of the integral operator $\mathcal{K}_B$ with kernel K_B are well understood [17, 14, 20, 22, 31].

To prepare for Theorem 3.5, we use the arclength parameterization, $\boldsymbol{\rho}(\sigma)$, of ℓ, introduced above, to rewrite (2.11) on $[-\Lambda,\Lambda]$. Putting $\boldsymbol{\omega} = \boldsymbol{\rho}(s)$ and $\boldsymbol{\omega}' = \boldsymbol{\rho}(\sigma)$ we obtain

$$(3.28)\quad (I+\widehat{\mathcal{L}}_B)\widehat{u}=\widehat{b}_B\ ,\quad \text{with}\ \ (\widehat{\mathcal{L}}_B\widehat{u})(s)=\int_{-\Lambda}^{\Lambda}\widehat{L}_B(s,\sigma)\widehat{u}(\sigma)d\sigma,\quad s\in[-\Lambda,\Lambda],$$

where $\widehat{u}(s)=u(\boldsymbol{\rho}(s))$. In the case of Dirichlet boundary data, using (2.12) and Lemma 3.1 we have

$$\widehat{b}_D(s):=-2g_0(\boldsymbol{\rho}(s),\boldsymbol{\omega}_0,\nu)\quad\text{and}$$
$$(3.29)\quad \widehat{L}_D(s,\sigma):=\frac{1}{2\cos(\pi\nu)}\ P'_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\rho}(s),\boldsymbol{\rho}(\sigma)))\ \boldsymbol{\rho}_s(\sigma)\cdot(\boldsymbol{\rho}(s)\wedge\boldsymbol{\rho}(\sigma)).$$

(Note that since $\boldsymbol{\rho}$ is the arclength parameterization, the Jacobian satisfies $|\boldsymbol{\rho}_s(\sigma)|=1$ and therefore does not appear explicitly in the kernel.) For Neumann boundary data, using (2.13) and Lemma 3.1 we obtain

$$\widehat{b}_N(s):=2\frac{\partial g_0}{\partial\mathbf{m}(s)}(\boldsymbol{\rho}(s),\boldsymbol{\omega}_0,\nu)\quad\text{and}$$
$$(3.30)\quad \widehat{L}_N(s,\sigma):=-\frac{1}{2\cos(\pi\nu)}\ P'_{\nu-\frac{1}{2}}(-\cos\theta(\boldsymbol{\rho}(s),\boldsymbol{\rho}(\sigma)))\ \boldsymbol{\rho}_s(s)\cdot(\boldsymbol{\rho}(\sigma)\wedge\boldsymbol{\rho}(s)),$$

where $\mathbf{m}(s)$ is the corresponding normal to ℓ at $\boldsymbol{\rho}(s)$.

If we now denote the arclength parameterization of w by $\boldsymbol{r}$, with $\boldsymbol{r}(-\Lambda)=(0,0,1)^T-\Lambda\mathbf{t}_c^-$, $\boldsymbol{r}(0)=\boldsymbol{\omega}_c$, and $\boldsymbol{r}(\Lambda)=(0,0,1)^T+\Lambda\mathbf{t}_c^+$, then we can also rewrite $\mathcal{K}_B$ as an operator:

$$(\widehat{\mathcal{K}}_B\widehat{u})(s)=\int_{-\Lambda}^{\Lambda}\widehat{K}_B(s,\sigma)\widehat{u}(\sigma)d\sigma,\qquad s\in[-\Lambda,\Lambda]\ ,\quad B=D,N\ ,$$

where, from (3.26), (3.27),

$$(3.31)\qquad \widehat{K}_D(s,\sigma):=-\frac{(\boldsymbol{r}(s)-\boldsymbol{r}(\sigma))\cdot\boldsymbol{n}(\sigma)}{\pi|\boldsymbol{r}(s)-\boldsymbol{r}(\sigma)|^2},$$
$$(3.32)\qquad \widehat{K}_N(s,\sigma):=-\frac{(\boldsymbol{r}(s)-\boldsymbol{r}(\sigma))\cdot\boldsymbol{n}(s)}{\pi|\boldsymbol{r}(s)-\boldsymbol{r}(\sigma)|^2}$$

for the Dirichlet and Neumann problems, respectively. Here $\boldsymbol{n}(\sigma)$ is the normal to w at $\mathbf{x}=\boldsymbol{r}(\sigma)$. The following theorem shows that $\widehat{K}_B$ contains the principal singularity of $\widehat{L}_B$ near the corner point $s=\sigma=0$ in both the Dirichlet and Neumann cases, $B=D,N$.

THEOREM 3.5. *Let $B=D$ or N. Then for $(s,\sigma)\in[-\Lambda,\Lambda]\times[-\Lambda,\Lambda]$, $\widehat{L}_B(s,\sigma)-\widehat{K}_B(s,\sigma)$ is a bounded function.*

Proof. We give the proof for the case $B=D$. The case $B=N$ is analogous. First we consider the kernel $\widehat{K}_D$. From Definition 3.4 the parametric equation, $\boldsymbol{r}$, for w is given by

$$(3.33)\qquad \boldsymbol{r}(\sigma)=\begin{cases}(-\sigma\cos(\lambda\pi),-\sigma\sin(\lambda\pi),1)^T, & \sigma\in[-\Lambda,0],\\ (\sigma,0,1)^T, & \sigma\in[0,\Lambda].\end{cases}$$

Notice that if $-\Lambda\le s,\sigma\le 0$ or $0\le s,\sigma\le\Lambda$, then $\boldsymbol{r}(s)$ and $\boldsymbol{r}(\sigma)$ lie on the same arm of w, and so it follows from (3.31) that $\widehat{K}_D(s,\sigma)=0$ and, by Theorem 3.3(ii),

$\widehat{L}_D(s,\sigma)$ is bounded. So we have to consider only the case when s and σ are on different sides of 0.

First consider the case $-\Lambda \le s \le 0 \le \sigma \le \Lambda$. Then (3.33) implies that $\boldsymbol{r}(s) - \boldsymbol{r}(\sigma) = (-s\cos(\lambda\pi) - \sigma, -s\sin(\lambda\pi), 0)^T$ and $\boldsymbol{n}(\sigma) = (0,1,0)^T$. Therefore $(\boldsymbol{r}(s) - \boldsymbol{r}(\sigma))\cdot \boldsymbol{n}(\sigma) = -s\sin(\lambda\pi)$ and $|\boldsymbol{r}(s) - \boldsymbol{r}(\sigma)|^2 = s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2$. So from (3.31),

$$\widehat{K}_D(s,\sigma) = \frac{1}{\pi}\frac{s\sin(\lambda\pi)}{(s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2)}, \quad -\Lambda \le s \le 0 \le \sigma \le \Lambda. \tag{3.34}$$

A similar calculation shows analogously that

$$\widehat{K}_D(s,\sigma) = -\frac{1}{\pi}\frac{s\sin(\lambda\pi)}{(s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2)}, \quad -\Lambda \le \sigma \le 0 \le s \le \Lambda. \tag{3.35}$$

Now we turn our attention to the kernel, $\widehat{L}_D(s,\sigma)$. Using Taylor's theorem we can write the parameterization $\boldsymbol{\rho}$ as

$$\boldsymbol{\rho}(\sigma) = \begin{cases} \boldsymbol{r}(\sigma) + \sigma^2(\alpha_1(-\sigma), \beta_1(-\sigma), \gamma_1(-\sigma))^T, & \sigma \in [-\Lambda, 0], \\ \boldsymbol{r}(\sigma) + \sigma^2(\alpha_2(\sigma), \beta_2(\sigma), \gamma_2(\sigma))^T, & \sigma \in [0, \Lambda], \end{cases} \tag{3.36}$$

where $\alpha_i(s), \beta_i(s)$, and $\gamma_i(s)$ are smooth functions on $[0,\Lambda]$ for $i = 1,2$. Thus, for $-\Lambda \le s \le 0 \le \sigma \le \Lambda$, we have, from (3.36),

$$\boldsymbol{\rho}(s) \wedge \boldsymbol{\rho}(\sigma) = (-s\sin(\lambda\pi), s\cos(\lambda\pi) + \sigma, 0)^T + O(\max\{|s|,|\sigma|\}^2)$$

as $\max\{|s|,|\sigma|\} \to 0$. Hence with $\boldsymbol{\omega} = \boldsymbol{\rho}(s)$ and $\boldsymbol{\omega}' = \boldsymbol{\rho}(\sigma)$, we have

$$\mathbf{t}' = \boldsymbol{\rho}_s(\sigma) = (1,0,0)^T + O(|\sigma|) \,, \tag{3.37}$$

$$-\mathbf{t}'\cdot(\boldsymbol{\omega}\wedge\boldsymbol{\omega}') = s\sin(\lambda\pi) + O(\max\{|s|,|\sigma|\}^2) \,,$$

$$\text{and } |\boldsymbol{\omega} - \boldsymbol{\omega}'|^2 = s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2 + O(\max\{|s|,|\sigma|\}^3) \tag{3.38}$$

as $\max\{|s|,|\sigma|\} \to 0$. Therefore we have, from (3.10),

$$\widehat{L}_D(s,\sigma) = \frac{\sin(\lambda\pi)}{\pi}\frac{s + \eta_2(s,\sigma)}{s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2 + \eta_3(s,\sigma)} + \widehat{F}_D(s,\sigma),$$

where $\widehat{F}_D(s,\sigma) = F_D(\boldsymbol{\rho}(s), \boldsymbol{\rho}(\sigma))$ and

$$\eta_i(s,\sigma) = O(\max\{|s|,|\sigma|\}^i), \quad i = 2,3. \tag{3.39}$$

Hence, for $-\Lambda \le s \le 0 \le \sigma \le \Lambda$,

$$(\widehat{L}_D - \widehat{K}_D)(s,\sigma) = \frac{\sin(\lambda\pi)}{\pi}\left\{\frac{s + \eta_2(s,\sigma)}{s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2 + \eta_3(s,\sigma)} - \frac{s}{s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2}\right\} + \widehat{F}_D(s,\sigma), \tag{3.40}$$

which is clearly continuous for $(s,\sigma) \ne (0,0)$.

In order to show that $(\widehat{L}_D - \widehat{K}_D)(s,\sigma)$ is bounded near $(s,\sigma) = (0,0)$ we need to show that the limit (as $(s,\sigma) \to (0,0)$) of the first term on the right-hand side of

(3.40) is bounded. We do this for $0 < -s \leq \sigma$. The case $0 < \sigma \leq -s$ is analogous. To obtain the result, write

$$\begin{aligned}
&\frac{s + \eta_2(s,\sigma)}{s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2 + \eta_3(s,\sigma)} - \frac{s}{s^2 + 2s\sigma\sin(\lambda\pi) + \sigma^2} \\
&\qquad = \frac{\eta_2(s,\sigma)(s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2) - \eta_3(s,\sigma)s}{(s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2)(s^2 + 2s\sigma\cos(\lambda\pi) + \sigma^2 + \eta_3(s,\sigma))} \\
&\qquad = \frac{\frac{\eta_2(s,\sigma)}{\sigma^2}((\frac{s}{\sigma})^2 + 2\frac{s}{\sigma}\cos(\lambda\pi) + 1) - \frac{\eta_3(s,\sigma)}{\sigma^3}\frac{s}{\sigma}}{((\frac{s}{\sigma})^2 + 2\frac{s}{\sigma}\cos(\lambda\pi) + 1)((\frac{s}{\sigma})^2 + 2\frac{s}{\sigma}\cos(\lambda\pi) + 1 + \frac{\eta_3(s,\sigma)}{\sigma^2})}. \qquad (3.41)
\end{aligned}$$

Now, when $0 < -s \leq \sigma$ we have $0 < |s| \leq |\sigma|$ and from (3.39) it follows that $\eta_2(s,\sigma)/\sigma^2 = O(1)$, $\eta_3(s,\sigma)/\sigma^3 = O(1)$, and $\eta_3(s,\sigma)/\sigma^2 \to 0$ as $(s,\sigma) \to (0,0)$. Moreover, since $\lambda \in (0,2)\backslash\{1\}$, we have

$$x^2 + 2x\cos(\lambda\pi) + 1 \ \geq \ \sin^2(\lambda\pi) \ > \ 0 \quad \text{for all } \ x \in \mathbb{R} \ .$$

Combining all these facts with (3.41) shows that the first term in (3.40) is bounded as $(s,\sigma) \to (0,0)$. Since $\widehat{F}_D$ is a bounded function, it follows that $\widehat{L}_D(s,\sigma) - \widehat{K}_D(s,\sigma)$ is bounded for $-\Lambda \leq s \leq 0 \leq \sigma \leq \Lambda$.

For $-\Lambda \leq \sigma \leq 0 \leq s \leq \Lambda$ the result follows analogously. □

We shall analyze (3.28) in the space $L^2[-\Lambda,\Lambda]$, equipped with the norm $\|v\|_{L^2[-\Lambda,\Lambda]} = \{\int_{-\Lambda}^{\Lambda} |v(\sigma)|^2 d\sigma\}^{1/2}$. This allows us to cover the Neumann and Dirichlet problems in a unified setting. (There is a corresponding theory in the space $L^\infty[-\Lambda,\Lambda]$ which applies to the Dirichlet problem but not to the Neumann problem.) The next result follows directly from Theorem 3.5, using, e.g., [26, p. 326].

COROLLARY 3.6. *For $B = D$ or N, $\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B$ is a compact operator on $L^2[-\Lambda,\Lambda]$.*

The remainder of this section is devoted to proving the well-posedness of (3.28) in $L^2[-\Lambda,\Lambda]$. This is done in Corollary 3.8. Since $\widehat{\mathcal{L}}_B$ is a compact perturbation of $\widehat{\mathcal{K}}_B$, the key part of the proof of Corollary 3.8 is contained in the following theorem, which is of key importance also when we come to the numerical analysis in section 4.

THEOREM 3.7. *For $B = D$ or N, $(I + \widehat{\mathcal{K}}_B)^{-1}$ exists and is bounded on $L^2[-\Lambda,\Lambda]$.*

Proof. Since the proof follows standard procedures for dealing with Mellin convolution operators, we will be brief. More detail is in [13]. The first step is to write the operator $v \mapsto (I + \widehat{\mathcal{K}}_B)v$ on $L^2[-\Lambda,\Lambda]$ as two coupled convolution operators on $[0,\Lambda]$. For $(w_1,w_2) \in L^2[0,\Lambda] \times L^2[0,\Lambda]$ we introduce the norm $\|(w_1,w_2)\| = \{\|w_1\|^2_{L^2[0,\Lambda]} + \|w_2\|^2_{L^2[0,\Lambda]}\}^{1/2}$. Also we define the map $\Pi : L^2[-\Lambda,\Lambda] \to L^2[0,\Lambda] \times L^2[0,\Lambda]$ by

$$\Pi v := (v_1, v_2), \text{ where } v_1(s) = v(-s) + v(s) \quad \text{and} \quad v_2(s) = v(-s) - v(s), \quad s \in [0,\Lambda].$$

Clearly Π is a bijection and $\|\Pi v\|^2 = 2\|v\|^2_{L^2[-\Lambda,\Lambda]}$. Moreover, an elementary calculation using (3.34) and (3.35) and the analogous kernels for $B = N$ (see [13] for details) shows that

$$\Pi\widehat{\mathcal{K}}_B = \tilde{\mathbb{K}}_B\Pi, \quad B = D \text{ or } N. \qquad (3.42)$$

Here $\tilde{\mathbb{K}}_B$ is the matrix operator

$$\tilde{\mathbb{K}}_B = \begin{pmatrix} \widetilde{\mathcal{K}}_B & 0 \\ 0 & -\widetilde{\mathcal{K}}_B \end{pmatrix},$$

and $\widetilde{\mathcal{K}}_B$ is the *Mellin convolution* operator on $L^2[0,\Lambda]$ defined by

$$(\widetilde{\mathcal{K}}_B v)(s) = \int_0^\Lambda \tilde{\kappa}_B(s/\sigma)v(\sigma)\frac{d\sigma}{\sigma},$$

with kernels

$$\tilde{\kappa}_D(s) = -\frac{\sin(\lambda\pi)}{\pi}\frac{s}{1-2s\cos\lambda\pi+s^2}, \qquad \tilde{\kappa}_N(s) = \frac{\sin(\lambda\pi)}{\pi}\frac{1}{1-2s\cos\lambda\pi+s^2} \ .$$

Hence, for all $v \in L^2[-\Lambda,\Lambda]$, we have

$$(I+\widehat{\mathcal{K}}_B)v = \Pi^{-1}(I+\mathbb{K}_B)\Pi v. \tag{3.43}$$

It can be shown, using Mellin integral transform techniques [20], that $\|\widetilde{\mathcal{K}}_B\|_{L^2[0,\Lambda]} < 1$ (see [13] for further details). Hence by Banach's lemma $I \pm \widetilde{\mathcal{K}}_B$ has a bounded inverse on $L_2[0,\Lambda]$ and the result follows from (3.43). □

Corollary 3.6 and Theorem 3.7 can now be combined to obtain the well-posedness of (2.11) via the Fredholm alternative. The proof requires the injectiveness of $(I+\hat{\mathcal{L}}_B)$; i.e., we need to show that for all ν on the contour γ (see Figure 2), the implication

$$(I+\widehat{\mathcal{L}}_B)\widehat{u} = 0 \ \Rightarrow\ \widehat{u} = 0 \quad \text{for} \ \ \widehat{u} \in L^2[-\Lambda,\Lambda] \tag{3.44}$$

holds.

This implication is established in a standard way using uniqueness results for boundary-value problems for the PDE $\Delta^* + \nu^2 - 1/4$ on the manifolds M and $S^2\backslash\{M \cup \ell\}$ and the jump relations for the corresponding layer potentials on ℓ. The uniqueness can be easily established because the contour γ is constructed to avoid the eigenvalues of $-\Delta^* + 1/4$, while the jump relations may be found in [19] or [4] for the case of smooth ℓ, and a standard local analysis at corners will provide the extension of the jump relations to corner domains.

Corollary 3.8. *For $B = D$ or N, $(I+\widehat{\mathcal{L}}_B)^{-1}$ exists and is bounded on $L^2[-\Lambda,\Lambda]$.*

Proof. Using Theorem 3.7, the left-hand equation in (3.28) can be rewritten as

$$(I+(I+\widehat{\mathcal{K}}_B)^{-1}(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B))\widehat{u} = (I+\widehat{\mathcal{K}}_B)^{-1}\widehat{b}_B. \tag{3.45}$$

Since, by Corollary 3.6, $(I+\widehat{\mathcal{K}}_B)^{-1}(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)$ is a compact operator, it follows from the Fredholm alternative and the injectiveness property (3.44) that (3.45) has a unique solution. It also follows that the operator on the left-hand side of (3.45) has a bounded inverse. Therefore,

$$\|\widehat{u}\|_{L^2[-\Lambda,\Lambda]} \le C\|(I+\widehat{\mathcal{K}}_B)^{-1}\widehat{b}\|_{L^2[-\Lambda,\Lambda]} \le C'\|\widehat{b}_B\|_{L^2[-\Lambda,\Lambda]}$$

for some constants C and C', and the result follows. □

Remark 3.9. If the cone Ξ contains more than one lateral edge, then the contour ℓ will contain several corners. All the results of this subsection remain true in this case. In particular the analogue of Corollary 3.8 ensures the well-posedness of (3.28), or equivalently (2.11) in the multiple corner case. The proof is entirely analogous to the proof above, except that a pair of coupled Mellin convolution equations local to each corner has to be considered. Such systems are standard—see, e.g., [14].

Remark 3.10. The operator $\hat{\mathcal{L}}_B$ depends on the parameter ν, and further analysis will be required if one wishes to obtain a "stability bound" (i.e., a bound on $\|(I-\hat{\mathcal{L}}_B)^{-1}\|_{L_2[-\Lambda,\Lambda]}$ as a function of ν). However, we note that for the case $\nu = i\tau$ the corresponding Helmholtz operator on the plane was analyzed in [30], where a stability bound independent of τ was proved. The case $\nu = i\tau$ is particularly important in our computations—see section 5.

4. Numerical method. In this section we shall discuss the piecewise polynomial collocation method for (3.28) and obtain its convergence, using the results of section 3. We also describe briefly its efficient implementation. The performance of this scheme will be illustrated in section 5.

The basic collocation scheme is entirely standard, so we will be brief. First introduce a mesh:

$$-\Lambda = x_0 < x_1 < \cdots < x_m < x_{m+1} < \cdots < x_n = \Lambda\ . \tag{4.1}$$

We assume here that ℓ has a single corner situated at $x_m = 0$ in parameter space and that $n = 2m$. The case of several corners is similar (see Remark 3.9 and the remarks below (4.5)), and that of a smooth boundary is straightforward (see [13]). We define $I_i = [x_{i-1}, x_i]$ and $h_i = x_i - x_{i-1}$ for $i = 1, \dots, n$. We assume that for each integer $r \geq 1$, we have chosen, a priori, r points: $0 < \xi_1^r < \xi_2^r < \cdots < \xi_r^r < 1$. Then we introduce the approximation space

$$S_n^r[-\Lambda,\Lambda] = \{v \in L^\infty[-\Lambda,\Lambda] : v|_{I_i} \in P_r\}\ , \tag{4.2}$$

where P_r denotes the set of polynomials of order $r \geq 1$ (i.e., of degree $r-1$). Also, on each I_i, we define the r collocation points $x_{ij}^r = x_{i-1} + h_i \xi_j^r$, and we define the basis functions of $S_n^r[-\Lambda,\Lambda]$ by

$$\phi_{ij}(x) = \begin{cases} \displaystyle\prod_{\substack{1\le k\le r \\ k\neq j}} \frac{x - x_{ik}^r}{x_{ij}^r - x_{ik}^r}\chi_i(x) & \text{when} \quad r > 1 \\ \chi_i(x) & \text{when} \quad r = 1 \end{cases}$$

for $j = 1, \dots, r$ and $i = 1, \dots, n$, where χ_i is the characteristic function on I_i. Clearly $\phi_{ij}|_{I_i} \in P_r$ and $\phi_{ij}(x_{i'j'}) = \delta_{ii'}\delta_{jj'}$.

In the collocation method for (3.28), we seek an approximate solution

$$\widehat{u}_n(s) := \sum_{i=1}^n \sum_{j=1}^r \mu_{ij}\phi_{ij}(s),$$

where μ_{ij} are chosen so that the residual vanishes at the collocation points:

$$\mu_{i'j'} + \sum_{i=1}^n \sum_{j=1}^r \mu_{ij} \int_{I_i} \widehat{L}_B(x_{i'j'}^r, \sigma)\phi_{ij}(\sigma)d\sigma = \widehat{b}_B(x_{i'j'}^r) \ \text{ for } \ i' = 1, \dots, n\ ,\ j' = 1, \dots, r\ . \tag{4.3}$$

Equivalently,

$$(I + \widehat{\mathcal{P}}_n\widehat{\mathcal{L}}_B)\widehat{u}_n = \widehat{\mathcal{P}}_n\widehat{b}_B\ , \tag{4.4}$$

where $\widehat{\mathcal{P}}_n$ denotes the operator onto $S_n^r[-\Lambda,\Lambda]$ defined by interpolation at the points $\{x_{i,j}\}$. Because the ξ_j^r are chosen interior to $[0,1]$, none of the points x_{ij} are corner points, and so $\widehat{\mathcal{P}}_n\widehat{\mathcal{L}}_B\widehat{u}_n$ and $\widehat{\mathcal{P}}_n\widehat{b}_B$ are well-defined in (4.3).

In the *h version* of the collocation method (with r fixed and $n \to \infty$), we adopt the usual a priori mesh grading:

$$x_{m\pm i} = \pm(i/m)^q \Lambda \qquad \text{for } i = 0, \dots, m, \tag{4.5}$$

where $q \geq 1$ is the grading exponent. Note that the corner point in parameter space ($x_m = 0$) is a mesh point. This is important. If ℓ has several corners, we would simply use meshes like (4.5) local to each corner, joined together with quasi-uniform refinement away from the corners in an obvious way.

To obtain the stability of the collocation scheme, we need the concept of a *modification parameter* $i* \geq 0$ (first introduced in [14]). The *modified collocation scheme* is exactly the same as (4.3) when $i* = 0$. But when $i* \geq 1$, $\widehat{u}_n$ is set to 0 on each of the subintervals $I_i, i = m - i* + 1, \dots, m + i*$, and equations (4.3) are required to hold only for $i' \notin \{m - i* + 1, \dots m + i*\}$. (In other words, the collocation solution is set to 0 on each of the $2i*$ subintervals nearest the corner and (4.3) is not required to hold on those subintervals.) For notational convenience we shall continue to write the collocation equations as (4.4), thus suppressing $i*$ from the notation.

THEOREM 4.1. *Let r and q be fixed and let $B = D$ or N. Then there exists a modification parameter $i* \geq 1$ independent of n, and a constant C which may depend on r, q, and $i*$ but not on n such that $\|(I + \widehat{\mathcal{P}}_n \widehat{\mathcal{L}}_B)^{-1}|_{S_n^r[-\Lambda,\Lambda]}\|_{L^2[-\Lambda,\Lambda]} \leq C$ for all sufficiently large n; i.e., the collocation method* (4.4) *is stable in* $L^2[-\Lambda, \Lambda]$.

Proof. We shall show that, for each $\epsilon > 0$, there exists a modification such that, for n sufficiently large,

$$\|(I - \widehat{\mathcal{P}}_n)\widehat{\mathcal{L}}_B v_n\|_{L^2[-\Lambda,\Lambda]} \leq \epsilon \|v_n\|_{L^2[-\Lambda,\Lambda]} \tag{4.6}$$

for all $v_n \in S_n^r[-\Lambda, \Lambda]$. Then, since

$$I + \widehat{\mathcal{P}}_n \widehat{\mathcal{L}}_B = (I + \widehat{\mathcal{L}}_B) - (I - \widehat{\mathcal{P}}_n)\widehat{\mathcal{L}}_B,$$

existence and stability of $(I + \widehat{\mathcal{P}}_n \widehat{\mathcal{L}}_B)^{-1}$ on $S_n^r[-\Lambda, \Lambda]$ follow from Corollary 3.8.

To obtain (4.6), note that by the triangle inequality,

$$\begin{aligned} \|(I - \widehat{\mathcal{P}}_n)\widehat{\mathcal{L}}_B v_n\|_{L^2[-\Lambda,\Lambda]} \leq{} & \|(I - \widehat{\mathcal{P}}_n)\widehat{\mathcal{K}}_B v_n\|_{L^2[-\Lambda,\Lambda]} \\ & + \|(I - \widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B) v_n\|_{L^2[-\Lambda,\Lambda]}. \end{aligned} \tag{4.7}$$

Now recall that $\widehat{\mathcal{P}}_n$ projects to zero on the $2i*$ intervals nearest 0. Thus

$$\begin{aligned} \|(I - \widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B) v_n\|^2_{L^2[-\Lambda,\Lambda]} \leq{} & \|(\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B) v_n\|^2_{L^2[x_{m-i*}, x_{m+i*}]} \\ & + \|(I - \widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B) v_n\|^2_{L^2([-\Lambda,\Lambda]\setminus[x_{m-i*}, x_{m+i*}])}. \end{aligned} \tag{4.8}$$

By Theorem 3.5, $\widehat{L}_B - \widehat{K}_B$ is a bounded function, and this implies that $(\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B)$ is compact from $L^2[-\Lambda, \Lambda]$ to $L^\infty[-\Lambda, \Lambda]$ (see [26, pp. 534–535]). Thus the first term on the right-hand side of (4.8) may be estimated by

$$\begin{aligned} \|(\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B) v_n\|^2_{L^2[x_{m-i*}, x_{m+i*}]} & \leq 2x_{m+i*} \|(\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B) v_n\|^2_{L^\infty[x_{m-i*}, x_{m+i*}]} \\ & \leq C n^{-q} \|v_n\|^2_{L^2[-\Lambda,\Lambda]}. \end{aligned} \tag{4.9}$$

(Throughout the proof, C denotes a generic constant which is independent of n but may depend on the other parameters.)

We now consider the second term on the right-hand side of (4.8). First we write

$$\begin{aligned}\|(I-\widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|^2_{L^2([-\Lambda,\Lambda]\setminus[x_{m-i*},x_{m+i*}])} &= \sum_{i\leq m-i*}\|(I-\widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|^2_{L^2(I_i)} \\ &\quad + \sum_{i\geq m+i*+1}\|(I-\widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|^2_{L^2(I_i)}.\end{aligned} \tag{4.10}$$

We will estimate the second sum in (4.10). (The first sum can be dealt with in a similar way.) To do this we recall the standard results for piecewise polynomial interpolation and write

$$\begin{aligned}\sum_{i\geq m+i*+1}\|(I-\widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|^2_{L^2(I_i)} &\leq C\sum_{i\geq m+i*+1} h_i^2\|D(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|^2_{L^2(I_i)} \\ &\leq C\sum_{i\geq m+i*+1} h_i^3\|s^{-1}sD(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|^2_{L^\infty(I_i)}.\end{aligned}$$

It can be shown, using the same argument as in the proof of Theorem 3.5, that the operator $sD(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)$ has a bounded kernel. Hence, noting that $h_i\leq Cn^{-1}$, we obtain

$$\begin{aligned}\sum_{i\geq m+i*+1}\|(I-\widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|^2_{L^2(I_i)} &\leq Cn^{-1}\sum_{i\geq m+i*+1}(h_i x_{i-1}^{-1})^2\|v_n\|^2_{L^2[-\Lambda,\Lambda]} \\ &\leq C\max_{i\geq m+i*+1}(h_i x_{i-1}^{-1})^2\|v_n\|^2_{L^2[-\Lambda,\Lambda]}.\end{aligned} \tag{4.11}$$

Now for $i\geq i*+1$, (4.5) implies

$$h_{m+i}=\Big(\frac{i}{m}\Big)^q\Lambda-\Big(\frac{i-1}{m}\Big)^q\Lambda\leq q\Lambda\frac{1}{m}\Big(\frac{i}{m}\Big)^{q-1}.$$

Hence, since $i*$ satisfies $i*\geq 1$,

$$h_{m+i}x_{m+i-1}^{-1}\leq q\frac{1}{m}\Big(\frac{i}{m}\Big)^{q-1}\Big(\frac{m}{i-1}\Big)^q\leq Cq\frac{1}{i-1}\leq Cq\frac{1}{i*}. \tag{4.12}$$

By substituting (4.12) into (4.11) it follows that

$$\sum_{i\geq m+i*+1}\|(I-\widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|^2_{L^2(I_i)}\leq C\Big(\frac{1}{i*}\Big)^2\|v_n\|^2_{L^2[-\Lambda,\Lambda]}.$$

A similar estimate holds for the first sum in (4.10) and so

$$\begin{aligned}\|(I-\widehat{\mathcal{P}}_n)(\widehat{\mathcal{L}}_B-\widehat{\mathcal{K}}_B)v_n\|_{L^2([-\Lambda,\Lambda]\setminus[x_{m-i*},x_{m+i*}])} &\leq C\frac{1}{i*}\|v_n\|_{L^2[-\Lambda,\Lambda]} \\ &\leq \frac{\epsilon}{2}\|v_n\|_{L^2[-\Lambda,\Lambda]}\end{aligned} \tag{4.13}$$

for sufficiently large $i*$.

By (4.7), (4.8), (4.9), and (4.13), we see that to prove (4.6) it is sufficient to prove it with $\widehat{\mathcal{L}}_B$ replaced by $\widehat{\mathcal{K}}_B$. However, this follows from now-classical results about numerical methods for Mellin convolution equations. To explain briefly, we first employ the operators Π and $\tilde{\mathbb{K}}_B$ defined in the proof of Theorem 3.7, as well as the fact that the mesh (4.1) is symmetric about 0, to obtain $\Pi\widehat{\mathcal{P}}_n\widehat{\mathcal{K}}_B = \tilde{\mathbb{P}}_n\tilde{\mathbb{K}}_B\Pi$, where

$$\tilde{\mathbb{P}}_n = \begin{pmatrix} \tilde{\mathcal{P}}_n & 0 \\ 0 & \tilde{\mathcal{P}}_n \end{pmatrix},$$

with $\tilde{\mathcal{P}}_n$ defined as the restriction of $\widehat{\mathcal{P}}_n$ to functions on $[0,\Lambda]$. Since $(I - \widehat{\mathcal{P}}_n)\widehat{\mathcal{K}}_B = \Pi^{-1}(I - \widetilde{\mathbb{P}}_n)\widetilde{\mathbb{K}}_B\Pi$, the result follows if, for all $\epsilon > 0$, there exists a modification $i*$ such that

$$\|(I - \tilde{\mathcal{P}}_n)\tilde{\mathcal{K}}_B v_n\|_{L^2[0,\Lambda]} \leq \epsilon\|v_n\|_{L^2[0,\Lambda]} \tag{4.14}$$

for all $v_n \in S_n^r[0,\Lambda]$ and for sufficiently large n. However, result (4.14) follows (even for all n) from the general results in the survey [22]. (See Theorem 3.1 there, and the remarks following it. Note that $\tilde{\kappa}_B$ and $\tilde{\kappa}_N$ both satisfy conditions (A1) and (A2) of [22], with $p = 2$.) See [23, 14, 20] and also [13] for more details. □

Remark 4.2. The introduction of the parameter $i*$ is solely a device to prove stability of the collocation method for (2.11) when ℓ contains a corner. No unmodified practical collocation method has ever been observed to be unstable. However, the proof that these methods are stable without modification has eluded researchers for 15 years. For this reason, and to simplify the presentation, we assume that Theorem 4.1 holds for $i* = 0$ (i.e., no modification) for the remainder of this section. All the following results also hold for $i* \geq 1$, but the proofs require slightly different technicalities.

Theorem 4.1 implies that the collocation equation (4.4) is uniquely solvable for all n sufficiently large. An easy manipulation using (3.28) and (4.4) shows that $(I + \widehat{\mathcal{P}}_n\widehat{\mathcal{L}}_B)(\widehat{\mathcal{P}}_n\widehat{u} - \widehat{u}_n) = -\widehat{\mathcal{P}}_n\widehat{\mathcal{L}}_B(I - \widehat{\mathcal{P}}_n)\widehat{u}$. Theorem 4.1 then implies

$$\|\widehat{\mathcal{P}}_n\widehat{u} - \widehat{u}_n\|_{L^2[-\Lambda,\Lambda]} \leq C\|\widehat{\mathcal{P}}_n\widehat{\mathcal{L}}_B(I - \widehat{\mathcal{P}}_n)\widehat{u}\|_{L^2[-\Lambda,\Lambda]}. \tag{4.15}$$

After some technical manipulations using properties of $\widehat{\mathcal{L}}_B = \widehat{\mathcal{K}}_B + (\widehat{\mathcal{L}}_B - \widehat{\mathcal{K}}_B)$ it can be shown that the right-hand side of (4.15) can be bounded by a constant multiple of $\|(I - \widehat{\mathcal{P}}_n)\widehat{u}\|_{L^2[-\Lambda,\Lambda]}$ (see [13]). Then the triangle inequality implies

$$\|\widehat{u} - \widehat{u}_n\|_{L^2[-\Lambda,\Lambda]} \leq \|\widehat{u} - \widehat{\mathcal{P}}_n\widehat{u}\|_{L^2[-\Lambda,\Lambda]} + \|\widehat{\mathcal{P}}_n\widehat{u} - \widehat{u}_n\|_{L^2[-\Lambda,\Lambda]} \leq C\|(I - \widehat{\mathcal{P}}_n)\widehat{u}\|_{L^2[-\Lambda,\Lambda]}. \tag{4.16}$$

Therefore to obtain convergence rates we need estimates for $\|(I - \widehat{\mathcal{P}}_n)\widehat{u}\|_{L^2[-\Lambda,\Lambda]}$. These of course depend on the regularity of the solution. To describe this regularity we introduce the weighted Sobolev space for an interval $J \subset \mathbb{R}$ and for $k \in \mathbb{N}$ and $\alpha \in \mathbb{R}$,

$$L_\alpha^{2,k}(J) = \{v : |x|^{j-\alpha}D^j v \in L^2(J),\ j = 0, 1, \dots, k\},$$

equipped with the norm $\|v\|_{L_\alpha^{2,k}(J)} = \sum_{j=0}^k \|x^{j-\alpha}D^j v\|_{L^2(J)}$ (see [20]).

Examples 4.3.

(i) The function

$$\widehat{u}(x) = C' + C''|x|^\theta, \quad \text{where} \quad 1/2 < \theta < 1, \tag{4.17}$$

satisfies $\widehat{u}(x) - C' \in L_\alpha^{2,k}[-\Lambda,\Lambda]$ for all $k \geq 0$ and $\alpha < \theta + 1/2$.

(ii) The function

$$\widehat{u}(x) = C|x|^{\theta-1}, \quad \text{where} \quad 1/2 < \theta < 1, \tag{4.18}$$

satisfies $\widehat{u}(x) \in L^{2,k}_{\alpha}[-\Lambda, \Lambda]$ for all $k \geq 0$ and $\alpha < \theta - 1/2$.

Remark 4.4. When we solve the Dirichlet problem for the Laplace equation in the region interior to a planar polygon using the indirect boundary integral method, the solution of the resulting integral equation has its principal singularity in the form (4.17), where the corner is at $x = 0$ and $\theta = 1/(1 + |\chi|)$, where$(1 - \chi)\pi$ is the angle between the tangents at the corner ($\chi \in (-1, 1)\backslash\{0\}$). When we solve the Neumann problem with the same geometry again using the indirect boundary integral method, the density has its principal singularity in the form (4.18), again with $\theta = 1/(1 + |\chi|)$ (see, e.g, [17, 23, 20]). It can be shown by standard local analysis (see, e.g., [32]) that the solutions of our integral equations have the same principal singularity as identified in Examples 4.3.

Estimates for $\|(I - \widehat{\mathcal{P}}_n)\widehat{u}\|_{L^2[-\Lambda,\Lambda]}$ under assumptions which encapsulate Examples 4.3(i) and (ii) are well known (see, e.g., [22]). Combining these with (4.16), we obtain the final result given below (see also [13] for more details).

Theorem 4.5. *Consider the collocation method* (4.4) *and assume that stability holds in the sense of Theorem* 4.1.

(i) *Suppose that $B = D$ and that the exact solution to* (3.28) *satisfies $\widehat{u} - C' \in L^{2,r}_{\alpha}[-\Lambda, \Lambda]$ with $1 < \alpha < 3/2$. Then for sufficiently large n the collocation method described by* (4.4) *converges with error*

$$\|\widehat{u} - \widehat{u}_n\|_{L^2[-\Lambda,\Lambda]} \leq Cn^{-r}\|\widehat{u} - C'\|_{L^{2,r}_{\alpha}[-\Lambda,\Lambda]} \quad \textit{as} \quad n \to \infty, \tag{4.19}$$

provided the grading parameter $q \geq \max\{r/\alpha, 1\}$.

(ii) *Suppose that $B = N$ and that the exact solution to* (3.28) *satisfies $\widehat{u} \in L^{2,r}_{\alpha}[-\Lambda, \Lambda]$ for some $0 < \alpha < 1/2$. Then for sufficiently large n the collocation method described by* (4.4) *converges with error*

$$\|\widehat{u} - \widehat{u}_n\|_{L^2[-\Lambda,\Lambda]} \leq Cn^{-r}\|\widehat{u}\|_{L^{2,r}_{\alpha}[-\Lambda,\Lambda]} \quad \textit{as} \quad n \to \infty, \tag{4.20}$$

provided the grading parameter $q \geq r/\alpha$.

The implementation of the collocation method (4.3) requires the efficient calculation of the stiffness matrix entries

$$\widehat{\mathbb{L}}_{i'j',ij} := \int_{I_i} \widehat{L}_B(x^r_{i'j'}, \sigma)\phi_{ij}(\sigma)d\sigma. \tag{4.21}$$

Each evaluation of the kernel $\widehat{L}_B$ in (4.21) requires an evaluation of (the derivative) of the Legendre function with complex index (see (3.29), (3.30)). We do this by integrating Legendre's differential equation using a Runge–Kutta method (cf. [4, 5, 6, 18]—details are in [13]). Thus efficient quadrature methods for (4.21) are of the utmost importance. This is especially true when we remember that (2.11) needs to be solved many times over (for different values of ν on the imaginary axis) in order to allow the approximate integration of (1.4). The main difficulty in evaluating (4.21) is the singularity which arises when $i' = i$. (This is strongest when I_i contains the origin in parameter space, corresponding to the corner on ℓ.) In [13] a detailed study of quadrature for (4.21) is carried out. Here we have room to mention only the most useful result from [13], as follows.

Theorem 4.6. *Suppose the collocation points x^r_{ij}, $j = 1,\ldots,r$, are chosen to be the r Gauss–Legendre points on* $[0,1]$*, shifted to I_i. Suppose that* (4.21) *is approximated by the Gauss–Legendre rule based at these points for all i, i' satisfying*

$$\mathrm{dist}(I'_i, I_i) \geq h_i^{1/(r+2)}, \tag{4.22}$$

and the remaining entries of (4.21) *are computed exactly. Then the $O(n^{-r})$ convergence rate reported in Theorem* 4.5 *continues to hold.*

Since ϕ_{ij} vanishes at all the points x^r_{ik}, except $k = j$, the implementation of the rule in Theorem 4.6 requires only one kernel evaluation, and (4.22) shows that this can be done for most of the matrix as the mesh is refined. It also turns out that even when (4.22) is not satisfied, rules with $O(\log(n))$ kernel evaluations can be employed and the $O(n^{-r})$ rate in Theorem 4.5 remains unperturbed—for more details see section 5 and also [13].

5. Numerical results. We shall illustrate the performance of the numerical method described above in the case of the diffraction of acoustic waves by a *trihedral cone*. In the diffraction literature this is an unsolved *canonical problem*; i.e., it is a relatively simple geometry which often occurs in applications, but there is no known closed form expression for the diffraction coefficients.

Our trihedral cone is determined by three edges which emanate from the origin and pass through the points $\boldsymbol{\omega}_{c_i} \in S^2$, $i = 1, 2, 3$, which are specified by spherical polar coordinates $(\theta^*, 0), (\theta^*, 2\pi/3)$, and $(\theta^*, 4\pi/3)$, respectively, where $\cos\theta^* = 1/\sqrt{3}$. Hence the edges are mutually perpendicular. The conical scatterer Ξ therefore has its surface composed of three mutually perpendicular planar segments determined by each pair of edges, and the contour ℓ is made up of three smooth geodesic curves in S^2, with each pair of smooth curves meeting at an angle of $\pi/2$ at one of the points $\boldsymbol{\omega}_{c_i}$. The geometry is depicted in Figure 4. The contour ℓ is drawn in bold. (This corresponds to the practically important case of the corner of a rectangular building.)

Throughout the computations we used collocation at the Gauss points of subintervals. For the evaluation of the boundary integrals (4.21), we used Gauss quadrature at the collocation points in the "far field," i.e., when i, i' satisfy (4.22). When (4.22) does not hold we increase the number of quadrature points, d, logarithmically. More precisely, we choose d to be the smallest integer satisfying

$$d \geq \frac{(r+1)\log(n)}{2\log(2)}.$$

This heuristic is motivated by an analysis in [13]. Note that for this geometry, when $\boldsymbol{\omega}, \boldsymbol{\omega}'$ lie on the same edge of the geodesic triangle ℓ, then $L_B(\boldsymbol{\omega}, \boldsymbol{\omega}') = 0$. Hence one-third of the matrix entries are zero. Included in these zero entries are the integrals that occur when the collocation point lies in the interval of integration. Note that our procedure uses only one kernel evaluation for most matrix entries, as mentioned in section 4. We shall see that our numerical results coincide with the theoretical predictions of Theorem 4.5.

Our first set of results illustrate the accuracy of methods for solving the integral equation (2.11) (equivalently (3.28)) arising from the boundary value problem (2.5), (2.6). For these tests we set $\boldsymbol{\omega}_0 = -\boldsymbol{\omega}_{c_1}$ and set the parameter $\nu = i$.

The density $\widehat{u}$ in (3.28) is not smooth near the corner. In fact, in the case of the Dirichlet problem, we expect from Remark 4.4 that there exists a constant C' such that $\widehat{u} - C' \in L^{2,r}_{\alpha}$, with $\alpha < 7/6$. (This is because for the corners in this example

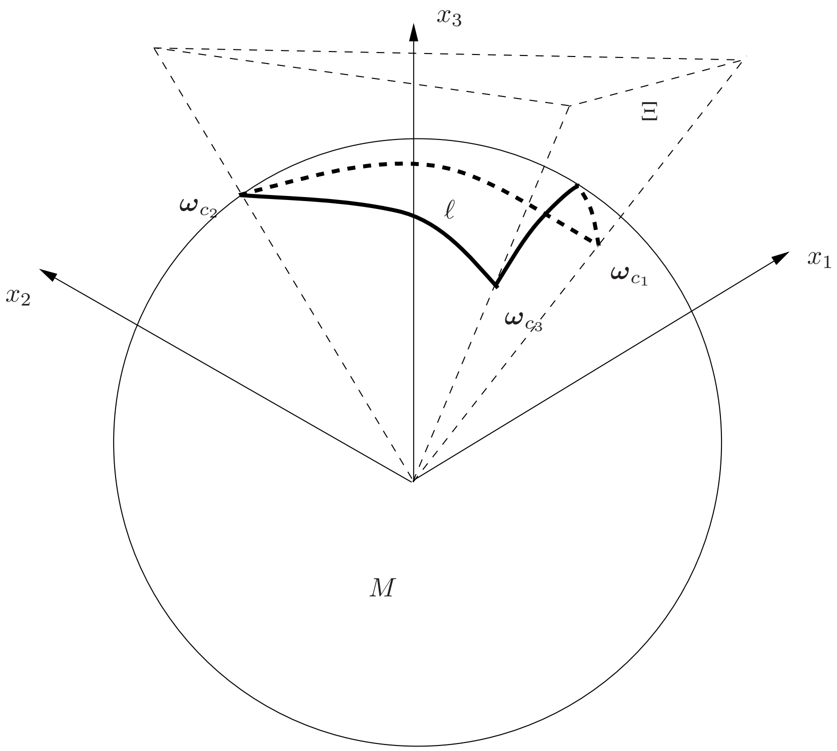

x_3
Ξ
$\boldsymbol{\omega}_{c_2}$
ℓ
x_1
x_2
$\boldsymbol{\omega}_{c_1}$
$\boldsymbol{\omega}_{c_3}$
M

Fig. 4. *The contour ℓ associated with a trihedral cone.*

Table 1

Estimated errors for densities $\hat{u}$ using the piecewise constant collocation method for (3.28) *on a uniform mesh, $q = 1$.*

	Dirichlet problem		Neumann problem	
n	err_n^1	Ratio	err_n^1	Ratio
24	9.957E-2		1.609E-3	
48	5.285E-2	1.88	1.530E-3	1.05
96	2.472E-2	2.14	1.229E-3	1.24
192	1.074E-2	2.30	1.077E-3	1.14
384	4.992E-3	2.15	9.589E-4	1.12

$\chi = 1/2$, so $\theta = 2/3$ and hence $\theta + 1/2 = 7/6$.) When Neumann boundary conditions are prescribed, we expect $\widehat{u} \in L^{2,r}_{\alpha}$, $\alpha < 1/6$. So, for the Dirichlet problem, piecewise constant approximation ($r = 1$) should (by Theorem 4.5) yield optimal convergence (i.e., $O(n^{-1})$ in the L^2 norm) on a uniform mesh ($q = 1$ in (4.5)). On the other hand for the Neumann problem we expect (by a generalization of Theorem 4.5) a rate of convergence close to $O(n^{-1/6})$ on a uniform mesh.

To illustrate convergence, for each case we have computed an "exact" solution $\widehat{u}^*$ by using piecewise linear collocation on a mesh with 498 nodes. (To obtain the "exact" Dirichlet solution we grade the mesh towards the corners with $q = 2$, and for the "exact" Neumann solution, since the grading required to obtain optimal convergence is rather severe, we use here a grading exponent $q = 3$.) We computed the approximate L^2 error $\mathrm{err}_n^1 := \|\widehat{u}^* - \widehat{u}_n\|_2$ using midpoint quadrature with respect to the mesh with n subintervals. The results are given in Table 1. As expected, a convergence rate of close

Table 2
Estimated errors for densities $\hat{u}$*, using the piecewise constant collocation method for* (3.28) *on a graded mesh,* $q = 3$*.*

	Dirichlet problem		Neumann problem	
n	err_n^1	Ratio	err_n^1	Ratio
24	1.257E-2		6.307E-3	
48	4.948E-3	2.54	6.106E-3	1.03
96	2.147E-3	2.30	4.744E-3	1.29
192	7.842E-4	2.74	3.553E-3	1.34
384	2.442E-4	3.21	2.738E-3	1.30

to $O(n^{-1})$ is observed for the Dirichlet problem and close to $O(n^{-1/6})$ for the Neumann problem. (In Tables 1–4, “ratio” is defined to be $\mathrm{err}_n^i/\mathrm{err}_{n-1}^i$ for $i = 1$ or 2.) As we have shown, mesh grading will improve suboptimal rates of convergence. Consider the integral equation arising from the Neumann problem. Because its solution satisfies $\widehat{u} \in L_\alpha^{2,r}$, with $\alpha < 1/6$, it can be shown (by the methods of Theorem 4.5) that with $q' \leq 6r$ a rate of convergence of $O(n^{-q'/6})$ in the L^2 norm can be attained when a graded mesh is used with grading exponent $q > q'$ for collocation onto piecewise polynomials of order r. We illustrate the correctness of this result with $q = 3$. The results are in Table 2. Here we find that the Neumann problem now converges with rate close to $O(n^{-1/2})$, as expected. The Dirichlet problem now appears to converge with a superoptimal rate, but this could be expected to subside back to $O(n^{-1})$ asymptotically. These results indicate that our integral equation solver is working as predicted by the theory.

Our next set of results illustrates the convergence of the approximate solutions to the spherical boundary-value problem (2.5), (2.6). We consider the same problem as above with $\boldsymbol{\omega}_0 = -\boldsymbol{\omega}_{c_1}$ and $\nu = i$. In Tables 3 and 4, we tabulate the errors in approximate solutions to (2.5), (2.6) obtained by substituting $u_n(\boldsymbol{\rho}(s), \nu) = \widehat{u}_n(s)$ into (2.7) (in the Dirichlet case) and (2.9) (in the Neumann case) and computing the resulting integrals by the Gauss quadrature rule based at the points used in the computation of $\widehat{u}_n$. For illustration we have chosen to observe the solution at the particular observation direction $\boldsymbol{\omega} = (0, 0, -1)$. The error err_n^2 is computed by $|g_n^r(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu) - \widetilde{g}^r(\boldsymbol{\omega}, \boldsymbol{\omega}_0, \nu)|$, where $\widetilde{g}^r$ is computed with a large n $(= 330)$ and $q = 3$.

The results illustrate the superconvergence of the method (well documented in the case of planar problems; see, e.g., [14, 3, 22]), with close to $O(n^{-2})$ convergence attained for $q = 3$. The extreme gradings needed for optimal convergence of the density may not be needed for the potential, and in fact better than optimal convergence may be obtained because of the smoothness of the fundamental solution away from the boundary ℓ.

We emphasize that the results in Tables 1–4 illustrate not only the convergence theory in section 4, but also show that the algorithm used to compute the Legendre functions with complex index (by applying a Runge–Kutta method to Legendre’s differential equation), which is described in detail in [13], is working in a stable manner.

The results given here involve approximation with piecewise constant basis functions. Results for piecewise linears are given in [13]. An important point is that, since only one kernel evaluation is needed for most matrix entries *independent* of the order of the basis functions, the cost of implementation does not increase much as the order of the basis functions is increased. This suggests that the h-p version of the boundary element method should be very competitive for this application, and our next set of

Table 3

Estimated errors for the potential (2.7) *using the piecewise constant collocation method (Dirichlet boundary conditions).*

	Uniform mesh, $q = 1$		Graded mesh, $q = 2$		Graded mesh, $q = 3$	
n	err_n^2	Ratio	err_n^2	Ratio	err_n^2	Ratio
12	3.12E-4		3.16E-4		4.70E-4	
24	1.36E-4	2.3	1.35E-4	2.4	1.62E-4	2.9
48	5.43E-5	2.5	4.21E-5	3.2	6.12E-5	2.7
96	2.13E-5	2.6	1.34E-5	3.2	2.02E-5	3.0
192	8.84E-6	2.4	4.09E-6	3.3	5.91E-6	3.4

Table 4

Estimated errors for the potential (2.9) *using the piecewise constant collocation method (Neumann boundary conditions).*

	Uniform mesh, $q = 1$		Graded mesh, $q = 2$		Graded mesh, $q = 3$	
n	err_n^2	Ratio	err_n^2	Ratio	err_n^2	Ratio
12	6.25E-5		3.74E-5		4.45E-5	
24	2.68E-5	2.3	1.02E-5	3.7	8.07E-6	5.5
48	1.11E-5	2.4	2.93E-6	3.5	3.08E-6	2.6
96	4.54E-6	2.5	7.72E-7	3.8	8.73E-7	3.5
192	1.82E-6	2.5	2.05E-7	3.8	2.35E-7	3.7

results concern this method.

For fixed $\sigma \in (0,1)$ we define a geometrically graded mesh on $[-\Lambda, \Lambda]$ by

$$(5.1) \qquad x_{m+i} = \sigma^{m-i}\Lambda, \qquad -x_{m-i} = \sigma^{m-i}\Lambda, \qquad i = 1, \dots, m, \qquad x_m = 0.$$

Instead of seeking an approximate solution in the space S_n^r of piecewise polynomials of fixed order r on each subinterval, we allow a variable order r_i on each subinterval I_i (see (4.1) and the remarks following). A typical distribution of orders would be

$$r = \lceil (m+1-i)\beta \rceil \quad \text{for } i < m, \qquad r = \lceil (i-m)\beta \rceil \quad \text{for } i > m+1$$

for some fixed parameter $\beta > 0$, where, for $x \in \mathbb{R}$, $\lceil x \rceil$ denotes the smallest integer which is strictly greater that x. On the intervals I_i, $i = m, m+1$, the approximate solution is set to zero. Thus, on intervals close to the corner we approximate the solution on small subintervals, using low order methods, while further away we use higher orders on larger subintervals. The maximum order increases linearly with m and hence also with n. This is a standard prescription (e.g., [21]).

By making use of the fundamental results of Elschner [21] for the Laplace case, and combining these with our results in section 3, it can be shown [13] that the h-p method is stable. By making further assumptions about the regularity of the solution to (3.28), it can be shown that the h-p method converges exponentially. In Figure 5 we illustrate the convergence of the h-p method, compared with the piecewise constant and piecewise linear cases for the potentials (2.7) arising from the Dirichlet problem with $\boldsymbol{\omega}_0 = -\boldsymbol{\omega}_{c_1}$, $\boldsymbol{\omega} = (0, 0, -1)$, and $\nu = i$.

In these computations, the parameter values $\sigma = 0.25$ and $\beta = 0.5$ were employed in the h-p method. For these results we naively used the r-point Gauss–Legendre rule to calculate the matrix entries $\widehat{\mathbb{L}}_{i'j',ij}$; i.e., in this case *all* of the matrix entries were computed using one kernel evaluation. Observe the exponential convergence of the h-p method in Figure 5. (For another way to obtain exponential convergence for this type of integral equation, see [29].)

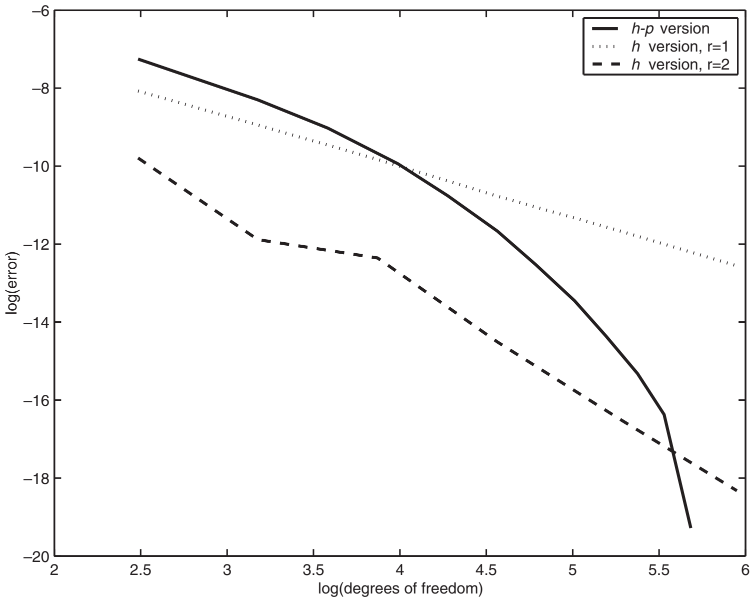


FIG. 5. *Errors for the potential* (2.7) *for the h version and h-p version of collocation.*

Finally, in order to illustrate the computations of the diffraction coefficients for this geometry, we shall show graphically how the computed $f(\boldsymbol{\omega}, \boldsymbol{\omega}_0)$ in (1.4) varies for three different incidence directions $\boldsymbol{\omega}_0$, and many observation directions $\boldsymbol{\omega}$ ranging over a subdomain of M. In this illustration we restrict ourselves to the Dirichlet problem, and we consider the incident directions given in spherical polar coordinates (θ, ϕ) by

$$\boldsymbol{\omega}_0 = (\pi, 0),\ (11\pi/12, 0),\ \text{and}\ (5\pi/6, 0), \tag{5.2}$$

and a range of observation directions

$$\boldsymbol{\omega} = ((\pi - \theta), \phi), \quad \text{with} \quad 0 \le \theta \le \pi/3,\ \ 0 \le \phi \le 2\pi. \tag{5.3}$$

In Figure 6 we illustrate how $|f(\boldsymbol{\omega}, \boldsymbol{\omega}_0)|$ varies as a function of θ and ϕ for each of the three different incident angles. The quantity $|f(\boldsymbol{\omega}, \boldsymbol{\omega}_0)|$ is plotted on the x_3 axis against the projection of $\boldsymbol{\omega}$ onto the x_1x_2-plane given by $\boldsymbol{\omega} = (\pi - \theta, \phi) \mapsto (\theta \cos\phi, \theta \sin\phi)$.

Observe in the first row of Figure 6 that when $\boldsymbol{\omega}_0 = (\pi, 0)$, i.e., the incident wave propagates in an "axial" direction, then the magnitude of the diffraction coefficients is smallest in the backscattering direction. This is in qualitative agreement with results for the circular cone [4]. Also note when $\boldsymbol{\omega}_0 = (\pi, 0)$ that if we fix $\theta > 0$, then the distance between $\boldsymbol{\omega} = (\pi - \theta, \phi)$ and the boundary of the nonsingular region, given by $\theta_1(\boldsymbol{\omega}, \boldsymbol{\omega}_0) = \pi$—see (2.1)—is smallest when $\phi = 0, 2\pi/3, 4\pi/3$. At the singular directions f is infinitely large; hence the three peaks appear in the first row of Figure 6.

As expected, the results are symmetric with respect to rotations by $\pm 2\pi/3$ about the axis. As we vary the angle of incidence, the symmetry breaks and the position of the singular directions will vary. In particular it can be shown from (2.1) that for $\boldsymbol{\omega}_0 = (11\pi/12, 0)$ and $(5\pi/6, 0)$ and fixed $\theta > 0$, the distance between $\boldsymbol{\omega}$ and the

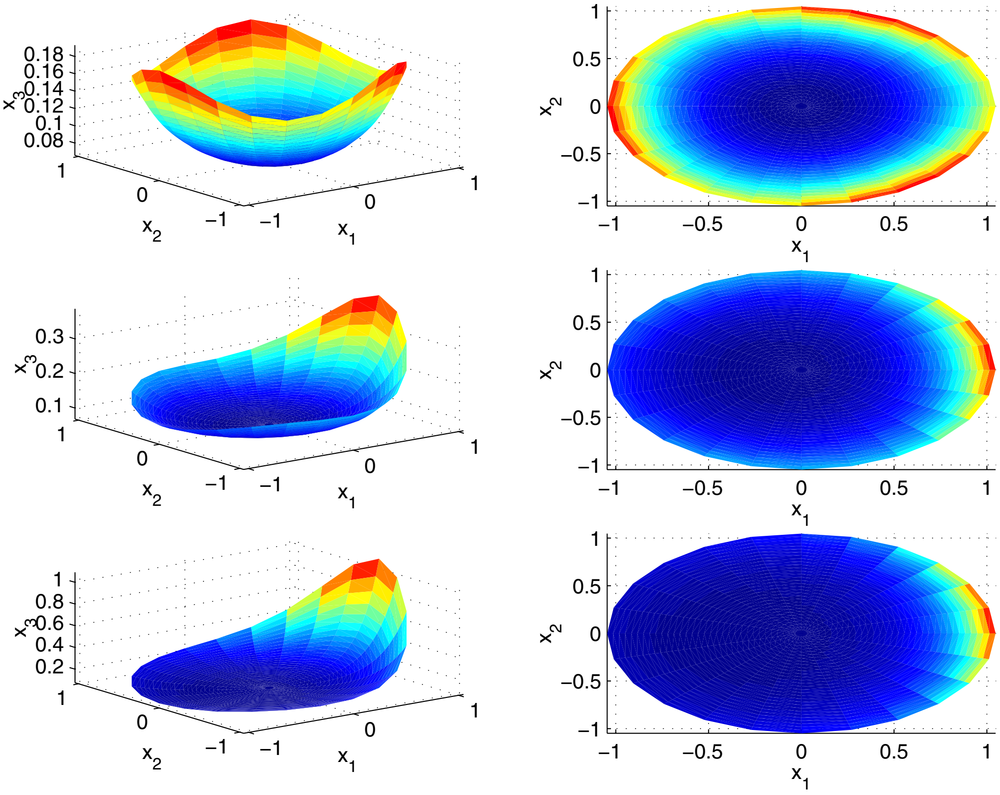


FIG. 6. *Diffraction coefficients for a trihedral cone.*

singular directions is smallest when $\phi = 0$. This explains the faster growth, as θ increases, of $|f(\boldsymbol{\omega}, \boldsymbol{\omega}_0)|$ along $\phi = 0$ (i.e., along the line $x_2 = 0$)—see the second and third rows of Figure 6.

The numerical method used for these computations was the piecewise constant collocation method with $n = 48$ subintervals on a uniform mesh (cf. Theorem 4.5). To produce each picture in Figure 6 the density in the integral equation (3.28) was approximated for 80 values of ν. Then using these densities we computed the solution to the boundary-value problem (2.5), (2.6) for the same 80 values of ν at ~ 800 observation points $\boldsymbol{\omega}$. Therefore $\sim$64,000 evaluations of the double layer potential were required. The diffraction coefficient f was computed from formula (1.4) by truncation to a finite domain of integration with respect to $\nu = i\tau$ and then applying the trapezoidal rule. The truncation points are chosen according to an analysis of the asymptotics of the integrand in (1.4) for large $|\tau|$ and are designed to yield an overall method which converges at the same rate as the method for computing g^r (see [13]). Clearly this is a very computationally intensive problem and the efficiency of our algorithm is of prime importance.

Acknowledgement. We thank Simon Chandler-Wilde for his helpful comments on this work.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1965.

[2] Y. A. Antipov, *Diffraction of a plane wave by a circular cone with an impedance boundary condition*, SIAM J. Appl. Math., 62 (2002), pp. 1122–1152.

[3] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[4] V. M. Babich, D. B. Dement'ev, and B. A. Samokish, *On the diffraction of high-frequency waves by a cone of arbitrary shape*, Wave Motion, 21 (1995), pp. 203–207.

[5] V. M. Babich, V. P. Smyshlyaev, D. B. Dement'ev, and B. A. Samokish, *Numerical calculation of the diffraction coefficients for an arbitrarily shaped perfectly conducting cone.* IEEE Trans. Antennas & Propagation, 44 (1996), pp. 740–747.

[6] V. M. Babich, D. B. Dement'ev, B. A. Samokish, and V. P. Smyshlyaev, *On evaluation of the diffraction coefficients for arbitrary "nonsingular" directions of a smooth convex cone*, SIAM J. Appl. Math., 60 (2000), pp. 536–573.

[7] V. M. Babič and V. S. Buldryev, *Short-Wavelength Diffraction Theory*, Springer-Verlag, Berlin, 1991.

[8] V. M. Babich and V. V. Kamotski, *Computation of the scattering amplitude of a wave diffracted by the vertex of a cone of arbitrary shape*, J. Math. Sci., 108 (2002), pp. 635–641.

[9] J. M. L. Bernard, *Méthode analytique et transformées functionnelles pour la diffraction d'ondes par une singularité conique*, Rapport CEA-R-5764 Editions Dist/Saslay, 1997.

[10] J. M. L. Bernard and M. A. Lyalinov, *Electromagnetic scattering by a smooth convex impedance cone*, IMA J. Appl. Math., 69 (2004), pp. 285–333.

[11] V. A. Borovikov, *Diffraction by Polygons and Polyhedrons*, Nauka, Moscow, 1966.

[12] J. J. Bowman, T. B. A. Senior, and P. L. E. Uslenghi, *Electromagnetic and Acoustic Scattering by Simple Shapes*, North-Holland, Amsterdam, 1969; revised 1987.

[13] B. D. Bonner, *Calculating Conical Diffraction Coefficients*, Ph.D. thesis, University of Bath, 2003; available online from http://www.maths.bath.ac.uk/∼igg/Bonner_Thesis.pdf.

[14] G. A. Chandler and I. G. Graham, *Product integration-collocation methods for noncompact integral operator equations*, Math. Comp., 50 (1988), pp. 125–138.

[15] J. Cheeger and I. M. Taylor, *On the diffraction of waves by conical singularities.* I, Comm. Pure Appl. Math., 35 (1982), pp. 275–331.

[16] J. Cheeger and I. M. Taylor, *On the diffraction of waves by conical singularities.* II, Comm. Pure Appl. Math., 35 (1982), pp. 487–529.

[17] M. Costabel and E. P. Stephan, *Boundary integral equations for mixed boundary value problems in polygonal domains and Galerkin approximation*, in Mathematical Models and Methods in Mechanics, Banach Center Publ. 15, Banach Centre Publications, PWN, Warsaw, 1985, pp. 175–251.

[18] D. B. Dement'ev, B. A. Samokish, and V. M. Babich, *Computation of the Legendre Function*, unpublished manuscript, 2000.

[19] R. Duduchava, *Boundary Value Problems on a Smooth Surface with Smooth Boundary*, Preprint 2002-5, Universität Stuttgart, 2002, pp. 1–19.

[20] J. Elschner, *On spline collocation for convolution equations*, Integral Equations Operator Theory, 12 (1989), pp. 486–510.

[21] J. Elschner, *The h-p-version of spline approximation methods for Mellin convolution equations*, J. Integral Equations Appl., 5 (1993), pp. 47–73.

[22] J. Elschner and I. G. Graham, *Numerical methods for integral equations of Mellin type*, J. Comp. Appl. Math., 125 (2000), pp. 423–437.

[23] I. G. Graham and G. A. Chandler, *High-order methods for linear functionals of solutions of second kind integral equations*, SIAM J. Numer. Anal., 25 (1988), pp. 1118–1137.

[24] E. W. Hobson, *The Theory of Spherical and Ellipsoidal Harmonics*, Cambridge University Press, Cambridge, UK, 1931.

[25] V. V. Kamotski, *Calculation of some integrals describing wave fields*, J. Math. Sci., 108 (5) (2002), pp. 665–673.

[26] L. V. Kantorovich and G. P. Akilov, *Functional Analysis*, Pergamon, Oxford, UK, 1982.

[27] J. B. Keller, *Diffraction by a convex cylinder*, IRE Trans. Ant. Prop., 4 (1956), pp. 312–321.

[28] J. B. Keller, *The geometrical theory of diffraction*, J. Opt. Soc. Amer., 52 (1962), pp. 116–130.

[29] R. Kress, *A Nyström method for boundary integral equations in domains with corners*, Numer. Math., 58 (1990), pp. 145–161.

[30] S. Langdon and I. G. Graham, *Boundary integral methods for singularly perturbed boundary value problems*, IMA J. Numer. Anal., 21 (2001), pp. 217–237.

[31] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.

[32] H. Schmitz, K. Volk, and W. L. Wendland, *Three-dimensional singularities of elastic fields near vertices*, Numer. Methods Partial Differential Equations, 9 (1993), pp. 323–337.

[33] V. P. Smyshlyaev, *Diffraction by conical surfaces at high frequencies*, Wave Motion, 12 (1990), pp. 329–339.

[34] V. P. Smyshlyaev, *The high-frequency diffraction of electromagnetic waves by cones of arbitrary cross sections*, SIAM J. Appl. Math., 53 (1993), pp. 670–688.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1231–1255

WELL-POSED BOUNDARY CONDITIONS FOR THE NAVIER–STOKES EQUATIONS*

JAN NORDSTRÖM[†] AND MAGNUS SVÄRD[‡]

Abstract. In this article we propose a general procedure that allows us to determine both the number and type of boundary conditions for time dependent partial differential equations. With those, well-posedness can be proven for a general initial-boundary value problem. The procedure is exemplified on the linearized Navier–Stokes equations in two and three space dimensions on a general domain.

Key words. well-posed problems, boundary conditions, Navier–Stokes equations, energy estimates, initial boundary value problems, stability

AMS subject classifications. 35M10, 65M99, 76N99

DOI. 10.1137/040604972

1. Introduction.

The problem of well-posed boundary conditions is an essential question in many areas of physics. In fluid dynamics, characteristic boundary conditions for the Euler equations have long been accepted as one way to impose boundary conditions since the specification of the ingoing variable at a boundary implies well-posedness. Often the Euler boundary conditions are used as a guidance when boundary conditions are chosen for the Navier–Stokes equations as well (see [1, 2, 3, 4, 5]). In [6] characteristic boundary conditions for the one-dimensional linearized Navier–Stokes equations were derived.

For the two- and three-dimensional Navier–Stokes equations, the number of boundary conditions implying well-posedness can be obtained using the Laplace transform technique. (See [7] for an introduction of the Laplace transform technique.) Although possible to use, the Laplace transform technique is usually a very complicated procedure for systems of partial differential equations such as the Navier–Stokes equations. However, the exact form of the boundary conditions that lead to a well-posed problem is still an open question and will be the issue addressed in this article.

In this paper we assume that we have unlimited access to accurate boundary data. We do not engage in the elaborate, difficult, and stimulating procedure of deriving artificial (or radiation or absorbing) boundary conditions. Examples of extensive research on these matters are given in [8, 9].

We propose a self-contained procedure to obtain both the number and type of boundary conditions for a general time dependent partial differential equation. The procedure is based on the energy method and has substantial similarities to the derivation of characteristic boundary conditions, since it involves a splitting of the boundary terms into ingoing and outgoing parts by a diagonalization. Compared to the Laplace transform technique, our procedure yields a much simpler analysis.

*Received by the editors March 10, 2004; accepted for publication (in revised form) December 14, 2004; published electronically September 23, 2005.

http://www.siam.org/journals/sinum/43-3/60497.html

[†]Department of Computational Physics, Division of Systems Technology, The Swedish Defence Research Agency, SE-164 90 Stockholm, Sweden (Jan.Nordstrom@foi.se), and Department of Information Technology, Uppsala University, SE-751 05, Uppsala, Sweden.

[‡]Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden (Magnus.Svard@it.uu.se).

As was already mentioned, boundary conditions for the Navier–Stokes equations have been the subject of many investigations, and still there is no theory for the general case. Hence, the linearized and symmetrized Navier–Stokes equations derived in [10] will serve as an example to which our proposed procedure is applied. Since the procedure involves a significant amount of work, we will not treat other equations in this article.

Well-posedness of the continuous problem is a necessary requirement for all numerical methods. Even for well-posed boundary conditions, numerous difficulties arise, and virtually all numerical schemes have their own way of handling boundary conditions. Hence, we will refrain from numerical calculations for a particular discretization technique and focus on the mathematical groundwork.

The contents of this article are divided as follows. In section 2 a general procedure for determining well-posed boundary conditions is presented. Section 3 applies the procedure to the three-dimensional Navier–Stokes equations on a general domain. In section 4 conclusions are drawn.

2. Well-posed boundary conditions.

Throughout this paper, the analysis will deal with linear constant coefficient equations. Frequently, the equations of interest are not linear constant coefficient equations but rather variable coefficient or even nonlinear equations (such as the Navier–Stokes equations). We will start with a brief discussion on the relevance of analyzing the constant coefficient case.

Consider a nonlinear initial-boundary value problem on a domain D with boundary ∂D. By linearizing around a solution u and freezing the coefficients, we obtain

$$\begin{aligned} \tilde{w}_t &= P(u)\tilde{w} + \delta F(x,t), \quad x \in D,\, t \geq 0, \\ \tilde{w} &= \delta f(x), \quad x \in D,\, t = 0, \\ L\tilde{w} &= \delta g(t), \quad x \in \partial D, t \geq 0, \end{aligned} \tag{1}$$

where P is the (nonlinear) differential operator and L a boundary operator. Here $\delta F, \delta f$, and δg are perturbations of the forcing, initial, and boundary functions. $\tilde{w}$ is the perturbation from the exact solution.

DEFINITION 2.1. *The linear problem* (1) *is well posed if there exists a unique solution bounded by the data $\delta F, \delta f$, and δg.*

Remark 1. There are many definitions of well-posedness. Our choice is sometimes referred to as strongly well-posed since it involves all types of data (see, for example, [7]).

Both existence and uniqueness are strongly coupled to the boundedness of the solution. In fact, it suffices to prove that a solution is bounded using a minimal number of boundary conditions; then both existence and uniqueness follow. (See, for example, [11].)

In short, the following principle holds: If (1) is well posed for all values of u, then the original nonlinear problem is well posed (see [12] for more details).

Before considering well-posedness of a problem of the type (1), we will briefly state some additional mathematical theory that is the basis of the forthcoming analysis. First we give a definition from [13].

DEFINITION 2.2. *Let A be a Hermitian matrix. The* inertia *of A is the ordered triple*

$$i(A) = (i_+(A), i_-(A), i_0(A)), \tag{2}$$

where $i_+(A)$ is the number of positive eigenvalues of A, $i_-(A)$ is the number of negative eigenvalues of A, and $i_0(A)$ is the number of zero eigenvalues of A, counting multiplicities.

We will also need the following theorem from [13], and we refer to that textbook for the proof. The theorem is also known as *Sylvester's law of inertia*.

THEOREM 2.3. *Let A, B be Hermitian matrices. There is a nonsingular matrix S such that $A = SBS^*$ if and only if A and B have the same inertia.*

S^* denotes the Hermitian adjoint of S. The following corollary is merely a rephrasing of Theorem 2.3.

COROLLARY 2.4. *Suppose that R is a nonsingular matrix and that A is a real symmetric matrix. Then the number of positive/negative eigenvalues of R^TAR is the same as the number of positive/negative eigenvalues of A.*

Proof. The claim follows immediately from Theorem 2.3 with $B = R^TAR$. □

Finally, we state another definition from [13].

DEFINITION 2.5. *If A is a real m-by-n matrix, we set $I(A) = [\mu_{ij}]$, where $\mu_{ij} = 1$ if $a_{ij} \neq 0$ and $\mu_{ij} = 0$ if $a_{ij} = 0$. The matrix $I(A)$ is called the* indicator matrix *of A.*

Now we turn to the main theory of this article. We will give general principles of how to determine boundary conditions such that the constant coefficient problem is well posed. Thus, assuming that linearization and freezing of coefficients have already been carried out, we consider a linear constant coefficient problem with n space dimensions and $\bar{x} = (x_1, \ldots, x_n)$,

$$\begin{aligned} \tilde{u}_t + \sum_{i=1}^{n} A_i \tilde{u}_{x_i} &= \sum_{i=1}^{n}\sum_{j=1}^{n} B_{ij}\tilde{u}_{x_i x_j} + F(\bar{x}, t), \quad \bar{x} \in D,\, t \geq 0, \\ \tilde{u}(\bar{x}, 0) &= f(\bar{x}), \quad \bar{x} \in D, \\ L\tilde{u}(\bar{x}, t) &= g(t), \quad \bar{x} \in \partial D,\, t \geq 0. \end{aligned} \tag{3}$$

The definition (3) of an initial-boundary value problem covers hyperbolic, parabolic, and incompletely parabolic partial differential equations depending on the rank of the matrices. Let $\|\cdot\|$ denote some norm for functions on D. Our approach of analyzing the well-posedness of (3) comprises the following steps.

(i) Symmetrize (3).

(ii) Apply the energy method. The energy estimate will have the structure

$$\|\tilde{u}\|_t^2 + c_i \sum_{i=1}^{n} \|\tilde{u}_{x_i}\|^2 + \oint_{\partial D} \tilde{v}^T \mathbf{A} \tilde{v} ds \leq 0, \tag{4}$$

where $c_i \geq 0$, $i = 1, \ldots, n$, are constants and $\tilde{v}$ a vector formed by combinations of $\tilde{u}$ and $\tilde{u}_{x_i}$. Further, $\mathbf{A}$ is reduced to a full rank matrix. The boundedness of $\tilde{u}$ now depends on the boundedness of $\tilde{v}^T \mathbf{A} \tilde{v}$ in boundary data.

(iii) Find a diagonalizing matrix, M, such that $M^T \mathbf{A} M = \Lambda$ is diagonal. ($\mathbf{A}$ is symmetric due to step (i) above.) Then we also have the variable transformation $M^{-1}\tilde{v} = \tilde{w}$.

(iv) Split $\Lambda = \Lambda^+ + \Lambda^-$ such that Λ^+ is positive semidefinite and Λ^- is negative semidefinite. Also, split $\tilde{w}$ into $\tilde{w} = \tilde{w}^+ + \tilde{w}^-$ corresponding to the nonzero entries of $\Lambda^{+,-}$. More precisely, let $\tilde{w}^- = I(\Lambda^-)\tilde{w}$ and $\tilde{w}^+ = \tilde{w} - \tilde{w}^-$.

(v) Supply boundary data to the negative part. That is, specify $\tilde{w}^-$ by g.

Remark 2. In step (iv) the number of boundary conditions is given as the number of negative eigenvalues of $\mathbf{A}$ or Λ. Further, the type of boundary conditions is given by the matrix M, derived in step (iii).

This implies boundedness of $\|\tilde{u}\|_t$ and hence of $\|\tilde{u}\|$. The difficult part of this scheme is step (iii). However, we know that $\mathbf{A}$ is symmetric, and we can prove the following proposition regarding steps (iii)–(v).

Proposition 2.6. *Assume that steps* (i) *and* (ii) *are fulfilled; then the matrix* $\mathbf{A}$ *and the vector* $\tilde{v}$ *can be split such that* $\tilde{v}^T\mathbf{A}\tilde{v} = \tilde{w}^{+T}\Lambda^+\tilde{w}^+ + \tilde{w}^{-T}\Lambda^-\tilde{w}^-$, *where* Λ^+ *is positive semidefinite,* Λ^- *is negative semidefinite, and* $M^{-1}\tilde{v} = \tilde{w} = \tilde{w}^+ + \tilde{w}^-$ *for some matrix* M^{-1}. *Further, by specifying* $\tilde{w}^- = I(\Lambda^-)w$ *at the boundary, we find that* (3) *is well posed.*

Proof. Since $\mathbf{A}$ is symmetric, the eigenvalues are real and there exists a full set of eigenvectors. If Z contains the eigenvectors, we have

$$\tilde{v}^T\mathbf{A}\tilde{v} = \tilde{v}^T ZZ^T\mathbf{A}ZZ^T\tilde{v} = \tilde{w}^T\Lambda_{\mathbf{Z}}\tilde{w} = \tilde{w}^{+T}\Lambda^+_{\mathbf{Z}}\tilde{w}^+ + \tilde{w}^{-T}\Lambda^-_{\mathbf{Z}}\tilde{w}^-, \tag{5}$$

where $\Lambda_{\mathbf{Z}}^{-/+}$ are diagonal negative/positive semidefinite. We define $\tilde{w}^- = I(\Lambda^-)\tilde{w}$ and $\tilde{w}^+ = \tilde{w} - \tilde{w}^-$. This proves the first part of Proposition 2.6.

Another way to prove the first part of Proposition 2.6 is to apply Corollary 2.4, to conclude that any nonsingular matrix R can be used as a transformation, $\mathbf{B} = R^T\mathbf{A}R$, such that $\mathbf{A}$ and $\mathbf{B}$ have the same inertia. By construction, $\mathbf{B}$ is symmetric. Then $\mathbf{B}$ may be diagonalized by its eigenvectors, and we have another diagonalization of $\mathbf{A}$. Denote by X the matrix containing the eigenvectors of $\mathbf{B}$ as columns such that

$$\begin{aligned}\tilde{v}^T\mathbf{A}\tilde{v} &= \tilde{v}^T R^{-1,T}R^T\mathbf{A}RR^{-1}\tilde{v} = \tilde{v}^T R^{-1,T}\mathbf{B}R^{-1}\tilde{v}\\ &= \tilde{v}^T R^{-1,T}X\Lambda_{\mathbf{M}}X^T R^{-1}\tilde{v} = \tilde{w}^T\Lambda^+_{\mathbf{M}}\tilde{w} + \tilde{w}^T\Lambda^-_{\mathbf{M}}\tilde{w}\end{aligned}$$

or

$$\tilde{v}^T M^{-1,T}M^T\mathbf{A}MM^{-1}\tilde{v} = \tilde{w}^T\Lambda_{\mathbf{M}}\tilde{w} = \tilde{w}^{+T}\Lambda^+_{\mathbf{M}}\tilde{w}^+ + \tilde{w}^{-T}\Lambda^-_{\mathbf{M}}\tilde{w}^-, \tag{6}$$

where $\tilde{w} = M^{-1}\tilde{v}$, $M = RX$, and $\Lambda_{\mathbf{M}}^{-/+}$ are diagonal negative/positive semidefinite. Further, $\tilde{w}^- = I(\Lambda^-_{\mathbf{M}})\tilde{w}$ and $\tilde{w}^+ = \tilde{w} - \tilde{w}^-$. We conclude that there are several different ways of diagonalizing $\mathbf{A}$, but in all cases $\Lambda_{\mathbf{Z}}$ and $\Lambda_{\mathbf{M}}$ have the same inertia. The fundamental difference between Z and another diagonalizing matrix, M, is that M is not orthogonal. We may regard Z as a specific M.

Next, we turn to the proof of the second part of the proposition. Specify $\tilde{w}^- = g$ at the boundary. Equation (4) can be rewritten as

$$\|\tilde{u}\|^2_t + \oint_{\partial D}\tilde{w}^{+T}\Lambda^+_{\mathbf{M}}\tilde{w}^+ ds + c_i\sum_{i=1}^{n}\|\tilde{u}_{x_i}\|^2 = -\oint_{\partial D} g^T\Lambda^-_{\mathbf{M}}g\,ds. \tag{7}$$

All the terms on the left-hand side of (7) are positive, implying that $\|\tilde{u}\|_t$, and hence $\|\tilde{u}\|$, are bounded. □

Remark 3. The assumption that steps (i) and (ii) in Proposition 2.6 can be fulfilled is true for many important partial differential equations. For example, it is true for the Euler, Navier–Stokes, and Maxwell equations.

Remark 4. The procedure that diagonalizes $\mathbf{A}$, with its eigenvectors and bounds the negative part, is what we mean by characteristic boundary conditions.

For Proposition 2.6 to be practically useful, a crucial point is to find a diagonalizing matrix. That is why we gave two examples of diagonalizing matrices. In the first example we used the eigenvalues and eigenvectors directly. For a system of equations, the matrix A can be large (9-by-9 for the Navier–Stokes equations in three dimensions). The eigenvalues of $\mathbf{A}$ are given as the roots of a polynomial of high degree, for which in general there do not exist roots in closed form.

In the second example, we can proceed in a different way. We will seek a diagonalizing matrix to $\mathbf{A}$ that is not orthogonal. By choosing R such that $\mathbf{B}$ has a simpler structure than $\mathbf{A}$, we may be able to find the eigenvalues and eigenvectors to $\mathbf{B}$. In fact, we will show that this is possible for the three-dimensional Navier–Stokes equations on general domains.

Certainly, not all of the points are novel in the above procedure. For example, in [10] a symmetrization of the linearized Navier–Stokes equations is presented. For the Euler equations, the whole procedure has been carried out when deriving the well-known characteristic boundary conditions. However, the idea of diagonalizing the boundary terms with a nonorthogonal matrix is, to the knowledge of the present authors, new. Furthermore, it is important to formalize the whole procedure since it should be possible to find well-posed boundary conditions to any problem of type (3).

3. The Navier–Stokes equations.

3.1. Step (i): Symmetrize the equations.

We will consider the Navier–Stokes equations as an example of how to use the procedure presented above to derive well-posed boundary conditions. We begin by rescaling the three-dimensional Navier–Stokes equations to nondimensional form. Consider the Navier–Stokes equations in primitive variables $\tilde{V} = [\tilde{\rho}, \tilde{u}_1, \tilde{u}_2, \tilde{u}_3, \tilde{p}]$ as stated in [10],

$$\begin{aligned} &\tilde{V}_t + \tilde{A}_1^p \tilde{V}_x + \tilde{A}_2^p \tilde{V}_y + \tilde{A}_3^p \tilde{V}_z \\ &= \tilde{B}_{11}^p \tilde{V}_{xx} + \tilde{B}_{22}^p \tilde{V}_{yy} + \tilde{B}_{33}^p \tilde{V}_{zz} + \tilde{B}_{xy}^p \tilde{V}_{xy} + \tilde{B}_{yz}^p \tilde{V}_{yz} + \tilde{B}_{zx}^p \tilde{V}_{zx}, \end{aligned} \tag{8}$$

where the tilde sign emphasizes that the entity depends on the solution. Further, $\tilde{\rho}$ is the density; $\tilde{u}_1, \tilde{u}_2, \tilde{u}_3$ are the velocities in the x, y, and z directions, respectively; and $\tilde{p}$ is the pressure. We will also use the ratio between the specific heat capacities, $\gamma = c_p/c_v$, and the speed of sound, c; μ the dynamic viscosity, λ the bulk viscosity, and $\nu = \frac{\mu}{\rho}$ the kinematic viscosity; $Pr = \frac{\nu}{\alpha}$ denoting the Prandtl number, where α is the thermal diffusivity. Let $Re = \frac{\rho_\infty U_\infty L}{\mu_\infty}$ denote the Reynolds number. The infinity subscript denotes free stream conditions, and L is some characteristic length scale.

The equations (8) are nondimensionalized and the coefficients are frozen, which corresponds to the linearization of the Navier–Stokes equations. The tilde signs are dropped on the matrices as they no longer depend on the solution. Using the parabolic symmetrizer S_p derived in [10] and letting $\epsilon = \frac{1}{Re}$ yields

$$\begin{aligned} &\tilde{u}_t + A_1 \tilde{u}_x + A_2 \tilde{u}_y + A_3 \tilde{u}_z \\ &= \epsilon (B_{11} \tilde{u}_{xx} + B_{22} \tilde{u}_{yy} + B_{33} \tilde{u}_{zz} + B_{xy} \tilde{u}_{xy} + B_{yz} \tilde{u}_{uz} + B_{zx} \tilde{u}_{zx}). \end{aligned} \tag{9}$$

The transformed nondimensionalized variables are

$$\begin{aligned} S_p^{-1} \tilde{V} &= \begin{pmatrix} \frac{c}{\sqrt{\gamma}\rho} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ -\frac{c}{\rho\sqrt{\gamma}\sqrt{\gamma-1}} & 0 & 0 & 0 & \sqrt{\frac{\gamma}{\gamma-1}}\frac{1}{\rho c} \end{pmatrix} \begin{pmatrix} \tilde{\rho} \\ \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \\ \tilde{p} \end{pmatrix} \\ &= \begin{pmatrix} \frac{c}{\sqrt{\gamma}\rho}\tilde{\rho} \\ \tilde{u}_1 \\ \tilde{u}_2 \\ \tilde{u}_3 \\ -\frac{c}{\sqrt{\gamma}\sqrt{\gamma-1}}\frac{\tilde{\rho}}{\rho} + \sqrt{\frac{\gamma}{\gamma-1}}\frac{1}{\rho c}\tilde{p} \end{pmatrix} = \tilde{u}. \end{aligned} \tag{10}$$

The symmetrized matrices are derived in [10] and are repeated here for convenience. Let $a = \sqrt{\frac{\gamma-1}{\gamma}}c$ and $b = \frac{c}{\sqrt{\gamma}}$. Then

$$A_1 = \begin{pmatrix} u_1 & b & 0 & 0 & 0 \\ b & u_1 & 0 & 0 & a \\ 0 & 0 & u_1 & 0 & 0 \\ 0 & 0 & 0 & u_1 & 0 \\ 0 & a & 0 & 0 & u_1 \end{pmatrix}, \qquad A_2 = \begin{pmatrix} u_2 & 0 & b & 0 & 0 \\ 0 & u_2 & 0 & 0 & 0 \\ b & 0 & u_2 & 0 & a \\ 0 & 0 & 0 & u_2 & 0 \\ 0 & 0 & a & 0 & u_2 \end{pmatrix}, \tag{11}$$

$$A_3 = \begin{pmatrix} u_3 & 0 & 0 & b & 0 \\ 0 & u_3 & 0 & 0 & 0 \\ 0 & 0 & u_3 & 0 & 0 \\ b & 0 & 0 & u_3 & a \\ 0 & 0 & 0 & a & u_3 \end{pmatrix}, \qquad B_{xy} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\lambda+\mu}{\rho} & 0 & 0 \\ 0 & \frac{\lambda+\mu}{\rho} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{12}$$

$$B_{yz} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\lambda+\mu}{\rho} & 0 \\ 0 & 0 & \frac{\lambda+\mu}{\rho} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \qquad B_{zx} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \frac{\lambda+\mu}{\rho} & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{\lambda+\mu}{\rho} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{13}$$

$$B_{11} = \operatorname{diag}\left(0, \frac{\lambda+2\mu}{\rho}, \frac{\mu}{\rho}, \frac{\mu}{\rho}, \frac{\gamma\mu}{Pr\rho}\right), \tag{14}$$

$$B_{22} = \operatorname{diag}\left(0, \frac{\mu}{\rho}, \frac{\lambda+2\mu}{\rho}, \frac{\mu}{\rho}, \frac{\gamma\mu}{Pr\rho}\right), \tag{15}$$

$$B_{33} = \operatorname{diag}\left(0, \frac{\mu}{\rho}, \frac{\mu}{\rho}, \frac{\lambda+2\mu}{\rho}, \frac{\gamma\mu}{Pr\rho}\right). \tag{16}$$

3.2. Step (ii): Apply the energy method. Next, we turn to the analysis of boundary conditions for the Navier–Stokes equations. Consider a general domain D with boundary ∂D in three space dimensions. From (9), the symmetrized and nondimensionalized Navier–Stokes equations are

$$\tilde{u}_t + (A_1\tilde{u} - \epsilon\tilde{F}_v)_x + (A_2\tilde{u} - \epsilon\tilde{G}_v)_y + (A_3\tilde{u} - \epsilon\tilde{H}_v)_z, \tag{17}$$

where

$$\tilde{F}_v = B_{11}\tilde{u}_x + B_{21}\tilde{u}_y + B_{31}\tilde{u}_z, \tag{18}$$

$$\tilde{G}_v = B_{22}\tilde{u}_y + B_{32}\tilde{u}_z + B_{12}\tilde{u}_x, \tag{19}$$

$$\tilde{H}_v = B_{33}\tilde{u}_z + B_{23}\tilde{u}_y + B_{13}\tilde{u}_x, \tag{20}$$

and

$$B_{21} = B_{12} = \frac{B_{xy}}{2}, \quad B_{32} = B_{23} = \frac{B_{yz}}{2}, \quad B_{31} = B_{13} = \frac{B_{zx}}{2}.$$

Applying the energy method (step (ii)),

$$\begin{aligned}\int_D \tilde{u}^T\tilde{u}_t dxdydz + \int_D \frac{\partial}{\partial x}\left(\frac{1}{2}\tilde{u}^T A_1\tilde{u} - \epsilon\tilde{u}^T\tilde{F}_v\right)&\\ + \frac{\partial}{\partial y}\left(\frac{1}{2}\tilde{u}^T A_2\tilde{u} - \epsilon\tilde{u}^T\tilde{G}_v\right) + \frac{\partial}{\partial z}\left(\frac{1}{2}\tilde{u}^T A_3\tilde{u} - \epsilon\tilde{u}^T\tilde{H}_v\right) dxdydz&\\ = -\epsilon\int_D (\tilde{u}_x^T\tilde{F}_v + \tilde{u}_y^T\tilde{G}_v + \tilde{u}_z^T\tilde{H}_v)dxdydz.&\end{aligned} \tag{21}$$

The right-hand side in (21) is negative definite and denoted by $-DI$.

Remark 5. It is easily verified that the last term in (21) is dissipation,

$$DI = \epsilon\int_D \left(\begin{array}{ccc} \tilde{u}_x^T & \tilde{u}_y^T & \tilde{u}_z^T \end{array}\right)\left(\begin{array}{ccc} B_{11} & B_{12} & B_{13}\\ B_{21} & B_{22} & B_{23}\\ B_{31} & B_{32} & B_{33}\end{array}\right)\left(\begin{array}{c}\tilde{u}_x\\ \tilde{u}_y\\ \tilde{u}_z\end{array}\right) dxdydz.$$

The matrix is symmetric with positive or zero diagonal entries. With $\lambda \leq \mu$, the matrix is diagonally dominant. Thus, it is positive semidefinite.

Denote by $\|\tilde{u}\|^2$ the integral $\int_D \tilde{u}^T\tilde{u}dxdydz$. Using Gauss' theorem, we obtain

$$\begin{aligned}\|\tilde{u}\|_t^2 + \oint_{\partial D}\left(\tilde{u}^T(A_1\tilde{u} - 2\epsilon\tilde{F}_v), \tilde{u}^T(A_2\tilde{u} - 2\epsilon\tilde{G}_v), \tilde{u}^T(A_3\tilde{u} - 2\epsilon\tilde{H}_v)\right)\cdot\hat{\mathbf{n}}\, ds&\\ = -2DI,&\end{aligned} \tag{22}$$

where $\hat{\mathbf{n}} = (n_1, n_2, n_3)$ is the outward-pointing unit normal on the surface ∂D and $ds = \sqrt{dx^2 + dy^2 + dz^2}$. Equation (22) can be rewritten as

$$\begin{aligned}\|\tilde{u}\|_t^2 + \oint_{\partial D}\left(\begin{array}{c}\tilde{u}\\ \tilde{F}^V\end{array}\right)^T\left(\begin{array}{cc} A_1n_1 + A_2n_2 + A_3n_3 & -\epsilon I_5\\ -\epsilon I_5 & 0_5\end{array}\right)\left(\begin{array}{c}\tilde{u}\\ \tilde{F}^V\end{array}\right)ds&\\ = -2DI,&\end{aligned} \tag{23}$$

where I_n denotes the n-by-n identity matrix, and similarly 0_n the n-by-n zero matrix and $\tilde{F}^V = \tilde{F}_v n_1 + \tilde{G}_v n_2 + \tilde{H}_v n_3$.

To prove well-posedness we have to split the matrix in the boundary integral into positive definite and negative definite parts. The negative part of the boundary term in (23) caused by

$$\mathbf{A_1} = \left(\begin{array}{cc} A_1n_1 + A_2n_2 + A_3n_3 & -\epsilon I_5\\ -\epsilon I_5 & 0_5\end{array}\right) \tag{24}$$

has to be supplied with boundary conditions, which in turn bounds the growth of $\|\tilde{u}\|_t^2$ in (21).

We note that the first component of $\tilde{F}^V$ is zero, and hence we can reduce the system by omitting that component and denoting the resulting vector by $\tilde{G}^V$. By this procedure $\mathbf{A_1}$ is also reduced from a 10-by-10 matrix to a 9-by-9 matrix by deleting the sixth row and column. With $\mathbf{u} = (u_1, u_2, u_3)$, we have

$$\left(\begin{array}{c}\tilde{u}\\ \tilde{F}^V\end{array}\right)^T\left(\begin{array}{cc} A_1n_1 + A_2n_2 + A_3n_3 & -\epsilon I_5\\ -\epsilon I_5 & 0_5\end{array}\right)\left(\begin{array}{c}\tilde{u}\\ \tilde{F}^V\end{array}\right) = \left(\begin{array}{c}\tilde{u}\\ \tilde{G}^V\end{array}\right)^T\mathbf{A}\left(\begin{array}{c}\tilde{u}\\ \tilde{G}^V\end{array}\right),$$

where

$$\mathbf{A} = \begin{pmatrix} \mathbf{u}\cdot\hat{\mathbf{n}} & bn_1 & bn_2 & bn_3 & 0 & 0 & 0 & 0 & 0 \\ bn_1 & \mathbf{u}\cdot\hat{\mathbf{n}} & 0 & 0 & an_1 & -\epsilon & 0 & 0 & 0 \\ bn_2 & 0 & \mathbf{u}\cdot\hat{\mathbf{n}} & 0 & an_2 & 0 & -\epsilon & 0 & 0 \\ bn_3 & 0 & 0 & \mathbf{u}\cdot\hat{\mathbf{n}} & an_3 & 0 & 0 & -\epsilon & 0 \\ 0 & an_1 & an_2 & an_3 & \mathbf{u}\cdot\hat{\mathbf{n}} & 0 & 0 & 0 & -\epsilon \\ 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} & 0_{14} \\ A_{21} & A_{22} & -\epsilon I_4 \\ 0_{41} & -\epsilon I_4 & 0_4 \end{pmatrix}, \tag{25}$$

using the notation 0_{nm} for the n-by-m zero matrix. We will also use the notation $u_n = \mathbf{u}\cdot\hat{\mathbf{n}}$. Since $\hat{\mathbf{n}}$ is the outward-pointing normal, $u_n < 0$ implies inflow. Further, note that A_{11} in (25) is a scalar.

3.3. Step (iii): Find a diagonalizing matrix.

Next, we state and prove the following proposition, where $M_n = u_n/c$ is the Mach number.

PROPOSITION 3.1. *If $|M_n| \neq 1, 0$ and $u_n < 0$, there are four positive and five negative eigenvalues of $\mathbf{A}$. If $|M_n| \neq 1, 0$ and $u_n > 0$, there are five positive and four negative eigenvalues of $\mathbf{A}$.*

Proposition 3.1 states that an inflow demands five and an outflow four boundary conditions. The number of boundary conditions can also be derived using the Laplace transform technique, which is shown in [14, 15]. However, to prove well-posedness of specific boundary conditions using the Laplace transform technique is algebraically very complex, as shown in [15]. In the proof of Proposition 3.1 we will continue with the procedure outlined in section 2 and find a diagonalizing matrix to A (step (iii)). However, finding the eigenvalues of A corresponds to solving a ninth degree polynomial. Besides the algebraic difficulty of finding roots to ninth degree polynomials, it is probable that the roots in this particular case do not exist in closed form. Instead, we will derive another diagonalizing matrix. That matrix gives the explicit form of the well-posed boundary conditions.

Proof of Proposition 3.1. Rotate $\mathbf{A}$ by

$$R^T\mathbf{A}R = \begin{pmatrix} 1 & 0_{14} & 0_{14} \\ \bar{\alpha}^T & I_4 & 0_4 \\ \bar{\beta}^T & \bar{\gamma}^T & I_4 \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} & 0_{14} \\ A_{21} & A_{22} & -\epsilon I_4 \\ 0_{41} & -\epsilon I_4 & 0_4 \end{pmatrix} \begin{pmatrix} 1 & \bar{\alpha} & \bar{\beta} \\ 0_{41} & I_4 & \bar{\gamma} \\ 0_{41} & 0_4 & I_4 \end{pmatrix} = \begin{pmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{pmatrix} = \mathbf{E}, \tag{26}$$

where

$$\begin{aligned} E_{11} &= A_{11}, \\ E_{12} &= A_{11}\bar{\alpha} + A_{12}, \\ E_{13} &= A_{11}\bar{\beta} + A_{12}\bar{\gamma}, \\ E_{21} &= \bar{\alpha}^T A_{11} + A_{21}, \\ E_{22} &= \bar{\alpha}^T(A_{11}\bar{\alpha} + A_{12}) + (A_{21}\bar{\alpha} + A_{22}), \end{aligned}$$

$$E_{23} = \bar{\alpha}^T(A_{11}\bar{\beta} + A_{12}\bar{\gamma}) + A_{21}\bar{\beta} + A_{22}\bar{\gamma} - \epsilon I_3,$$
$$E_{31} = \bar{\beta}^T A_{11} + \bar{\gamma}^T A_{21},$$
$$E_{32} = \bar{\beta}^T(A_{11}\bar{\alpha} + A_{12}) + \bar{\gamma}^T(A_{21}\bar{\alpha} + A_{22}) - \epsilon I_3,$$
$$E_{33} = \bar{\beta}^T(A_{11}\bar{\beta} + A_{12}\bar{\gamma}) + \bar{\gamma}^T(A_{21}\bar{\beta} + A_{22}\bar{\gamma} - \epsilon I_3) - \epsilon I_3\bar{\gamma}.$$

Using $A_{12}^T = A_{21}$, we cancel the off-diagonal blocks and solve for $\bar{\alpha}, \bar{\beta}$, and $\bar{\gamma}$. We obtain

$$\text{(27)} \qquad \bar{\alpha} = -A_{11}^{-1}A_{12}, \quad \bar{\beta} = \epsilon A_{11}^{-1}A_{12}E_{22}^{-1}, \quad \bar{\gamma} = -\epsilon E_{22}^{-1},$$

$$\text{(28)} \qquad \mathbf{E} = \begin{pmatrix} A_{11} & 0_{14} & 0_{14} \\ 0_{41} & E_{22} & 0_4 \\ 0_{41} & 0_4 & -\epsilon^2 E_{22}^{-1} \end{pmatrix}, \quad E_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}.$$

The conditions for this procedure to hold are that $\det(A_{11}) \neq 0$ and $\det(E_{22}) \neq 0$.

We know from Corollary 2.4 that $i(\mathbf{A}) = i(\mathbf{E})$. Thus, we can instead determine the sign of the eigenvalues of $\mathbf{E}$. Note that the upper-left entry of $\mathbf{E}$ is a scalar and hence an eigenvalue. We denote that by

$$\text{(29)} \qquad \lambda_1 = A_{11} = u_n.$$

If $\det(E_{22}) \neq 0$, we know that E_{22} has four real nonzero eigenvalues, since E_{22} is symmetric by construction. The signs of those do not change as E_{22} is inverted such that from the second and third block there are always four negative and four positive eigenvalues of $\mathbf{E}$. Including λ_1, we have for $u_n > 0$ four negative and five positive eigenvalues, and for $u_n < 0$, five negative and four positive eigenvalues of $\mathbf{E}$, as stated in the proposition (assuming that $\det(A_{11}) \neq 0$ and $\det(E_{22}) \neq 0$).

We will now show that $\det(A_{11}) \neq 0$ and $\det(E_{22}) \neq 0$ for $M_n \neq \pm 1, 0$. Since $A_{11} = u_n$, we have $\det(A_{11}) \neq 0$ for $M_n \neq 0$. To evaluate the second condition, we compute the eigenvalues of E_{22} explicitly. From (25) and (28) we have

$$\text{(30)} \qquad E_{22} = \begin{pmatrix} -\frac{b^2n_1^2}{u_n} + u_n & -\frac{b^2n_1n_2}{u_n} & -\frac{b^2n_1n_3}{u_n} & an_1 \\ -\frac{b^2n_1n_2}{u_n} & -\frac{b^2n_2^2}{u_n} + u_n & -\frac{b^2n_2n_3}{u_n} & an_2 \\ -\frac{b^2n_1n_3}{u_n} & -\frac{b^2n_2n_3}{u_n} & -\frac{b^2n_3^2}{u_n} + u_n & an_3 \\ an_1 & an_2 & an_3 & u_n \end{pmatrix},$$

and the eigenvalues are

$$\text{(31)} \qquad \lambda_{2,3} = \frac{-b^2 + 2u_n^2 \pm \sqrt{b^4 + 4a^2u_n^2}}{2u_n},$$

$$\text{(32)} \qquad \lambda_4 = \lambda_5 = u_n,$$

where $n_1^2 + n_2^2 + n_3^2 = 1$ has been used to simplify the expressions. λ_4 and λ_5 obviously shift sign at $u_n = 0$. Also, since $\lambda_4 = \lambda_5 = 0$ with $M_n = u_n = 0$, we have that $\det(E_{22}) = 0$. Thus, to rotate $\mathbf{A}$ by R we once more need $M_n \neq 0$. λ_2 and λ_3 can be expressed as

$$\text{(33)} \qquad \lambda_{2,3} = \frac{c}{2\gamma M_n}\left(-1 + 2\gamma M_n^2 \pm \sqrt{1 + 4(\gamma - 1)\gamma M_n^2}\right).$$

Consider λ_2, and note that $\gamma \geq 1$. Then $\sqrt{1-4\gamma M_n^2+4\gamma^2 M_n^2} \geq 1$ such that the sign of λ_2 is the same as the sign of the denominator, i.e., M_n or u_n. This means that $\lambda_2 \neq 0$ for all $M_n \neq 0$, and $\lambda_2 = 0$ for $M_n = 0$.

At last, λ_3 is considered. λ_3 shifts sign when

$$2\gamma M_n^2 - 1 - \sqrt{1-4\gamma M_n^2+4\gamma^2 M_n^2} = 0.$$

Alternatively, $(2\gamma M_n^2-1)^2 = (1-4\gamma M_n^2+4\gamma^2 M_n^2)$, which has the solutions $M_n = 0, 1, -1$, but $M_n = 0$ is discarded due to the original equality. Thus, $\lambda_3 \neq 0$, and hence $\det(E_{22}) \neq 0$ for $|M_n| \neq 1$. Note that, λ_3 is singular for $M_n = 0$. □

We have now derived the number of positive and negative eigenvalues of $\mathbf{A}$, and hence the number of boundary conditions, and their dependence on M_n. This was done by calculating the eigenvalues of $\mathbf{E}$ explicitly.

To obtain a set of boundary conditions we also need the eigenvectors of $\mathbf{E}$. Given the eigenvectors of $\mathbf{E}$, it is a simple task to derive a diagonalizing matrix to $\mathbf{A}$. The eigenvectors of E_{22} are able to be explicitly derived since the eigenvalues are explicitly given and they are $Y = (y_2, y_3, y_4, y_5)$, where

$$\begin{aligned} y_2 &= \left(n_1, n_2, n_3, -\frac{-b^4-\sqrt{b^2+4a^2u_n^2}}{2au_n}\right)^T \\ &= \left(n_1, n_2, n_3, \frac{-\lambda_3+u_n}{a}\right)^T, \qquad (34) \\ y_3 &= \left(n_1, n_2, n_3, -\frac{-b^4+\sqrt{b^2+4a^2u_n^2}}{2au_n}\right)^T \\ &= \left(n_1, n_2, n_3, \frac{-\lambda_2+u_n}{a}\right)^T, \qquad (35) \\ y_4 &= (-n_2, n_1, 0, 0)^T, \qquad (36) \\ y_5 &= (-n_3, 0, n_1, 0). \qquad (37) \end{aligned}$$

Remark 6. We omit the normalization of the eigenvectors to keep the expressions (34)–(37) simple.

Now, we can derive a specific diagonalizing matrix M and conclude step (iii). For convenience, we restate (6),

$$\tilde{v}^T M^{-1,T} M^T \mathbf{A} M M^{-1} \tilde{v} = \tilde{w}^T \Lambda_{\mathbf{M}} \tilde{w},$$

where $M = RX$ and $\tilde{v} = (\tilde{u}^T\, (\tilde{G}^V)^T)^T$. R is given in (26), (27), and (28). Further,

$$X = \begin{pmatrix} 1 & 0_{14} & 0_{14} \\ 0_{41} & Y & 0_4 \\ 0_{41} & 0_4 & Y \end{pmatrix}, \qquad \Lambda_{\mathbf{M}} = \begin{pmatrix} u_n & 0_{14} & 0_{14} \\ 0_{41} & \Lambda & 0_4 \\ 0_{41} & 0_4 & -\epsilon^2\Lambda^{-1} \end{pmatrix},$$

where $\Lambda = \operatorname{diag}(\lambda_2, \lambda_3, \lambda_4, \lambda_5)$. Inverting R and M yields

$$R^{-1} = \begin{pmatrix} 1 & -\bar{\alpha} & 0_{14} \\ 0_{41} & I_4 & -\bar{\gamma} \\ 0_{41} & 0_4 & I_4 \end{pmatrix}, \qquad M^{-1} = X^T R^{-1} = \begin{pmatrix} 1 & -\bar{\alpha} & 0_{14} \\ 0_{41} & Y^T & -Y^T\bar{\gamma} \\ 0_{41} & 0_4 & Y^T \end{pmatrix}. \qquad (38)$$

To simplify the computation of M^{-1} we use (27) and obtain

$$-Y^T\gamma = \epsilon Y^T E_{22}^{-1} = \epsilon Y^T Y \Lambda^{-1} Y^T = \epsilon \Lambda^{-1} Y^T = \epsilon \begin{pmatrix} \lambda_2^{-1} y_2^T \\ \lambda_3^{-1} y_3^T \\ \lambda_4^{-1} y_4^T \\ \lambda_5^{-1} y_5^T \end{pmatrix}, \tag{39}$$

yielding

$$M^{-1} = \begin{pmatrix} 1 & -\bar{\alpha} & 0_{14} \\ 0_{41} & Y^T & \epsilon\Lambda^{-1}Y^T \\ 0_{41} & 0_4 & Y^T \end{pmatrix}, \quad \text{where } \bar{\alpha} = \left(-\frac{b}{u_n}\hat{\mathbf{n}}, 0\right). \tag{40}$$

We proceed by computing the variables, $\tilde{w} = X^T R^{-1} \tilde{v} = M^{-1}\tilde{v}$, to which boundary conditions should be applied. Let $\tilde{G}_i^V$ be the ith component of $\tilde{G}^V$. Define $\tilde{v}_{i\ldots j} = (\tilde{v}_i, \ldots, \tilde{v}_j)^T$ and $\tilde{u}_n = (\tilde{u}_1, \tilde{u}_2, \tilde{u}_3) \cdot \hat{\mathbf{n}}$. For convenience, we restate $\tilde{v}$,

$$\tilde{v} = \left(\frac{b}{\rho}\tilde{\rho}, \tilde{u}_1, \tilde{u}_2, \tilde{u}_3, -\frac{b}{\sqrt{\gamma-1}}\frac{\tilde{\rho}}{\rho} + \frac{1}{\rho a}\tilde{p}, \tilde{G}_1^V, \tilde{G}_2^V, \tilde{G}_3^V, \tilde{G}_4^V\right)^T. \tag{41}$$

Then,

$$\tilde{w} = M^{-1}\tilde{v} = \begin{pmatrix} \tilde{v}_1 - \bar{\alpha}\cdot\tilde{v}_{2\ldots5} \\ y_2^T(\tilde{v}_{2\ldots5} - \epsilon\lambda_2^{-1}\tilde{G}^V) \\ y_3^T(\tilde{v}_{2\ldots5} - \epsilon\lambda_3^{-1}\tilde{G}^V) \\ y_4^T(\tilde{v}_{2\ldots5} - \epsilon\lambda_4^{-1}\tilde{G}^V) \\ y_5^T(\tilde{v}_{2\ldots5} - \epsilon\lambda_5^{-1}\tilde{G}^V) \\ y_2^T\tilde{G}^V \\ y_3^T\tilde{G}^V \\ y_4^T\tilde{G}^V \\ y_5^T\tilde{G}^V \end{pmatrix}, \tag{42}$$

by using (34)–(37).

For completeness we also give the reverse transformation. It is $\tilde{v} = RX\tilde{w} = M\tilde{w}$,

$$M = \begin{pmatrix} 1 & \bar{\alpha}Y & \bar{\beta}Y \\ 0_{31} & Y & \bar{\gamma}Y \\ 0_{31} & 0_3 & Y \end{pmatrix} = \begin{pmatrix} 1 & \bar{\alpha}Y & \bar{\alpha}\Lambda^{-1}Y^T \\ 0_{31} & Y & -\epsilon Y\Lambda \\ 0_{31} & 0_3 & Y \end{pmatrix}. \tag{43}$$

The corresponding diagonalizing matrices in the two-dimensional case are given in Appendix A.

Remark 7. Note that we have found one of possibly several diagonalizing matrices. M is not orthogonal, which means that $\Lambda_{\mathbf{M}}$ does not hold the eigenvalues of $\mathbf{A}$.

Remark 8. Note that the only condition involved with finding a diagonalizing matrix M is that $\mathbf{A}$ be nonsingular. Then we can choose to rotate $\mathbf{A}$ to block diagonal form with blocks of arbitrary size. If the blocks are small enough, we can derive their eigenvalues analytically.

3.4. Step (iv) and (v): Split $\Lambda_{\mathbf{M}}$ and $\tilde{w}$. In order to know which components of $\tilde{w}$ to bound with boundary conditions we need to investigate the sign of the diagonal entries of $\Lambda_{\mathbf{M}}$, i.e., the eigenvalues of $\mathbf{E}$ (step (iv)).

Table 1
The sign of the eigenvalues for different Mach numbers.

Eigenvalue	$M_n < -1$	$-1 < M_n < 0$	$0 < M_n < 1$	$M_n > 1$
λ_1	$-$	$-$	$+$	$+$
λ_2	$-$	$-$	$+$	$+$
λ_3	$-$	$+$	$-$	$+$
λ_4	$-$	$-$	$+$	$+$
λ_5	$-$	$-$	$+$	$+$
λ_6	$+$	$+$	$-$	$-$
λ_7	$+$	$-$	$+$	$-$
λ_8	$+$	$+$	$-$	$-$
λ_9	$+$	$+$	$-$	$-$

Table 2
The number of boundary conditions to be specified at different flow cases for the three-dimensional Navier–Stokes equations.

Supersonic inflow	5
Subsonic inflow	5
Subsonic outflow	4
Supersonic outflow	4

Table 3
The number of boundary conditions to be specified at different flow cases for the three-dimensional Euler equations.

Supersonic inflow	5
Subsonic inflow	4
Subsonic outflow	1
Supersonic outflow	0

In the proof of Proposition 3.1, λ_3 given by (33) was analyzed. It was shown that λ_3 changes sign at $M_n = 0$ and $|M_n| = 1$. The eigenvalues $\lambda_1, \lambda_2, \lambda_4$, and λ_5 only change signs at $M_n = 0$. Thus, the different cases are inflow or outflow and sub- or supersonic flow. A consequence is that sub- or supersonic flow affects which boundary conditions to choose but not the number of them. In fact, only the boundary condition corresponding to λ_3 (and hence $-\epsilon^2\lambda_3^{-1} \equiv \lambda_7$) changes sign at $|M_n| = 1$. With $\Lambda = \mathrm{diag}(\lambda_2, \lambda_3, \lambda_4, \lambda_5)$, the diagonal form of $\mathbf{E}$ is $\Lambda_{\mathbf{M}} = \mathrm{diag}(\lambda_1, \Lambda, -\epsilon^2\Lambda^{-1})$. In Table 1 the signs of the different eigenvalues are summarized, where $\lambda_6, \ldots, \lambda_9$ denotes the diagonal entries of $-\epsilon^2\Lambda^{-1}$. Those with negative signs have to be supplied with boundary conditions. As mentioned above, since $\hat{\mathbf{n}}$ is the outward-pointing normal, negative values of M_n indicate inflow and positive values mean outflow.

In Table 2 the numbers of boundary conditions deduced from Table 1 for different flow cases are shown. They are in full agreement with the results from the Laplace transform technique derived in [14] and also in [15]. Note that in the Euler limit, i.e., $\epsilon \to 0$, the last four eigenvalues will become zero, and there are five nontrivial eigenvalues. In Table 3 the numbers of boundary conditions are displayed for the Euler case, $\epsilon \to 0$. The result agrees with the well-known theory for Euler equations.

At last, we can split $\tilde{w}$ given by (42) into $\tilde{w}^+$ and w^- corresponding to the positive and negative eigenvalues and perform step (v), such that well-posedness follows.

Remark 9. Though there are no numerical computations in this article, we would like to comment on some computational aspects. We assume that we know the exact boundary data ahead of time. This implies knowledge of the type of boundary (inflow/outflow, subsonic/supersonic) that we have at each point on the boundary as well as when one boundary type changes to another.

However, in computations, the numerical result might indicate that the assumed data are erroneous. In such a case, this procedure as well as other boundary condition procedures require an adjustment of the given data or location of the boundary for better accuracy.

3.5. Special case: $\boldsymbol{u_n = 0}$. The above derivation gives a set of boundary conditions that leads to a well-posed mathematical problem. However, it is assumed that $u_n \neq 0$, which excludes two cases: tangential flow and the important solid wall condition. We will treat the case $u_n = 0$ separately and redo the steps (iii)–(v). Throughout this paper, we have considered the Navier–Stokes equations linearized around the solution at the boundary, in this case $u_n = 0$. We obtain

$$\mathbf{A} = \begin{pmatrix} 0 & bn_1 & bn_2 & bn_3 & 0 & 0 & 0 & 0 & 0 \\ bn_1 & 0 & 0 & 0 & an_1 & -\epsilon & 0 & 0 & 0 \\ bn_2 & 0 & 0 & 0 & an_2 & 0 & -\epsilon & 0 & 0 \\ bn_3 & 0 & 0 & 0 & an_2 & 0 & 0 & -\epsilon & 0 \\ 0 & an_1 & an_2 & an_3 & 0 & 0 & 0 & 0 & -\epsilon \\ 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 \end{pmatrix}, \tag{44}$$

to which the previous rotation does not apply, since $\mathbf{A}$ is now singular. This leaves us with no other choice but to seek the eigenvalues and eigenvectors of this matrix. It turns out that it is now a simpler task than with $u_n \neq 0$. The result is presented below, and the details of the derivation are found in Appendix B.

Define $\mathbf{m_1}$ and $\mathbf{m_2}$ such that $\hat{\mathbf{n}}^T\mathbf{m_1} = 0$, $\hat{\mathbf{n}}^T\mathbf{m_2} = \mathbf{m_1}^T\mathbf{m_2} = 0$, and

$$\mu_{1,2} = -\frac{c^2}{2} \pm \sqrt{\frac{c^4}{4} + a^2\epsilon^2}.$$

Then,

$$\begin{aligned}
\lambda_1 &= -\epsilon, & e_1 &= (0, \mathbf{m_1}^T, 0, \mathbf{m_1}^T, 0)^T, \\
\lambda_2 &= -\epsilon, & e_2 &= (0, \mathbf{m_2}^T, 0, \mathbf{m_2}^T, 0)^T, \\
\lambda_3 &= \epsilon, & e_3 &= (0, \mathbf{m_1}^T, 0, -\mathbf{m_1}^T, 0)^T, \\
\lambda_4 &= \epsilon, & e_4 &= (0, \mathbf{m_2}^T, 0, -\mathbf{m_2}^T, 0)^T, \\
\lambda_5 &= 0, & e_5 &= \left(1, 0, 0, 0, 0, \frac{b}{\epsilon}\hat{\mathbf{n}}, 0^T\right), \\
\lambda_6 &= \sqrt{\epsilon^2 - \mu_1}, & e_6 &= \left(b, \lambda_6\hat{\mathbf{n}}^T, -\frac{a\lambda_6^2}{\mu_1^2}, -\epsilon\hat{\mathbf{n}}^T, \frac{\epsilon a\lambda_6}{\mu_1^2}\right)^T, \\
\lambda_7 &= -\sqrt{\epsilon^2 - \mu_1}, & e_7 &= \left(b, \lambda_7\hat{\mathbf{n}}^T, -\frac{a\lambda_7^2}{\mu_1^2}, -\epsilon\hat{\mathbf{n}}^T, \frac{\epsilon a\lambda_7}{\mu_1^2}\right)^T, \\
\lambda_8 &= \sqrt{\epsilon^2 - \mu_2}, & e_8 &= \left(b, \lambda_8\hat{\mathbf{n}}^T, -\frac{a\lambda_8^2}{\mu_2^2}, -\epsilon\hat{\mathbf{n}}^T, \frac{\epsilon a\lambda_8}{\mu_2^2}\right)^T, \\
\lambda_9 &= -\sqrt{\epsilon^2 - \mu_2}, & e_9 &= \left(b, \lambda_9\hat{\mathbf{n}}^T, -\frac{a\lambda_9^2}{\mu_2^2}, -\epsilon\hat{\mathbf{n}}^T, \frac{\epsilon a\lambda_9}{\mu_2^2}\right)^T.
\end{aligned} \tag{45}$$

Remark 10. With some algebra one can show that $\epsilon^2 \geq \mu_{1,2}$ such that the eigenvalues $\lambda_6, \dots, \lambda_9$ are real. In fact, since $\mathbf{A}$ is symmetric and the vectors $e_1, \dots, e_9$ are orthogonal and diagonalize $\mathbf{A}$, $\lambda_1, \dots, \lambda_9$ have to be real.

Above, step (iii) is performed and we turn to step (iv). We have

$$\begin{aligned}\Lambda^- &= \mathrm{diag}(\lambda_1, \lambda_2, 0, 0, 0, 0, \lambda_7, 0, \lambda_9),\\ \Lambda^+ &= \mathrm{diag}(0, 0, \lambda_3, \lambda_4, 0, \lambda_6, 0, \lambda_8, 0).\end{aligned}$$

Remark 11. Note that we have four negative eigenvalues. This means that a boundary with $u_n = 0$ is classified as an outflow boundary.

Further, $\tilde{w} = X^T v$, where the column vectors of X are the eigenvectors. With $\tilde{\mathbf{u}} = (\tilde{u}_1, \tilde{u}_2, \tilde{u}_3)^T$, $\tilde{G}^V_{i \dots j} = (\tilde{G}^V_i, \dots, \tilde{G}^V_j)^T$, and the ith component of $\tilde{v}$ denoted by $\tilde{v}_i$, we obtain

$$\tilde{w} = \begin{pmatrix} \mathbf{m_1}^T(\tilde{\mathbf{u}} + \tilde{G}^V_{1\dots 3}) \\ \mathbf{m_2}^T(\tilde{\mathbf{u}} + \tilde{G}^V_{1\dots 3}) \\ \mathbf{m_1}^T(\tilde{\mathbf{u}} - \tilde{G}^V_{1\dots 3}) \\ \mathbf{m_2}^T(\tilde{\mathbf{u}} - \tilde{G}^V_{1\dots 3}) \\ v_1 + \frac{b}{\epsilon}\hat{\mathbf{n}}^T(\tilde{G}^V)_{1\dots 3} \\ bv_1 + \hat{\mathbf{n}}^T(\lambda_6\tilde{\mathbf{u}} - \epsilon\tilde{G}^V_{1\dots 3}) - \frac{a\lambda_6}{\mu_1^2}(\lambda_6 v_4 - \epsilon\tilde{G}^V_4) \\ bv_1 + \hat{\mathbf{n}}^T(\lambda_7\tilde{\mathbf{u}} - \epsilon\tilde{G}^V_{1\dots 3}) - \frac{a\lambda_7}{\mu_1^2}(\lambda_7 v_4 - \epsilon\tilde{G}^V_4) \\ bv_1 + \hat{\mathbf{n}}^T(\lambda_8\tilde{\mathbf{u}} - \epsilon\tilde{G}^V_{1\dots 3}) - \frac{a\lambda_8}{\mu_2^2}(\lambda_8 v_4 - \epsilon\tilde{G}^V_4) \\ bv_1 + \hat{\mathbf{n}}^T(\lambda_9\tilde{\mathbf{u}} - \epsilon\tilde{G}^V_{1\dots 3}) - \frac{a\lambda_9}{\mu_2^2}(\lambda_9 v_4 - \epsilon\tilde{G}^V_4) \end{pmatrix}. \tag{46}$$

Finally, we can split $\tilde{w}$ into $\tilde{w}^+$ and $\tilde{w}^-$ as before and perform step (v), i.e., supply $\tilde{w}^-$ with boundary conditions to obtain a well-posed system.

Remark 12. There are two more cases where $u_n \neq 0$. Those are tangential flows with $|M_n| = 1$. To find the eigenvalues of $\mathbf{A}$ directly for $M_n = 1, -1$ is equally difficult as the general case, and we did not find roots in closed form.

3.6. Curvilinear coordinates. Until now, we have analyzed well-posed boundary conditions for the Navier–Stokes equations in a Cartesian coordinate system and a general domain. Considering numerical computations, that derivation suffices when using unstructured methods such as finite volume schemes. However, for structured methods, such as finite difference schemes, the Navier–Stokes equations are usually expressed in a curvilinear coordinate system. We have included a brief analysis in Appendix C showing that the Cartesian results are directly applicable in the curvilinear case through metric transformations.

4. Conclusions. We have proposed a step-by-step procedure to analyze a general time dependent partial differential equation in terms of well-posedness including boundary conditions. The procedure applied to the Euler equations results in the well-known characteristic boundary conditions. In this article we have applied the procedure to the three-dimensional Navier–Stokes equations on a general domain and obtained a novel set of well-posed boundary conditions.

Appendix A. The two-dimensional matrices. With very few comments and leaving out most details, we show the differences of the derivation in section 3 for the two-dimensional case.

With

$$B_{21} = B_{12} = \frac{B_{xy}}{2},$$

the symmetrized equations are

$$\tilde{u}_t + A_1\tilde{u}_x + A_2\tilde{u}_y = \epsilon(B_{11}\tilde{u}_{xx} + B_{22}\tilde{u}_{yy} + B_{12}\tilde{u}_{xy} + B_{21}\tilde{u}_{yx}).$$

The matrices are obtained by deleting the row and column referring to the u_3 component (see [10]). We introduce

$$\tilde{F}_v = B_{11}\tilde{u}_x + B_{21}\tilde{u}_y, \quad \tilde{G}_v = B_{22}\tilde{u}_y + B_{12}\tilde{u}_x,$$

such that

$$\frac{1}{2}\|\tilde{u}\|_t^2 + \oint_{\partial D} \frac{1}{2}\begin{pmatrix} \tilde{u} \\ \tilde{F}^V \end{pmatrix}^T \begin{pmatrix} A_1 n_1 + A_2 n_2 & -\epsilon I_4 \\ -\epsilon I_4 & 0_4 \end{pmatrix}\begin{pmatrix} \tilde{u} \\ \tilde{F}^V \end{pmatrix} = DI,$$

where $\hat{\mathbf{n}} = [n_1, n_2]$, $ds = \sqrt{dx^2 + dy^2}$, and $\tilde{F}^V = \tilde{F}_v n_1 + \tilde{G}_v n_2$.

By deleting the first component of $\tilde{F}^V$ yielding $\tilde{G}^V$, the matrix is reduced from an 8-by-8 matrix to a 7-by-7 matrix. With $\mathbf{u} = (u_1, u_2)$, we obtain

$$\begin{pmatrix} \tilde{u} \\ \tilde{F}^V \end{pmatrix}^T \begin{pmatrix} A_1 n_1 + A_2 n_2 & -\epsilon I_4 \\ -\epsilon I_4 & 0_4 \end{pmatrix}\begin{pmatrix} \tilde{u} \\ \tilde{F}^V \end{pmatrix} = \begin{pmatrix} \tilde{u} \\ \tilde{G}^V \end{pmatrix}^T \mathbf{A} \begin{pmatrix} \tilde{u} \\ \tilde{G}^V \end{pmatrix},$$

where

$$\mathbf{A} = \begin{pmatrix} \mathbf{u}\cdot\hat{\mathbf{n}} & bn_1 & bn_2 & 0 & 0 & 0 & 0 \\ bn_1 & \mathbf{u}\cdot\hat{\mathbf{n}} & 0 & an_1 & -\epsilon & 0 & 0 \\ bn_2 & 0 & \mathbf{u}\cdot\hat{\mathbf{n}} & an_2 & 0 & -\epsilon & 0 \\ 0 & an_1 & an_2 & \mathbf{u}\cdot\hat{\mathbf{n}} & 0 & 0 & -\epsilon \\ 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\epsilon & 0 & 0 & 0 \end{pmatrix}$$
$$= \begin{pmatrix} A_{11} & A_{12} & 0_{14} \\ A_{21} & A_{22} & -\epsilon I_4 \\ 0_{41} & -\epsilon I_4 & 0_4 \end{pmatrix}.$$

The rotation of $\mathbf{A}$ is precisely similar,

$$\mathbf{R}^T\mathbf{A}\mathbf{R} = \begin{pmatrix} 1 & 0_{13} & 0_{13} \\ \bar{\alpha}^T & I_3 & 0_3 \\ \bar{\beta}^T & \bar{\gamma}^T & I_3 \end{pmatrix}\begin{pmatrix} A_{11} & A_{12} & 0_{13} \\ A_{21} & A_{22} & -\epsilon I_3 \\ 0_{31} & -\epsilon I_3 & 0_3 \end{pmatrix}\begin{pmatrix} 1 & \bar{\alpha} & \bar{\beta} \\ 0_{31} & I_3 & \bar{\gamma} \\ 0_{31} & 0_3 & I_3 \end{pmatrix}$$
$$= \begin{pmatrix} E_{11} & E_{12} & E_{13} \\ E_{21} & E_{22} & E_{23} \\ E_{31} & E_{32} & E_{33} \end{pmatrix} = \mathbf{E}.$$

The same solution is obtained,

$$\bar{\alpha} = -A_{11}^{-1}A_{12}, \quad \bar{\beta} = A_{11}^{-1}A_{12}E_{22}^{-1}, \quad \bar{\gamma} = -\epsilon E_{22}^{-1},$$

$$\mathbf{E} = \begin{pmatrix} A_{11} & 0_{14} & 0_{14} \\ 0_{41} & E_{22} & 0_4 \\ 0_{41} & 0_4 & -\epsilon^2 E_{22}^{-1} \end{pmatrix}, \qquad E_{22} = A_{22} - A_{21}A_{11}^{-1}A_{12}.$$

The first eigenvalue of $\mathbf{E}$ is $\lambda_1 = A_{11} = u_n$, and the others are given by the eigenvalues of E_{22},

$$E_{22} = \begin{pmatrix} -\frac{b^2 n_1^2}{u_n} + u_n & -\frac{b^2 n_1 n_2}{u_n} & a n_1 \\ -\frac{b^2 n_1 n_2}{u_n} & -\frac{b^2 n_2^2}{u_n} + u_n & a n_2 \\ a n_1 & a n_2 & u_n \end{pmatrix},$$

$$\lambda_{2,3} = \frac{-b^2 + 2u_n^2 \pm \sqrt{b^4 + 4a^2 u_n^2}}{2u_n}, \quad \lambda_4 = u_n,$$

where $n_1^2 + n_2^2 = 1$ and $u_n = \mathbf{u} \cdot \hat{\mathbf{n}}$. These can be simplified similarly as for the three-dimensional case.

The eigenvectors $Y = (y_2, y_3, y_4)$ are

$$y_2 = \begin{pmatrix} n_1 \\ n_2 \\ \frac{-\lambda_3 + u_n}{a} \end{pmatrix}, \quad y_3 = \begin{pmatrix} n_1 \\ n_2 \\ \frac{-\lambda_2 + u_n}{a} \end{pmatrix}, \quad y_4 = \begin{pmatrix} -n_2, \\ n_1 \\ 0 \end{pmatrix}. \tag{47}$$

Introduce the block matrix, $X = \operatorname{diag}(1, Y, Y)$, such that $X^T \mathbf{E} X = \Lambda$, where $\Lambda = \operatorname{diag}(u_n, \Lambda, -\epsilon^2\Lambda)$. Let $\tilde{v} = (\tilde{u}^T, (\tilde{G}^V)^T)^T$; then $\tilde{v}^T \mathbf{A} \tilde{v} = \tilde{w}^T \Lambda \tilde{w}$, where $\tilde{w} = X^T R^{-1} \tilde{v} = M^{-1}\tilde{v}$ and $\Lambda = M^T \mathbf{A} M$. The matrices are

$$R^{-1} = \begin{pmatrix} 1 & -\bar{\alpha} & 0_{13} \\ 0_{31} & I_3 & -\bar{\gamma} \\ 0_{31} & 0_3 & I_3 \end{pmatrix}, \qquad M^{-1} = \begin{pmatrix} 1 & -\bar{\alpha} & 0_{14} \\ 0_{41} & Y^T & \epsilon\Lambda^{-1}Y^T \\ 0_{41} & 0_4 & Y^T \end{pmatrix},$$

where

$$\Lambda^{-}Y^T = \begin{pmatrix} \lambda_2^{-1} y_2 \\ \lambda_3^{-1} y_3 \\ \lambda_4^{-1} y_4 \end{pmatrix}, \quad \bar{\alpha} = \left(-\frac{b}{u_n}\hat{\mathbf{n}}, 0\right), \quad M = \begin{pmatrix} 1 & \bar{\alpha}Y & \bar{\alpha}\Lambda^{-1}Y^T \\ 0_{31} & Y & -\epsilon Y \Lambda \\ 0_{31} & 0_3 & Y \end{pmatrix}.$$

In two dimensions, $\tilde{v}$ is

$$\tilde{v} = \left(\frac{b}{\rho}\frac{\tilde{\rho}}{\rho}, \tilde{u}_1, \tilde{u}_2, -\frac{b}{\sqrt{\gamma - 1}}\tilde{\rho} + \frac{1}{\rho a}\tilde{p}, \tilde{G}_1^V, \tilde{G}_2^V, \tilde{G}_3^V\right).$$

Then,

$$\tilde{w} = M^{-1}\tilde{v} = \begin{pmatrix} \tilde{v}_1 - \bar{\alpha} \cdot \tilde{v}_{2\ldots4} \\ y_2^T(\tilde{v}_{2\ldots4} - \epsilon\lambda_2^{-1}\tilde{G}^V) \\ y_3^T(\tilde{v}_{2\ldots4} - \epsilon\lambda_3^{-1}\tilde{G}^V) \\ y_4^T(\tilde{v}_{2\ldots4} - \epsilon\lambda_4^{-1}\tilde{G}^V) \\ y_2^T\tilde{G}^V \\ y_3^T\tilde{G}^V \\ y_4^T\tilde{G}^V \end{pmatrix}. \tag{48}$$

Appendix B. Diagonalization with $\boldsymbol{u_n = 0}$. Consider

$$\mathbf{A}e = \lambda e, \tag{49}$$

where $\mathbf{A}$ is given by (44), repeated here for convenience,

$$\mathbf{A} = \begin{pmatrix} 0 & bn_1 & bn_2 & bn_3 & 0 & 0 & 0 & 0 & 0 \\ bn_1 & 0 & 0 & 0 & an_1 & -\epsilon & 0 & 0 & 0 \\ bn_2 & 0 & 0 & 0 & an_2 & 0 & -\epsilon & 0 & 0 \\ bn_3 & 0 & 0 & 0 & an_2 & 0 & 0 & -\epsilon & 0 \\ 0 & an_1 & an_2 & an_3 & 0 & 0 & 0 & 0 & -\epsilon \\ 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -\epsilon & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{50}$$

The structure of $\mathbf{A}$ suggests the following ansatz:

$$e_1 = (0, m_1, m_2, m_3, 0, m_1, m_2, m_3, 0)^T, \tag{51}$$

$$e_2 = (0, m_1, m_2, m_3, 0, -m_1, -m_2, -m_3, 0)^T, \tag{52}$$

$$e_3 = (m_4, m_5n_1, m_5n_2, m_5n_3, m_6, m_7n_1, m_7n_2, m_7n_3, m_8). \tag{53}$$

We will use the notation $\mathbf{m} = (m_1, m_2, m_3)^T$. With (51), equation (49) becomes

$$\begin{pmatrix} b\mathbf{\hat{n}^T m} \\ -\epsilon m_1 \\ -\epsilon m_2 \\ -\epsilon m_3 \\ a\mathbf{\hat{n}^T m} \\ -\epsilon m_1 \\ -\epsilon m_2 \\ -\epsilon m_3 \\ 0 \end{pmatrix} = \lambda \begin{pmatrix} 0 \\ m_1 \\ m_2 \\ m_3 \\ 0 \\ m_1 \\ m_2 \\ m_3 \\ 0 \end{pmatrix}. \tag{54}$$

With $\lambda = \lambda_1$ and $\mathbf{m} = \mathbf{m}_1$, the following choice satisfies the above equation, $\mathbf{\hat{n}}^T\mathbf{m_1} = 0$, and $\lambda_1 = -\epsilon$. Further, we may also choose a second solution $\mathbf{m} = \mathbf{m}_2$ and $\lambda_2 = -\epsilon$ such that $\mathbf{\hat{n}}^T\mathbf{m_2} = 0$ and $\mathbf{m_2}^T\mathbf{m_1} = 0$. Similarly, ansatz (52) yields

$$\lambda_3 = \epsilon, \quad \mathbf{\hat{n}}^T\mathbf{m_3} = 0, \tag{55}$$

$$\lambda_4 = \epsilon, \quad \mathbf{\hat{n}}^T\mathbf{m_4} = 0, \quad \mathbf{m_3}^T\mathbf{m_4} = 0. \tag{56}$$

In fact, we can let $\mathbf{m}_1 = \mathbf{m}_3$ and $\mathbf{m}_2 = \mathbf{m}_4$. It is obvious that the vectors (51) and (52) will be orthogonal, and, by definition, they are orthogonal to (53). So far, four eigenvalues and eigenvectors out of nine are derived when we turn to the last ansatz (53). In this case (49) becomes

$$\begin{pmatrix} m_5b \\ (bm_4 + am_6 - \epsilon m_7)n_1 \\ (bm_4 + am_6 - \epsilon m_7)n_2 \\ (bm_4 + am_6 - \epsilon m_7)n_3 \\ am_5 - \epsilon m_8 \\ -\epsilon m_5n_1 \\ -\epsilon m_5n_2 \\ -\epsilon m_5n_3 \\ -\epsilon m_6 \end{pmatrix} = \lambda \begin{pmatrix} m_4 \\ m_5n_1 \\ m_5n_2 \\ m_5n_3 \\ m_6 \\ m_7n_1 \\ m_7n_2 \\ m_7n_3 \\ m_8 \end{pmatrix}, \tag{57}$$

where $n_1^2+n_2^2+n_3^3$ has been used. Note that the above system of equations reduces to only five equations by the choice of the eigenvector. Further, we have five unknowns, including λ. (One of the unknowns of the eigenvector drops out since it should only enter as a scaling.) We have

$$m_5 b = \lambda m_4, \tag{58}$$

$$b m_4 + a m_6 - \epsilon m_7 = \lambda m_5, \tag{59}$$

$$a m_5 - \epsilon m_8 = \lambda m_6, \tag{60}$$

$$-\epsilon m_5 = \lambda m_7, \tag{61}$$

$$-\epsilon m_6 = \lambda m_8. \tag{62}$$

In this case it turns out that the ansatz was satisfactory since five solutions to the system (58)–(62) exist.

The case we examine is the marginal case with $u_n = 0$, which leads us to expect one eigenvalue to be zero. Thusly, with $\lambda_5 = 0$ the following eigenvector is obtained:

$$e_5 = \left(1, 0, 0, 0, 0, \frac{b}{\epsilon} n_1, \frac{b}{\epsilon} n_2, \frac{b}{\epsilon} n_3, 0\right)^T. \tag{63}$$

Next, we solve full system (58)–(62) without assumptions on the solution. With $\mu = \epsilon^2 - \lambda^2$, a second degree equation in μ is obtained,

$$\mu^2 + (b + a^2)\mu - a^2\epsilon^2 = 0, \tag{64}$$

with the solutions

$$\mu_{1,2} = -\frac{b + a^2}{2} \pm \sqrt{\frac{(b + a^2)^2}{4} + a^2\epsilon^2} = -\frac{c^2}{2} \pm \sqrt{\frac{c^4}{4} + a^2\epsilon^2} \tag{65}$$

such that $\lambda_{6,7} = \pm\sqrt{\epsilon^2 - \mu_1}$ and $\lambda_{8,9} = \pm\sqrt{\epsilon^2 - \mu_2}$. For any of these λ's the eigenvector is given by

$$e = \begin{pmatrix} b \\ \lambda n_1 \\ \lambda n_2 \\ \lambda n_3 \\ -\frac{a\lambda^2}{\epsilon^2 - \lambda^2} \\ -\epsilon n_1 \\ -\epsilon n_2 \\ -\epsilon n_3 \\ \frac{\epsilon a \lambda}{\epsilon^2 - \lambda^2} \end{pmatrix}. \tag{66}$$

Next, we have to show that the different eigenvectors obtained from (66) are orthogonal to each other. We distinguish between two cases: 1. any of the eigenvalues derived from μ_1, denoted by ξ_1, and another eigenvalue ξ_2 derived from μ_2; 2. both eigenvalues $\xi_{1,2}$ derived from the same μ.

The scalar product is

$$e(\xi_1)^T \cdot e(\xi_2) = b + \xi_1\xi_2 + \frac{a^2\xi_1^2\xi_2^2}{(\epsilon^2 - \xi_1^2)(\epsilon^2 - \xi_2^2)} + \epsilon^2 + \frac{\epsilon^2 a^2 \xi_1 \xi_2}{(\epsilon^2 - \xi_1^2)(\epsilon^2 - \xi_2^2)}. \tag{67}$$

Case 1. For a general quadratic equation $x^2+px+q=0$ the roots fulfill $x_1x_2=q$ and $x_1+x_2=-p$. When applied to (64) this implies

$$\mu_1\mu_2=(\epsilon^2-\xi_1^2)(\epsilon^2-\xi_2^2)=-a^2\epsilon^2, \tag{68}$$

$$\mu_1+\mu_2=-(b+a^2). \tag{69}$$

Thus, (67) is

$$\begin{aligned} & b+\xi_1\xi_2-\frac{\xi_1^2\xi_2^2}{\epsilon^2}+\epsilon^2-\xi_1\xi_2 \\ &= b+\epsilon^2+\frac{(\epsilon^2-\mu_1)(\epsilon^2-\mu_2)}{\epsilon^2} \\ &= b+\epsilon^2-(\epsilon^2-(\mu_1+\mu_2)-a^2) \\ &= b-(b+a^2)+a^2=0. \end{aligned} \tag{70}$$

Case 2. In this case the following relations hold:

$$\lambda^2=\xi_1^2=\xi_2^2, \tag{71}$$

$$\lambda=\xi_1=-\xi_2, \tag{72}$$

$$\lambda^2=-\xi_1\xi_2=(\mu-\epsilon^2),$$

$$(\epsilon^2-\xi_{1,2}^2)=\mu. \tag{73}$$

Then (67) becomes, after multiplying by $(\epsilon^2-\lambda^2)^2$,

$$\begin{aligned} & (\epsilon^2-\lambda^2)^2(b-\lambda^2+\epsilon^2)+a^2\lambda^4-\epsilon^2a^2\lambda^2 \\ &= (b-\lambda^2+\epsilon^2)(\epsilon^2-\lambda^2)^2+a^2\lambda^2(\lambda^2-\epsilon^2) \\ &= (\lambda^2-\epsilon^2)((b+(\epsilon-\lambda^2))(\epsilon^2-\lambda^2)+a^2\lambda^2) \\ &= -\mu((b+\mu)\mu+a^2(\mu-\epsilon^2)) \\ &= -\mu(\mu^2+(b+a^2)\mu-a^2\epsilon^2)=0, \end{aligned}$$

where the last equality is due to (64).

One should also normalize these vectors to formally obtain the eigenvectors of the matrix $\mathbf{A}$. With this done, we conclude that in the case of neither inflow nor outflow, the above derivation gives the eigenvalues and eigenvectors of the linearized Navier–Stokes equations in three dimensions.

Appendix C. Curvilinear coordinates.

C.1. Metric relations.

Let x, y, z denote the usual Cartesian coordinates. Consider the following coordinate transformation:

$$\xi=\xi(x,y,z), \quad \eta=\eta(x,y,z), \quad \zeta=\zeta(x,y,z).$$

The Jacobian is defined as

$$\mathbf{J}=\begin{pmatrix} x_\xi & x_\eta & x_\zeta \\ y_\xi & y_\eta & y_\zeta \\ z_\xi & z_\eta & z_\zeta \end{pmatrix}. \tag{74}$$

Let $\bar{x}=(x,y,z)=(x_1,x_2,x_3)$ and $\bar{\xi}=(\xi,\eta,\zeta)=(\xi_1,\xi_2,\xi_3)$. Then we can formally express the Jacobian as $\mathcal{D}_{\bar{\xi}}\bar{x} = \mathbf{J}$. The following relation holds:

$$I = \mathcal{D}_{\bar{x}}\bar{x}(\bar{\xi}) = \mathcal{D}_{\bar{\xi}}\bar{x}\mathcal{D}_{\bar{x}}\bar{\xi}. \tag{75}$$

Hence,

$$\mathbf{J}^{-1} = \mathcal{D}_{\bar{x}}\xi = \begin{pmatrix} \xi_x & \xi_y & \xi_z \\ \eta_x & \eta_y & \eta_z \\ \zeta_x & \zeta_y & \zeta_z \end{pmatrix}. \tag{76}$$

However, $\mathbf{J}^{-1}$ can also be obtained directly by inverting (74),

$$\mathbf{J}^{-1} = \mathcal{D}_{\bar{x}}\xi = \frac{1}{J}\begin{pmatrix} y_\eta z_\zeta - y_\zeta z_\eta & -(x_\eta z_\zeta - x_\zeta z_\eta) & x_\eta y_\zeta - x_\zeta y_\eta \\ -(y_\xi z_\zeta - y_\zeta z_\xi) & x_\xi z_\zeta - x_\zeta z_\xi & -(x_\xi y_\zeta - x_\zeta y_\xi) \\ y_\xi z_\eta - y_\eta z_\xi & -(x_\xi z_\eta - x_\eta z_\xi) & x_\xi y_\eta - x_\eta y_\xi \end{pmatrix}, \tag{77}$$

where J denotes the determinant of the Jacobian. Then (76) and (77) give relations between the different metric coefficients. For example, we note that

$$\begin{aligned} (J\xi_x)_\xi + (J\eta_x)_\eta + (J\zeta_x)_\zeta &= (y_\eta z_\zeta - y_\zeta z_\eta)_\xi - (y_\xi z_\zeta - y_\zeta z_\xi)_\eta + (y_\xi z_\eta - y_\eta z_\xi)_\zeta = 0, \\ (J\xi_y)_\xi + (J\eta_y)_\eta + (J\zeta_y)_\zeta &= -(x_\eta z_\zeta - x_\zeta z_\eta)_\xi + (x_\xi z_\zeta - x_\zeta z_\xi)_\eta - (x_\xi z_\eta - x_\eta z_\xi)_\zeta = 0, \\ (J\xi_z)_\xi + (J\eta_z)_\eta + (J\zeta_z)_\zeta &= (x_\eta y_\zeta - x_\zeta y_\eta)_\xi - (x_\xi y_\zeta - x_\zeta y_\xi)_\eta + (x_\xi y_\eta - x_\eta y_\xi)_\zeta = 0, \end{aligned} \tag{78}$$

which will be used below.

C.2. Curvilinear Navier–Stokes equations. Consider the linearized and symmetrized Navier–Stokes equations (9), restated here for convenience,

$$\begin{aligned} \tilde{u}_t + \quad & (A_1\tilde{u} - \epsilon(B_{11}\tilde{u}_x + B_{12}\tilde{u}_y + B_{13}\tilde{u}_z))_x \\ & + (A_2\tilde{u} - \epsilon(B_{22}\tilde{u}_y + B_{23}\tilde{u}_z + B_{12}\tilde{u}_x))_y \\ & + (A_3\tilde{u} - \epsilon(B_{33}\tilde{u}_z + B_{32}\tilde{u}_y + B_{13}\tilde{u}_x))_z = 0 \end{aligned}$$

or

$$\begin{aligned} \tilde{u}_t + (F^I - \epsilon\tilde{F}_v)_x + (G^I - \epsilon\tilde{G}_v)_y + (H^I - \epsilon\tilde{H}_v)_z \\ = \tilde{u}_t + F_x + G_y + H_z = 0. \end{aligned} \tag{79}$$

Multiply (79) by J and make the change of coordinates,

$$\begin{aligned} 0 &= (J\tilde{u})_t + JF_x + JG_y + JH_z \\ &= (J\tilde{u})_t + J\xi_x F_\xi + J\eta_x F_\eta + J\zeta_x F_\zeta \\ &\quad + J\xi_y G_\xi + J\eta_y G_\eta + J\zeta_y G_\zeta \\ &\quad + J\xi_z H_\xi + J\eta_z H_\eta + J\zeta_z H_\zeta. \end{aligned} \tag{80}$$

Reformulating (80) yields

$$\begin{aligned} (J\tilde{u})_t &+ (J\xi_x F + J\xi_y G + J\xi_z H)_\xi - R_1 \\ &+ (J\eta_x F + J\eta_y G + J\eta_z H)_\eta - R_2 \\ &+ (J\zeta_x F + J\zeta_y G + J\zeta_z H)_\zeta - R_3, \end{aligned}$$

where

$$\begin{aligned}R_1 &= (J\xi_x)_\xi F + (J\xi_y)_\xi G + (J\xi_z)_\xi H,\\ R_2 &= (J\eta_x)_\eta F + (J\eta_y)_\eta G + (J\eta_z)_\eta H,\\ R_3 &= (J\zeta_x)_\zeta F + (J\zeta_y)_\zeta G + (J\zeta_z)_\zeta H.\end{aligned}$$

By using the metric relations in (78), we obtain

$$\begin{aligned}R_1 + R_2 + R_3 = F(&(J\xi_x)_\xi + (J\eta_x)_\eta + (J\zeta_x)_\zeta)\\ + G(&(J\xi_y)_\xi + (J\eta_y)_\eta + (J\zeta_y)_\zeta)\\ + H(&(J\xi_z)_\xi + (J\eta_z)_\eta + (J\zeta_z)_\zeta) = 0.\end{aligned}$$

Define

$$\begin{aligned}\hat{F} &= (J\xi_x F + J\xi_y G + J\xi_z H),\\ \hat{G} &= (J\eta_x F + J\eta_y G + J\eta_z H),\\ \hat{H} &= (J\zeta_x F + J\zeta_y G + J\zeta_z H)\end{aligned}$$

such that

$$0 = (J\tilde{u})_t + JF_x + JG_y + JH_z = (J\tilde{u})_t + \hat{F}_\xi + \hat{G}_\eta + \hat{H}_\zeta. \tag{81}$$

Next, we express the new fluxes in curvilinear coordinates. We obtain

$$\begin{aligned}\hat{F}^I &= (J\xi_x F^I + J\xi_y G^I + J\xi_z H^I) = J(\xi_x A_1 + \xi_y A_2 + \xi_z A_3)u,\\ \hat{G}^I &= (J\eta_x F^I + J\eta_y G^I + J\eta_z H^I) = J(\eta_x A_1 + \eta_y A_2 + \eta_z A_3)u,\\ \hat{H}^I &= (J\zeta_x F^I + J\zeta_y G^I + J\zeta_z H^I) = J(\zeta_x A_1 + \zeta_y A_2 + \zeta_z A_3)u,\end{aligned} \tag{82}$$

and

$$\begin{aligned}\hat{F}_v &= (J\xi_x \tilde{F}_v + J\xi_y \tilde{G}_v + J\xi_z \tilde{H}_v),\\ \hat{G}_v &= (J\eta_x \tilde{F}_v + J\eta_y \tilde{G}_v + J\eta_z \tilde{H}_v),\\ \hat{H}_v &= (J\zeta_x \tilde{F}_v + J\zeta_y \tilde{G}_v + J\zeta_z \tilde{H}_v),\end{aligned} \tag{83}$$

where

$$\begin{aligned}\tilde{F}_v &= \tilde{B}_{11}\tilde{u}_\xi + \tilde{B}_{12}\tilde{u}_\eta + \tilde{B}_{13}\tilde{u}_\zeta,\\ \tilde{G}_v &= \tilde{B}_{22}\tilde{u}_\eta + \tilde{B}_{23}\tilde{u}_\zeta + \tilde{B}_{12}\tilde{u}_\xi,\\ \tilde{H}_v &= \tilde{B}_{33}\tilde{u}_\zeta + \tilde{B}_{32}\tilde{u}_\eta + \tilde{B}_{13}\tilde{u}_\xi,\end{aligned}$$

and

$$\begin{aligned}\tilde{B}_{11} &= B_{11}\xi_x + B_{12}\xi_y + B_{13}\xi_z, & \tilde{B}_{12} &= B_{11}\eta_x + B_{12}\eta_y + B_{13}\eta_z,\\ \tilde{B}_{13} &= B_{11}\zeta_x + B_{12}\zeta_y + B_{13}\zeta_z, & \tilde{B}_{22} &= B_{22}\xi_y + B_{23}\xi_z + B_{12}\xi_x,\\ \tilde{B}_{23} &= B_{22}\eta_y + B_{23}\eta_z + B_{12}\eta_x, & \tilde{B}_{21} &= B_{22}\zeta_y + B_{23}\zeta_z + B_{12}\zeta_x,\\ \tilde{B}_{33} &= B_{33}\xi_z + B_{32}\xi_y + B_{13}\xi_x, & \tilde{B}_{32} &= B_{33}\eta_z + B_{32}\eta_y + B_{13}\eta_x,\\ \tilde{B}_{31} &= B_{33}\zeta_z + B_{32}\zeta_y + B_{13}\zeta_x.\end{aligned}$$

C.3. Energy estimate. Next, we turn to the well-posedness of (81). We apply the energy method and derive the boundary terms. Our aim is to relate the boundary terms in curvilinear coordinates to those derived in $\bar{x}$-space.

First we note that

$$dxdydz = Jd\xi d\eta d\zeta. \tag{84}$$

Further, we use the notation $D_{\bar{\xi}}$ in $\bar{\xi}$-space for the image of the domain $D_{\bar{x}}$ in $\bar{x}$-space.

Apply the energy method to (81) to obtain

$$\begin{aligned} 0 = \int_{D_{\bar{\xi}}} \tilde{u}^T \tilde{u}_t J d\xi d\eta d\zeta + \int_{D_{\bar{\xi}}} \tilde{u}^T (\hat{F}^I_\xi + \hat{G}^I_\eta + \hat{H}^I_\zeta) d\xi d\eta d\zeta \\ -\epsilon \int_{D_{\bar{\xi}}} \tilde{u}^T ((\hat{F}_v)_\xi + (\hat{G}_v)_\eta + (\hat{H}_v)_\zeta) d\xi d\eta d\zeta = \int_{D_{\bar{x}}} \tilde{u}^T \tilde{u}_t dxdydz + I_1 - \epsilon I_2, \end{aligned} \tag{85}$$

$$\begin{aligned} I_2 &= \int_{D_{\bar{\xi}}} (\tilde{u}^T \hat{F}_v)_\xi + (\tilde{u}^T \hat{G}_v)_\eta + (\tilde{u}^T \hat{H}_v)_\zeta d\xi d\eta d\zeta \\ &\quad - \int_{D_{\bar{\xi}}} \tilde{u}^T_\xi (\hat{F}_v)_\xi + \tilde{u}^T_\eta (\hat{G}_v)_\eta + \tilde{u}^T_\zeta (\hat{H}_v)_\zeta d\xi d\eta d\zeta \\ &= \int_{D_{\bar{\xi}}} (\tilde{u}^T \hat{F}_v)_\xi + (\tilde{u}^T \hat{G}_v)_\eta + (\tilde{u}^T \hat{H}_v)_\zeta d\xi d\eta d\zeta - DI \\ &= \oint_{\Gamma_{\bar{\xi}}} (\tilde{u}^T \hat{F}_v, \tilde{u}^T \hat{G}_v, \tilde{u}^T \hat{H}_v) \cdot \mathbf{n}_{\bar{\xi}} ds_{\bar{\xi}} - DI \\ &= \oint_{\Gamma_{\bar{\xi}}} \tilde{u}^T \hat{F}^V ds_{\bar{\xi}} - DI, \end{aligned} \tag{86}$$

where $\mathbf{n}_{\bar{\xi}} = (n_\xi, n_\eta, n_\zeta)$ and $ds_{\bar{\xi}}$ denote the outward-pointing normal and surface element in $\bar{\xi}$-space, respectively. Further, $\hat{F}^V = \hat{F}_v n_\xi + \hat{G}_v n_\eta + \hat{H}_v n_\zeta$. DI denotes a dissipative term and is equal to DI defined in subsection 3.2.

$$\begin{aligned} I_1 &= \int_{D_{\bar{\xi}}} \tilde{u}^T (\hat{F}^I_\xi + \hat{G}^I_\eta + \hat{H}^I_\zeta) d\xi d\eta d\zeta \\ &= \int_{D_{\bar{\xi}}} \tilde{u}^T (J\xi_x A_1 \tilde{u} + J\xi_y A_2 \tilde{u} + J\xi_z A_3 \tilde{u})_\xi \\ &\quad + \tilde{u}^T (J\eta_x A_1 \tilde{u} + J\eta_y A_2 \tilde{u} + J\eta_z A_3 \tilde{u})_\eta \\ &\quad + \tilde{u}^T (J\zeta_x A_1 \tilde{u} + J\zeta_y A_2 \tilde{u} + J\zeta_z A_3 \tilde{u})_\zeta d\xi d\eta d\zeta. \end{aligned}$$

Next, we use relations of the type

$$\begin{aligned} \tilde{u}^T (J\xi_x A_1 \tilde{u})_\xi &= (J\xi_x)_\xi \tilde{u}^T A_1 \tilde{u} + (J\xi_x) \left(\frac{1}{2} \tilde{u}^T A_1 \tilde{u}\right)_\xi \\ &= (J\xi_x)_\xi \tilde{u}^T A_1 \tilde{u} + \left(J\xi_x \frac{1}{2} \tilde{u}^T A_1 \tilde{u}\right)_\xi - (J\xi_x)_\xi \left(\frac{1}{2} \tilde{u}^T A_1 \tilde{u}\right) \\ &= \left(J\xi_x \frac{1}{2} \tilde{u}^T A_1 \tilde{u}\right)_\xi + (J\xi_x)_\xi \left(\frac{1}{2} \tilde{u}^T A_1 \tilde{u}\right) \end{aligned}$$

to obtain

$$\begin{aligned} I_1 = \int_{D_{\bar{\xi}}} & \left(J\xi_x \frac{1}{2}\tilde{u}^T A_1 \tilde{u}\right)_\xi + \left(J\xi_y \frac{1}{2}\tilde{u}^T A_2 \tilde{u}\right)_\xi + \left(J\xi_z \frac{1}{2}\tilde{u}^T A_3 \tilde{u}\right)_\xi \\ & + \left(J\eta_x \frac{1}{2}\tilde{u}^T A_1 \tilde{u}\right)_\eta + \left(J\eta_y \frac{1}{2}\tilde{u}^T A_2 \tilde{u}\right)_\eta + \left(J\eta_z \frac{1}{2}\tilde{u}^T A_3 \tilde{u}\right)_\eta \\ & + \left(J\zeta_x \frac{1}{2}\tilde{u}^T A_1 \tilde{u}\right)_\zeta + \left(J\zeta_y \frac{1}{2}\tilde{u}^T A_2 \tilde{u}\right)_\zeta + \left(J\zeta_z \frac{1}{2}\tilde{u}^T A_3 \tilde{u}\right)_\zeta \\ & + \frac{1}{2}\tilde{u}^T A_1 \tilde{u}(J\xi_x)_\xi + \frac{1}{2}\tilde{u}^T A_1 \tilde{u}(J\eta_x)_\eta + \frac{1}{2}\tilde{u}^T A_1 \tilde{u}(J\zeta_x)_\zeta \\ & + \frac{1}{2}\tilde{u}^T A_2 \tilde{u}(J\xi_y)_\xi + \frac{1}{2}\tilde{u}^T A_2 \tilde{u}(J\eta_y)_\eta + \frac{1}{2}\tilde{u}^T A_2 \tilde{u}(J\zeta_y)_\zeta \\ & + \frac{1}{2}\tilde{u}^T A_3 \tilde{u}(J\xi_z)_\xi + \frac{1}{2}\tilde{u}^T A_3 \tilde{u}(J\eta_z)_\eta + \frac{1}{2}\tilde{u}^T A_3 \tilde{u}(J\zeta_z)_\zeta \, d\xi d\eta d\zeta. \end{aligned} \tag{87}$$

Hence, by using (78), the last three rows of (87) are identically zero:

$$I_1 = \oint_{\Gamma_{\bar{\xi}}} \frac{1}{2}(\tilde{u}^T(\hat{A}_1)\tilde{u}, \tilde{u}^T(\hat{A}_2)\tilde{u}, \tilde{u}^T(\hat{A}_3)\tilde{u}) \cdot \mathbf{n}_{\bar{\xi}} ds_{\bar{\xi}}, \tag{88}$$

where

$$\begin{aligned} \hat{A}_1 &= (A_1 J\xi_x + A_2 J\xi_y + A_3 J\xi_z), \\ \hat{A}_2 &= (A_1 J\eta_x + A_2 J\eta_y + A_3 J\eta_z), \\ \hat{A}_3 &= (A_1 J\zeta_x + A_2 J\zeta_y + A_3 J\zeta_z). \end{aligned}$$

By inserting (86) and (88) into (85), we obtain

$$\begin{aligned} & 2\int_{D_{\bar{x}}} \tilde{u}^T \tilde{u}_t dxdydz \\ & + \oint_{\Gamma_{\bar{\xi}}} (\tilde{u}^T(\hat{A}_1)\tilde{u}, \tilde{u}^T(\hat{A}_2)\tilde{u}, \tilde{u}^T(\hat{A}_3)\tilde{u}) \cdot \mathbf{n}_{\bar{\xi}} ds_{\bar{\xi}} - \epsilon\left(\oint_{\Gamma_{\bar{\xi}}} 2\tilde{u}^T \hat{F}^V ds_{\bar{\xi}} - DI\right) \\ & = \|\tilde{u}\|_t^2 + \oint_{\Gamma_{\bar{\xi}}} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \begin{pmatrix} (\hat{A}_1, \hat{A}_2, \hat{A}_3) \cdot \mathbf{n}_{\bar{\xi}} & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} - DI \\ & = \|\tilde{u}\|_t^2 + \oint_{\Gamma_{\bar{\xi}}} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \hat{\mathbf{A}} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} - DI = 0. \end{aligned} \tag{89}$$

The form (89) is completely similar to the one in the $\bar{x}$-system. As mentioned earlier, the domain in $\bar{\xi}$-space is a cube. Hence, $\mathbf{n}_{\bar{\xi}}$ is particularly simple. It is a unit vector in the coordinate directions, $\pm e_\xi, \pm e_\eta, \pm e_\zeta$, on the boundary of the computational domain, $0 \leq \xi \leq 1$, $0 \leq \eta \leq 1$, $0 \leq \zeta \leq 1$. The full formulation for the cube is

$$\begin{aligned}
\|\tilde{u}\|_t^2 - & \int_{\xi=0} \begin{pmatrix} \tilde{u} \\ \hat{\mathbf{F}}_{\mathbf{v}} \end{pmatrix} \begin{pmatrix} \hat{A}_1 & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} \\
& + \int_{\xi=1} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \begin{pmatrix} \hat{A}_1 & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} \\
& - \int_{\eta=0} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \begin{pmatrix} \hat{A}_2 & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} \\
& + \int_{\eta=1} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \begin{pmatrix} \hat{A}_2 & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} \\
& - \int_{\zeta=0} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \begin{pmatrix} \hat{A}_3 & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} \\
& + \int_{\zeta=1} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \begin{pmatrix} \hat{A}_3 & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} = DI.
\end{aligned} \tag{90}$$

Note that $ds_{\bar{\xi}}$ is different in the different coordinate directions. As a last step we will express one of the integrals in (90) in $\bar{x}$- space. Consider, for example,

$$\begin{aligned}
& - \int_{\xi=0} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \begin{pmatrix} \hat{A}_1 & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} \\
&= \int_{\xi=0} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} \begin{pmatrix} -A_1 J\xi_x - A_2 J\xi_y - A_3 J\xi_z & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \hat{F}^V \end{pmatrix} ds_{\bar{\xi}} \\
&= \int_{\xi=0} \begin{pmatrix} \tilde{u} \\ \frac{\hat{F}^V}{JT_1} \end{pmatrix} \begin{pmatrix} -A_1 \frac{\xi_x}{T_1} - A_2 \frac{\xi_y}{T_1} - A_3 \frac{\xi_z}{T_1} & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \frac{\hat{F}^V}{JT_1} \end{pmatrix} JT_1 ds_{\bar{\xi}} \\
&= \int_{\xi=0} \begin{pmatrix} \tilde{u} \\ \frac{\hat{F}^V}{JT_1} \end{pmatrix} \begin{pmatrix} A_1 n_1 + A_2 n_2 + A_3 n_3 & -\epsilon I \\ -\epsilon I & 0 \end{pmatrix} \begin{pmatrix} \tilde{u} \\ \frac{\hat{F}^V}{JT_1} \end{pmatrix} JT_1 ds_{\bar{\xi}},
\end{aligned} \tag{91}$$

where $T_1 = \sqrt{(\xi_x)^2 + (\xi_y)^2 + (\xi_z)^2}$ and $n_1^2 + n_2^2 + n_3^2 = 1$. In fact, (n_1, n_2, n_3) is equal to the normal in the $\bar{x}$- system. This is easily seen by the following. Denote by $\mathbf{r} = (x, y, z)$ a position vector in space. The unnormalized normal vector at $\xi = 0$ is

$$\begin{aligned}
& \frac{\partial \mathbf{r}}{\partial \eta} \times \frac{\partial \mathbf{r}}{\partial \zeta} = (x_\eta, y_\eta, z_\eta) \times (x_\zeta, y_\zeta, z_\zeta) \\
&= (y_\eta z_\zeta - z_\eta y_\zeta, -(x_\eta z_\zeta - z_\eta x_\zeta), x_\eta y_\zeta - y_\eta x_\zeta) = JT_1(n_1, n_2, n_3),
\end{aligned} \tag{92}$$

where (76) and (77) have been used. Hence the matrices appearing in (91) and (23) are equal. Next, we will show that the vectors in (91) and (23) are also equal. We have

$$\begin{aligned}
\frac{\hat{F}^V}{JT_1} &= \frac{\hat{F}_v \cdot 1 + \hat{G}_v \cdot 0 + \hat{H}_v \cdot 0}{JT_1} = \frac{\hat{F}_v}{JT_1} \\
&= \frac{(\xi_x \tilde{F}_v + \xi_y \tilde{G}_v + \xi_z \tilde{H}_v)}{T_1} = \tilde{F}_v n_1 + \tilde{G}_v n_2 + \tilde{H}_v n_3 = \tilde{F}^V.
\end{aligned}$$

At last, we find

$$ds_{\bar{x}} = \left| \frac{\partial \mathbf{r}}{\partial \eta} \times \frac{\partial \mathbf{r}}{\partial \zeta} \right| ds_{\bar{\xi}} = JT_1 ds_{\bar{\xi}}, \tag{93}$$

implying that (91) and (23) are equal.

The other boundaries can be treated similarly. To summarize, we have shown that the relations in $\bar{x}$-space are completely equivalent to those in $\bar{\xi}$-space.

REFERENCES

[1] J. S. Hesthaven and D. Gottlieb, *A stable penalty method for the compressible Navier–Stokes equations*: I. *Open boundary conditions*, SIAM J. Sci. Comput., 17 (1996), pp. 579–612.

[2] J. S. Hesthaven, *A stable penalty method for the compressible Navier–Stokes equations*: II. *One-dimensional domain decomposition schemes*, SIAM J. Sci. Comput., 18 (1997), pp. 658–685.

[3] J. S. Hesthaven, *A stable penalty method for the compressible Navier–Stokes equations*: III. *Multi-dimensional domain decomposition schemes*, SIAM J. Sci. Comput., 20 (1998), pp. 62–93.

[4] J. Nordström, *The influence of open boundary conditions on the convergence to steady state for the Navier–Stokes equations*, J. Comput. Phys., 85 (1989), pp. 210–244.

[5] B. Gustafsson and A. Sundström, *Incompletely parabolic systems in fluid dynamics*, SIAM J. Appl. Math., 35 (1978), pp. 343–357.

[6] J. Nordström, *The use of characteristic boundary conditions for the Navier-Stokes equations*, Comput. & Fluids, 24 (1995), pp. 609–623.

[7] B. Gustafsson, H.-O. Kreiss, and J. Oliger, *Time Dependent Problems and Difference Methods*, John Wiley & Sons, New York, 1995.

[8] T. Hagstrom, *Radiation boundary conditions for the numerical simulation of waves*, Acta Numer., 8 (1999), pp. 47–106.

[9] S. V. Tsynkov, *Numerical solution of problems on unbounded domains. A review*, Appl. Numer. Math., 27 (1998), pp. 465–532.

[10] S. Abarbanel and D. Gottlieb, *Optimal time splitting for two- and three-dimensional Navier-Stokes equations with mixed derivatives*, J. Comput. Phys., 41 (1981), pp. 1-43.

[11] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1996.

[12] H.-O. Kreiss and J. Lorenz, *Initial Boundary Value Problems and the Navier–Stokes Equations*, Academic Press, New York, 1989.

[13] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1990.

[14] J. C. Strikwerda, *Initial boundary value problems for incompletely parabolic systems*, Comm. Pure Appl. Math., 30 (1977), pp. 797–822.

[15] M. Svärd and J. Nordström, *Well Posed Boundary Conditions for the Navier–Stokes Equations*, Technical report 2003-052, Department of Information Technology, Uppsala University, Uppsala, Sweden, 2003.

SIAM J. Numer. Anal.
Vol. 43, No. 3, pp. 1256–1275

SEQUENTIAL CONTROL VARIATES FOR FUNCTIONALS OF MARKOV PROCESSES*

EMMANUEL GOBET[†] AND SYLVAIN MAIRE[‡]

Abstract. Using a sequential control variates algorithm, we compute Monte Carlo approximations of solutions of linear partial differential equations connected to linear Markov processes by the Feynman–Kac formula. It includes diffusion processes with or without absorbing/reflecting boundary and jump processes. We prove that the bias and the variance decrease geometrically with the number of steps of our algorithm. Numerical examples show the efficiency of the method on elliptic and parabolic problems.

Key words. sequential Monte Carlo, Feynman–Kac formula, variance reduction

AMS subject classifications. 65C, 65B, 65M70

DOI. 10.1137/040609124

1. Introduction.

We are concerned with the numerical evaluation of $\mathbf{E}(\Psi(X_s : s \geq t)|X_t = x)$, where $(X_t)_t$ is a Markov process (with linear dynamics) and where Ψ belongs to a class of functionals related to Feynman–Kac representations. These issues arise, for example, in physics in the computations of the solution of diffusion equations (see [CDL$^+$89]), or in finance in the pricing of European options (see [DG95] and the references therein). Monte Carlo methods are usually used to evaluate these expectations for high-dimensional problems or when the functionals are complex. They give a rather poor approximation because of a slow convergence as $\sigma/\sqrt{M}$, M being the number of simulations and σ^2 the relative variance. A better accuracy can nevertheless be reached by using relevant variance-reduction tools like, for instance, the control variates method or importance sampling [Hal70, New94]. One of the most performing tools is the sequential Monte Carlo approach which consists in using iteratively these variance-reduction ideas [Hal62, Hal70, Boo89]. Using, respectively, importance sampling and control variates, this approach has been recently developed in [BCP00] for Markov chains and in [Mai03] for the numerical integration of multivariate smooth functions. We have introduced in [GM04] a sequential Monte Carlo method to solve the Poisson equation with Dirichlet boundary conditions over square domains. This method was based on Feynman–Kac computations of pointwise solutions combined with a global approximation on Tchebychef polynomials [BM97]. Pointwise solutions were computed using walk-on-spheres (WOS) simulations of stopped Brownian motion, which induces a simulation error due to the absorption layer thickness. We have nevertheless observed a geometric reduction up to a limit of both the simulation error and the variance with the number of steps of the algorithm. The global error was comparable to standard deterministic spectral methods [BM97] while avoiding the resolution of a linear system. Our goal here is twofold:

- to extend the scope of the approach to general Markov processes connected to linear elliptic and parabolic Dirichlet problems;

*Received by the editors May 27, 2004; accepted for publication (in revised form) January 20, 2005; published electronically September 23, 2005.
http://www.siam.org/journals/sinum/43-3/60912.html

[†]Ecole Polytechnique, Centre de Mathématiques Appliquées, 91128 Palaiseau Cedex, France (emmanuel.gobet@polytechnique.fr).

[‡]ISITV, Université de Toulon et du Var, Avenue G. Pompidou, BP56, 83262 La Valette du Var Cedex, France (maire@univ-tln.fr).

- to analyze mathematically the convergence of both the bias and the variance. This will be achieved for general discretization schemes for the stochastic processes and also for general global approximations of the solution (not only using Tchebychef polynomials). We also emphasize two major improvements compared to [BCP00] where analogous geometric convergences are proved for Markov chains. First we incorporate in our analysis the influence of the discretization error on the underlying process. Second we allow the global solution not to be in the right approximation space.

In section 2, we make a complete study of the algorithm on elliptic problems with general boundary conditions. At each step of the algorithm, the Monte Carlo computation of $\mathbf{E}(\Psi(X_s : s \geq 0)|X_0 = x)$ at some points x in the domain is required. Then, we build a global approximation using the values at these points. This approximation is used as a control variate at the next step and so on. If the discretization step is small enough, we first prove that the error on the mean value of the global solution reduces geometrically up to a limit directly linked to the approximation error of the exact solution. If furthermore the number of drawings at each step is large enough, we also prove that the variance of the solution reduces in the same way. The proofs of convergence mainly rely on independence properties of the different simulations, on the connection with a linear partial differential equation (PDE), and on the linearity of the functionals with respect to (w.r.t.) the data. This means that the algorithm can be used for Brownian stochastic differential equations (SDEs) with or without absorbing/reflecting boundary, or for Lévy-driven SDEs. The last two sections describe the practical implementation of the algorithm. We first make a discussion on the discretization schemes and on the approximation problems. We then give numerical examples on elliptic and parabolic problems after having precisely studied the speed of convergence of the algorithm on the relative approximation bases. The numerical results confirm the efficiency of the method and the phenomenon of geometric convergence on both the bias and the variance up to a limit.

2. Statement of the problem.

2.1. Elliptic problems. Before giving a general formulation, we prefer listing relevant examples. The Markov process underlying our study is denoted by $X(x) = [X_t(x)]_{t\geq 0}$ and its initial value x belongs to a domain $D \subset \Re^d$. The functionals $\Psi(X_t(x) : t \geq 0) := \Psi(f, g, X(x))$ are related to Feynman–Kac formulas and represented by two continuous functions f and g, respectively defined on $\bar{D}$ and its boundary ∂D. We especially consider the following.

Ex. 1: Brownian SDEs [RY94]: $X_t = x + \int_0^t b(X_s)ds + \int_0^t \sigma(X_s)dW_s$, where W is a Brownian motion. Set $Z_t = e^{-\int_0^t c(X_r)dr}$ and $\tau = \inf\{t \geq 0 : X_t \notin D\}$: we can take $\Psi(f, g, X(x)) = g(X_\tau)Z_\tau - \int_0^\tau f(X_s)Z_s ds$ (provided that $\tau < +\infty$ a.s.). If $D = \Re^d$, we may consider $\Psi(f, g, X(x)) = -\int_0^\infty f(X_s)Z_s ds$ $(g \equiv 0)$ (see [Fre85]).

Ex. 2: SDEs with reflection on ∂D in the nontangential direction γ [Fre85]: $X_t = x+\int_0^t b(X_r)dr+\int_0^t \sigma(X_r)dW_r+\int_0^t \gamma(X_r)d\Lambda_r$, where X_t takes values in $\bar{D}$ and where the so-called local time $(\Lambda_t)_t$ is increasing only when X_t is on ∂D. Set $Z_s = e^{-\int_0^s c(X_r)dr-\int_0^s \lambda(X_r)d\Lambda_r}$: we can take $\Psi(f, g, X(x)) = -\int_0^\infty f(X_s)Z_s ds - \int_0^\infty g(X_s)Z_s d\Lambda_s$.

Ex. 3: Lévy-driven SDEs (see [BL84, GM92]): $X_t = x+\int_0^t b(X_r)dr+\int_0^t \sigma(X_r)dW_r+ \int_0^t \int_{\Re^p} \Upsilon(X_r, z)\mu(dr, dz)$, where μ is a martingale normalized Poisson measure on $\Re^p$, with Lévy measure $m(dz)$. We can take $\Psi(f, g, X(x)) =$

$-\int_0^\tau f(X_s)Z_s ds$ ($g \equiv 0$) where τ and Z are defined as in the Brownian case. Note that reflections could be included [BL84]. Similar situations also occur with transport equations [CDL$^+$89].

The processes above are well defined under, for example, Lipschitz assumptions on their coefficients. For the first two cases, the relative infinitesimal generator is given by $L_0\phi = \sum_{i=1}^d b_i(x)\partial_{x_i}\phi + \frac{1}{2}\sum_{i,j=1}^d [\sigma\sigma^*(x)]_{i,j}\partial^2_{x_i,x_j}\phi$. For the last one, jumps are taken into account by an extra integral kernel. Thus, the associated infinitesimal generator is defined by $L_0\phi(y) + \int_{\Re^p}\left[\phi(y+z) - \phi(y) - z\cdot\nabla\phi(y)\mathbf{1}_{|z|\leq 1}\right]M(y,dz)$ (here the measure $M(y,\cdot)$ is defined by $M(y,A) = m\{z : \Upsilon(y,z) \in A\}$). The key point is that these operators are linear. The goal of this section is to describe how to evaluate efficiently the quantity

$$u(x) = \mathbf{E}\big(\Psi(f,g,X(x))\big). \tag{2.1}$$

We assume the following.

(H1) The process X, the domain D, and the data (f,g) are such that $\Psi(f,g,X(x))$ is a linear map w.r.t. the data (f,g) and that $\mathbf{V}\mathrm{ar}(\Psi(f,g,X(x))) < +\infty$.

This assumption is natural in view of the previous examples. Usually, it imposes restrictions on the domain, on the sign of c arising in Z. See Proposition 4.1 for explicit conditions about absorbed Brownian SDEs. We now assume that u solves an elliptic PDE with appropriate boundary conditions, using the connection between Markov processes and PDEs.

(H2) The process X, the domain D, and the data (f,g) are such that u is a classic[1] solution of

$$\begin{cases}\mathcal{A}u = f & \text{in } D,\\ \mathcal{B}u = g & \text{on } \partial D,\end{cases} \tag{2.2}$$

where $\mathcal{A}$ and $\mathcal{B}$ are second order linear operators.

The domain may be bounded or unbounded in some cases, the diffusion processes may be elliptic or hypo-elliptic. We refer the reader to [Fri75, BL84, Fre85] for details and references. For the first example, we have $\mathcal{A}u = L_0u - cu$ and $\mathcal{B}u = u$ (Dirichlet boundary condition), for the second example $\mathcal{B}u = \nabla_x u.\gamma - \lambda u$ (Neumann boundary condition). Second order operators $\mathcal{B}$ arise with Ventcel's boundary conditions corresponding to processes having a diffusion part on ∂D (see [Cat92]). These different boundary conditions can also be mixed.

2.2. Parabolic problems. The extension to problems with a terminal time T is straightforward. Formally, it is achieved by considering the time-space process $(t, X_t)_t$ in the domain $]0,T[\times D$. Then the operator $\mathcal{A}$ has to replaced by $\partial_t + \mathcal{A}$. In that case, we can take $D = \Re^d$ (this is the so-called Cauchy problem). The coefficients of X and the domain D may also be time-dependent [BL84, Lie96]. The reader can check that the following algorithm and its convergence proof are derived in the same way.

3. Study of the algorithm.

3.1. Description. We now describe our algorithm, which computes iterative approximations $(u_n)_{n\geq 0}$ of the global solution u. These approximations rely on the computations of $\mathbf{E}[\Psi(\tilde{f},\tilde{g},X(x))]$ (for data $\tilde{f}$ and $\tilde{g}$ possibly different from f and g) at some points $(x_i)_{1\leq i\leq N}$.

[1] The regularity of u depends on the type of the boundary condition.

Initialization. We begin with $u_0 \equiv 0$.

***Iteration n*, step 1.** Assume that an approximated solution u_{n-1} of class $C^2(\bar{D})$ is built at stage $n-1$ and that the representation $u_{n-1}(x)=\mathbf{E}(\Psi[\mathcal{A}u_{n-1},\mathcal{B}u_{n-1},X(x)])$ holds (which simply means that u_{n-1} solves (2.2) with $f=\mathcal{A}u_{n-1}$ and $g=\mathcal{B}u_{n-1}$). The idea is to compute a correction $y_n = u - u_{n-1}$ on this approximation. Using (2.1), we have

$$y_n(x) = \mathbf{E}(\Psi[f-\mathcal{A}u_{n-1}, g-\mathcal{B}u_{n-1}, X(x)]). \tag{3.1}$$

In the above equation, the expectation is relative to the law of X and not to the law of u_{n-1}, which can be random. We intend to compute a Monte Carlo approximation of $y_n(x_i)$. For this, we replace the simulations of the random variable $\Psi[f-\mathcal{A}u_{n-1}, g-\mathcal{B}u_{n-1}, X(x_i)]$ by $\Psi[f-\mathcal{A}u_{n-1}, g-\mathcal{B}u_{n-1}, X^{\Delta}(x_i)]$ using a suitable discretization procedure $X^{\Delta}(x_i)$ for the stochastic process $X(x_i)$. For the moment, we prefer keeping quite abstract notations concerning the discretization scheme, since mild assumptions are required (see assumption (H4) below). We just mention that Δ usually represents the discretization parameter which tends to 0 (for instance, for the WOS procedure [Sab91], Δ is the space step; for the Euler procedure [CPS98, Gob01], Δ is the time step). Consequently, $y_n(x_i)$ is approximated by

$$\bar{y}_n(x_i) = \frac{1}{M}\sum_{m=1}^{M} \Psi[f-\mathcal{A}u_{n-1}, g-\mathcal{B}u_{n-1}, X^{\Delta,n,m}(x_i)] \tag{3.2}$$

using M independent simulations of the paths $X^{\Delta,n,m}(x_i)$. They are also generated independently of everything else. In fact, the independence of simulations at different points is not crucial to ensure the convergence of the algorithm. Nevertheless, we think that dependent drawings slow down the convergence of the method and are less adapted to parallel computations.

***Iteration n*, step 2.** In order to build a global approximation y_n based on the values $(\bar{y}_n(x_i))_i$, we use a linear approximation [CHQZ88, BM97]. The linear approximation of a function $w(\cdot)$ at some points $(x_j)_j$ can always be written

$$\mathcal{P}w(x) = \sum_{j=1}^{N} w(x_j)\mathcal{C}_j(x) \tag{3.3}$$

for some functions $(\mathcal{C}_j)_j$. In addition we assume a stability property:

$$\mathcal{P}[\mathcal{P}w] = \mathcal{P}w \quad \text{for any function } w. \tag{3.4}$$

If we use an interpolation, the functions $(\mathcal{C}_j)_j$ simply verify $\mathcal{C}_j(x_i) = \delta_{i,j}$. This is, for example, the case of interpolation in dimension one on Lagrange polynomials $L_j(x) = \frac{\prod_{i\neq j}(x-x_j)}{\prod_{i\neq j}(x_i-x_j)}$. This approximation can also come from a problem of fitting an approximation model $\sum_{k=1}^{K}\alpha_k\varphi_k$ on some basis functions $(\varphi_k)_k$ to the values $(w(x_i))_i$. This leads to a discrete least-squares problem [Bjö96], using the norm associated to the discrete inner product $\langle u,v\rangle_\mu = \sum_{j=1}^{N}\mu_j v(x_j)u(x_j)$ for some positive weights $(\mu_j)_j$, which consists in the minimization of the squared norm $\|\sum_{k=1}^{K}\alpha_k\varphi_k - w\|_\mu^2$. The optimal coefficients $(\alpha_k)_k$ are hence solutions of a linear system $A\alpha = b$ with $A_{ik} =$

$\langle \varphi_i, \varphi_k \rangle_\mu$, $b_k = \langle w, \varphi_k \rangle_\mu$. As $\alpha = A^{-1}b$, we get $\alpha_k = \sum_{i=1}^N A^{-1}_{ik} \sum_{j=1}^N \mu_j \varphi_k(x_j) w(x_j)$ and we are still in a linear form of type (3.3) in letting

$$\mathcal{C}_j(x) = \mu_j \sum_{k=1}^K \sum_{i=1}^N A^{-1}_{ik} \varphi_k(x_j) \varphi_k(x).$$

If the $(\varphi_k)_k$ are, moreover, orthonormal w.r.t. $\langle \cdot, \cdot \rangle_\mu$, then we simply have $\mathcal{C}_j(x) = \mu_j \sum_{k=1}^K \varphi_k(x_j)\varphi_k(x)$. A slightly different situation is the computation of the projection of the function w on orthonormal polynomials $(T_n)_n$ w.r.t. the inner product $\langle u, v \rangle_\nu = \int_{[a,b]} v(x)u(x)\nu(x)dx$, where ν is a positive weight function on the interval $D = [a, b]$. We have $w(x) \simeq \sum_{k=0}^N \alpha_k T_k(x)$ with $\alpha_k = \langle w, T_k \rangle_\nu$ and the points $(x_i)_i$ are used to build quadrature formulas to compute accurately the coefficients $(\alpha_k)_k$. These points are usually chosen as the zeros of T_{N+1} which makes the quadrature formula exact for all polynomials of degree $\leq 2N+1$. Note that in this case, this approximation is equal to the interpolation at the same points. Another possibility is the Gauss–Lobatto formulas where the boundaries of the interval are chosen as quadrature points. In higher dimensions, the approximations are built using tensor products. In any case the approximations are still linear and they will be described in detail in section 4. Note that the stability property (3.4) holds for all of these approximations. Once one of the above approximations has been chosen we just write

$$u_n = u_{n-1} + \mathcal{P}\bar{y}_n = u_{n-1} + \sum_{j=1}^N \bar{y}_n(x_j)\mathcal{C}_j \tag{3.5}$$

and we can proceed to the next iteration. Furthermore, we assume the following throughout the remainder of the paper.

(H3) The functions $(\mathcal{C}_j)_{1 \leq j \leq N}$ are of class $C^2(\bar{D})$. Furthermore, for any $x \in D$ we have $\mathbf{V}\mathrm{ar}(\Psi(\mathcal{A}\mathcal{C}_j, \mathcal{B}\mathcal{C}_j, X(x))) < +\infty$ and

$$\mathbf{E}(\Psi(\mathcal{A}\mathcal{C}_j, \mathcal{B}\mathcal{C}_j, X(x))) = \mathcal{C}_j(x). \tag{3.6}$$

In other words, $\mathcal{C}_j$ (formally) solves (2.2) with data $f = \mathcal{A}\mathcal{C}_j$ and $g = \mathcal{B}\mathcal{C}_j$. Hence, $\mathcal{P}w \in C^2(\bar{D})$ for any function w. In particular, u_n is of class $C^2(\bar{D})$ for any n and satisfies the representation $u_n(x) = \mathbf{E}(\Psi[\mathcal{A}u_n, \mathcal{B}u_n, X(x)])$, which makes our algorithm valid. Note also that the stability property (3.4) written for u_n gives

$$\mathcal{P}u_n = u_n \quad \text{for any } n \geq 0. \tag{3.7}$$

Note that the numerical computations of $\mathcal{A}u_{n-1}$ in (3.2) (and analogously those of $\mathcal{B}u_{n-1}$) are performed using the evaluations of $(\mathcal{A}\mathcal{C}_j)_j$ through the equality $\mathcal{A}u_{n-1} = \sum_{j=1}^N u_{n-1}(x_j)\mathcal{A}\mathcal{C}_j$ (see [GM04]).

3.2. Convergence results.

Our goal is now to estimate the convergence of

$$m_n := \sup_{1 \leq i \leq N} |\mathbf{E}(u_n(x_i) - u(x_i))|, \quad v_n := \sup_{1 \leq i \leq N} \mathbf{V}\mathrm{ar}(u_n(x_i)). \tag{3.8}$$

It is possible to derive other measures of the error, like $\frac{1}{N}\sum_{i=1}^N |\mathbf{E}(u_n(x_i) - u(x_i))|$, without major differences. However, in this work, rather than finding the optimal way to measure the error, we prefer studying in detail the convergence. We need to introduce some extra notation regarding the scheme X^{Δ}. For a deterministic smooth function $\tilde{g}$ we set

$$\begin{aligned} e(\tilde{g},\Delta,x) &= \mathbf{E}\big(\Psi[\mathcal{A}\tilde{g},\mathcal{B}\tilde{g},X^{\Delta}(x)]\big)-\mathbf{E}\big(\Psi[\mathcal{A}\tilde{g},\mathcal{B}\tilde{g},X(x)]\big),\\ V(\tilde{g},\Delta,x) &= \mathbf{V}\mathrm{ar}\big(\Psi[\mathcal{A}\tilde{g},\mathcal{B}\tilde{g},X^{\Delta}(x)]\big). \end{aligned}$$

We state mild assumptions on the discretization scheme X^{Δ}, which allows great generality on the procedures that can be used.

(H4) (1) The map $(\tilde{f},\tilde{g})\mapsto\Psi(\tilde{f},\tilde{g},X^{\Delta}(x))$ is linear.
(2) The discretization errors $[e(\mathcal{C}_j,\Delta,x_i)]_{1\le i,j\le N}$ converge to 0 as $\Delta\to 0$.
(3) The variances $[V(u,\Delta,x_i)]_i$ and $[V(\mathcal{C}_k,\Delta,x_i)]_{i,k}$ are finite.
(4) The latter are uniformly bounded for Δ close to 0:
$\limsup_{\Delta\to 0} V(\mathcal{C}_k,\Delta,x_i)<\infty$ for any i and k.

The first assumption is natural since the initial map $(\tilde{f},\tilde{g})\mapsto\Psi(\tilde{f},\tilde{g},X(x))$ is linear. The second one is minimal since it requires only the weak convergence of the discretization scheme for the $C^2(\bar{D})$-functions $(\mathcal{C}_k)_k$. The last two ones are also very natural since they are satisfied for X (see (H1)–(H3)). A practical verification of (H4) will be given in section 4.1. It easily follows from statement (3) that $V(u-\mathcal{P}u,\Delta,x_i)<+\infty$ for any i. This justifies the finiteness of the terms which appear in Theorem 3.2.

We first state a convergence result for the bias.

Theorem 3.1. *Assume* (H1)–(H2)–(H3)–(H4). *Then, for any $n\ge 1$, one has*

$$m_n \le \rho_m\ m_{n-1}+\sup_{1\le i\le N}|[\mathcal{P}u-u](x_i)+\mathcal{P}[e(u-\mathcal{P}u,\Delta,\cdot)](x_i)|, \tag{3.9}$$

where $\rho_m=\sup_{1\le i\le N}\big[\sum_{j=1}^N|\mathcal{P}[e(\mathcal{C}_j,\Delta,\cdot)](x_i)|\big]$. For Δ small enough, one has $\rho_m<1$. Thus, the convergence of $(m_n)_n$ is geometric at rate ρ_m, up to a threshold equal to

$$\limsup_n m_n \le \frac{1}{1-\rho_m}\sup_{1\le i\le N}|[\mathcal{P}u-u](x_i)+\mathcal{P}[e(u-\mathcal{P}u,\Delta,\cdot)](x_i)|. \tag{3.10}$$

The upper limit for the bias strongly depends on the quality of the approximation of u by the operator $\mathcal{P}$. Note that if u is in the right approximation space ($\mathcal{P}u\equiv u$), the first term on the right-hand side of (3.10) cancels and the bias m_n converges geometrically to 0. In other words, even if the simulations are biased because of Δ, the bias vanishes at the limit. This is a surprising and very interesting phenomenon. However, unlike the direct Monte Carlo procedure, there is no guarantee that $\lim_{\Delta\to 0}\limsup_n m_n=0$, except in the case of the interpolation operator $\mathcal{P}$ (i.e., $\mathcal{P}u(x_i)=u(x_i)$ for any x_i). We now state the convergence of the variance $(v_n)_n$.

Theorem 3.2. *Assume* (H1)–(H2)–(H3)–(H4) *and set*

$$\begin{aligned} C(\Delta,N) &= 2\sup_{1\le i\le N}\sum_{j=1}^N \mathcal{C}_j^2(x_i)\Big[\sum_{k=1}^N\sqrt{V(\mathcal{C}_k,\Delta,x_j)}\ \Big]^2,\\ \rho_v &= \sup_{1\le i\le N}\left(\sum_{j=1}^N|\mathcal{P}[e(\mathcal{C}_j,\Delta,\cdot)](x_i)|\right)^2+\frac{C(\Delta,N)}{M}. \end{aligned}$$

Then, for any $n\ge 1$, one has

$$v_n\le\rho_v v_{n-1}+\frac{1}{M}\Big\{2\sup_{1\le i\le N}\sum_{j=1}^N\mathcal{C}_j^2(x_i)V(u-\mathcal{P}u,\Delta,x_j)+C(\Delta,N)m_{n-1}^2\Big\}. \tag{3.11}$$

For Δ small enough and M large enough, one has $\rho_v < 1$. Thus, the convergence of $(v_n)_n$ is geometric at rate ρ_v, up to a threshold equal to

$$\limsup_n v_n \le \frac{1}{(1-\rho_v)M}\left\{2\sup_{1\le i\le N}\sum_{j=1}^{N}\mathcal{C}_j^2(x_i)V(u-\mathcal{P}u,\Delta,x_j)+C(\Delta,N)\limsup_n m_n^2\right\}. \tag{3.12}$$

Note that when $\rho_v < 1$, $\rho_m < 1$, so that the convergence holds simultaneously for the bias and for the variance. As for the bias, if $\mathcal{P}u = u$, then $\limsup_n m_n = 0$ and thus $\limsup_n v_n = 0$: the variance v_n converges geometrically to 0, provided that $1/\Delta$ and M are large enough.

3.3. Proofs of convergence. To make the distinction between what is simulated before stage n and at stage n, we define the usual conditional expectations and variances

$$\mathbf{E}^{n-1}(Y) = \mathbf{E}\big(Y\big|\sigma(X^{\Delta,n',m}(x_i) : 1 \le n' \le n-1; 1 \le m \le M; 1 \le i \le N)\big)$$

and $\mathbf{V}\text{ar}^{n-1}(Y) = \mathbf{E}^{n-1}(Y^2) - [\mathbf{E}^{n-1}(Y)]^2$. Note that the construction of the algorithm yields that the discretized processes $[X^{\Delta,n,m}(x_i)]_{m,i,n}$ are independent.

3.3.1. Proof of Theorem 3.1. Formula (3.10) is a straightforward consequence of (3.9). Before proving (3.9), we transform the expression of $u_n(x_i)$ for a fixed x_i. Using (3.5), (3.2), and the PDE solved by u, we get $u_n(x_i) = u_{n-1}(x_i) + \frac{1}{M}\sum_{j=1}^{N}\sum_{m=1}^{M}\Psi[\mathcal{A}(u-u_{n-1}),\mathcal{B}(u-u_{n-1}),X^{\Delta,n,m}(x_j)]\mathcal{C}_j(x_i)$. In view of (3.7), note that

$$u - u_{n-1} = u - \mathcal{P}u + \sum_{k=1}^{N}(u-u_{n-1})(x_k)\mathcal{C}_k \tag{3.13}$$

and that $\Psi[\mathcal{A}(u-u_{n-1}),\mathcal{B}(u-u_{n-1}),X^{\Delta,n,m}(x_j)]$ equals

$$\Psi[\mathcal{A}(u-\mathcal{P}u),\mathcal{B}(u-\mathcal{P}u),X^{\Delta,n,m}(x_j)] + \sum_{k=1}^{N}(u-u_{n-1})(x_k)\Psi[\mathcal{A}\mathcal{C}_k,\mathcal{B}\mathcal{C}_k,X^{\Delta,n,m}(x_j)],$$

because of the linearity of $\Psi[(\cdot,\cdot),X^{\Delta}(x)]$ under (H4). Thus, we obtain

$$\begin{aligned} u_n(x_i) = u_{n-1}(x_i) + \frac{1}{M}\sum_{j=1}^{N}\sum_{m=1}^{M}&\Bigg[\Psi[\mathcal{A}(u-\mathcal{P}u),\mathcal{B}(u-\mathcal{P}u),X^{\Delta,n,m}(x_j)] \\ &+ \sum_{k=1}^{N}(u-u_{n-1})(x_k)\Psi[\mathcal{A}\mathcal{C}_k,\mathcal{B}\mathcal{C}_k,X^{\Delta,n,m}(x_j)]\Bigg]\mathcal{C}_j(x_i). \end{aligned} \tag{3.14}$$

Computation of $\mathbf{E}^{n-1}$ $(u_n(x_i))$. As $[X^{\Delta,n,m}(x_j)]_{m,j}$ is independent of u_{n-1}, we readily get

$$\begin{aligned} \mathbf{E}^{n-1}(u_n(x_i)) = u_{n-1}(x_i) &+ \sum_{j=1}^{N}\mathbf{E}\left(\Psi[\mathcal{A}(u-\mathcal{P}u),\mathcal{B}(u-\mathcal{P}u),X^{\Delta}(x_j)]\right)\mathcal{C}_j(x_i) \\ &+ \sum_{k=1}^{N}(u-u_{n-1})(x_k)\sum_{j=1}^{N}\mathbf{E}\left(\Psi[\mathcal{A}\mathcal{C}_k,\mathcal{B}\mathcal{C}_k,X^{\Delta}(x_j)]\right)\mathcal{C}_j(x_i). \end{aligned} \tag{3.15}$$

Note that

(a) using (2.1) and (3.6), we have $\mathbf{E}\big(\Psi[\mathcal{A}(u-\mathcal{P}u),\mathcal{B}(u-\mathcal{P}u),X^{\Delta}(x_j)]\big) = e(u-\mathcal{P}u,\Delta,x_j)+(u-\mathcal{P}u)(x_j)$;

(b) using (3.6), we have $\mathbf{E}\left(\Psi[\mathcal{A}\mathcal{C}_k,\mathcal{B}\mathcal{C}_k,X^{\Delta}(x_j)]\right) = e(\mathcal{C}_k,\Delta,x_j)+\mathcal{C}_k(x_j)$;

(c) owing to (3.4), we have $\sum_{j=1}^{N}\mathcal{C}_k(x_j)\mathcal{C}_j(x_i)=\mathcal{C}_k(x_i)$ (indeed, a choice of $w(\cdot)$ such that $w(x_i)=\delta_{i,k}$ leads to $\mathcal{P}w=\mathcal{C}_k$, and thus $\mathcal{P}\mathcal{C}_k=\mathcal{C}_k$ by (3.4)).

Plugging these identities in (3.15), it readily follows that

$$\begin{aligned}\mathbf{E}^{n-1}(u_n(x_i)) &= u_{n-1}(x_i)+\sum_{j=1}^{N}\big[e(u-\mathcal{P}u,\Delta,x_j)+(u-\mathcal{P}u)(x_j)\big]\mathcal{C}_j(x_i)\\ &\quad+\sum_{k=1}^{N}(u-u_{n-1})(x_k)\left[\mathcal{C}_k(x_i)+\sum_{j=1}^{N}e(\mathcal{C}_k,\Delta,x_j)\mathcal{C}_j(x_i)\right]\\ &= \mathcal{P}u(x_i)+\mathcal{P}[e(u-\mathcal{P}u,\Delta,\cdot)](x_i)+\sum_{k=1}^{N}(u-u_{n-1})(x_k)\mathcal{P}[e(\mathcal{C}_k,\Delta,\cdot)](x_i), \qquad (3.16)\end{aligned}$$

simplifications at the last line arising from (3.4).

Computation of $\mathbf{E}(u_n(x_i))$. Taking the expectation in (3.16) we obtain

$$\begin{aligned}\mathbf{E}\big(u_n(x_i)-u(x_i)\big) &= \mathcal{P}u(x_i)-u(x_i)+\mathcal{P}[e(u-\mathcal{P}u,\Delta,\cdot)](x_i)\\ &\quad+\sum_{k=1}^{N}\mathbf{E}((u-u_{n-1})(x_k))\mathcal{P}[e(\mathcal{C}_k,\Delta,\cdot)](x_i).\end{aligned}$$

It remains to take absolute values and the supremum over i on both sides to complete the proof of (3.9). □

3.3.2. Proof of Theorem 3.2. Note that the inequality $\rho_v<1$ holds for Δ small enough and M large enough. Indeed, under (H4) $C(\Delta,N)$ remains uniformly bounded w.r.t. Δ close to 0. We prove only (3.11). Taking some fixed x_i, we have

$$\mathbf{V}\mathrm{ar}(u_n(x_i)) = \mathbf{V}\mathrm{ar}\big[\mathbf{E}^{n-1}(u_n(x_i))\big]+\mathbf{E}\big[\mathbf{V}\mathrm{ar}^{n-1}(u_n(x_i))\big]. \qquad (3.17)$$

Computation of $\mathbf{Var}[\mathbf{E}^{n-1}(u_n(x_i))]$. In view of (3.16), we have

$$\begin{aligned}\mathbf{V}\mathrm{ar}\big[\mathbf{E}^{n-1}(u_n(x_i))\big] &= \mathbf{V}\mathrm{ar}\left[\sum_{j=1}^{N}u_{n-1}(x_j)\mathcal{P}[e(\mathcal{C}_j,\Delta,\cdot)](x_i)\right]\\ &\le v_{n-1}\left(\sum_{j=1}^{N}|\mathcal{P}[e(\mathcal{C}_j,\Delta,\cdot)](x_i)|\right)^2, \qquad (3.18)\end{aligned}$$

where we have used the standard inequality

$$\begin{aligned}\mathbf{V}\mathrm{ar}\left(\sum_{j=1}^{N}\alpha_jY_j\right) &= \sum_{j_1,j_2=1}^{N}\alpha_{j_1}\alpha_{j_2}\mathbf{C}\mathrm{ov}(Y_{j_1},Y_{j_2})\\ &\le \sum_{j_1,j_2=1}^{N}|\alpha_{j_1}|\,|\alpha_{j_2}|\sqrt{\mathbf{V}\mathrm{ar}(Y_{j_1})}\sqrt{\mathbf{V}\mathrm{ar}(Y_{j_2})} = \left[\sum_{j=1}^{N}|\alpha_j|\sqrt{\mathbf{V}\mathrm{ar}(Y_j)}\right]^2 \qquad (3.19)\end{aligned}$$

for any real numbers $(\alpha_j)_j$ and any square integrable real random variables $(Y_j)_j$.

Computation of $\mathbf{Var}^{n-1}(u_n(x_i))$. We invoke the independence of $[X^{\Delta,n,m}(x_i)]_{m,i}$ and u_{n-1} in (3.14) to derive

$$\begin{aligned}\mathbf{V}\mathrm{ar}^{n-1}(u_n(x_i)) &= \sum_{j=1}^{N}\frac{\mathcal{C}_j^2(x_i)}{M}\mathbf{V}\mathrm{ar}^{n-1}\Bigg[\Psi[\mathcal{A}(u-\mathcal{P}u),\mathcal{B}(u-\mathcal{P}u),X^{\Delta}(x_j)]\\ &\qquad+\sum_{k=1}^{N}(u-u_{n-1})(x_k)\Psi[\mathcal{A}\mathcal{C}_k,\mathcal{B}\mathcal{C}_k,X^{\Delta}(x_j)]\Bigg]\\ &\le \sum_{j=1}^{N}\frac{\mathcal{C}_j^2(x_i)}{M}\Bigg[\sqrt{V(u-\mathcal{P}u,\Delta,x_j)}+\sum_{k=1}^{N}|u-u_{n-1}|(x_k)\sqrt{V(\mathcal{C}_k,\Delta,x_j)}\Bigg]^2,\end{aligned}$$

using (3.19) at the last inequality. Applying the inequality $\mathbf{E}(\alpha_0+\sum_{k=1}^N\alpha_kY_k)^2\le 2\alpha_0^2+2\big(\sum_{k=1}^N|\alpha_k|\sqrt{\mathbf{E}(Y_k^2)}\big)^2$, which can be proved as (3.19), we get

$$\begin{aligned}\mathbf{E}(\mathbf{V}\mathrm{ar}^{n-1}(u_n(x_i))) &\le 2\sup_{1\le i\le N}\sum_{j=1}^{N}\frac{\mathcal{C}_j^2(x_i)}{M}\Bigg\{V(u-\mathcal{P}u,\Delta,x_j)\\ &\qquad+\sup_{1\le k\le N}\mathbf{E}((u-u_{n-1})^2(x_k))\Bigg[\sum_{k=1}^{N}\sqrt{V(\mathcal{C}_k,\Delta,x_j)}\Bigg]^2\Bigg\}.\end{aligned}$$

Combine this estimate with (3.18) and (3.17), use $\sup_{1\le k\le N}\mathbf{E}((u-u_{n-1})^2(x_k))\le v_{n-1}+m_{n-1}^2$, and take the supremum over i to complete the proof of (3.11). □

4. Influence of parameters of the algorithm.

We focus mainly on the first example of Brownian SDEs with a Dirichlet boundary condition. This means (see section 2) that $X_t=x+\int_0^t b(X_s)ds+\int_0^t\sigma(X_s)dW_s$, $\Psi(f,g,X(x))=g(X_\tau)Z_\tau-\int_0^\tau f(X_s)Z_sds$ (with $Z_t=e^{-\int_0^t c(X_r)dr}$ and $\tau=\inf\{t\ge 0: X_t\notin D\}$), $\mathcal{A}u=\sum_{i=1}^d b_i\partial_{x_i}u+\frac{1}{2}\sum_{i,j=1}^d[\sigma\sigma^*]_{i,j}\partial^2_{x_i,x_j}u-cu$, and $\mathcal{B}u=u$. We first give a set of explicit assumptions implying (H1)–(H2) and (3.6) in this case.

(H2′) (i) The functions b and σ are Lipschitz continuous on $\bar{D}$ and $\sigma\sigma^*(x)\ge\epsilon_0\mathrm{I}_d$ uniformly w.r.t. $x\in D$ ($\epsilon_0>0$).

(ii) D is a bounded domain and each point of its boundary ∂D satisfies the *exterior cone condition*: for any $x\in\partial D$, there exists a finite right circular cone $\mathcal{K}$, with vertex x, such that $\bar{\mathcal{K}}\cap\bar{D}=\{x\}$.

(iii) The function g is continuous on ∂D, f and $c\ge 0$ are uniformly Hölder continuous in $\bar{D}$ (with exponent $\alpha\in]0,1[$).

PROPOSITION 4.1. *Under* (H2′), (H1) *and* (H2) *are fulfilled. Furthermore, if* $\mathcal{C}_j$ *is of class* $C^2(\bar{D})$, (H3) *holds.*

Proof. The variance in (H1) is finite. Indeed, on the one hand the functions g and f are bounded and c is nonnegative. On the other hand, we have $\mathbf{E}_x(\tau)<+\infty$ (which automatically induces exponential moments for τ; see [Fre85, section 3.3]).

The proof of (H2) is somewhat classic except that D is not smooth here. We just recall the two main steps. First, the existence of a solution to (2.2) follows from the remark after [GT83, Theorem 6.13], noting that under (H2′) every point of the boundary has a barrier (see Problem 6.3 in [GT83]). Second, [Fre85, Theorem 2.1, section 2.2] states that the solution is given by (2.1). This is achieved by applying Itô's formula to u and using some careful localization procedures because derivatives of u explode near the boundary (see also Appendix B). (H3) is proved analogously. □

4.1. Verification of assumption (H4). We propose to check it when we use the so-called *discrete Euler scheme* [Gob00], which is the simplest procedure that can be used for general stopped diffusions. An alternative is the WOS scheme, which is especially efficient when we are dealing with the Brownian motion (see [HMG03] and the references therein). Some refinements to the discrete Euler scheme are also possible, using Brownian bridge simulations [Bal95, Gob00, Gob01].

For a given time step Δ and discretization times $t_k = k\Delta$, the Euler scheme is defined by $X_0^\Delta = x$ and $X_{t_{k+1}}^\Delta - X_{t_k}^\Delta = b(X_{t_k}^\Delta)(t_{k+1} - t_k) + \sigma(X_{t_k}^\Delta)(W_{t_{k+1}} - W_{t_k})$, which can be written in continuous time as

$$X_t^\Delta = x + \int_0^t b(X_{\phi(s)}^\Delta)ds + \int_0^t \sigma(X_{\phi(s)}^\Delta)dW_s. \tag{4.1}$$

Here, $\phi(s) = t_k$ for $t_k \leq s < t_{k+1}$. The approximated exit time is defined by $\tau^\Delta = \inf\{t_k \geq 0 : X_{t_k}^\Delta \notin D\}$. Thus, to approximate $\Psi(f, g, X(x))$, we simply propose

$$\Psi(f, g, X^\Delta(x)) = g(X_{\tau^\Delta}^\Delta)Z_{\tau^\Delta}^\Delta - \int_0^{\tau^\Delta} f(X_{\phi(s)}^\Delta)Z_{\phi(s)}^\Delta ds, \quad Z_s^\Delta = e^{-\int_0^s c(X_{\phi(u)}^\Delta)du}. \tag{4.2}$$

Here, g is evaluated at $X_{\tau^\Delta}^\Delta$ which is not a priori on ∂D: hence, in (4.2) g has to be understood as a bounded continuous function on the whole space. In view of (4.2), (H4)-(1) is clearly fulfilled. To verify (2)–(4) of (H4), our main tool is the following theorem, which is original in this context of elliptic problems and whose proof is postponed to Appendix A.

THEOREM 4.2. *Assume* (H2′). *Then, the following assertions hold.*

(a) *For any* $p \geq 1$, $\sup_{x\in\bar{D}} \mathbf{E}_x(\tau^p) + \limsup_{\Delta\to 0} \sup_{x\in\bar{D}} \mathbf{E}_x([\tau^\Delta]^p) < \infty$.

(b) $\lim_{\Delta\to 0} \tau^\Delta = \tau$ *in* $\mathbf{L}_p$ *for any* $p \geq 1$.

(c) $\lim_{\Delta\to 0} X_{\tau^\Delta}^\Delta = X_\tau$ *in probability.*

(d) *For any bounded continuous* γ, $\limsup_{\Delta\to 0} \int_0^{\tau^\Delta\wedge\tau} |\gamma(X_s) - \gamma(X_{\phi(s)}^\Delta)|ds = 0$ *in* $\mathbf{L}_p$ *for any* $p \geq 1$.

Since $|\Psi[\mathcal{A}u, u, X^\Delta(x)]| = |\Psi[f, u, X^\Delta(x)]| \leq C(1+\tau^\Delta)$ and $|\Psi[\mathcal{A}\mathcal{C}_k, \mathcal{C}_k, X^\Delta(x)]| \leq C(k)(1 + \tau^\Delta)$, we easily get that (H4)-(3) and (H4)-(4) are fulfilled in view of (a). It remains to prove that $e(\mathcal{C}_k, \Delta, x) = \mathbf{E}\big(\Psi[\mathcal{A}\mathcal{C}_k, \mathcal{C}_k, X^\Delta(x)] - \Psi[\mathcal{A}\mathcal{C}_k, \mathcal{C}_k, X(x)]\big)$ converges to 0 when $\Delta \to 0$, for any $\mathcal{C}_k$ of class $C^2(\bar{D})$. Using (c), we get the convergence of $\mathcal{C}_k(X_{\tau^\Delta}^\Delta)$ to $\mathcal{C}_k(X_\tau)$ in probability, and thus in $\mathbf{L}_1$ since $\mathcal{C}_k$ is bounded. Since $\exp(\cdot)$ is 1-Lipschitz on $\Re^-$ and $c(\cdot)$ is nonnegative, we have $|Z_{\tau^\Delta}^\Delta - Z_\tau| \leq |\int_0^{\tau^\Delta} c(X_{\phi(u)}^\Delta)du - \int_0^\tau c(X_u)du|$, which converges to 0 in $\mathbf{L}_1$, using (b) and (d). For the convergence in $\mathbf{L}_1$ of $\int_0^{\tau^\Delta} f(X_{\phi(s)}^\Delta)Z_{\phi(s)}^\Delta ds$ to $\int_0^\tau f(X_s)Z_s ds$, the previous arguments apply and this completes the verification of (H4)-(2).

4.2. Impact of the approximation operator. The discretization parameter Δ has to be chosen small enough to ensure the convergence of the bias. This convergence depends on the approximation operator as we must have $\rho_m < 1$. As it is mainly described by the sensitivity of regular functions to the discretization error, it actually depends very little on the approximation operator. The convergence of the variance (described by the condition $\rho_v < 1$) depends a lot more on the choice of the approximation, but in the same way concerning the discretization parameter. In order to study this convergence, we can hence focus on the case $\Delta = 0$. In this ideal case, we have

$$v_n \leq v_{n-1}\frac{C(0, N)}{M} + \frac{2}{M}\sup_{1\leq i\leq N}\sum_{j=1}^N \mathcal{C}_j^2(x_i)V(u - \mathcal{P}u, 0, x_j)$$

with

$$C(0,N) = 2 \sup_{1\le i\le N} \sum_{j=1}^{N} \mathcal{C}_j^2(x_i) \left(\sum_{k=1}^{N} \sqrt{V(\mathcal{C}_k, 0, x_j)} \right)^2 .$$

The quantities $V(u-\mathcal{P}u, 0, x_j)$ and $V(\mathcal{C}_k, 0, x_j)$ can be computed as

$$V(u-\mathcal{P}u,0,x_j) = \mathbf{E}_{x_j}\left[\int_0^\tau Z_s^2 \left|\nabla_x(u-\mathcal{P}u)\sigma\right|^2 (X_s)ds\right],$$
$$V(\mathcal{C}_k,0,x_j) = \mathbf{E}_{x_j}\left[\int_0^\tau Z_s^2 \left|\nabla_x \mathcal{C}_k\sigma\right|^2 (X_s)ds\right],$$

using Lemma B.1 given in Appendix B. The first term enables us to control the final error and the second one the speed of convergence of the algorithm. Both depend only on the gradient of the basis functions and not on the second derivatives of these functions. It is quite difficult to make a general discussion on the optimal choice of the approximation in a general domain. We prefer focusing on polynomial interpolations on square domains and give explicit computations of the convergence parameters in this case. The process X remains general.

4.3. Gauss–Tchebychef interpolation on $]-1,1[^d$. The Tchebychef polynomials $T_n(x) = \cos(n \arccos(x))$ are the orthogonal polynomials on $]-1,1[$ with respect to the inner product $\langle P,Q\rangle_w = \int_{-1}^{1} P(x)Q(x)w(x)dx$, where $w(x) = \frac{1}{\sqrt{1-x^2}}$. We have $\|T_0\|_w^2 = \pi$ and $\|T_n\|_w^2 = \frac{\pi}{2}$ if $n \ge 1$. In dimension one, the interpolation polynomial $\mathcal{P}_N(u)$ of the function u at the Tchebychef abscissae

$$x_k = \cos\left(\frac{2k+1}{N+1}\frac{\pi}{2}\right), \qquad k = 0,1,\dots,N,$$

is given by [Bjö96]

$$\mathcal{P}_N(u) = \sum_{n=0}^{N} \alpha_n T_n$$

with

$$\alpha_n = \frac{\pi}{\|T_n\|_w^2 (N+1)} \sum_{k=0}^{N} u(x_k) T_n(x_k).$$

This interpolation is optimal w.r.t. the sup norm. The control of u and of its derivative is given in the following theorem (see [CHQZ88, p. 298]).

THEOREM 4.3. *Denote by H_w^m the w-weighted Sobolev space with regularity $m \in \mathbf{N}^*$. Then $\exists\, c_1, c_2 > 0$ such that $\forall u \in H_w^m$, we have*

$$\|u - \mathcal{P}_N(u)\|_w \le c_1 N^{-m} \|u\|_{H_w^m}, \qquad \|u - \mathcal{P}_N(u)\|_{H_w^1} \le c_2 N^{2-m} \|u\|_{H_w^m} .$$

The Tchebychef interpolation of a function $u : D =]-1,1[^d \mapsto \mathbf{R}$ is built using the same process as in dimension one. The interpolation polynomial $\mathcal{P}_N(u)$ at the N^d points of a tensored Tchebychef grid and evaluated at $z = (z_1,\dots,z_d)$ is

$$\mathcal{P}_N(u)(z) = \sum_{n_1,\dots,n_d=1}^{N} \alpha_{n_1,\dots,n_d} T_{n_1}(z_1)\cdots T_{n_d}(z_d),$$

where the $\alpha_{n_1,\dots,n_d}$ are defined by

$$\alpha_{n_1,\dots,n_d} = \prod_{i=1}^{d}\left(\frac{\pi}{\|T_{n_i}\|_w^2(N+1)}\right)\sum_{m_1,\dots,m_d=1}^{N} u(x_{m_1},\dots,x_{m_d})T_{n_1}(x_{m_1})\cdots T_{n_d}(x_{m_d}).$$

The quality of this interpolation is exactly the same as that in Theorem 4.3. In dimension one, the basis functions $\mathcal{C}_k$ write

$$\mathcal{C}_k(x) = \sum_{n=0}^{N}\frac{\pi}{\|T_n\|_w^2(N+1)}T_n(x_k)T_n(x).$$

As $T_n^{'}(x) = \frac{-n\sin(n\arccos(x))}{\sqrt{1-x^2}}$ and because $\frac{\pi}{\|T_n\|_w^2(N+1)}T_n(x_k) \leq \frac{2}{N+1}$, we have $|\nabla_x\mathcal{C}_k(x)|^2 \leq [\sum_{n=0}^{N}\frac{2}{N+1}\frac{n}{\sqrt{1-x^2}}]^2 = \frac{N^2}{1-x^2} \leq \frac{N^2}{1-|x|}$. Using $Z_s^2 \leq 1$, the occupation time formula [RY94], and Lemma 4.4 (which is proved in Appendix C), we finally have $V(\mathcal{C}_k, 0, x_j) \leq \int_{-1}^{1}|\nabla_x\mathcal{C}_k(y)|^2\mathbf{E}_{x_j}(L_\tau^y(X))dy \leq CN^2$.

Lemma 4.4. *For $D =]-1,1[$ and under* (H2$'$)*, we have* $\mathbf{E}_x(L_\tau^y(X)) \leq C(1-|y|)$ *uniformly in $x \in D$.*

As furthermore $\mathcal{C}_k(x_j) = \delta_{k,j}$, we have

$$C(0,N) = 2\sup_{0\leq j\leq N}\left(\sum_{k=0}^{N}\sqrt{V(\mathcal{C}_k,0,x_j)}\right)^2 \leq 2\left(\sum_{k=0}^{N}\sqrt{C}N\right)^2 = O(N^4).$$

This means that we need to take $M \geq CN^4$ (with C large enough) to ensure the convergence. Using the same tools, we can also prove that if $u \in H_w^m$,

$$V(u-\mathcal{P}_N(u),0,x_j) \leq C\int_{-1}^{1}|\nabla_x(u-\mathcal{P}_N(u))|^2(y)dy \leq C\|u-\mathcal{P}_N(u)\|_{H_w^1}^2 = O(N^{-2m+4})$$

and so that the final error on the solution is a

$$O\left(\sqrt{\sup_{1\leq i\leq N}\sum_{j=1}^{N}\mathcal{C}_j^2(x_i)\frac{V(u-\mathcal{P}u,0,x_j)}{M}}\right) = O(N^{-m}).$$

We now go back to the interpolation on the multidimensional Tchebychef grid. In this case, the basis functions simply write $\mathcal{C}_{k_1,\dots,k_d}(x_1,\dots,x_d) = \mathcal{C}_{k_1}(x_1)\cdots\mathcal{C}_{k_d}(x_d)$. As $|\nabla_x\mathcal{C}_{k_1,\dots,k_d}(x)|^2 \leq C\sum_{i=1}^{d}|\nabla_x\mathcal{C}_{k_i}(x_i)|^2$, we deduce immediately that $V(\mathcal{C}_{k_1,\dots,k_d},0,x) = O(N^2)$ and that $C(0,N) = O(N^{2+2d})$. The error estimates, using the order m of regularity of u, are the same in dimension d as they are in dimension one. As the convergence is geometric and the solution is computed at N^d points with a source term constituted of N^d terms, the complexity of the algorithm is essentially $C(0,N)N^{2d}$. The upper bound on $C(0,N)$ may not be tight. We shall especially see that in all the numerical experiments $C(0,N)$ is a lot smaller than N^{2+2d}, and we should also keep in mind that the spectral methods are used for very smooth solutions so that N is usually small. As a comparison, the usual spectral method requires us to solve a linear system of size N^d which involves a complexity of a $O(N^{3d})$ using a direct method and of N^{2d} at each step of an iterative method. The resolution induces moreover an additional error on the solution due to the condition number of the matrix which can grow very quickly with N [BM97]. This can also make the speed of convergence of the iterative method quite slow. A big advantage of our method is that it keeps all the advantages of the Monte Carlo method in terms of parallel computing. One can, for example, use one processor for each of the N^d points of the grid.

4.4. Gauss–Lobatto–Tchebychef interpolation. To reach a slightly better accuracy, one can also use the Gauss–Lobatto points [BM97]

$$y_k = \cos\left(\frac{N-k}{N}\pi\right), \qquad k = 0, 1, \dots, N,$$

where the boundaries of the interval are taken as interpolation points. The coefficients (β_n) of the interpolation polynomial $\mathcal{Q}_N(u) = \sum_{n=0}^{N} \beta_n T_n$ satisfy the relation $\beta_n \|T_n\|_w^2 = \int_{-1}^{1} \mathcal{Q}_N(u)(x) T_n(x) w(x) dx$. For $n < N$, the integral equals

$$\frac{\pi}{N}\left(\frac{u(-1)T_n(-1) + u(1)T_n(1)}{2} + \sum_{k=1}^{N-1} u(y_k) T_n(y_k)\right)$$

using the relative quadrature formula, which is exact on polynomials of degree smaller than $2N-1$. This gives the values of β_n for $n < N$. Moreover, as $T_n(1) = 1$, we have $\beta_N = u(1) - \sum_{n=0}^{N-1} \beta_n$. Hence, the basis functions write

$$\mathcal{C}_N = T_N + \sum_{n=0}^{N-1} \frac{\pi(T_n - T_N)}{2N\|T_n\|_w^2}, \qquad \mathcal{C}_0 = \sum_{n=0}^{N-1} \frac{\pi(-1)^n (T_n - T_N)}{2N\|T_n\|_w^2}$$

and if $j \neq 0, N$,

$$\mathcal{C}_j = \sum_{n=0}^{N-1} \frac{\pi T_n(y_j)}{N\|T_n\|_w^2}(T_n - T_N).$$

As for the Gauss–Tchebychef case, the d-dimensional extension on tensored domains is obtained by setting $\mathcal{C}_{k_1,\dots,k_d}(x_1, \dots, x_d) = \mathcal{C}_{k_1}(x_1) \cdots \mathcal{C}_{k_d}(x_d)$. This interpolation enables a better control of the derivative of u than the previous one, as we can see in the following theorem [BM97] (valid in any dimension).

THEOREM 4.5. *For all $m \in \mathbf{N}^*$, $\exists c_1, c_2 > 0$ such that $\forall u \in H_w^m$, we have*

$$\|u - \mathcal{Q}_N(u)\|_w \leq c_1 N^{-m} \|u\|_{H_w^m}, \qquad \|u - \mathcal{Q}_N(u)\|_{H_w^1} \leq c_2 N^{1-m} \|u\|_{H_w^m}.$$

Using the previous tools, one can easily prove that $C(0, N) = O(N^{2+2d})$ and $V(u - \mathcal{P}u, 0, x_j) = O(N^{-2m+2})$ so that the final error on the solution is an $O(N^{-m-1})$ which is compared to an $O(N^{-m})$ for the previous interpolation.

5. Numerical results.

The aim of this numerical part is not to give the optimal way to solve a general problem using our algorithm. We study various classic situations only to illustrate the convergence and the accuracy of our algorithm. Different approximations and discretization schemes are tested to confirm our theoretical estimates and the great efficiency and generality of our approach.

5.1. Poisson equation in dimension one. Our first example is the numerical resolution of the Poisson equation in dimension one using a Monte Carlo scheme with no discretization error [GM04] and Tchebychef interpolations on either the $(x_k)_k$ or the $(y_k)_k$. The solution of this Poisson equation

$$\frac{1}{2}u'' = f \qquad \text{in }]-1, 1[$$

with boundary conditions $u(-1) = a, u(1) = b$ is $u(x) = a\mathbf{P}_x(W_{\tau_D} = -1) + b(1 - \mathbf{P}_x(W_{\tau_D} = -1)) - \mathbf{E}_x(\int_0^{\tau_D} f(W_s)ds)$. As $\mathbf{P}_x(W_{\tau_D} = -1) = \frac{1-x}{2}$ the contribution of the boundary conditions to the solution can be easily simulated. To achieve the contribution of the source term with no discretization error, we use the representation $\mathbf{E}_x(\int_0^{\tau_D} f(W_s)ds) = (1-x^2)\mathbf{E}(f(Y_x))$ where the density of the random variable Y_x is $\frac{1+r}{1+x}1_{-1\leq r\leq x} + \frac{1-r}{1-x}1_{x\leq r\leq 1}$.

We study the example $f(x) = (x+2)\exp(x)/2, a = -\frac{1}{e}, b = e$, so that the solution of this equation is $u(x) = x\exp(x)$. We give in the following table the error

$$e(j) = \sup_{0\leq k\leq N} |u(x_k) - u_j(x_k)|$$

as a function of the number j of steps and of the number M of sample values to compute the pointwise approximation at the x_k. Even if our algorithm is based on independent random drawings, we have observed in [GM04] that one could use low-discrepancy sequences to speed up the convergence of the algorithm. We hence use here a version of the algorithm based on Halton sequences, which is twice as fast as the Monte Carlo version. The accuracy of the crude quasi–Monte Carlo method with M sample values is given by $e(1)$. L is the number of steps until convergence for a given value of M. All the CPU times are less than one second.

N	M	L	$e(1)$	$e(5)$	$e(L)$
5	80	16	2×10^{-1}	2×10^{-2}	5×10^{-4}
7	200	30	1×10^{-1}	9×10^{-3}	4×10^{-6}
10	800	45	9×10^{-2}	2×10^{-3}	8×10^{-10}

M has to be chosen large enough with respect to N to make the algorithm converge but is significantly smaller than N^4. The error at convergence $e(L)$ corresponds exactly to the interpolation error of the interpolation polynomial $\mathcal{P}_N(u)$ of the exact solution at the Tchebychef abscissae. We now study the same example using the Gauss–Lobatto–Tchebychef interpolation.

N	M	L	$e(1)$	$e(5)$	$e(L)$
5	40	10	4×10^{-1}	1×10^{-3}	7×10^{-4}
7	100	26	2×10^{-1}	2×10^{-2}	2×10^{-6}
10	400	35	1×10^{-1}	1×10^{-2}	2×10^{-9}

We can see that we can take half as many drawings to achieve the same final accuracy for similar number of steps. This means that the convergence is twice as fast using this kind of interpolation, maybe because there is no error on the correction at the two boundary points. We do not notice any major difference on the final error, as we might have expected.

5.2. The bidimensional case. We consider the Poisson equation on the square domain $D =]-1,1[^2$ with Dirichlet boundary conditions. We use an interpolation at the bidimensional Tchebychef grid, two types of discretization schemes, and Monte Carlo simulations. The first one is the modified WOS method [HMG03, GM04] which can take into account the source term f. This walk goes from one sphere to another until the motion reaches the ε-absorption layer. The second one is based on the continuous Euler scheme with parameter $\triangle t$ [Gob01]. We study the equation $\frac{1}{2}\triangle u = \frac{1}{2}(4x^2+3)\exp(x^2+y)$ with Dirichlet boundary conditions chosen so that the solution of this equation is $u(x, y) = \exp(x^2 + y)$. We begin with the WOS scheme taking, respectively, $\varepsilon_1 = 10^{-2}$ and $\varepsilon_2 = 10^{-3}$ for the absorption layer thickness and use the same notation as in the previous examples.

N	M	L	$e_1(1)$	$e_1(L)$	CPU	$e_2(1)$	$e_2(L)$	CPU
6	200	11	0.24	1×10^{-3}	1.4	0.18	1×10^{-3}	2.4
8	400	13	0.17	8×10^{-5}	6.7	0.14	7×10^{-5}	10.5
10	800	17	0.13	3×10^{-6}	28	0.14	4×10^{-6}	44

We observe a geometric convergence on both the bias and the variance up to the interpolation error of the exact solution. We do not notice any difference on the final accuracy for the two values ε_1 and ε_2. CPU times are of course smaller for ε_1 and also about eight times smaller than in [GM04] by using a recursive computation of the Tchebychef polynomials instead of their trigonometric expression. We now use the corrected Euler scheme with discretization parameters $\triangle t_1 = 0.005$ and $\triangle t_2 = 0.002$.

N	M	L	$e_1(1)$	$e_1(L)$	CPU	L	$e_2(1)$	$e_2(L)$	CPU
6	30	10	0.7	3×10^{-3}	1.4	12	0.6	2×10^{-3}	4
8	100	40	0.37	2×10^{-4}	28	10	0.38	1×10^{-4}	16
10	200	10	0.4	5	27	50	0.3	3×10^{-5}	300

We observe the same kind of convergence as in the previous examples except in the case $\triangle t_1 = 0.005$ and $N = 10$ because the discretization parameter is not small enough (see the condition $\rho_v < 1$). Nevertheless, we do not achieve the same accuracy at the limit. A bias due to the discretization scheme still remains. When $\triangle t$ decreases this bias decreases and the convergence is faster. Using the naive Euler scheme on the same example, the bias at the limit was twice as big and the convergence twice as slow. We can take M a lot smaller than using the WOS method to achieve the convergence.

5.3. Parabolic problem. We consider a regular up and out call option with maturity $T = 1$, corresponding to the domain $D =]0, 2[$ (actually, 0 is a natural boundary) and the Cauchy–Dirichlet boundary condition $g(t, x) = (x - 1)_+$ if $t = 1$ and $g(t, x) = 1$ if $x = 2$. The dynamics is given by $X_t = x \exp(\sigma W_t - \frac{1}{2}\sigma^2 t)$. The quantity $u(0, x) = \mathbf{E}_x((X_\tau - 1)_+)$ (with $\tau = \inf\{t : X_t \notin]0, 2[\} \wedge 1$) gives the risk-neutral price of the option at time 0, when the interest rate equals 0. The solution u can be computed by a closed formula [Zha97]. The derivatives of u have singularities around $(t, x) = (1, 1)$. Thus, the solution is less smooth than in previous examples and the numerical results below show that the benefit of our method is less important in that case. For the interpolation procedure, we propose a piecewise linear interpolation w.r.t. the time variable and a Tchebychef interpolation w.r.t. the space variable. The interpolation times are $(t_i = iT/(K - 1))_{0 \leq i < K}$ and the $N + 1$ Tchebychef points at each interpolation time are the (x_n) or (y_n) (on the interval $[0, 2]$ instead of $[-1, 1]$).

Notation relative to the errors remains the same. We first compare the accuracy of the Gauss–Tchebychef (GT) and Gauss–Lobatto–Tchebychef (GLT) interpolations, in the case $K = 5$, $N = 5$. The simulation of X can be exactly performed at discretization times $(k\triangle t)$, from which we can derive a naive approximation of τ. Here, we take $\triangle t = 0.05$ and $M = 10$ simulated paths. All the CPU times are less than one second.

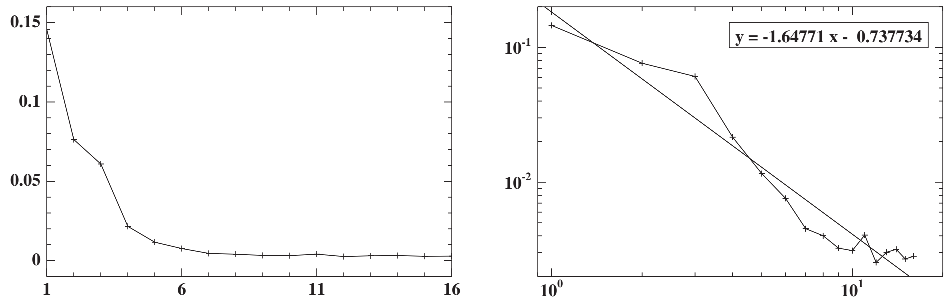


FIG. 5.1. *On the left, error $e(j)$ w.r.t. the number of iterations. On the right, the same in logarithmic scales.*

	K	N	M	$\triangle t$	L	$e(1)$	$e(2)$	$e(L)$
GT	5	5	10	0.05	4	1.32×10^{-1}	7.92×10^{-2}	3.88×10^{-2}
GLT	5	5	10	0.05	4	1.72×10^{-1}	4.7×10^{-2}	1.44×10^{-2}

The GLT interpolation converges faster and the relative final error is slightly smaller. Note also that from the first iteration, the accuracy of our method compared to a crude Monte Carlo method is improved by a factor of 3.6 using the GLT interpolation: this accuracy could be achieved using 13 times more crude Monte Carlo simulations. To diminish the bias, we now use Brownian bridge corrections to simulate the exit event. To obtain more evidence of the better convergence of the GLT interpolation, we increase the value of N to 15: the contraction constant ρ_v should increase (from Theorem 3.2 and sections 4.2 and 4.3, one has $\rho_v \approx O(\frac{N^4}{M}))$ and may become larger than 1 (to recover $\rho_v < 1$, one can increase M). This is confirmed by the following result, where for the GLT interpolation, the convergence still holds but not for the GT interpolation.

	K	N	M	$\triangle t$	L	$e(1)$	$e(2)$	$e(L)$
GT	5	15	10	0.01	–	1.98×10^{-1}	5.64×10^{-1}	2.7 for $L = 15$
GLT	5	15	10	0.01	10	2.13×10^{-1}	1.24×10^{-1}	5.65×10^{-3}

In view of this nice behavior, the following experiments are done using GLT interpolation. Figure 5.1 illustrates the geometric convergence till the interpolation error. It has been obtained with $K = 10$, $N = 20$, $M = 20$, and $\triangle t = 0.001$. The optimal choice of (K, N) will be analyzed in future research.

6. Conclusion. We developed and studied a sequential Monte Carlo method for the numerical solution of linear PDEs. This method provides a regular global approximation of this solution by combining pointwise approximations via the Feynman–Kac formula and a linear approximation on some basis functions. As the pointwise approximations are computed by means of a Monte Carlo method, statistical and discretization errors occur. We have proved the geometric reduction of these two kinds of errors up to a limit linked to the linear approximation. Numerical experiments on simple diffusion equations using various discretization schemes and different kinds of approximations have confirmed this geometric reduction and the efficiency of our method. Further numerical examples should be developed on more complex domains, in higher dimensions, or for less regular solutions. In higher dimensions, one needs to diminish the dimensional effect by making a good choice of the basis functions [Mai03]. For more complex domains or less regular solutions, new versions of the algorithm based on finite elements approximations or domain decomposition methods can certainly be developed. In all those situations, our algorithm could at least be a variance reduction tool by computing a rather poor approximation on few basis functions. As a final remark, we think that our method could advantageously replace the usual deterministic methods in many situations, especially if it is used in a parallel version.

Appendix A. Proof of Theorem 4.2.

Statement (a). Since D is bounded, it is sufficient to prove the result when $D = D_R = [-R, R]^d$ for an arbitrary R. We proceed in several steps.

Step 1. τ^Δ has exponential moments, not necessarily uniformly bounded in Δ. If we use the Markov property at time $(t_k)_k$ for the Euler scheme, we get the rough estimate $\mathbf{P}(\tau^\Delta > m\Delta) \le [\sup_{x\in\bar{D}} \mathbf{P}_x(X^\Delta_\Delta \in D)]^m$. Under (H2′), $\sup_{x\in\bar{D}} \mathbf{P}_x(X^\Delta_\Delta \in D) < 1$, and thus τ^Δ has exponential moments. The point is to prove that they are uniformly bounded w.r.t. Δ, which is not clear from the computations above.

Step 2. The first moment $\sup_{x\in\bar{D}} \mathbf{E}_x(\tau^\Delta)$ is uniformly bounded w.r.t. Δ close to 0. We adapt the arguments from [Fre85], where it is proved that $\sup_{x\in\bar{D}} \mathbf{E}_x(\tau) < \infty$. Take $x \in D$ and set $\mathbf{1} = (1, \ldots, 1)^*$; the ellipticity assumption combined with Itô's formula applied to $e^{\lambda\mathbf{1}.x}$ (for λ large enough such that $-\lambda|b|_\infty d + \frac{1}{2}\lambda^2\epsilon_0 d \ge 1$) gives

$$\begin{aligned}\mathbf{E}_x(e^{\lambda\mathbf{1}.X^\Delta_{\tau^\Delta}}) &\ge e^{\lambda\mathbf{1}.x} + \mathbf{E}_x\left(\int_0^{\tau^\Delta} e^{\lambda\mathbf{1}.X^\Delta_s} ds\right)\\ &\ge e^{\lambda\mathbf{1}.x} + \min_{z\in D_{R+1}} e^{\lambda\mathbf{1}.z}\left[\mathbf{E}_x(\tau^\Delta) - \mathbf{E}_x\left(\int_0^{\tau^\Delta} \mathbf{1}_{X^\Delta_s \notin D_{R+1}} ds\right)\right]. \qquad \text{(A.1)}\end{aligned}$$

On the one hand, we have

$$\mathbf{E}_x(e^{\lambda\mathbf{1}.X^\Delta_{\tau^\Delta}}) \le \sup_{z\in D_{R+1}} e^{\lambda\mathbf{1}.z} + \sum_{k=0}^{\infty} \mathbf{E}_x\big[\mathbf{1}_{t_k<\tau^\Delta}\mathbf{1}_{X^\Delta_{t_{k+1}}\notin D_{R+1}} e^{\lambda\mathbf{1}.X^\Delta_{t_{k+1}}}\big].$$

Standard large deviation estimates (see Lemma 4.1 in [Gob00]) give

$$\mathbf{E}\left(e^{\lambda\mathbf{1}.X^\Delta_{t_{k+1}}}\mathbf{1}_{X^\Delta_{t_{k+1}}\notin D_{R+1}}|\mathcal{F}_{t_k}\right) \le C(\lambda)e^{-c/\Delta}$$

for some constants $c > 0$ and $C(\lambda)$ uniform w.r.t. $\Delta \le 1$ and $X^\Delta_{t_k} \in D_R$. For Δ small enough such that $\frac{C(\lambda)e^{-c/\Delta}}{\Delta} \le \frac{1}{3}\min_{z\in D_{R+1}} e^{\lambda\mathbf{1}.z}$, we obtain

$$\mathbf{E}_x(e^{\lambda\mathbf{1}.X^\Delta_{\tau^\Delta}}) \le \sup_{z\in D_{R+1}} e^{\lambda\mathbf{1}.z} + \mathbf{E}_x(\tau^\Delta)\,\frac{1}{3}\min_{z\in D_{R+1}} e^{\lambda\mathbf{1}.z}.$$

On the other hand, from Fubini's theorem, we have $\mathbf{E}_x(\int_0^{\tau^\Delta} \mathbf{1}_{X^\Delta_s\notin D_{R+1}} ds) = \int_0^\infty \mathbf{P}_x(\phi(s) < \tau^\Delta; X^\Delta_s \notin D_{R+1}) ds$. The previous arguments give $\mathbf{E}_x(\int_0^{\tau^\Delta} \mathbf{1}_{X^\Delta_s\notin D_{R+1}} ds) \le Ce^{-c/\Delta}\int_0^\infty \mathbf{P}_x(\phi(s) < \tau^\Delta) ds \le \frac{1}{3}\mathbf{E}_x\tau^\Delta$ for Δ small enough. Plugging all these estimates into (A.1), we get

$$\sup_{z\in D_{R+1}} e^{\lambda\mathbf{1}.z} \ge e^{\lambda\mathbf{1}.x} + \mathbf{E}_x(\tau^\Delta)\,\frac{1}{3}\min_{z\in D_{R+1}} e^{\lambda\mathbf{1}.z}$$

uniformly in $x \in D$ for Δ small enough. This proves our assertion.

Step 3. The pth moment $\sup_{x\in\bar{D}} \mathbf{E}_x([\tau^\Delta]^p)$ $(p \geq 1)$ is uniformly bounded w.r.t. Δ close to 0. This is a standard consequence of the boundedness of $\sup_{x\in\bar{D}} \mathbf{E}_x(\tau^\Delta)$ and of the Markov property at times $(k\Delta)_k$. We refer to section 3.3 in [Fre85] for a proof in the diffusion case, which can be adapted to the discrete Euler scheme in a straightforward way. Statement (a) is proved.

Statement (b). In view of the uniform integrability conditions (a), it is enough to prove that τ^Δ converges in probability to τ, which follows from the weak convergence of (τ^Δ, τ) to (τ, τ) (stable convergence). Thus, we aim at showing that for any s_1, s_2, we have

$$\lim_{\Delta\to 0} \mathbf{P}_x(\tau^\Delta \leq s_1, \tau \leq s_2) = \mathbf{P}_x(\tau \leq s_1, \tau \leq s_2). \tag{A.2}$$

We introduce the signed distance to ∂D, defined by $F(x) = d(x, \partial D)$ if $x \in D$ and $F(x) = -d(x, \partial D)$ if $x \notin D$. Without additional regularity on D, F is at least a Lipschitz continuous function. Note that $\{\tau^\Delta \leq s_1\} = \{\inf_{t\leq s_1} F(X^\Delta_{\phi(t)}) \leq 0\}$ and $\{\tau \leq s_1\} = \{\inf_{t\leq s_1} F(X_t) \leq 0\}$. From the a.s. uniform convergence of $(X^\Delta_{\phi(t)})_t$ to $(X_t)_t$ on $[0, s_1]$ (see [Gob00], for instance), we have $\lim_{\Delta\to 0} \inf_{t\leq s_1} F(X^\Delta_{\phi(t)}) = \inf_{t\leq s_1} F(X_t)$ a.s. Thus, (A.2) holds true if $0 = \mathbf{P}(\inf_{t\leq s_1} F(X_t) = 0)$, and we obtain, using the strong Markov property,

$$0 = \mathbf{E}(\mathbf{1}_{\tau<s_1} \mathbf{P}_{X_\tau}(\forall t \leq s_1 - \tau : X_t \in \bar{D})) + \mathbf{P}(\tau = s_1). \tag{A.3}$$

In fact, for any $r > 0$ and $x \in \partial D$, under (H2$'$) we have $\mathbf{P}_x(\forall t \leq r : X_t \in \bar{D}) \leq \mathbf{P}_x(X_r \in \bar{D}) < 1$ (in [Gob00], see inequality (68) and the comments before Remark 5.1). Thus, by the Blumenthal zero-one law, the probability $\mathbf{P}_x(\forall t \leq r : X_t \in \bar{D})$ must be equal to 0. Hence, (A.3) is reduced to $\mathbf{P}(\tau = s_1) = 0$. This equality is true except for the countable number of points of discontinuity of $s_1 \mapsto \mathbf{P}(\tau \leq s_1)$, which is enough to derive (A.2) for any s_1.

Statements (c) and (d). Both statements easily follow from the a.s. uniform convergence of X^Δ to X on compact sets, from the uniform integrability in (a) and from the convergence (b). □

Appendix B. A technical lemma.

LEMMA B.1. *Assume* (H2$'$) *and that $\hat{u}$ is the $C^0(\bar{D}) \cap C^2(D)$-solution of $L_0\hat{u} - c\hat{u} = \hat{f}$ in D and $\hat{u} = \hat{g}$ on ∂D, where $\hat{f}$ and $\hat{g}$ are bounded continuous functions. Then*

$$V(\hat{u}, 0, x) = \mathbf{E}_x \left[\int_0^\tau Z_s^2 \ |\nabla_x \hat{u}\ \sigma|^2(X_s) ds \right] < +\infty.$$

Proof. The technical difficulty comes from the fact that $\hat{u}$ may have derivatives exploding near the boundary. To circumvent this problem, set $D_\varepsilon = \{x \in D : d(x, \partial D) > \varepsilon\}$ for $\varepsilon > 0$ and denote by τ_ε the associated exit time. By standard interior estimates [GT83], $\hat{u}$ has a bounded gradient in D_ε. Furthermore, it is straightforward to see that $\tau_\varepsilon \uparrow \tau$ a.s. as $\varepsilon \downarrow 0$. An application of Itô's formula gives

$$\hat{u}(X_{\tau_\varepsilon}) Z_{\tau_\varepsilon} + \int_0^{\tau_\varepsilon} Z_s(-L_0\hat{u} + c\hat{u})(X_s) ds = \hat{u}(x) + \int_0^{\tau_\varepsilon} Z_s\ [\nabla_x \hat{u}\ \sigma](X_s) dW_s.$$

Owing to the localization in D_ε, it is easy to see that $\mathbf{E}_x\big[\int_0^{\tau_\varepsilon} Z_s^2\ |\nabla_x\hat{u}\ \sigma|^2(X_s)ds\big] < \infty$. Hence, by the isometry property, we obtain

$$\mathbf{E}_x\left[\int_0^{\tau_\varepsilon} Z_s^2\ |\nabla_x\hat{u}\ \sigma|^2(X_s)ds\right] = \mathbf{V}\mathrm{ar}_x\left[\hat{u}(X_{\tau_\varepsilon})Z_{\tau_\varepsilon} + \int_0^{\tau_\varepsilon} Z_s(-L_0\hat{u}+c\hat{u})(X_s)ds\right].$$

Take the limit when ε goes to 0: the left-hand side converges using the monotone convergence theorem and the right-hand side using the dominated convergence theorem. The limit writes $\mathbf{E}_x\big[\int_0^{\tau} Z_s^2\ |\nabla_x\hat{u}\ \sigma|^2(X_s)ds\big] = V(\hat{u},0,x)$, which is our statement. □

Appendix C. Proof of Lemma 4.4. We can assume $y \geq 0$. Tanaka's formula [RY94] yields

$$\frac{1}{2}\mathbf{E}_x(L_\tau^y(X)) = \mathbf{E}_x(X_\tau - y)_+ - (x-y)_+ - \mathbf{E}_x\left(\int_0^\tau b(X_s)\mathbf{1}_{X_s\geq y}ds\right).$$

Hence, we get $\mathbf{E}_x(L_\tau^y(X)) \leq C$ uniformly in x, y. Using the occupation time formula in the equality above and the previous uniform estimate, we obtain $\frac{1}{2}\mathbf{E}_x(L_\tau^y(X)) \leq (1-y)+C\int_y^1 \mathbf{E}_x(L_\tau^z(X))dz \leq C(1-y)$. If $y<0$, then the same arguments apply. □

REFERENCES

[Bal95] P. Baldi, *Exact asymptotics for the probability of exit from a domain and applications to simulation,* Ann. Probab., 23 (1995), pp. 1644–1670.

[BCP00] K. Baggerly, D. Cox, and R. Picard, *Exponential convergence of adaptive importance sampling for Markov chains,* J. Appl. Probab., 37 (2000), pp. 342–358.

[Bjö96] Å. Björck, *Numerical Methods for Least Squares Problems,* SIAM, Philadelphia, PA, 1996.

[BL84] A. Bensoussan and J. L. Lions, *Impulse Control and Quasivariational Inequalities μ,* J. M. Cole, transl., Gauthier-Villars, Montrouge, 1984 (in French).

[BM97] C. Bernardi and Y. Maday, *Spectral methods,* in Handbook of Numerical Analysis, Vol. V, North-Holland, Amsterdam, 1997, pp. 209–485.

[Boo89] T. E. Booth, *Zero-variance solutions for linear Monte Carlo,* Nuclear Sci. Engrg., 102 (1989), pp. 332–340.

[Cat92] P. Cattiaux, *Stochastic calculus and degenerate boundary value problems,* Ann. Inst. Fourier, 42 (1992), pp. 541–624.

[CDL+89] M. Cessenat, R. Dautray, G. Ledanois, P. L. Lions, E. Pardoux, and R. Sentis, *Méthodes probabilistes pour les équations de la physique,* Collect. CEA, Eyrolles, 1989.

[CHQZ88] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics,* Springer Ser. Comput. Phys., Springer-Verlag, New York, 1988.

[CPS98] C. Costantini, B. Pacchiarotti, and F. Sartoretto, *Numerical approximation for functionals of reflecting diffusion processes,* SIAM J. Appl. Math., 58 (1998), pp. 73–102.

[DG95] D. Duffie and P. Glynn, *Efficient Monte Carlo simulation of security prices,* Ann. Appl. Probab., 5 (1995), pp. 897–905.

[Fre85] M. Freidlin, *Functional Integration and Partial Differential Equations,* Ann. of Math. Stud. 109, Princeton University Press, Princeton, NJ, 1985.

[Fri75] A. Friedman, *Stochastic Differential Equations and Applications,* Vol. 1, Academic Press, New York, 1975.

[GM92] M. G. Garroni and J. L. Menaldi, *Green Functions for Second Order Parabolic Integro-differential Problems,* Pitman Res. Notes Math. Ser. 275, Longman Scientific & Technical, Harlow, 1992.

[GM04] E. Gobet and S. Maire, *A spectral Monte Carlo method for the Poisson equation,* Monte Carlo Methods Appl., 10 (2004), pp. 275–285.

[Gob00] E. Gobet, *Euler schemes for the weak approximation of killed diffusion,* Stochastic Process. Appl., 87 (2000), pp. 167–197.

[Gob01] E. Gobet, *Euler schemes and half-space approximation for the simulation of diffusions in a domain,* ESAIM Probab. Stat., 5 (2001), pp. 261–297.

[GT83] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order,* 2nd ed., Springer-Verlag, New York, 1983.

[Hal62] J. H. Halton, *Sequential Monte Carlo,* Proc. Cambridge Philos. Soc., 58 (1962), pp. 57–78.

[Hal70] J. H. Halton, *A retrospective and prospective survey of the Monte Carlo method,* SIAM Rev., 12 (1970), pp. 1–63.

[HMG03] C. O. Hwang, M. Mascagni, and J. A. Given, *A Feynman-Kac path-integral implementation for Poisson's equation using an h-conditioned Green's function,* Math. Comput. Simulation, 62 (2003), pp. 347–355.

[Lie96] G. M. Lieberman, *Second Order Parabolic Differential Equations,* World Scientific, River Edge, NJ, 1996.

[Mai03] S. Maire, *An iterative computation of approximations on Korobov-like spaces,* J. Comput. Appl. Math., 157 (2003), pp. 261–281.

[New94] N. J. Newton, *Variance reduction for simulated diffusions,* SIAM J. Appl. Math., 54 (1994), pp. 1780–1805.

[RY94] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion,* 2nd ed., Grundlehren Math. Wiss. 293, Springer-Verlag, Berlin, 1994.

[Sab91] K. K. Sabelfeld, *Monte Carlo Methods in Boundary Value Problems,* Springer Ser. Comput. Phys., Springer-Verlag, Berlin, 1991 (translated from the Russian).

[Zha97] P. G. Zhang, *Exotic Options. A Guide to Second Generation Options,* World Scientific, Singapore, 1997.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1276–1294

ON THE CONVERGENCE OF MORTAR EDGE ELEMENT METHODS IN $\mathbb{R}^3$*

XUEJUN XU[†] AND R. H. W. HOPPE[‡]

Abstract. In this paper, we are concerned with mortar element methods for the numerical solution of the eddy currents equations based on domain decompositions on nonmatching grids using individual subdomain discretizations by the lowest order edge elements of Nédélec's first family. The main results are optimal a priori error estimates of the global discretization error and the Lagrange multipliers that take care of the weak continuity constraints on the tangential traces across interior subdomain boundaries. These estimates are derived under moderate regularity assumptions.

Key words. mortar edge elements, domain decomposition on nonmatching grids, eddy currents equations

AMS subject classifications. 65F10, 65N30

DOI. 10.1137/S0036142903438094

1. Introduction.

Mortar element methods have attracted considerable attention in recent years, since they can handle situations where meshes on different subdomains need not align across interfaces, and the matching of discretizations on adjacent subdomains is only enforced weakly. In [8], Bernardi, Maday, and Patera first introduced basic concepts of general mortar element methods, including the coupling of spectral elements with finite elements. Subsequently, they have been extensively used and analyzed by many authors. In [4], Ben Belgacem studied the mortar element method within a primal hybrid finite element formulation. Some extensions and convergence results in three dimensions have been considered in [5], [10], and [22].

In the framework of edge element discretizations, the mortar element method has been studied for two-dimensional problems in [3] and [6]. However, similar to second order elliptic problems (cf., e.g., [5], [10], [22]), the situation in the three-dimensional case is much more complicated, since it particularly requires a subtle specification of the multiplier space. Recently, the second author of this paper considered a mortar element method for three-dimensional Maxwell equations in [20], where the edge element of the first family has been studied (see also [21]). Related work for mortar edge elements has been proposed by Ben Belgacem, Buffa, and Maday in [7], but their result holds only for the lowest order edge elements of Nédélec's second family [26]. Furthermore, their error estimate of order $O(hlog(h))$ is not optimal and requires a somewhat high regularity of the solution, i.e., the solution is assumed to belong to $H^2(\mathbf{curl};\Omega)$.

In this paper, we will give an optimal error estimate for the mortar edge element method based on the lowest order edge elements of Nédélec's first family. Our convergence results are established under a weaker regularity assumption, i.e., the solution

[*] Received by the editors November 25, 2003; accepted for publication (in revised form) February 7, 2005; published electronically September 23, 2005.
http://www.siam.org/journals/sinum/43-3/43809.html

[†] LSEC, Institute of Computational Mathematics, Chinese Academy of Sciences, P.O. Box 2719, Beijing, 100080, People's Republic of China (xxj@lsec.cc.ac.cn). The work of this author was supported by the Alexander von Humboldt Foundation and special funds for major state basic research projects under 2005CB321701 as well as a grant from the National Science Foundation (NSF) of China (10471144).

[‡] Institut für Mathematik, Universität Augsburg, D-86159, Augsburg, Germany, and Department of Mathematics, University of Houston, Houston, TX 77204-3008 (rohop@math.uh.edu).

is assumed to belong to $H^1(\mathbf{curl};\Omega)$. On the other hand, on the basis of the discrete inf-sup condition constructed in [20], we also obtain an optimal error estimate for the Lagrange multiplier.

The paper is organized as follows. Section 2 describes the model problem under consideration. Section 3 introduces the mortar edge element method followed by the derivation of the optimal energy error estimate in section 4. Finally, section 5 is devoted to an optimal error estimate for the Lagrange multiplier.

2. Model problem. Given a bounded simply connected domain Ω in R^3 with polyhedral boundary $\partial\Omega$, we consider the following elliptic boundary value problem:

$$\begin{cases} \mathbf{curl}\ \mathbf{A}\ \mathbf{curl}\ \mathbf{j} + \mathbf{B}\ \mathbf{j} = \mathbf{f} & \text{in } \Omega, \\ \mathbf{j} \wedge \mathbf{n} = \mathbf{g} & \text{on } \partial\Omega, \end{cases} \tag{2.1}$$

where $\mathbf{n}$ denotes the exterior unit normal vector on $\partial\Omega$. We note that the above problem arises, for instance, in the computation of eddy currents and can be deduced from the time-dependent equations by using an implicit finite difference scheme (cf. [9], [18], [23]).

We assume $\mathbf{A} = \{a_{ij}\}_{i,j=1}^3$ and $\mathbf{B} = \{b_{ij}\}_{i,j=1}^3$ to be symmetric matrix-valued functions, with $a_{ij} \in C^1(\bar{\Omega}),\ b_{ij} \in L^\infty(\Omega),\ 1 \le i,j \le 3$, satisfying

$$c|\xi|^2 \le \sum_{i,j=1}^3 a_{ij}(x)\xi_i\xi_j \le C|\xi|^2, \qquad c|\xi|^2 \le \sum_{i,j=1}^3 b_{ij}(x)\xi_i\xi_j \le C|\xi|^2, \quad \xi \in R^3,$$

for almost all $x \in \Omega$. In this paper, the constants c and C with or without subscript always denote general positive constants independent of the mesh size. Moreover, we assume $\mathbf{f} \in L^2(\Omega)^3$ and suppose, for simplicity, that $\mathbf{g} = 0$.

We denote by $H(\mathbf{curl};\Omega)$ the Hilbert space

$$H(\mathbf{curl};\Omega) := \{\mathbf{q} \in L^2(\Omega)^3 \mid \mathbf{curl}\mathbf{q} \in L^2(\Omega)^3\}$$

equipped with the norm

$$\|\mathbf{q}\|_{\mathbf{curl},\Omega} := (\|\mathbf{q}\|_{0,\Omega}^2 + \|\mathbf{curl}\mathbf{q}\|_{0,\Omega}^2)^{\frac{1}{2}}.$$

Here and in what follows, $\|\cdot\|_{k,\Omega}, k \in \mathbb{N}_0$, stands for the norm of the Sobolev space $H^k(\Omega)^3$. Moreover, we define the space

$$H^1(\mathbf{curl};\Omega) := \{\mathbf{q} \in H^1(\Omega)^3 | \mathbf{curl}\mathbf{q} \in H^1(\Omega)^3\}$$

equipped with the norm

$$\|\mathbf{q}\|_{1,\mathbf{curl},\Omega} := (\|\mathbf{q}\|_{1,\Omega} + \|\mathbf{curl}\mathbf{q}\|_{1,\Omega})^{\frac{1}{2}}.$$

Similarly, if G is a subdomain of Ω, we can define the space $H^1(\mathbf{curl};G)$ over the subdomain G. The corresponding norm is denoted by $\|\mathbf{q}\|_{1,\mathbf{curl},G}$.

We refer to

$$\mathbf{V} := H_0(\mathbf{curl};\Omega) = \{\mathbf{q} \in H(\mathbf{curl};\Omega) \mid \mathbf{n} \wedge (\mathbf{q} \wedge \mathbf{n})|_{\partial\Omega} = 0\}$$

as the subspace of vector fields with vanishing tangential components trace on $\partial\Omega$.

Then, the variational formulation of (2.1) is to find $\mathbf{j} \in \mathbf{V}$ such that

$$a_\Omega(\mathbf{j},\mathbf{q}) \ = \ l(\mathbf{q}) \quad \forall \mathbf{q} \in \mathbf{V}, \tag{2.2}$$

where the bilinear form $a_\Omega(\cdot,\cdot) : H(\mathbf{curl};\Omega) \times H(\mathbf{curl};\Omega) \to \mathbb{R}$ and the functional $l(\cdot) : H(\mathbf{curl};\Omega) \to R$ are given by

$$a_\Omega(\mathbf{j},\mathbf{q}) := \int_\Omega (\mathbf{A}\ \mathbf{curl}\ \mathbf{j} \cdot \mathbf{curl}\ \mathbf{q} + \mathbf{B}\ \mathbf{j} \cdot \mathbf{q})\, dx,$$
$$l(\mathbf{q}) := \int_\Omega \mathbf{f} \cdot \mathbf{q}\, dx.$$

We further have to introduce the tangential traces of $H(\mathbf{curl};\Omega)$. In particular, we denote by div_τ and curl_τ the surfacic divergence and the adjoint of the surfacic rotational $\mathbf{curl}_\tau$ (cf. [1]). For $B \subset \partial\Omega$, the space $H_{00}^{\frac{1}{2}}(B)$ is the subspace of functions $u \in H^{\frac{1}{2}}(\Omega)$ whose extension $\tilde{u}$ by zero to $\partial\Omega\backslash B$ belongs to $H^{\frac{1}{2}}(\partial\Omega)$ with norm $\|u\|_{H_{00}^{\frac{1}{2}}(B)} := \|\tilde{u}\|_{\frac{1}{2},\partial\Omega}$. We refer to $H^{-\frac{1}{2}}(B)$ as the dual space of $H_{00}^{\frac{1}{2}}(B)$ (cf. [19] for details).

The tangential trace $(\mathbf{q} \wedge \mathbf{n})|_B$ belongs to the Hilbert space

$$H^{-\frac{1}{2}}(\mathrm{div}_\tau; B) \ := \ \{\mathbf{q} \in H^{-\frac{1}{2}}(B)^3 \mid \mathbf{n} \cdot \mathbf{q}|_B = 0 \text{ and } \mathrm{div}_\tau \mathbf{q} \in H^{-\frac{1}{2}}(B)\}$$

equipped with the norm

$$\|\mathbf{q}\|_{-\frac{1}{2},\mathrm{div}_\tau,B} \ := \ (\|\mathbf{q}\|^2_{-\frac{1}{2},B} + \|\mathrm{div}_\tau \mathbf{q}\|^2_{-\frac{1}{2},B})^{1/2},$$

whereas the tangential components trace $(\mathbf{n} \wedge (\mathbf{q} \wedge \mathbf{n}))|_B$ lives in the Hilbert space

$$H^{-\frac{1}{2}}(\mathrm{curl}_\tau; B) \ := \ \{\mathbf{q} \in H^{-\frac{1}{2}}(B)^3 \mid \mathbf{n} \cdot \mathbf{q}|_B = 0 \quad \text{and} \quad \mathrm{curl}_\tau \mathbf{q} \in H^{-\frac{1}{2}}(B)\}$$

equipped with the norm

$$\|\mathbf{q}\|_{-\frac{1}{2},\mathrm{curl}_\tau,B} \ := \ (\|\mathbf{q}\|^2_{-\frac{1}{2},B} + \|\mathrm{curl}_\tau \mathbf{q}\|^2_{-\frac{1}{2},B})^{1/2}.$$

The spaces $H^{-\frac{1}{2}}(\mathrm{div}_\tau; B)$ and $H^{-\frac{1}{2}}(\mathrm{curl}_\tau; B)$ are dual to each other with $\mathbf{L}^2_\mathbf{t}(B) := \{\mathbf{q} \in L^2(B)^3 \mid \mathbf{n} \cdot \mathbf{q}|_B = 0\}$ as the pivot space (cf. [13], [14], and [15] for details).

3. The mortar edge element method. We now introduce a mortar finite element method for the solution of (2.1). First, we partition Ω into nonoverlapping subdomains such that

$$\overline{\Omega} = \bigcup_{i=1}^N \overline{\Omega}_i \quad \text{and} \quad \Omega_i \cap \Omega_j = \emptyset, \quad i \neq j.$$

We assume this decomposition to be geometrically conforming in the sense that the intersection of $\bar{\Omega}_i \cap \bar{\Omega}_j$ for $i \neq j$ is either empty, a vertex, an edge, or a face. The skeleton of the decomposition

$$S = \bigcup_{i=1}^N \partial\Omega_i \backslash \partial\Omega$$

is partitioned into a set of disjoint open faces γ_m $(1 \leq m \leq M)$ called mortars, i.e.,

$$S = \bigcup_{m=1}^M \bar{\gamma}_m, \quad \gamma_m \cap \gamma_n = \emptyset \ \text{ if } \ m \neq n.$$

We denote the common interface between Ω_i and Ω_j by γ_m. We refer to $\gamma_{m(i)}$ as the mortar associated with subdomain Ω_i, while the other face, which geometrically occupies the same place, is denoted by $\delta_{m(j)}$ and is called the nonmortar.

Let $\mathcal{T}_i$ be a regular and quasi-uniform triangulation of the subdomain Ω_i with mesh size $h_i := \max_{K\in\mathcal{T}_i} h_K$ made of tetrahedra. The triangulations generally do not align at the interfaces. We denote the global mesh $\cup_i \mathcal{T}_i$ by $\mathcal{T}_h$ with mesh size $h := \max_i h_i$. We refer to $\mathcal{T}_{\gamma_{m(i)}}$ and $\mathcal{T}_{\delta_{m(j)}}$ as the triangulations which are inherited from the triangulations $\mathcal{T}_i$ and $\mathcal{T}_j$ on the mortar and nonmortar sides, respectively. We further denote by $h_{\gamma_{m(i)}}$ and $h_{\delta_{m(j)}}$ the global mesh sizes with respect to the triangulations $\mathcal{T}_{\gamma_{m(i)}}$ and $\mathcal{T}_{\delta_{m(j)}}$. Moreover, for $\Sigma_i \subset \bar{\Omega}_i$ we define $\mathcal{F}_h(\Sigma_i)$ and $\mathcal{E}_h(\Sigma_i)$ as the sets of faces, respectively, edges, of $\mathcal{T}_i$ in Σ_i. Likewise, for $\Sigma_{\gamma_{m(i)}}$ and $\Sigma_{\delta_{m(j)}} \subset \gamma_m$ we refer to $\mathcal{E}_h(\Sigma_{\gamma_{m(i)}})$ and $\mathcal{E}_h(\Sigma_{\delta_{m(j)}})$ as the set of edges of $\mathcal{T}_{\gamma_{m(i)}}$, respectively, $\mathcal{T}_{\delta_{m(j)}}$, in $\Sigma_{\gamma_{m(i)}}$, respectively, $\Sigma_{\delta_{m(j)}}$.

We assume that there exist constants c, C independent of $h_{\gamma_{m(i)}}$ and $h_{\delta_{m(j)}}$ such that

$$c\, h_{\gamma_{m(i)}} \;\le\; h_{\delta_{m(j)}} \;\le\; C\, h_{\gamma_{m(i)}}. \tag{3.1}$$

For the discretization of $H(\mathbf{curl};\Omega_i)$, we introduce Nédélec's **curl**-conforming edge elements of the first family as described in [25], i.e., for a tetrahedron $K \in \mathcal{T}_i$ the lowest order edge element $\mathrm{ND_1(K)}$ is defined as

$$\mathrm{ND_1(K)} \;:=\; \{\mathbf{q} = \mathbf{a} + \mathbf{b} \wedge \mathbf{x} \mid \mathbf{a},\ \mathbf{b} \in R^3,\ \mathbf{x} \in K\}.$$

Note that any $\mathbf{q} \in \mathrm{ND_1(K)}$ is uniquely determined by the degrees of freedom

$$l_e(\mathbf{q}) \;:=\; \int_e \mathbf{t_e} \cdot \mathbf{q}\, ds, \quad e \in \mathcal{E}_h(K), \tag{3.2}$$

where $\mathbf{t_e}$ stands for the tangential unit vector along e.

Then, the spaces $\mathrm{ND}_1(\Omega_i;\mathcal{T}_i)$ are given as follows:

$$\mathrm{ND}_1(\Omega_i;\mathcal{T}_i) \;:=\; \{\mathbf{q_h} \in H(\mathbf{curl};\Omega_i) \mid \mathbf{q_h}|_K \in \mathrm{ND}_1(K),\ K \in \mathcal{T}_i\}.$$

On the basis of the above definition, we consider the product space

$$\tilde{\mathbf{V}}_\mathbf{h} \;:=\; \{\mathbf{q_h} \in L^2(\Omega)^3 \mid \mathbf{q_h}|_{\Omega_i} \in \mathrm{ND}_{1,0}(\Omega_i;\mathcal{T}_i),\ 1 \le i \le n\},$$

where we refer to $\mathrm{ND}_{1,0}(\Omega_i;\mathcal{T}_i)$ as the subspace of vector fields with vanishing tangential component traces on $\partial\Omega \cap \partial\Omega_i$.

It is clear that we cannot expect $\tilde{\mathbf{V}}_\mathbf{h}$ to be a subspace of $H_0(\mathbf{curl};\Omega)$, since the tangential traces $(\mathbf{q_h} \wedge \mathbf{n})|_F, \mathbf{q_h} \in \tilde{\mathbf{V}}_\mathbf{h}$, are not continuous across the common face F of two adjacent subdomains. Therefore, in order to guarantee consistency of the approximation, we have to impose some weak continuity constraints on the tangential traces. We note that $(\mathbf{q_h} \wedge \mathbf{n})|_{\gamma_{m(i)}}$ and $(\mathbf{q_h} \wedge \mathbf{n})|_{\delta_{m(j)}}$ are elements of the lowest order Raviart–Thomas finite element spaces $\mathrm{RT}_0(\gamma_{m(i)};\mathcal{T}_{\gamma_{m(i)}})$ and $\mathrm{RT}_0(\delta_{m(j)};\mathcal{T}_{\delta_{m(j)}})$. We recall the definition of the lowest order Raviart–Thomas conforming finite element (cf. [12], [27]). For a triangle $T \in \mathcal{T}_{\gamma_{m(i)}}$, we define $\mathrm{RT_0(T)}$ by means of

$$\mathrm{RT_0(T)} \;:=\; \{\mathbf{q} = \mathbf{a} + b\mathbf{x} \mid \mathbf{a} \in \mathbb{R}^2,\ b \in \mathbb{R},\ \mathbf{x} \in T\}.$$

Any $\mathbf{q} \in \mathrm{RT_0(T)}$ is uniquely defined by the degrees of freedom

$$l_e(\mathbf{q}) \;:=\; \int_e \mathbf{n_e} \cdot \mathbf{q}\, ds, \quad e \in \mathcal{E}_h(T), \tag{3.3}$$

where $\mathbf{n_e}$ stands for the exterior unit normal vector with respect to e.

Then, $\mathrm{RT}_0(\gamma_{m(i)}; \mathcal{T}_{\gamma_{m(i)}})$ is given as

$$\mathrm{RT}_0(\gamma_{m(i)}; \mathcal{T}_{\gamma_{m(i)}}) \ := \ \{\mathbf{q_h} \in H(\mathrm{div}; \gamma_{m(i)}) \mid \mathbf{q_h}|_T \in \mathrm{RT}_0(T),\ T \in \mathcal{T}_{\gamma_{m(i)}}\},$$

and we can similarly define $\mathrm{RT}_0(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$.

For the Lagrange multiplier space we choose

$$\mathbf{M_h} := \prod_{\delta_{m(j)}} \mathbf{M_h}(\delta_{m(j)})$$

with

$$\dim\ \mathbf{M_h}(\delta_{m(j)}) = \dim\ \mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}}),$$

where $\mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$ denotes the subspace of vector fields with vanishing normal components along the boundary $\partial\delta_{m(j)}$.

For the proper definition of $\mathbf{M_h}(\delta_{m(j)})$ we need a more detailed specification of the basis fields of $\mathrm{RT}_0(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$. In view of (3.3), we specify the basis field $\mathbf{q}_\gamma$ associated with the edge $e_\gamma \in \mathcal{E}_h(\bar{\delta}_{m(j)})$ according to

$$\int_{e_\mu} \mathbf{n}_\mu \cdot \mathbf{q}_\gamma \ ds \ = \ h_{\delta_{m(j)}} \delta_{\gamma\mu}, \quad e_\mu \in \mathcal{E}_h(\bar{\delta}_{m(j)}). \tag{3.4}$$

We now define $\mathbf{M_h}(\delta_{m(j)})$ by an extension of the basis field $\mathbf{q_e} \in \mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$ with respect to those edges in $\delta_{m(j)}$ that have at least one neighboring edge on the boundary $\partial\delta_{m(j)}$. The precise specification requires some notation:

1. For an interior edge $e \in \mathcal{E}_h(\delta_{m(j)})$, we denote by

$$\mathcal{E}_h^{\partial\delta_{m(j)}}(e) \ := \ \{f \in \mathcal{E}_h(\partial\delta_{m(j)}) \mid f \subset \mathrm{supp}\ \mathbf{q}_e\} \tag{3.5}$$

the set of the neighboring edges on $\partial\delta_{m(j)}$.

2. For a boundary edge $f \in \mathcal{E}_h(\partial\delta_{m(j)})$, we refer to

$$\mathcal{E}_h^{\delta_{m(j)}}(f) \ := \ \{e \in \mathcal{E}_h(\delta_{m(j)}) \mid e \subset \mathrm{supp}\ \mathbf{q}_f\} \tag{3.6}$$

as the set of neighboring edges in the interior of $\delta_{m(j)}$.

Finally we define

$$\mathcal{E}_h^{\delta_{m(j)}}(\partial\delta_{m(j)}) \ := \bigcup_{f \in \mathcal{E}_h(\partial\delta_{m(j)})} \mathcal{E}_h^{\delta_{m(j)}}(f) \tag{3.7}$$

as the set of interior edges with a neighboring edge on $\partial\delta_{m(j)}$.

Then, for $e \in \mathcal{E}_h^{\delta_{m(j)}}(\partial\delta_{m(j)})$, we choose appropriate weighting factors $\lambda_{e,f} \in \mathbb{R}$, $f \in \mathcal{E}_h^{\partial\delta_{m(j)}}(e)$, and define the basis field $\tilde{\mathbf{q}}_\mathbf{e}$, $e \in \mathcal{E}_h(\delta_{m(j)})$, according to

$$\tilde{\mathbf{q}}_e = \begin{cases} \mathbf{q}_e, & e \in \mathcal{E}_h(\delta_{m(j)}) \backslash \mathcal{E}_h^{\delta_{m(j)}}(\partial\delta_{m(j)}) \\ \mathbf{q}_e + \sum_{f \in \mathcal{E}_h^{\partial\delta_{m(j)}}(e)} \lambda_{e,f} \mathbf{q}_f, & e \in \mathcal{E}_h^{\delta_{m(j)}}(\partial\delta_{m(j)}), \end{cases} \tag{3.8}$$

where the weighting factors are assumed to satisfy

$$\begin{cases} \lambda_{e,f} \geq 0, \\ \sum_{e \in \mathcal{E}_h^{\delta_{m(j)}}(f)} \lambda_{e,f} = 1, \quad f \in \mathcal{E}_h(\partial\delta_{m(j)}). \end{cases} \tag{3.9}$$

The thus specified basis fields define

$$\mathbf{M_h}(\delta_{m(j)}) \ := \ \operatorname{span}\{\tilde{\mathbf{q}}_e | e \in \mathcal{E}_h(\delta_{m(j)})\}. \tag{3.10}$$

Remark 3.1. In view of (3.9) it is easy to check that $\mathbf{M_h}(\delta_{m(j)})$ contains the constant vectors.

Next, we introduce the L^2-projection $Q_h^{\delta_{m(j)}} : L^2(\gamma_m)^2 \rightarrow \mathbf{M_h}(\delta_{m(j)})$ as follows:

$$(Q_h^{\delta_{m(j)}}\mathbf{q}, \mathbf{w}) = (\mathbf{q}, \mathbf{w}), \quad \mathbf{w} \in \mathbf{M_h}(\delta_{m(j)}). \tag{3.11}$$

LEMMA 3.1. *Let* $Q_h^{\delta_{m(j)}}$ *be given by* (3.11). *Then there holds*

$$\|\mathbf{q} - Q_h^{\delta_{m(j)}}\mathbf{q}\|_{0,\gamma_m} \leq C \ h_{\delta_{m(j)}}^{\frac{1}{2}} \ |\mathbf{q}|_{\frac{1}{2},\delta_{m(j)}}, \quad \mathbf{q} \in (H^{\frac{1}{2}}(\delta_{m(j)}))^2.$$

Proof. Let $\mathbf{I_h}$ denote the global interpolation operator associated with the space $\mathbf{M_h}(\delta_{m(j)})$, i.e.,

$$\mathbf{I_h q} = \sum_{e \in \mathcal{E}_h(\delta_{m(j)})} l_e(\mathbf{q})\tilde{\mathbf{q}}_e,$$

where $l_e(\mathbf{q}) = \int_e \mathbf{n}_e \cdot \mathbf{q}\, ds \ \forall \mathbf{q} \in (H^1(\delta_{m(j)}))^2$.

In view of Remark 3.1 we know that $\mathbf{I_h}$ preserves constant vectors, i.e., for any $\mathbf{C} \in R^2$,

$$\mathbf{I}_h\mathbf{C} = \mathbf{C}.$$

Consequently, by the standard Bramble–Hilbert lemma and scaling argument we get

$$\begin{aligned} \|(\mathbf{I} - \mathbf{I_h})\mathbf{q}\|_{0,\gamma_m}^2 &= \|(\mathbf{I} - \mathbf{I_h})(\mathbf{q} + \mathbf{C})\|_{0,\gamma_m}^2 \\ &= \sum_{T \in \mathcal{T}_{\delta_{m(j)}}} \|(\mathbf{I} - \mathbf{I_h})(\mathbf{q} + \mathbf{C})\|_{0,T}^2 \\ &\leq \ Ch_{\delta_{m(j)}}^2 |\mathbf{q}|_{1,\delta_{m(j)}}^2, \qquad \mathbf{q} \in (H^1(\delta_{m(j)}))^2, \end{aligned}$$

whence

$$\|(\mathbf{I} - \mathbf{I_h})\mathbf{q}\|_{0,\gamma_m} \ \leq \ Ch_{\delta_{m(j)}} |\mathbf{q}|_{1,\delta_{m(j)}}, \quad \mathbf{q} \in (H^1(\delta_{m(j)}))^2.$$

It follows from the definition of $Q_h^{\delta_{m(j)}}$ that

$$\|(\mathbf{I} - Q_h^{\delta_{m(j)}})\mathbf{q}\|_{0,\gamma_m} \leq \|(\mathbf{I} - \mathbf{I_h})\mathbf{q}\|_{0,\gamma_m} \ \leq \ Ch_{\delta_{m(j)}} |\mathbf{q}|_{1,\delta_{m(j)}}, \quad \mathbf{q} \in (H^1(\delta_{m(j)}))^2.$$

On the other hand,

$$\|(\mathbf{I} - Q_h^{\delta_{m(j)}})\mathbf{q}\|_{0,\gamma_m} \leq 2\|\mathbf{q}\|_{0,\delta_{m(j)}}.$$

The assertion then follows from a standard interpolation of the preceding inequalities. □

We now introduce the following mortar edge element space:

$$\begin{aligned} \mathbf{V_h} = \{\mathbf{q_h} \mid \mathbf{q_h} \in \tilde{\mathbf{V}}_\mathbf{h}, \text{ and for any } \gamma_m = \gamma_{m(i)} = \delta_{m(j)}, \\ Q_h^{\delta_{m(j)}}(\mathbf{q_h} \wedge \mathbf{n}|_{\gamma_{m(i)}}) = Q_h^{\delta_{m(j)}}(\mathbf{q_h} \wedge \mathbf{n}|_{\delta_{m(j)}})\}. \end{aligned} \tag{3.12}$$

We define the bilinear form $a_h(\cdot,\cdot) : \mathbf{V_h} \times \mathbf{V_h} \rightarrow \mathbb{R}$ by means of

$$a_h(\mathbf{j_h}, \mathbf{q_h}) \;=\; \sum_{i=1}^{N} \int_{\Omega_i} (\mathbf{A\ curl\ j_h} \cdot \mathbf{curl\ q_h} + \mathbf{B\ j_h} \cdot \mathbf{q_h})\ dx. \tag{3.13}$$

Then the mortar finite element method for the solution of (2.4) can be stated as follows: Find $\mathbf{j_h} \in \mathbf{V_h}$ such that

$$a_h(\mathbf{j_h}, \mathbf{q_h}) \;=\; l(\mathbf{q_h}), \quad \mathbf{q_h} \in \mathbf{V_h}. \tag{3.14}$$

4. Error estimates.

We first recall the well-known Strang lemma (cf., e.g., [17]).

LEMMA 4.1 (Strang's lemma). *Let* $\mathbf{j}, \mathbf{j_h}$ *be the solutions of* (2.2) *and* (3.14), *respectively. Then there holds*

$$\begin{aligned}\|\mathbf{j} - \mathbf{j_h}\|_{a_h} \;\le\; &\left(\inf_{\mathbf{q_h} \in \mathbf{V_h}/\{0\}} \|\mathbf{j} - \mathbf{q_h}\|_{a_h} + \sup_{\mathbf{q_h} \in \mathbf{V_h}\setminus\{0\}} \frac{|a_h(\mathbf{j}, \mathbf{q_h}) - (f, \mathbf{q_h})|}{\|\mathbf{q_h}\|_{a_h}} \right) \\ &:= C(E_a + E_c),\end{aligned}$$

where $\|\cdot\|_{a_h} = a_h(\cdot,\cdot)^{\frac{1}{2}}$.

We are now in a position to estimate the two terms on the right side of the above inequality. As usual, we refer to the first one as the approximation error and to the second one as the consistency error.

4.1. Consistency error.

For $\mathbf{curl\ j} \in (H^1(\Omega_i))^3$, $\mathbf{q_h} \in ND_1(\Omega_i; \mathcal{T}_i)$, by Stokes' theorem we get

$$\begin{aligned}&\int_{\Omega_i} \mathbf{curl} \cdot \mathbf{Acurl\ j} \cdot \mathbf{q_h}\ dx \\ &\quad - \int_{\Omega_i} \mathbf{Acurl\ j} \cdot \mathbf{curl\ q_h}\ dx \;=\; (\mathbf{n} \wedge (\mathbf{Acurl\ j} \wedge \mathbf{n}), \mathbf{q_h} \wedge \mathbf{n})_{0,\partial\Omega_i},\end{aligned}$$

where $\mathbf{n} \wedge (\mathbf{Acurl\ j} \wedge \mathbf{n})$ is the tangential components trace of $\mathbf{Acurl\ j}$. Rearranging the right-hand term in the above equality, for any $\mathbf{q_h} \in \tilde{\mathbf{V}}_\mathbf{h}$, and $\mathbf{curl\ j} \in (H^1(\Omega_i))^3$, $i = 1, \ldots, N$, we have (cf. [7] for details)

$$\begin{aligned}&\sum_{i=1}^{N} \left(\int_{\Omega_i} \mathbf{curl} \cdot \mathbf{Acurl\ j} \cdot \mathbf{q_h}\ dx \;-\; \int_{\Omega_i} \mathbf{Acurl\ j} \cdot \mathbf{curl\ q_h}\ dx \right) \\ &\quad = \sum_{m=1}^{M} (\mathbf{n} \wedge (\mathbf{Acurl\ j} \wedge \mathbf{n}), [\mathbf{q_h} \wedge \mathbf{n}])_{0,\gamma_m},\end{aligned} \tag{4.1}$$

where $[\cdot]$ denotes the jump across the interface γ_m, i.e.,

$$[\mathbf{q_h} \wedge \mathbf{n}] = \mathbf{q_h} \wedge \mathbf{n}|_{\delta_{m(j)}} - \mathbf{q_h} \wedge \mathbf{n}|_{\gamma_{m(i)}}.$$

On the basis of the above equality, we can easily show that

$$E_c \;=\; \sup_{\mathbf{q_h} \in \mathbf{V_h}\setminus\{0\}} \left| \sum_{m=1}^{M} \frac{(\mathbf{n} \wedge (\mathbf{Acurl\ j} \wedge \mathbf{n}), [\mathbf{q_h} \wedge \mathbf{n}])_{0,\gamma_m}}{\|\mathbf{q_h}\|_{a_h}} \right|.$$

THEOREM 4.1. *Assume* $\mathbf{j} \in H^1(\mathbf{curl};\Omega)$. *Then the consistency error can be estimated as follows*:

$$E_c \leq C \left(\sum_{j=1}^{N} h_j^2 \ \|\mathbf{curl}\,\mathbf{j}\|_{1,\Omega_j}^2 \right)^{\frac{1}{2}}.$$

Proof. It follows from Lemma 3.1, (3.12), and the trace inequality that

$$\begin{aligned}
&|(\mathbf{n} \wedge (\mathbf{A}\mathbf{curl}\,\mathbf{j} \wedge \mathbf{n}), [\mathbf{q_h} \wedge \mathbf{n}])_{0,\gamma_m}| \\
&\quad = |(\mathbf{n} \wedge (\mathbf{A}\mathbf{curl}\,\mathbf{j} \wedge \mathbf{n}) - Q_h^{\delta_{m(j)}}(\mathbf{n} \wedge (\mathbf{A}\mathbf{curl}\,\mathbf{j} \wedge \mathbf{n})), [\mathbf{q_h} \wedge \mathbf{n}])_{0,\gamma_m}| \\
&\quad \leq \|\mathbf{n} \wedge (\mathbf{A}\mathbf{curl}\,\mathbf{j} \wedge \mathbf{n}) - Q_h^{\delta_{m(j)}}(\mathbf{n} \wedge (\mathbf{A}\mathbf{curl}\,\mathbf{j} \wedge \mathbf{n}))\|_{0,\gamma_m} \ \|[\mathbf{q_h} \wedge \mathbf{n}]\|_{0,\gamma_m} \\
&\quad \leq Ch_{\delta_{m(j)}}^{\frac{1}{2}} \ |\mathbf{n} \wedge (\mathbf{A}\mathbf{curl}\,\mathbf{j} \wedge \mathbf{n})|_{\frac{1}{2},\delta_{m(j)}} \ \|[\mathbf{q_h} \wedge \mathbf{n}]\|_{0,\gamma_m} \\
&\quad \leq Ch_j^{\frac{1}{2}} \ \|\mathbf{curl}\,\mathbf{j}\|_{1,\Omega_j} \ \|[\mathbf{q_h} \wedge \mathbf{n}]\|_{0,\gamma_m}.
\end{aligned}$$

On the other hand, for $\mathbf{q_h} \in \mathbf{V_h}$, Theorem 3.2 in [20] yields

$$\|[\mathbf{q_h} \wedge \mathbf{n}]\|_{0,\gamma_m} \ \leq \ C \ h_{\delta_{m(j)}}^{\frac{1}{2}} \ (\|\mathbf{curl}\,\mathbf{q_h}\|_{0,\Omega_i} + \|\mathbf{curl}\,\mathbf{q_h}\|_{0,\Omega_j}). \tag{4.2}$$

On the basis of the preceding inequalities, we get

$$\begin{aligned}
E_c &\leq \left[\sum_{j=1}^{N} C \ h_j \ \|\mathbf{curl}\,\mathbf{j}\|_{1,\Omega_j} (\|\mathbf{curl}\,\mathbf{q_h}\|_{0,\Omega_i} + \|\mathbf{curl}\,\mathbf{q_h}\|_{0,\Omega_j}) \right] / \|\mathbf{q_h}\|_{a_h} \\
&\leq C \ \left[\|\mathbf{curl}\,\mathbf{q_h}\|_{0,\Omega} \left(\sum_{j=1}^{N} h_j^2 \ \|\mathbf{curl}\,\mathbf{j}\|_{1,\Omega_j}^2 \right)^{\frac{1}{2}} \right] / \|\mathbf{q_h}\|_{a_h} \\
&\leq C \ \left(\sum_{j=1}^{N} h_j^2 \ \|\mathbf{curl}\,\mathbf{j}\|_{1,\Omega_j}^2 \right)^{\frac{1}{2}}. \qquad \square
\end{aligned}$$

4.2. Approximation error. We first introduce the extension operator $E_h^{\delta_{m(j)}}$: $\mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}}) \rightarrow \mathrm{ND}_1(\Omega_j; \mathcal{T}_j)$, defined according to

$$(E_h^{\delta_{m(j)}} \lambda_h^j) \wedge \mathbf{n} = \lambda_h^j \text{ on } \delta_{m(j)}, \quad \lambda_h^j \in \mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}}),$$

where all degrees of freedom that are not located on $\delta_{m(j)}$ are set equal to zero.

In order to estimate $E_h^{\delta_{m(j)}} \lambda_h^j$, $\lambda_h^j \in \mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$, we need some auxiliary results.

LEMMA 4.2. *For any* $\mathbf{q_h} \in \mathrm{ND}_1(\Omega_i; \mathcal{T}_i)$, *there holds*

$$ch_i^3 \sum_{T \in \mathcal{F}_h(\bar{\Omega}_i)} |(\mathbf{n_T} \cdot \mathbf{curl}\mathbf{q_h})|_T|^2 \leq \|\mathbf{curl}\mathbf{q_h}\|_{0,\Omega_i}^2 \leq Ch_i^3 \sum_{T \in \mathcal{F}_h(\bar{\Omega}_i)} |(\mathbf{n_T} \cdot \mathbf{curl}\mathbf{q_h})|_T|^2,$$

and

$$ch_i^3 \sum_{e \in \mathcal{E}_h(\bar{\Omega}_i)} |(\mathbf{t_e} \cdot \mathbf{q_h})(x_e^M)|^2 \leq \|\mathbf{q_h}\|_{0,\Omega_i}^2 \leq Ch_i^3 \sum_{e \in \mathcal{E}_h(\bar{\Omega}_i)} |(\mathbf{t_e} \cdot \mathbf{q_h})(x_e^M)|^2,$$

where $\mathbf{n_T}$ *denotes the exterior unit normal vector with respect to* $T \in \mathcal{F}_h(\bar{\Omega}_i)$, *and* x_e^M *is the midpoint of the edge e. Similarly, for any* $\delta_{m(j)} \subset S$, *and any* $\mathbf{q_h} \in \mathrm{RT}_0(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$, *we have*

$$ch^2_{\delta_{m(j)}} \sum_{T \in \mathcal{T}_{\delta_{m(j)}}} |(\mathrm{div}_\tau \mathbf{q_h})|_T|^2 \leq \|\mathrm{div}_\tau \mathbf{q_h}\|^2_{0,\delta_{m(j)}} \leq Ch^2_{\delta_{m(j)}} \sum_{T \in \mathcal{T}_{\delta_{m(j)}}} |(\mathrm{div}_\tau \mathbf{q_h})|_T|^2,$$

and

$$ch^2_{\delta_{m(j)}} \sum_{e \in \mathcal{E}_h(\bar{\delta}_{m(j)})} |(\mathbf{n_e} \cdot \mathbf{q_h})(x_e^M)|^2 \leq \|\mathbf{q_h}\|^2_{0,\delta_{m(j)}} \leq Ch^2_{\delta_{m(j)}} \sum_{e \in \mathcal{E}_h(\bar{\delta}_{m(j)})} |(\mathbf{n_e} \cdot \mathbf{q_h})(x_e^M)|^2.$$

Proof. We first prove the second inequality. In the reference tetrahedron $\hat{K}$, it is easy to see that

$$\|\hat{\mathbf{q}}_h\|_{0,\hat{K}} \quad \text{and} \quad \left(\sum_{e \in \mathcal{E}_h(\bar{K})} |(\mathbf{t}_e \cdot \hat{\mathbf{q}}_\mathbf{h})(x_e^M)|^2 \right)^{\frac{1}{2}}$$

are equivalent norms over the finite dimension space. By a scaling argument and summing up all $e \in \mathcal{E}_h(\bar{\Omega}_i)$, we can get the second inequality. Similarly, the fourth inequality can be verified. Moreover, the first and third inequalities are easy consequences of the following fact:

$$\mathbf{curl}\ \mathbf{q_h}|_K \in P_0(K)^3,\ K \in \mathcal{T}_i, \quad \text{and} \quad \mathrm{div}_\tau \mathbf{q_h}|_T \in P_0(T),\ T \in \mathcal{T}_{\delta_{m(j)}}. \qquad \square$$

On the basis of Lemma 4.2 we can derive the following lemma.

LEMMA 4.3. *For* $\lambda_h^j \in \mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$ *there holds*

$$\|E_h^{\delta_{m(j)}} \lambda_h^j\|_{\mathbf{curl},\Omega_j} \leq C\ h^{\frac{1}{2}}_{\delta_{m(j)}}\ \|\lambda_h^j\|_{\mathrm{div}_\tau,\delta_{m(j)}},$$

where $\|\mathbf{v}\|_{\mathrm{div}_\tau,\delta_{m(j)}} := (\|\mathbf{v}\|^2_{0,\delta_{m(j)}} + \|\mathrm{div}_\tau \mathbf{v}\|^2_{0,\delta_{m(j)}})^{\frac{1}{2}}, \quad \forall \mathbf{v} \in \mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$.

Proof. It follows from the definition of the extension operator $E_h^{\delta_{m(j)}}$ and Lemma 4.2 that

$$\begin{aligned} \|\mathbf{curl}(E_h^{\delta_{m(j)}} \lambda_h^j)\|^2_{0,\Omega_j} &\leq Ch_j^3 \sum_{T \in \mathcal{T}_{\delta_{m(j)}}} |\mathbf{n_T} \cdot \mathbf{curl}(E_h^{\delta_{m(j)}} \lambda_h^j)|_T|^2 \\ &= Ch_j^3 \sum_{T \in \mathcal{T}_{\delta_{m(j)}}} |\mathrm{div}_\tau (E_h^{\delta_{m(j)}} \lambda_h^j \wedge \mathbf{n})|_T|^2 \\ &= Ch_j^3 \sum_{T \in \mathcal{T}_{\delta_{m(j)}}} |\mathrm{div}_\tau (\lambda_h^j)|_T|^2 \\ &\leq Ch_j\ \|\mathrm{div}_\tau (\lambda_h^j)\|^2_{0,\delta_{m(j)}}. \end{aligned}$$

Using Lemma 4.2 again, we have

$$\begin{aligned} \|E_h^{\delta_{m(j)}} \lambda_h^j\|^2_{0,\Omega_j} &\leq Ch_j^3 \sum_{e \in \mathcal{E}_h(\bar{\Omega}_j)} |(\mathbf{t_e} \cdot \mathbf{E_h^{\delta_{m(j)}}} \boldsymbol{\lambda}_\mathbf{h}^\mathbf{j})(x_e^M)|^2 \\ &= Ch_j^3 \sum_{e \in \mathcal{E}_h(\bar{\Omega}_j)} |\mathbf{n_e} \cdot (\mathbf{E_h^{\delta_{m(j)}}} \boldsymbol{\lambda}_\mathbf{h}^\mathbf{j} \wedge \mathbf{n})(x_e^M)|^2 \end{aligned}$$

$$
\begin{aligned}
&= Ch_j^3 \sum_{e\in\mathcal{E}_h(\delta_{m(j)})} |(\mathbf{n_e}\cdot\lambda_h^j)(x_e^M)|^2 \\
&\le h_j\|\lambda_h^j\|^2_{0,\delta_{m(j)}}.
\end{aligned}
$$

Then, Lemma 4.3 follows from the above two inequalities. □

LEMMA 4.4. *Let $\Pi_h^j : H^1(\mathbf{curl};\Omega_j)\to \mathrm{ND}_1(\Omega_j;\mathcal{T}_j)$ be the standard interpolation operator associated with subdomain Ω_j. Then there holds*

(i)$\|\mathbf{n_T}\cdot(\mathbf{curl}\ \Pi_h^j\mathbf{j}-\mathbf{curl}\ \mathbf{j})\|_{0,T}\le Ch_K^{\frac12}\|\mathbf{curl}\ \mathbf{j}\|_{1,K},\quad K\in\mathcal{T}_j,$

(ii) $\|\Pi_h^j\mathbf{j}-\mathbf{j}\|_{0,T}\le Ch_K^{\frac12}\|\mathbf{j}\|_{1,\mathbf{curl},K},\quad T\in\partial K.$

Proof. We first prove (i). For $K\in\mathcal{T}_i$ and $T\in\partial K$ let $F_K(\hat x)=B_K\hat x+b_K,\ \hat x\in\hat K$, be the affine transformation mapping the reference element $\hat K$ onto K. Further, choose $\hat T\in\partial\hat K$ such that $T=F_K(\hat T)$ and denote by $F_T=F_K|_{\hat T}$ the associated affine transformation $F_T(\hat x)=B_T\hat x+b_T$, $\hat x\in\hat T$, mapping $\hat T$ onto T. Setting $\hat{\mathbf{j}}=B_K^*\mathbf{j}$, it is easy to check that

$$\mathbf{n_T}\cdot(\mathbf{curl}\ \Pi_h^j\mathbf{j}-\mathbf{curl}\ \mathbf{j})|_T=\mathrm{curl}_\tau\Pi_h^j\mathbf{j}|_T-\mathrm{curl}_\tau\mathbf{j}|_T.$$

We note (cf. Lemma 3.57 of [24] for details) that

$$\mathrm{curl}_\tau\ \mathbf{j}|_T\ =\ (B_T^*)^{-1}\ \mathrm{curl}_\tau\ \hat{\mathbf{j}}|_{\hat T}\ B_T^{-1},$$

where $\mathrm{curl}_\tau\ \mathbf{u}$ denotes the 2×2 matrix with entries

$$[\mathrm{curl}_\tau\ \mathbf{u}]_{i,j}=\frac{\partial u_i}{\partial x_j}-\frac{\partial u_j}{\partial x_i},\quad \mathbf{u}:=(u_1,u_2).$$

It follows that

$$
\begin{aligned}
(4.3)\qquad &\|\mathbf{n_T}\cdot(\mathbf{curl}\ \Pi_h^j\mathbf{j}-\mathbf{curl}\ \mathbf{j})\|^2_{0,T}\\
&=\|\mathrm{curl}_\tau\Pi_h^j\mathbf{j}|_T-\mathrm{curl}_\tau\mathbf{j}|_T\|^2_{0,T}\\
&\le C\ |\det\ B_T|\|B_T^{-1}\|^4\ \|\mathbf{n}_{\hat{\mathbf{T}}}\cdot\mathbf{curl}(\hat\Pi_h^j\hat{\mathbf{j}}-\hat{\mathbf{j}})\|^2_{0,\hat T}\\
&\le C\ |\det\ B_T|\|B_T^{-1}\|^4\ \|\mathbf{curl}(\hat\Pi_h^j\hat{\mathbf{j}}-\hat{\mathbf{j}})\|^2_{0,\hat T}\\
&\le C|\det\ B_T|\|B_T^{-1}\|^4\ \|(I-\hat W_h^j)\mathbf{curl}\ \hat{\mathbf{j}}\|^2_{0,\hat T}.
\end{aligned}
$$

Here, we have used $\mathbf{curl}\ \hat\Pi_h^j\hat{\mathbf{j}}=\hat W_h^j\mathbf{curl}\ \hat{\mathbf{j}}$ with $\hat W_h^j$ being the L^2-projection onto the space of elementwise constants. It follows that

$$(4.4)\qquad \|(I-\hat W_h^j)\mathbf{curl}\ \hat{\mathbf{j}}\|^2_{0,\hat T}\ \le\ C\ |\mathbf{curl}\ \hat{\mathbf{j}}|^2_{1,\hat K}.$$

We note that

$$\mathbf{curl}\ \hat{\mathbf{j}}\ =\ B_K^*\ \mathbf{curl}\ \mathbf{j}\ B_K,$$

where $\mathbf{curl}\ \mathbf{j}$ stands for the 3×3 matrix with entries

$$[\mathbf{curl}\ \mathbf{j}]_{i,j}=\frac{\partial j_i}{\partial x_j}-\frac{\partial j_j}{\partial x_i},\quad \mathbf{j}:=(j_1,j_2,j_3).$$

Hence, by backtransformation we obtain (cf. Lemma 5.5 in [1] for details)

$$(4.5)\qquad |\mathbf{curl}\ \hat{\mathbf{j}}|^2_{1,\hat K}\ \le\ C\ |\det\ B_K|^{-2}\ \|B_K\|^7\|B_K^*\|^2\ |\mathbf{curl}\ \mathbf{j}|^2_{1,K}.$$

Summarizing (4.3), (4.4), and (4.5), it follows that

$$\begin{aligned}(4.6)\quad &\|\mathbf{n_T}\cdot(\mathbf{curl}\,\Pi_h^j\,\mathbf{j}-\mathbf{curl}\,\mathbf{j})\|_{0,T}^2\\ &\leq\; C\frac{|\det\,B_T|}{|\det\,B_K|}\,(\|B_T^{-1}\|\;\|B_K\|)^4\;\|B_K\|^3\|B_K^*\|^2|\det\,B_K|^{-1}\;|\mathbf{curl}\,\mathbf{j}|_{1,K}^2.\end{aligned}$$

Finally, taking into account that $\mathcal{T}_i$ is a regular triangulation, we have

$$(4.7)\qquad \|B_T^{-1}\|\;\|B_K\|\;\leq\;C,\quad \|B_K\|,\|B_K^*\|\;\leq\;C\;h_K.$$

Moreover,

$$(4.8)\qquad |\det\,B_T|\;=\;\frac{\text{meas}(T)}{\text{meas}(\hat{T})},\quad |\det\,B_K|\;=\;\frac{\text{meas}(K)}{\text{meas}(\hat{K})}.$$

Using (4.7) and (4.8) in (4.6) gives the assertion.

We now prove (ii). Observing

$$\mathbf{j}|_T=(B_T^*)^{-1}\hat{\mathbf{j}}|_{\hat{T}},$$

we have

$$\|\Pi_h^j\mathbf{j}-\mathbf{j}\|_{0,T}^2\leq|\text{det}B_T|\|(B_T^*)^{-1}\|^2\|\hat{\Pi}_h^j\hat{\mathbf{j}}-\hat{\mathbf{j}}\|_{0,\hat{T}}^2.$$

Using the trace inequality and similar arguments as in the proof of Theorem 5.41 of [24], we can derive that

$$\|\hat{\Pi}_h^j\hat{\mathbf{j}}-\hat{\mathbf{j}}\|_{0,\hat{T}}^2\leq C(|\hat{\mathbf{j}}|_{1,\hat{K}}+|\mathbf{curl}\,\hat{\mathbf{j}}|_{1,\hat{K}}).$$

On the other hand,

$$|\hat{\mathbf{j}}|_{1,\hat{K}}^2\leq\|B_K\|^5\|B_K^*\|^2|\text{det}B_K^{-1}|^2|\mathbf{j}|_{1,K}^2.$$

Combining the above three inequalities with (4.5), (4.7), and (4.8) yields Lemma 4.4(ii). □

We further introduce a special projection operator $\pi_h^{\delta_{m(j)}}$ which will play an important role in analyzing the approximate error of the mortar edge element method. We define $\pi_h^{\delta_{m(j)}}:L^2(\gamma_m)^2\to\text{RT}_{0,0}(\delta_{m(j)};\mathcal{T}_{\delta_{m(j)}})$ according to

$$(4.9)\qquad \int_{\delta_{m(j)}}\pi_h^{\delta_{m(j)}}(\mathbf{p})\cdot\mathbf{q_h}\;dx\;=\;\int_{\delta_{m(j)}}\mathbf{p}\cdot\mathbf{q_h}\;dx,\quad \mathbf{q_h}\in\mathbf{M_h}(\delta_{m(j)}).$$

The boundedness of $\pi_h^{\delta_{m(j)}}$ is a direct consequence of the following result.

Lemma 4.5. *The following* inf-sup *condition holds true:*

$$\inf_{\mathbf{q_h}\in RT_0(\delta_{m(j)};\mathcal{T}_{\delta_{m(j)}})}\;\sup_{\mu_h\in\mathbf{M_h}(\delta_{m(j)})}\;\frac{(\mathbf{q_h},\mu_h)_{0,\delta_{m(j)}}}{\|\mathbf{q_h}\|_{0,\delta_{m(j)}}\;\|\mu_h\|_{0,\delta_{m(j)}}}\;\geq\;C\;>\;0.$$

Proof. Taking the construction (3.8) on the basis of $\mathbf{M_h}(\delta_{m(j)})$ into account, for $\mathbf{q_h}\in RT_0(\delta_{m(j)};\mathcal{T}_{\delta_{m(j)}})$ we determine $\mu_h\in\mathbf{M_h}(\delta_{m(j)})$ by specifying its degrees of freedom according to

$$\ell_e(\mu_h)=\begin{cases}\ell_e(\mathbf{q_h}), & e\in\mathcal{E}_h(\delta_{m(j)})\setminus\mathcal{E}_h^{\delta_{m(j)}}(\partial\delta_{m(j)}),\\ \ell_e(\mathbf{q_h})+\sum_{f\in\mathcal{E}_h^{\delta_{m(j)}}(e)}\lambda_{e,f}\;\ell_f(\mathbf{q_h}), & e\in\mathcal{E}_h^{\delta_{m(j)}}(\partial\delta_{m(j)}).\end{cases}$$

The assertion can then be verified by following lines of proof analogous to those of [20, Lemma 3.2]. □

Furthermore, by Lemma 3.2 in [20], we know that the following inf-sup condition also true

COROLLARY 4.6. *There holds*

$$\inf_{\mu_h \in \mathbf{M_h}(\delta_{m(j)})} \sup_{\mathbf{q_h} \in RT_{0,0}(\delta_{m(j)};\mathcal{T}_{\delta_{m(j)}})} \frac{(\mathbf{q_h}, \mu_h)_{0,\delta_{m(j)}}}{\|\mathbf{q_h}\|_{0,\delta_{m(j)}} \; \|\mu_h\|_{0,\delta_{m(j)}}} \;\geq\; C \;>\; 0.$$

On the basis of Lemma 4.5, we have the following.

COROLLARY 4.7. *Let $\pi_h^{\delta_{m(j)}}$ be given by* (4.9)*. Then there holds*

$$\|\pi_h^{\delta_{m(j)}}(\mathbf{p})\|_{0,\delta_{m(j)}} \;\leq\; C\;\|\mathbf{p}\|_{0,\gamma_m}, \quad \mathbf{p} \in L^2(\gamma_m)^2.$$

Proof. Using Lemma 4.5, straightforward computation reveals

$$\begin{aligned}
\|\pi_h^{\delta_{m(j)}}(\mathbf{p})\|_{0,\delta_{m(j)}} &\leq C \sup_{\mu_h \in \mathbf{M_h}(\delta_{m(j)})} \frac{(\pi_h^{\delta_{m(j)}}(\mathbf{p}), \mu_h)_{0,\delta_{m(j)}}}{\|\mu_h\|_{0,\delta_{m(j)}}} \\
&= C \sup_{\mu_h \in \mathbf{M_h}(\delta_{m(j)})} \frac{(\mathbf{p}, \mu_h)_{0,\delta_{m(j)}}}{\|\mu_h\|_{0,\delta_{m(j)}}} \\
&\leq C\;\|\mathbf{p}\|_{0,\gamma_m}. \qquad □
\end{aligned}$$

As a further consequence of the inf-sup condition in Lemma 4.5, we obtain the following.

LEMMA 4.8. *Let $\Pi_h \,:\, H^1(\mathbf{curl};\Omega) \cap \mathbf{V} \;\to\; \tilde{\mathbf{V}}_\mathbf{h}$ be the standard interpolation operator. Then we have*

$$\|\mathrm{div}_\tau \pi_h^{\delta_{m(j)}}[\Pi_h \mathbf{j} \wedge \mathbf{n}]\|_{0,\gamma_m} \;\leq\; C\;\|\mathrm{div}_\tau[\Pi_h \mathbf{j} \wedge \mathbf{n}]\|_{0,\gamma_m}.$$

Proof. We denote by $P_h^{\delta_{m(j)}}$ the $\mathrm{RT}_0(\delta_{m(j)};\mathcal{T}_{\delta_{m(j)}})$-interpolation operator. Observing that $P_h^{\delta_{m(j)}}|_T,\; T \in \mathcal{T}_{\delta_{m(j)}}$, preserves constant tangential traces, by a Bramble–Hilbert argument we obtain

$$\begin{aligned}
&\|(I - P_h^{\delta_{m(j)}})[\Pi_h \mathbf{j} \wedge \mathbf{n}]\|^2_{0,\gamma_m} \\
&\quad \leq C h^2_{\delta_{m(j)}} \sum_{T \in \mathcal{T}_{\delta_{m(j)}}} \; \sum_{T' \cap T \neq \emptyset, T' \in \mathcal{T}_{\gamma_{m(i)}}} |[\Pi_h \mathbf{j} \wedge \mathbf{n}]|^2_{1,T' \cap T} \\
&\quad = C h^2_{\delta_{m(j)}} \sum_{T \in \mathcal{T}_{\delta_{m(j)}}} \|\mathrm{div}_\tau[\Pi_h \mathbf{j} \wedge \mathbf{n}]\|^2_{0,T} \\
&\quad = C\; h^2_{\delta_{m(j)}}\;\|\mathrm{div}_\tau[\Pi_h \mathbf{j} \wedge \mathbf{n}]\|^2_{0,\gamma_m},
\end{aligned}$$

where we have used the fact that $\Pi_h \mathbf{j} \wedge \mathbf{n}|_{\gamma_m}$ belongs to the lowest order Raviart–Thomas space. Similar arguments for the proof of the first inequality can be found in [16] . So we get

$$\|(I - P_h^{\delta_{m(j)}})[\Pi_h \mathbf{j} \wedge \mathbf{n}]\|_{0,\gamma_m} \leq C\; h_{\delta_{m(j)}}\;\|\mathrm{div}_\tau[\Pi_h \mathbf{j} \wedge \mathbf{n}]\|_{0,\gamma_m}. \tag{4.10}$$

Moreover, in view of

$$\mathrm{div}_\tau P_h^{\delta_{m(j)}}[\Pi_h \mathbf{j} \wedge \mathbf{n}] \;=\; W_h^{\delta_{m(j)}}\;\mathrm{div}_\tau[\Pi_h \mathbf{j} \wedge \mathbf{n}],$$

where $W_h^{\delta_{m(j)}}$ is the L^2-projection onto the elementwise constants, we obtain

$$\|\mathrm{div}_\tau P_h^{\delta_{m(j)}}[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m} \;\leq\; C\;\|\mathrm{div}_\tau[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m}. \tag{4.11}$$

We have $(\pi_h^{\delta_{m(j)}} - P_h^{\delta_{m(j)}})[\Pi_h\mathbf{j}\wedge\mathbf{n}] \in \mathrm{RT}_0(\delta_{m(j)};\mathcal{T}_{\delta_{m(j)}})$, and hence, by Lemma 4.5 and (4.10),

$$\begin{aligned}\|(\pi_h^{\delta_{m(j)}} - P_h^{\delta_{m(j)}})[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m} &\leq\; C \sup_{\psi\in\mathbf{M_h}(\delta_{m(j)})}\frac{((\pi_h^{\delta_{m(j)}} - P_h^{\delta_{m(j)}})[\Pi_h\mathbf{j}\wedge\mathbf{n}],\psi)}{\|\psi\|_{0,\delta_{m(j)}}}\\ &=\; C \sup_{\psi\in\mathbf{M_h}(\delta_{m(j)})}\frac{((I - P_h^{\delta_{m(j)}})[\Pi_h\mathbf{j}\wedge\mathbf{n}],\psi)}{\|\psi\|_{0,\delta_{m(j)}}}\\ &\leq\; C\;h_{\delta_{m(j)}}\;\|\mathrm{div}_\tau[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m}.\end{aligned} \tag{4.12}$$

Combining (4.11) and (4.12), we get

$$\begin{aligned}&\|\mathrm{div}_\tau\pi_h^{\delta_{m(j)}}[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m}\\ &\quad\leq\; \|\mathrm{div}_\tau(\pi_h^{\delta_{m(j)}} - P_h^{\delta_{m(j)}})[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m} + \|\mathrm{div}_\tau P_h^{\delta_{m(j)}}[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m}\\ &\quad\leq C\;h_{\delta_{m(j)}}^{-1}\;\|(\pi_h^{\delta_{m(j)}} - P_h^{\delta_{m(j)}})[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m} + \|\mathrm{div}_\tau[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m}\\ &\quad\leq\; C\;\|\mathrm{div}_\tau[\Pi_h\mathbf{j}\wedge\mathbf{n}]\|_{0,\gamma_m}. \qquad \square\end{aligned}$$

We are now in a position to estimate the discretization error of the mortar edge element method.

THEOREM 4.2. *For any* $\mathbf{j}\in H^1(\mathbf{curl};\Omega)$ *there exists a function* $\mathbf{q_h}\in\mathbf{V_h}$ *such that*

$$\|\mathbf{j}-\mathbf{q_h}\|_{a_h} \leq C\left(\sum_{j=1}^{N} h_j^2\;\|\mathbf{j}\|_{1,\mathbf{curl},\Omega_j}^2\right)^{\frac{1}{2}}.$$

Proof. We define $\mathbf{q_h}$ as

$$\mathbf{q_h} = \Pi_h\mathbf{j} - \sum_{m=1}^{M} E_h^{\delta_{m(j)}}\{\pi_h^{\delta_{m(j)}}[(\Pi_h^j\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\Pi_h^i\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}}]\}$$

and remark that $\mathbf{q_h}\in\mathbf{V_h}$ can be easily seen.

For each $\delta_{m(j)}$, by Lemma 4.3, Corollary 4.7, and Lemma 4.8, we get

$$\begin{aligned}&\|E_h^{\delta_{m(j)}}(\pi_h^{\delta_{m(j)}}((\Pi_h^j\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\Pi_h^i\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}}))\|_{\mathbf{curl},\Omega_j}\\ &\quad\leq C\;h_{\delta_{m(j)}}^{\frac{1}{2}}\;\|\mathrm{div}_\tau\big(\pi_h^{\delta_{m(j)}}((\Pi_h^j\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\Pi_h^i\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}})\big)\|_{0,\gamma_m}\\ &\qquad + C\;h_{\delta_{m(j)}}^{\frac{1}{2}}\;\|\pi_h^{\delta_{m(j)}}\big((\Pi_h^j\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\Pi_h^i\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}}\big)\|_{0,\gamma_m}\\ &\quad\leq\; C\;h_{\delta_{m(j)}}^{\frac{1}{2}}\;\|\mathrm{div}_\tau\big((\Pi_h^j\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\Pi_h^i\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}}\big)\|_{0,\gamma_m}\\ &\qquad + C\;h_{\delta_{m(j)}}^{\frac{1}{2}}\;\|(\Pi_h^j\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\Pi_h^i\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}}\|_{0,\gamma_m}\\ &\quad:=\; I_1\;+\;I_2.\end{aligned} \tag{4.13}$$

As far as the first term I_1 is concerned, applying Lemma 4.4 results in

$$\begin{aligned} I_1 \le& \; C\ h_j^{\frac{1}{2}}\ \big(\|\mathrm{div}_\tau\big((\Pi_h^j \mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}}\big)\|_{0,\gamma_m} \\ &+\|\mathrm{div}_\tau\big((\Pi_h^i \mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}} - (\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}}\big)\|_{0,\gamma_m}\big) \\ \le& \; C\ h_j^{\frac{1}{2}}\ \left(\sum_{T\in\mathcal{T}(\delta_{m(j)})} (\|\mathbf{n}_T\cdot(\mathbf{curl}\ \Pi_h^j\ \mathbf{j} - \mathbf{curl}\ \mathbf{j})|_T\|_{0,T}^2\right)^{\frac{1}{2}} \\ &+\ \left(\sum_{T\in\mathcal{T}(\gamma_{m(i)})} (\|\mathbf{n}_T\cdot(\mathbf{curl}\ \Pi_h^i\ \mathbf{j} - \mathbf{curl}\ \mathbf{j})|_T\|_{0,T}^2)^{\frac{1}{2}}\right) \\ \le& \; C\ h_j^{\frac{1}{2}}\ \big(h_j^{\frac{1}{2}}\|\mathbf{curl}\ \mathbf{j}\|_{1,\Omega_j} + h_i^{\frac{1}{2}}\|\mathbf{curl}\ \mathbf{j}\|_{1,\Omega_i}\big). \end{aligned} \tag{4.14}$$

For the second term I_2, using Lemma 4.4, we obtain

$$\begin{aligned} I_2 \le& \; C\ h_{\delta_{m(j)}}^{\frac{1}{2}}\ \big(\|(\Pi_h^j\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}}\|_{0,\gamma_m}\big) \\ &+\|(\Pi_h^i\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}} - (\mathbf{j}\wedge\mathbf{n})|_{\gamma_{m(i)}}\|_{0,\gamma_m} \\ \le& \; C\ h_j^{\frac{1}{2}}\ \big(h_j^{\frac{1}{2}}\|\mathbf{j}\|_{1,\mathbf{curl},\Omega_j} + h_i^{\frac{1}{2}}\|\mathbf{j}\|_{1,\mathbf{curl},\Omega_i}\big). \end{aligned} \tag{4.15}$$

Observing the standard approximation property

$$\|\mathbf{j}-\Pi_h\mathbf{j}\|_{a_h} \le C\left(\sum_{j=1}^N h_j^2\ \|\mathbf{j}\|_{1,\mathbf{curl},\Omega_j}^2\right)^{\frac{1}{2}}$$

and using (4.13), (4.14), and (4.15) results in

$$\begin{aligned} \|\mathbf{j}-\mathbf{q_h}\|_{a_h}^2 \ \le&\ C\ \big(\|\mathbf{j}-\Pi_h\mathbf{j}\|_{a_h}^2 \\ &+\ \sum_{m=1}^{m}\|E_h^{\delta_{m(j)}}(\pi_h^{\delta_{m(j)}}((\Pi_h^j\mathbf{j}\wedge\mathbf{n})|_{\delta_{m(j)}} - (\Pi_h^i\mathbf{j}\wedge\mathbf{n}))\|_{\mathbf{curl},\Omega_j}^2\big) \\ \le&\ C\ \sum_{j=1}^N h_j^2\ \|\mathbf{j}\|_{1,\mathbf{curl},\Omega_j}^2. \qquad \square \end{aligned}$$

Finally, Theorems 4.1 and 4.2 imply the main result of this paper.

THEOREM 4.3. *Let* $\mathbf{j}\in H^1(\mathbf{curl};\Omega)$ *and* $\mathbf{j_h}\in\mathbf{V_h}$ *be the solutions of* (2.2) *and* (3.14), *respectively. Then there holds*

$$\|\mathbf{j}-\mathbf{j_h}\|_{a_h}\ \le\ C\ \left(\sum_{j=1}^N h_j^2\ \|\mathbf{j}\|_{1,\mathbf{curl},\Omega_j}^2\right)^{\frac{1}{2}}.$$

5. Saddle point formulation.

A saddle point formulation for mortar element methods associated with second order elliptic problems has been introduced in [4]. In particular, an a priori estimate for the Lagrange multiplier in the $(H_{00}^{\frac{1}{2}})'$-norm has been established there, whereas related estimates in mesh-dependent norms have been given in [28], [29], [30]. In this section, we will derive an a priori estimate for the Lagrange multiplier of the mortar edge element method.

First, we introduce a macrohybrid variational formulation for the continuous problem (2.1).

Using the domain decomposition as presented in the preceding section, we introduce the product space

$$\mathbf{X} \;:=\; \{\mathbf{q} \in L^2(\Omega)^3 \mid \mathbf{q}|_{\Omega_i} \in H(\mathbf{curl};\Omega_i),\; (\mathbf{n} \wedge (\mathbf{q} \wedge \mathbf{n}))|_{\partial\Omega_i \cap \partial\Omega} = \mathbf{0}\}$$

equipped with the norm

$$\|\mathbf{q}\|_{\mathbf{X}} \;:=\; \left(\sum_{i=1}^{N} \|\mathbf{q}\|^2_{\mathbf{curl},\Omega_i} \right)^{\frac{1}{2}}.$$

We further consider the subspace

$$\tilde{\mathbf{V}} \;:=\; \left\{ \mathbf{q} \in \mathbf{X} \mid [\mathbf{q} \wedge \mathbf{n}]|_{\gamma_m} \in (H^{\frac{1}{2}}_{00}(\gamma_m))^2 \right\}$$

provided with the norm

$$\|\mathbf{q}\|_{\tilde{\mathbf{V}}} \;:=\; \left(\|\mathbf{q}\|^2_{\mathbf{X}} + \|[\mathbf{q} \wedge \mathbf{n}]\|^2_{\frac{1}{2},S} \right)^{\frac{1}{2}},$$

where

$$\|[\mathbf{q} \wedge \mathbf{n}]\|_{\frac{1}{2},S} \;:=\; \left(\sum_{\gamma_m \in S} \|[\mathbf{q} \wedge \mathbf{n}]\|^2_{(H^{\frac{1}{2}}_{00}(\gamma_m))^2} \right)^{\frac{1}{2}}.$$

A natural candidate for the multiplier space is then

$$\mathbf{M} \;:=\; \prod_{\gamma_m} (H^{-\frac{1}{2}}(\delta_{m(j)}))^2$$

equipped with the norm

$$\|\mu\|_{\mathbf{M}} \;:=\; \left(\sum_{\delta_{m(j)} \in S} \|\mu|_{\delta_{m(j)}}\|^2_{-\frac{1}{2},\delta_{m(j)}} \right)^{\frac{1}{2}},$$

where $H^{-\frac{1}{2}}(\delta_{m(j)}) := (H^{\frac{1}{2}}_{00}(\delta_{m(j)}))'$.

We introduce the bilinear form $a(\cdot,\cdot)\mathbf{X} \times \mathbf{X} \to \mathbb{R}$ as the sum of the bilinear forms associated with the subdomain problems according to

$$a(\mathbf{j},\mathbf{q}) \;:=\; \sum_{i=1}^{N} a_{\Omega_i}(\mathbf{j}|_{\Omega_i},\mathbf{q}|_{\Omega_i}) = \sum_{i=1}^{N} \int_{\Omega_i} \left[\mathbf{A}\mathbf{curl}\;\mathbf{j} \cdot \mathbf{curl}\;\mathbf{q} + \mathbf{B}\mathbf{j} \cdot \mathbf{q} \right] dx.$$

Furthermore, we define the bilinear form $b(\cdot,\cdot) : \tilde{\mathbf{V}} \times \mathbf{M} \to \mathbf{R}$ by means of

$$b(\mathbf{q},\mu) := \langle [\mathbf{q} \wedge \mathbf{n}], \mu \rangle_{\frac{1}{2},S},$$

where $\langle \cdot,\cdot \rangle_{\frac{1}{2},S} := \sum_{\delta_{m(j)} \in S} \langle \cdot,\cdot \rangle_{\frac{1}{2},\delta_{m(j)}}$.

Then the appropriate macrohybrid variational formulation of (2.1) can be formulated as follows:

Find $(\mathbf{j},\lambda) \in \tilde{\mathbf{V}} \times \mathbf{M}$ such that

$$\begin{aligned} a(\mathbf{j},\mathbf{q}) \;+\; b(\mathbf{q},\lambda) &= l(\mathbf{q}), \quad \mathbf{q} \in \tilde{\mathbf{V}}, \\ b(\mathbf{j},\mu) &= 0, \quad \mu \in \mathbf{M}. \end{aligned} \tag{5.1}$$

Denote by $B : \tilde{\mathbf{V}} \to \mathbf{M}$ the operator associated with the bilinear form $b(\cdot,\cdot)$, i.e.,

$$\langle B\mathbf{q}, \mu \rangle_{\frac{1}{2},S} = b(\mathbf{q}, \mu), \quad \mu \in \mathbf{M}.$$

It is proved in Theorem 2.1 of [20] that the bilinear form $a(\cdot,\cdot)$ is Ker$B$-elliptic and the bilinear form $b(\cdot,\cdot)$ satisfies the LBB condition. So the saddle point problem (5.1) admits a unique solution. For $\mathbf{q} \in \mathbf{V} \subset \tilde{\mathbf{V}}$, the first equation of (5.1) reduces to (2.2). Hence, the solution $\mathbf{j}$ of (5.1) is also the solution of (2.2). Finally, by (4.1) we know that $\lambda|_{\gamma_m} = \mathbf{n} \wedge (\mathbf{A}\ \mathbf{curl}\ \mathbf{j} \wedge \mathbf{n})|_{\gamma_m}$.

Next, we consider the discrete version of (5.1). On $\tilde{\mathbf{V}}_\mathbf{h}$, we define the norm

$$\|\mathbf{q_h}\|_{\tilde{\mathbf{V}}_\mathbf{h}} \ := \ \left(\|\mathbf{q_h}\|^2_{\mathbf{X}} \ + \ \|[\mathbf{q_h} \wedge \mathbf{n}]|_S\|^2_{\frac{1}{2},h,S}\right)^{\frac{1}{2}}, \quad \mathbf{q_h} \in \tilde{\mathbf{V}}_\mathbf{h},$$

where $\|\cdot\|_{\frac{1}{2},h,S}$ is given by

$$\|[\mathbf{q_h} \wedge \mathbf{n}]|_S\|_{\frac{1}{2},h,S} \ := \ \left(\sum_{\gamma_m \subset S} \|[\mathbf{q_h} \wedge \mathbf{n}]\|^2_{\frac{1}{2},h,\gamma_m}\right)^{\frac{1}{2}}$$

and $\|\cdot\|_{\frac{1}{2},h,\gamma_m}$ stands for the mesh-dependent norm:

$$\|[\mathbf{q_h} \wedge \mathbf{n}]\|_{\frac{1}{2},h,\gamma_m} \ := \ h^{-\frac{1}{2}}_{\delta_{m(j)}} \ \|[\mathbf{q_h} \wedge \mathbf{n}]\|_{0,\gamma_m}.$$

The Lagrange multiplier space $\mathbf{M_h}$ will be provided with the following mesh-dependent norm:

$$\|\mu_h\|_{\mathbf{M_h}} \ := \ \|\mu_h\|_{-\frac{1}{2},h,S}, \quad \mu_h \in \mathbf{M_h},$$

where

$$\|\mu_h\|_{-\frac{1}{2},h,S} \ := \ \left(\sum_{\delta_{m(j)} \subset S} \|\mu_h\|^2_{-\frac{1}{2},h,\delta_{m(j)}}\right)^{\frac{1}{2}}$$

and $\|\cdot\|_{-\frac{1}{2},h,\delta_{m(j)}}$ is given by

$$\|\mu_h|_{\delta_{m(j)}}\|_{-\frac{1}{2},h,\delta_{m(j)}} \ := \ h^{\frac{1}{2}}_{\delta_{m(j)}} \ \|\mu_h\|_{0,\delta_{m(j)}}.$$

In addition to the bilinear form $a_h(\cdot,\cdot) : \tilde{\mathbf{V}}_\mathbf{h} \times \tilde{\mathbf{V}}_\mathbf{h} \to \mathbb{R}$ as defined by (3.13), we introduce the bilinear form $b_h(\cdot,\cdot) : \tilde{\mathbf{V}}_\mathbf{h} \times \mathbf{M_h} \to \mathbb{R}$ according to

$$b_h(\mathbf{q_h}, \mu_h) \ := \ \sum_{\gamma_m \in S} ([\mathbf{q_h} \wedge \mathbf{n}]|_{\gamma_m}, \mu_h)_{0,\delta_{m(j)}}.$$

Then the mortar edge element approximation of (5.1) amounts to the solution of the following problem: Find $(\mathbf{j_h}, \lambda_h) \in \tilde{\mathbf{V}}_\mathbf{h} \times \mathbf{M_h}$ such that

$$\begin{aligned} a_h(\mathbf{j_h}, \mathbf{q_h}) \ + \ b_h(\mathbf{q_h}, \lambda_h) &= \ l(\mathbf{q_h}), \quad \mathbf{q_h} \in \tilde{\mathbf{V}}_\mathbf{h}, \\ b_h(\mathbf{j_h}, \mu_h) &= \ 0, \quad \mu_h \in \mathbf{M_h}. \end{aligned} \tag{5.2}$$

The saddle point problem (5.2) admits a unique solution which follows from the following LBB condition for the bilinear form $b_h(\cdot,\cdot)$.

LEMMA 5.1. *The bilinear form* $b_h(\cdot,\cdot) : \tilde{\mathbf{V}}_\mathbf{h} \times \mathbf{M}_\mathbf{h} \to \mathbf{R}$ *satisfies a discrete* inf-sup *condition (LBB condition) uniformly in* h_i, *i.e., there exists a constant* $c > 0$ *independent of the mesh size* h_i *such that*

$$\sup_{\mathbf{q_h} \in \tilde{\mathbf{V}}_\mathbf{h}} \frac{b_h(\mathbf{q_h}, \mu_h)}{\|\mathbf{q_h}\|_{\tilde{\mathbf{V}}_\mathbf{h}}} \;\geq\; c\;\|\mu_h\|_{\mathbf{M_h}}.$$

Proof. For any $\mu_h \in \mathbf{M_h}(\delta_{m(j)})$ we define $\mathbf{p_h^j} \in \mathrm{RT}_{0,0}(\delta_{m(j)}; \mathcal{T}_{\delta_{m(j)}})$ according to

$$\ell_e(\mathbf{p_h^j}) \;=\; \ell_e(\mu_h), \quad e \in \mathcal{E}_h(\delta_{m(j)})$$

and refer to $\mathbf{q_h^j} \in \mathrm{ND}_1(\Omega_j; \mathcal{T}_j)$ as the trivial extension, i.e.,

$$\mathbf{q_h^j} \wedge \mathbf{n} \;=\; \mathbf{p_h^j} \quad \text{on} \quad \delta_{m(j)},$$

where all degrees of freedom that are not located on $\delta_{m(j)}$ are set equal to zero, especially $[\mathbf{q_h^j} \wedge \mathbf{n}] = \mathbf{p_h^j}$. On the basis of Lemma 4.3, we have

$$\begin{aligned}
\|\mathbf{q_h^j}\|_{\mathbf{curl},\Omega_j} &\leq C\; h_j^{\frac{1}{2}}\;\|\mathbf{p_h^j}\|_{\mathrm{div}_\tau,\delta_{m(j)}} \\
&\leq C\; h_j^{-\frac{1}{2}}\;\|\mathbf{p_h^j}\|_{0,\delta_{m(j)}} \\
&= C\; h_j^{-\frac{1}{2}}\;\|[\mathbf{q_h^j} \wedge \mathbf{n}]\|_{0,\delta_{m(j)}}.
\end{aligned}$$

By Corollary 4.6 and the above inequality, we obtain

$$\begin{aligned}
(\mu_h, [\mathbf{q_h^j} \wedge \mathbf{n}]|_{\delta_{m(j)}})_{0,\delta_{m(j)}} &\geq C\;\|\mu_h\|_{0,\delta_{m(j)}} \|[\mathbf{q_h^j} \wedge \mathbf{n}]\|_{0,\delta_{m(j)}} \\
&\geq C\; h_j^{\frac{1}{2}}\;\|\mu_h\|_{0,\delta_{m(j)}} \|\mathbf{q_h^j}\|_{\mathbf{curl},\Omega_j} \\
&\geq C\;\|\mu_h\|_{-\frac{1}{2},h,\delta_{m(j)}}\;\|\mathbf{q_h^j}\|_{\mathbf{curl},\Omega_j}.
\end{aligned}$$

On the other hand,

$$\begin{aligned}
(\mu_h, [\mathbf{q_h^j} \wedge \mathbf{n}]|_{\delta_{m(j)}})_{0,\delta_{m(j)}} &\geq C\;\|\mu_h\|_{0,\delta_{m(j)}}\;\|[\mathbf{n} \wedge \mathbf{q_h^j}]\|_{0,\delta_{m(j)}} \\
&= C\; h_j^{\frac{1}{2}}\;\|\mu_h\|_{0,\delta_{m(j)}}\; h_j^{-\frac{1}{2}}\;\|[\mathbf{q_h^j} \wedge \mathbf{n}]\|_{0,\delta_{m(j)}} \\
&= C\;\|\mu_h\|_{-\frac{1}{2},h,\delta_{m(j)}}\;\|[\mathbf{q_h^j} \wedge \mathbf{n}]\|_{\frac{1}{2},h,\delta_{m(j)}}.
\end{aligned}$$

Adding the above inequalities and summing over all $\delta_{m(j)} \subset \Gamma$ gives the assertion. □

Finally, we obtain the following.

THEOREM 5.2. *Let* $\mathbf{j} \in H^1(\mathbf{curl}; \Omega)$ *and* $(\mathbf{j_h}, \lambda_h) \in \tilde{\mathbf{V}}_\mathbf{h} \times \mathbf{M_h}$ *be the solutions of* (2.2) *and* (5.2), *respectively. Then there holds*

$$\|\lambda - \lambda_h\|_{-\frac{1}{2},h,S} \leq C \left(\sum_{j=1}^{N} h_j^2\;\|\mathbf{j}\|^2_{1,\mathbf{curl},\Omega_j} \right)^{\frac{1}{2}}.$$

Proof. On the basis of the inf-sup condition developed in Lemma 5.1 and arguments similar to those in [12] for the mixed finite element methods and [30] for the saddle point method for mortar element methods, we get

$$\|\lambda - \lambda_h\|_{-\frac{1}{2},h,S} \leq C(\|\mathbf{j} - \mathbf{j_h}\|_{a_h} + \inf_{\mu_h \in \mathbf{M_h}} \|\lambda - \mu_h\|_{-\frac{1}{2},h,S}).$$

By Theorem 4.3, we have

$$\|\mathbf{j}-\mathbf{j_h}\|_{a_h} \le C\left(\sum_j^N h_j^2\|\mathbf{j}\|^2_{1,\mathbf{curl},\Omega_j}\right)^{\frac{1}{2}}. \tag{5.3}$$

Moreover, by Lemma 3.1

$$\begin{aligned}\inf_{\mu_h\in\mathbf{M_h}(\delta_{m(j)})}\|\lambda-\mu_h\|_{-\frac{1}{2},h,\delta_{m(j)}} &= h^{\frac{1}{2}}_{\delta_{m(j)}}\inf_{\mu_h\in\mathbf{M_h}(\delta_{m(j)})}\|\lambda-\mu_h\|_{0,\delta_{m(j)}}\\ &\le C\,h_j\;\|\mathbf{n}\wedge(\mathbf{A}\ \mathbf{curl}\ \mathbf{j}\wedge\mathbf{n})\|_{\frac{1}{2},\delta_{m(j)}}\\ &\le C\,h_j\;\|\mathbf{curl}\ \mathbf{j}\|_{1,\Omega_j}.\end{aligned}$$

Summing over all $\delta_{m(j)}$ results in

$$\inf_{\mu_h\in\mathbf{M_h}}\|\lambda-\mu_h\|_{-\frac{1}{2},h,S} \le C\left(\sum_j^N h_j^2\|\mathbf{curl}\ \mathbf{j}\|^2_{1,\Omega_j}\right)^{\frac{1}{2}}. \tag{5.4}$$

Finally, combining (5.3) and (5.4) gives the assertion. □

Acknowledgment. We thank the anonymous referees who meticulously read through the paper and made many helpful suggestions which led to an improved presentation of this paper.

REFERENCES

[1] A. Alonso and A. Valli, *Some remarks on the characterization of the space of tangential traces of $H(rot;\Omega)$ and the construction of an extension operator*, Manuscr. Math., 89 (1986), pp. 159–178.

[2] A. Alonso and A. Valli, *An optimal domain decomposition preconditioner for low-frequency time harmonic Maxwell equations*, Math. Comp., 68 (1999), pp. 607–631.

[3] A. Ben Abdallah, F. Ben Belgacem, and Y. Maday, *Mortaring the two-dimensional Nédélec finite elements for the discretization of the Maxwell equations*, Math Models Methods Appl. Sci., to appear.

[4] F. Ben Belgacem, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–198.

[5] F. Ben Belgacem and Y. Maday, *The mortar element method for three-dimensional finite elements*, RAIRO Modél. Math. Anal. Numér., 31 (1997), pp. 289–302.

[6] F. Ben Belgacem and Y. Maday, *The Mortar and Primal Hybrid Mortar Finite Element Method of the Class $H(curl)$*, preprint, Universite Paul Sabatier, Toulouse, France, 1997.

[7] F. Ben Belgacem, A. Buffa, and Y. Maday, *The mortar finite element method for 3D Maxwell equations: First results*, SIAM J. Numer. Anal., 39 (2001), pp. 880–901.

[8] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Nonlinear Partial Differential Equations and Their Applications, College de France Seminar, Pitman Res. Notes in Math. 9, H. Brézis and J. L. Lions, eds., Longman Scientific and Technical, Harlow, UK, 1994.

[9] A. Bossavit, *Electromagnétism, en vue de la modélisation*, Springer, New York, 1993.

[10] D. Braess and W. Dahmen, *Stability estimates of the mortar finite element method for 3-dimensional problems*, East-West J. Numer. Math., 6 (1998), pp. 249–264.

[11] D. Braess, W. Dahmen, and C. Wieners, *A multigrid algorithm for the mortar finite element method*, SIAM J. Numer. Anal., 37 (2000), pp. 48–69.

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

[13] A. Buffa and Ph. Ciarlet Jr., *On traces for functional spaces related to Maxwell's equations. Part* I: *An integration by parts formula in Lipschitz polyhedra*, Math. Methods Appl. Sci., 24 (2001), pp. 9–30.

[14] A. Buffa and Ph. Ciarlet Jr., *On traces for functional spaces related to Maxwell's equations. Part* II: *Hodge decompositions on the boundary of Lipschitz polyhedra and applications*, Math. Methods Appl. Sci., 24 (2001), pp. 31–48.

[15] A. Buffa, M. Costabel, and D. Sheen, *On traces for* $\mathbf{H}(\mathbf{curl},\Omega)$ *in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.

[16] T. F. Chan, B. F. Smith, and J. Zou, *Overlapping Schwarz methods on unstructured meshes using nonmatching coarse grids*, Numer. Math., 73 (1996), pp. 149–167.

[17] Ph. G. Ciarlet, *The Finite Element Method for Elliptic Problem*, North-Holland, Amsterdam, 1978.

[18] J. P. Ciarlet and J. Zou, *Fully discrete finite element approaches for time-dependent Maxwell's equation*, Numer. Math., 82 (1999), pp. 193–219.

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

[20] R. H. W. Hoppe, *Mortar edge element methods in* $\mathbb{R}^3$, East-West J. Numer. Math., 7 (1999), pp. 159–173.

[21] R. H. W. Hoppe, S. Petrova, and V. Schulz, 3*D structural optimization in electromagnetics*, in Proceedings of the 13th International Conference on Domain Decomposition Methods and Applications, Lyon, 2000, N. Debit et al., eds., CIMNE, Barcelona, 2002, pp. 479–486.

[22] C. Kim, R. Lazarov, J. Pasciak, and P. Vassilevski, *Multiplier spaces for the mortar finite element method in three dimensions*, SIAM J. Numer. Anal., 39 (2001), pp. 519–538.

[23] P. Monk, *A finite element method for approximating the time harmonic Maxwell equations*, Numer. Math., 63 (1992), pp. 243–261.

[24] P. Monk, *Finite Element Methods for Maxwell's Equations*, Oxford University Press, New York, 2003.

[25] J.-C. Nédélec, *Mixed finite element in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[26] J.-C. Nédélec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 50 (1986), pp. 57–81.

[27] P. Raviart and J. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Method, Lectures Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer, New York, 1977, pp. 292–315.

[28] B. Wohlmuth, *A mortar finite element method using dual spaces for the Lagrange multiplier*, SIAM J. Numer. Anal., 38 (2000), pp. 989–1012.

[29] B. Wohlmuth, *A residual based error estimator for mortar finite element discretisations*, Numer. Math., 84 (1999), pp. 143–171.

[30] B. Wohlmuth, *Hierarchical a posteriori error estimators for mortar finite element methods with Lagrange multipliers*, SIAM J. Numer. Anal., 36 (1999), pp. 1636–1658.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1295–1320

MIXED FINITE ELEMENT METHODS FOR SMOOTH DOMAIN FORMULATION OF CRACK PROBLEMS*

Z. BELHACHMI†, J. M. SAC-EPÉE†, AND J. SOKOLOWSKI‡

Abstract. The discretization by finite element methods of a new variational formulation of crack problems is considered. The new formulation, called the smooth domain method, is derived for crack problems in the case of an elastic membrane. Inequality type boundary conditions are prescribed at the crack faces. The resulting model takes the form of a unilateral contact problem on the crack. We study and implement various mixed finite element methods for the numerical approximation of the model. A priori error estimates are derived, and results of computations are provided. The convergence rates obtained from the numerical simulations are in agreement with the theoretical estimates.

Key words. crack problems, variational inequalities, mixed finite element methods, a priori error estimates

AMS subject classifications. 65N30, 74M15, 35J85

DOI. 10.1137/S0036142903429729

1. Introduction.

The smooth domain method is based on a new approach [25] to the crack modelling of linear elastic bodies with inequality type boundary conditions prescribed on the crack faces. By a crack problem we mean the boundary value problem defined in the geometrical domain $\Omega_c = \Omega \setminus \Gamma_c$ with the cut Γ_c. The cut is called a crack provided that some boundary conditions are specified on both sides $\Gamma_c^{\pm}$ of the set Γ_c [24]. In general, such conditions are only an approximation of the exact contact conditions derived for the displacement and stress fields in the framework of the elasticity theory. We restrict ourselves to a model problem with the unilateral conditions for scalar unknown functions. More realistic boundary conditions for the elasticity boundary value problem are the subject of further investigations from the numerical point of view in a forthcoming work. The boundary conditions of unilateral type on $\Gamma_c^{\pm}$ describe the mutual nonpenetration between the crack faces. In the smooth domain method the elements of the convex cone of admissible displacements and stresses used in the mixed variational formulation are extended to the crack surface. Therefore, the admissible displacements and stresses are defined in the smooth domain Ω, with the removed cut Γ_c. However, the restrictions imposed on the admissible functions are still present and can be considered as internal constraints prescribed on the given subset Γ_c of the smooth geometrical domain Ω. The resulting weak formulation of the nonlinear boundary value problem is defined over the smooth domain of integration. Such a formulation includes integral equations and inequalities. Applying this new approach to the elastic membrane problem in the domain with a cut, we analyze and implement the discretization by mixed finite element methods. The internal constraint in the model, which requires the nonpositivity of normal derivatives at crack faces, is expressed by means of a Lagrange multiplier.

*Received by the editors June 5, 2003; accepted for publication (in revised form) September 27, 2004; published electronically October 14, 2005.

http://www.siam.org/journals/sinum/43-3/42972.html

†Laboratoire de Mathématiques LMAM UMR 7122, Université de Metz, ISGMP, Batiment A, Ile du Saulcy, 57045 Metz, France (belhach@math.univ-metz.fr, jmse@math.univ-metz.fr).

‡Institut Elie Cartan, Laboratoire de Mathématiques, Université Henri Poincaré Nancy I, BP 239, 54506 Vandoeuvre lès Nancy Cedex, France (sokolows@iecn.u-nancy.fr).

The analysis of contact problems and their approximation by finite elements, or mixed finite elements methods, were performed by several authors (e.g., in [21], [22], [26], [9], [10], [19], [35], and the references therein). More recently, significant progress in the numerical analysis of such problems has been made. We refer the reader to, for example, [23], [3], [4] for the discretization by the mortar affine finite element method. In [14], [29] the mixed finite elements method is analyzed and implemented. We also refer to [34], [33] for a general setting to study some mixed variational formulations and the application to the discretization by the Raviart–Thomas elements of lowest order. In [33], an extension of the analysis of some mixed formulations arising in contact problems to the case where the first bilinear form is not coercive on the whole space is proposed and the numerical analysis is performed. Even if the problem we consider is set in a non-Lipschitzian domain, the mixed variational formulation that we obtain in the entire domain fits within the framework of [33]. However, our approach is different since we prefer to work with the hybrid mixed formulation. The tools of the numerical analysis and the results that we obtain in this article for our specific problem appear to be a complementary contribution to the analysis and discretization of contact problems already deeply performed in [5], [14], [34], [33].

The important feature of these approaches is the specific approximation of the nonpenetration conditions in the discrete model. For mixed finite element methods, the unilateral conditions can be expressed by introduction of either a piecewise constant [29], [33] or a piecewise continuous Lagrange multiplier [5], [14]. Such a construction is also crucial for the convergence analysis and numerical solution of the crack problem with the smooth domain formulation.

In this paper, we perform the convergence analysis and derive a priori error estimates for some mixed finite element methods that are also implemented numerically. Each approach among those considered in this work is characterized by the specific space of approximation for the Lagrangian multipliers. We obtain for the smooth domain method the convergence rate $O(h^{\frac{3}{4}})$ for the best choice of continuous piecewise affine approximation of the multipliers, exactly the same as the rate derived for the classical unilateral problems arising in contact mechanics. In fact, the smooth domain method can be considered as a mixed variational formulation, which takes into consideration the inequality conditions for the normal derivatives on the crack faces rather than the unilateral conditions for the jump of the displacement over the crack.

The paper is organized as follows. In section 2, the crack problem for an elastic membrane is introduced. The smooth domain formulation is given, and its well-posedness is established. Section 3 is devoted to the discretization of the continuous problem. First of all, a new formulation, based on the mixed variational method, is introduced and analyzed. The convergence analysis is performed in section 4. The error estimates and the convergence rates are established for the proposed approximations. In section 5, the numerical implementation is described in details, and some numerical examples are presented. Finally, some concluding remarks and perspectives are given.

2. Model problem.

We present the smooth domain method for a scalar model problem. First, appropriate notations are introduced.

Let Ω be a bounded domain in $\mathbb{R}^2$ with smooth boundary Γ, and $\Gamma_c \subset \Omega$ be a smooth curve without self-intersections. We assume that Γ_c can be extended to a closed smooth curve $\Sigma \subset \Omega$, with Σ of class $C^{1,1}$, and $\Omega = \Omega^1 \cup \Sigma \cup \Omega^2$ divided into two subdomains Ω^1, Ω^2 (see Figure 1). In this case, $\Sigma = \partial\Omega^1$ is the boundary of

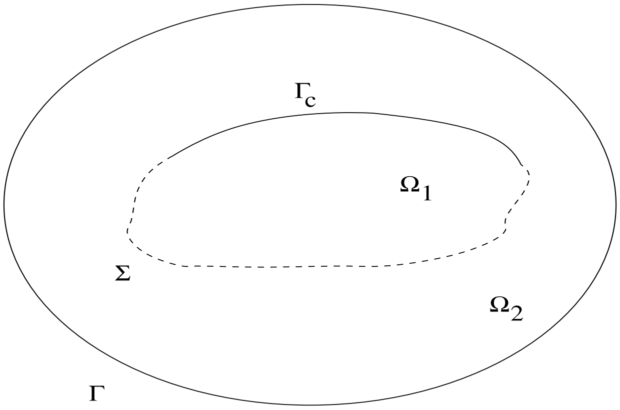


FIG. 1. *Domain with a cut.*

Ω^1 and $\Sigma \cup \Gamma = \partial\Omega^2$ is the boundary of Ω^2. Let Ω_c be the domain $\Omega \setminus \overline{\Gamma_c}$; then Γ_c is called a crack in the elastic body of the reference configuration Ω_c (see Figure 2). The static equilibrium problem for the elastic membrane in the domain Ω_c with the interior crack Γ_c can be formulated as follows.

Find u such that

$$-\Delta u = f \qquad \text{in } \Omega_c, \tag{1}$$

$$u = 0 \qquad \text{on } \Gamma, \tag{2}$$

$$[u] \geq 0, \qquad \left[\frac{\partial u}{\partial \nu}\right] = 0, \qquad [u]\,\frac{\partial u}{\partial \nu} = 0 \qquad \text{on } \Gamma_c, \tag{3}$$

$$\frac{\partial u}{\partial \nu} \leq 0 \qquad \text{on } \Gamma_c^{\pm}, \tag{4}$$

where f is a given function in $L^2(\Omega_c)$. The jump of the function u on Γ_c is denoted by $[u] = u^+ - u^-$, where $u^\pm = u|_{\Gamma_c^\pm}$ are the traces of u on $\Gamma_c^\pm$.

Boundary value problem (1)–(4) can be considered as a free boundary problem since the coincidence set $\Xi = \{x \in \Gamma_c | \, [u] = 0\}$ is an unknown part of Γ_c. When using the modelling in the framework of linear elasticity, similar boundary value problems arise in the crack theory for elastic bodies [25], [28]. In such a case the inequality type boundary conditions are imposed on Γ_c to describe mutual nonpenetration between the crack faces. It is the so-called frictionless contact problem on the crack. It is well-known that problem (1)–(4) admits a unique weak solution that minimizes the energy functional

$$\frac{1}{2}\int_{\Omega_c} |\nabla v|^2\, dx - \int_{\Omega_c} fv\, dx$$

over the closed convex cone

$$\mathbb{C}(\Gamma_c) = \{v \in H^1(\Omega_c) | v = 0 \quad \text{on} \quad \Gamma, \ [v] \geq 0 \quad \text{on} \quad \Gamma_c\}$$

including the functions in the Sobolev space $H^1(\Omega_c)$ which vanish on Γ and satisfy the unilateral condition $[v] \geq 0$ on Γ_c.

We denote by $V = V(\Omega_c)$ the space $L^2(\Omega_c)$ and consider the vector space $\mathbf{X} = \mathbf{X}(\Omega_c)$,

$$\mathbf{X} = \left\{\mathbf{q} \in L^2(\Omega_c)^2, \ \operatorname{div} \mathbf{q} \in L^2(\Omega_c)\right\}$$

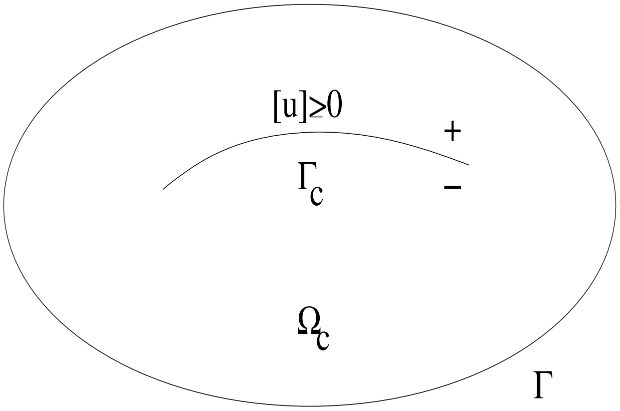


FIG. 2. *Elastic membrane.*

equipped with the norm

$$\|\mathbf{q}\|_{\mathbf{X}} = \left(\|\mathbf{q}\|^2_{(L^2(\Omega_c))^2} + \|\operatorname{div} \mathbf{q}\|^2_{L^2(\Omega_c)}\right)^{\frac{1}{2}}.$$

Remark 2.1. For the sake of simplicity, in the smooth domain formulation over Ω, we use the same notation for $V = V(\Omega)$ and for $\mathbf{X} = \mathbf{X}(\Omega)$, i.e., the domain Ω_c is replaced in both function spaces by the smooth domain Ω.

The closed convex cone $\mathbf{K} \subset \mathbf{X}$ can be defined using the dual order in the Sobolev space $(H^{\frac{1}{2}}_{00}(\Gamma_c))'$ as

$$\mathbf{K} = \left\{\mathbf{q} \in \mathbf{X},\ [\mathbf{q}.\nu] = 0,\ \text{on}\, \Gamma_c,\ (\mathbf{q}.\nu)^{\pm} \le 0,\ \text{on}\, \Gamma_c^{\pm}\right\},$$

where $[.]$ denotes the jump across Γ_c and ν is the unit normal vector pointing to the exterior of Ω_1 (see Figure 1). More precisely, we can use the integral inequalities and the cone $\mathbb{C}(\Gamma_c)$ in order to define $\mathbf{K}$ in an equivalent way,

$$\mathbf{K} = \left\{\mathbf{q} \in \mathbf{X},\ \int_{\Omega_c} \mathbf{q} \cdot \nabla v + v \operatorname{div}\mathbf{q} \ge 0 \quad \forall v \in \mathbb{C}(\Gamma_c)\right\}.$$

Under our assumptions, functions in $H^1(\Omega_c)$ admit traces on the curve Σ, and therefore on the crack Γ_c. In order to characterize such traces as elements of fractional Sobolev spaces and introduce the dual order in the dual spaces, we need some notations. We introduce the space $H^{\frac{1}{2}}(\Sigma)$ equipped with the norm

$$\|\varphi\|^2_{H^{\frac{1}{2}}(\Sigma)} = \|\varphi\|^2_{L^2(\Sigma)} + \int_\Sigma \int_\Sigma \frac{|\varphi(x) - \varphi(y)|^2}{|x-y|^2}\, dx\, dy,$$

and we denote by $H^{-\frac{1}{2}}(\Sigma)$ its dual space. For $\mathbf{q} \in \mathbf{X}$ the traces $(\mathbf{q}.\nu)^{\pm}$ on $\Sigma^{\pm}$ can be defined as elements of $H^{-\frac{1}{2}}(\Sigma)$ and the trace operator is continuous from $\mathbf{X}$ into $H^{-\frac{1}{2}}(\Sigma)$. The space $H^{\frac{1}{2}}_{00}(\Gamma_c)$ is the subspace of $H^{\frac{1}{2}}(\Gamma_c)$ which can be identified [30] with the subspace of functions in $H^{\frac{1}{2}}(\Sigma)$ vanishing on the set $\Sigma \setminus \Gamma_c$. So, formally $H^{\frac{1}{2}}_{00}(\Gamma_c)$ includes the functions vanishing at the endpoints of Γ_c (see [30]). We can define the traces $(\mathbf{q}.\nu)^{\pm} \in (H^{\frac{1}{2}}_{00}(\Gamma_c))'$ and the constraints in the definition of $\mathbf{K}$ can be formulated using the duality, i.e., $[\mathbf{q}.\nu]=0$ means that

$$\langle [\mathbf{q}.\nu]\,,\varphi\rangle_{\frac{1}{2},\Gamma_c}=0 \quad \forall\varphi\in H_{00}^{\frac{1}{2}}(\Gamma_c),$$

and if the jump condition $[\mathbf{q}.\nu]=0$ is combined with the sign condition $(\mathbf{q}.\nu)^{\pm}\leq 0$ it is simply required that

$$\langle (\mathbf{q}.\nu)^{\pm},\varphi\rangle_{\frac{1}{2},\Gamma_c}\leq 0 \quad \forall\varphi\in H_{00}^{\frac{1}{2}}(\Gamma_c),\ \varphi\geq 0,\ \text{a.e. on } \Gamma_c,$$

where $\langle .,.\rangle_{\frac{1}{2},\Gamma_c}$ stands for the duality pairing between $H_{00}^{\frac{1}{2}}(\Gamma_c)$ and its dual $(H_{00}^{\frac{1}{2}}(\Gamma_c))'$.

We can consider the following equalities in the sense of distributions, i.e., in $\mathcal{D}'(\Omega_c)$,

$$\mathbf{p}=\mathbf{grad}\,u \quad \text{in } \Omega_c, \tag{5}$$

and then we have

$$-\mathrm{div}\,\mathbf{p}=f \quad \text{in } \Omega_c. \tag{6}$$

The mixed formulation of boundary value problem (1)–(4) can be written as follows. Find $(\mathbf{p},u)\in\mathbf{K}\times V$ such that

$$\left\{\begin{array}{l} \int_{\Omega_c}\mathbf{p}\,(\mathbf{q}-\mathbf{p})\,dx+\int_{\Omega_c}u\,(\mathrm{div}\,\mathbf{q}\,-\,\mathrm{div}\,\mathbf{p})\,dx\geq 0 \quad \forall\mathbf{q}\in\mathbf{K},\\ -\int_{\Omega_c}\mathrm{div}\,\mathbf{p}\,v\,dx=\int_{\Omega_c}f\,v\,dx \quad \forall v\in V. \end{array}\right. \tag{7}$$

Note that $\mathbf{p}=\mathbf{grad}u$ and u is the solution of problem (1)–(4).

Remark 2.2. The mixed formulation (7) can also be applied in the case of the set $\overline{\Gamma}_c$ crossing the external boundary Γ as well as in the case of Γ_c, which is less regular and belongs only to the class $C^{0,1}$ (see, e.g., [27]). In the case of Γ_c, which divides Ω into two disjoint Lipschitz subdomains Ω^1 and Ω^2, and for the inequality type boundary conditions (3)–(4) prescribed on Γ_c, the mixed formulation of the contact problem for two elastic bodies occupying Ω^1 and Ω^2 is obtained.

Remark 2.3. The smooth domain method can also be used for the modelling of a large class of crack problems with inequality type conditions prescribed on the crack including the most difficult, from a mathematical point of view, contact problem with Coulomb friction law. In particular, the field of applications includes the linear elastic bodies and the Kirchhoff plate models [25]. As an example of modelling in the framework of elasticity, we can consider the frictionless contact problem that takes the form of an equilibrium boundary value problem for the linear elastic body occupying the domain of reference Ω_c with the interior crack Γ_c. Such a problem can be formulated as follows.

Find $\mathbf{u}=(u_1,u_2)$ and $\sigma=(\sigma)_{ij}$, $i,j=1,2$, such that

$$-\mathbf{div}\,\sigma=\mathbf{f} \qquad \text{in } \Omega_c, \tag{8}$$

$$C\sigma-\varepsilon(\mathbf{u})=0 \qquad \text{in } \Omega_c, \tag{9}$$

$$\mathbf{u}=0 \qquad \text{on } \Gamma, \tag{10}$$

$$[\mathbf{u}]\,\nu\geq 0, \qquad [\sigma_\nu]=0, \qquad \sigma_\nu\,[\mathbf{u}]\cdot\nu=0 \qquad \text{on } \Gamma_c, \tag{11}$$

$$\sigma_\nu\leq 0, \qquad \sigma_\tau=0 \qquad \text{on } \Gamma_c^{\pm}. \tag{12}$$

Here

$$\sigma_\nu = \sigma_{ij}\nu_j\nu_i,\ \sigma_\tau = \sigma\nu - \sigma_\nu = \left\{\sigma_\tau^i\right\}_{i=1}^2,\ \sigma\nu = \left\{\sigma_{ij}\nu_j\right\}_{i=1}^2,$$
$$\varepsilon_{ij}(\mathbf{u}) = \frac{1}{2}(u_{i,j} + u_{j,i}),\ i,j = 1,2,\ \varepsilon(\mathbf{u}) = (\varepsilon_{ij})_{i,j=1}^2,$$
$$\left\{C\sigma\right\}_{ij} = c_{ijk\ell}\sigma_{k\ell},\ c_{ijk\ell} = c_{jik\ell} = c_{k\ell ij},\ c_{ijk\ell} \in L^\infty(\Omega).$$

The tensor C satisfies the ellipticity condition

$$c_{ijk\ell}\xi_{ji}\xi_{k\ell} \geq c_0|\xi|^2 \quad \forall\xi_{ji} = \xi_{ij},\ c_0 > 0, \tag{13}$$

and we have used the summation convention over repeated indices.

This problem can also be reformulated in the framework of the smooth domain method. Then the complete numerical analysis performed in the paper can be extended to the boundary value problem for frictionless contact conditions on the crack. The most challenging task is to extend the numerical results to the case of contact problems with friction.

Now we are in position to present the smooth domain method that allows us to solve numerically the crack problem in the smooth domain Ω. An application of the method means that the functions $\mathbf{p}$ and u are extended to the entire domain $\Omega = \Omega_c \cup \Gamma_c$, and it results in the closed problem formulation (14) obtained by replacing Ω_c with Ω in (7), with the obvious modification of the function spaces defined now in Ω, which are still denoted by $\mathbf{X} = \mathbf{X}(\Omega)$ and $V = V(\Omega)$,

$$\left\{\begin{array}{l} \int_\Omega \mathbf{p}\,(\mathbf{q} - \mathbf{p})\,dx + \int_\Omega u\,(\operatorname{div}\mathbf{q}\, -\, \operatorname{div}\mathbf{p})\,dx \geq 0 \quad \forall\mathbf{q} \in \mathbf{K}, \\ -\int_\Omega \operatorname{div}\mathbf{p}\,v\,dx = \int_\Omega f\,v\,dx \quad \forall v \in V. \end{array}\right. \tag{14}$$

$\mathbf{K} = \mathbf{K}(\Omega)$ denotes the closed convex set

$$\mathbf{K} = \left\{\mathbf{q} \in L^2(\Omega)^2,\ \operatorname{div}\mathbf{q} \in L^2(\Omega),\ \mathbf{q}.\nu \leq 0,\ \text{on}\ \ \Gamma_c\right\}.$$

The well-posedness of smooth domain formulation (14) is proved in [25] with arguments based on the regularization technique (a similar argument used in [33] yields the same result). We briefly recall the main idea of the regularization technique that we also use in the discretization of the problem. Actually, for the regularization parameter $\delta > 0$, we consider the equation

$$\delta u_\delta - \operatorname{div}\mathbf{p}_\delta = f \quad \text{in } \Omega, \tag{15}$$

which leads to the following regularized formulation: Find $(\mathbf{p}_\delta, u_\delta) \in \mathbf{K} \times V$ such that

$$\left\{\begin{array}{l} a_\delta(u_\delta, v_\delta) + b(v_\delta, \mathbf{p}_\delta) = (f, v_\delta) \quad \forall v_\delta \in V, \\ -b(u_\delta, \mathbf{q}_\delta - \mathbf{p}_\delta) + g(\mathbf{p}_\delta, \mathbf{q}_\delta - \mathbf{p}_\delta) \geq 0 \quad \forall\mathbf{q}_\delta \in \mathbf{K}, \end{array}\right. \tag{16}$$

where we have

$$a_\delta(u, v) = \delta\int_\Omega u\,v\,dx,$$

$$b(v, \mathbf{q}) = -\int_\Omega v\operatorname{div}\mathbf{q}\,dx,$$

and

$$g(\mathbf{p}, \mathbf{q}) = \int_\Omega \mathbf{p}\,\mathbf{q}\,dx.$$

The bilinear form $a_\delta(.,.)$ is continuous and V-elliptic, and $b(.,.)$ is continuous and satisfies the inf-sup condition. This yields the following result.

THEOREM 2.4. *Problem* (16) *admits a unique solution* $(u_\delta, \mathbf{p}_\delta)$ *for any* $f \in L^2(\Omega)$. *Moreover, we have the following stability estimate:*

$$\|u_\delta\|_V + \|\mathbf{p}_\delta\|_X \leq c\|f\|_{L^2(\Omega)}, \tag{17}$$

and the sequence of the regularized solutions $(u_\delta, \mathbf{p}_\delta)_\delta$ *converges to* $(u, \mathbf{p})$ *the solution of problem* (14) *when* δ *goes to zero.*

Remark 2.5. Note that the interpretation of problem (16) leads to

$$\mathbf{p}_\delta = \mathbf{grad}\, u_\delta,$$

and also it gives $u_\delta = 0$ on Γ; thus u_δ belongs to $H_0^1(\Omega)$. It follows also from (15) that if f is in $H^1(\Omega)$, then $\mathbf{p}_\delta$ belongs to the space

$$\mathbf{Z} = \left\{\mathbf{q} \in (H^1(\Omega))^2;\ \operatorname{div} \mathbf{q} \in H^1(\Omega)\right\}.$$

However, the convergence of u_δ to u is to be understood in the weak $H^1(\Omega)$-norm and the convergence of $\mathbf{p}_\delta$ holds in the weak $H(\operatorname{div}, \Omega)$-norm [25].

There are only very few results on the regularity of the solutions to crack problems with unilateral conditions. However, proceeding as for the unilateral contact problems [31], it can readily be checked that the unilateral conditions may also generate some singularities of solutions in the vicinity of the tips of Γ_c even for a regular datum f and the smooth exterior boundary $\partial\Omega$. For example, if $f \in H^1(\Omega)$, the solution u may not be of class H^3 in the vicinity of the crack Γ_c (see [31]). The reason for the lack of regularity can be explained in the following way. Let $\mathbf{m}$ be a point on Γ_c where the unilateral constraints change from binding to nonbinding; then the Dirichlet–Neumann singular function $S_\mathbf{m}(r, \vartheta) = r^{\frac{3}{2}} \sin(\frac{3}{2}\vartheta)\varphi(r)$ appears in the decomposition of the solution. Here (r, ϑ) are the polar coordinates with the origin at $\mathbf{m}$, and φ is a smooth function with the compact support, which equals 1 in the vicinity of $\mathbf{m}$. We refer the reader to [20] for the details on the decompositions of solutions to elliptic equations into singular and regular parts. The first Dirichlet–Neumann singular function $r^{\frac{1}{2}} \sin(\frac{1}{2}\vartheta)\psi(r)$ does not appear in the decomposition, since it fails to satisfy the required unilateral conditions—i.e., the nonnegativity of $S_\mathbf{m}$ and $\frac{\partial S_\mathbf{m}}{\partial n}$, simultaneously. Therefore, following [31] (see also [17]), we can apply the results derived for the Signorini problem and show that under appropriate symmetry conditions for f and Ω, it can be expected at most that $u \in H^\sigma(\mathcal{O}(\Gamma_c))$ with $\sigma < \frac{5}{2}$, where $\mathcal{O}(\Gamma_c)$ is an open set containing $\overline{\Gamma_c}$.

3. Discrete variational formulation.

3.1. The discrete regularized problem. We propose a discretization of the continuous variational formulation. First, we note that the bilinear form $g(.,.)$ is not elliptic on $\mathbf{X}$; therefore, the regularized problem (16) is considered for the approximation analysis. In fact, the ellipticity of $g(.,.)$ on the subspace of divergence free vector fields is a sufficient condition in order to apply the saddle-point theory. However, the regularization technique leads to a simple numerical method and, acting as a stabilized formulation, allows more flexibility in the choice of discretization spaces. In what follows, the regularization parameter $\delta > 0$ is omitted in our notation; e.g., we write u_h instead of $u_{\delta,h}$ for the sake of simplicity.

The finite element family that we have chosen in the discretization is based on the element called the Taylor–Hood mini element coming from the fluid dynamic. It consists of the approximation of the displacement u_h by piecewise affine polynomials and the pressure p_h by the P_1-bubble element. Other (more standard) choices, such as elements based on the Raviart–Thomas finite element, are currently considered in the implementation. The reason of our choice is that we expect the solutions of our problem to have the same regularity as those of the usual Signorini problem; thus, we choose an element that requires few degrees of freedom and provides “high” accuracy. In fact this element is an intermediary choice between RT0 and RT1.

We assume, to avoid curved elements, that Ω is polygonal and the crack Γ_c is a straight segment. We denote by $\mathcal{T}_h$ a triangulation of Ω. $\mathcal{T}_h$ is a family of elements which are triangles (or quadrilaterals), and the maximal size of elements is the parameter of approximation denoted by $h > 0$, in addition satisfying the usual admissibility assumptions. That is, the intersection of two different elements is either empty, a vertex, or a whole edge. Furthermore, $\mathcal{T}_h$ is assumed to be regular; i.e., the ratio of the diameter of any element $T \in \mathcal{T}_h$ to the diameter of its largest inscribed ball is bounded by a constant σ independent of T and h. We also assume that the endpoints of Γ_c are vertices of the triangulation. The nodes on Γ_c are denoted by $c_1 = x_0, x_1, \ldots, x_{I-1}, x_I = c_2$, and we set $t_i =]x_{i-1}, x_i[$ and $|t_i| = |x_i - x_{i-1}|$. We will assume for simplicity that the triangulation $\mathcal{T}_h$ is quasi-uniform, i.e., there is a constant $\tau > 0$ such that

$$\frac{\max_T \ h_T}{\min_T \ h_T} \leq \tau.$$

Remark 3.1. This assumption could be weakened. Indeed, if some general meshes are considered (for example in some adaptivity process), we can avoid such assumption and only assume that the 1D triangulation on Γ_c satisfies the following criterion (due to Crouzeix and Thomée [13]):

$$\frac{|t_i|}{|t_j|} \leq C\beta^{|i-j|} \quad \forall i, j \, (0 \leq i, j \leq I-1), \tag{18}$$

with $1 \leq \beta < 4$ and a constant C independent of h. In fact, this last condition is sufficient to prove the continuity property of the projection operators (for the appropriate norms), which is equivalent to the inf-sup conditions [11], [5]. In the case of quasi-uniform meshes, this continuity property results directly from inverse inequalities.

We introduce the following finite dimensional space for $h > 0$:

$$V_h = \left\{ v_h \in C(\overline{\Omega}), \ v_{h|T} \in P_1(T) \right\}.$$

For each element T in $\mathcal{T}_h$, the associated bubble function φ_T is defined by

$$\varphi_T(x) = \prod_{i=1}^{3} \lambda_i(x) \quad \forall x \in T,$$

where λ_i denotes the ith barycentric coordinate in T. $\mathcal{P}_1(T)$ stands for the space of the first-order polynomials over T, and $P_B(T) = P_1(T) \oplus \mathbb{R}\varphi_T$ is selected as the local approximation space for the vector fields in $\mathbf{X}$. Hence the global approximation space takes the form

$$\mathbf{X}_h = \left\{ \mathbf{q}_h \in C(\overline{\Omega})^2,\ \mathbf{q}_h \in P_B(T)^2 \right\},$$

and the following closed convex set is introduced for approximation of $\mathbf{K}$:

$$\mathbf{K}_h = \{ \mathbf{q}_h \in \mathbf{X}_h,\ \mathbf{q}_h.\nu \leq 0,\ \text{on}\, \Gamma_c \}.$$

Due to the lack of regularity for solutions of crack problems, we restrict ourselves to affine finite elements. The numerical simulations show in some cases an additional regularity of solutions, and therefore the use of higher order finite elements could be advantageous [2]. We point out that the choice of the discrete spaces V_h and $\mathbf{X}_h$ can be modified by choosing any other classical lower order finite elements spaces. Some possible choices, such as Raviart–Thomas elements, will be considered in a forthcoming paper. Note also that, for the choice presented in this paper, the discrete spaces satisfy the usual inf-sup condition that we recall now (see [8]): There exists $\gamma > 0$ independent of h such that

$$\forall v_h \in V_h,\ \sup_{\mathbf{q}_h \in \mathbf{X}_h} \frac{|b(v_h, \mathbf{q}_h)|}{\|\mathbf{q}_h\|_{\mathbf{X}}} \geq \gamma \|v_h\|_V. \tag{19}$$

By $V_h(\Omega^\ell)$ and $\mathbf{X}_h(\Omega^\ell)$, $\ell = 1, 2$, we denoted the finite dimensional spaces of functions of V_h and $\mathbf{X}_h$, respectively, restricted to the subdomains Ω^ℓ.

The discrete problem is defined in the following way.
Find $(u_h, \mathbf{p}_h) \in V_h \times \mathbf{K}_h$ such that

$$\left\{ \begin{array}{l} a_\delta(u_h, v_h) + b(v_h, \mathbf{p}_h) = (f, v_h) \quad \forall v_h \in V_h, \\ \\ -b(u_h, \mathbf{q}_h - \mathbf{p}_h) + g(\mathbf{p}_h, \mathbf{q}_h - \mathbf{p}_h) \geq 0 \quad \forall \mathbf{q}_h \in \mathbf{K}_h. \end{array} \right. \tag{20}$$

THEOREM 3.2. *Problem* (20) *admits a unique solution* $(u_h, \mathbf{p}_h)$ *for* $h > 0$. *Moreover, we have the following stability estimate, uniform with respect to h:*

$$\|u_h\|_V + \|\mathbf{p}_h\|_X \leq c \|f\|_{L^2(\Omega)}. \tag{21}$$

Proof. The proof in the discrete case is similar to the continuous problem (see [33, Theorem 2.3]). Indeed, the existence is based on a perturbation technique applied to the variational inequality obtained by summing the two lines of (20), then applying Stampacchia's theorem and, thanks to a uniform a priori estimate (with respect to h), passing to the limit.

It remains to prove the uniqueness of the solution of problem (20), which follows from direct computations. Assume there exist two solutions $(u_{hi}, \mathbf{p}_{hi})$, $i = 1, 2$. Choosing $\mathbf{q}_h = \mathbf{p}_{h2}$ in the second line of (20) with the first solution, respectively, $\mathbf{q}_h = \mathbf{p}_{h1}$ with the second solution, and summing the resulting inequalities yield

$$-b(u_{h2} - u_{h1}, \mathbf{p}_{h2} - \mathbf{p}_{h1}) + g(\mathbf{p}_{h2} - \mathbf{p}_{h1}, \mathbf{p}_{h2} - \mathbf{p}_{h1}) \leq 0.$$

Subtracting the first lines of problem (20),

$$\begin{array}{l} a(u_{h1}, v_h) + b(v_h, \mathbf{p}_{h1}) = (f, v_h), \\ a(u_{h2}, v_h) + b(v_h, \mathbf{p}_{h2}) = (f, v_h), \end{array}$$

and choosing $v_h = u_{h2} - u_{h1}$ lead to

$$-b(u_{h2} - u_{h1}, \mathbf{p}_{h2} - \mathbf{p}_{h1}) = a_\delta(u_{h2} - u_{h1}, u_{h2} - u_{h1}) \geq 0;$$

thus $g(\mathbf{p}_{h2} - \mathbf{p}_{h1}, \mathbf{p}_{h2} - \mathbf{p}_{h1}) \leq 0$, which implies $\mathbf{p}_{h1} = \mathbf{p}_{h2}$, and we deduce that $u_{h1} = u_{h2}$. □

For the convergence analysis and the numerical simulations we will introduce a new discrete formulation of problem (20) in the framework of the saddle-point approach.

3.2. New discrete formulation. The numerical analysis of mixed variational formulations for unilateral frictionless contact problems is performed, e.g., in [14], [29], [5], [33]. There are some similarities between modelling of contact problems and the smooth domain models of crack problems considered in this paper. Therefore, the same saddle-point framework, as in the case of contact problems, can be used for the convergence analysis of proposed finite element approximations for the smooth domain method.

First, the following energy functional is defined over the convex set $V \times \mathbf{K}$:

$$\mathcal{J}(v, \mathbf{q}) = \frac{1}{2} a_\delta(v, v) - (f, v) - b(v, \mathbf{q}) + \frac{1}{2} g(\mathbf{q}, \mathbf{q}).$$

Clearly, solution of problem (20) is equivalent to the minimization problem for $\mathcal{J}$,

$$\mathcal{J}(u, \mathbf{p}) = \min_{(v,\mathbf{q}) \in V \times \mathbf{K}} \mathcal{J}(v, \mathbf{q}). \tag{22}$$

The constraints in $\mathbf{K}$ can be defined by duality using the closed convex cone

$$M = \left\{ \mu \in H^{\frac{1}{2}}(\Gamma_c);\ \mu \geq 0 \right\};$$

thus, we can write equivalently problem (16) as follows: Find $(u, \mathbf{p}, \lambda) \in V \times \mathbf{X} \times M$ such that

$$\left\{ \begin{array}{l} a_\delta(u, v) + b(v, \mathbf{p}) = (f, v) \quad \forall v \in V, \\ \\ -b(u, \mathbf{q}) + g(\mathbf{p}, \mathbf{q}) + \langle \lambda, \mathbf{q}.\nu \rangle_{\frac{1}{2}, \Gamma_c} = 0 \quad \forall \mathbf{q} \in \mathbf{X}, \\ \\ \langle \mu - \lambda, \mathbf{p}.\nu \rangle_{\frac{1}{2}, \Gamma_c} \leq 0 \quad \forall \mu \in M, \end{array} \right. \tag{23}$$

where the elements $(u, \mathbf{p}, \lambda)$ are given by a saddle-point of the Lagrangian

$$\mathcal{L}(v, \mathbf{q}, \mu) = \mathcal{J}(v, \mathbf{q}) + \langle \mu, \mathbf{q}.\nu \rangle_{\frac{1}{2}, \Gamma_c},$$

defined over the product $V \times \mathbf{X} \times M$. Note that λ is the Lagrange multiplier associated to the inequality constraint $\mathbf{p}.\nu \leq 0$ on Γ_c.

PROPOSITION 3.3. *Problem* (23) *admits a unique solution* $(u, \mathbf{p}, \lambda) \in V \times \mathbf{X} \times M$ *for any* $\delta > 0$. *Moreover,*

$$\lambda = [u] \ \text{ on } \Gamma_c,$$

and $(u, \mathbf{p})$ *is the solution of problem* (16).

In order to define a finite dimensional approximation of problem (23) we need an approximation of the inequality constraints in $\mathbf{K}$. For this purpose, we introduce two finite dimensional spaces of scalar functions on Γ_c,

$$W_h^0(\Gamma_c) = \left\{\mu_h,\ \mu_{h|t_i} \in \mathbb{P}_0(t_i),\ 0 \le i \le I-1\right\},$$
$$W_h^1(\Gamma_c) = \left\{\mu_h \in C(\overline{\Gamma}_c),\ \exists \mathbf{q}_h \in \mathbf{X}_h,\ \text{such that } \mathbf{q}_h.\nu = \mu_h \text{ on } \Gamma_c\right\}.$$

Now, we are in position to define the finite dimensional approximations of the set M. The first set M_h^0 is an external approximation of M since piecewise constant functions are not elements of the fractional Sobolev space $H^{\frac{1}{2}}$. Thus, with the choice of the space $W_h^0(\Gamma_c)$ we have the definition

$$M_h^0 = \left\{\mu_h \in W_h^0(\Gamma_c),\ \mu_h \ge 0,\ \text{on } \Gamma_c\right\}, \tag{24}$$

and the choice of $W_h^1(\Gamma_c)$ leads to the following two approximation sets:

$$M_h^1 = \left\{\mu_h \in W_h^1(\Gamma_c),\ \mu_h \ge 0 \text{ on } \Gamma_c\right\} \tag{25}$$

and

$$M_h^{1,*} = \left\{\mu_h \in W_h^1(\Gamma_c),\ \int_{\Gamma_c} \mu_h \psi_h \, d\Gamma \ge 0 \,\forall \psi_h \in M_h^1\right\}. \tag{26}$$

We note that the set M_h^1 is well defined since μ_h is a continuous function, piecewise P_1 so that the nonnegativity condition can only be imposed at the nodes of the 1D mesh of Γ_c.

Remark 3.4. It can be expected that the choice of the set M_h^0, for the approximation of Lagrangian multipliers as elements of $H^{\frac{1}{2}}(\Gamma_c)$, leads to worse results compared to the two other possibilities listed above. This fact is confirmed by the convergence analysis, at least when we compare with the choice of $M_h^{1,*}$. We consider such an approximation (with M_h^0) for the sake of completeness (since it is sometimes used in computations in contact mechanics).

The solution $(u_h, \mathbf{p}_h, \lambda_h) \in V_h \times \mathbf{X}_h \times M_h$ of the finite dimensional mixed approximation of (23) satisfies the following discrete system of equations and inequalities:

$$\begin{cases} a_\delta(u_h, v_h) + b(v_h, \mathbf{p}_h) = (f, v_h) \quad \forall v_h \in V_h, \\ -b(u_h, \mathbf{q}_h) + g(\mathbf{p}_h, \mathbf{q}_h) + \int_{\Gamma_c} (\lambda_h)(\mathbf{q}_h.\nu)\, d\sigma = 0 \quad \forall \mathbf{q}_h \in \mathbf{X}_h, \\ \int_{\Gamma_c} (\mu_h - \lambda_h)(\mathbf{p}_h.\nu)\, d\sigma \le 0 \quad \forall \mu_h \in M_h, \end{cases} \tag{27}$$

where M_h is a specific multiplier set defined by (24), (25), or in (26).

NOTATION . *$d(.,.)$ is a bilinear form defined on $M \times (V \times \mathbf{X})$ by*

$$d(\mu, \mathbf{V}) = \langle (\mathbf{q}.\nu), \mu \rangle_{\frac{1}{2},\Gamma_c},$$

where $\mathbf{V} = (v, \mathbf{q}) \in V \times \mathbf{X}$.

The uniqueness of $(u_h, \mathbf{p}_h)$ as a solution of (27) follows directly from the uniqueness argument in the proof of Theorem 3.2.

Therefore, to prove the existence and uniqueness of a saddle-point in (27), it is sufficient to verify that

$$\{\mu_h \in M_h,\ d(\mu_h, \mathbf{V}) = 0 \quad \forall \mathbf{V} \in V_h \times \mathbf{X}_h\} = \{0\},$$

which is straightforward in the case $M_h = M_h^1$ or $M_h = M_h^{1,*}$ and which still holds for $M_h = M_h^0$ by the inf-sup condition (29) under condition (18). Thus, the following result is obtained.

PROPOSITION 3.5. *Assume that the set M_h is given by $M_h = M_h^0$, $M_h = M_h^1$, or $M_h = M_h^{1,*}$. For each of the choices, problem* (27) *admits a unique solution.*

In order to perform the convergence analysis of finite element methods, we first check the compatibility condition (inf-sup condition) between the spaces $W_h^1(\Gamma_c)$ and $V_h \times \mathbf{X}_h$, respectively, the compatibility condition between W_h^0 and $V_h \times \mathbf{X}_h$. These conditions are necessary to obtain optimal stability results.

The first inf-sup condition. This condition between W_h^0 and $V_h \times \mathbf{X}_h$ is obtained under the assumption (18) in [5, Lemma 6.3] with W_h^0 equipped with the $H^{-\frac{1}{2}}$-norm. For our purpose, we need the inf-sup condition in the mesh-dependent norm

$$\|\mu_h\|_L = \left(\sum_{i=1}^{I-1} h_{t_i} \|\mu_h\|^2_{L^2(t_i)} \right)^{\frac{1}{2}}.$$

We denote by $\|\mu_h\|_{L^{-1}} = \left(\sum_{i=1}^{I-1} h_{t_i}^{-1} \|\mu_h\|^2_{L^2(t_i)}\right)^{\frac{1}{2}}$ its dual norm. Therefore, it is standard [11] that the inf-sup condition follows from the stability with respect to the norm $\|.\|_{L^{-1}}$ of the L^2-projection operator $\pi_h^0 : L^2(\Gamma_c) \longrightarrow W_h^0$:

$$\int_{\Gamma_c} v\psi_h \, d\sigma = \int_{\Gamma_c} \pi_h^0(v)\psi_h \, d\sigma, \quad \psi_h \in W_h^0. \tag{28}$$

Noting that (28) is a linear system of the form $y = Dx$, with a diagonal (positive definite) matrix D, the stability property for π_h^0 follows by straightforward computations. Thus, we derive the following inf-sup condition.

PROPOSITION 3.6. *There exists $\beta_0 > 0$ such that*

$$\forall \mu_h \in W_h^0, \quad \sup_{\mathbf{V}_h \in V_h \times \mathbf{X}_h} \frac{d(\mu_h, \mathbf{V}_h)}{\|\mathbf{V}_h\|_{V\times\mathbf{X}}} \geq \gamma_0 \,\|\mu_h\|_L \geq \beta_0 \, h^{\frac{1}{2}} \|\mu_h\|_{L^2(\Gamma_c)}. \tag{29}$$

The operator π_h^0 satisfies the following estimates proved in [29]. Namely, for the functions $\varphi \in H^\nu(\Gamma_c)$ with $\nu = \frac{1}{2}$, or with $\nu = 1$, there exists a constant $c > 0$ independent of h such that

$$\|\varphi - \pi_h^0\varphi\|_{L^2(\Gamma_c)} \leq ch^\nu \|\varphi\|_{H^\nu(\Gamma_c)}. \tag{30}$$

Moreover, if $\varphi \in L^2(\Gamma_c)$, then

$$\|\varphi - \pi_h^0\varphi\|_{H^{-\frac{1}{2}}(\Gamma_c)} \leq ch^{\frac{1}{2}} \|\varphi - \pi_h^0\varphi\|_{L^2(\Gamma_c)}. \tag{31}$$

In addition, we have the following property of the operator: For $\varphi \geq 0$ it follows that $\pi_h^0\varphi \in M_h^0$.

The second inf-sup condition. The second condition is stated in Proposition 3.7 and requires us to introduce some additional tools. We define the projection operator $\pi_h^1 : L^2(\Gamma_c) \mapsto W_h^1(\Gamma_c)$, with respect to the scalar product in $L^2(\Gamma_c)$, which satisfies the following properties proved in [7].

Given $\mu \in [0,1]$ and $\nu \in \left]\frac{1}{2}, 2\right]$, there exists a constant $c > 0$ that is independent of h, such that for all functions $\varphi \in H^\nu(\Gamma_c)$,

$$\|\varphi - \pi_h^1\varphi\|_{H^{-\mu}(\Gamma_c)} + h^{\mu+\frac{1}{2}}\|\varphi - \pi_h^1\varphi\|_{H^{\frac{1}{2}}(\Gamma_c)} \leq ch^{\mu+\nu}\|\varphi\|_{H^\nu(\Gamma_c)}. \tag{32}$$

Proposition 3.7. *There exists $\beta > 0$, independent of h, such that the following inf-sup condition holds:*

$$\sup_{\mathbf{V}\in V_h\times\mathbf{X}_h} \frac{d(\mu_h, \mathbf{V})}{\|\mathbf{V}\|} \geq \beta\|\mu_h\|_{H^{\frac{1}{2}}(\Gamma_c)} \quad \forall \mu_h \in W_h^1. \tag{33}$$

Proof. Let $\mu_h \in W_h^1(\Gamma_c)$. We want to construct $\mathbf{q}_h \in \mathbf{X}_h$ such that

$$d(\mu_h, (0, \mathbf{q}_h)) \geq \|\mu_h\|^2_{H^{\frac{1}{2}}(\Gamma_c)} \quad \text{and} \quad \beta\|\mathbf{q}_h\|_{H^1(\Omega)^2} \leq \|\mu_h\|_{H^{\frac{1}{2}}(\Gamma_c)}. \tag{34}$$

Let us consider $\mathbf{q} \in H^1(\Omega)^2$ such that

$$\int_\Omega \mathbf{q}\cdot\mathbf{w}\,d\mathbf{x} + \int_\Omega (\mathbf{grad}\,\mathbf{q}) : (\mathbf{grad}\,\mathbf{w})\,d\mathbf{x} = d(\mu_h, (0, \mathbf{w})) \quad \forall \mathbf{w} \in H^1(\Omega)^2, \tag{35}$$

where

$$\mathbf{grad}\,\mathbf{q} : \mathbf{grad}\,\mathbf{w} = \sum_{i,j=1}^{2} \frac{\partial \mathbf{q}_j}{\partial x_i}\frac{\partial \mathbf{w}_j}{\partial x_i}.$$

The following stability inequalities hold:

$$c_1\|\mu_h\|_{H^{\frac{1}{2}}(\Gamma_c)} \leq \|\mathbf{q}\|_{H^1(\Omega)^2} \leq c_2\|\mu_h\|_{H^{\frac{1}{2}}(\Gamma_c)}. \tag{36}$$

Note that the first inequality comes from the continuous inf-sup condition of $d(.,.)$ and (35).

Then we set $\mathbf{q}_h \in \mathbf{X}_h$ such that $(\mathbf{q}_h.\nu)_{\Gamma_c} = \pi_h^1((\mathbf{q}.\nu)_{\Gamma_c})$ and

$$\|\mathbf{q}_h\|_{H^1(\Omega)^2} \leq c\|\pi_h^1(\mathbf{q}.\nu)\|_{H^{\frac{1}{2}}(\Gamma_c)} \leq c\|\mathbf{q}.\nu\|_{H^{\frac{1}{2}}(\Gamma_c)}. \tag{37}$$

Such $\mathbf{q}_h$ is built using a stable finite element extension operator studied in [6]. Next, observe that we have

$$d(\mu_h, (0, \mathbf{q}_h)) = d(\mu_h, (0, \mathbf{q})) = \|\mathbf{q}\|^2_{H^1(\Omega)^2} \geq c_1^2\|\mu_h\|^2_{H^{\frac{1}{2}}(\Gamma_c)},$$

which yields the first statement of (34). The second one is obtained from (36) and (37). □

4. Convergence analysis.

In this section we perform the convergence analysis and give the error estimates and the rates of convergence. Let us define a bilinear form $A(.,.)$ on the product space $V \times \mathbf{X}$ by

$$\begin{aligned} A(\mathbf{U} - \mathbf{U}_h, \mathbf{U} - \mathbf{V}_h) = {} & a_\delta(u - u_h, u - v_h) + b(u - v_h, \mathbf{p} - \mathbf{p}_h) \\ & -b(u - u_h, \mathbf{p} - \mathbf{q}_h) + g(\mathbf{p} - \mathbf{p}_h, \mathbf{p} - \mathbf{q}_h), \end{aligned}$$

where $\mathbf{U} = (u, \mathbf{p})$, $\mathbf{U}_h = (u_h, \mathbf{p}_h)$ and $\mathbf{V}_h = (v_h, \mathbf{q}_h)$. In the following lemma we establish an abstract error expression for the solution $(u, \mathbf{p})$ of problem (27); since the proof is rather long we give it in the appendix.

Lemma 4.1. *Let $(u, \mathbf{p})$ be the solution of problem* (16) *and $(u_h, \mathbf{p}_h)$ the solution of problem* (27)*; then the following estimate holds: For small $\delta > 0$, and for any $(\mathbf{V}_h, \mu_h) \in L^2(\Omega) \times \mathbf{X}_h \times M_h$,*

(38)
$$\|u-u_h\|_V^2+\|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2}^2 \le c\big(|A(\mathbf{U}-\mathbf{U}_h, \mathbf{U}-\mathbf{V}_h)|+|d(\mu_h-\lambda, \mathbf{p}-\mathbf{p}_h)+d(\lambda-\lambda_h, \mathbf{p}-\mathbf{q}_h) + d(\lambda-\mu_h, \mathbf{p}) + d(\lambda_h, \mathbf{p}) + d(\mu_h, \mathbf{p}_h)| + \inf_{v_h \in V_h} \|u-v_h\|_V^2 + \|\lambda-\lambda_h\|^2\big).$$

We also need some related results on the approximation by means of the Lagrange interpolation. We refer the reader to [12] for the proofs of the results. Denote by I_h^ℓ and i_h the Lagrange interpolation operators with values in the spaces $V_h(\Omega^\ell)$ and $W_h^1(\Gamma_c)$, respectively. Then, there exists a constant $C > 0$ such that for all $v^\ell \in H^2(\Omega^\ell)$ and $v \in H^{\frac{3}{2}}(\Gamma_c)$,

$$\|v^\ell - I_h^\ell v^\ell\|_{L^2(\Omega^\ell)} \le ch^2 \|v^\ell\|_{H^2(\Omega^\ell)}; \ \|v - i_h v\|_{L^2(\Gamma_c)} \le ch^{\frac{3}{2}} \|v\|_{H^{\frac{3}{2}}(\Gamma_c)}. \tag{39}$$

Let Π_h^ℓ be the projection operator from $\mathbf{X}(\Omega^\ell)$ into $(V_h(\Omega^\ell))^2$ introduced in [8]; then we have for $\mathbf{q} \in \mathbf{X}$ such that $\mathbf{q}^\ell = \mathbf{q}_{|\Omega^\ell} \in H^{s_\ell}(\Omega^\ell)$, $1 \le s_\ell \le 2$, $\ell = 1, 2$ [8, Theorem 3.4],

$$\|\mathbf{q}^\ell - \Pi_h^\ell \mathbf{q}^\ell\|_{L^2(\Omega^\ell)^2} \le ch^{s_\ell} \|\mathbf{q}^\ell\|_{(H^{s_\ell}(\Omega^\ell))^2}, \tag{40}$$

$$\|\operatorname{div} \mathbf{q}^\ell - \operatorname{div} \Pi_h^\ell \mathbf{q}^\ell\|_{L^2(\Omega^\ell)} \le ch^{s_\ell - 1} \|\mathbf{q}^\ell\|_{(H^{s_\ell - 1}(\Omega^\ell))^2}. \tag{41}$$

In order to simplify the error analysis we will assume that the solution $\mathbf{p}$ of problem (23) is such that

$$\operatorname{div} \mathbf{p}_\ell \in H^1(\Omega^\ell), \quad \ell = 1, 2. \tag{42}$$

It is readily checked that under assumption (42) the trace $\mathbf{p}.\nu$ belongs to $H^{\frac{3}{2}}(\Gamma_c)$. Note that assumption (42) is not stringent since it requires only local regularity in each subdomain (which holds in general—see the end of section 1). If we assume only $\mathbf{p}^\ell \in H^1(\Omega^\ell)$, we derive, similarly to the rest of this section, the complete error estimate (in this case the error behaves as $O(h^{\frac{1}{4}})$).

The following lemma states the first error estimates; for the convenience of the reader the proof is also given in the appendix.

Lemma 4.2. *Let $(\mathbf{U}, \lambda)$, with $\mathbf{U} = (u, \mathbf{p})$, be the solution of problem* (23). *Suppose that $u_{|\Omega^1} \in H^2(\Omega^1)$, $u_{|\Omega^2} \in H^2(\Omega^2)$, that also $\mathbf{p}_{|\Omega^1} \in H^1(\Omega^1)^2$, $\mathbf{p}_{|\Omega^2} \in H^1(\Omega^2)^2$, and that assumption* (42) *holds.*

(i) *Let $(\mathbf{U}_h, \lambda_h)$ be the solution of* (27) *with $M_h = M_h^0$; then the following estimate holds:*

(43)
$$\|u-u_h\|_V^2 + \|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2}^2 \le C(\delta, u, \mathbf{p})\big(h\|\lambda-\lambda_h\|_{L^2(\Gamma_c)} + \|\lambda-\lambda_h\|_{L^2(\Gamma_c)}^2 + h^{\frac{3}{2}}\big).$$

(ii) *In the case of $M_h = M_h^{1,*}$ the following estimate holds:*

(44)
$$\|u-u_h\|_V^2+\|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2}^2 \le C(\delta, u, \mathbf{p})\big(h^{\frac{3}{2}}\|\lambda-\lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)}+\|\lambda-\lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)}^2+h^{\frac{3}{2}}\big),$$

where the (generic) constant $C(\delta, u, \mathbf{p})$ *depends linearly on the norms* $\|u_{|\Omega^\ell}\|_{H^2(\Omega^\ell)}$, $\|\mathbf{p}_{|\Omega^\ell}\|_{H^1(\Omega^\ell)^2}$, *and* $\|\operatorname{div} \mathbf{p}_{|\Omega^\ell}\|_{H^1(\Omega^\ell)^2}$, $\ell = 1, 2$.

Proof. First, the result is established for $M_h = M_h^{1,*}$, and we denote by π_h the associated projection operator π_h^1 given in (32). The proof is divided into small steps.

Step 1. Recall the definition of $A(.,.)$:

$$\begin{aligned} A(\mathbf{U} - \mathbf{U}_h, \mathbf{U} - \mathbf{V}_h) = a_\delta(u - u_h, u - v_h) + b(u - v_h, \mathbf{p} - \mathbf{p}_h) \\ -b(u - u_h, \mathbf{p} - \mathbf{q}_h) + g(\mathbf{p} - \mathbf{p}_h, \mathbf{p} - \mathbf{q}_h). \end{aligned}$$

Then, for particular choices of the test functions as $v_h^\ell = I_h^\ell u^\ell$ and $\mathbf{q}_h^\ell = \Pi_h \mathbf{p}^\ell$, $\ell = 1, 2$, and using inequalities (39) and (40), we derive

$$|a_\delta(u - u_h, u - v_h)| \leq \delta c(u) h^2 \|u - u_h\|_V,$$

$$|b(u - v_h, \mathbf{p} - \mathbf{p}_h)| \leq c(u) h^2 \|\operatorname{div} \mathbf{p} - \operatorname{div} \mathbf{p}_h\|_{L^2(\Omega)},$$

$$|b(u - u_h, \mathbf{p} - \mathbf{q}_h)| \leq c(\mathbf{p}) h \|u - u_h\|_V,$$

$$|g(\mathbf{p} - \mathbf{p}_h, \mathbf{p} - \mathbf{q}_h)| \leq c(\mathbf{p}) h \|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)^2},$$

where the (generic) constants are linear with respect to the appropriate norms, i.e., with

$$c(u) \leq c \max\{\|u_{|\Omega^1}\|_{H^2(\Omega^1)}, \|u_{|\Omega^2}\|_{H^2(\Omega^2)}\},$$

and

$$c(\mathbf{p}) \leq c \max\{\|\mathbf{p}_{|\Omega^1}\|_{H^1(\Omega^1)^2}, \|\mathbf{p}_{|\Omega^2}\|_{H^1(\Omega^2)^2}, \|\operatorname{div} \mathbf{p}_{|\Omega^1}\|_{H^1(\Omega^1)}, \|\operatorname{div} \mathbf{p}_{|\Omega^2}\|_{H^1(\Omega^2)}\}.$$

Note that we have (see [8, Theorem 3.4 and condition (3.10)])

$$\begin{aligned} (45) \quad \|\operatorname{div}\mathbf{p} - \operatorname{div}\mathbf{p}_h\|_{L^2(\Omega)} &= \|\operatorname{div}\mathbf{p} - \operatorname{div}\mathbf{q}_h\|_{L^2(\Omega)} \\ &= \left(\sum_{T \in \mathcal{T}_h} \|\operatorname{div}\mathbf{p} - \operatorname{div}\mathbf{q}_h\|_{L^2(T)}^2 \right)^{\frac{1}{2}} \leq c \left(\sum_{T \in \mathcal{T}_h} h_T^2 \|\operatorname{div}\mathbf{p}\|_{H^1(T)}^2 \right)^{\frac{1}{2}} \leq c(\mathbf{p}) h. \end{aligned}$$

Thus we can write

$$|b(u - v_h, \mathbf{p} - \mathbf{p}_h)| \leq c(u, \mathbf{p}) h^3.$$

The above estimates yield

$$(46) \qquad A(\mathbf{U} - \mathbf{U}_h, \mathbf{U} - \mathbf{V}_h) \leq C(u, \mathbf{p}, \delta) \left\{ h(\|u - u_h\|_V + \|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)^2}) + h^3 \right\}.$$

Step 2. Taking $\mu_h = i_h(\lambda)$, by an application of the Cauchy–Schwarz inequality, combined with estimates (39), it follows that

$$\begin{aligned} d(\lambda - \mu_h, \mathbf{U}_h - \mathbf{U}) &\leq \|\lambda - \mu_h\|_{L^2(\Gamma_c)} \|(\mathbf{p}_h - \mathbf{p}) \cdot \nu\|_{L^2(\Gamma_c)} \\ &\leq c h^{\frac{3}{2}} \|\lambda\|_{H^{\frac{3}{2}}(\Gamma_c)} \|(\mathbf{p}_h - \mathbf{p}) . \nu\|_{L^2(\Gamma_c)}. \end{aligned}$$

By a scaling we derive, for each edge e of a triangle $T \in \mathcal{T}_h \cap \Gamma_c$, the following trace theorem:

$$\|(\mathbf{p}-\mathbf{p}_h)\cdot\nu\|_{L^2(e)} \le c\big(h_T^{-\frac{1}{2}}\|\mathbf{p}-\mathbf{p}_h\|_{L^2(T)^2} + h_T^{\frac{1}{2}}\|(\operatorname{div}\mathbf{p} - \operatorname{div}\mathbf{p}_h)\|_{L^2(T)}\big). \tag{47}$$

Applying this inequality and (45) yields

$$d(\lambda-\mu_h, \mathbf{U}_h-\mathbf{U}) \le C(u,\mathbf{p})\big(h\|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2} + h^2\big). \tag{48}$$

The term $d(\mu_h, \mathbf{p}_h)$ is bounded similarily and yields

$$d(\mu_h, \mathbf{U}_h-\mathbf{U}) \le C(u,\mathbf{p})\big(h\|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2} + h^2\big). \tag{49}$$

Step 3. For $\mathbf{q}_h.\nu = \pi_h(\mathbf{q}.\nu)$, an application of the Cauchy–Schwarz inequality combined with (32) leads to

$$\begin{aligned} d(\lambda-\lambda_h, \mathbf{U}-\mathbf{V}_h) &\le \|\lambda-\lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)}\|(\mathbf{p}-\mathbf{q}_h).\nu\|_{H^{-\frac{1}{2}}(\Gamma_c)} \\ &\le C(\mathbf{p})h^2\|\lambda-\lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)}. \end{aligned} \tag{50}$$

Step 4. We have

$$\begin{aligned} d(\lambda-\mu_h, \mathbf{U}) &= \int_{\Gamma_c}(\lambda-\mu_h)(\mathbf{p}.\nu)\,d\Gamma \\ &= \int_{\Gamma_c}(\lambda - i_h(\lambda))(\mathbf{p}.\nu - \pi_h(\mathbf{p}.\nu))\,d\Gamma \quad (\pi_h(i_h(\varphi)) = i_h(\varphi)) \\ &\le \|\lambda - i_h(\lambda)\|_{L^2(\Gamma_c)}\|\mathbf{p}.\nu - \pi_h(\mathbf{p}.\nu)\|_{L^2(\Gamma_c)}. \end{aligned}$$

Using (39) and the appropriate approximation property of the projection $\pi_h = \pi_h^1$ [7], we obtain

$$d(\lambda-\mu_h, \mathbf{U}) \le C(u,\mathbf{p})h^3. \tag{51}$$

Step 5. We establish an estimate for the last term in (38) in order to apply estimate (38). The term is rewritten as follows:

$$\begin{aligned} d(\lambda_h, \mathbf{U}) &= \int_{\Gamma_c}(\lambda_h)(\mathbf{p}.\nu)\,d\Gamma \\ &= \int_{\Gamma_c}(\lambda_h)(\mathbf{p}.\nu - i_h(\mathbf{p}.\nu))\,d\Gamma + \int_{\Gamma_c}\lambda_h i_h(\mathbf{p}.\nu)\,d\Gamma. \end{aligned}$$

Since $i_h(\mathbf{p}.\nu) \le 0$, and $\lambda_h \in M_h^{1,*}$, we have

$$\int_{\Gamma_c}\lambda_h i_h(\mathbf{p}.\nu)\,d\Gamma \le 0.$$

Therefore

$$\begin{aligned} d(\lambda_h, \mathbf{U}) &\le \int_{\Gamma_c}(\lambda_h)(\mathbf{p}.\nu - i_h(\mathbf{p}.\nu))\,d\Gamma \\ &\le \int_{\Gamma_c}(\lambda_h-\lambda)(\mathbf{p}.\nu - i_h(\mathbf{p}.\nu))\,d\Gamma \\ &\quad + \int_{\Gamma_c}(\lambda)(\mathbf{p}.\nu - i_h(\mathbf{p}.\nu))\,d\Gamma \\ &\le \|\lambda-\lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)}\|(\mathbf{p}.\nu - i_h(\mathbf{p}.\nu))\|_{L^2(\Gamma_c)} \\ &\quad + \|\lambda\|_{L^2(\Gamma_c)}\|(\mathbf{p}.\nu - i_h(\mathbf{p}.\nu))\|_{L^2(\Gamma_c)} \end{aligned}$$

and by (39)

$$d(\lambda_h, \mathbf{U}) \leq C(u, \mathbf{p}) h^{\frac{3}{2}} \|\lambda - \lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)} + C(u, \mathbf{p}) h^{\frac{3}{2}}. \tag{52}$$

Finally, assembling estimates (46)–(52) in the right-hand side of (38) yields

$$\begin{aligned}\|u - u_h\|_V^2 + \|\mathbf{p} - \mathbf{p}_h\|^2_{L^2(\Omega)^2} \leq C(\delta, u, \mathbf{p})\big(h(\|u - u_h\|_V + \|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)^2}) \\ + h^{\frac{3}{2}} \|\lambda - \lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)} + \|\lambda - \lambda_h\|^2_{H^{\frac{1}{2}}(\Gamma_c)} + h^{\frac{3}{2}}\big).\end{aligned}$$

Writing

$$\begin{aligned}&C(\delta, u, \mathbf{p}) h(\|u - u_h\|_V + \|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)^2}) \\ &\qquad \leq \ \gamma(\|u - u_h\|_V + \|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)^2})^2 + \frac{C(\delta, u, \mathbf{p})^2}{4\gamma} h^2\end{aligned}$$

with $\gamma > 0$ leads to the desired estimate for sufficiently small γ. This completes the proof for $M_h = M_h^{1,*}$.

The case of $M_h = M_h^0$ can be treated in the same way, with the projection π_h in Steps 3 and 4 replaced by the appropriate L^2-projection π_h^0 on M_h (see [29] for the properties of the projection operator), and in view of $d(\lambda_h, \mathbf{U}) \leq 0$, Step 5 can be neglected. We briefly list the results that can be obtained for the particular case of $M_h = M_h^0$; the details are left to the reader. Actually, in Step 3 the following estimate is established:

$$d(\lambda - \lambda_h, \mathbf{U} - \mathbf{V}_h) \leq \|\lambda - \lambda_h\|_{L^2(\Gamma_c)} \|(\mathbf{p} - \mathbf{q}_h).\nu\|_{L^2(\Gamma_c)} \leq C(\mathbf{p}) h \|\lambda - \lambda_h\|_{L^2(\Gamma_c)},$$

and accordingly, in Step 4 the resulting inequality takes the form

$$d(\lambda - \mu_h, \mathbf{U}) \leq C(u, \mathbf{p}) h^{\frac{5}{2}}. \qquad \square$$

Remark 4.3. The same rate of convergence as in Lemma 4.2 cannot be derived for the case of $M_h = M_h^1$. However, the same procedure as in [5] results in the convergence rate $O(h^{\frac{1}{2}})$.

Remark 4.4. Estimates (43) and (44) can be formulated in terms of the mesh-dependent norm $\|.\|_h$ on $\mathbf{X}$ defined by

$$\|\mathbf{q}\|_h = \left(\sum_{T \in \mathcal{T}_h} \|\mathbf{q}\|^2_{L^2(T)^2} + h_T^2 \|\operatorname{div} \mathbf{q}\|^2_{L^2(T)} \right)^{\frac{1}{2}}.$$

Lemma 4.5. *Let $(u, \mathbf{p}, \lambda)$ be the solution of problem* (23). *As usual we denote $\mathbf{U} = (u, \mathbf{p})$. Suppose that $u_{|\Omega^1} \in H^2(\Omega^1)$, $u_{|\Omega^2} \in H^2(\Omega^2)$ and also $\mathbf{p}_{|\Omega^1} \in H^1(\Omega^1)^2$, $\mathbf{p}_{|\Omega^2} \in H^1(\Omega^2)^2$, and assumption* (42) *holds. Let $(\mathbf{U}_h, \lambda_h)$ be the solution of* (27) *for $M_h = M_h^0$; then the following estimate holds:*

$$\|\lambda - \lambda_h\|_{L^2(\Gamma_c)} \leq C h^{-\frac{1}{2}} (\|u - u_h\|_V + \|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)^2}) + C(u) h^{\frac{1}{2}}. \tag{53}$$

If $M_h = M_h^{1,*}$, then the following estimate is obtained:

$$\|\lambda - \lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)} \leq C(\|u - u_h\|_V + \|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)^2} + C(u, \mathbf{p}) h), \tag{54}$$

where $C(u,\mathbf{p}) \leq c \max\{\|u_{|\Omega^\ell}\|_{H^2(\Omega^\ell)}, \|\mathbf{p}_{|\Omega^\ell}\|_{H^1(\Omega^\ell)^2}, \|\mathrm{div}\,\mathbf{p}_{|\Omega^\ell}\|_{H^1(\Omega^\ell)^2},\ \ell = 1,2\}$.

Proof. We set $\mathbf{q} = \mathbf{q}_h \in \mathbf{X}_h \subset \mathbf{X}$ in the second equation of (23) and subtract the resulting equation from the second equation in (27), which leads to

$$-b(u-u_h,\mathbf{q}_h) + g(\mathbf{p}-\mathbf{p}_h,\mathbf{q}_h) + d(\lambda - \lambda_h(0,\mathbf{q}_h)) = 0$$

and

$$\begin{aligned} d(\lambda_h - \pi_h(\lambda),(0,\mathbf{q}_h)) &= -b(u-u_h,\mathbf{q}_h) + g(\mathbf{p}-\mathbf{p}_h,\mathbf{q}_h) + d(\lambda - \pi_h(\lambda),(0,\mathbf{q}_h)) \\ &\leq C(\|u-u_h\|_V + \|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2} + C\|\lambda - \pi_h(\lambda)\|_{H^s(\Gamma_c)}\|\mathbf{q}_h\|_{\mathbf{X}}), \end{aligned}$$

where $s = 0$ or $\frac{1}{2}$. We consider separately two particular cases of projection operators and derive the resulting error estimates. Namely, $s = 0$ and $\pi_h := \pi_h^0$ in the first case of $M_h = M_h^0$; otherwise, $s = \frac{1}{2}$ and $\pi_h := \pi_h^1$ if $M_h = M_h^{1,*}$ in the second case.

Case of projection π_h^0 *in* M_h^0. In view of (30) we have

$$\|\lambda - \pi_h(\lambda)\|_{L^2(\Gamma_c)} \leq Ch\|\lambda\|_{H^{\frac{3}{2}}(\Gamma_c)}. \tag{55}$$

On the other hand, using (29) we derive the following estimate:

$$\begin{aligned} \|\lambda_h - \pi_h(\lambda)\|_{L^2(\Gamma_c)} &\leq \beta_0^{-1} h^{-\frac{1}{2}} \sup_{\mathbf{V}_h \in V_h \times \mathbf{X}_h} \frac{d(\lambda_h - \pi_h(\lambda_h),\mathbf{V}_h)}{\|\mathbf{V}_h\|_{V_h\times\mathbf{X}_h}} \\ &\leq Ch^{-\frac{1}{2}}(\|u-u_h\|_V + \|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2} + Ch\|\lambda\|_{H^{\frac{3}{2}}(\Gamma_c)}). \end{aligned} \tag{56}$$

Case of projection π_h^1 *in* $M_h^{1,*}$. Using (32), we obtain

$$\|\lambda - \pi_h(\lambda)\|_{H^{\frac{1}{2}}(\Gamma_c)} \leq Ch\|\lambda\|_{H^{\frac{3}{2}}(\Gamma_c)}, \tag{57}$$

which together with the inf-sup condition (33) leads to

$$\begin{aligned} \beta\|\lambda_h - \pi_h(\lambda)\|_{H^{\frac{1}{2}}(\Gamma_c)} &\leq \sup_{\mathbf{V}_h \in V_h \times \mathbf{X}_h} \frac{d(\lambda_h - \pi_h(\lambda_h),\mathbf{V}_h)}{\|\mathbf{V}_h\|_{V_h\times\mathbf{X}_h}} \\ &\leq C(\|u-u_h\|_V + \|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2} + Ch\|\lambda\|_{H^{\frac{3}{2}}(\Gamma_c)}). \end{aligned} \tag{58}$$

By the triangular inequality

$$\|\lambda - \lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)} \leq \|\lambda - \pi_h(\lambda)\|_{H^{\frac{1}{2}}(\Gamma_c)} + \|\lambda_h - \pi_h(\lambda)\|_{H^{\frac{1}{2}}(\Gamma_c)}$$

and the estimates (56) and (55) (respectively, (58) and (57)), we obtain the inequality (53) (respectively, (54)). □

Assembling all the estimates given in Lemmas 4.2 and 4.5, we obtain the main result on error estimates.

THEOREM 4.6. *Let* $(u,\mathbf{p},\lambda)$ *be the solution of problem* (23). *Suppose that* $u_{|\Omega^1} \in H^2(\Omega^1)$, $u_{|\Omega^2} \in H^2(\Omega^2)$ *and also* $\mathbf{p}_{|\Omega^1} \in H^1(\Omega^1)^2$, $\mathbf{p}_{|\Omega^2} \in H^1(\Omega^2)^2$, *and assumption* (42) *holds. Let* $(\mathbf{U}_h,\lambda_h)$ *be the solution of* (27) *with* $M_h = M_h^0$. *Then the following estimate holds:*

$$\|u-u_h\|_V + \|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2} + \|\lambda - \lambda_h\|_{L^2(\Gamma_c)} \leq C(\delta,u,\mathbf{p})h^{\frac{1}{2}}. \tag{59}$$

When $M_h = M_h^{1,*}$ *the following estimate holds:*

$$\|u-u_h\|_V + \|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2} + \|\lambda - \lambda_h\|_{H^{\frac{1}{2}}(\Gamma_c)} \leq C(\delta,u,\mathbf{p})h^{\frac{3}{4}}, \tag{60}$$

where $C(\delta, u, \mathbf{p})$ *depends linearly on* $\|u_{|\Omega^\ell}\|_{H^2(\Omega^\ell)}$, $\|\mathbf{p}_{|\Omega^\ell}\|_{H^1(\Omega^\ell)^2}$, *and* $\|\operatorname{div}\mathbf{p}_{|\Omega^\ell}\|_{H^1(\Omega^\ell)^2}$, $\ell = 1, 2$.

Remark 4.7. The global error estimate (59) is only $O(h^{\frac{1}{2}})$ because the approximation of the Lagrange multiplier λ by functions of M_h^0 cannot provide better results even if λ is more regular, as can be seen in (55).

5. Numerical experiments.

In order to perform the computations, the matrix formulation of problem (27) is derived. It is readily checked that $(u_h, \mathbf{p}_h, \lambda_h) \in V_h \times \mathbf{X}_h \times M_h$ is a solution of (27) if and only if $(u_h, \mathbf{p}_h, \lambda_h)$ is a saddle-point of the Lagrangian defined on $V_h \times \mathbf{X}_h \times M_h$ by

$$\mathcal{L}(v_h, \mathbf{q}_h, \mu_h) = \mathcal{J}(v_h, \mathbf{q}_h) + \int_{\Gamma_c} \mu_h.(\mathbf{q}_h.\nu)\, d\sigma, \tag{61}$$

which means that $(u_h, \mathbf{p}_h, \lambda_h)$ satisfies

$$\mathcal{L}(u_h, \mathbf{p}_h, \mu_h) \le \mathcal{L}(u_h, \mathbf{p}_h, \lambda_h) \le \mathcal{L}(v_h, \mathbf{q}_h, \lambda_h) \quad \forall (v_h, \mathbf{q}_h) \in V_h \times \mathbf{X}_h,\ \forall \mu_h \in M_h.$$

Let $\mathbf{V}$, $\mathbf{U}$ denote the vectors with the entries given by the nodal values of the functions $(v_h, \mathbf{q}_h)$ and $(u_h, \mathbf{p}_h)$, respectively. Let M and Λ be the vectors with the entries given by the nodal values of μ_h and λ_h, respectively, for the three different choices of the space M_h, namely $M_h = M_h^1$, $M_h = M_h^{1,*}$, or $M_h = M_h^0$. Therefore, the saddle-point problem for Lagrangian (61) can be rewritten in finite dimensional setting:
Find $\mathbf{U} = (u_h, \mathbf{p}_h)$ and Λ, defined by the following max-min condition:

$$\max_{SM \ge 0} \left(\min_{\mathbf{V}} \frac{1}{2} {}^t\mathbf{V}\mathbf{K}\mathbf{V} - {}^t\mathbf{V}\mathbf{F} + ({}^t\mathbf{V}\,\mathbf{L})SM \right), \tag{62}$$

where $\mathbf{K}$ denotes the stiffness matrix, $\mathbf{F}$ is the vector corresponding to the external loading, and the matrix S expresses the sign conditions for multipliers (24)–(26).

Given a triangularization $\mathcal{T}_h$ of Ω, let N denote the number of nodes in Ω and N_T the number of elements in $\mathcal{T}_h$. Denote by $(w_i)_{i=1}^N$ the Lagrange finite element basis of V_h and let (Φ_i) stand for the basis in the space $\mathbf{X}_h$. Each vector function Φ_i either is of the form $(w_i, 0)$, $1 \le i \le N$, $(b_i, 0)$, $1 \le i \le N_T$, or is given by $(0, w_i)$, $1 \le i \le N$, $(0, b_i)$, $1 \le i \le N_T$, respectively, where b_i denotes a bubble function. Then the matrix K is defined by

$$\mathbf{K} = \begin{pmatrix} A_\delta & {}^tB_1 & {}^tB_2 \\ -B_1 & G_1 & 0 \\ -B_2 & 0 & G_2 \end{pmatrix},$$

and the right-hand side takes the form

$$\mathbf{F} = \begin{pmatrix} D\,F \\ 0 \\ 0 \end{pmatrix},$$

with $F = (f_i)_i$, $D = (\int_\Omega w_i\, w_j\, dx)_{ij}$, $i, j = 1, \dots, N$, and $A_\delta = \delta\, D$. The matrices B_1 and B_2 are defined by $(B_1)_{ij} = (\int_\Omega w_j \partial_1 \Phi_i^{(1)}\, dx)_{ij}$ and $(B_2)_{ij} = (\int_\Omega w_j \partial_2 \Phi_i^{(2)}\, dx)_{ij}$, $j = 1, \dots, N$, $i = 1, \dots, N + N_T$. Finally,

$$G_1 = G_2 = \begin{pmatrix} D & (\int_\Omega w_i\, b_j\, dx)_{ij} \\ (\int_\Omega b_i\, w_j\, dx)_{ij} & (\int_\Omega b_i^2\, dx)\delta_{ij} \end{pmatrix},$$

where δ_{ij} is the Kronecker symbol.

Let N_c denote the number of nodes on Γ_c and let us denote by $(\psi_i)_i$, $1 \le i \le N_c$ the basis in the space $W_h^1(\Gamma_c)$ and by $(\varphi_i)_i$ the basis of $W_h^0(\Gamma_c)$, $1 \le i \le N_c - 1$. We have a specific form of S for each particular choice of M_h, namely

- if M_h is M_h^0 or M_h^1 then S is given by the identity matrix,
- if $M_h = M_h^{1,*}$, then S is given by $S_{ij} = \int_{\Gamma_c} \psi_i \, \psi_j \, d\Gamma$, $1 \le i, j \le N_c$.

Finally, the *coupling* matrix

$$L = \begin{pmatrix} 0 \\ L_1 \\ L_2 \end{pmatrix}$$

is defined in the following way:

- If $M_h = M_h^1$ or $M_h = M_h^{1,*}$, then

$$(L^1)_{ij} = \begin{cases} \int_{\Gamma_c} \psi_j \, ((w_i, 0).\nu) \, d\Gamma, & 1 \le i \le N, \ 1 \le j \le N_c, \\ 0, & 1 \le i \le N_T, \ 1 \le j \le N_c, \end{cases}$$

and

$$(L^2)_{ij} = \begin{cases} \int_{\Gamma_c} \psi_j \, ((0, w_i).\nu) \, d\Gamma, & 1 \le i \le N, \ 1 \le j \le N_c, \\ 0, & 1 \le i \le N_T, \ 1 \le j \le N_c. \end{cases}$$

- If $M_h = M_h^0$, then

$$(L^1)_{ij} = \begin{cases} \int_{\Gamma_c} \varphi_j \, ((w_i, 0).\nu) \, d\Gamma, & 1 \le i \le N, \ 1 \le j \le N_c - 1, \\ 0, & 1 \le i \le N_T, \ 1 \le j \le N_c, \end{cases}$$

and

$$(L^2)_{ij} = \begin{cases} \int_{\Gamma_c} \varphi_j \, ((0, w_i).\nu) \, d\Gamma, & 1 \le i \le N, \ 1 \le j \le N_c - 1, \\ 0, & 1 \le i \le N_T, \ 1 \le j \le N_c. \end{cases}$$

The solution $(\mathbf{U}, \Lambda)$ of (62) satisfies the saddle-point conditions and we have

$$\mathbf{U} = \mathbf{K}^{-1}(\mathbf{F} - L \, S\Lambda). \tag{63}$$

Therefore, for $\Phi = S\Lambda$, the saddle-point problem (62) can be rewritten as a quadratic programming problem

$$\min_{\Phi \ge 0} \left(\frac{1}{2} {}^t\Phi {}^tL\mathbf{K}^{-1}L\Phi - {}^t\Phi {}^tL\mathbf{K}^{-1}\mathbf{F} + \frac{1}{2} {}^t\mathbf{F}\mathbf{K}^{-1}\mathbf{F} \right). \tag{64}$$

If $\overline{\Phi}$ is the solution of (64) then $\Lambda = S^{-1}\overline{\Phi}$. The solution $\mathbf{U}$ is obtained by solving (63).

We note that eliminating the bubbles yields problems (64) and (63) with reduced matrices to be solved and, therefore, with reduced number of unknowns.

In all the numerical experiments, the domain $\Omega =]0, 4[\times]0, 1[$ is fixed, and data are selected as follows: $f = 1$, $\delta = 0.1$.

In the first example the numerical comparison of the three methods corresponding to $M_h = M_h^0$, $M_h = M_h^1$, and $M_h = M_h^{1,*}$ is performed. This comparison is limited to the case of only one global triangulation in the whole domain Ω, which is in fact our goal. It can be expected, as for classical unilateral contact problems, that different

Table 1
$\|u - u_h\|_{L^2(\Omega)}$.

dof	160	610	934	1233
$M_h = M_h^{1,*}$	2.32×10^{-2}	0.621×10^{-3}	4.61×10^{-3}	3.31×10^{-3}
$M_h = M_h^1$	2.30×10^{-2}	6.17×10^{-3}	4.58×10^{-3}	3.28×10^{-3}
$M_h = M_h^0$	2.30×10^{-2}	6.18×10^{-3}	4.58×10^{-3}	3.29×10^{-3}

Table 2
$\|p_1 - p_{1h}\|_{L^2(\Omega)}$.

dof	160	610	934	1233
$M_h = M_h^{1,*}$	1.20×10^{-1}	5.83×10^{-2}	3.987×10^{-2}	2.902×10^{-2}
$M_h = M_h^1$	1.18×10^{-1}	5.79×10^{-2}	3.955×10^{-2}	2.75×10^{-2}
$M_h = M_h^0$	1.18×10^{-1}	5.79×10^{-2}	3.955×10^{-2}	2.754×10^{-2}

triangulations of Ω^1 and Ω^2 with nonmatching grids at the crack Γ_c lead to the same numerical results. For $\Gamma_c =]0,4[\times \left\{\frac{1}{2}\right\}$ the relative errors $\|u - u_h\|_{L^2(\Omega)}$, $\|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)}$, and $\|\lambda - \lambda_h\|_{L^2(\Gamma_c)}$ are computed. Since the exact solution is not available, we use (as usual) a reference solution obtained for the sufficiently fine mesh, namely for the triangulation which is made of 3308 elements, with 1725 degrees of freedom and 114 vertices on Γ_c. In Tables 1–4, the relative errors are reported as a function of the degrees of freedom (dof). The respective convergence curves, which represent the logarithms of the relative errors in function of the logarithms of dof, are given by Figures 3–5. The three methods lead to similar approximation results, and this fact agrees with the theoretical estimates and it is also in accordance with previous numerical studies for classical unilateral contact problems (see, e.g., [14]). Since the solution is sufficiently smooth (in particular, u is continuous), the norm of λ is relatively small, which explains the fact that the relative error for λ is quite large (compared to the others errors). In addition, the approximation results for $\mathbf{p}$ could be significantly improved by employing finite elements other than the stabilized linear elements ($\mathcal{P}_B$), which are known to be less efficient for the approximation of functions in $\mathbf{X}$.

In the second example, we present the isolines of u, p_1, and p_2 with various cuts. The plots are obtained by the interpolation of the computed solution $(u_h, \mathbf{p}_h)$ on a new regular and coarse mesh. Figures 6–8 correspond to $\Gamma_c =]1,3[\times \left\{\frac{3}{4}\right\}$; as expected from the symmetry of the domain (with respect to Γ_c) and of the data, the solution $(u, \mathbf{p})$ is smooth. In Figures 9–11, the crack line $\Gamma_c =]0,1[\times \left\{\frac{1}{4}\right\}$ touches the boundary at one of its endpoints, and the domain is not symmetric.

6. Conclusion. We performed the convergence analysis and derived the error estimates. The variational formulation was constructed and implemented. The numerical results are in full agreement with the theoretical estimates.

The further applications of the proposed numerical methods concern the contact problems in elasticity with the Coulomb friction.

TABLE 3
$\|p_2 - p_{2h}\|_{L^2(\Omega)}$.

dof	160	610	934	1233
$M_h = M_h^{1,*}$	7.6×10^{-2}	3.7×10^{-2}	3.2×10^{-2}	2.9×10^{-2}
$M_h = M_h^1$	5.40×10^{-2}	2.50×10^{-2}	2.17×10^{-2}	1.51×10^{-2}
$M_h = M_h^0$	5.40×10^{-2}	2.50×10^{-2}	2.17×10^{-2}	1.50×10^{-2}

TABLE 4
$\|\lambda - \lambda_h\|_{L^2(\Gamma_c)}$.

dof	160	610	934	1233
$M_h = M_h^{1,*}$	5.0240	1.6019	1.3745	1.2922
$M_h = M_h^1$	17.2834	2.6886	1.6801	1.4406
$M_h = M_h^0$	13.9451	2.9765	1.7065	1.5474

Appendix. In this appendix, we give the proof of Lemma 4.1, which states the following abstract error:

$$\|u-u_h\|_V^2+\|\mathbf{p}-\mathbf{p}_h\|_{L^2(\Omega)^2}^2 \le c\big(|A(\mathbf{U}-\mathbf{U}_h,\mathbf{U}-\mathbf{V}_h)|+|d(\mu_h-\lambda,\mathbf{p}-\mathbf{p}_h)+d(\lambda-\lambda_h,\mathbf{p}-\mathbf{q}_h) \\ + d(\lambda-\mu_h,\mathbf{p})+d(\lambda_h,\mathbf{p})+d(\mu_h,\mathbf{p}_h)| + \inf_{v_h\in V_h}\|u-v_h\|_V^2 + \|\lambda-\lambda_h\|^2\big).$$

Proof. We have

$$\delta\|u-u_h\|_V^2 = a_\delta(u-u_h,u-u_h) = a_\delta(u-u_h,u-v_h)+a_\delta(u-u_h,v_h-u_h), \quad v_h\in V_h.$$

Noting that $V_h \subset V$ and subtracting the first lines of (23) and (27) (with $v = u_h - v_h$), we obtain

$$\delta\|u-u_h\|_V^2 = a_\delta(u-u_h,u-v_h)+b(u_h-v_h,\mathbf{p}-\mathbf{p}_h) \\ = a_\delta(u-u_h,u-v_h)+b(u-v_h,\mathbf{p}-\mathbf{p}_h)+b(u_h-u,\mathbf{p}-\mathbf{q}_h)+b(u_h-u,\mathbf{p}_h-\mathbf{q}_h) \quad \forall \mathbf{q}_h\in\mathbf{X}_h.$$

Since $\mathbf{X}_h \subset \mathbf{X}$, subtracting the second lines of (23) and (27) (with $\mathbf{q}$ and $\mathbf{q}_h$ replaced by $\mathbf{p}_h - \mathbf{q}_h$) yields

$$\begin{aligned}\delta\|u-u_h\|_V^2 &= a_\delta(u-u_h,u-v_h)+b(u_h-v_h,\mathbf{p}-\mathbf{p}_h)\\ &= a_\delta(u-u_h,u-v_h)+b(u-v_h,\mathbf{p}-\mathbf{p}_h)+b(u_h-u,\mathbf{p}-\mathbf{q}_h)\\ &\quad + g(\mathbf{q}_h-\mathbf{p}_h,\mathbf{p}-\mathbf{p}_h)+d(\lambda-\lambda_h,\mathbf{q}_h-\mathbf{p}_h), \quad \mathbf{q}_h\in\mathbf{X}_h\\ &= a_\delta(u-u_h,u-v_h)+b(u-v_h,\mathbf{p}-\mathbf{p}_h)+b(u_h-u,\mathbf{p}-\mathbf{q}_h)\\ &\quad + g(\mathbf{q}_h-\mathbf{p},\mathbf{p}-\mathbf{p}_h)+g(\mathbf{p}-\mathbf{p}_h,\mathbf{p}-\mathbf{p}_h)+d(\lambda-\lambda_h,\mathbf{p}_h-\mathbf{q}_h).\end{aligned}$$

Noting that $d(\lambda_h,\mathbf{p}_h)=0$, we deduce that for all $\mu_h\in M_h$

$$\begin{aligned}d(\lambda-\lambda_h,\mathbf{p}_h-\mathbf{q}_h) &= d(\lambda,\mathbf{p}_h-\mathbf{q}_h)+d(\lambda_h,\mathbf{q}_h)\\ &= d(\mu_h-\lambda,\mathbf{p}-\mathbf{p}_h)+d(\lambda-\lambda_h,\mathbf{p}-\mathbf{q}_h)+d(\lambda-\mu_h,\mathbf{p})\\ &\quad + d(\lambda_h,\mathbf{p})+d(\mu_h,\mathbf{p}_h).\end{aligned}$$

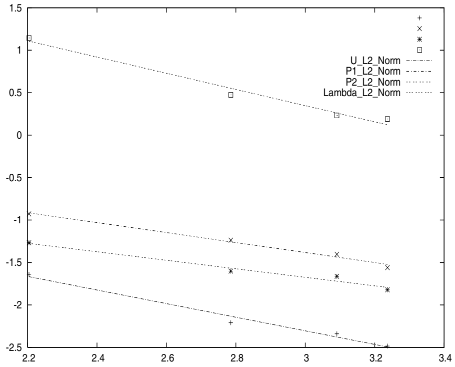


FIG. 3. *Convergence rate with $M_h = M_h^0$.*

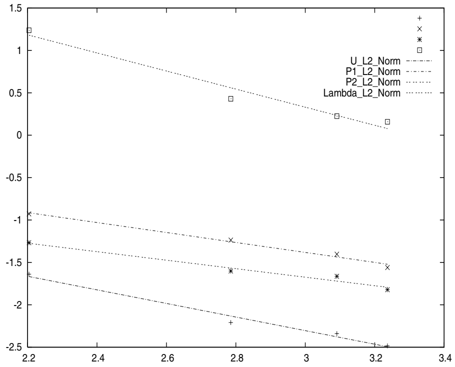


FIG. 4. *Convergence rate with $M_h = M_h^1$.*

Thus, we obtain (from the definition of $g(.,.)$)

$$\begin{aligned}\|\mathbf{p}-\mathbf{p}_h\|^2_{L^2(\Omega)^2} \le \delta\|u-u_h\|_V^2 &+ \big(|a_\delta(u-u_h,u-v_h)| + b(u-v_h,\mathbf{p}-\mathbf{p}_h)\\ &+ b(u_h-u,\mathbf{p}-\mathbf{q}_h) + g(\mathbf{q}_h-\mathbf{p},\mathbf{p}-\mathbf{p}_h)| + |d(\mu_h-\lambda,\mathbf{p}-\mathbf{p}_h)\\ &d(\lambda-\lambda_h,\mathbf{p}-\mathbf{q}_h) + d(\lambda-\mu_h,\mathbf{p}) + d(\lambda_h,\mathbf{p}) + d(\mu_h,\mathbf{p}_h)|\big),\end{aligned}$$

and this inequality is written as

$$\begin{aligned}\|\mathbf{p}-\mathbf{p}_h\|^2_{L^2(\Omega)^2} \le \delta\|u-u_h\|_V^2 &+ \big(|A(\mathbf{U}-\mathbf{U}_h,\mathbf{U}-\mathbf{V}_h)| + |d(\mu_h-\lambda,\mathbf{p}-\mathbf{p}_h)\\ &d(\lambda-\lambda_h,\mathbf{p}-\mathbf{q}_h) + d(\lambda-\mu_h,\mathbf{p}) + d(\lambda_h,\mathbf{p}) + d(\mu_h,\mathbf{p}_h)|\big).\end{aligned} \tag{65}$$

Next, we have

$$\begin{aligned}b(u_h-v_h,\mathbf{q}_h) &= b(u_h-u,\mathbf{q}_h) + b(u-v_h,\mathbf{q}_h)\\ &= g(\mathbf{p}-\mathbf{p}_h,\mathbf{q}_h) + d(\lambda-\lambda_h,\mathbf{q}_h) + b(u-v_h,\mathbf{q}_h)\\ &\le g(\mathbf{p}-\mathbf{p}_h,\mathbf{q}_h) + d(\lambda-\lambda_h,\mathbf{q}_h) + b(u-v_h,\mathbf{q}_h),\end{aligned}$$

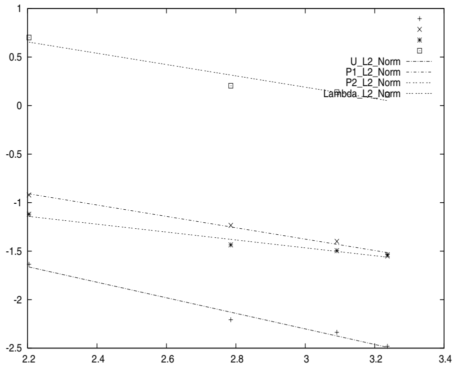


FIG. 5. *Convergence rate with* $M_h = M_h^{1,*}$.

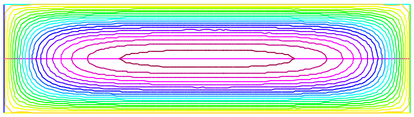


FIG. 6. *Isolines of* u.

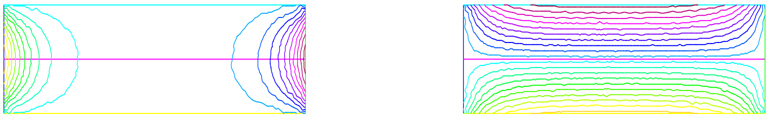


FIG. 7. *Isolines of* p_1 *(left) and* p_2 *(right).*

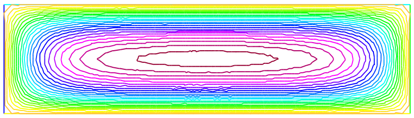


FIG. 8. *Isolines of* u.

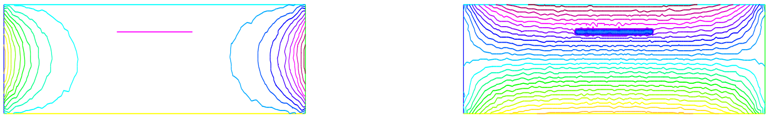


FIG. 9. *Isolines of* p_1 *(left) and* p_2 *(right).*

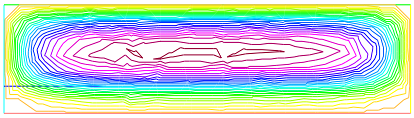


FIG. 10. *Isolines of* u.

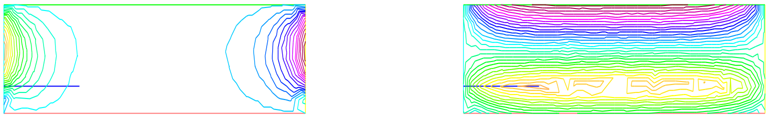


FIG. 11. *Isolines of* p_1 *(left) and* p_2 *(right).*

and using the inf-sup condition (19), we derive

$$\begin{aligned}\|u_h - v_h\|_V &\leq \gamma^{-1} \sup_{\mathbf{q}_h \in \mathbf{X}_h} \frac{|b(u - v_h, \mathbf{q}_h) + g(\mathbf{p} - \mathbf{p}_h, \mathbf{q}_h) + d(\lambda - \lambda_h, \mathbf{q}_h)|}{\|\mathbf{q}_h\|_{\mathbf{X}}} \\ &\leq \gamma^{-1}\big(\|b\| \|u - v_h\|_V + \|g\|\, \|\mathbf{p} - \mathbf{q}_h\|_{L^2(\Omega)^2} + \|d\|\, \|\lambda - \lambda_h\|\big),\end{aligned}$$

where $\|b\|$, $\|g\|$, and $\|d\|$ denote the norms of the bilinear forms, from which, combining with the triangle inequality, we have

(66)

$$\|u - u_h\|_V \leq \big((1 + \gamma^{-1}\|b\|) \inf_{v_h \in V_h} \|u - v_h\|_V + \gamma^{-1}\|g\|\, \|\mathbf{p} - \mathbf{p}_h\|_{L^2(\Omega)^2} + \gamma^{-1}\|d\|\, \|\lambda - \lambda_h\|\big).$$

Inserting in (65) yields

$$\begin{aligned}(1 - \delta\gamma^{-1}\|g\|)\, \|\mathbf{p} - \mathbf{p}_h\|^2_{L^2(\Omega)^2} &\leq \delta\,\big((1 + \gamma^{-1}\|b\|) \inf_{v_h \in V_h} \|u - v_h\|^2_V + \gamma^{-1}\|d\|\, \|\lambda - \lambda_h\|^2\big) \\ &\quad + \big(|A(\mathbf{U} - \mathbf{U}_h, \mathbf{U} - \mathbf{V}_h)| + |d(\mu_h - \lambda, \mathbf{p} - \mathbf{p}_h) d(\lambda - \lambda_h, \mathbf{p} - \mathbf{q}_h) \\ &\qquad + d(\lambda - \mu_h, \mathbf{p}) + d(\lambda_h, \mathbf{p}) + d(\mu_h, \mathbf{p}_h)|\big),\end{aligned}$$

from which

(67)

$$\begin{aligned}\|\mathbf{p} - \mathbf{p}_h\|^2_{L^2(\Omega)^2} &\leq C(\delta, \gamma, \|b\|, \|g\|, \|d\|)\left(\inf_{v_h \in V_h} \|u - v_h\|^2_V + \|\lambda - \lambda_h\|^2\right) \\ &\quad + \big(|A(\mathbf{U} - \mathbf{U}_h, \mathbf{U} - \mathbf{V}_h) + |d(\mu_h - \lambda, \mathbf{p} - \mathbf{p}_h) + d(\lambda - \lambda_h, \mathbf{p} - \mathbf{q}_h) \\ &\qquad + d(\lambda - \mu_h, \mathbf{p}) + d(\lambda_h, \mathbf{p}) + d(\mu_h, \mathbf{p}_h)|\big).\end{aligned}$$

The inequality (38) follows easily from (67) and (66). □

REFERENCES

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

[2] Z. Belhachmi and F. Ben Belgacem, *Quadratic finite element approximation of the Signorini problem*, Math. Comp., 72 (2003), pp. 83–104.

[3] F. Ben Belgacem, *Numerical simulation of some variational inequalities arisen from unilateral contact problems by the finite element methods*, SIAM J. Numer. Anal, 37 (2000), pp. 1198–1216.

[4] F. Ben Belgacem, P. Hild, and P. Laborde, *Extention of the mortar finite element method to a variational inequality modeling unilateral contact*, Math. Models Methods Appl. Sci., 9 (1999), pp. 287–303.

[5] F. Ben Belgacem and Y. Renard, *Hybrid finite element methods for the Signorini problem*, Math. Comp., 72 (2003), pp. 1117–1145.

[6] C. Bernardi and V. Girault, *A local regularization operator for triangular and quadrilateral finite elements*, SIAM. J. Numer. Anal., 35 (1998), pp. 1893–1916.

[7] C. Bernardi, Y. Maday, and A. T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Collège de France Seminar, Vol. XI, H. Brézis and J. L. Lions, eds., Longman, Harlow, UK, 1994, pp. 13–51.

[8] F. Brezzi, J. Douglas Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[9] F. Brezzi, W. W. Hager, and P. A. Raviart, *Error estimates for the finite element solution of variational inequalities*, Numer. Math., 28 (1977), pp. 431–443.

[10] F. Brezzi, W. W. Hager, and P. A. Raviart, *Error estimates for the finite element solution of variational inequalities, Part* 2: *Mixed methods*, Numer. Math., 31 (1978), pp. 1–16.

[11] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[12] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in the Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J.-L. Lions eds., North-Holland, 1991, pp. 17–351.

[13] M. Crouzeix and V. Thomée, *The stability in L^p and $W^{1,p}$ of the L^2-projection on finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.

[14] P. Coorevits, P. Hild, K. Lhalouani, and T. Sassi, *Mixed finite element methods for unilateral problems: Convergence analysis and numerical studies*, Math. Comp., 71 (2001), pp. 1–25.

[15] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains: Smoothness and Asymptotics of Solutions*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.

[16] G. Duvaut and J.-L. Lions, *Les inéquations en mécanique et en physique*, Dunod, Paris, 1972.

[17] C. Eck, A. Nazarov, and W. L. Wendland, *Asymptotic analysis for a mixed boundary value contact problem*, Arch. Ration. Mech. Anal., 156 (2001), pp. 274–316.

[18] V. Girault and P.-A. Raviart, *Finite Element Methods for the Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.

[19] R. Glowinski, *Lectures on Numerical Methods for Nonlinear Variational Problems*, Springer-Verlag, Berlin, 1980.

[20] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monographs and Studies in Mathematics 24, Pitman, Boston, 1985.

[21] J. Haslinger and I. Hlaváček, *Contact between elastic bodies* II. *Finite element analysis*, Aplikace Matematiky, 26 (1981), pp. 263–290.

[22] J. Haslinger, I. Hlaváček, and J. Nečas, *Numerical methods for unilateral problems in solid mechanics*, in Handbook of Numerical Analysis, Vol. IV, Part 2, P. G. Ciarlet and J.-L. Lions eds., North-Holland, Amsterdam, 1996, pp. 313–485.

[23] P. Hild, *Problèmes de contact unilatéral et maillages incompatibles*, Thèse de l'Université Paul Sabatier, Toulouse 3, 1998.

[24] A. M. Khludnev and V. A. Kovtunenko, *Analysis of Cracks in Solids*, WIT Press, Southampton, Boston, 2000.

[25] A. M. Khludnev and J. Sokolowski, *Smooth domain method for crack problems*, Quarterly Appl. Math., to appear.

[26] N. Kikuchi and J. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM, Philadelphia, 1988.

[27] D. Kinderlehrer and G. Stamppachia, *An Introduction to Variational Inequalities and Their Applications*, Academic Press, 1980.

[28] T. Lewinski and J. J. Telega, *Plates, Laminates, and Shells. Asymptotic Analysis and Homogenization*, Ser. Adv. Math. Appl. Sci. 52, World Scientific, River Edge, NJ, 2000.

[29] K. Lhalouani and T. Sassi, *Nonconforming mixed variational formulation and domain decomposition for unilateral problems*, East-West J. Numer. Math., 7 (1999), pp. 23–30.

[30] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, Dunod, Paris, 1968.

[31] M. Moussaoui and K. Khodja, *Régularité des solutions d'un problème mêlé Dirichlet-Signorini dans un domaine polygonal plan*, Commun. Partial Differential Equations, 17 (1992), pp. 805–826.

[32] J. Nečas, *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

[33] L. Slimane, A. Bendali, and P. Laborde, *Mixed formulations for a class of variational inequalities*, M2AN, 38 (2004), pp. 177–201.

[34] L. Slimane, *Méthodes mixtes et traitement du verouillage numérique pour la résolution des inéquations variationnelles*, Ph.D. thesis, INSA Toulouse, France, 2001.

[35] Z.-H. Zhong, *Finite Element Procedures for Contact-Impact Problems*, Oxford University Press, Oxford, UK, 1993.

SIAM J. Numer. Anal.
Vol. 43, No. 3, pp. 1321–1347

TENSOR-KRYLOV METHODS FOR SOLVING LARGE-SCALE SYSTEMS OF NONLINEAR EQUATIONS*

BRETT W. BADER†

Abstract. This paper develops and investigates iterative tensor methods for solving large-scale systems of nonlinear equations. Direct tensor methods for nonlinear equations have performed especially well on small, dense problems where the Jacobian matrix at the solution is singular or ill-conditioned, which may occur when approaching turning points, for example. This research extends direct tensor methods to large-scale problems by developing three tensor-Krylov methods that base each iteration upon a linear model augmented with a limited second-order term, which provides information lacking in a (nearly) singular Jacobian. The advantage of the new tensor-Krylov methods over existing large-scale tensor methods is their ability to solve the local tensor model to a specified accuracy, which produces a more accurate tensor step. The performance of these methods in comparison to Newton-GMRES and tensor-GMRES is explored on three Navier–Stokes fluid flow problems. The numerical results provide evidence that tensor-Krylov methods are generally more robust and more efficient than Newton-GMRES on some important and difficult problems. In addition, the results show that the new tensor-Krylov methods and tensor-GMRES each perform better in certain situations.

Key words. nonlinear systems, tensor methods, Newton's method, linesearch, curvilinear linesearch, Krylov subspace methods

AMS subject classification. 65H10

DOI. 10.1137/040607095

1. Introduction.

This paper describes a new class of methods for solving the nonlinear equations problem

$$\text{given } F : \mathbb{R}^n \to \mathbb{R}^n, \quad \text{find } x_* \in \mathbb{R}^n \text{ such that } F(x_*) = 0, \tag{1.1}$$

where it is assumed that $F(x)$ is at least once continuously differentiable. Large-scale systems of nonlinear equations defined by (1.1) arise in many practical situations, including systems produced by finite-difference or finite-element discretizations of boundary value problems for ordinary and partial differential equations.

Standard direct methods, such as Newton's method, are impractical on large-scale problems because of their high linear algebra costs and large memory requirements. Thus, most current practical approaches for solving large problems involve approximately solving a local linear model and then using these "inexact" steps to locate the next point.

*Received by the editors April 21, 2004; accepted for publication (in revised form) November 2, 2004; published electronically October 14, 2005. This work was supported by Air Force Office of Scientific Research grant F49620-00-1-0162 and Army Research Office grant DAAG55-98-1-0176 while the author was at the Department of Computer Science, University of Colorado, Boulder. While at Sandia, the author also received support from DOE's Office of Science MICS program and Sandia's Laboratory Directed Research and Development program under grant LDRD 04-1024. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sinum/43-3/60709.html

†Computational Sciences Department, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-0316 (bwbader@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Two inexact versions of tensor methods already exist for solving large problems. Bouaricha [5] describes an implementation of a tensor method using Krylov subspace methods for linear equations, which involves constructing an inexact tensor step from the approximate solutions of two linear systems (with the same Jacobian matrix). In addition, Feng and Pulliam [18] have developed a "tensor-GMRES" method, which first finds the Newton-GMRES step and then solves for an approximate tensor step.

We propose three variants of a new approach for solving the large-scale nonlinear equations problem (1.1). These new methods are an extension of the class of standard tensor methods [28], which base each iteration on a simplified quadratic model of $F(x)$ such that the quadratic term is a low-rank secant approximation that augments the standard linear model. Specifically, the new algorithms are an amalgamation of various techniques, including tensor methods for nonlinear equations [28], Krylov subspace techniques [8], and an inexact solver framework [14], that make them well-suited for large-scale problems. Given the parallels to Newton-Krylov methods [8], we call the new algorithms "tensor-Krylov" methods. In a manner similar to Newton-GMRES, the tensor-Krylov methods calculate an inexact tensor step from a specially chosen Krylov subspace that facilitates the solution of a minimization subproblem at each step. The Krylov subspace generated is different from the standard implementation of GMRES and requires some modifications to a block-Krylov solver.

The key feature of these new methods is that the step satisfies the local tensor model to within a specified tolerance, making it possible to control the quality of the step. In addition, the new tensor-Krylov methods are aptly suited to target problems where the Jacobian at the root is singular or, at least, very ill-conditioned. Newton-based methods do not handle singular problems well because they converge linearly to the solution and, in some cases, with poor accuracy [11, 12, 13, 21]. On the other hand, tensor methods are superlinearly convergent on singular problems under mild conditions [17].

Because the tensor-Krylov methods borrow elements from both direct tensor methods and linear Krylov subspace methods, these topics are reviewed before introducing the tensor-Krylov formulations. Section 2 reviews direct tensor methods for solving small-scale systems of nonlinear equations and includes background on pertinent large-scale methods, namely linear Krylov subspace methods, Newton-GMRES, and existing large-scale tensor methods. Then section 3 describes three different approaches for solving the local tensor model using a Krylov-based method and wraps these local Krylov solvers into a large-scale method with options for various global strategies. With the complete tensor-Krylov nonlinear solver fully discussed, section 4 describes several fluid flow benchmark problems that serve as ambitious test problems. Finally, section 5 makes some concluding remarks and discusses directions for future research.

Throughout this paper, a subscript k refers to the current iterate of a nonlinear solver. We denote the Jacobian $F'(x)$ by $J(x)$ and abbreviate $J(x_k)$ as J_k. Similarly, $F(x_k)$ is abbreviated often as F_k. When the context is clear, we may drop the subscript k on J_k, F_k, a_k, and s_k while still referring to the current values at an iteration.

2. Background and review.

In this section, we introduce standard methods for solving systems of nonlinear equations. We provide a brief review of standard methods in section 2.1 and a short introduction to tensor methods in section 2.2. We extend these methods to large-scale problems in sections 2.3 and 2.4 by reviewing Newton-Krylov methods and existing large-scale tensor methods, respectively. General references for topics in nonlinear solvers include [15], [20], and [25].

2.1. Standard methods. In this paper, we will refer to a class of methods, which we will call standard methods, for solving (1.1) that are based on a linear local model. Most notable among these methods is Newton's method, which bases each iteration upon a linear local model $M_N(x_k + d)$ of the function $F(x)$ around the current iterate $x_k \in \mathbb{R}^n$:

$$M_N(x_k + d) = F(x_k) + J(x_k)d, \tag{2.1}$$

where $d \in \mathbb{R}^n$ is the step and $J(x_k) \in \mathbb{R}^{n\times n}$ is either the current Jacobian matrix or an approximation to it. A root of this local model provides the Newton step

$$d_N = -J(x_k)^{-1}F(x_k),$$

which is used to reach the next trial point. Thus, Newton's method is defined when J_k is nonsingular and consists of updating the current point with the Newton step

$$x_{k+1} = x_k + d_N. \tag{2.2}$$

Due to large arithmetic and storage costs, implementations of Newton's method using direct factorizations of $J(x_k)$ are not practical for large-scale problems.

2.2. Tensor methods. Tensor methods solve (1.1) by including more information in the local model than Newton's method. By solving this augmented local model, tensor methods tend to generate steps of better quality than standard methods, thus reaching the solution faster. The local tensor model has the generic form

$$M_T(x_k + d) = F_k + J_k d + \tfrac{1}{2}T_k dd, \tag{2.3}$$

where $T_k \in \mathbb{R}^{n\times n\times n}$ is a tensor, which includes second-order information and is where these methods get their name. This term is selected so that the model interpolates $p \le \sqrt{n}$ previous function values in the recent history of iterates, which makes T_k a rank p tensor. Most often p is 1 or 2, but computational evidence in [28] suggests that $p > 1$ actually adds little to the computational performance of the direct method.

For this paper, we focus on the case of $p = 1$ because the tensor-Krylov methods only use one secant update. In this case, the tensor model about x_k reduces to

$$M_T(x_k + d) = F_k + J_k d + \tfrac{1}{2}a_k(s_k^T d)^2, \tag{2.4}$$

where

$$a_k \in \mathbb{R}^n = \frac{2(F_{k-1} - F_k - J_k s_k)}{(s_k^T s_k)^2}, \tag{2.5}$$

$$s_k \in \mathbb{R}^n = x_{k-1} - x_k. \tag{2.6}$$

After forming the model, we use it to determine the step to the next trial point. Because (2.4) may not have a root, one solves the minimization subproblem

$$\min_{d\in\mathbb{R}^n} \|M_T(x_k + d)\|_2\,, \tag{2.7}$$

and a root or minimizer of the model is the tensor step. Due to the special form of (2.4), the solution of (2.7) in the nonsingular case reduces to solving a quadratic equation followed by solving a system of $n - 1$ linear equations in as many unknowns.

A practical approach for solving (2.7), which relates to the presentation of tensor-Krylov methods in section 3, uses two orthogonal transformations to reduce the problem to two subproblems that are more easily solved. Briefly, the first transformation finds an orthogonal $Q_1 \in \mathbb{R}^{n\times n}$ such that $s/\|s\|$ is the last column and permits a change in variables

$$d = Q_1\hat{d}.$$

The second transformation finds an orthogonal $Q_2 \in \mathbb{R}^{n\times n}$ such that $Q_2 J_k Q_1$ is upper triangular. Thus, applying the two transformations to (2.4) and setting it equal to zero yields the following triangular system of n equations in n unknowns

$$Q_2 F + Q_2 J Q_1 \hat{d} + \tfrac{1}{2} Q_2 a \|s\|^2 \hat{d}_n^2 = 0, \tag{2.8}$$

where $\hat{d}_n \in \mathbb{R}$ is the quadratic variable.

Then, breaking (2.8) into two smaller problems, the solution to (2.7) continues by first solving for $\hat{d}_n$ by minimizing the quadratic equation appearing in the last row of (2.8) and choosing the smaller magnitude minimizer if there are two. Using the value of $\hat{d}_n$ in (2.8), a triangular linear system of size $(n-1)\times(n-1)$ is revealed. Finally, the complete solution to (2.7) is found by solving this resultant system for the remaining components of $\hat{d}$ and then reversing the variable space transformation from the first step, $d = Q_1\hat{d}$.

2.3. Newton-Krylov methods. Up to this point, this review has discussed direct methods for the solution of small, dense problems such that the local model is solved using direct factorizations of the Jacobian matrix. Large, sparse systems often are solved successfully using a class of "inexact" Newton methods:

$$x_{k+1} = x_k + d_k, \quad \text{where} \quad J(x_k)d_k = -F(x_k) + r_k, \quad \|r_k\| \le \eta_k \|F(x_k)\|, \tag{2.9}$$

where the local model typically is solved only approximately at each step using a less expensive approach. Successively better approximations at each iteration preserve the rapid convergence behavior of Newton's method when nearing the solution. The computational savings reflected in this less expensive inner iteration are usually partially offset with more outer iterations, but the overall savings still are quite significant on large-scale problems.

The most common methods for approximately solving the local Newton model in (2.9) are Krylov-based methods. Newton-Krylov methods have the appeal of requiring almost no matrix storage due to their exclusive use of Jacobian-vector products, which may be calculated by a finite-difference directional derivative.

A linear Krylov subspace method is a projection method that seeks an approximate solution x_m to the linear system $Ax = b$ from an m-dimensional affine subspace $x_0 + \mathcal{K}_m$. Here, $\mathcal{K}_m$ is the Krylov subspace

$$\mathcal{K}_m(A, r_0) = span\{r_0, Ar_0, A^2r_0, \ldots, A^{m-1}r_0\},$$

where $r_0 = b - Ax_0$ is the residual at an initial guess x_0. A popular Krylov subspace method is the Generalized Minimum Residual method (GMRES) [26], which computes a solution $x_m \in x_0 + \mathcal{K}_m$ such that the residual norm over all vectors in $x_0 + \mathcal{K}_m$ is minimized. That is, at the mth step, GMRES finds x_m such that $\|b - Ax_m\|_2$ is minimized for all $x_m \in x_0 + \mathcal{K}_m$. One drawback of GMRES is the storage requirement of an orthogonal basis, which could be larger than a sparse Jacobian matrix unless m is kept small. Other Krylov methods, such as BiCGSTAB and TFQMR (see, e.g., [25]), do not have these additional storage requirements but may not be as robust. Newton-GMRES is a popular algorithm for solving large-scale problems and will be the standard algorithm in our numerical experiments.

2.4. Previous large-scale, sparse tensor methods. The large-scale tensor methods described in this paper are not the first tensor methods aimed at large-scale problems. Two other methods have been proposed, and we will discuss them now.

Bouaricha [5] describes a large-scale implementation of a tensor method using Krylov subspace methods for linear equations (GMRES and FOM), which involves constructing an inexact tensor step from the approximate solutions of $J^{-1}F_k$ and $J^{-1}F_{k-1}$. The approach involves finding the values $s^TJ^{-1}F_k$ and $s^TJ^{-1}a_k$, which are used to calculate an approximate value of s^Td_T, which multiplies $J^{-1}a_k$ in the final computation of the step. More precise details regarding an efficient implementation may be found in [1] or [5].

Despite some favorable results in [5], we have found Bouaricha's method to be not as competitive on more practical problems. The two main disadvantages of this method stem from the fact that two linear systems must be solved for each outer iteration and that an accurate value of s^Td_T is not calculated, which may lead to spurious steps. Due to these theoretical disadvantages and based upon our own numerical experience with the algorithm in [1], we will not consider any numerical comparisons with Bouaricha's algorithm.

Feng and Pulliam [18] describe another large-scale tensor method that they call "tensor-GMRES." It uses a Krylov subspace projection technique, namely GMRES, to find the approximate Newton step $d_N = d_0 + V_m y_m$. The Arnoldi process in GMRES generates a Hessenberg matrix H_m and an orthonormal basis for the Krylov subspace $\mathcal{K}_m$ in the columns of V_m. Given these key matrices, their algorithm proceeds to solve a projected version of the tensor model (2.4) along a subspace that spans the Newton step direction (i.e., the approximate tensor step is in the span of the Krylov subspace $\mathcal{K}_m^N$ and d_0, or equivalently the span of the matrix $[V_m, d_0]$). Thus, their algorithm solves the least-squares problem

$$\min_{d \in \{d_0\} \cup \mathcal{K}_m^N} \left\| F_k + J_k d + \tfrac{1}{2} P a (s^T d)^2 \right\|, \tag{2.10}$$

where P is the projection matrix $P = Y(Y^TY)^{-1}Y^T$ and $Y = J_k[V_m, d_0]$.

The algorithm has some difficult algebra (details may be found in [18]), but the design is actually rather straightforward. The algorithm may be viewed as an extension of Newton-GMRES, where the inexact Newton step is calculated via GMRES in the standard way. The tensor step is calculated subsequently using the Krylov subspace information generated for the Newton step. In this way, the method is also consistent with preconditioning techniques and a matrix-free implementation, which makes it appealing for general use.

The extra work and storage beyond the GMRES method is actually quite small, and the analysis in [18] shows that the same superlinear convergence properties for the unprojected tensor model considered in [17] also hold for the projected tensor model in (2.10). These properties are evident in the numerical results of [18], which show the superlinear convergence behavior of tensor-GMRES on the singular and nearly singular problems, where the Newton-GMRES method exhibits linear convergence due to a lack of sufficient first-order information. The margin of improvement (in terms of reduction of nonlinear iterations over Newton's method) spanned 20–55 percent on the simpler problems and 32–60 percent on the more difficult Euler problem.

However, there are a few potential disadvantages related to the Feng and Pulliam method. The variable space restriction on d in the minimization problem (2.10) illustrates a possible disadvantage, particularly when using preconditioners or restarted GMRES. The norm of the projected tensor model is only minimized to the extent that the Krylov subspace for the Newton-GMRES step is large enough to capture important directions in the tensor step. For example, consider using an *exact* preconditioner, i.e., J_k itself. One iteration of GMRES solves the Newton equations exactly, and the Newton step direction is along v_1, the first basis vector in V_m. Then, according to the Feng–Pulliam method, the approximate tensor step that solves (2.10) could only be a scalar multiple of the direction v_1 (assuming that $d_0 = 0$). A similar example may be developed when using restarted GMRES in the Feng–Pulliam method—if GMRES converges soon after a restart, then the orthonormal basis V_m is smaller than before the restart. A smaller basis may lead to a tensor step that solves (2.10) with more error due to fewer degrees of freedom.

Despite these hypothetical examples, it is unclear whether solving (2.10) in a smaller variable space will adversely affect the practical performance of this method when using preconditioners or restarted GMRES. Because the Newton step tends to undershoot (or overshoot) when first-order information is lacking in the local model, the solution to the tensor model is often nearly along the Newton direction, so the subspace restriction on d might not be a problem. The fact remains, however, that the Feng–Pulliam method solves the *projected* tensor model (2.10), which loses some information in the projection. In addition, the relative stopping tolerance η_k in the Newton-GMRES step has no direct relationship with the error in the tensor model.

3. Tensor-Krylov methods.

The new tensor-Krylov methods differ from previous large-scale tensor methods in their ability to solve the local tensor model to a specified tolerance. Using either the methods of Bouaricha or Feng and Pulliam, the residual error $\|M_T(x_k + d)\|$ must be computed explicitly, making it difficult to assess the quality of the approximate tensor step that they compute. In addition, the new methods avoid the costly solution of two linear systems (as in Bouaricha's method) and compute the solution to the full tensor model, as opposed to a projected tensor model (as in the Feng–Pulliam method).

In the same manner that GMRES is an algorithm for solving linear systems and Newton-GMRES is the nonlinear solver, we make a distinction between the solver for the local tensor model and the nonlinear solver. In this section, we describe three procedures for iteratively solving the local tensor model that use the concepts from linear Krylov subspace methods. We restrict ourselves to the rank-one tensor model in (2.4)–(2.6), which only interpolates the function value at the previous iterate. Because (2.4) may or may not have a root, we seek a solution to the minimization problem

$$\min_{d \in \mathcal{K}_m} \|M_T(x_k + d)\|_2 = \min_{d \in \mathcal{K}_m} \left\| F_k + J_k d + \tfrac{1}{2} a_k (s_k^T d)^2 \right\|_2, \tag{3.1}$$

where $\mathcal{K}_m$ is a specially chosen Krylov subspace that facilitates the solution of the quadratic model. The three tensor-Krylov methods differ in their choice of $\mathcal{K}_m$, which becomes their signature difference and dictates the algorithm. We differentiate the three variants by the size of their initial block subspace, identifying them as block-2, block-2+, and block-3. The reason for considering three variants is related to their complexity and usefulness as a block algorithm. The block-3 method is the most straightforward and most likely the best block implementation, while the block-2 methods are more complex but may work better in scalar implementations.

In sections 3.1–3.3, we start with a description of the three Krylov-based techniques for solving the local tensor model. Due to space considerations, we describe only the block-2 method in detail and refer to [1] for more detailed information on the other two methods. Section 3.1 covers all aspects of the block-2 method that are important to a nonlinear equations solver, including block-Krylov subspace issues, residual calculation, stopping conditions, preconditioning and scaling techniques, computation of the Newton step, and cost. Section 3.4 wraps the local solver into a complete tensor-Krylov algorithm for solving large-scale systems of nonlinear equations, and section 3.5 discusses the global strategies for the tensor-Krylov algorithm.

3.1. Block-2 method. The block-2 algorithm proceeds in a block-Krylov-like fashion, operating on a matrix of initial vectors V instead of the single residual vector of a linear system. In particular, this method uses a block-Krylov subspace composed of two initial vectors.

We begin by rearranging the local tensor model and noting that it looks like a linear system involving a linear combination of two right-hand sides:

$$Jd = -F_k - \tfrac{1}{2}a\beta^2, \tag{3.2}$$

where $\beta \equiv s^T d$. The right-hand side spans only the directions F_k and a; the vector s appears in the inner product $\beta = s^T d$, which is a scalar multiple for a that is unknown. Thus, the premise of the block-2 method is that we start with the initial block Krylov subspace $\mathcal{K}_0 = span\{a, F_k\}$ and build $\mathcal{K}_m = span\{a, F_k, Ja, JF_k, J^2a, J^2F_k, \dots\}$ to solve (3.1). Specifically, we consider the block of initial vectors

$$R_0 = [(Jd_0 + F_k),\ a], \tag{3.3}$$

where $d_0 \in \mathbb{R}^n$ is some initial guess for the step. Because the starting matrix R_0 uses the residual $F_k + Jd_0$, which depends on d_0, the block-2 method may be restarted with successively better initial guesses in a manner similar to restarted GMRES.

The first step of the algorithm computes the QR-factorization of R_0,

$$R_0 = VR = [v_1, v_2]R, \tag{3.4}$$

where $V \in \mathbb{R}^{n\times 2} = [v_1, v_2]$ is unitary and $R \in \mathbb{R}^{2\times 2}$ is upper triangular. A block-Arnoldi process then creates additional columns of an orthonormal basis V_m that spans the block-Krylov subspace

$$span\{V, JV, J^2V, J^3V, \dots\}. \tag{3.5}$$

There are several block-Arnoldi versions available for implementation, and the particular variant is not critical to the implementation of the tensor-Krylov method. The standard procedure works on a whole block $V \in \mathbb{R}^{n\times t}$ and adds t vectors—$t = 2$ in this case—to the subspace at a time. This procedure may work well when considering cache memory performance and may be considered in future research. However, we decided to implement the single-vector version of block-Arnoldi to more closely correspond with the scalar implementation of GMRES. The version in Algorithm 3.1 is very similar to the standard Arnoldi algorithm, which operates on a single vector at a time and is due to Ruhe [22] (see also [25]) for the symmetric case (block Lanczos).

Algorithm 3.1. Block Arnoldi Process—Ruhe's variant.

```
1. Choose $t$ initial orthonormal vectors $\{v_i\}_{i=1,\dots,t}$.
2. Choose a number of Arnoldi iterations to perform and set to $m$.
3. For $k = 1, \dots, m$ :
    (a) Set $j := k + t - 1$
    (b) Compute $w := Jv_k$
    (c) For $i = 1, 2, \dots, j$
        i. $h_{ik} := (w, v_i)$
        ii. $w := w - h_{ik} v_i$
    (d) $h_{j+1,k} := \|w\|_2$
    (e) If $h_{j+1,k} \neq 0$, then set $v_{j+1} := w/h_{j+1,k}$;
        Else if $t = 1$, then Stop;
        Else set $t := t - 1$ and continue.
```

The first step of the algorithm is to multiply a single vector, v_1, by the Jacobian matrix J and orthonormalize the resulting vector w against all j vectors $v_1, \dots, v_j$ ($j = t$ at the first iteration) in the orthonormal basis, building the subspace one vector at a time. Thus, a vector from the initial block $\{v_i\}_{i=1,\dots,t}$ is multiplied by J every t steps. Step 3(e) avoids a division by zero and is commonly referred to as the breakdown condition. In the scalar case ($t = 1$), a breakdown condition indicates that the solution is in the subspace spanned by the k basis vectors computed thus far. Here in the block case, we modify the usual condition to reduce the block dimension by one until it eventually reduces to the scalar case.

After m steps on the initial matrix $V \in \mathbb{R}^{n\times 2}$ defined in (3.4), the block-Arnoldi process produces an orthogonal matrix $V_{m+2} \in \mathbb{R}^{n\times(m+2)}$ and a matrix $\bar{H}_m \in \mathbb{R}^{(m+2)\times m}$ whose nonzero entries are the elements h_{ik} computed in the process. It is important to note that $\bar{H}_m$ is banded upper Hessenberg with two subdiagonals. The orthonormal basis V_{m+2} and the matrix $\bar{H}_m$ have an important relationship,

$$JV_m = V_{m+2}\bar{H}_m. \tag{3.6}$$

Continuing with the solution to (3.1), let the approximate solution at the mth iteration be

$$d = d_0 + V_m y, \tag{3.7}$$

where $V_m \in \mathbb{R}^{n\times m}$ is an orthonormal basis for the Krylov subspace generated in (3.5) and $y \in \mathbb{R}^m$ is unknown. Substituting (3.7) into the tensor model yields

$$\begin{aligned} M_T(x_k + d) &= F_k + Jd + \tfrac{1}{2}a(s^T d)^2 \\ &= F_k + Jd_0 + JV_m y + \tfrac{1}{2}a(s^T d_0 + s^T V_m y)^2 \\ &= r_0 + JV_m y + \tfrac{1}{2}a(s^T d_0 + s^T V_m y)^2, \end{aligned} \tag{3.8}$$

where $r_0 = F_k + Jd_0$, which is also the residual for the Newton model when using the initial guess d_0. Having r_0 permits the calculation of the approximate Newton step later in the algorithm.

Let $Q_1 \in \mathbb{R}^{m\times m}$ be an orthogonal matrix that has $V_m^T s / \left\|V_m^T s\right\|$ in the last column, and let the vector $\hat{y} \in \mathbb{R}^m$ be defined by the following transformation:

$$y = Q_1 \hat{y}. \tag{3.9}$$

Then, using Q_1 and $\hat{y}$, the simplification of (3.8) continues:

$$\begin{aligned} M_T(x_k+d) &= r_0 + JV_m y + \tfrac{1}{2}a(s^T d_0 + s^T V_m y)^2 \\ &= r_0 + JV_m Q_1 \hat{y} + \tfrac{1}{2}a(s^T d_0 + (V_m^T s)^T Q_1 \hat{y})^2 \\ &= r_0 + JV_m Q_1 \hat{y} + \tfrac{1}{2}a(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2, \end{aligned} \tag{3.10}$$

where $\hat{y}_m$ is the mth element of $\hat{y}$ and limits the quadratic part to a single unknown. After the next step, we will discuss a good choice for efficiently constructing the orthogonal matrix Q_1 that retains a desirable structure for solving this problem.

From (3.3) and (3.4), let $\bar{r}_1 \in \mathbb{R}^{m+2}$ and $\bar{a} \in \mathbb{R}^{m+2}$ denote the first and second columns of R, respectively, padded with m zeros to a length $m+2$. These definitions, along with (3.6), permit a change in the function space of (3.10):

$$\begin{aligned} M_T(x_k+d) &= r_0 + JV_m Q_1 \hat{y} + \tfrac{1}{2}a(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2 \\ &= r_0 + V_{m+2}\bar{H}_m Q_1 \hat{y} + \tfrac{1}{2}a(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2 \\ &= V_{m+2}\left(\bar{r}_1 + \bar{H}_m Q_1 \hat{y} + \tfrac{1}{2}\bar{a}(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2\right). \end{aligned} \tag{3.11}$$

Because the column-vectors of V_{m+2} are orthonormal, the original least-squares problem of (3.1) may be simplified:

$$\min_{d\in\mathcal{K}_m} \left\|M_T(x_k+d)\right\|_2 = \min_{d\in\mathcal{K}_m} \left\|V_{m+2}^T M_T(x_k+d)\right\|_2, \tag{3.12}$$

where

$$V_{m+2}^T M_T(x_k+d) = \bar{r}_1 + \bar{H}_m Q_1 \hat{y} + \tfrac{1}{2}\bar{a}(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2. \tag{3.13}$$

At this point, we want to preserve some structure of the problem by requiring that the product $\bar{H}_m Q_1$ does not need expensive updates to transform it to upper triangular form. In other words, we want the banded Hessenberg structure of $\bar{H}_m$ to be retained after multiplication with Q_1, adding at most another subdiagonal. That restriction may be accomplished with an orthonormal Q_1 that is itself a Hessenberg matrix. Efficiently constructing such an orthogonal matrix Q_1 in this algorithm involves some careful algebra, which we now discuss.

Let the orthonormal matrix Q_1 be represented by an orthogonal matrix times a diagonal scaling matrix:

$$Q_1 = \hat{Q}_1 D.$$

The diagonal scaling matrix has diagonal entries $D_{ii} = \frac{1}{\|\hat{Q}_1[:,i]\|}$ so that the columns in Q_1 have unit length. The last column in $\hat{Q}_1$ is the m-dimensional vector $V_m^T s$, and constructing an orthogonal Hessenberg matrix using this last column involves only two updates on the mth iteration: the (m,m) and $(m,m-1)$ elements. We define $\hat{Q}_1$ recursively:

$$\hat{Q}_1 = \begin{pmatrix} \hat{Q}_1[1:m-1;1:m-2] & \hat{Q}_1[1:m-1;m-1] & V_{m-1}^T s \\ 0 & \frac{-(V_{m-1}^T s)^2}{v_m^T s} & v_m^T s \end{pmatrix}, \tag{3.14}$$

where the initial $\hat{Q}_1[1;1] = v_1^T s$. The matrix Q_1 may be represented by three vectors: the mth column of $\hat{Q}_1$ holding $V_m^T s$, the subdiagonal of $\hat{Q}_1$, and the entries in the diagonal scaling matrix. In addition, because only the two elements in the last row are new on the mth iteration, the product $\bar{H}_m Q_1$ has only two newly updated columns, which can be computed in $2(m+1)+2$ multiplications. Using a simple example with $m=4$, we may represent this graphically:

$$\bar{H}_m Q_1 = \begin{pmatrix} x & x & x & \star \\ x & x & x & \star \\ x & x & x & \star \\ & x & x & \star \\ & & x & \star \\ & & & \star \end{pmatrix} \begin{pmatrix} x & x & x & x \\ x & x & x & x \\ & x & x & x \\ & & \star & \star \end{pmatrix} = \begin{pmatrix} x & x & (x+\star) & (x+\star) \\ x & x & (x+\star) & (x+\star) \\ x & x & (x+\star) & (x+\star) \\ x & x & (x+\star) & (x+\star) \\ & x & (x+\star) & (x+\star) \\ & & \star & \star \end{pmatrix},$$

where a $\star$ represents a new number on the mth iteration and an x represents a nonzero from a previous iteration.

Given that the matrix $\bar{H}_m$ has two subdiagonals and that the matrix product $\bar{H}_m Q_1$ has three subdiagonals, the structure of (3.13) is

$$V_{m+2}^T M_T(x_k+d) = \begin{pmatrix} \star \\ \\ \\ \\ \\ \\ \\ \\ \end{pmatrix} + \begin{pmatrix} \star & \star & \star & \cdots & \star & \star \\ \star & \star & \star & & \star & \star \\ \star & \star & \star & & \star & \star \\ \star & \star & \star & & \star & \star \\ & \star & \star & & \star & \star \\ & & \star & \cdots & \star & \star \\ & & & \ddots & \vdots & \vdots \\ & & & & \star & \star \end{pmatrix} \hat{y} + \begin{pmatrix} \star \\ \star \\ \\ \\ \\ \\ \\ \\ \end{pmatrix} (s^T d_0 + \left\| V_m^T s \right\| \hat{y}_m)^2.$$

Eliminating the subdiagonals of $\bar{H}_m Q_1$ may be accomplished with a series of Givens rotations or Householder reflections. Let $Q_2 \in \mathbb{R}^{(m+2)\times(m+2)}$ be the product of all Givens rotations or Householder reflections applied to the system, and let the variables $\tilde{r}_1$, $\tilde{a}$, and $\tilde{H}_m$ denote the following transformed vectors:

$$\tilde{r}_1 \equiv Q_2 \bar{r}_1, \tag{3.15}$$

$$\tilde{a} \equiv Q_2 \bar{a}, \tag{3.16}$$

$$\tilde{H}_m \equiv Q_2 \bar{H}_m Q_1. \tag{3.17}$$

Using (3.15)–(3.17), we may premultiply (3.13) by Q_2 as a step toward the least-squares solution:

$$Q_2 V_{m+2}^T M_T(x_k+d) = \tilde{r}_1 + \tilde{H}_m \hat{y} + \tfrac{1}{2}\tilde{a}(s^T d_0 + \left\| V_m^T s \right\| \hat{y}_m)^2, \tag{3.18}$$

which has the following structure:

$$Q_2 V_{m+2}^T M_T(x_k+d) = \begin{pmatrix} \star \\ \star \\ \star \\ \vdots \\ \star \\ \star \\ \star \end{pmatrix} + \begin{pmatrix} \star & \star & \star & \cdots & \star \\ & \star & \star & & \star \\ & & \star & \cdots & \star \\ & & & \ddots & \vdots \\ & & & & \star \\ \\ \\ \end{pmatrix} \hat{y} + \begin{pmatrix} \star \\ \star \\ \star \\ \vdots \\ \star \\ \star \\ \star \end{pmatrix} (s^T d_0 + \left\| V_m^T s \right\| \hat{y}_m)^2.$$

This system has $m+2$ equations in m unknowns, and the last three rows are quadratic equations in the variable $\hat{y}_m$.

By the orthonormality of Q_2 and V_{m+2}, minimizing $\left\|Q_2 V_{m+2}^T M_T(x_k+d)\right\|$ is the same as minimizing $\|M_T(x_k+d)\|$. Thus, the solution to (3.1) involves minimizing the last three rows of (3.18), which requires finding the optimum value for $\hat{y}_m$:

(3.19)

$$\min_{\hat{y}_m\in\mathbb{R}} \left\|\tilde{r}_1[m:m+2] + \tilde{H}_m[m:m+2,m]\hat{y}_m + \tfrac{1}{2}\tilde{a}[m:m+2](s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2\right\|.$$

Problem (3.19), which has a closed-form solution, involves minimizing a quartic equation in a single unknown. The minimizers correspond to the critical points of the quartic equation in the objective function of (3.19) are thus among the real roots of a cubic equation found by differentiating the quartic equation with respect to $\hat{y}_m$. Thus, (3.19) can have one or two minimizers, in which case we choose the minimizer that makes $\beta = (s^T d_0 + \|s\|\,\hat{y}_m)$ have smaller magnitude. This choice is not necessarily the global minimizer of (3.19). Justification for choosing the smaller magnitude minimizer comes both from the step itself and from considering the sequence of iterates. Choosing the smaller magnitude minimizer is consistent with the approach used in direct tensor methods and results in the inexact tensor step being closer to the inexact Newton step. Also, if we consider the sequence of values $\{\beta_j\}$ for the first j iterations, then this sequence converges to a single number when choosing the smaller magnitude minimizer. If we were to choose the global minimizer, then the sequence $\{\beta_j\}$ could oscillate between two numbers, and thus the residual error would not necessarily be monotonically decreasing (β enters into the residual calculation via $\frac{1}{2}a\beta^2$). Hence, the smaller magnitude minimizer has more theoretical appeal and is used here.

As a simpler alternative for determining $\hat{y}_m$, we mention an approach that would approximately minimize $\left\|Q_2 V_{m+2}^T M_T(x_k+d)\right\|$. Instead of solving a quartic equation, we solve the single quadratic equation in the mth row of (3.18), choosing the root such that $\beta = (s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)$ has smaller magnitude. If the equation does not have a root, then we choose the value of $\hat{y}_m$ that minimizes the quadratic equation. The difference between $\hat{y}_m$ found in this manner and $\hat{y}_m$ found by minimizing (3.19) is usually negligible once the relative residual decreases by about two orders of magnitude.

Once the minimizer $\hat{y}_m$ is determined, the remaining elements of $\hat{y}$ may be found by computing a single right-hand side using the value $\hat{y}_m$ and solving the resultant $(m-1)\times(m-1)$ linear system found by neglecting the last 3 rows of (3.18).

At this stage, the vector $\hat{y}$ contains the coefficients for the linear combination of basis vectors $\{v_i\}$. Thus, the approximate tensor step that solves (3.1) is

$$d_T = d_0 + V_m Q_1 \hat{y}.$$

The decision for stopping the Arnoldi process so that the approximate step solves the tensor model to a specified tolerance appears before the computation of the explicit step, which is at an inconvenient location. GMRES has a similar dilemma but uses an efficient approach in its least-squares solution. With GMRES, the least-squares error $\|b-Ax\|_2$ is equal to the last element of $Qe_1\|b\|$, where Q is the product of all Givens rotations to transform the Hessenberg matrix to upper triangular form and e_1 is the unit vector $(1,0,0,\dots)^T$. Similarly, the last 3 rows of (3.18) pertain to the least-squares error of the local tensor model and may be used in stopping conditions.

There are two possible implementations for computing a stopping condition in these Krylov-based methods, and they are fundamentally similar. Both may be checked without explicitly computing the approximate step d_m after each step in the Arnoldi process.

The practical stopping condition that is used in our numerical tests is similar to GMRES in that it involves computing the norm of the remaining rows below the triangular part of $\tilde{H}_m$. That is, we neglect the contribution from the quadratic equation in row m of (3.18) and only calculate the norm of the last two rows when computing the least-squares error. This computation does not include any contribution from $\tilde{H}_m$ because its last two rows contain only zeroes. Thus, the practical stopping condition may be simplified to

$$\left\| \tilde{r}_1[m+1:m+2] + \tilde{a}[m+1:m+2]\tilde{\beta}_m^2 \right\| \leq \eta_k \left\| F(x_k) \right\|_2, \tag{3.20}$$

where η_k is the relative stopping tolerance and the norm covers only the last two rows of the vectors $\tilde{r}_3$ and $\tilde{a}$. We point out that (3.20) requires the computation of $\tilde{\beta}_m$ at each iteration m. Calculating $\tilde{\beta}_m$ is an $\mathcal{O}(1)$ calculation using the first-order condition for a minimizer of $|q(\tilde{\beta})|$.

Another stopping condition, which is briefly mentioned here but covered in more detail in [1], considers how close the residual norm at the approximate step d_m comes to the minimum residual norm at the exact step d_T. In other words, the comparison is

$$\| M_T(x_k + d_m) \| - \| M_T(x_k + d_T) \| \leq \eta_k \| F(x_k) \| . \tag{3.21}$$

A practical implementation of (3.21) is straightforward. The residual error $\|M_T(x_k + d_m)\|$ equals the minimum value calculated in (3.19), and the current estimate of $\|M_T(x_k + d_T)\|$ at the mth iteration equals the value of the quadratic equation on the mth row.

Of the two conditions, (3.20) is a more demanding test than (3.21), and an implementation using (3.20) may require more iterations before satisfying the same relative tolerance.

Algorithm 3.2 describes the whole process for computing the approximate tensor step d_T. The algorithm is a basic implementation that progressively updates $\bar{H}_m$ to $\tilde{H}_m$ after each step in the Arnoldi process.

Algorithm 3.2. Block-2 Iterative Tensor Method.

```
1. Choose a relative residual tolerance $\eta \in [0,1)$ and maximum subspace dimension $m_{max}$.
2. Given the local tensor model $M_T(x_k + d) = F_k + Jd + \frac{1}{2}a(s^T d)^2$, previous function value $F_{k-1}$, and initial guess $d_0$, form the block of initial vectors $R_0 = [(Jd_0 + F_k), a]$, where $a = \frac{2(F_{k-1} - F_k - Js)}{(s^T s)^2}$ and $s = x_{k-1} - x_k$.
3. Perform a partial QR-factorization on $R_0$ such that $R_0 = [v_1, v_2]R = VR$.
4. For $m = 1, 2, \ldots, m_{max}$ do:
    (a) Let the two columns of $R$, appended with $m$ zeroes to a length $m+2$, be labeled $\bar{r}_1$ and $\bar{r}_2$, respectively.
    (b) Form the vector $Jv_m$ and orthogonalize it against the previous $v_1, \ldots, v_{m+1}$
```

vectors via the block Arnoldi process, Algorithm 3.1:

$$\begin{aligned} &w := Jv_m \\ &h_{i,m} := (w, v_i),\ i = 1, 2, \ldots, m+1 \\ &w := w - \sum_{i=1}^{m+1} h_{i,m} v_i \\ &h_{m+2,m} := \|w\|_2 \\ &v_{m+2} := w/h_{m+2,m}. \end{aligned}$$

(c) Define $\bar{H}_m$ to be the $(m+2)\times m$ upper banded Hessenberg matrix whose nonzero entries are the coefficients $h_{ij}, 1 \le i \le m+3, 1 \le j \le m$, and define $V_m = [v_1, v_2, \ldots, v_m]$.

(d) Let $Q_1 \in \mathbb{R}^{m\times m}$ be an orthogonal Hessenberg matrix that has $V_m^T s / \left\|V_m^T s\right\|$ in the last column and be computed via (3.14), and let the vector $\hat{y} \in \mathbb{R}^m$ be defined by the following transformation $y = Q_1 \hat{y}$.

(e) Let $\bar{h}_{m-1}$ and $\bar{h}_m$ denote the two newly updated columns of the matrix-matrix product $\bar{H}_m Q_1$.

(f) Using Householder reflections, transform the $(m+2)\times m$ system $\bar{r}_1 + \bar{H}_m Q_1 \hat{y} + \frac{1}{2}\bar{a}(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2$ into $\tilde{r}_1 + \tilde{H}_m \hat{y} + \frac{1}{2}\tilde{a}(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2$, which involves applying all previous reflections to $\bar{h}_{m-1}$ and $\bar{h}_m$, followed by two new reflections to zero three elements in $\bar{h}_{m-1}$ and two elements in $\bar{h}_m$. Apply these reflections to the vectors $\tilde{r}_1$ and $\tilde{a}$.

(g) Find the minimizer $\hat{y}_m$ of (3.19) such that $\beta = (s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)$ has smaller magnitude and minimizes the least-squares error in (3.19).

(h) Let ρ_m represent the error estimate of solving the local tensor model. Either set ρ_m equal to the norm of the last 2 rows of (3.18), or set ρ_m equal to the norm of the last 3 rows of (3.18) minus the absolute value of the mth row.

(i) If $\rho_m \le \eta \|F_k\|$, then proceed to step 5 to calculate the approximate step.

5. Form the approximate solution:

(a) Find the remaining $m-1$ elements of the vector $\hat{y}$ by solving the first $m-1$ rows of the linear system

$$\tilde{H}_m \hat{y} = -\tilde{r}_1 - \tilde{h}_m \hat{y}_m - \tfrac{1}{2}\tilde{a}(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2,$$

where $\tilde{h}_m$ is the mth column of $\tilde{H}_m Q_1$.

(b) Form the approximate step $d_T = d_0 + V_m Q_1 \hat{y}$.

Just as with the Newton-GMRES algorithm, Algorithm 3.2 may be implemented matrix-free. Jacobian-vector products may be approximated by

$$J(x)v \approx \frac{F(x+\sigma v) - F(x)}{\sigma}.$$

In addition, we may apply preconditioning to accelerate convergence of the iterative methods. Consider a matrix M that approximates the current Jacobian J in some manner and is simple enough to permit inexpensive solutions to linear systems of the form $Mx = b$. Then, given M, the following left-preconditioned tensor model can be formed and solved:

$$\min_{d\in\mathcal{K}_m} \left\|M^{-1}F_k + M^{-1}Jd + \tfrac{1}{2}M^{-1}a(s^T d)^2\right\|. \tag{3.22}$$

The iterative tensor algorithms outlined above requires only minor modifications to incorporate left preconditioning—replace the call to Jv with $M^{-1}Jv$ in the Arnoldi process and premultiply all occurrences of F_k and a by M^{-1}. A separate subroutine that computes the action of M^{-1} times a vector is all that is needed.

Right preconditioning transforms the variable space. Given a preconditioner M, the following right-preconditioned tensor model can be formed and solved:

$$\min_{u \in \mathcal{K}_m} \left\| F_k + JM^{-1}u + \tfrac{1}{2}a(s^T M^{-1}u)^2 \right\|, \tag{3.23}$$

where the approximate step d is found from the solution of $Md = u$. Once again, the iterative tensor algorithm requires only minor modifications—replace the call to Jv with $JM^{-1}v$ in the Arnoldi process and replace the starting vector s in R_0 with $M^{-T}s$. If the matrix M^{-1} is not explicitly stored, then $M^{-T}s$ may be difficult to compute. The algorithm may be modified to avoid this step, however.

Of the two forms, right preconditioning is mildly preferred over left preconditioning because the norm of the residual $\|M_T(x_k + d)\|$, which enters directly into the stopping conditions, is unaffected with right preconditioning. In addition, right preconditioning guarantees that the inexact Newton step is a descent direction on the function, which is of paramount importance to a linesearch global strategy. With left preconditioning (with either GMRES or the Krylov-based tensor methods), the Newton step $d_N \in \mathcal{K}$ is no longer guaranteed to be a descent direction on $F(x)$, which has dire consequences in linesearch global strategies because backtracking along a step presupposes that the direction is a descent direction for eventual step acceptance. An advantage of left preconditioning is that the relative residual reduction can be a better indicator of true error reduction, thereby producing a more accurate step.

Scaling is of particular importance when solving systems of nonlinear equations, as noted in [15], and is a subject that is closely related to preconditioning. We only mention here that variable and function scaling is possible in the Krylov-based tensor method and may be implemented in a manner similar to preconditioning. Often it is desirable to scale the system on the left and precondition on the right to address the issues above.

An important remark about Algorithm 3.2 is that the Newton model is carried through the procedure, so the Newton step is calculated readily at the end of the algorithm and permits greater flexibility with the global strategy. In a manner similar to GMRES, we solve the linear least-squares problem

$$\min_{\hat{y} \in \mathbb{R}^m} \left\| \tilde{r}_1 + \tilde{H}_m \hat{y} \right\|,$$

which involves a back substitution with the upper triangular matrix $\tilde{H}_m$ and right-hand side $-\tilde{r}_1$. Then the approximate Newton step is given by

$$d_N = d_0 + V_m Q_1 \hat{y}.$$

Calculating the Newton step in addition to the tensor step adds a minimal cost. It involves a back substitution ($\frac{1}{2}m^2$ multiplications), matrix-vector product (m^2 multiplications), and a linear combination of basis vectors (nm multiplications), which is the dominant cost.

The cost of Algorithm 3.2 is very similar to the cost of GMRES. The extra work beyond GMRES involves the following:

1. Computation of one extra Jacobian-vector product to get a,

2. Partial QR-factorization of the $n \times 2$ matrix of initial vectors R_0 ($4n$ multiplications to get the second column of both Q and R),
3. Orthogonalization against one extra vector in the Gram–Schmidt process ($2nm$ multiplications),
4. Formation of an orthogonal Q_1 (nm multiplications for $V_m^T s$),
5. Computation of the new columns in the matrix-matrix product $\bar{H}_m Q_1$ (m^2 multiplications),
6. Orthogonal transformations involving Q_2 to form $\tilde{H}_m$ ($8m^2$ multiplications, if using Householder reflections),
7. Matrix-vector multiplication $Q_1 \hat{y}$ (m^2 multiplications).

Thus, when considering only the leading terms, the total cost beyond GMRES is $4n + 3nm + 6m^2$ multiplications plus one Jacobian-vector product. The bulk of the cost of GMRES is due to Gram–Schmidt orthogonalization in the Arnoldi process, $\mathcal{O}(nm^2)$, so the extra cost of the iterative Krylov-based tensor method is minor. This method compares favorably with the tensor-GMRES of Feng and Pulliam, which costs $5n + 4nm + 2m^2$ multiplications plus one extra Jacobian-vector product beyond GMRES.

As a final remark, extending these methods to run in parallel should be straightforward. Because the tensor-Krylov method uses the same basic vector and matrix operations as GMRES, we don't expect any implementation issues.

3.2. Block-2+ method. It is conceivable that the block-2 method above could generate a Krylov subspace for a step that minimizes (3.1) but does not include any information in the direction s, thereby neglecting any contribution from the second-order term $\frac{1}{2}a(s^T d)^2$. The aim of the the block-2+ method is to explicitly include the direction s in the subspace so that the inner product $s^T d$ is fully captured in the Krylov subspace while still working with a block of dimension two. Here we only mention the discriminating feature of this method and refer to [1] or [2] for the algorithmic details.

The problem that we are solving changes to finding the step d that solves the minimization problem

$$\min_{d \in \{s\} \cup \mathcal{K}_m} \left\| F_k + Jd + \tfrac{1}{2}a(s^T d)^2 \right\|_2 . \tag{3.24}$$

The procedure is basically the same as the block-2 method in section 3.1 but with some extra algebra and a special technique for augmenting the standard block-Krylov subspace with the new direction s at *each* Arnoldi iteration. This approach contrasts with the usual implementation of augmented Krylov subspace methods [10] and is discussed in [1].

3.3. Block-3 method. The block-3 algorithm for solving (3.1) proceeds in a block-Krylov-like fashion, operating on a matrix of three initial vectors instead of the single residual r_0 of a linear system. By choosing three vectors, we may include information on the three known vectors in the local tensor model (s, a, and F_k) and allow a transformation of the variable space and function space in a manner similar to the method of orthogonal transformations of section 2.2. To that end, we consider the block of initial vectors

$$R_0 = [s,\ (Jd_0 + F_{k-1}),\ (Jd_0 + F_k)]. \tag{3.25}$$

The rationale for choosing these specific vectors is as follows. The vector s is listed first in order to isolate the inner product $s^T d$ (via $\|s\|\, v_1^T d$) and later create a single quadratic equation in a single unknown. The second vector is the residual involving the previous function value F_{k-1} and is needed for computing the tensor term a, (2.5). The third vector is the residual of the Newton equations, and it may be placed as the second or third column in R_0. Collectively, these three vectors are chosen specifically to compute the tensor term a later in the algorithm in addition to fully characterizing the local tensor model (i.e., the three known vectors F_k, a, s) with this initial subspace.

The block-3 algorithm is procedurally different from the block-2 algorithm because it uses orthogonal transformations and permutation matrices to switch rows and columns to isolate a quadratic equation in the mth row. After the block-Arnoldi process adds a basis vector and an extra column to $\bar{H}_m$, we perform a series of plane rotations to put the matrix $\bar{H}_m$ in upper triangular form. In its current ordering, the quadratic equation would not be isolated to a single variable in the first row and should be switched to the mth row. So we permute the first row and column with the mth row and column to facilitate an easier solution. After all of the orthogonal transformations and row/column permutations, the structure of the simplified problem is

$$QP_L V_{m+3}^T M_T(x_k + d) = \begin{pmatrix} \star \\ \star \\ \star \\ \vdots \\ \star \\ \star \\ \star \\ \star \end{pmatrix} + \begin{pmatrix} \star & \star & \star & \cdots & \star \\ & \star & \star & & \star \\ & & \star & \cdots & \star \\ & & & \ddots & \vdots \\ & & & & \star \\ & & & & \\ & & & & \\ & & & & \end{pmatrix} \hat{y} + \begin{pmatrix} \star \\ \star \\ \star \\ \vdots \\ \star \\ \star \\ \star \\ \star \end{pmatrix} (s^T d_0 + \|s\|\, \hat{y}_m)^2.$$

Otherwise, the block-3 algorithm is similar to the block-2 algorithm above.

3.4. Tensor-Krylov methods. With the introduction of the Krylov-based iterative methods for solving the local tensor model in sections 3.1–3.3, we return to solving the general nonlinear equations problem (1.1). The following algorithm outlines the tensor-Krylov algorithm, which at every outer iteration calls a Krylov-based iterative method for solving the local tensor model.

Algorithm 3.3. Tensor-Krylov Method.

1. Given the nonlinear equations problem $F(x)$, choose a starting point x_0 and set the maximum iteration counter k_{max}.
2. For $k = 0, 1, 2, \ldots, k_{max}$, do:
   (a) Choose a forcing term tolerance $\eta_k \in [0, 1)$.
   (b) If $k = 0$, then calculate the Newton-GMRES step d_N according to the relative tolerance η_k and proceed to step 2e.
   (c) Form the local tensor model $M_T(x_k + d) = F_k + Jd + \frac{1}{2}a(s^T d)^2$, where $F_k = F(x_k)$, $F_{k-1} = F(x_{k-1})$, $J = F'(x_k)$, $s = x_{k-1} - x_k$, and $a = \frac{2(F_{k-1} - F_k - Js)}{(s^T s)^2}$.
   (d) Compute the inexact tensor step d_T according to the relative tolerance η_k by approximately solving the local tensor model according to the methods of sections 3.1, 3.2, or 3.3.
   (e) Set $x_{k+1} = x_k + \lambda d$, where d and λ are chosen according to a linesearch strategy that uses the directions d_T and/or d_N.
   (f) If x_{k+1} is an acceptable approximation to a root of $F(x)$, then stop and signal a success.

When referring to Algorithm 3.3 that uses a specific Krylov-based local solver in step 2d as defined in sections 3.1, 3.2, or 3.3 (i.e., the block-2, block-2+, or block-3 methods), we will abbreviate the method as TK2, TK2+, and TK3, respectively.

The main advantage of this method over previous inexact tensor methods is that the inexact tensor step d_T satisfies the local tensor model to within the specified tolerance η_k. The tensor-GMRES method, on the other hand, computes the solution of a *projected* tensor model, which is missing second-order information in the direction of s and may compute a less desirable step. The Bouaricha method [5] uses the exact model, but the relationship $\|M(x + d_T)\| < \eta_k \|F_k\|$ is not guaranteed, thereby raising the possibility of less accurate steps.

3.5. Global strategy and step selection. Algorithm 3.3 needs a robust strategy for global convergence if neither the full tensor step nor Newton step is satisfactory in step 2e. While step 2e uses a linesearch strategy, a trust region strategy is still viable, albeit less straightforward. Here we discuss details regarding a linesearch implementation in the tensor-Krylov method.

The standard tensor linesearch of [28] and the TENSOLVE linesearch of [5, 6] are straightforward applications of backtracking along the tensor step, if it is a descent direction, or otherwise along the Newton direction. The curvilinear linesearch implementation of [4] requires a little adaptation. Because the curvilinear linesearch for tensor methods has posted encouraging results and has a nice theoretical basis, we will focus primarily on this linesearch implementation in the tensor-Krylov algorithm.

The curvilinear step $d_T(\lambda)$ is the solution of the modified tensor model $\lambda F + Jd + \frac{1}{2}a(s^T d)^2$, where λ is the linesearch parameter. Thus, in the tensor-Krylov algorithm, the local tensor model is likewise changed and recomputed. Fortunately, the scalar λ is carried through the process in a straightforward manner, irrespective of method, as will now be discussed. For this discussion, we focus on the block-2 method in Algorithm 3.2, but the procedure applies to the block-2+ and block-3 methods in the obvious way. The only trick to this implementation involves the scaling of the initial guess d_0; all other aspects are intuitive.

The derivation involves changing the block of initial vectors in (3.3) to include λ in the Newton residual,

$$R_0 = [\lambda(Jd_0 + F_k),\ a]. \tag{3.26}$$

The scalar λ follows through the steps of Algorithm 3.2 and changes (3.18) to

$$Q_2 V_{m+2}^T M_T(x_k + d) = \lambda \tilde{r}_1 + \tilde{H}_m \hat{y} + \tfrac{1}{2}\tilde{a}(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2, \tag{3.27}$$

which only differs by λ multiplying $\tilde{r}_1$. Up to this point in the algorithm, the presence of λ does not require any new computations. To be more precise, the basis vectors in V and the matrix $\tilde{H}_m$ are unchanged. The presence of λ in (3.27) does change the calculation of the vector $\hat{y}_m$, and the corresponding change in (3.19) is

$$\min_{\hat{y}_m \in \mathbb{R}} \left\| \lambda \tilde{r}_1 + \tilde{H}_m \hat{y} + \tfrac{1}{2}\tilde{a}(s^T d_0 + \left\|V_m^T s\right\| \hat{y}_m)^2 \right\|.$$

The remaining elements of $\hat{y}$ are found by solving the triangular system with a right-hand side modified by λ and the newly computed $\hat{y}_m$. Finally, the change in (3.26) corresponds to scaling both F_k and d_0 by λ, so the curvilinear step changes to

$$d_T(\lambda) = \lambda d_0 + V_m Q_1 \hat{y}, \tag{3.28}$$

where $\hat{y}$ is also a function of λ, as noted above.

We reiterate that the scalar λ may multiply $\tilde{r}_1$ after all orthogonal transformations, so the initial work in generating a Krylov basis and performing the subsequent orthogonal transformations to calculate d_T is not repeated for computing $d_T(\lambda)$. That is, the matrix of basis vectors V_m used in (3.28) contains the same vectors from the original computation of d_T; only the vector $\hat{y}$ depends on λ.

The additional cost of the curvilinear linesearch per trial is an extra backsolve per λ-value (an extra $\frac{1}{2}m^2$ multiplications) plus a linear combination of basis vectors (nm multiplications), which is the dominant cost. However, if using right preconditioning, one application of the preconditioner must be used, which increases the cost further. While these costs are more than in the other linesearches, they are probably still less than the cost of evaluating $F(x)$ and certainly less than the cost of evaluating $J(x)$ or the total cost of generating the Krylov subspace for the original computation of d_T. Alternative implementations that use other simplifications or approximations are discussed in [1].

It should be noted that other large-scale tensor methods, such as the tensor-GMRES method of Feng and Pulliam [18], could employ the curvilinear linesearch even though these other methods have subtle differences in calculating an inexact tensor step. This is because the curvilinear step is calculated from a simple scalar multiplication of the function value in the local tensor model and may be carried through the algebra of the step calculation to arrive at a parametric form of the curvilinear step.

As a final note, it is unclear how best to employ the adaptive forcing terms of Eisenstat and Walker [16] in the curvilinear linesearch or in tensor methods, in general. The theory behind adaptive forcing is to reduce oversolving of the linear system, especially when far from the nonlinear solution, which tends to prevent the approximate Newton step from being excessively long (e.g., if the Jacobian has small singular values) or from being nearly orthogonal to the gradient of $\|F\|$ (e.g., if the Jacobian is ill-conditioned); see [31]. It is not clear how the same theory should be applied to a tensor method. The simplified quadratic term augmenting the linear system provides good directional information when dealing with ill-conditioned systems, so a more accurate solution provides a qualitatively better step than the Newton step. On the other hand, we also wish to avoid oversolving the tensor model when possible, which argues for a larger η_k. Thus, appropriate forcing terms for tensor methods, or perhaps just more reasonable safeguards, are still open questions at this stage.

4. Computational results and discussion.

This section describes numerical tests aimed at comparing the tensor-Krylov methods with Newton-GMRES and tensor-GMRES [18]. Other numerical tests may be found in [1], and results on several ill-conditioned problems are included in [3].

4.1. Test results for the Chan problem.

To first highlight some basic differences among the various classes of methods with a numerical example, we present results of solving a simplified model of combustion phenomena using Matlab implementations. The problem is an elliptic PDE considered by Chan [9] that has multiple turning points and is given by

$$\Delta u + \lambda \left(1 + \frac{u + u^2/2}{1 + u^2/100} \right) = 0 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{4.1}$$

where Δ is the Laplace operator, $\lambda \in \mathbb{R}$ is a parameter, Ω is the domain $[0,1] \times [0,1]$, and $\partial\Omega$ is the boundary of Ω. For our tests, we set $\lambda = 7.978735$, which is very close

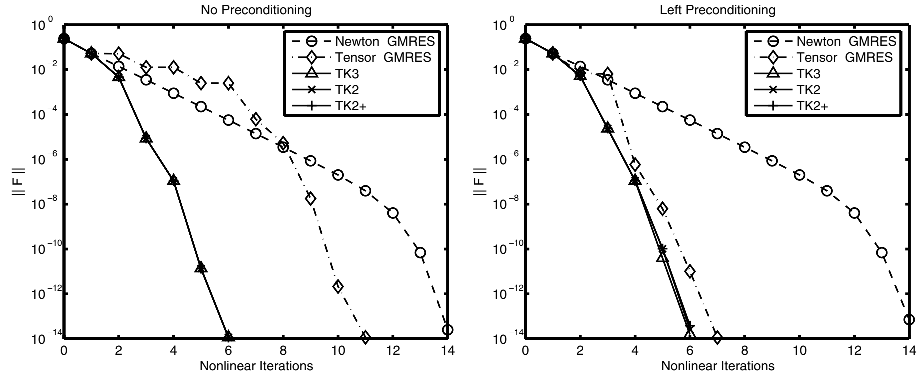


Fig. 4.1. *Iteration histories on the Chan problem with and without preconditioning.*

to a turning point in the problem. We used a centered differences discretization with 31×31 uniformly spaced grid points and started from the zero vector. We used the Laplacian as our preconditioner and used a constant forcing term of $\eta_k = 10^{-4}$. All methods converged to the same point.

Figure 4.1 presents the iteration history of all methods when solving with no preconditioning and with left preconditioning. Both plots show that Newton-GMRES has a long stretch of linear convergence rate due to the nearly singular Jacobian at the solution. The three tensor-Krylov methods have roughly the same quality of steps (as evidenced by their nearly identical trajectories). Tensor-GMRES does not share the same trajectory as the tensor-Krylov methods, which indicates that the projection of the tensor term in (2.10) loses some critical directional information. Similar numerical behavior is observed for different starting points, preconditioners, and/or values of the forcing term η_k on this problem, indicating that tensor-GMRES has some inherent difficulty with this type of problem.

4.2. Test results on fluid flow benchmark problems. For a broader comparison, we consider a couple of CFD benchmark problems described in [30] that are used for verification of fluid flow codes and solution algorithms: the 2D backward-facing step problem and the 2D and 3D thermal convection problem.

We implemented the algorithms in a software package called NOX [19], which is a C++ object-oriented nonlinear solver package being developed at Sandia National Laboratories. For objective comparisons, all of the methods, including Newton-GMRES and tensor-GMRES, used the same Arnoldi process (modified Gram–Schmidt) as the tensor-Krylov methods. That choice granted us more control over the algorithm and assured us of a controlled experiment. Thus, the results in this section do not reflect the most efficient and optimized implementations available.

We set up the numerical experiments to closely correspond to those in [30], using many of the same conditions and parameters. A successful termination was declared when both of the following stopping conditions were satisfied:

$$\|F(x_k)\| \leq \varepsilon_F \, \|F(x_0)\| \tag{4.2}$$

and

$$\frac{1}{\sqrt{n}} \, \|W d_k\| < 1, \tag{4.3}$$

where n is the total number of unknowns, d_k is the full Newton or tensor step, and W is a diagonal scaling matrix with entries

$$W_{ii} = \frac{1}{\varepsilon_r \left|x_{k_i}\right| + \varepsilon_a},$$

in which x_{k_i} is the ith element of the current solution x_k. We used the same parameters as in [30]: $\varepsilon_F = 10^{-2}$, $\varepsilon_r = 10^{-3}$, and $\varepsilon_a = 10^{-8}$.

In practice, the step length criterion (4.3) is more stringent than (4.2) and is necessary to resolve finer details of the fluid flow and transport by requiring that each ith element of the Newton or tensor step be small relative to its current value x_{k_i}. All successful runs, except for two noted in section 4.2.1, satisfied (4.2) with at least $\varepsilon_F = 10^{-8}$ and converged to the same solution. The last several iterations in each run were needed to satisfy (4.3).

If the test problem required more than 200 nonlinear (outer) iterations or if there was a linesearch failure (i.e., $f(x_c + \lambda d) \leq f(x_c) + \alpha\lambda\nabla f(x_c)^T d$, where $f(x) \equiv \frac{1}{2}\left\|F(x)\right\|_2$ and $\alpha = 10^{-4}$, could not be satisfied with $\lambda > 10^{-12}$ in at most 40 backtracks), then we declared a failure for the run.

The tests in [30] used a variety of forcing terms, particularly the adaptive forcing terms of Eisenstat and Walker [16]. As mentioned in section 3.5, more research is needed to determine how best to apply adaptive forcing terms to tensor methods. Consequently, we have used a constant forcing term of $\eta_k = 10^{-4}$ in the 2D problems and $\eta_k = 10^{-2}$ in the 3D problem. As in [30], we allowed the local solver (i.e., GMRES or its tensor-Krylov equivalent) a restart value of 200 with a maximum of 600 total iterations. If the local solver did not satisfy the desired tolerance within the 600 iterations, then we used the step computed thus far and tested for step acceptance with our global strategies. Restarting became more of an issue as the problem difficulty increased.

We used an explicit Jacobian, which our PDE code computed efficiently by a combination of analytic evaluation and numerical differentiation, and enabled the option for maximum accuracy in the Jacobian. We employed right preconditioning in all cases using an ILUT preconditioner [24], and we performed no variable or function scaling in the problems. The initial approximation was the zero vector for all cases.

We used a standard backtracking linesearch procedure for Newton-GMRES and used the complete tensor-GMRES algorithm in [18], including their globalization. For the tensor-Krylov methods, we used the curvilinear linesearch due to favorable theoretical and performance considerations in [4]. For selecting the linesearch parameter at each trial step, we used the λ-halving procedure (dividing λ by two at each inner iteration). Quadratic backtracking was an option, but it generally required more iterations and function evaluations than λ-halving across all methods in preliminary tests on these problems, so it was not used.

All tests were performed on a dual 3GHz Pentium Xeon desktop computer with 2GB of RAM, which was more than sufficient for these problems. However, the computer was not dedicated to these tests, so the timing statistics provided are only approximate and could be off by 10 percent or more relative to each other.

The fluid flow problems are set up using a particular spatial discretization of the governing steady-state transport equations for momentum and heat transfer in flowing fluids. These governing PDEs are given below. The unknown quantities in these equations are the fluid velocity vector ($\mathbf{u}$), the hydrodynamic pressure (P), and the temperature (T).

$$\text{Conservation of mass: } \nabla \cdot \mathbf{u} = 0 \tag{4.4}$$

$$\text{Momentum transport: } \rho\, \mathbf{u} \cdot \nabla \mathbf{u} - \nabla \cdot \mathbf{T} - \rho \mathbf{g} = 0 \tag{4.5}$$

$$\text{Energy transport: } \rho C_p \mathbf{u} \cdot \nabla T + \nabla \cdot \mathbf{q} = 0 \tag{4.6}$$

In these equations, $\mathbf{g}$ is the gravity vector, and ρ and C_p are the density and specific heat at constant pressure of the bulk fluid, respectively. The constitutive equations for the stress tensor $\mathbf{T}$ and heat flux $\mathbf{q}$ are

$$\begin{aligned} \mathbf{T} &= -P\mathbf{I} + \mu(\nabla \mathbf{u} + \nabla \mathbf{u}^T), \\ \mathbf{q} &= -\kappa \nabla T, \end{aligned}$$

where μ is the dynamic viscosity and κ is the thermal conductivity of the fluid.

The particular spatial discretization of (4.4)–(4.6) that we use is from a finite element reacting flow code called MPSalsa [27] developed at Sandia National Laboratories. MPSalsa generates an algebraic system of equations by a pressure-stabilized Petrov–Galerkin finite element formulation of the low Mach number Navier–Stokes equations with heat transport. This scheme uses equal-order interpolation of velocity and pressure, and we enabled the option for streamline upwinding to limit oscillations due to high grid Reynolds numbers. Since the publication of [30], the pressure-stabilized streamline upwinding Petrov–Galerkin formulation in MPSalsa has been changed to a Galerkin least squares–type method [29]. This stabilization method is slightly less dissipative, and the nonlinear convergence behavior for difficult problems can be less robust at higher Reynolds numbers. Consequently, this change precludes direct comparisons with results in [30].

To complete a problem's specification, boundary conditions are imposed on the governing PDEs, which we discuss in the subsections that follow. The problems differ only in their boundary conditions and in whether they use (4.4)–(4.6) or only (4.4)–(4.5). The next three subsections describe the test problems and their results.

4.2.1. Backward-facing step problem. This problem consists of a rectangular channel with a 1×30 aspect ratio in which a reentrant backward-facing step (i.e., a sudden expansion in the channel width) is simulated by injecting fluid with a fully developed parabolic velocity profile in the upper half of the inlet boundary and imposing a zero velocity on the lower half. The channel geometry and flowing fluid produce recirculation zones beneath the entering flow on the lower wall and, for sufficiently fast flow, farther downstream on the upper wall. This problem requires the solution of (4.4)–(4.5) on the unit square with the following Dirichlet boundary conditions:

$$\begin{aligned} \mathbf{u} = 24y(\tfrac{1}{2} - y)U_0\hat{x} \text{ at } x = 0,\ 0 \le y \le \tfrac{1}{2}, \\ \mathbf{u} = 0 \text{ at } x = 0,\ -\tfrac{1}{2} \le y < 0, \\ \mathbf{u} = 0 \text{ at } y = -\tfrac{1}{2}, \tfrac{1}{2}, \\ \mathbf{T}_{xx} = \mathbf{T}_{xy} = 0 \text{ at } x = 30, \end{aligned}$$

where $\hat{x}$ is the unit vector in the x-direction. Once the governing equations and boundary conditions are nondimensionalized, the Reynolds number (Re) appears, which is a measure of inertial forces to viscous forces. In our experiments, we increased the Reynolds number up to 800, which increases the nonlinear inertial terms in the momentum equation and makes the solution more difficult to obtain. Beyond Re = 800, it is not clear that the problem is stable and admits a physical solution. All solutions for this problem were computed on a 20 × 400 unequally spaced mesh, which has 25,263 unknowns.

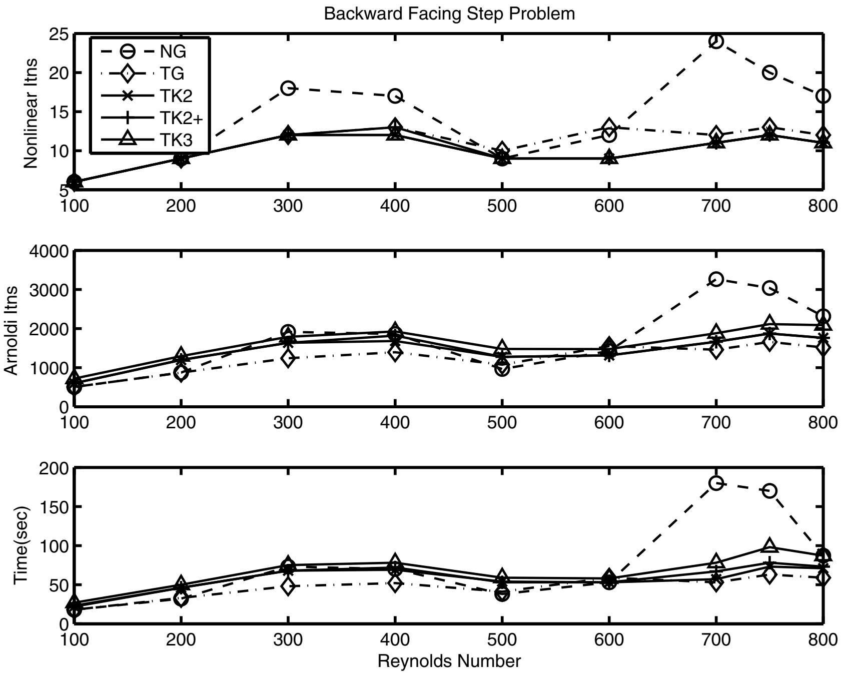


Fig. 4.2. *Results of the backward-facing step problem for the following methods: Newton-GMRES (○, dashed line), tensor-GMRES (◇, dash-dotted line), TK2 (×, solid line), TK2+ (+, solid line), and TK3 (△, solid line).*

The plots in Figure 4.2 show that all of the methods require about 10–12 iterations, on average, to solve, with Newton-GMRES requiring considerably more iterations in some cases. Newton's method tended to be more erratic, having slight difficulty at Re = 300 and 400, improvements at Re = 500 and 600, and then more difficulty on the three hardest problems. The two Newton solutions at Re = 700 and 750 actually converged to a local minimizer of the linesearch merit function yet still satisfied the relative residual reduction criterion of 10^{-2}. If ε_F in (4.2) were 10^{-3}, then these two runs would have been linesearch failures.

All of the tensor-Krylov methods share almost the exact same level of performance in terms of nonlinear iterations. Tensor-GMRES requires slightly more nonlinear iterations than the tensor-Krylov methods, but its local solve with GMRES is more efficient. Thus, for this problem, tensor-GMRES is more efficient than the tensor-Krylov methods by a small margin, and TK2 and TK2+ are more efficient than TK3 by about the same amount. There appears to be no distinct difference between TK2 and TK2+.

4.2.2. Thermal convection problem. This problem consists of the thermal convection (or buoyancy-driven) flow of a fluid in a differentially heated square cavity in the presence of gravity. It requires the solution of (4.4)–(4.6) on the unit square with the following Dirichlet and Neumann boundary conditions:

$$T = T_{cold},\ \mathbf{u} = 0 \text{ at } x = 0, \tag{4.7}$$

$$T = T_{hot},\ \mathbf{u} = 0 \text{ at } x = 1, \tag{4.8}$$

$$\frac{\partial T}{\partial y} = 0,\ \mathbf{u} = 0 \text{ at } y = 0, 1. \tag{4.9}$$

Once the governing equations and boundary conditions are nondimensionalized, two parameters appear: the Prandtl number (Pr) and the Rayleigh number (Ra). In our experiments, we fixed Pr = 1 and increased the Rayleigh number from Ra = 10^4 up to 2×10^7, which increases the nonlinear effects of the convection terms and makes the solution more difficult to obtain. The range in [30] is Ra = 10^3 to 10^6, but we shifted the range to explore the effectiveness of tensor methods on more difficult problems. We used a 100 × 100 equally spaced mesh, which has 40,804 unknowns. On this size mesh, it is unclear whether the choice of Ra $> 10^6$ admits a physically accurate and/or stable solution, but we are interested only in the relative performance of the numerical methods on this problem, which remain valid comparisons.

Figure 4.3 shows the overall performance of the methods on this problem. While difficult to see, Newton-GMRES performs a little better than the other methods on the easiest problem difficulties (Ra $\leq 10^5$). Yet as the Rayleigh number increases, Newton-GMRES requires increasingly more work to solve the problems. For Ra $\geq 10^7$, Newton-GMRES fails to solve the problem in 200 iterations. Thus, the trend for Newton-GMRES is a clear degradation in performance as the problem becomes more difficult to solve. In contrast, the tensor-Krylov methods with the curvilinear linesearch are much less affected by the transition and see a much smaller increase in execution time. Results of the tensor-Krylov methods with the old standard tensor linesearch (not shown) are less impressive but are still better than Newton-GMRES and tensor-GMRES on the harder problems.

Among the tensor-Krylov methods, TK2 is virtually identical to TK2+ in terms of nonlinear iterations, Arnoldi iterations, and execution time. Both are more efficient than TK3, which required more Arnoldi iterations at all Rayleigh numbers. This was due in part to restarts of the TK3 method. At Ra $\geq 3 \times 10^6$, restarts contributed to an increasing number of Arnoldi iterations for TK3. For both TK2 and TK2+, the local model was solved in less than 200 Arnoldi iterations.

Tensor-GMRES only does well on this problem at the easier difficulties (Ra $\leq 10^5$) due to fewer nonlinear iterations and an efficient use of GMRES. However, tensor-GMRES is unable to solve the most difficult problems in 200 nonlinear iterations or less at Ra $> 5 \times 10^6$, whereas all tensor-Krylov methods are able to solve the hardest problems. Even when employing the standard linesearch, the tensor-Krylov methods can solve up to Ra = 10^7 but fail on the hardest problem at Ra = 2×10^7.

4.2.3. 3D Thermal convection problem. This final problem uses slightly different test conditions to help assess the applicability for very large problems. The 3D version of this problem has the same boundary conditions as (4.7)–(4.9) but with an additional boundary condition in 3D

$$\frac{\partial T}{\partial z} = 0,\ \mathbf{u} = 0 \text{ at } z = 0, 1. \tag{4.10}$$

All solutions were computed on a $32 \times 32 \times 32$ equally spaced grid, resulting in 179,685 unknowns for the discretized problem. To force more restarts, we reduced the restart

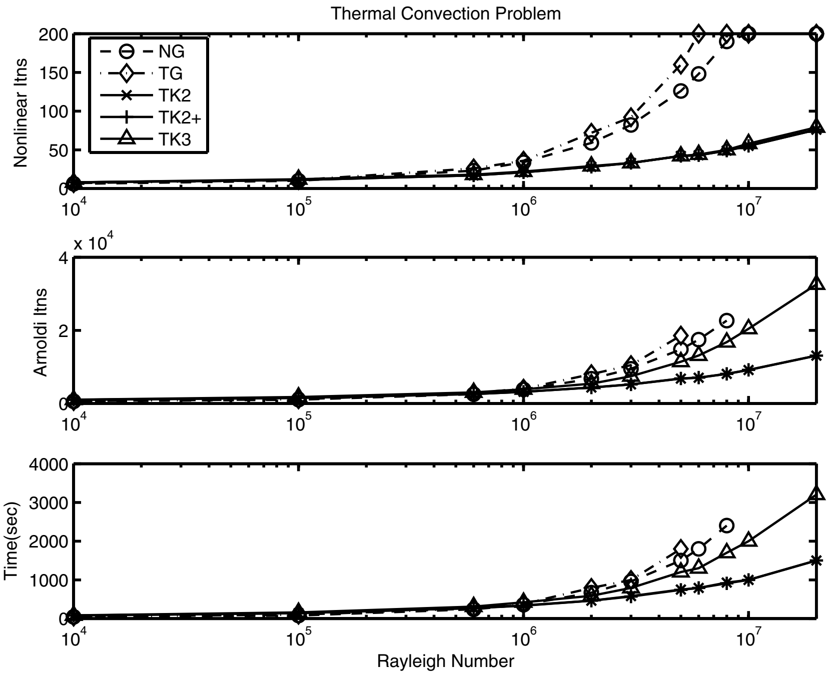


FIG. 4.3. *2D Thermal convection problem results for the following methods: Newton-GMRES (○, dashed line), tensor-GMRES (◇, dash-dotted line), TK2 (×, solid line), TK2+ (+, solid line), and TK3 (△, solid line). Only results for the successful runs are included in the bottom two plots.*

value to 75 and allowed a maximum of 225 total iterations. The quality of the preconditioner did not permit a restart value much less than 75 across all problem difficulties in this experiment.

Figure 4.4 shows the results for this problem. As the Rayleigh number increases, restarts become increasingly more important, and all methods start to suffer, especially the tensor methods. Without restarts, the results for all methods more closely resemble the 2D results in Figure 4.3. Up until about Ra $= 6 \times 10^6$, Newton-GMRES is the best method. At Ra $= 10^7$, TK2 and TK2+ outperform Newton-GMRES, but at harder difficulties, they produce less accurate steps and fail to solve in 200 nonlinear iterations. TK3 is adversely affected by restarts from the beginning.

Here, TK2 and TK2+ are better at restarting than TK3. We believe this is because in a single restart cycle, the block-2 methods have a polynomial expansion of the block-Krylov subspace that contains higher orders of the Jacobian matrix. That is, after m Arnoldi iterations, the block-2 methods have terms in the Krylov subspace up to $J^{\frac{m}{2}-1}$, whereas the block-3 method has terms up to $J^{\frac{m}{3}-1}$. For comparison, GMRES includes terms up to J^{m-1}.

Figure 4.4 also shows some evidence that tensor-GMRES loses effectiveness when restarting. A smaller subspace provides less information in the projection of the tensor term (i.e., $Pa(s^T d)^2$) as well as including a smaller basis for solving the minimization problem in (2.10), where $Pa(s^T d)^2$ acts as another right-hand side but the Krylov subspace generated by GMRES starts with the residual $F_k + J_k d_0$.

Other numerical tests investigating restarts that were performed in [1] support these findings. In addition, the tensor-Krylov methods tended to stall more frequently

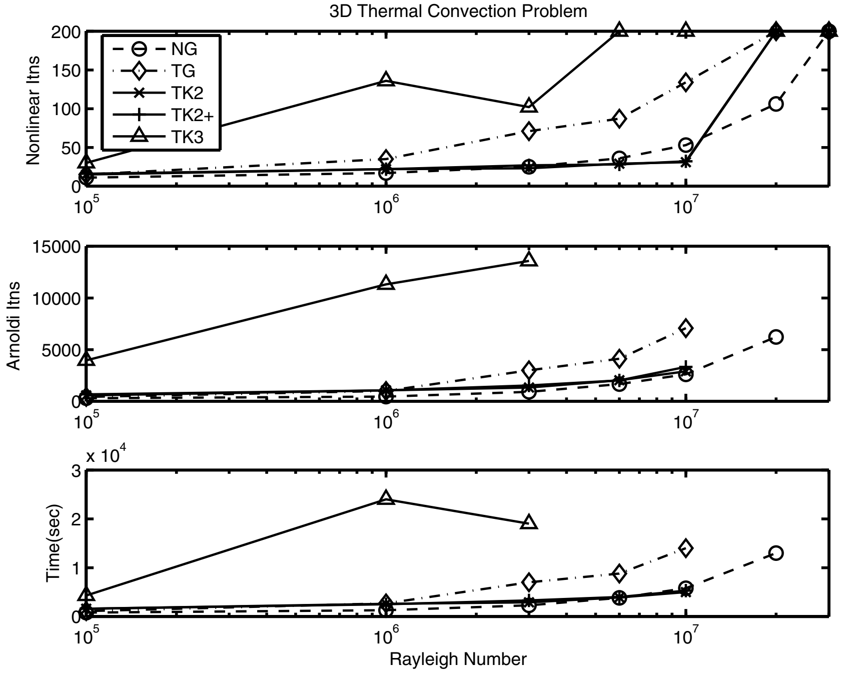


FIG. 4.4. *3D Thermal convection problem results for the following methods: Newton-GMRES (○, dashed line), tensor-GMRES (◇, dash-dotted line), TK2 (×, solid line), TK2+ (+, solid line), and TK3 (△, solid line). Only results for the successful runs are included in the bottom two plots.*

than restarted GMRES. That is, restarting the method did not always appreciably improve the step after another m iterations. This behavior may be attributed to the block-Krylov style of the Arnoldi process, which retains a constant vector in R_0 at each restart (i.e., s and/or a, depending upon the algorithm), keeping part of the subspace unchanged. Restarts rely on a new and different subspace to make progress.

5. Summary and conclusions.

The main objective of this research was to combine approaches based on direct tensor methods and Krylov subspace methods into an effective large-scale nonlinear equations solver. We developed three Krylov-based methods for iteratively solving the local tensor model, and we incorporated these three local solvers into an inexact nonlinear solver framework for different versions of a "tensor-Krylov" method, which we denoted TK2, TK2+, and TK3. The new tensor-Krylov methods are especially effective at solving large-scale problems that possess Jacobians at the solution that are highly ill-conditioned or singular. Algorithms based on Newton's method exhibit very slow convergence on such problems.

The new methods proposed in this paper solve the local tensor model in a novel fashion. Their costs per iteration are similar to GMRES, requiring only one Jacobian-vector product at each iteration and $O(nm)$ additional arithmetic operations beyond GMRES per solve. Relative to previous iterative tensor methods, they are the only methods that produce an approximate tensor step that solves the local tensor model to within a specified accuracy. In addition, these methods can compute an exact solution to the tensor model in at most n iterations (in exact arithmetic). The new tensor-Krylov methods can also utilize much of the technology developed for Newton-Krylov methods, including preconditioning and restarting.

Our numerical results suggest that the new tensor-Krylov methods clearly have some advantages over Newton-GMRES, especially as the problem becomes more difficult to solve or more ill-conditioned. In addition, the tensor-Krylov methods have some potential advantages over tensor-GMRES that make them likely to be beneficial on some important problems. Overall, tensor-GMRES and the tensor-Krylov methods are fairly similar—sometimes tensor-GMRES is better due to its efficient use of restarted GMRES(m), and sometimes tensor-Krylov methods are better due to a more accurate tensor step.

There are many different research questions at this point to explore. We mention four future extensions here. First, adaptive forcing terms like the form by Eisenstat and Walker [16] may help improve robustness. Second, an improved restart strategy may be necessary for some difficult problems, as witnessed with the 3D thermal convection problem results. One alternative to restarting is incomplete orthogonalization [7, 23], which would require modifications to the algorithms but may be a better strategy for coping with difficult problems. Third, changing the current scalar implementation of the block-Arnoldi method to a true block implementation (i.e., simultaneously multiplying a block of vectors by the Jacobian) may improve memory efficiency and make the tensor-Krylov methods even more economical and attractive. Fourth, the current tensor-Krylov and tensor-GMRES implementations need to manipulate data structures that are inaccessible in many linear solver packages, so we would like to simplify these methods and investigate better ways to incorporate standalone linear algebra packages. These changes would make the methods possibly more efficient, more robust, and more accessible than their current implementations.

Acknowledgments. Special thanks is owed to Prof. Robert Schnabel for his guidance and supervision of this research. The author is grateful to Michele Benzi and the anonymous referee for their constructive comments in the review process that helped improve this paper.

REFERENCES

[1] B. W. Bader, *Tensor-Krylov Methods for Solving Large-Scale Systems of Nonlinear Equations*, Ph.D. thesis, University of Colorado, Boulder, Department of Computer Science, 2003.

[2] B. W. Bader and A. H. Baker, *Implicitly augmented GMRES*, in preparation.

[3] B. W. Bader and R. B. Schnabel, *On the performance of tensor methods for solving ill-conditioned problems*, SIAM J. Sci. Comput., submitted.

[4] B. W. Bader and R. B. Schnabel, *Curvilinear linesearch for tensor methods*, SIAM J. Sci. Comput., 25 (2003), pp. 604–622.

[5] A. Bouaricha, *Solving Large Sparse Systems of Nonlinear Equations and Nonlinear Least Squares Problems Using Tensor Methods on Sequential and Parallel Computers*, Ph.D. thesis, University of Colorado, Boulder, 1992.

[6] A. Bouaricha and R. B. Schnabel, *Algorithm* 768: *TENSOLVE: A software package for solving systems of nonlinear equations and nonlinear least-squares problems using tensor methods*, ACM Trans. Math. Software, 23 (1997), pp. 174–195.

[7] P. N. Brown and A. C. Hindmarsh, *Reduced storage methods in stiff ODE systems*, J. Appl. Math. Comp., 31 (1989), pp. 40–91.

[8] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 450–481.

[9] T. F. Chan, *Newton-like pseudo-arclength methods for computing simple turning points*, SIAM J. Sci. Statist. Comput., 5 (1984), pp. 135–148.

[10] A. Chapman and Y. Saad, *Deflated and augmented Krylov subspace techniques*, Numer. Linear Algebra Appl., 4 (1997), pp. 43–66.

[11] D. W. Decker, H. B. Keller, and C. T. Kelley, *Convergence rate for Newton's method at singular points*, SIAM J. Numer. Anal., 20 (1983), pp. 296–314.

[12] D. W. Decker and C. T. Kelley, *Newton's method at singular points* I, SIAM J. Numer. Anal., 17 (1980), pp. 66–70.

[13] D. W. Decker and C. T. Kelley, *Newton's method at singular points* II, SIAM J. Numer. Anal., 17 (1980), pp. 465–471.

[14] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[15] J. E. Dennis, Jr. and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice-Hall, Englewood Cliffs, NJ, 1983.

[16] S. C. Eisenstat and H. F. Walker, *Choosing the forcing terms in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[17] D. Feng, P. D. Frank, and R. B. Schnabel, *Local convergence analysis of tensor methods for nonlinear equations*, Math. Program., 62 (1993), pp. 427–459.

[18] D. Feng and T. H. Pulliam, *Tensor-GMRES method for large systems of nonlinear equations*, SIAM J. Optim., 7 (1997), pp. 757–779.

[19] T. Kolda and R. Pawlowski, *NOX: An Object-oriented Nonlinear Solver Package.* http://software.sandia.gov/nox/.

[20] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer-Verlag, New York, 1999.

[21] G. W. Reddien, *On Newton's method for singular problems*, SIAM J. Numer. Anal., 15 (1978), pp. 993–996.

[22] A. Ruhe, *Implementation aspects of band Lanczos algorithms for computation of eigenvalues of large sparse symmetric matrices*, Math. Comput., 33 (1979), pp. 680–687.

[23] Y. Saad, *Krylov subspace methods for solving unsymmetric linear systems*, Math. Comp., 37 (1981), pp. 105–126.

[24] Y. Saad, *ILUT: A dual threshold incomplete ILU preconditioner*, Numer. Linear Algebra Appl., 1 (1994), pp. 387–402.

[25] Y. Saad, *Analysis of augmented Krylov subspace methods*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 435–449.

[26] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.

[27] A. G. Salinger, K. D. Devine, G. L. Hennigan, H. K. Moffat, S. A. Hutchinson, and J. N. Shadid, *MPSalsa, A Finite Element Computer Program for Reacting Flow Problems, Part 2—User's Guide*, Technical report SAND96-2331, Sandia National Laboratories, Albuquerque, NM, 1996.

[28] R. B. Schnabel and P. D. Frank, *Tensor methods for nonlinear equations*, SIAM J. Numer. Anal., 21 (1984), pp. 815–843.

[29] J. N. Shadid, *A fully-coupled Newton-Krylov solution method for parallel unstructured finite element fluid flow, heat and mass transfer simulations*, Int. J. Comput. Fluid Dyn., 12 (1999), pp. 199–211.

[30] J. N. Shadid, R. S. Tuminaro, and H. F. Walker, *An inexact Newton method for fully coupled solution of the Navier–Stokes equations with heat and mass transport*, J. Comput. Phys., 137 (1997), pp. 155–185.

[31] R. S. Tuminaro, H. F. Walker, and J. N. Shadid, *On backtracking failure in Newton-GMRES methods with a demonstration for the Navier–Stokes equations*, J. Comput. Phys., 180 (2002), pp. 549–558.

SIAM J. NUMER. ANAL.
Vol. 43, No. 3, pp. 1348–1361

EXISTENCE TESTS FOR SOLUTIONS OF NONLINEAR EQUATIONS USING BORSUK'S THEOREM[*]

ANDREAS FROMMER[†] AND BRUNO LANG[†]

Abstract. We show how interval arithmetic can be used in connection with Borsuk's theorem to computationally prove the existence of a solution of a system of nonlinear equations. It turns out that this new test, which can be checked computationally in several different ways, is more general than an existing test based on Miranda's theorem in the sense that it is successful for a larger set of situations. A numerical example is included.

Key words. nonlinear systems, Miranda's existence theorem, Borsuk's theorem, computational verification, interval analysis

AMS subject classifications. 47H10, 65G20, 65G40, 65H10

DOI. 10.1137/S0036142903438148

1. Introduction.

One of the most common problems in numerical analysis is to find a zero $\mathbf{x}^*$ of a nonlinear mapping

$$\mathbf{f} : D \subseteq \mathbb{R}^n \to \mathbb{R}^n, \qquad \mathbf{x} \mapsto (f_1(\mathbf{x}), \ldots, f_n(\mathbf{x})).$$

Any numerical method will usually only deliver an approximation $\mathbf{x}^0$ to $\mathbf{x}^*$, i.e., $\mathbf{f}(\mathbf{x}^0)$ will be fairly small, but not equal to zero. The question therefore arises whether there really exists an "exact" zero $\mathbf{x}^*$ of $\mathbf{f}$ in a (sufficiently small) neighborhood of $\mathbf{x}^0$ and whether there is a computational method to obtain such a neighborhood. Among the most successful such methods are Moore's existence test [11] and an existence test based on Miranda's theorem [12]. Both these tests rely on the use of interval arithmetic. We assume that the reader is familiar with the basics of interval arithmetic as described in [2], for example.

To fix our notation let us use square brackets to denote interval quantities. So $\mathbb{IR} = \{[a] = [\underline{a}, \overline{a}] : \underline{a} \le \overline{a}\}$ is the space of all (compact) intervals, $[\mathbf{x}] = ([x_1], \ldots, [x_n])^T$ with $[x_i] = [\underline{x}_i, \overline{x}_i] \in \mathbb{IR}$ is an interval vector, and, similarly, $[A] = ([a_{ij}])$ is an interval matrix. Given $[a] = [\underline{a}, \overline{a}] \in \mathbb{IR}$, it will sometimes be useful to denote the bounds of $[a]$ in a different way as

$$\underline{a} = \inf[a], \qquad \overline{a} = \sup[a].$$

For any interval $[a]$, let

$$\mathrm{mid}[a] = \frac{\underline{a} + \overline{a}}{2}$$

denote its midpoint; this is similar for interval vectors and interval matrices.

[*] Received by the editors November 26, 2003; accepted for publication (in revised form) November 17, 2004; published electronically October 14, 2005. The research of the first author was supported by VolkswagenStiftung within the project "Konstruktive Methoden der Nichtlinearen Dynamik zum Entwurf verfahrenstechnischer Prozesse," Geschäftszeichen I/79 288.

http://www.siam.org/journals/sinum/43-3/43814.html

[†] Faculty of Mathematics and Science, Department of Mathematics, Bergische Universität Wuppertal, Gaußstr. 20, D-42097 Wuppertal, Germany (frommer@math.uni-wuppertal.de, lang@math.uni-wuppertal.de).

Assume now that for a given interval vector $[\mathbf{x}]$ and a fixed $\hat{\mathbf{x}} \in [\mathbf{x}]$ we know an interval matrix (*slope matrix*) $[Y]$ such that

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}}) \in [Y] \cdot (\mathbf{x} - \hat{\mathbf{x}}) \quad \text{for all } \mathbf{x} \in [\mathbf{x}]; \tag{1.1}$$

i.e., for each $\mathbf{x} \in [\mathbf{x}]$ there is a matrix $Y_{\mathbf{x}} \in [Y]$ such that

$$\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}}) = Y_{\mathbf{x}} \cdot (\mathbf{x} - \hat{\mathbf{x}}).$$

If $\mathbf{f}$ is (Fréchet-)differentiable, $[Y]$ could be taken to be the interval hull $\Box D\mathbf{f}$ of the set $D\mathbf{f} = \{\mathbf{f}'(\mathbf{z}) : \mathbf{z} \in [\mathbf{x}]\}$, i.e., the intersection of all interval matrices containing $D\mathbf{f}$. In a more constructive manner, if $\mathbf{f}'(\mathbf{x})$ admits an interval arithmetic evaluation $\mathbf{f}'([\mathbf{x}])$ in the sense of [2], we can take $[Y] = \mathbf{f}'([\mathbf{x}]) \supseteq \Box D\mathbf{f}$.

Now, let $A \in \mathbb{R}^{n\times n}$ be some nonsingular matrix. Then Moore's existence test means checking the inclusion (1.2) given in the following theorem.

THEOREM 1.1. *Let* $\hat{\mathbf{x}} \in [\mathbf{x}]$ *and assume that*

$$[\mathbf{x}]^1 := \hat{\mathbf{x}} - A\mathbf{f}(\hat{\mathbf{x}}) + (I - A[Y])([\mathbf{x}] - \hat{\mathbf{x}}) \subseteq [\mathbf{x}]. \tag{1.2}$$

Then $[\mathbf{x}]^1$ *contains a zero* $\mathbf{x}^*$ *of* $\mathbf{f}$.

This theorem dates back to [11]. It follows from the fact that, due to the inclusion property of interval arithmetic, the interval vector $[\mathbf{x}]^1$ contains the range of the continuous function $\mathbf{x} - A\mathbf{f}(\mathbf{x}) = \hat{\mathbf{x}} - A \cdot \mathbf{f}(\hat{\mathbf{x}}) + (\mathbf{x} - \hat{\mathbf{x}}) - A \cdot (\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}})) = \hat{\mathbf{x}} - A\mathbf{f}(\hat{\mathbf{x}}) + (I - AY_{\mathbf{x}})(\mathbf{x} - \hat{\mathbf{x}})$ for $\mathbf{x} \in [\mathbf{x}]$. By Brouwer's theorem, this function therefore has a fixed point which is a zero of $\mathbf{f}$. Often A is taken as (an approximation to) the inverse of $\mathbf{f}'(\mathrm{mid}[\mathbf{x}])$.

The Miranda existence test is based on Miranda's existence theorem [10], which generalizes the intermediate value theorem to higher dimensions. The key idea is to consider the signs of components of $\mathbf{f}$ on opposite facets of an interval vector $[\mathbf{x}]$. Let us denote these facets by $[\mathbf{x}]^{i,+}, [\mathbf{x}]^{i,-}$, i.e.,

$$\left.\begin{aligned} [\mathbf{x}]^{i,+} &= ([x_1], \dots, [x_{i-1}], \overline{x}_i, [x_{i+1}], \dots, [x_n])^T, \\ [\mathbf{x}]^{i,-} &= ([x_1], \dots, [x_{i-1}], \underline{x}_i, [x_{i+1}], \dots, [x_n])^T \end{aligned}\right\} \; i = 1, \dots, n.$$

Miranda's existence theorem reads as follows.

THEOREM 1.2. *Assume that* $\mathbf{f} : [\mathbf{x}] \subseteq \mathbb{R}^n \to \mathbb{R}^n$ *is continuous and that*

$$f_i(\mathbf{x}) \ge 0 \textit{ for } \mathbf{x} \in [\mathbf{x}]^{i,+}, \;\; f_i(\mathbf{x}) \le 0 \textit{ for } \mathbf{x} \in [\mathbf{x}]^{i,-}, \quad i = 1, \dots, n. \tag{1.3}$$

Then $[\mathbf{x}]$ *contains a zero* $\mathbf{x}^*$ *of* $\mathbf{f}$.

Miranda's theorem essentially requires that each component f_i is of constant and opposite sign on the opposite faces $[\mathbf{x}]^{i,\pm}$. In practice, the theorem is not applied to the function $\mathbf{f}$ under consideration but to a *preconditioned* (affinely transformed) function $\mathbf{g} : [\mathbf{x}] \to \mathbb{R}^n$, where

$$\mathbf{g}(\mathbf{x}) = A \cdot \mathbf{f}(\mathbf{x}).$$

Again, A denotes some nonsingular matrix, and often A is taken as an approximation to the inverse of $\mathbf{f}'(\mathrm{mid}[\mathbf{x}])$.

Let $\mathrm{Rg}(g_i, [\mathbf{z}])$ denote the range of g_i over some interval vector $[\mathbf{z}]$. The continuity of g_i implies that $\mathrm{Rg}(g_i, [\mathbf{z}])$ is an interval. Then, verifying (1.3) for the function $\mathbf{g}$ is equivalent to checking, for each $i \in \{1, \dots, n\}$, the conditions

$$\sup \mathrm{Rg}(g_i, [\mathbf{x}]^{i,-}) \le 0 \le \inf \mathrm{Rg}(g_i, [\mathbf{x}]^{i,+}). \tag{1.4}$$

There are several ways to use interval arithmetic for enclosing the ranges in (1.4) into intervals $[m]^{i,\pm}$.

Naive: In a naive implementation, the exact ranges are replaced with the interval arithmetic evaluations of the functions over the facets,

$$[m_{\mathrm{N}}]^{i,\pm} := g_i([\mathbf{x}]^{i,\pm}).$$

Centered: A more sophisticated technique is due to Moore and Kioustelidis [12]. It works as follows: As before, let $\hat{\mathbf{x}} \in [\mathbf{x}]$ and let $[Y]$ be a slope matrix fulfilling (1.1). Since

$$\begin{aligned}\mathbf{g}(\mathbf{x}) - \mathbf{g}(\hat{\mathbf{x}}) &= A \cdot (\mathbf{f}(\mathbf{x}) - \mathbf{f}(\hat{\mathbf{x}})) \\ &= (A \cdot Y_{\mathbf{x}}) \cdot (\mathbf{x} - \hat{\mathbf{x}}),\end{aligned}$$

$A \cdot [Y]$ is a slope matrix for $\mathbf{g}$, and therefore we have, for $i = 1, \dots, n$,

$$\mathrm{Rg}(g_i, [\mathbf{x}]^{i,\pm}) \subseteq g_i(\hat{\mathbf{x}}) + (A \cdot [Y])_i \cdot ([\mathbf{x}]^{i,\pm} - \hat{\mathbf{x}}) =: [m_{\mathrm{C}}]^{i,\pm},$$

where $(\cdots)_i$ denotes the ith row of a matrix.

Facet-centered: Similar to [12], suppose now that for each facet $[\mathbf{x}]^{i,\pm}$ a point $\hat{\mathbf{x}}^{i,\pm} \in [\mathbf{x}]^{i,\pm}$ is given and an interval row vector (*slope vector*) $[\mathbf{y}]^{i,\pm}$ is known such that

$$g_i(\mathbf{x}) - g_i(\hat{\mathbf{x}}^{i,\pm}) \in [\mathbf{y}]^{i,\pm} \cdot (\mathbf{x} - \hat{\mathbf{x}}^{i,\pm}) \text{ for all } \mathbf{x} \in [\mathbf{x}]^{i,\pm}. \tag{1.5}$$

For example, $\hat{\mathbf{x}}^{i,\pm}$ might be the midpoint of the facet $[\mathbf{x}]^{i,\pm}$, and $[\mathbf{y}]^{i,+}$ and $[\mathbf{y}]^{i,-}$ might be taken as the ith row of $A \cdot \mathbf{f}'([\mathbf{x}]^{i,\pm})$ or $A \cdot \mathbf{f}'([\mathbf{x}])$. Of course, other choices are also possible. Then, for $i = 1, \dots, n$,

$$\mathrm{Rg}(g_i, [\mathbf{x}]^{i,\pm}) \subseteq g_i(\hat{\mathbf{x}}^{i,\pm}) + [\mathbf{y}]^{i,\pm} \cdot ([\mathbf{x}]^{i,\pm} - \hat{\mathbf{x}}^{i,\pm}) =: [m_{\mathrm{F}}]^{i,\pm}.$$

(Note that in the original paper [12], $\hat{\mathbf{x}}$ and $\hat{\mathbf{x}}^{i,\pm}$ are assumed to be the midpoints of $[\mathbf{x}]$ and $[\mathbf{x}]^{i,\pm}$, respectively, and that $[Y]$ and $[\mathbf{y}]^{i,\pm}$ are enclosures for the derivatives $\mathbf{f}'$ over $[\mathbf{x}]$ and of g_i' over $[\mathbf{x}]^{i,\pm}$, respectively. These requirements can be relaxed as formulated above.) The resulting computational tests are summarized in the following definition.

Definition 1.3. *The term* Miranda test *means checking, for each* $i \in \{1, \dots, n\}$, *one of the conditions*

$$\sup[m_{\mathrm{N}}]^{i,-} \le 0 \le \inf[m_{\mathrm{N}}]^{i,+}\,, \tag{1.6a}$$

$$\sup[m_{\mathrm{C}}]^{i,-} \le 0 \le \inf[m_{\mathrm{C}}]^{i,+}\,, \tag{1.6b}$$

$$\sup[m_{\mathrm{F}}]^{i,-} \le 0 \le \inf[m_{\mathrm{F}}]^{i,+}\,. \tag{1.6c}$$

Note that we do *not* require the same condition to be checked for all i.

By the inclusion property of interval arithmetic we know that each of the conditions (1.6a)–(1.6c) implies (1.4). So $\mathbf{g}$ and therefore $\mathbf{f}$ has a zero $\mathbf{x}^*$ in $[\mathbf{x}]$.

The question of how the Miranda criterion compares to the Moore criterion was raised in [3] and settled in [14, 8]. Indeed in [8] the Miranda test was shown to be more powerful than the Moore test in the sense that if (1.2) holds for some $[\mathbf{x}]$, $\hat{\mathbf{x}}$, $[Y]$, and A then (1.6b) also holds for $i = 1, \dots, n$ with the same $[\mathbf{x}]$, $\hat{\mathbf{x}}$, $[Y]$, and A.

As a further result in this direction we now show that, at the cost of some additional function evaluations, formulation (1.6c) of the Miranda test is more powerful than (1.6b).

THEOREM 1.4. *If* (1.6b) *is satisfied for some* i*,* $[\mathbf{x}]$*,* $\hat{\mathbf{x}}$*,* $[Y]$ *and* A*, then with the same* i*,* $[\mathbf{x}]$ *and* A *also* (1.6c) *holds true provided that we choose* $\hat{\mathbf{x}}^{i,\pm}$ *as the orthogonal projection of* $\hat{\mathbf{x}}$ *onto the facet* $[\mathbf{x}]^{i,\pm}$*,*

$$\hat{\mathbf{x}}^{i,\pm} = (\hat{x}_1, \dots, \hat{x}_{i-1}, [x_i]^{i,\pm}, \hat{x}_{i+1}, \dots, \hat{x}_n)$$

and that $[\mathbf{y}]^{i,\pm}$ *is a subset of the ith row of* $A \cdot [Y]$*.*

Proof. First note that the $[x_i]^{i,\pm}$ are point intervals, $[x_i]^{i,+} = \sup[x_i]$, and $[x_i]^{i,-} = \inf[x_i]$. Since

$$\begin{aligned} \hat{\mathbf{x}}^{i,\pm} - \hat{\mathbf{x}} &= (0, \dots, 0, [x_i]^{i,\pm} - \hat{x}_i, 0, \dots, 0)^T \qquad \text{and} \\ [\mathbf{x}]^{i,\pm} - \hat{\mathbf{x}}^{i,\pm} &= ([x_1]^{i,\pm} - \hat{x}_1, \dots, [x_{i-1}]^{i,\pm} - \hat{x}_{i-1}, 0, \\ &\qquad [x_{i+1}]^{i,\pm} - \hat{x}_{i+1}, \dots, [x_n]^{i,\pm} - \hat{x}_n)^T, \end{aligned}$$

by the definition of $[Y]$ we have

$$\begin{aligned} g_i(\hat{\mathbf{x}}^{i,\pm}) &\in g_i(\hat{\mathbf{x}}) + (A[Y])_i \cdot (\hat{\mathbf{x}}^{i,\pm} - \hat{\mathbf{x}}) \\ &= g_i(\hat{\mathbf{x}}) + (A[Y])_{ii} \cdot ([x_i]^{i,\pm} - \hat{x}_i), \end{aligned}$$

and therefore, making use of $[\mathbf{y}]^{i,\pm} \subseteq (A \cdot [Y])_i$,

$$\begin{aligned} [m_{\mathrm{F}}]^{i,\pm} &= g_i(\hat{\mathbf{x}}^{i,\pm}) + [\mathbf{y}]^{i,\pm} \cdot ([\mathbf{x}]^{i,\pm} - \hat{\mathbf{x}}^{i,\pm}) \\ &\subseteq g_i(\hat{\mathbf{x}}^{i,\pm}) + \sum_{\ell \neq i} (A[Y])_{i\ell} \cdot ([x_\ell]^{i,\pm} - \hat{x}_\ell) \\ &\subseteq g_i(\hat{\mathbf{x}}) + (A[Y])_{ii} \cdot ([x_i]^{i,\pm} - \hat{x}_i) + \sum_{\ell \neq i} (A[Y])_{i\ell} \cdot ([x_\ell]^{i,\pm} - \hat{x}_\ell) \\ &= g_i(\hat{\mathbf{x}}) + (A[Y])_i \cdot ([\mathbf{x}]^{i,\pm} - \hat{\mathbf{x}}) \\ &= [m_{\mathrm{C}}]^{i,\pm}. \end{aligned}$$

Thus, (1.6b) implies (1.6c). □

Note. (i) The compatibility condition $[\mathbf{y}]^{i,\pm} \subseteq (A \cdot [Y])_i$ for the slopes is not a triviality since $[Y]$ contains slopes relative to the point $\hat{\mathbf{x}}$, whereas the slopes in $[\mathbf{y}]^{i,\pm}$ are relative to the point $\hat{\mathbf{x}}^{i,\pm}$. The condition is, however, certainly fulfilled for two simple choices, $[Y] = \mathbf{f}'([\mathbf{x}])$ and $[\mathbf{y}]^{i,\pm} = (A \cdot [Y])_i$, or $[Y] = \mathbf{f}'([\mathbf{x}])$ and $[\mathbf{y}]^{i,\pm} = (A \cdot \mathbf{f}'([\mathbf{x}]^{i,\pm}))_i$.

(ii) If $\hat{\mathbf{x}}$ is the midpoint of $[\mathbf{x}]$, then the orthogonal projections $\hat{\mathbf{x}}^{i,\pm}$ are the midpoints of the respective facets.

Another fixed point test, which will not be discussed further in the present paper, is based on Kantorovich's theorem. As pointed out in [1], Kantorovich's theorem is weaker than Borsuk's in the sense that the prerequisites of Kantorovich's theorem being fulfilled implies those of Borsuk's theorem to be fulfilled, too. Note that the cited paper ranks different fixed point *theorems* with respect to their strength without

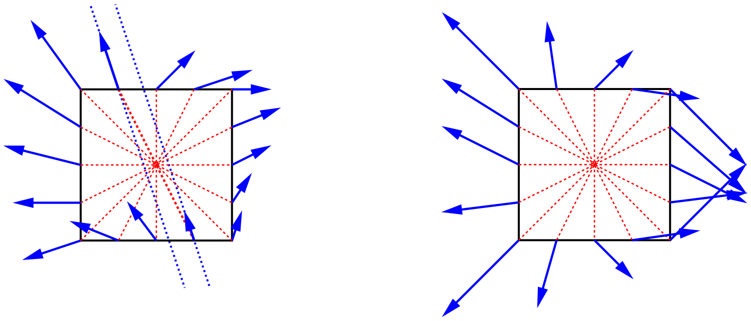


Fig. 2.1. *Illustration of Borsuk's theorem. In the left picture, there is a pair of opposite points on $\partial\Omega$ such that* $\mathbf{f}$ *points in the same direction, i.e., condition* (2.1) *is* not *fulfilled. By contrast, there is no such pair of points in the right picture, and therefore Borsuk's theorem guarantees the existence of a zero in this case.*

addressing the issue of their implementation in computational tests. Concerning such tests, [13] concludes that a computational test based on Kantorovich's theorem is not much stronger, but significantly more costly, than the Moore test.

2. Tests based on Borsuk's theorem.

Let $\Omega \subseteq \mathbb{R}^n$ be an open, bounded, convex set that is symmetric with respect to its center $\mathbf{x}^0$, i.e., $\mathbf{x}^0+\mathbf{y} \in \Omega$ iff $\mathbf{x}^0-\mathbf{y} \in \Omega$. Then Borsuk's theorem [6, 7] can be formulated as follows.

Theorem 2.1. *Assume that* $\mathbf{f} : \overline{\Omega} \to \mathbb{R}^n$ *is continuous and that for all* $\mathbf{x} = \mathbf{x}^0 + \mathbf{y} \in \partial\Omega$*, the topological boundary of* Ω*, we have*

$$\mathbf{f}(\mathbf{x}^0+\mathbf{y}) \neq \lambda \mathbf{f}(\mathbf{x}^0-\mathbf{y}) \text{ for all } \lambda > 0. \tag{2.1}$$

Then $\mathbf{f}$ *has a zero in* $\overline{\Omega}$.

This theorem can in particular be applied in the case when $\overline{\Omega} = [\mathbf{x}]$ is an interval vector; cf. Figure 2.1.

Similarly to the Miranda existence test, we now derive three range-based criteria for (2.1). Later on we will show how these criteria can be checked computationally using interval arithmetic. Note that the boundary of $[\mathbf{x}]$ is given by the facets $[\mathbf{x}]^{i,\pm}$ defined before, and $\mathbf{x}^0+\mathbf{y} \in [\mathbf{x}]^{i,+}$ iff $\mathbf{x}^0-\mathbf{y} \in [\mathbf{x}]^{i,-}$. We again consider the preconditioned function $\mathbf{g}(\mathbf{x}) = A \cdot \mathbf{f}(\mathbf{x})$, $\mathbf{x} \in [\mathbf{x}]$, with a nonsingular matrix A, and for easier exposition we assume that $\mathbf{f}$ (and therefore $\mathbf{g}$) has no zero on $\partial[\mathbf{x}]$, the boundary of $[\mathbf{x}]$.

First we note that, for any vectors $\mathbf{u}, \mathbf{v} \neq \mathbf{0}$, the Cauchy–Schwarz inequality yields

$$-1 \leq \frac{\langle\, \mathbf{u}, \mathbf{v}\, \rangle}{\|\mathbf{u}\| \cdot \|\mathbf{v}\|} \leq +1,$$

the value $+1$ being attained iff $\mathbf{u}$ and $\mathbf{v}$ are collinear and point into the same direction. Thus, condition (2.1) from Borsuk's theorem is equivalent to

$$\sup \mathrm{Rg}(\sigma, [\mathbf{x}]^{i,+}) < 1 \tag{2.2}$$

being fulfilled for each $i \in \{1, \dots, n\}$, where for $\mathbf{x}^0 + \mathbf{y} \in [\mathbf{x}]^{i,+}$ we define

$$\sigma(\mathbf{x}^0 + \mathbf{y}) = \frac{\langle\, \mathbf{g}(\mathbf{x}^0 + \mathbf{y}), \mathbf{g}(\mathbf{x}^0 - \mathbf{y})\, \rangle}{\|\mathbf{g}(\mathbf{x}^0 + \mathbf{y})\| \cdot \|\mathbf{g}(\mathbf{x}^0 - \mathbf{y})\|}. \tag{2.3}$$

So one criterion is (2.2) to be fulfilled for $i = 1, \dots, n$.

To derive the remaining two (sufficient) range-based criteria, we assume that for some $\mathbf{x}^0 + \mathbf{y} \in \partial[\mathbf{x}]$ the function $\mathbf{g}$ does *not* fulfill (2.1). This means that for some $i \in \{1, \dots, n\}$ with $\mathbf{x}^0 + \mathbf{y} \in [\mathbf{x}]^{i,+}$ and for some $\lambda > 0$ we would have

$$g_j(\mathbf{x}^0 + \mathbf{y}) = \lambda g_j(\mathbf{x}^0 - \mathbf{y}), \qquad j = 1, \dots, n. \tag{2.4}$$

This would imply

$$\lambda = \frac{g_j(\mathbf{x}^0 + \mathbf{y})}{g_j(\mathbf{x}^0 - \mathbf{y})} \in \frac{\mathrm{Rg}(g_j, [\mathbf{x}]^{i,+})}{\mathrm{Rg}(g_j, [\mathbf{x}]^{i,-})} \quad \text{for all } j = 1, \dots, n,$$

invoking extended interval arithmetic [15] in the case $0 \in \mathrm{Rg}(g_j, [\mathbf{x}]^{i,-})$. Therefore (2.1) is certainly fulfilled if

$$\bigcap_{j=1}^{n} \frac{\mathrm{Rg}(g_j, [\mathbf{x}]^{i,+})}{\mathrm{Rg}(g_j, [\mathbf{x}]^{i,-})} \cap (0, \infty) = \emptyset \tag{2.5}$$

for each $i \in \{1, \dots, n\}$, which is our second criterion.

The third criterion is also obtained from (2.4). Given a slope matrix $[Y]$ according to (1.1), condition (2.4) would imply

$$g_j(\mathbf{x}^0) + (A \cdot Y_+)_j \cdot \mathbf{y} = \lambda \left(g_j(\mathbf{x}^0) - (A \cdot Y_-)_j \cdot \mathbf{y} \right) \tag{2.6}$$

with matrices $Y_+, Y_- \in [Y]$. Transforming further, (2.6) yields

$$\begin{aligned} (1 - \lambda) \cdot g_j(\mathbf{x}^0) &= -(A \cdot (\lambda Y_- + Y_+))_j \cdot \mathbf{y} \\ &\in -(A \cdot (\lambda[Y] + [Y]))_j \cdot ([\mathbf{x}]^{i,+} - \mathbf{x}^0) \\ &= -\,(\lambda + 1)\,(A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,+} - \mathbf{x}^0), \end{aligned}$$

the last equality holding because $\lambda > 0$. Thus, we finally obtain

$$\frac{\lambda - 1}{\lambda + 1} g_j(\mathbf{x}^0) \in (A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,+} - \mathbf{x}^0). \tag{2.7}$$

This gives us a condition on the range of $(\lambda - 1)/(\lambda + 1)$, because (2.7) implies

$$\frac{\lambda - 1}{\lambda + 1} \in \frac{(A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,+} - \mathbf{x}^0)}{g_j(\mathbf{x}^0)}$$

with the convention that the division of an interval by 0 is defined as

$$\frac{[a]}{0} = \begin{cases} \mathbb{R} & \text{if } 0 \in [a], \\ \emptyset & \text{else.} \end{cases}$$

For $\lambda > 0$, the range of $(\lambda-1)/(\lambda+1)$ is contained in $(-1, 1)$. Putting things together, we see that if (2.4) holds, we have

$$\frac{\lambda - 1}{\lambda + 1} \in \bigcap_{j=1}^{n} \frac{(A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,+} - \mathbf{x}^0)}{g_j(\mathbf{x}^0)} \cap (-1, 1).$$

Since the existence of a zero of $\mathbf{g}$ in $[\mathbf{x}]$ is guaranteed by Borsuk's theorem if (2.4) does *not* hold for $i = 1, \ldots, n$, we obtain the following, third, criterion: If

$$\underbrace{\bigcap_{j=1}^{n} \frac{(A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,+} - \mathbf{x}^0)}{g_j(\mathbf{x}^0)}}_{=:\, [b_{\mathrm{M}}]^i} \cap (-1, 1) = \emptyset \tag{2.8}$$

for each $i \in \{1, \ldots, n\}$, then $\mathbf{g}$ (and thus $\mathbf{f} = A^{-1}\mathbf{g}$) has a zero $\mathbf{x}^*$ in $[\mathbf{x}]$.

At first glance it may be surprising that the opposite facets $[\mathbf{x}]^{i,-}$ do not at all enter criterion (2.8). But since $[\mathbf{x}]^{i,-} - \mathbf{x}^0 = -([\mathbf{x}]^{i,+} - \mathbf{x}^0)$, we have

$$\bigcap_{j=1}^{n} \frac{(A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,-} - \mathbf{x}^0)}{g_j(\mathbf{x}^0)} = -\bigcap_{j=1}^{n} \frac{(A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,+} - \mathbf{x}^0)}{g_j(\mathbf{x}^0)}$$

so that substituting $[\mathbf{x}]^{i,+}$ by $[\mathbf{x}]^{i,-}$ in (2.8) just produces an equivalent condition to (2.8).

We will now discuss several ways for checking the above three criteria computationally with the use of interval arithmetic.

While the formulation (2.8) is directly amenable to interval evaluation, the exact ranges in (2.2) and (2.5) must be replaced with quantities that can be computed effectively.

For the criterion (2.5), the naive, centered, and facet-centered techniques described in the context of the Miranda test lead to the interval quantities

$$[b_{\mathrm{N}}]^i := \bigcap_{j=1}^{n} \frac{g_j([\mathbf{x}]^{i,+})}{g_j([\mathbf{x}]^{i,-})},$$

$$[b_{\mathrm{C}}]^i := \bigcap_{j=1}^{n} \frac{g_j(\hat{\mathbf{x}}) + (A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,+} - \hat{\mathbf{x}})}{g_j(\hat{\mathbf{x}}) + (A \cdot [Y])_j \cdot ([\mathbf{x}]^{i,-} - \hat{\mathbf{x}})},$$

$$[b_{\mathrm{F}}]^i := \bigcap_{j=1}^{n} \frac{g_j(\hat{\mathbf{x}}^{i,+}) + [\mathbf{y}_j]^{i,+} \cdot ([\mathbf{x}]^{i,+} - \hat{\mathbf{x}}^{i,+})}{g_j(\hat{\mathbf{x}}^{i,-}) + [\mathbf{y}_j]^{i,-} \cdot ([\mathbf{x}]^{i,-} - \hat{\mathbf{x}}^{i,-})},$$

where again $[Y]$ is a slope matrix fulfilling (1.1), $\hat{\mathbf{x}}$ is a given point in the box $[\mathbf{x}]$ (not necessarily its midpoint), the $\hat{\mathbf{x}}^{i,\pm}$ are given points on the facets ($\hat{\mathbf{x}}^{i,\pm} \in [\mathbf{x}]^{i,\pm}$), and the $[\mathbf{y}_j]^{i,\pm}$ are row vectors such that $g_j(\mathbf{x}) - g_j(\hat{\mathbf{x}}^{i,\pm}) \in [\mathbf{y}_j]^{i,\pm} \cdot (\mathbf{x} - \hat{\mathbf{x}}^{i,\pm})$ for all $\mathbf{x} \in [\mathbf{x}]^{i,\pm}$.

In (2.2) the exact range $\mathrm{Rg}(\sigma, [\mathbf{x}]^{i,+})$ must be enclosed in an interval $[b_{\mathrm{S}}]^i$. Since a naive interval-arithmetic evaluation of the function σ defined in (2.3) suffers severely from the dependence problem [9, p. 4], higher-order methods should be applied to obtain a reasonably narrow enclosure of the range, e.g., by decomposing the domain into smaller subdomains or by using Taylor models [4, 5].

The computational tests resulting from the criteria (2.8), (2.5), and (2.2) are summarized in the following definition; note that $[b_{\mathrm{M}}]^i$ is defined in (2.8).

Definition 2.2. *The term* Borsuk test *denotes checking, for each* $i \in \{1, \ldots, n\}$, *one of the conditions*

$$[b_{\mathrm{M}}]^i \cap (-1, 1) = \emptyset, \tag{2.9}$$

$$[b_{\mathrm{N}}]^i \cap (0, \infty) = \emptyset, \tag{2.10a}$$

$$[b_{\mathrm{C}}]^i \cap (0, \infty) = \emptyset, \tag{2.10b}$$

$$[b_{\mathrm{F}}]^i \cap (0, \infty) = \emptyset, \tag{2.10c}$$

$$\sup [b_{\mathrm{S}}]^i < 1. \tag{2.11}$$

Again, we do not have to require the same condition to be checked for all i.

With the same arguments as in the proof to Theorem 1.4 we can show that (2.10c) is more powerful than (2.10b), provided that $\hat{\mathbf{x}}^{i,\pm}$ is chosen as the orthogonal projection of $\hat{\mathbf{x}}$ onto the facet $[\mathbf{x}]^{i,\pm}$ and that $[\mathbf{y}_j]^{i,\pm}$ is a subset of the jth row of $A \cdot [Y]$. We do not explicitly formulate this as a theorem here.

We can also show that the Borsuk test compares favorably to the Miranda test.

Theorem 2.3. *Each variant* (2.10a)–(2.10c) *of the Borsuk test is more powerful than the corresponding variant* (1.6a)–(1.6c) *of the Miranda test in the sense that if* (1.6a) *(respectively,* (1.6b), *respectively,* (1.6c)*) is satisfied for some* i, $[\mathbf{x}]$, $\hat{\mathbf{x}}$, $\hat{\mathbf{x}}^{i,\pm}$, A, $[Y]$ *and* $[\mathbf{y}]^{i,\pm}$, *then* (2.10a) *(respectively,* (2.10b), *respectively,* (2.10c)*) also holds true with the same* i, $[\mathbf{x}]$, $\hat{\mathbf{x}}$, $\hat{\mathbf{x}}^{i,\pm}$, A *and* $[Y]$, *provided that the* $[\mathbf{y}_j]^{i,\pm}$ *are chosen such that* $[\mathbf{y}_i]^{i,\pm} \subseteq [\mathbf{y}]^{i,\pm}$ *is fulfilled.*

Proof. We will only consider the naive variant, the arguments for the centered and face-centered cases being the same. If (1.6a) is fulfilled, then

$$\frac{g_i([\mathbf{x}]^{i,+})}{g_i([\mathbf{x}]^{i,-})} \subseteq (-\infty, 0],$$

extended interval arithmetic coming into play if $\sup g_i([\mathbf{x}]^{i,-}) = 0$. This immediately implies

$$\bigcap_{j=1}^{n} \frac{g_j([\mathbf{x}]^{i,+})}{g_j([\mathbf{x}]^{i,-})} \cap (0, \infty) = \emptyset,$$

with the empty intersection being already enforced by the single component $j = i$. Thus (2.10a) also holds true. □

Note. If A is an approximation to the inverse of $\mathbf{f}'(\hat{\mathbf{x}})$, if $\hat{\mathbf{x}}$ is a good approximation to a zero of $\mathbf{f}$, and if the box $[\mathbf{x}]$ is *sufficiently small*, then the function $\mathbf{g}(\mathbf{x}) + \hat{\mathbf{x}}$ is a small perturbation of the identity on $[\mathbf{x}]$, and therefore empty intersections in (2.10a)–(2.10c) will typically be caused by the component $j = i$, i.e., the corresponding condition of the Miranda test would be fulfilled as well. Under these circumstances, the variants (2.10a)–(2.10c) of the Borsuk test are hardly more powerful than the Miranda variants (1.6a)–(1.6c). The situation can be substantially different for larger boxes.

The variants (2.9) and (2.11) are completely different. In particular, (2.11) may be significantly more powerful than any variant of the Miranda test *provided that narrow enclosures for the range of* σ *are available.*

3. An example. In this section we present an example showing that the Borsuk test may be able to verify the existence of a zero in situations where Miranda-based tests as well as the Moore test must fail. Consider the function

$$\mathbf{f}(u,v) = \begin{pmatrix} 4 - 2(u-1)^2 \\ (2-(u+1)^2)\cdot(2-(v-1)^2) \end{pmatrix},$$

which has a unique zero $\mathbf{x}^* = (u^*, v^*) = (1-\sqrt{2}, 1-\sqrt{2})$ in the box $[\mathbf{x}] = [u] \times [v] = [-1,1] \times [-1,1]$.

For future reference, we note that

$$[\mathbf{x}]^{1,\pm} = \{\pm 1\} \times [-1,1], \qquad [\mathbf{x}]^{2,\pm} = [-1,1] \times \{\pm 1\},$$

$$\mathbf{x}^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \mathbf{f}(\mathbf{x}^0) = \begin{pmatrix} 2 \\ 1 \end{pmatrix},$$

$$\mathbf{f}(1,1) = \begin{pmatrix} 4 \\ -4 \end{pmatrix}, \qquad \mathbf{f}(1,-1) = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \tag{3.1}$$

$$\mathbf{f}(-1,1) = \begin{pmatrix} -4 \\ 4 \end{pmatrix}, \qquad \mathbf{f}(-1,-1) = \begin{pmatrix} -4 \\ -4 \end{pmatrix}, \tag{3.2}$$

$$\mathbf{f}'(u,v) = \begin{pmatrix} -4(u-1) & 0 \\ -2(u+1)\left(2-(v-1)^2\right) & -2\left(2-(u+1)^2\right)(v-1) \end{pmatrix}.$$

In this two-dimensional example the facets of the rectangle $[\mathbf{x}]$ are just its sides.

3.1. Miranda-based tests. We first show that *no* test based on Miranda's theorem is applicable to $\mathbf{f}$ and $[\mathbf{x}]$, *even if arbitrary preconditioning is allowed.* To this end consider the function

$$\mathbf{g}(u,v) = \underbrace{\begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}}_{A} \cdot \mathbf{f}(u,v), \quad \alpha, \beta, \gamma, \delta \in \mathbb{R}. \tag{3.3}$$

We will now show that for any nonsingular A, the function g does not satisfy the assumptions of Miranda's theorem. Since $\mathbf{g}$'s property of having—or not having—constant (different) sign on opposite sides of $[\mathbf{x}]$ is invariant to row scaling, we may assume that, at the upper right corner $(1,1)$, each component g_j takes either a specified nonzero value, say 4, or is zero.

Case 1. $g_1(1,1) = g_2(1,1) = 0$. By (3.3) and (3.1), this implies $\mathbf{0} = A \cdot (4,-4)^T$, i.e., A must be singular, which is a contradiction.

Case 2. $g_1(1,1) = g_2(1,1) = 4$. Then, from (3.3) and (3.1) we obtain

$$\begin{aligned} 4\alpha - 4\beta &= 4, \\ 4\gamma - 4\delta &= 4, \end{aligned}$$

or $\beta = \alpha - 1$, $\delta = \gamma - 1$. Combining these relations with (3.3) and (3.2) yields

$$\mathbf{g}(-1,1) = \begin{pmatrix} \alpha & \alpha - 1 \\ \gamma & \gamma - 1 \end{pmatrix} \cdot \begin{pmatrix} -4 \\ 4 \end{pmatrix} = \begin{pmatrix} -4 \\ -4 \end{pmatrix},$$

implying that both g_1 and g_2 change sign on the upper side $[\mathbf{x}]^{2,+}$ of the box.

Case 3. $g_1(1,1) = 4$, $g_2(1,1) = 0$. Then, (3.3) and (3.1) give

$$4\alpha - 4\beta = 4,$$
$$4\gamma - 4\delta = 0,$$

or $\beta = \alpha - 1$, $\delta = \gamma$. As in Case 2, the first of these two equations implies that g_1 changes sign on the upper side $[\mathbf{x}]^{2,+}$. Noting that $\gamma \neq 0$ because A is nonsingular, a short calculation shows that g_2 takes the constant sign $\operatorname{sgn}(\gamma)$ on the upper side but changes sign on the lower side $[\mathbf{x}]^{2,-}$ since, according to (3.3), (3.1), and (3.2),

$$g_2(1,-1) = \begin{pmatrix} \gamma & \gamma \end{pmatrix} \cdot \begin{pmatrix} 4 \\ 4 \end{pmatrix} = 8\gamma,$$
$$g_2(-1,-1) = \begin{pmatrix} \gamma & \gamma \end{pmatrix} \cdot \begin{pmatrix} -4 \\ -4 \end{pmatrix} = -8\gamma.$$

Case 4. $g_1(1,1) = 0$, $g_2(1,1) = 4$. This case is completely symmetric to Case 3, with g_2 changing sign on $[\mathbf{x}]^{2,+}$ and g_1 changing sign on $[\mathbf{x}]^{2,-}$.

To summarize, in all cases none of the components g_1 and g_2 takes (constant) different sign on the opposite sides $[\mathbf{x}]^{2,+}$ and $[\mathbf{x}]^{2,-}$. Thus, no nonsingular preconditioning matrix A leads to the prerequisites of Miranda's theorem being fulfilled. Therefore the Miranda test as well as the Moore test cannot be successful.

3.2. Applying Borsuk's theorem. Now we show that the Borsuk test according to Definition 2.2 is successful for this example, demonstrating that the assumptions for Borsuk's theorem can be checked computationally. Unless explicitly stated otherwise, plain interval arithmetic is used for evaluating expressions over a box. No preconditioning is necessary, i.e., $A = I_2$ and $\mathbf{g} \equiv \mathbf{f}$.

With the simplest choice $[Y] = \mathbf{f}'([\mathbf{x}])$ we obtain $A \cdot [Y] = \mathbf{f}'([\mathbf{x}]) = \mathbf{g}'([\mathbf{x}])$, leading to $(A \cdot [Y])_j = g'_j([\mathbf{x}])$.

Interval evaluation of $\mathbf{g}' \equiv \mathbf{f}'$ on $[\mathbf{x}] = [u] \times [v] = [-1,1] \times [-1,1]$ yields

$$\begin{aligned}\mathbf{g}'([\mathbf{x}]) &= \begin{pmatrix} -4([u]-1) & 0 \\ -2([u]+1)(2-([v]-1)^2) & -2(2-([u]+1)^2)([v]-1) \end{pmatrix} \\ &= \begin{pmatrix} [0,8] & 0 \\ [-8,8] & [-8,8] \end{pmatrix}.\end{aligned}$$

In all our computations we assume that the formula $[a]^2 = [\langle [a] \rangle^2, |[a]|^2]$ is used to compute the range of the square function over some interval $[a]$, where the *magnitude* $|[a]|$ and the *mignitude* $\langle [a] \rangle$ denote the maximum and minimum, respectively, absolute values of elements in $[a]$.

$i = 1$ *(right and left side)*: Evaluating $[b_{\mathrm{M}}]^i$ from (2.8) for $i = 1$ yields

$$\begin{aligned}[b_{\mathrm{M}}]^1 &= \frac{g'_1([\mathbf{x}]) \cdot ([\mathbf{x}]^{1,+} - \mathbf{x}^0)}{g_1(\mathbf{x}^0)} \cap \frac{g'_2([\mathbf{x}]) \cdot ([\mathbf{x}]^{1,+} - \mathbf{x}^0)}{g_2(\mathbf{x}^0)} \\ &= \frac{([0,8], 0) \cdot (1, [-1,1])^T}{2} \cap \frac{([-8,8],[-8,8]) \cdot (1,[-1,1])^T}{1} \\ &= [0,4].\end{aligned}$$

Since $[b_{\mathrm{M}}]^1 \cap (-1,1) \neq \emptyset$, (2.9) could not be verified for $i = 1$. By contrast,

$$\begin{aligned}
[b_{\mathrm{N}}]^1 &= \frac{g_1([\mathbf{x}]^{1,+})}{g_1([\mathbf{x}]^{1,-})} \cap \frac{g_2([\mathbf{x}]^{1,+})}{g_2([\mathbf{x}]^{1,-})} \\
&= \frac{4 - 2(1-1)^2}{4 - 2(-1-1)^2} \cap \frac{(2-(1+1)^2)(2-([-1,1]-1)^2)}{(2-(-1+1)^2)(2-([-1,1]-1)^2)} \\
&= \frac{4}{-4} \cap \frac{[-4,4]}{[-4,4]} \\
&= -1 \cap \mathbb{R} \\
&= -1\,,
\end{aligned}$$

showing (2.10a) for $i = 1$.

$i = 2$ *(upper and lower side)*: Choosing $\hat{\mathbf{x}}$ as the midpoint of $[\mathbf{x}]$, $\hat{\mathbf{x}}^{i,\pm}$ as the midpoint of the facet $[\mathbf{x}]^{i,\pm}$, and the slope vector $[\mathbf{y}_j]^{i,\pm}$ as the jth row of $A \cdot [Y] = \mathbf{g}'([\mathbf{x}])$, short calculations show that

$$\begin{aligned}
[b_{\mathrm{M}}]^2 &= \frac{g_1'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,+} - \mathbf{x}^0)}{g_1(\mathbf{x}^0)} \cap \frac{g_2'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,+} - \mathbf{x}^0)}{g_2(\mathbf{x}^0)} \\
&= \frac{([0,8],0) \cdot ([-1,1],1)^T}{2} \cap \frac{([-8,8],[-8,8]) \cdot ([-1,1],1)^T}{1} \\
&= \frac{[-8,8]}{2} \cap \frac{[-16,16]}{1} \\
&= [-4,4],
\end{aligned}$$

implying $[b_{\mathrm{M}}]^2 \cap (-1,1) \neq \emptyset$,

$$\begin{aligned}
[b_{\mathrm{N}}]^2 &= \frac{g_1([\mathbf{x}]^{2,+})}{g_1([\mathbf{x}]^{2,-})} \cap \frac{g_2([\mathbf{x}]^{2,+})}{g_2([\mathbf{x}]^{2,-})} \\
&= \frac{4 - 2([-1,1]-1)^2}{4 - 2([-1,1]-1)^2} \cap \frac{(2-([-1,1]+1)^2) \cdot (2-(1-1)^2)}{(2-([-1,1]+1)^2) \cdot (2-((-1)-1)^2)} \\
&= \frac{[-4,4]}{[-4,4]} \cap \frac{[-4,4]}{[-4,4]} \\
&= \mathbb{R},
\end{aligned}$$

meaning $[b_{\mathrm{N}}]^2 \cap (0,\infty) \neq \emptyset$, and similarly,

$$\begin{aligned}
[b_{\mathrm{C}}]^2 &= \frac{g_1(\hat{\mathbf{x}}) + g_1'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,+} - \hat{\mathbf{x}})}{g_1(\hat{\mathbf{x}}) + g_1'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,-} - \hat{\mathbf{x}})} \cap \frac{g_2(\hat{\mathbf{x}}) + g_2'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,+} - \hat{\mathbf{x}})}{g_2(\hat{\mathbf{x}}) + g_2'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,-} - \hat{\mathbf{x}})} \\
&= \frac{2 + ([0,8],0) \cdot ([-1,1],1)^T}{2 + ([0,8],0) \cdot ([-1,1],-1)^T} \cap \frac{1 + ([-8,8],[-8,8]) \cdot ([-1,1],1)}{1 + ([-8,8],[-8,8]) \cdot ([-1,1],-1)} \\
&= \frac{[-6,10]}{[-6,10]} \cap \frac{[-15,17]}{[-15,17]} \\
&= \mathbb{R}
\end{aligned}$$

and

$$\begin{aligned}[b_{\mathrm{F}}]^2 &= \frac{g_1(\mathrm{mid}[\mathbf{x}]^{2,+}) + g_1'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,+} - \mathrm{mid}[\mathbf{x}]^{2,+})}{g_1(\mathrm{mid}[\mathbf{x}]^{2,-}) + g_1'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,-} - \mathrm{mid}[\mathbf{x}]^{2,-})} \\ &\qquad \cap \frac{g_2(\mathrm{mid}[\mathbf{x}]^{2,+}) + g_2'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,+} - \mathrm{mid}[\mathbf{x}]^{2,+})}{g_2(\mathrm{mid}[\mathbf{x}]^{2,-}) + g_2'([\mathbf{x}]) \cdot ([\mathbf{x}]^{2,-} - \mathrm{mid}[\mathbf{x}]^{2,-})} \\ &= \frac{g_1(0,1) + ([0,8],0) \cdot ([-1,1],0)^T}{g_1(0,-1) + ([0,8],0) \cdot ([-1,1],0)^T} \\ &\qquad \cap \frac{g_2(0,1) + ([-8,8],[-8,8]) \cdot ([-1,1],0)^T}{g_2(0,-1) + ([-8,8],[-8,8]) \cdot ([-1,1],0)^T} \\ &= \frac{[-6,10]}{[-6,10]} \cap \frac{[-6,10]}{[-10,6]} \\ &= \mathbb{R}.\end{aligned}$$

So none of the conditions (2.9) and (2.10a)–(2.10c) is fulfilled for $i = 2$. To check the remaining condition (2.11), we enclose the range of the function σ over $[\mathbf{x}]^{2,+}$. Note that, in order to have $\mathbf{x}^0 + \mathbf{y} \in [\mathbf{x}]^{2,+}$, the offset $\mathbf{y}$ must be of the form $\mathbf{y} = (u,1)^T$ with $u \in [-1,1]$. Therefore $\mathbf{x}^0 + \mathbf{y} = (u,1)^T$ and $\mathbf{x}^0 - \mathbf{y} = (-u,-1)^T$. This gives

$$\begin{aligned}\sigma(\mathbf{x}^0 + \mathbf{y}) &= \frac{\langle\, \mathbf{g}(u,1), \mathbf{g}(-u,-1)\,\rangle}{\|\mathbf{g}(u,1)\| \cdot \|\mathbf{g}(-u,-1)\|} \\ &= \frac{\left\langle \begin{pmatrix} 4-2(u-1)^2 \\ (2-(u+1)^2)\cdot 2 \end{pmatrix}, \begin{pmatrix} 4-2(-u-1)^2 \\ (2-(-u+1)^2)\cdot(-2) \end{pmatrix} \right\rangle}{\left\| \begin{pmatrix} 4-2(u-1)^2 \\ (2-(u+1)^2)\cdot 2 \end{pmatrix} \right\| \cdot \left\| \begin{pmatrix} 4-2(-u-1)^2 \\ (2-(-u+1)^2)\cdot(-2) \end{pmatrix} \right\|} \\ &= \frac{\nu}{\delta}\end{aligned}$$

with

$$\begin{aligned}\nu &= (4-2(u-1)^2)\cdot(4-2(-u-1)^2) \\ &\qquad - 4(2-(u+1)^2)\cdot(2-(-u+1)^2)\ , \\ \delta &= \sqrt{(4-2(u-1)^2)^2 + 4(2-(u+1)^2)^2} \\ &\qquad \cdot \sqrt{(4-2(-u-1)^2)^2 + 4(2-(-u+1)^2)^2}\ .\end{aligned}$$

Using these expressions for the numerator and denumerator, respectively, and plain interval arithmetic for their evaluation over $[u] = [-1,1]$, again causes too much overestimation:

$$\frac{[\nu]}{[\delta]} = \frac{[-32,32]}{\sqrt{[0,32]}\cdot\sqrt{[0,32]}} \ = \ \mathbb{R}.$$

Other ways for bounding the range, however, yield the required result:

- A symbolic term optimizer detects that $\nu \equiv 0$ and therefore replaces the term ν/δ representing σ with 0. Evaluating this term immediately yields

$$\sup_{\mathbf{x}^0+\mathbf{y}\in[\mathbf{x}]^{2,+}} \sigma(\mathbf{y}) = 0 < 1.$$

  This extreme situation is, of course, not very likely to occur in practice.

- Taylor arithmetic [5] can be used to reduce the heavy dependence of the variables within the term ν/δ. Indeed, in our example an order-4 Taylor model—implemented, e.g., using the COSY INFINITY software [4]—shows that the range of σ is contained in the interval $[b_{\mathrm{S}}]^2 = [-10^{-12}, +10^{-12}]$, which is bounded away from $+1$.
- Subdividing the domain for $\mathbf{y}$ is another way to reduce the overestimation. In our example, dividing $[u]$ into 8 equally sized subintervals and using naive interval evaluation of the expression ν/δ over these subintervals shows that the range of σ is a subset of $[b_{\mathrm{S}}]^2 = [-0.8394, +0.8394]$, which is also bounded below 1.

Since (2.10a) holds for $i = 1$ and (2.11) is fulfilled for $i = 2$, Borsuk's theorem guarantees the existence of a zero $\mathbf{x}^*$ of $\mathbf{g}$ (and $\mathbf{f}$) in the (rather large) box $[\mathbf{x}]$.

4. Concluding remarks. We have presented a new method for *automatically* verifying the existence of a zero of a nonlinear function $\mathbf{f}$ within some interval vector $[\mathbf{x}]$. Our test is based on a theorem by Borsuk and can be implemented in several different ways. For some of these, similar criteria based on a theorem by Miranda are known. We could show that in these cases our test is at least as powerful as the corresponding Miranda test.

The Borsuk test also allows for formulations that have no Miranda analogue and may be applicable in situations where no Miranda-based test can be successful. This fact has been illustrated by a numerical example.

From a complexity point of view, it is important that the Borsuk test can be implemented to complement the Miranda test in the following way: If we first apply the Miranda test and one of the conditions (1.6a)–(1.6c) can be verified for some i then, by Theorem 2.3, we know that the respective condition (2.10a)–(2.10c) is fulfilled as well. Thus, the (more expensive) Borsuk test must be applied only to those components i for which the Miranda test has failed.

As Borsuk's theorem only requires that the function does not point in the same direction at opposite boundary points, *different preconditioning* $\mathbf{g}^{(i)} = A_i \cdot \mathbf{f}$ might be used on each pair $[\mathbf{x}]^{i,\pm}$ of opposite facets. The potential benefits of this additional flexibility, which is not available in Miranda-based tests, will be explored in the future.

REFERENCES

[1] G. Alefeld, A. Frommer, G. Heindl, and J. Mayer, *On the existence theorems of Kantorovich, Miranda and Borsuk*, Electron. Trans. Numer. Anal., 17 (2004), pp. 102–111.

[2] G. Alefeld and J. Herzberger, *Introduction to Interval Computation*, Academic Press, New York, 1983.

[3] G. Alefeld, F. Potra, and Z. Shen, *On the existence theorems of Kantorovich, Moore and Miranda*, Computing Suppl., 15 (2001), pp. 21–28.

[4] M. Berz, *COSY INFINITY Version* 8.1 *Programming Manual*, Tech. Rep. MSUHEP-20703, National Superconductor Cyclotron Laboratory, Michigan State University, East Lansing, MI, 2002.

[5] M. Berz and G. Hofstätter, *Computation and application of Taylor polynomials with interval remainder bounds*, Reliable Comput., 4 (1998), pp. 83–97.

[6] K. Borsuk, *Drei Sätze über die n-dimensionale Sphäre*, Fund. Math., 20 (1933), pp. 177–190.

[7] K. Deimling, *Nonlinear Functional Analysis*, Springer, Berlin, Heidelberg, New York, 1985.

[8] A. Frommer, B. Lang, and M. Schnurr, *A comparison of the Moore and Miranda existence tests*, Computing, 72 (2004), pp. 349–354.

[9] R. B. Kearfott, *Rigorous Global Search: Continuous Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1996.

[10] C. Miranda, *Un' osservazione su un theorema di Brouwer*, Bolletino Unione Mathematica Italiana, 3 (1940), pp. 5–7.

[11] R. E. Moore, *A test for existence of solutions to nonlinear systems*, SIAM J. Numer. Anal., 14 (1972), pp. 611–615.

[12] R. E. Moore and J. B. Kioustelidis, *A simple test for accuracy of approximate solutions to nonlinear (or linear) systems*, SIAM J. Numer. Anal., 17 (1980), pp. 521–529.

[13] L. B. Rall, *A comparison of the existence theorems of Kantorovich and Moore*, SIAM J. Numer. Anal., 17 (1980), pp. 148–161.

[14] M. Schnurr, *On the proofs of some statements concerning the theorems of Kantorovich, Moore and Miranda*, Reliab. Comput., 11 (2005), pp. 77–85.

[15] W. Walster, *The Extended Real Interval System*, http://www.mscs.mu.edu/∼globsol/walster-papers.html, 1998.

SIAM J. Numer. Anal.
Vol. 43, No. 4, pp. 1363–1384

GALERKIN FINITE ELEMENT METHODS FOR STOCHASTIC PARABOLIC PARTIAL DIFFERENTIAL EQUATIONS*

YUBIN YAN[†]

Abstract. We study the finite element method for stochastic parabolic partial differential equations driven by nuclear or space-time white noise in the multidimensional case. The discretization with respect to space is done by piecewise linear finite elements, and in time we apply the backward Euler method. The noise is approximated by using the generalized L_2-projection operator. Optimal strong convergence error estimates in the L_2 and $\dot{H}^{-1}$ norms with respect to the spatial variable are obtained. The proof is based on appropriate nonsmooth data error estimates for the corresponding deterministic parabolic problem. The computational analysis and numerical example are given.

Key words. stochastic parabolic partial differential equations, finite element method, additive noise

AMS subject classifications. 60H15, 65C30, 65M65

DOI. 10.1137/040605278

1. Introduction.

We study the finite element approximation of the stochastic parabolic partial differential equation

$$du + Au\,dt = \sigma(u)dW \quad \text{for } 0 < t \le T, \quad \text{with } u(0) = u_0, \tag{1.1}$$

in a Hilbert space H, with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$, where $u(t)$ is an H-valued random process; A is a linear, self-adjoint, positive definite, not necessarily bounded operator with a compact inverse, densely defined in $\mathcal{D}(A) \subset H$; and σ is a nonlinear operator-valued function defined on H which we will specify later. Here $W(t)$ is a Wiener process defined on a filtered probability space $(\Omega, \mathcal{F}, \mathbf{P}, \{\mathcal{F}_t\}_{t\ge 0})$ and $u_0 \in H$.

For the sake of simplicity, we shall concentrate on the case $A = -\Delta$, where Δ stands for the Laplacian operator subject to homogeneous Dirichlet boundary conditions, and $H = L_2(\mathcal{D})$, where $\mathcal{D}$ is a bounded convex domain in $\mathbf{R}^d$, $d = 1, 2, 3$, with a sufficiently smooth boundary $\partial\mathcal{D}$.

Many models in physics, chemistry, biology, population dynamics, neurophysiology, etc., are described by stochastic partial differential equations; see Da Prato and Zabczyk [8], Walsh [29], etc. The existence, uniqueness, and properties of the solutions of such equations have been well studied; see Curtain and Falb [4], [5], Da Prato [6], Da Prato and Lunardi [7], Da Prato and Zabczyk [8], Dawson [10], Gozzi [12], Peszat and Zabczyk [23], Walsh [29], etc. However, numerical approximation of such equations has not been studied thoroughly.

We assume that $W(t)$ is a Wiener process with covariance operator Q. This process may be considered in terms of its Fourier series. Suppose that Q is a bounded, linear, self-adjoint, positive definite operator on H, with eigenvalues $\gamma_l > 0$ and corresponding eigenfunctions e_l. Let β_l, $l = 1, 2, \ldots,$ be a sequence of real-valued

[*] Received by the editors March 16, 2004; accepted for publication (in revised form) November 26, 2004; published electronically October 19, 2005.
http://www.siam.org/journals/sinum/43-4/60527.html

[†] Department of Automatic Control and Systems Engineering, The University of Sheffield, Mappin Street, Sheffield, S1 3JD, UK (Y.Yan@sheffield.ac.uk).

independently and identically distributed Brownian motions. Then

$$W(t) = \sum_{l=1}^{\infty} \gamma_l^{1/2} e_l \beta_l(t)$$

is a Wiener process with covariance operator Q.

If Q is in trace class, then $W(t)$ is an H-valued process. If Q is not in trace class, for example, $Q = I$, then $W(t)$ does not belong to H, in which case $W(t)$ is called a cylindrical Wiener process.

Let $L_2^0 = HS(Q^{1/2}(H), H)$ denote the space of Hilbert–Schmidt operators from $Q^{1/2}(H)$ to H, i.e.,

$$L_2^0 = \left\{ \psi \in L(H) : \sum_{l=1}^{\infty} \|\psi Q^{1/2} e_l\|^2 < \infty \right\},$$

with norm $\|\psi\|_{L_2^0} = (\sum_{l=1}^{\infty} \|\psi Q^{1/2} e_l\|^2)^{1/2}$, where $L(H)$ is the space of bounded linear operators from H to H.

Let $\mathbf{E}$ denote the expectation. Let $\psi \in L_2^0$. Then $\int_0^t \psi(s)\, dW(s)$ can be defined, and the following isometry property holds:

$$\mathbf{E}\left\| \int_0^t \psi(s)\, dW(s) \right\|^2 = \int_0^t \|\mathbf{E}\psi(s)\|_{L_2^0}^2\, ds. \tag{1.2}$$

Following Da Prato and Zabczyk [8, Chapter 7], we assume that $\sigma : H \to L_2^0$ satisfies the following global Lipschitz and growth conditions:

(i) $\|\sigma(x) - \sigma(y)\|_{L_2^0} \leq C\|x - y\| \quad \forall x, y \in H$,

(ii) $\|\sigma(x)\|_{L_2^0} \leq C\|x\| \quad \forall x \in H$.

Then (1.1) admits a unique mild solution which has the form

$$u(t) = E(t)u_0 + \int_0^t E(t-s)\sigma(u(s))\, dW(s), \tag{1.3}$$

where $E(t) = e^{-tA}$ is the analytic semigroup generated by $-A$. Moreover,

$$\sup_{t \in [0,T]} \mathbf{E}\|u(t)\|^2 \leq C(1 + \mathbf{E}\|u_0\|^2).$$

Note that if $\mathrm{Tr}(Q) = \infty$, then the identity mapping $\sigma(u) = I$ does not satisfy the condition (ii). In order to cover this important case, we introduce a modified version of (ii), i.e.,

(ii′) $\|A^{(\beta-1)/2}\sigma(x)\|_{L_2^0} \leq C\|x\|$ for some $\beta \in [0,1]$, $\forall x \in H$.

We see that (ii) is the special case $\beta = 1$ of (ii′). If $\sigma(\cdot) = I$, the condition (ii′) reduces to $\|A^{(\beta-1)/2}\|_{L_2^0} \leq C$.

The numerical approximation for (1.1) started with the work by Greksch and Kloeden [13] and Gyöngy and Nualart [16]. Further contributions include Allen, Novosel, and Zhang [1], Benth and Gjerde [2], Davie and Gaines [9], Du and Zhang [11], Gyöngy [14], [15], Hausenblas [17], [18], Kloeden and Shott [19], Lord and Rougemont [20], Printems [24], Shardlow [25], Theting [26], [27], Müller-Gronbach and Ritter [21], and Yan [31], [30].

In this paper we will consider error estimates for approximations of (1.1) based on the finite element method in space and the backward Euler method in time. The noise will be approximated by using generalized the L_2-projection operator defined below.

Let S_h be a family of finite element spaces, where S_h consists of continuous piecewise polynomials of degree ≤ 1 with respect to the triangulation $\mathcal{T}_h$ of Ω. For simplicity, we always assume that $\{S_h\} \subset H_0^1 = H_0^1(\mathcal{D}) = \{v \in L_2(\mathcal{D}),\ \nabla v \in L_2(\mathcal{D}),\ v|_{\partial\mathcal{D}} = 0\}$.

To introduce the finite element formulation of (1.1), we will use the generalized L_2-projection operator $P_h : \dot{H}^{-1} \to S_h$ defined by (see Chrysafinos and Hou [3])

$$(P_h v, \chi) = \langle v, \chi \rangle \quad \forall \chi \in S_h \subset \dot{H}^1,\ \forall v \in \dot{H}^{-1}, \tag{1.4}$$

where $\langle \cdot, \cdot \rangle$ denotes the pairing between $\dot{H}^{-1}$ and $\dot{H}^1$. One can easily show that P_h is well defined by introducing a basis $\{\varphi_i\}_{i=1}^{N_h}$ and solving for $P_h v = \sum_{j=1}^{N_h} \alpha_j \varphi_j$ from the equations $(\sum_{j=1}^{N_h} \alpha_j \varphi_j, \varphi_i) = \langle v, \varphi_i \rangle$. Also it is evident that when $v \in L_2(\mathcal{D})$, $P_h v$ is the standard L_2-projection operator; see Thomée [28].

The semidiscrete problem corresponding to (1.1) is to find the process $u_h(t) = u_h(\cdot, t) \in S_h$ such that

$$du_h + A_h u_h \, dt = P_h \sigma(u_h) \, dW \quad \text{for } 0 < t \leq T, \quad \text{with } u_h(0) = P_h u_0, \tag{1.5}$$

where $A_h : S_h \to S_h$ is the discrete analogue of A, defined by

$$(A_h \psi, \chi) = A(\psi, \chi) \quad \forall \psi, \chi \in S_h. \tag{1.6}$$

Here $A(\cdot, \cdot)$ is the bilinear form obtained from the operator A.

Let $E_h(t) = e^{-tA_h}$, $t \geq 0$. Then (1.5) admits a unique mild solution

$$u_h(t) = E_h(t) P_h u_0 + \int_0^t E_h(t-s) P_h \sigma(u_h(s)) \, dW(s).$$

Let $\dot{H}^s = \dot{H}^s(\mathcal{D}) = \mathcal{D}(A^{s/2})$ with norm $|v|_s = \|A^{s/2} v\|$ for any $s \in \mathbf{R}$. For any Hilbert space H, we define

$$L_2(\Omega; H) = \left\{ v : \mathbf{E}\|v\|_H^2 = \int_\Omega \|v(\omega)\|_H^2 \, d\mathbf{P}(\omega) < \infty \right\},$$

with norm $\|v\|_{L_2(\Omega;H)} = (\mathbf{E}\|v\|_H^2)^{1/2}$.

Let k be a time step and $t_n = nk$ with $n \geq 1$. We define the backward Euler method

$$\frac{U^n - U^{n-1}}{k} + A_h U^n = \frac{1}{k} \int_{t_{n-1}}^{t_n} P_h \sigma(U^{n-1}) \, dW(s), \quad n \geq 1, \quad U^0 = P_h u_0. \tag{1.7}$$

With $r(\lambda) = 1/(1-\lambda)$, we can rewrite (1.7) in the form

$$\begin{aligned} U^n &= r(kA_h) U^{n-1} + \int_{t_{n-1}}^{t_n} r(kA_h) P_h \sigma(U^{n-1}) \, dW(s), \quad n \geq 1, \\ U^0 &= P_h u_0. \end{aligned} \tag{1.8}$$

Our main results in this paper are the following theorems.

Theorem 1.1. *Let U^n and $u(t_n)$ be the solutions of* (1.8) *and* (1.1)*, respectively. Assume that σ satisfies* (i) *and* (ii′)*. Assume that $u_0 \in L_2(\Omega; \dot{H}^\beta)$, $0 \leq \beta \leq 1$. Then there exists a constant $C = C(T)$ such that, for $t_n \in [0,T]$ and $0 \leq \gamma < \beta \leq 1$,*

$$\|U^n - u(t_n)\|_{L_2(\Omega;H)} \leq C(k^{\gamma/2} + h^\beta)\Big(\|u_0\|_{L_2(\Omega;\dot{H}^\beta)} + \sup_{0\leq s\leq T}\|u(s)\|_{L_2(\Omega;H)}\Big). \tag{1.9}$$

In particular, if σ satisfies (i) *and* (ii)*, then we have, for $u_0 \in L_2(\Omega;\dot{H}^1)$ and $0 \leq \gamma < 1$,*

$$\|U^n - u(t_n)\|_{L_2(\Omega;H)} \leq C(k^{\gamma/2} + h)\Big(\|u_0\|_{L_2(\Omega;\dot{H}^1)} + \sup_{0\leq s\leq T}\|u(s)\|_{L_2(\Omega;H)}\Big). \tag{1.10}$$

When $\sigma(\cdot) = I$, we have the following error estimates.

Theorem 1.2. *Let U^n and $u(t_n)$ be the solutions of* (1.8) *and* (1.1)*, respectively. Assume that $\sigma(\cdot) = I$. Further assume that $u_0 \in L_2(\Omega;\dot{H}^\beta)$, $0 \leq \beta \leq 1$. If $\|A^{(\beta-1)/2}\|_{L_2^0} < \infty$ for some $\beta \in [0,1]$, then we have, for $l = 0,1$ with $\ell_k = \log(T/k)$, where $T = t_n$,*

$$\|U^n - u(t_n)\|_{L_2(\Omega;\dot{H}^{-l})} \leq C(k^{(\beta+l)/2} + h^{\beta+l})\Big(\|u_0\|_{L_2(\Omega;\dot{H}^\beta)} + \ell_k^l\|A^{(\beta-1)/2}\|_{L_2^0}\Big). \tag{1.11}$$

In particular, if $W(t)$ is an H-valued Wiener process with $\mathrm{Tr}(Q) < \infty$*, then we have, for $u_0 \in L_2(\Omega;\dot{H}^1)$,*

$$\|U^n - u(t_n)\|_{L_2(\Omega;\dot{H}^{-l})} \leq C(k^{(1+l)/2} + h^{1+l})\Big(\|u_0\|_{L_2(\Omega;\dot{H}^1)} + \ell_k^l\,\mathrm{Tr}(Q)^{1/2}\Big). \tag{1.12}$$

This paper is organized as follows. In section 2, we give some regularity results for the mild solution of (1.1) and some error estimates of the corresponding deterministic problem. In section 3, we prove main theorems. In section 4, we consider how to compute the approximate solution U^n numerically in the additive noise case. Finally, in section 5, we give the numerical simulations.

2. Regularity of (1.1) and error estimates for the deterministic problem.

In this section, we give some results in order to prove the main theorems.

2.1. Regularity of the mild solution of (1.1).

In this section we will consider the regularity of the mild solution of (1.1). We have the following theorem.

Theorem 2.1. *Assume that σ satisfies* (i) *and* (ii′)*. Let $u(t)$ be the mild solution* (1.3) *of* (1.1)*. Then we have, for $u_0 \in L_2(\Omega;\dot{H}^\beta)$,*

$$\|u(t)\|_{L_2(\Omega;\dot{H}^\beta)} \leq C\Big(\|u_0\|_{L_2(\Omega;\dot{H}^\beta)} + \sup_{0\leq s\leq t}\|u(s)\|_{L_2(\Omega;H)}\Big). \tag{2.1}$$

In particular, if σ satisfies (i) *and* (ii)*, then we have, for $u_0 \in L_2(\Omega;\dot{H}^1)$,*

$$\|u(t)\|_{L_2(\Omega;\dot{H}^1)} \leq C\Big(\|u_0\|_{L_2(\Omega;\dot{H}^1)} + \sup_{0\leq s\leq t}\|u(s)\|_{L_2(\Omega;H)}\Big). \tag{2.2}$$

To prove this theorem, we need some regularity results which are related to the fact that $E(t) = e^{-tA}$ is an analytic semigroup on H. For later use, we collect some results in the next two lemmas; see Thomée [28] or Pazy [22].

LEMMA 2.2. *For any $\mu, \nu \in \mathbf{R}$ and $l \geq 0$, there is a $C > 0$ such that*

$$|D_t^l E(t)v|_\nu \leq Ct^{-(\nu-\mu)/2-l}|v|_\mu \quad \text{for } t > 0, \quad 2l + \nu \geq \mu, \tag{2.3}$$

and

$$\int_0^t s^\mu |D_t^l E(s)v|_\nu^2 \, ds \leq C|v|_{2l+\nu-\mu-1}^2 \quad \text{for } t \geq 0, \quad \mu \geq 0. \tag{2.4}$$

LEMMA 2.3. *For any $\mu \geq 0$, $0 \leq \nu \leq 1$, there is a $C > 0$ such that*

$$\|A^\mu E(t)\| \leq Ct^{-\mu} \quad \text{for } t > 0, \tag{2.5}$$

and

$$\|A^{-\nu}(I - E(t))\| \leq Ct^\nu \quad \text{for } t \geq 0. \tag{2.6}$$

Proof. Recall that the mild solution has the form

$$u(t) = E(t)u_0 + \int_0^t E(t-s)\sigma(u(s)) \, dW(s).$$

Thus, for any $\beta \geq 0$, using the stability of $E(t)$ and the isometry (1.2),

$$\begin{aligned}
\mathbf{E}|u(t)|_\beta^2 &\leq 2\mathbf{E}|E(t)u_0|_\beta^2 + 2\mathbf{E}\left\| \int_0^t A^{\beta/2}E(t-s)\,\sigma(u(s))dW(s) \right\|^2 \\
&\leq 2\mathbf{E}|u_0|_\beta^2 + 2\mathbf{E}\int_0^t \|A^{\beta/2}E(t-s)\sigma(u(s))\|_{L_2^0}^2 \, ds \\
&= 2\mathbf{E}|u_0|_\beta^2 + 2\mathbf{E}\int_0^t \|A^{1/2}E(t-s)A^{(\beta-1)/2}\sigma(u(s))\|_{L_2^0}^2 \, ds.
\end{aligned}$$

By (ii′) and Lemma 2.2, we have

$$\begin{aligned}
&\mathbf{E}\int_0^t \|A^{1/2}E(t-s)A^{(\beta-1)/2}\sigma(u(s))\|_{L_2^0}^2 \, ds \\
&\quad \leq \left(\int_0^t \|A^{1/2}E(t-s)\|^2 \, ds \right) \sup_{0 \leq s \leq t} \mathbf{E}\|u(s)\|^2 \leq C \sup_{0 \leq s \leq t} \mathbf{E}\|u(s)\|^2.
\end{aligned}$$

Thus we get

$$\mathbf{E}|u(t)|_\beta^2 \leq C\left(\mathbf{E}|u_0|_\beta^2 + \sup_{0 \leq s \leq t} \mathbf{E}\|u(s)\|^2 \right),$$

which implies (2.1) by noting that

$$\left(\sup_{0 \leq s \leq t} \mathbf{E}\|u(s)\|^2 \right)^{1/2} \leq \sup_{0 \leq s \leq t} \left(\mathbf{E}\|u(s)\|^2 \right)^{1/2} = \sup_{0 \leq s \leq t} \|u(s)\|_{L_2(\Omega;H)}.$$

In particular, if (ii) holds, then $\beta = 1$, and we get (2.2). □

We remark that in Theorem 2.1, if $\sigma(\cdot) = I$, the condition (ii′) reduces to the condition $\|A^{(\beta-1)/2}\|_{L_2^0} \leq C$.

In the case of $\sigma(\cdot) = I$, by the proof of Theorem 2.1, we obtain the following.

Corollary 2.4. *Let $u(t)$ be the mild solution* (1.3) *of* (1.1). *Assume that $\sigma(\cdot) = I$. If $\|A^{(\beta-1)/2}\|_{L_2^0} < \infty$ for some $\beta \in [0,1]$, then we have, for fixed $t \in [0,T]$,*

$$\|u(t)\|_{L_2(\Omega;\dot{H}^\beta)} \le C\Big(\|u_0\|_{L_2(\Omega;\dot{H}^\beta)} + \|A^{(\beta-1)/2}\|_{L_2^0}\Big) \quad \text{for } u_0 \in L_2(\Omega;\dot{H}^\beta). \tag{2.7}$$

In particular, if $W(t)$ is an H-valued Wiener process with covariance operator Q, $\mathrm{Tr}(Q) < \infty$, *then we have*

$$\|u(t)\|_{L_2(\Omega;\dot{H}^1)} \le C\Big(\|u_0\|_{L_2(\Omega;\dot{H}^1)} + \mathrm{Tr}(Q)^{1/2}\Big) \quad \text{for } u_0 \in L_2(\Omega;\dot{H}^1). \tag{2.8}$$

If $\sigma(\cdot) = I$ and $d = 1$, then we may specialize to $Q = I$.

Corollary 2.5. *Assume that $\sigma(\cdot) = I$. Let $u(t)$ be the solution of* (1.1) *and $A = -\frac{\partial^2}{\partial x^2}$ with $\mathcal{D}(A) = H_0^1(0,1) \cap H^2(0,1)$. Assume that $W(t)$ is a cylindrical Wiener process with $Q = I$. Then we have, for every $\beta \in [0,1/2)$,*

$$\|u(t)\|_{L_2(\Omega;\dot{H}^\beta)} \le C(1 + \|u_0\|_{L_2(\Omega;\dot{H}^\beta)}) \quad \text{for } u_0 \in L_2(\Omega;\dot{H}^\beta).$$

Proof. By (2.7), it suffices to check in what case $\|A^{(\beta-1)/2}\|_{L_2^0} < \infty$. It is well known that A has eigenvalues $\lambda_j = j^2\pi^2$, $j = 1, 2, \dots$, and corresponding eigenfunctions $\varphi_j = \sqrt{2}\sin j\pi x$, $j = 1, 2, \dots$, which form an orthonormal basis in $H = L_2(0,1)$. Thus, we have

$$\|A^{(\beta-1)/2}\|_{L_2^0}^2 = \sum_{j=1}^\infty \|A^{(\beta-1)/2}\varphi_j\|^2 = \sum_{j=1}^\infty \lambda_j^{\beta-1},$$

which is convergent if $\beta \in [0,1/2)$. The proof is complete. □

We note that in Corollary 2.4 we require the condition $\|A^{(\beta-1)/2}\|_{L_2^0} < \infty$ for $\beta \in [0,1]$. The following lemma shows that this condition is equivalent to having that $W(t)$ is $\dot{H}^{\beta-1}$-valued. In particular, $W(t) \in \dot{H}^{-1}$, which is important when applying the finite element method.

Lemma 2.6. *Assume that $W(t)$ is a Wiener process with covariance operator Q. Assume that A and Q have the same eigenvectors. Then the following statements hold.*

(i) *If $\|A^{(\beta-1)/2}\|_{L_2^0} < \infty$ for some $\beta \in [0,1]$, then*

$$W(t) = \sum_{l=1}^\infty Q^{1/2} e_l \beta_l(t), \quad t \ge 0,$$

defines an $\dot{H}^{\beta-1}$-valued Wiener process with covariance operator $\tilde{Q}$, $\mathrm{Tr}(\tilde{Q}) < \infty$. *In particular, $\tilde{Q} = Q$ if* $\mathrm{Tr}(Q) < \infty$.

(ii) *If $W(t) = \sum_{l=1}^\infty Q^{1/2} e_l \beta_l(t)$, $t \ge 0$, is an $\dot{H}^{\beta-1}$-valued Wiener process with the covariance operator $\tilde{Q}$,* $\mathrm{Tr}(\tilde{Q}) < \infty$, *then*

$$\|A^{(\beta-1)/2}\|_{L_2^0} < \infty \quad \text{for some } \beta \in [0,1].$$

Proof. We first prove (i). With $\{\gamma_l, e_l\}_{l=1}^\infty$ the eigensystem of Q in H, it is easy to show that $g_l = Q^{1/2} e_l = \gamma_l^{1/2} e_l$ is an orthogonal basis of $Q^{1/2}(H)$. In fact,

$$(g_l, g_k)_{Q^{1/2}(H)} = (Q^{-1/2} g_l, Q^{1/2} g_k) = (e_l, e_k) = \delta_{l,k}.$$

Note that

$$\sum_{l=1}^{\infty} |g_l|_{\beta-1}^2 = \sum_{l=1}^{\infty} \|A^{(\beta-1)/2} Q^{1/2} e_l\|^2 = \|A^{(\beta-1)/2}\|_{L_2^0} < \infty,$$

which means that the embedding of $Q^{1/2}(H)$ into $\dot{H}^{\beta-1}$ is Hilbert–Schmidt. By Lemma 4.11 in Da Prato and Zabczyk [8], $W(t)$ defines an $\dot{H}^{\beta-1}$-valued Wiener process with covariance operator $\tilde{Q}$, $\mathrm{Tr}(\tilde{Q}) < \infty$. It is obvious that $\tilde{Q} = Q$ if $\mathrm{Tr}(Q) < \infty$.

We now turn to (ii). Since $W(t) = \sum_{l=1}^{\infty} Q^{1/2} e_l \beta_l(t)$, $t \geq 0$, is an $\dot{H}^{\beta-1}$-valued Wiener process with the covariance operator $\tilde{Q}$, $\mathrm{Tr}(\tilde{Q}) < \infty$, we have

$$\mathbf{E}|W(t)|_{\beta-1}^2 < \infty.$$

With $\{\lambda_l, e_l\}_{l=1}^{\infty}$ the eigensystem of A in H, we have

$$\mathbf{E}|W(t)|_{\beta-1}^2 = \mathbf{E}\left|\sum_{l=1}^{\infty} Q^{1/2} e_l \beta_l(t)\right|_{\beta-1}^2 = \mathbf{E}\sum_{l=1}^{\infty} \lambda_l^{\beta-1} \gamma_l \beta_l(t)^2 = t\|A^{(\beta-1)/2}\|_{L_2^0},$$

which implies that $\|A^{(\beta-1)/2}\|_{L_2^0} < \infty$ for $\beta \in [0,1]$. The proof is complete. □

We also need regularity in time of the solution of (1.1); see Printems [24, Proposition 3.4].

LEMMA 2.7. *Assume that* (ii′) *holds. Let u be the mild solution of* (1.1). *Then we have, for $0 \leq \gamma < \beta \leq 1$,*

$$\begin{aligned} \mathbf{E}\|u(t_2) - u(t_1)\|^2 &\leq C(t_2 - t_1)^{\gamma} \mathbf{E}|u_0|_{\gamma}^2 \\ &\quad + C(t_2 - t_1)^{\gamma} \sup_{0 \leq s \leq T} \mathbf{E}\|u(s)\|^2. \end{aligned} \tag{2.9}$$

2.2. Error estimates for the deterministic problem. In order to prove error estimates for the stochastic parabolic partial differential equation in the fully discrete case, we need some error estimates for the corresponding deterministic parabolic problem.

We first introduce some operators. Consider the stationary problem

$$-\Delta u = f \text{ in } \mathcal{D}, \quad \text{with } u = 0 \text{ on } \partial\mathcal{D}, \tag{2.10}$$

where $f \in \dot{H}^{-1}$.

The variational form of (2.10) is to find $u \in H_0^1$ such that

$$(\nabla u, \nabla \phi) = \langle f, \phi \rangle \quad \forall \phi \in H_0^1, \tag{2.11}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing between $\dot{H}^{-1}$ and H_0^1.

Let $S_h \subset H_0^1$ be the finite element space. The semidiscrete problem of (2.11) is to find $u_h \in S_h$ such that

$$(\nabla u_h, \nabla \chi) = \langle f, \chi \rangle \quad \forall \chi \in S_h. \tag{2.12}$$

By the Lax–Milgram lemma, there exist unique solutions $u \in H_0^1$ and $u_h \in S_h$ such that (2.11) and (2.12) hold. Moreover the following stability result holds:

$$|u|_1 \leq C|f|_{-1} \quad \forall f \in \dot{H}^{-1}. \tag{2.13}$$

The standard error estimates read

$$\|u_h - u\| \le Ch^s |u|_s, \quad s = 1, 2. \tag{2.14}$$

Let $G : \dot{H}^{-1} \to H_0^1$ denote the exact solution operator of (2.10), i.e., $u = Gf$. We define the linear operator $G_h : \dot{H}^{-1} \to S_h$ by $G_h f = u_h$, so that $u_h = G_h f \in S_h$ is the approximate solution of (2.11). It is easy to see that G_h is self-adjoint, positive semidefinite on H, and positive definite on S_h. Introducing the elliptic projection $R_h : H_0^1 \to S_h$ by

$$(\nabla R_h v, \nabla \chi) = (\nabla v, \nabla \chi) \quad \forall v \in H_0^1,$$

we see that $G_h = R_h G$, and $R_h v$ is the finite element approximation of the solution of the corresponding elliptic problem with exact solution v. By (2.14), we get

$$\|R_h v - v\| \le Ch^s |v|_s, \quad s = 1, 2.$$

Hence, using (2.13) and the elliptic regularity estimate, we have

$$\big\| \big(G_h - G\big) f \big\| = \|(R_h - I) G f\| \le Ch^s |Gf|_s = Ch^s |f|_{s-2} \quad \text{for } s = 1, 2, \tag{2.15}$$

which we need below.

Let $E_{kh} = r(kA_h)$ and $E(t_n) = e^{-t_n A}$, where $r(\lambda) = 1/(1+\lambda)$ is introduced in (1.8). We have

Lemma 2.8. *Let $F_n = E_{kh}^n P_h - E(t_n)$. Then*

$$\|F_n v\| \le C(k^{\beta/2} + h^{\beta}) |v|_\beta \quad \textit{for } v \in \dot{H}^\beta,\ 0 \le \beta \le 1, \tag{2.16}$$

and

$$\left(k \sum_{j=1}^n \|F_j v\|^2 \right)^{1/2} \le C(k^{\beta/2} + h^\beta) |v|_{\beta-1} \quad \textit{for } v \in \dot{H}^{\beta-1},\ 0 \le \beta \le 1. \tag{2.17}$$

Furthermore, in the weak norm,

$$|F_n v|_{-1} \le C(k^{\beta/2} + h^\beta) |v|_{\beta-1} \quad \textit{for } v \in \dot{H}^{\beta-1},\ 1 \le \beta \le 2, \tag{2.18}$$

and, with $\ell_k = \log(T/k)$, where $T = t_n$,

$$\left(k \sum_{j=1}^n |F_j v|_{-1}^2 \right)^{1/2} \le C(k^{\beta/2} + h^\beta) \ell_k |v|_{\beta-2} \quad \textit{for } v \in \dot{H}^{\beta-2},\ 1 \le \beta \le 2. \tag{2.19}$$

Proof. Here we prove only (2.17) in detail. Other proofs are similar. We define $u(t_n) = u^n = E(t_n) v$, $U^n = E_{kh}^n P_h v$, and $e^n = F_n v$. By interpolation theory, it suffices to show that

$$\left(k \sum_{j=1}^n \|F_j v\|^2 \right)^{1/2} \le C |v|_{-1} \tag{2.20}$$

and

$$\left(k\sum_{j=1}^{n}\|F_j v\|^2\right)^{1/2} \le C(k^{1/2}+h)\|v\|. \tag{2.21}$$

With $\partial_t e^n = (e^n - e^{n-1})/k$, we have the following error equation:

$$G_h\partial_t e^n + e^n = \rho^n + G_h\tau^n, \tag{2.22}$$

where $\rho^n = (G_h - G)u_t(t_n)$ and $\tau^n = u_t(t_n) - \partial_t u^n$.

Taking the inner product of (2.22) with e^n, we get

$$(G_h\partial_t e^n, e^n) + (e^n, e^n) = (\rho^n, e^n) + (G_h\tau^n, e^n).$$

By summation on n, using the inequality $(\rho^n, e^n) \le \frac{1}{2}(\|\rho^n\|^2 + \|e^n\|^2)$, and noting that $G_h e^0 = 0$, we have

(2.23)

$$(G_h e_n, e_n) + k\sum_{j=1}^{n}\|e_j\|^2 \le Ck\sum_{j=1}^{n}\|\rho^j\|^2 + Ck\sum_{j=1}^{n}\|G\tau^j\|^2 + Ck\sum_{j=1}^{n}\|(G_h - G)\tau^j\|^2.$$

Here, using Lemma 2.2, we have, since $\rho^j = \rho(s) + \int_s^{t_j}\rho_t(\tau)\,d\tau$,

$$\begin{aligned} k\sum_{j=1}^{n}\|\rho^j\|^2 &= k\|\rho\|^2 + \sum_{j=2}^{n}\int_{t_{j-1}}^{t_j}\|\rho^j\|^2\,ds \\ &\le k\|\rho\|^2 + 2\sum_{j=2}^{n}\int_{t_{j-1}}^{t_j}\left(\|\rho(s)\|^2 + \left\|\int_s^{t_j}\rho_t(\tau)\,d\tau\right\|^2\right)ds \\ &\le k\|\rho\|^2 + 2\int_{t_1}^{t_n}\|\rho(s)\|^2\,ds + 2\sum_{j=2}^{n}\int_{t_{j-1}}^{t_j}\left((t_j - s)\int_s^{t_j}\|\rho_t(\tau)\|^2\,d\tau\right)ds \\ &\le k\|\rho\|^2 + 2\int_{t_1}^{t_n}\|\rho(s)\|^2\,ds + 2k\sum_{j=2}^{n}\int_{t_{j-1}}^{t_j}\tau\|\rho_t(\tau)\|^2\,d\tau \\ &\le Ck\|u\|^2 + Ch^2\int_0^{t_n}|u(s)|_1^2\,ds + Ck\int_0^{t_n}\tau\|u_t(\tau)\|^2\,d\tau \le C(k+h^2)\|v\|^2, \end{aligned} \tag{2.24}$$

and, by Taylor's formula,

$$\begin{aligned} k\sum_{j=1}^{n}\|(G_h - G)\tau^j\|^2 &\le Ckh^2|\tau^1|_{-1}^2 + Ckh^2\sum_{j=2}^{n}|\tau^j|_{-1}^2 \\ &= Ckh^2\left|u_t(k) - \frac{1}{k}\int_0^k u_t(\tau)\,d\tau\right|_{-1}^2 + Ckh^2\sum_{j=2}^{n}\left|\frac{1}{k}\int_{t_{j-1}}^{t_j}(s - t_{j-1})u_{tt}(s)\,ds\right|_{-1}^2 \\ &\le Ch^2\|v\|^2 + Ch^2\sum_{j=2}^{n}\int_{t_{j-1}}^{t_j}k(s - t_{j-1})|u_{tt}(s)|_{-1}^2\,ds \\ &\le Ch^2\|v\|^2 + Ch^2\sum_{j=2}^{n}\int_{t_{j-1}}^{t_j}s^2|u_{tt}(s)|_{-1}^2\,ds \le Ch^2\|v\|^2, \end{aligned}$$

and

$$k\sum_{j=1}^{n}\|G\tau^j\|^2 = k\sum_{j=1}^{n}\left\|\frac{1}{k}\int_{t_{j-1}}^{t_j}(s-t_{j-1})u_t(s)\,ds\right\|^2$$
$$\le k\sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}(s-t_{j-1})\|u_t(s)\|^2\,ds \le Ck\int_0^{t_n} s\|u_t(s)\|^2\,ds \le k\|v\|^2.$$

We therefore obtain

$$(G_h e^n, e^n)^{1/2} + \left(k\sum_{j=1}^{n}\|e^j\|^2\right)^{1/2} \le C(k^{1/2}+h)\|v\|, \tag{2.25}$$

which implies that (2.21) holds. □

3. Proofs of Theorems 1.1 and 1.2. In this section, we will consider the proofs of Theorems 1.1 and 1.2.

Proof of Theorem 1.1. We have, by (1.8), with $E_{kh}^n = r(kA_h)^n$,

$$U^n = E_{kh}^n P_h u_0 + \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j} E_{kh}^{n-j}P_h\sigma(U^{j-1})\,dW(s),$$

and, by the definition of the mild solution of (1.1), with $E(t)=e^{-tA}$,

$$u(t_n) = E(t_n)u_0 + \int_0^{t_n} E(t_n-s)\sigma(u(s))\,dW(s).$$

Defining $e^n = U^n - u(t_n)$ and $F_n = E_{kh}^n P_h - E(t_n)$, we write

$$\begin{aligned} e^n &= F_n u_0 + \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j} r(kA_h)^{n-j}P_h\Big(\sigma(U^{j-1}) - \sigma(u(t_{j-1}))\Big)\,dW(s) \\ &\quad + \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j} r(kA_h)^{n-j}P_h\Big(\sigma(u(t_{j-1})) - \sigma(u(s))\Big)\,dW(s) \\ &\quad + \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\Big(r(kA_h)^{n-j}P_h - E(t_n-t_j)\Big)\sigma(u(s))\,dW(s) \\ &\quad + \sum_{j=1}^{n}\int_{t_{j-1}}^{t_j}\Big(E(t_n-t_j) - E(t_n-s)\Big)\sigma(u(s))\,dW(s) \\ &= \sum_{j=1}^{5} I_j. \end{aligned}$$

Thus

$$\|e^n\|_{L_2(\Omega;H)} \le C\sum_{j=1}^{5}\|I_j\|_{L_2(\Omega;H)}.$$

For I_1, we have, by (2.16) with $v=u_0$,

$$\|I_1\| = \|F_n u_0\| \le C(k^{\beta/2}+h^\beta)|u_0|_\beta,$$

which implies that $\|I_1\|_{L_2(\Omega;H)} \le C(k^{\beta/2}+h^\beta)\|u_0\|_{L_2(\Omega;\dot H^\beta)}$.

For I_2, we have, by isometry, the stability of $r(\lambda)$, and the Lipschitz condition (i),

$$\begin{aligned}
\|I_2\|^2_{L_2(\Omega;H)} &= \mathbf{E}\left\|\sum_{j=1}^n \int_{t_{j-1}}^{t_j} r(kA_h)^{n-j}P_h\Big(\sigma(U^{j-1})-\sigma(u(t_{j-1}))\Big)\,dW(s)\right\|^2\\
&= k\sum_{j=1}^n \mathbf{E}\Big\|r(kA_h)^{n-j}P_h\Big(\sigma(U^{j-1})-\sigma(u(t_{j-1}))\Big)\Big\|^2_{L_2^0}\\
&\le k\sum_{j=1}^n \|r(kA_h)^{n-j}P_h\|^2\,\mathbf{E}\|\sigma(U^{j-1})-\sigma(u(t_{j-1}))\|^2_{L_2^0}\\
&\le Ck\sum_{j=1}^n \mathbf{E}\|U^{j-1}-u(t_{j-1})\|^2 = C\sum_{j=1}^n\int_{t_{j-1}}^{t_j}\mathbf{E}\|e^{j-1}\|^2\,ds\\
&\le Ck\sum_{j=1}^n \|e^j\|^2.
\end{aligned}$$

For I_3, we have, by Lemma 2.7, for $0\le\gamma<\beta\le 1$,

$$\begin{aligned}
\|I_3\|^2_{L_2(\Omega;H)} &= \sum_{j=1}^n\int_{t_{j-1}}^{t_j}\mathbf{E}\Big\|r(kA_h)^{n-j}P_h\Big(\sigma(u(t_{j-1}))-\sigma(u(s))\Big)\Big\|^2_{L_2^0}\,ds\\
&\le C\sum_{j=1}^n\int_{t_{j-1}}^{t_j}\mathbf{E}\|u(t_{j-1})-u(s)\|^2\,ds\\
&\le C\left(\sum_{j=1}^n\int_{t_{j-1}}^{t_j}(s-t_{j-1})^\gamma\,ds\right)\left(\mathbf{E}|u_0|_\gamma^2+\sup_{0\le s\le T}\mathbf{E}\|u(s)\|^2\right)\\
&\le Ck^\gamma\left(\mathbf{E}|u_0|_\gamma^2+\sup_{0\le s\le T}\mathbf{E}\|u(s)\|^2\right).
\end{aligned}$$

For I_4, we have

$$\begin{aligned}
\|I_4\|^2_{L_2(\Omega;H)} &= \mathbf{E}\left\|\sum_{j=1}^n\int_{t_{j-1}}^{t_j}F_{n-j}\sigma(u(s))\,dW(s)\right\|^2\\
&= \sum_{j=1}^n\int_{t_{j-1}}^{t_j}\mathbf{E}\|F_{n-j}A^{(1-\beta)/2}A^{(\beta-1)/2}\sigma(u(s))\|^2_{L_2^0}\,ds\\
&\le C\left(k\sum_{j=1}^n\|F_jA^{(1-\beta)/2}\|^2\right)\sup_{0\le s\le T}\mathbf{E}\|u(s)\|^2.
\end{aligned}$$

We will show that

$$k\sum_{j=1}^n\|F_jA^{(1-\beta)/2}\|^2\le C(k^\beta+h^{2\beta}). \tag{3.1}$$

Assuming this for the moment, we get

$$\|I_4\|^2_{L_2(\Omega;H)}\le C(k^\beta+h^{2\beta})\sup_{0\le s\le T}\mathbf{E}\|u(s)\|^2.$$

For I_5, we have

$$\begin{aligned}\|I_5\|^2_{L_2(\Omega;H)} &= \mathbf{E}\left\|\sum_{j=1}^n \int_{t_{j-1}}^{t_j} (E(t_n - t_j) - E(t_n - s))\sigma(u(s))\,dW(s)\right\|^2 \\ &= \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \mathbf{E}\|(E(t_n - t_j) - E(t_n - s))A^{(1-\beta)/2}A^{(\beta-1)/2}\sigma(u(s))\|^2_{L_2^0}\,ds \\ &\le C\left(\sum_{j=1}^n \int_{t_{j-1}}^{t_j} \|(E(t_n - t_j) - E(t_n - s))A^{(1-\beta)/2}\|^2\,ds\right) \sup_{0\le s\le T} \mathbf{E}\|u(s)\|^2.\end{aligned}$$

Noting that, by Lemmas 2.2 and 2.3,

$$\begin{aligned}&\sum_{j=1}^n \int_{t_{j-1}}^{t_j} \|(E(t_n - t_j) - E(t_n - s))A^{(1-\beta)/2}\|^2\,ds \\ &= \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \|A^{1/2}E(t_n - t_j)A^{-\beta/2}(I - E(t_j - s))\|^2\,ds \\ &\le Ck^\beta \sum_{j=1}^n \int_{t_{j-1}}^{t_j} \|A^{1/2}E(t_n - t_j)\|^2\,ds \\ &= Ck^\beta \left(\sum_{j=1}^n k\|A^{1/2}E(t_n - t_j)\|^2\right) \le Ck^\beta,\end{aligned}$$

we have

$$\|I_5\|^2_{L_2(\Omega;H)} \le Ck^\beta \sup_{0\le s\le T} \mathbf{E}\|u(s)\|^2.$$

It remains to show (3.1). In fact, by (2.17),

$$\begin{aligned}k\sum_{j=1}^n \|F_j A^{(1-\beta)/2}\|^2 &= k\sum_{j=1}^n \left(\sup_{v\ne 0} \frac{\|F_j A^{(1-\beta)/2}v\|}{\|v\|}\right)^2 \\ &= \sup_{v\ne 0} \frac{k\sum_{j=1}^n \|F_j A^{(1-\beta)/2}v\|^2}{\|v\|^2} \\ &\le \sup_{v\ne 0} \frac{C(k^\beta + h^{2\beta})|A^{(1-\beta)/2}v|^2_{\beta-1}}{\|v\|^2} \le C(k^\beta + h^{2\beta}).\end{aligned}$$

Together these estimates show, for $0 \le \gamma < \beta \le 1$,

$$\begin{aligned}\mathbf{E}\|e^n\|^2 \le C(k^\gamma + h^{2\beta})\mathbf{E}|u_0|^2_\beta + Ck\sum_{j=1}^n \mathbf{E}\|e^j\|^2 \\ + C(k^\gamma + h^{2\beta}) \sup_{0\le s\le T} \mathbf{E}\|u(s)\|^2.\end{aligned} \tag{3.2}$$

By the discrete Gronwall lemma, we get

$$\mathbf{E}\|e^n\|^2 \le C(k^\gamma + h^{2\beta})\Big(\mathbf{E}|u_0|^2_\beta + \sup_{0\le s\le T} \mathbf{E}\|u(s)\|^2\Big), \tag{3.3}$$

which implies that

$$\|e^n\|_{L_2(\Omega;H)} \leq C(k^{\gamma/2} + h^{\beta})\Big(\mathbf{E}|u_0|_{L_2(\Omega;\dot{H}^{\beta})} + \sup_{0\leq s\leq T} \|u(s)\|_{L_2(\Omega;H)}\Big). \tag{3.4}$$

The proof is now complete. □

Now we turn to the proof of Theorem 1.2.

Proof of Theorem 1.2. We first consider the case $l = 0$. We have, by (1.8), with $E_{kh}^n = r(kA_h)^n$, noting that $\sigma(\cdot) = I$,

$$U^n = E_{kh}^n P_h u_0 + \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} E_{kh}^{n-j+1} P_h \, dW(s),$$

and, by the definition of the mild solution of (1.1), with $E(t) = e^{-tA}$,

$$u(t_n) = E(t_n)u_0 + \int_0^{t_n} E(t_n - s)\, dW(s).$$

Defining $e^n = U^n - u(t_n)$ and $F_n = E_{kh}^n P_h - E(t_n)$, we write

$$\begin{aligned} e^n &= F_n u_0 + \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} F_{n-j+1}\, dW(s) \\ &\quad + \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \Big(E(t_n - t_{j-1}) - E(t_n - s)\Big)\, dW(s) \\ &= I + II + III. \end{aligned}$$

Thus

$$\|e^n\|_{L_2(\Omega;H)} \leq C\Big(\|I\|_{L_2(\Omega;H)} + \|II\|_{L_2(\Omega;H)} + \|III\|_{L_2(\Omega;H)}\Big).$$

For I, we have, by (2.16) with $v = u_0$,

$$\|I\| = \|F_n u_0\| \leq C(k^{\beta/2} + h^{\beta})|u_0|_{\beta},$$

which implies that $\|I\|_{L_2(\Omega;H)} \leq C(k^{\beta/2} + h^{\beta})\|u_0\|_{L_2(\Omega;\dot{H}^{\beta})}$.

For II, we have, by the isometry property,

$$\begin{aligned} \|II\|_{L_2(\Omega;H)}^2 &= \mathbf{E}\left\|\sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} F_{n-j+1}\, dW(s)\right\|^2 = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \|F_{n-j+1}\|_{L_2^0}^2\, ds \\ &= \sum_{l=1}^{\infty} \left(k \sum_{j=1}^{n} \|F_{n-j+1} Q^{1/2} e_l\|^2\right), \end{aligned}$$

where $\{e_l\}$ is any orthonormal basis in H. Using (2.17) with $v = Q^{1/2}e_l$, we obtain

$$\begin{aligned} \|II\|_{L_2(\Omega;H)}^2 &\leq C \sum_{l=1}^{\infty} (k^{\beta} + h^{2\beta})|Q^{1/2}e_l|_{\beta-1}^2 \\ &= C \sum_{l=1}^{\infty} (k^{\beta} + h^{2\beta})\|A^{(\beta-1)/2}Q^{1/2}e_l\|^2 = C(k^{\beta} + h^{2\beta})\|A^{(\beta-1)/2}\|_{L_2^0}^2. \end{aligned}$$

For III, we have, by the isometry property,

$$\|III\|^2_{L_2(\Omega;H)} = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \left\| \Big(E(t_n - t_{j-1}) - E(t_n - s) \Big) \right\|^2_{L_2^0} ds$$
$$= \sum_{l=1}^{\infty} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \left\| A^{-\beta/2} \Big(E(s - t_{j-1}) - I \Big) A^{\beta/2} E(t_n - s) Q^{1/2} e_l \right\|^2 ds.$$

Using (2.6), and (2.4) with $v = A^{(\beta-1)/2} Q^{1/2} e_l$, we obtain

$$(3.5) \qquad \|III\|^2_{L_2(\Omega;H)} \le Ck^\beta \sum_{l=1}^{\infty} \int_0^{t_n} \|A^{1/2} E(t_n - s) A^{(\beta-1)/2} Q^{1/2} e_l\|^2 \, ds$$
$$\le Ck^\beta \sum_{l=1}^{\infty} \|A^{(\beta-1)/2} Q^{1/2} e_l\|^2 = Ck^\beta \|A^{(\beta-1)/2}\|^2_{L_2^0},$$

which completes the proof of (1.9).

In particular, if $W(t)$ is a Wiener process with $\mathrm{Tr}(Q) < \infty$, then we can choose $\beta = 1$ since $\|I\|_{L_2^0} = \mathrm{Tr}(Q)$.

Now we turn to the case $l = 1$. We have, by (2.18),

$$\|I\|_{L_2(\Omega;\dot{H}^{-1})} \le Ch^{\beta+1} \|u_0\|_{L_2(\Omega;\dot{H}^\beta)} \quad \text{for } 0 \le \beta \le 1.$$

For II, we have, by the isometry property and (2.19) with $v = Q^{1/2} e_l$,

$$\|II\|^2_{L_2(\Omega;\dot{H}^{-1})} = \mathbf{E} \left\| \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} A^{-1/2} F_{n-j+1} \, dW(s) \right\|^2$$
$$= \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \|A^{-1/2} F_{n-j+1}\|^2_{L_2^0} \, ds$$
$$= \sum_{l=1}^{\infty} \left(k \sum_{j=1}^{n} \|A^{-1/2} F_{n-j+1} Q^{1/2} e_l\|^2 \right)$$
$$\le C(k^{\beta+1} + h^{2(\beta+1)}) \ell_k^2 \sum_{l=1}^{\infty} \|A^{(\beta-1)/2} Q^{1/2} e_l\|^2$$
$$\le C(k^{\beta+1} + h^{2(\beta+1)}) \ell_k^2 \|A^{(\beta-1)/2}\|^2_{L_2^0}.$$

For III, we have, by the isometry property,

$$\|III\|^2_{L_2(\Omega;\dot{H}^{-1})} = \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \left\| A^{-1/2} \Big(E(t_n - t_{j-1}) - E(t_n - s) \Big) \right\|^2_{L_2^0} ds$$
$$= \sum_{l=1}^{\infty} \sum_{j=1}^{n} \int_{t_{j-1}}^{t_j} \left\| A^{-(\beta+1)/2} \Big(E(s - t_{j-1}) - I \Big) A^{1/2} E(t_n - s) A^{(\beta-1)/2} Q^{1/2} e_l \right\|^2 ds.$$

Following the proof of (3.5), we get

$$\|III\|^2_{L_2(\Omega;\dot{H}^{-1})} \le Ck^\beta \|A^{(\beta-2)/2}\|^2_{L_2^0},$$

which completes the proof of (1.11).

Similarly, if $W(t)$ is a Wiener process with $\mathrm{Tr}(Q) < \infty$, then we can choose $\beta = 1$. □

If $\sigma(\cdot) = I$ and $d = 1$, we can specialize to the case $Q = I$.

Corollary 3.1. *Let U^n and $u(t_n)$ be the solutions of* (1.8) *and* (1.1)*, respectively. Assume that $\sigma(\cdot) = I$. Assume that $A = -\frac{\partial^2}{\partial x^2}$ and $\mathcal{D}(A) = H_0^1(0,1) \cap H^2(0,1)$. If $W(t)$ is a cylindrical Wiener process with $Q = I$, then we have, for $u_0 \in L_2(\Omega; \dot{H}^\beta)$, for $l = 0, 1$, with $\ell_k = \log(T/k)$, where $T = t_n$,*

$$\|U^n - u(t_n)\|_{L_2(\Omega;\dot{H}^{-l})} \le C(k^{(\beta+l)/2} + h^{(\beta+l)})\big(\|u_0\|_{L_2(\Omega;\dot{H}^\beta)} + \ell_k\big) \quad \text{for } 0 \le \beta < \frac{1}{2}.$$

4. Computational analysis.

In this section we consider how to compute the approximate solution U^n of the solution u of (1.1). For simplicity, we assume that $\sigma(\cdot) = I$. Recall that the Wiener process $W(t)$ with covariance operator Q has the form (see Da Prato and Zabczyk [8, Chapter 4])

$$W(t) = \sum_{j=1}^{\infty} \gamma_j^{1/2} e_j \beta_j(t), \tag{4.1}$$

where $\{\gamma_j, e_j\}_{j=1}^\infty$ is an eigensystem of Q, and $\{\beta_j(t)\}_{j=1}^\infty$ are independently and identically distributed (iid) real-valued Brownian motions. If $\mathrm{Tr}(Q) < \infty$, then $W(t)$ is an H-valued process. In fact,

$$\mathbf{E}\|W(t)\|^2 = \mathbf{E}\sum_{j=1}^{\infty} \gamma_j \beta_j(t)^2 = \sum_{j=1}^{\infty} \gamma_j \big(\mathbf{E}\beta_j(t)^2\big) = t\,\mathrm{Tr}(Q) < \infty.$$

If $\mathrm{Tr}(Q) = \infty$, for example $Q = I$, then $W(t)$ is not H-valued.

Let U^n be the approximation in S_h of $u(t)$ at $t = t_n = nk$. The backward Euler method is to find $U^n \in S_h$ such that, with $\bar{\partial}U^n = (U^n - U^{n-1})/k$, $n \ge 1$, $U^0 = P_h u_0$,

$$(\bar{\partial}U^n, \chi) + (A_h U^n, \chi) = \left(\frac{1}{k}\int_{t_{n-1}}^{t_n} P_h\, dW(s), \chi\right) \quad \forall \chi \in S_h, \tag{4.2}$$

where A_h, P_h are defined in the introduction.

Since $W(t)$ is $\dot{H}^{\beta-1}$-valued with $\beta \in [0,1]$, $P_h W(t)$ is well defined. We therefore can write

$$\int_{t_{n-1}}^{t_n} P_h\, dW(s) = P_h\big(W(t_n) - W(t_{n-1})\big) = P_h \sum_{j=1}^{\infty} \gamma_j^{1/2}\big(\beta_j(t_n) - \beta_j(t_{n-1})\big).$$

Here

$$\frac{1}{\sqrt{k}}\Big(\beta_j(t_n) - \beta_j(t_{n-1})\Big) = \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ is the real-valued Gaussian random variable.

Thus the right-hand side of (4.2) can be computed by truncating the following series to J terms:

$$\left(\frac{1}{k}\int_{t_{n-1}}^{t_n} P_h\,dW(s),\chi\right) = \left(\frac{1}{k}\sum_{j=1}^{\infty}\gamma_j^{1/2}e_j\big(\beta_j(t_n)-\beta_j(t_{n-1})\big),\chi\right) \tag{4.3}$$
$$= \frac{1}{k}\sum_{j=1}^{\infty}\gamma_j^{1/2}\big(\beta_j(t_n)-\beta_j(t_{n-1})\big)(e_j,\chi)$$
$$\approx \frac{1}{k}\sum_{j=1}^{J}\gamma_j^{1/2}\big(\beta_j(t_n)-\beta_j(t_{n-1})\big)(e_j,\chi).$$

Denote by N_h the dimension of S_h. Below we will show that it is sufficient to choose $J = N_h$ in order to achieve the required convergence order. To see this, let us consider the semidiscrete approximation solution u_h of u of (1.1). Recall that the semidiscrete solution u_h satisfies

$$u_h(t) = E_h(t)P_hu_0 + \int_0^t E_h(t-s)P_h\,dW(s) \tag{4.4}$$
$$= E_h(t)P_hu_0 + \sum_{j=1}^{\infty}\int_0^t E_h(t-s)P_he_j\gamma_j^{1/2}\,d\beta_j(s).$$

Truncating the series in the right-hand side of (4.4), we have

$$u_h^J(t) = E_h(t)P_hu_0 + \sum_{j=1}^{J}\int_0^t E_h(t-s)P_he_j\gamma_j^{1/2}\,d\beta_j(s). \tag{4.5}$$

We then have the following lemma with respect to the L_2 norm in space.

LEMMA 4.1. *Let u_h^J and u_h be defined by (4.4) and (4.5), respectively. Assume that $\{S_h\}$ is defined on a quasi-uniform family of triangulations, and let N_h be the dimension of S_h. Assume that $\|A^{(\beta-1)/2}\|_{L_2^0} < \infty$ for some $\beta \in [0,1]$. If $J \geq N_h$, then we have, for $t > 0$,*

$$\|u_h^J(t) - u_h(t)\|_{L_2(\Omega,H)} \leq Ch^{\beta}\|A^{(\beta-1)/2}\|_{L_2^0}. \tag{4.6}$$

Proof. We have, by the isometry property, with $F_h(t) = E_h(t)P_h - E(t)$,

$$\mathbf{E}\|u_h^J(t)-u_h(t)\|^2 = \mathbf{E}\left\|\sum_{j=J+1}^{\infty}\int_0^t E_h(t-s)P_he_j\gamma_j^{1/2}\,d\beta_j(s)\right\|^2$$
$$= \sum_{j=J+1}^{\infty}\gamma_j\int_0^t\|E_h(t-s)P_he_j\|^2\,ds$$
$$\leq 2\sum_{j=J+1}^{\infty}\gamma_j\int_0^t\|E(t-s)e_j\|^2\,ds$$
$$+ 2\sum_{j=J+1}^{\infty}\gamma_j\int_0^t\|F_h(t-s)e_j\|^2\,ds$$
$$= I + II.$$

For I, we have

$$\begin{aligned} I &= 2\sum_{j=J+1}^{\infty}\gamma_j\int_0^t e^{-2(t-s)\lambda_j}\,ds \le \sum_{j=J+1}^{\infty}\gamma_j\lambda_j^{-1} \\ &= \sum_{j=J+1}^{\infty}\lambda_j^{-\beta}\lambda_j^{\beta-1}\gamma_j \le \lambda_{J+1}^{-\beta}\|A^{(\beta-1)/2}\|_{L_2^0}^2. \end{aligned}$$

For II, we have, noting that $\int_0^t \|F_h(t)v\|^2\,dt \le Ch^{\beta}|v|_{\beta-1}^2$ (see Thomée [28]),

$$II \le Ch^{2\beta}\sum_{j=J+1}^{\infty}\gamma_j|e_j|_{\beta-1}^2 \le Ch^{2\beta}\sum_{j=1}^{\infty}|Q^{1/2}e_j|_{\beta-1}^2 = Ch^{2\beta}\|A^{(\beta-1)/2}\|_{L_2^0}^2.$$

Thus we get

$$\mathbf{E}\|u_h^J(t)-u_h(t)\|^2 \le C(\lambda_{J+1}^{-\beta}+h^{2\beta})\|A^{(\beta-1)/2}\|_{L_2^0}^2.$$

Hence (4.6) follows from the following obvious facts: with some constant C which may be different in different inequalities,

$$\lambda_{J+1}^{-1} \le CJ^{-2/d} \le CN_h^{-2/d} \le Ch^2,$$

where d is the dimension of the spatial domain $\mathcal{D}$. □

Under the same assumptions as in Lemma 4.1, we can also show the following results with respect to the weak norm in space:

$$\|u_h^J(t)-u_h(t)\|_{L_2(\Omega,\dot{H}^{-1})} \le Ch^{\beta+1}\ell_h\|A^{(\beta-1)/2}\|_{L_2^0}.$$

5. Numerical illustrations. In this section, we will show some numerical experiments. We consider the following one-dimensional stochastic partial differential equation driven by white noise (see Allen, Novosel, and Zhang [1] and also Du and Zhang [11]):

$$\begin{cases} \frac{\partial u}{\partial t}(t,x) - \frac{\partial^2 u}{\partial^2 x}(t,x) + bu(t,x) = \frac{\partial^2 W}{\partial t\partial x}(t,x) + g(t,x), \quad t>0, \\ u(0,x) = 10x^2(1-x)^2, \quad 0\le x\le 1, \\ u(t,0)=u(t,1)=0, \quad t\ge 0, \end{cases} \tag{5.1}$$

where $\frac{\partial^2 W}{\partial t\partial x}$ denotes the mixed second order derivative of the Brownian sheet and $b=0.5$, and

$$g(t,x) = 10(1+b)x^2(1-x)^2e^t - 10(2-12x+12x^2)e^t.$$

Let U^n be the approximation in S_h of $u(t)$ at $t=t_n=nk$. Define the following backward Euler method of (5.1), with $\bar\partial U^n = (U^n-U^{n-1})/k$, $n\ge 1$, $U^0 = P_h u_0$:

(5.2)
$$(\bar\partial U^n,\chi) + ((U^n)',\chi') + b(U^n,\chi) = \left(\frac{1}{k}\int_{t_{n-1}}^{t_n} P_h\,dW(s),\chi\right) + (g^n,\chi) \quad \forall\chi\in S_h,$$

where $(U^n)',\chi'$ denote the derivatives with respect to the spatial variable, where $W(t)$ is given by (4.1) with $\gamma_j=1$.

Table 5.1
Ratios of error by using different time steps.

i	k_i	$S(k_i)/S(k_{i+1})$	$D(k_i)/D(k_{i+1})$
1	2^{-1}	1.1705	2.3468
2	2^{-2}	1.1801	2.2778
3	2^{-3}	1.2041	2.2562
4	2^{-4}	1.2322	2.1845
5	2^{-5}	1.2441	2.1025
6	2^{-6}	1.2049	2.0938
7	2^{-7}		

We approximate the stochastic integral $\int_{t_{n-1}}^{t_n} P_h dW(s)$ by

$$\left(\int_{t_{n-1}}^{t_n} P_h dW, \chi\right) \approx \left(\sum_{j=1}^{N_h} \gamma_j^{1/2} e_j(\beta_j(t) - \beta_{j-1}(t)), \chi\right),$$

where $e_j = \sqrt{2}\sin j\pi x$. By Corollary 3.1 and section 4, we see that there exists a constant $C > 0$ such that

$$\|U^n - u(t_n)\|_{L_2(\Omega,H)} \le C(k^{\beta/2} + h^{\beta}) \quad \text{for } 0 \le \beta < \frac{1}{2}. \tag{5.3}$$

In our experiment, we want to see how the error estimates in (5.3) depend on k. To do this, we choose fixed small $h > 0$ and a sequence of moderate time steps k_i. In Table 5.1, we choose $h = 2^{-8}$ and different time steps $k_i = 2^{-i}$, $i = 1, \dots, 7$. We consider $M = 100$ simulations. For each simulation ω_j, $j = 1, 2, \dots, M$, we generate N_h independent Brownian motions $\beta_l(t)$, $l = 1, 2, \dots, N_h$, and compute $U^n \approx u(t_n)$ at time $t_n = 1$ by using the different time step k_i. We then compute the following L_2 norm of the error at $t_n = 1$ for each simulation ω_j, $j = 1, 2, \dots, M$,

$$\epsilon(k_i, \omega_j) = \epsilon(k_i, \omega_j, t_n) = \|U^n(\omega_j) - u(t_n, \omega_j)\|^2,$$

where the "true" solution $u(t_n, \omega_j)$ is approximated by a solution computed by small time step $k = 2^{-10}$ and space step $h = 2^{-8}$. We then average $\epsilon(k_i, \omega_j)$ with respect to ω_j to obtain the following approximation of $\|U^n - u(t_n)\|_{L_2(\Omega,H)}$ for fixed time step k_i:

$$\begin{aligned} S(k_i) &= \left(\frac{1}{M}\sum_{j=1}^{M} \epsilon(k_i, \omega_j)\right)^{1/2} = \left(\frac{1}{M}\sum_{j=1}^{M} \|U^n(\omega_j) - u(t_n, \omega_j)\|^2\right)^{1/2} \\ &\approx \|U^n - u(t_n)\|_{L_2(\Omega,H)}, \end{aligned}$$

where

$$\|U^n - u(t_n)\|_{L_2(\Omega,H)} = \left(\int_{\Omega} \|U^n(\omega) - u(t_n, \omega)\|^2 \, d\mathbf{P}(\omega)\right)^{1/2}.$$

Since the convergence order is almost $O(k^{1/4})$, we expect that

$$\frac{S(k_i)}{S(k_{i+1})} \approx \left(\frac{k_i}{k_{i+1}}\right)^{1/4} = 2^{1/4} \approx 1.19$$

Table 5.2
The approximation of $\mathbf{E}u(1,0.5)$ *and* $\mathbf{E}(u(1,0.5))^2$.

h	k	$\mathbf{E}U(1,0.5)$	$\mathbf{E}(U(1,0.5))^2$
1/4	1/4	1.5281	2.4691
1/4	1/6	1.6364	2.6367
1/4	1/8	1.6473	2.6701
1/4	1/16	1.6977	2.7525
1/4	1/32	1.7011	2.8376
1/8	1/4	1.5278	2.5276
1/8	1/6	1.6308	2.6317
1/8	1/8	1.6579	2.7046
1/8	1/16	1.6802	2.7959
1/8	1/32	1.6993	2.8752
1/16	1/4	1.5259	2.4978
1/16	1/6	1.6364	2.6117
1/16	1/8	1.6492	2.6918
1/16	1/16	1.6889	2.7572
1/16	1/32	1.6951	2.8875
1/32	1/4	1.5357	2.4896
1/32	1/6	1.6674	2.6017
1/32	1/8	1.6894	2.6818
1/32	1/16	1.6951	2.7972
1/32	1/32	1.7003	2.8975

for sufficiently small h such that the error estimates are dominated by k_i. Table 5.1 shows that the numerical result is consistent with the above analysis. To compare with the deterministic convergence order, which is $O(k)$ for the backward Euler method, in Table 5.1, we also compute $D(k_i) = \|U^n - u(t_n)\|$, the L_2 norm of the error at time $t_n = 1$ for fixed time step k_i in the deterministic case. We see that $\frac{D(k_i)}{D(k_{i+1})} \approx 2$, as expected.

In Allen, Novosel, and Zhang [1] and Du and Zhang [11], they show the numerical approximation of $\mathbf{E}(u(t,x))$ and $\mathbf{E}(u(t,x)^2)$ at time $t = 1$ and $x = 0.5$ for (5.1). In Table 5.2, we obtain approximation values similar to those in their papers for different pairs k, h. In our experiment, for each pair (k,h), 1000 runs are performed and N_h independent Brownian motions are generated in each run. We then calculate the average. In Table 5.2, $U(1,0.5)$ denotes the approximation of $u(t,x)$ at $t = 1$ and $x = 0.5$. The computational results converge as k and h approach 0.

Let us review the numerical methods in [1] and [11]. They consider both finite element and finite difference methods for (5.1). They approximate the space-time white noise by using piecewise constant functions on a partition $[t_{n-1}, t_n] \times [x_{j-1}, x_j]$, $1 \le n \le N$, $1 \le j \le J$, of $[0,T] \times [0,1]$. More precisely, with $k = t_n - t_{n-1}$ and $h = x_j - x_{j-1}$,

$$dW(t,x) \approx d\hat{W}(t,x) = \frac{\partial^2 \hat{W}(t,x)}{\partial t \partial x} dt dx = \frac{1}{kh} \sum_{n=1}^{N} \sum_{j=1}^{J} \eta_{nj} \sqrt{kh} \chi_n(t) \chi_j(x) dt dx,$$

where

$$\chi_n(t) = \begin{cases} 1, & t_{n-1} \le t \le t_n, \\ 0 & \text{otherwise}, \end{cases} \qquad \chi_j(x) = \begin{cases} 1, & x_{j-1} \le x \le x_j, \\ 0 & \text{otherwise}, \end{cases}$$

and

$$\eta_{nj} = \frac{1}{kh} \int_{t_{n-1}}^{t_n} \int_{x_{j-1}}^{x_j} dW(t,x) = \mathcal{N}(0,1),$$

where $\mathcal{N}(0,1)$ is the standard real-valued Gaussian random variable and η_{nj} are independently and identically distributed. It is obvious that $\frac{\partial^2 \hat{W}}{\partial t \partial x} \in L_2(0,1)$ for fixed $t \in [0,T]$, $\omega \in \Omega$. Applying the standard finite element and finite difference methods for the new "simpler" problems, they obtain the approximate solution $U^n \approx u(t_n)$ and the corresponding error estimates. For example, using the backward Euler method, the finite element approximate solution U^n satisfies, with $\ell_k = 1 + \log(T/k)$,

$$\|U^n - u(t_n)\|_{L_2(\Omega;H)} \leq C\ell_k(k^{1/4} + h^{1/2}),$$

which is consistent with our estimates in Corollary 3.1.

We can also approximate the stochastic integral by

$$\left(\int_{t_{n-1}}^{t_n} d\hat{W}_h, \chi \right) = \left(\sum_{j=1}^{N_h} \alpha_j^{1/2} \varphi_j (\beta_j(t) - \beta_{j-1}(t)), \chi \right),$$

where

$$\hat{W}_h = \sum_{j=1}^{N_h} \alpha_j^{1/2} \varphi_j \beta_j(t),$$

where φ_j is the finite element basis and α_j is decided specially by solving the following linear system in order to compare the approximation (4.3):

$$\sum_{j=1}^{N_h} \alpha_j (\varphi_j, \chi)^2 = \sum_{j=1}^{N_h} \gamma_j (e_j, \chi)^2 \quad \forall \chi \in S_h.$$

The right-hand side of the above system is related to the Wiener process $P_h W$. Recall that $P_h W$ is an S_h-valued Wiener process with covariance operator $Q_h = P_h Q$ and

$$(Q_h \chi, \chi) = (P_h Q \chi, \chi) = \sum_{j=1}^{\infty} \gamma_j (e_j, \chi)^2 \quad \forall \chi \in S_h.$$

Below we will give a lemma to show the property of the covariance operator of $\hat{W}_h$.

LEMMA 5.1. *$\hat{W}_h$ is an S_h-valued Wiener process with covariance operator $\hat{Q}_h$, where $\hat{Q}_h$ satisfies*

$$(\hat{Q}_h \chi, \chi) = \sum_{j=1}^{N_h} \gamma_j (e_j, \chi)^2.$$

Proof. Noting that α_j is decided by the linear system

$$\sum_{j=1}^{N_h} \alpha_j (\varphi_j, \chi)^2 = \sum_{j=1}^{N_h} \gamma_j (e_j, \chi)^2,$$

we have

$$\begin{aligned}(\mathbf{E}(\hat{W}_h \otimes \hat{W}_h)\chi, \chi) &= (t\hat{Q}_h \chi, \chi) = \mathbf{E}(\hat{W}_h, \chi)^2 \\ &= t \sum_{j=1}^{N_h} \alpha_j (\varphi_j, \chi)^2 = t \sum_{j=1}^{N_h} \gamma_j (e_j, \chi)^2 \quad \forall \chi \in S_h. \qquad \square\end{aligned}$$

Acknowledgments. I wish to express my sincere gratitude to my supervisor Dr. Stig Larsson, Chalmers University of Technology, who suggested the topic of this paper, for his support and valuable criticism. I thank Dr. Tony Shardlow for helpful discussions for the numerical illustrations, and also the referees for a number of valuable comments.

REFERENCES

[1] E. J. Allen, S. J. Novosel, and Z. Zhang, *Finite element and difference approximation of some linear stochastic partial differential equations*, Stochastics Stochastics Rep., 64 (1998), pp. 117–142.

[2] F. E. Benth and J. Gjerde, *Convergence rates for finite element approximations of stochastic partial differential equations*, Stochastics Stochastics Rep., 63 (1998), pp. 313–326.

[3] K. Chrysafinos and L. S. Hou, *Error estimates for semidiscrete finite element approximations of linear and semilinear parabolic equations under minimal regularity assumptions*, SIAM J. Numer. Anal., 40 (2002), pp. 282–306.

[4] R. F. Curtain and P. L. Falb, *Ito's lemma in infinite dimensions*, J. Math. Anal. Appl., 31 (1970), pp. 434–448.

[5] R. F. Curtain and P. L. Falb, *Stochastic differential equations in Hilbert space*, J. Differential Equations, 10 (1971), pp. 412–430.

[6] G. Da Prato, *Some results on linear stochastic evolution equations in Hilbert spaces by the semigroups method*, Stochastic Anal. Appl., 1 (1983), pp. 57–88.

[7] G. Da Prato and A. Lunardi, *Maximal regularity for stochastic convolutions in L^p spaces*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 9 (1998), pp. 25–29.

[8] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Cambridge University Press, Cambridge, UK, 1992.

[9] A. M. Davie and J. G. Gaines, *Convergence of numerical schemes for the solution of parabolic stochastic partial differential equations*, Math. Comp., 70 (2001), pp. 121–134.

[10] D. A. Dawson, *Stochastic evolution equations*, Math. Biosci., 15 (1972), pp. 287–316.

[11] Q. Du and T. Zhang, *Numerical approximation of some linear stochastic partial differential equations driven by special additive noises*, SIAM J. Numer. Anal., 40 (2002), pp. 1421–1445.

[12] F. Gozzi, *Regularity of solutions of a second order Hamilton–Jacobi equation and application to a control problem*, Comm. Partial Differential Equations, 20 (1995), pp. 775–826.

[13] W. Grecksch and P. E. Kloeden, *Time-discretised Galerkin approximations of parabolic stochastic PDEs*, Bull. Austral. Math. Soc., 54 (1996), pp. 79–85.

[14] I. Gyöngy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise.* I, Potential Anal., 9 (1998), pp. 1–25.

[15] I. Gyöngy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise.* II, Potential Anal., 11 (1999), pp. 1–37.

[16] I. Gyöngy and D. Nualart, *Implicit scheme for stochastic parabolic partial differential equations driven by space-time white noise*, Potential Anal., 7 (1997), pp. 725–757.

[17] E. Hausenblas, *Numerical analysis of semilinear stochastic evolution equations in Banach spaces*, J. Comput. Appl. Math., 147 (2002), pp. 485–516.

[18] E. Hausenblas, *Approximation for semilinear stochastic evolution equation*, Potential Anal., 18 (2003), pp. 141–186.

[19] P. E. Kloeden and S. Shott, *Linear-implicit strong schemes for Itô–Galerkin approximations of stochastic PDEs*, J. Appl. Math. Stochastic Anal., 14 (2001), pp. 47–53.

[20] G. J. Lord and J. Rougemont, *A numerical scheme for stochastic PDEs with Gevrey regularity*, IMA J. Numer. Anal., 24 (2004), pp. 587–604.

[21] T. Müller-Gronbach and K. Ritter, *Non-uniform time discretization and lower bounds for approximation of stochastic heat equations*, Found. Comput. Math., (2005), to appear.

[22] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer-Verlag, New York, 1983.

[23] S. Peszat and J. Zabczyk, *Stochastic evolution equations with a spatially homogeneous Wiener process*, Stochastic Process. Appl., 72 (1997), pp. 187–204.

[24] J. Printems, *On the discretization in time of parabolic stochastic partial differential equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 1055–1078.

[25] T. Shardlow, *Numerical methods for stochastic parabolic PDEs*, Numer. Funct. Anal. Optim., 20 (1999), pp. 121–145.

[26] T. G. Theting, *Solving Wick-stochastic boundary value problems using a finite element method*, Stochastics Stochastics Rep., 70 (2000), pp. 241–270.

[27] T. G. Theting, *Solving parabolic Wick-stochastic boundary value problems using a finite element method*, Stochastics Stochastics Rep., 75 (2003), pp. 49–77.

[28] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, Berlin, 1997.

[29] J. B. Walsh, *An introduction to stochastic partial differential equations*, in École d'été de probabilités de Saint-Flour. XIV—1984, Lecture Notes in Math. 1180, Springer-Verlag, Berlin, 1986, pp. 265–439.

[30] Y. Yan, *Semidiscrete Galerkin approximation for a linear stochastic parabolic partial differential equation driven by an additive noise*, BIT, 44 (2004), pp. 829–847.

[31] Y. Yan, *Error Analysis and Smoothing Properties of Discretized Deterministic and Stochastic Parabolic Problems*, Ph.D. thesis, Department of Mathematics, Chalmers University of Technology and Göteborg University, Göteborg, Sweden, 2003.

SIAM J. Numer. Anal.
Vol. 43, No. 4, pp. 1385–1406

PSEUDO-TRANSIENT CONTINUATION FOR NONSMOOTH NONLINEAR EQUATIONS [*]

K. R. FOWLER[†] AND C. T. KELLEY[‡]

Abstract. Pseudo-transient continuation is a Newton-like iterative method for computing steady-state solutions of differential equations in cases where the initial data are far from a steady state. The iteration mimics a temporal integration scheme, with the time step being increased as steady state is approached. The iteration is an inexact Newton iteration in the terminal phase.

In this paper we show how steady-state solutions to certain ordinary and differential algebraic equations with nonsmooth dynamics can be computed with the method of pseudo-transient continuation. An example of such a case is a discretized PDE with a Lipschitz continuous, but nondifferentiable, constitutive relation as part of the nonlinearity. In this case we can approximate a generalized derivative with a difference quotient.

The existing theory for pseudo-transient continuation requires Lipschitz continuity of the Jacobian. Newton-like methods for nonsmooth equations have been globalized by trust-region methods, smooth approximations, and splitting methods in the past, but these approaches are not designed to find steady-state solutions of time-dependent problems. The method in this paper synthesizes the ideas from nonsmooth calculus and the method of pseudo-transient continuation.

Key words. pseudo-transient continuation, nonlinear equations, semismooth functions, Clarke differential

AMS subject classifications. 65H10, 65H20, 65L05

DOI. 10.1137/S0036142903431298

1. Introduction.

In this paper we show how pseudo-transient continuation (Ψtc) can be used to solve a class of nonsmooth nonlinear equations. Ψtc is a predictor-corrector method for efficient integration of a time-dependent differential equation to steady state. The objective of the method is not temporal accuracy but rather to resolve the transient behavior of the solution until the iteration is close to steady state, and then to increase the "time step" and transition to a fast Newton-like method.

In this paper we extend the theoretical convergence results of [8, 19] to problems with certain nonsmooth nonlinearities and, thereby, partially explain the results reported in [9, 11]. We also show how generalized derivatives can be approximated by finite differences, and how these approximate derivatives can be used effectively both in locally convergent iterations, such as those which arise in temporal integration, and in the context of Ψtc. This aspect of the work is motivated by several papers on simulation of unsaturated flow [11, 15, 16, 25, 32, 33], in which Lipschitz continuous spline approximations to the non-Lipschitz continuous van Genuchten [35] and Mualem [26] constitutive laws are used. These nonsmooth functions are then differentiated with finite differences as if they were smooth. Our results explain the success reported

[*] Received by the editors July 8, 2003; accepted for publication (in revised form) January 20, 2005; published electronically October 19, 2005. This research was supported in part by Army Research Office grants DAAD19-02-1-0391 and DAAD19-02-1-0111, National Science Foundation grants DMS-0070641, DMS-0209695, and DMS-0404537, and a United States Department of Education GAANN fellowship.

http://www.siam.org/journals/sinum/43-4/43129.html

[†] Department of Mathematics and Computer Science, Clarkson University, Potsdam, NY 13699-5815 (kfowler@clarkson.edu).

[‡] Center for Research in Scientific Computation and Department of Mathematics, North Carolina State University, Box 8205, Raleigh, NC 27695-8205 (Tim_Kelley@ncsu.edu).

in those papers. Another aspect of the paper is an extension of the local results in [10, 22, 28, 30].

Ψtc methods are particularly appropriate for the types of nonsmooth nonlinearities which we discuss in this paper. Traditional methods, such as line searches, for globalizing iterative methods for nonlinear equations can fail as commonly implemented in practice for both smooth and nonsmooth equations [8, 9, 19]. The existing global convergence results for nonsmooth nonlinear equations are based on either line searches for an inexact Newton formulation [10, 23], a sequence of smooth approximations [5, 31], methods based on the one-dimensional secant method [29], or explicit treatment of the nonsmoothness [21]. Only the latter two approaches admit approximation of the generalized Jacobian by a difference, and we use the ideas of [29] in this paper. Ψtc allows one to deal with the nonsmoothness directly and exploits the dynamics to guarantee convergence to x^*, the solution one wants.

In section 2 we review the relevant results from nonsmooth analysis (section 2.1) and Ψtc (section 2.2). Then we describe the setting for the new results. In section 3 we show how finite difference approximations of generalized Jacobians affect the local convergence of inexact Newton methods for nonsmooth problems. We use those results in section 4, where we state and prove our local and global convergence results for Ψtc. We present a numerical example in section 5.

Some extensions of our results to infinite dimensions are possible, using ideas from [6,12,13,20,34] if the appropriate compactness conditions hold. These extensions will be explored in a subsequent paper.

2. Previous results.

In this section we review the prior results about Ψtc and nonsmooth analysis that we will need for this paper.

In this paper the norm will be the scaled discrete l^2 norm on R^N,

$$\|w\| = \frac{1}{\sqrt{N}}\|w\|_2,$$

unless stated otherwise. The ball of radius ϵ about a point $x \in R^N$ will be denoted by

$$\mathcal{B}(x, \epsilon) = \{z \mid \|x - z\| < \epsilon\}.$$

As is standard, we will let x^* denote the solution of $F(x) = 0$, and let

$$e = x - x^*$$

denote the error. We will let $(x)_i$ denote the ith component of the vector x.

2.1. Nonsmooth analysis.

In this section we review the concepts from nonsmooth analysis [7, 24] that we will need for our convergence results. We then state the local convergence result from [22, 30] that we extend in this paper.

Let $F : R^N \to R^N$ be locally Lipschitz continuous. This implies that F is Fréchet differentiable almost everywhere. We let D_F denote the set of points where F is Fréchet differentiable. The generalized Jacobian [7] of F at $x \in R^N$ is the set

$$\partial F(x) = \mathrm{co}\left\{ \lim_{x_j \to x; x_j \in D_F} F'(x_j) \right\}, \tag{2.1}$$

where co denotes the convex hull.

We will consider extensions of the Newton-like iteration

$$x_{n+1} = x_n - V_n^{-1} F(x_n), \tag{2.2}$$

where $V_n \in \partial F(x_n)$. Our results will be stated in terms of an inexact formulation,

$$x_{n+1} = x_n + s, \tag{2.3}$$

where

$$\|V_n s + F(x_n)\| \le \eta_n \|F(x_n)\| \tag{2.4}$$

and $V_n \in \partial F(x_n)$.

The concept of semismoothness [24, 30] is critical for the results in this paper.

Definition 2.1. *F is semismooth at $x \in R^N$ if F is locally Lipschitz continuous and, for all $w \in R^N$, the limit*

$$\lim_{V \in \partial F(x+tw'), w' \to w, t \downarrow 0} \{Vw'\} \tag{2.5}$$

exists.

Semismoothness is a useful concept [6, 30, 34] in the proofs of convergence and local convergence rates of the iteration (2.2). In the standard theory for Lipschitz continuously differentiable F, local quadratic convergence follows from nonsingularity of the Jacobian $F'(x^*)$ at the solution and the Lipschitz continuity of F'. To obtain convergence rates, in the nonsmooth case, one must prove that the Newton iteration is well defined and quantify the degree of nonsmoothness.

Lemma 2.2, taken from [28], and Lemma 2.4 are the results that are needed to prove local superlinear convergence of (2.2).

Lemma 2.2. *F is semismooth at $x \in R^N$ if and only if*

$$\lim_{w \to 0, V \in \partial F(x+w)} \frac{\|F(x+w) - F(x) - Vw\|}{\|w\|} = 0. \tag{2.6}$$

If F is semismooth, then [30] the directional derivatives

$$F'(x : w) = \lim_{h \to 0} \frac{F(x+hw) - F(x)}{h}$$

exist for all $x, w \in R^N$.

To obtain convergence rates one needs a stronger condition than semismoothness [30].

Definition 2.3. *F is semismooth of order p at x if for all $w \in R^N$ and $V \in \partial F(x+w)$*

$$F(x+w) - F(x) - Vw = O(\|w\|^{1+p}) \tag{2.7}$$

as $w \to 0$.

Lemma 2.4. *Let F be semismooth, let $F(x^*) = 0$, and assume that all matrices in $\partial F(x^*)$ are nonsingular. Then there are M and Δ such that if $x \in \mathcal{B}(x^*, \Delta)$ and $V \in \partial F(x)$, then $\|V^{-1}\| \le M$.*

These results have been used to prove several convergence theorems [10, 22, 28, 30] for (2.2) and (2.3). Theorem 2.5 is a combination of these local convergence results and is the basis for the new algorithms in this paper.

THEOREM 2.5. *Let $F : R^N \to R^N$ with $F(x^*) = 0$. Assume that F is semismooth at x^* and that all matrices in $\partial F(x^*)$ are nonsingular. Then there are $\bar{\eta}, \bar{\delta}, K > 0$ such that if $x_0 \in \mathcal{B}(x^*, \bar{\delta})$ and $\eta_n \le \bar{\eta}$, then the inexact Newton iteration* (2.3) *converges to x^* and*

$$\|e_{n+1}\| \le K\eta_n \|e_n\| + o(\|e_n\|).$$

Moreover, if F is semismooth of order p at x^*, then

$$\|e_{n+1}\| \le K(\eta_n \|e_n\| + \|e_n\|^{1+p}).$$

In previous work [12, 17, 20, 21] on nonsmooth nonlinear equations in function spaces and their discretizations, we used properties of the solution to isolate a smooth component of the nonlinearity. Each problem required a different approach, and all assumed that the nonsmooth component was small. In those papers, mesh-independent convergence results were obtained, and standard implementations of matrix-free Newton–Krylov methods worked well.

The formulation we consider in this paper is different. Here one does not have to explicitly split the operator into smooth and nonsmooth parts, a significant advantage in complicated applications [32]. However, we know of no general proofs of mesh-independent convergence rates, a problem also mentioned in [34]. In fact, the numerical results in section 5 show mesh-dependent performance of the iteration, especially in the midrange. Mesh-dependent convergence was also reported in [5]. In section 5.4 we illustrate this phenomenon and show how a nested iteration can overcome it.

Numerical differentiation, a topic we consider in section 3.1, is a simple matter if the smooth and nonsmooth parts of the nonlinearity are split. Here, we use methods from [29], which can prove accuracy only if one is differentiating in coordinate directions, and only then for special classes of operators. Since the directions in a matrix-free Newton–Krylov solver are not predictable, neither the results in [29] nor our results apply to those methods.

2.2. Pseudo-transient continuation. The objective of Ψtc, as we present it here, is to find the steady-state solution of the semiexplicit index-one differential algebraic equation (DAE)

$$D \begin{pmatrix} u \\ v \end{pmatrix}' = - \begin{pmatrix} f(u,v) \\ g(u,v) \end{pmatrix} = -F(x), \quad x(0) = x_0. \tag{2.8}$$

Here $x = (u, v)^T \in C([0, \infty], R^{N_1+N_2})$. The functions $u : [0, \infty] \to R^{N_1}$ and $v : [0, \infty] \to R^{N_2}$ are to be found. The differential variables u and the algebraic variables v are clearly separated in the semiexplicit case where

$$D = \begin{pmatrix} D_{11} & 0 \\ 0 & 0 \end{pmatrix},$$

where D_{11} is a nonsingular scaling matrix. A good general reference for DAEs is [4].

We assume that the initial data for (2.8) are consistent (i.e., $g(u(0), v(0)) = 0$) and seek the solution x^* to $F(x^*) = 0$ that satisfies

$$\lim_{t\to\infty} x(t) = x^*.$$

If (2.8) is a discretization in space of a PDE and the initial data are far from the desired steady state, the application of a conventional method, such as a line search [18], to the time-independent equation

$$F(x) = 0$$

may fail to converge. Possible failure modes [9] are stagnation of the iteration at a singularity of F', the Jacobian of F, and finding a solution other than x^*.

We formulate Ψtc as

$$x_{n+1} = x_n - (\delta_n^{-1} D + F'(x_n))^{-1} F(x_n). \tag{2.9}$$

In (2.9), $\{\delta_n\}$ is adjusted to efficiently find the steady-state solution rather than to enforce temporal accuracy.

The convergence results in [8, 19] assume that the time step was updated with "switched evolution relaxation" (SER) [27], i.e.,

$$\delta_n = \max\left(\delta_{n-1} \frac{\|F(x_{n-1})\|}{\|F(x_n)\|}, \delta_{max}\right). \tag{2.10}$$

In [8] the authors prove convergence for smooth F under the assumptions that the DAE has index one in a certain uniform sense, that it has a global solution in time, and that the solution converges to a steady state. The convergence result for the exact $\eta = 0$ case is

$$\|x_{n+1} - x^*\| = O(\|x_n - x^*\|(\delta_{max}^{-1} + \|x_n - x^*\|)) \tag{2.11}$$

as $n \to \infty$.

In this paper we relax the smoothness assumptions and consider the iteration

$$x_{n+1} = x_n + s, \tag{2.12}$$

where

$$\|(\delta_n^{-1} D + V(x_n))s + F(x_n)\| \leq \eta_n \|F(x_n)\|, \tag{2.13}$$

where $V(x_n)$ is near to the set $\partial F(x_n)$ in the sense that

$$V(x_n) \in \mathcal{D}(F, x_n, C, h), \text{ for some small } h, \tag{2.14}$$

where, for $x \in R^N$ and $h \geq 0$,

$$\mathcal{D}(F, x, C, h) = \{V \mid \|V - \bar{V}\| \leq Ch \text{ for some } \bar{V} \in \partial F(\bar{x}) \text{ and } \|x - \bar{x}\| \leq h\}. \tag{2.15}$$

The sense in which $V(x_n)$ is close to $\partial F(x_n)$ is technical because $\partial F(x)$ is a set-valued function of x. The requirement that $V(x_n) \in \mathcal{D}(F, x_n, C, h)$ is, in a sense, a requirement that a combination of the forward and backward errors be small.

In [8,9,11,19] we report on computational results that show that both the local and global phases of the Ψtc iteration perform as (2.11) predicts even if the nonlinearity is nonsmooth [8, 9, 11] and the derivative is approximated by differencing [11, 14, 15, 32].

3. Local convergence theory. In this section we analyze the accuracy of a finite difference approximation of a generalized Jacobian in the case where the nonsmoothness arises from a substitution operator. The approximation is accurate in a combined forward and backward sense, and this affects not only the convergence speed of an inexact Newton iteration but also the limiting accuracy.

3.1. Finite difference approximations. The results in this paper are motivated in part by our experience with nonsmooth nonlinear substitution operators. A substitution operator on R^N has the form

$$\Phi(x) = (\phi(x_1), \ldots, \phi(x_N))^T, \tag{3.1}$$

where $\phi : R \rightarrow R$. The generalized Jacobian of Φ is the set of diagonal matrices

$$\partial\Phi(x) = (\partial\phi(x_1), \ldots, \partial\phi(x_N))^T.$$

In this section, we consider maps that are compositions of smooth maps with substitution operators. Let

$$G(x) = S(\Phi(x)),$$

where S is continuously differentiable. S is then strictly differentiable in the sense of [7], and hence the definition (2.1) of ∂G implies (see Theorem 2.6.6 and its corollary in [7]) that

$$\partial G(x) = S'(\Phi(x))\partial\Phi(x),$$

where S' is the Jacobian of S. Of interest here is the approximation of ∂G with a finite difference approximation using the coordinate directions.

Let $\partial_h^F G(x)$ be the matrix whose ith column is

$$\frac{G(x + h1_i) - G(x)}{h},$$

where 1_i is the unit vector in the ith coordinate direction. We show that there is $C > 0$ such that the forward difference $\partial_h^F G$ approximates $\partial G(x)$ in the sense described by (2.15).

Theorem 3.1 can be derived from Lemma 2.3 of [29], which considers one-sided difference approximations and derivatives by taking averages. We give a direct proof which allows us to exhibit $\bar{x}$ as a function of x.

THEOREM 3.1. *Let $\phi : R \rightarrow R$ be Lipschitz continuous and differentiable except at finitely many points $\{\xi_i\}_{i=1}^M$. Let S be Lipschitz continuously differentiable in R^N. Then there is $C > 0$ such that for all h sufficiently small*

$$\partial_h^F G(x) \in \mathcal{D}(G, x, C, h). \tag{3.2}$$

Proof. We begin by showing that it suffices to prove the result for scalar functions. Differentiability of S and the Lipschitz continuity of ϕ imply that

$$\begin{aligned} G(x + h1_i) - G(x) &= S'(\Phi(x))(\Phi(x + h1_i) - \Phi(x)) + O(h^2) \\ &= S'(\Phi(\bar{x}))(\Phi(x + h1_i) - \Phi(x)) + O(h^2) \end{aligned}$$

for all $\bar{x}$ such that $\|\bar{x} - x\| \leq h$. Hence we need only prove the result for substitution operators.

Since Φ is a substitution operator, the ith component of $\Phi(x + h1_i) - \Phi(x)$ is

$$\phi((x)_i + h) - \phi((x)_i),$$

and we need only consider scalar functions. Now let

$$h < \min_{i,j} \|\xi_i - \xi_j\|_\infty;$$

then at most one ξ is in the interval $[(x)_i, (x)_i + h]$. If ϕ is differentiable in the interval $[(x)_i, (x)_i + h]$, then

$$\frac{\phi((x)_i + h) - \phi((x)_i)}{h} = \phi'((x)_i) + O(h)$$

and we let the ith component of $\bar{x}$ be $(x)_i$.

Now assume that $\xi_j \in [(x)_i, (x)_i + h]$ for some j. Let $\phi'_+(\xi_j)$ and $\phi'_-(\xi_j)$ be the right- and left-handed derivatives at ξ_j:

$$\phi'_\pm(\xi_j) = \lim_{h\to 0} \frac{\phi(\xi_j \pm h) - \phi(\xi_j)}{\pm h}.$$

Let $\xi_j - (x)_i = \nu h$ for $\nu \in [0, 1]$. Then

$$\begin{aligned} \phi((x)_i + h) - \phi((x)_i) &= \phi(\xi_j + (1-\nu)h) - \phi(\xi_j - \nu h) \\ &= \phi(\xi_j + (1-\nu)h) - \phi(\xi_j) + \phi(\xi_j) - \phi(\xi_j - \nu h) \\ &= (1-\nu)\phi'_+(\xi_j) + \nu\phi'_-(\xi_j) + O(h^2). \end{aligned}$$

Since

$$(1-\nu)\phi'_+(\xi_j) + \nu\phi'_-(\xi_j) \in \partial\phi(\xi_j)$$

for all $\nu \in [0, 1]$, the proof is complete with $(\bar{x})_i = \xi_j$. □

A similar result holds for central differences. Let $\partial_h^C G(x)$ be the matrix whose ith column is

$$\frac{G(x + h1_i) - G(x - h1_i)}{2h}.$$

If S is Lipschitz continuously twice differentiable and ϕ is piecewise Lipschitz continuously twice differentiable, then the statement of Theorem 3.1 with

$$\|V - \bar{V}\| \le Ch$$

in (2.15) is replaced by

$$\|V - \bar{V}\| \le Ch^2. \tag{3.3}$$

3.2. Local convergence. If the generalized Jacobian is approximated by a finite difference, one cannot expect asymptotic convergence, because the accuracy in the terminal phase of the iteration will be limited by the accuracy in the derivative. We quantify this in Theorem 3.2, which extends the existing local convergence theorems for inexact Newton methods for semismooth equations. The new assumption that $V(x) \in \mathcal{D}(F, x, C, h)$ is motivated by the results in section 3.1.

THEOREM 3.2. *Assume that F is semismooth at x^*, that $F(x^*) = 0$, and that all matrices in $\partial F(x^*)$ are nonsingular. Assume that there is $C > 0$ such that*

$$V(x) \in \mathcal{D}(F, x, C, h) \tag{3.4}$$

for all x sufficiently near x^*.

Then there is ϵ such that if $x_0 \in \mathcal{B}(x^*, \epsilon)$, $\{\eta_n\}$, and h are sufficiently small, then the iteration

$$x_{n+1} = x_n + s, \tag{3.5}$$

where

$$\|V(x_n)s + F(x_n)\| \le \eta \|F(x_n)\|, \tag{3.6}$$

converges to x^*. Moreover, there is $K > 0$ such that

$$\|e_{n+1}\| \le K((\eta_n + h)\|e_n\| + h) + o(\|e_n\|), \tag{3.7}$$

or, if F is semismooth of order p at x^*, then

$$\|e_{n+1}\| \le K((\eta_n + h)\|e_n\| + \|e_n\|^{1+p} + h). \tag{3.8}$$

Proof. The plan of the proof is to compare x_{n+1} with the Newton iteration from $\bar{x}_n$, where $\bar{x}_n$ is the point specified in the definition of $\mathcal{D}$. We can then apply Theorem 2.5.

Let ϵ and h be small enough so that

$$\|V^{-1}\| \le M \text{ for all } V \in \partial F(x) \text{ and } x \in \mathcal{B}(x^*, h + \epsilon), \tag{3.9}$$

which we can do by Lemma 2.4. We assume that $x_n \in \mathcal{B}(x^*, \epsilon)$ and will show, reducing ϵ and h if necessary, that (3.8) holds and therefore that $x_{n+1} \in \mathcal{B}(x^*, \epsilon)$.

By assumption, there are $\overline{x}_n \in \mathcal{B}(x_n, h)$ and $\bar{V}_n \in \partial F(\overline{x}_n)$ such that

$$\|V(x_n) - \bar{V}_n\| \le Ch.$$

Hence the step s is nearly an inexact Newton step from $\bar{x}_n$.

By (3.9), for h sufficiently small,

$$\|V(x_n)^{-1}\| \le 1/(M^{-1} - Ch) \le 2M,$$

and hence

$$\|s\| \le 2M(\eta_n + 1)\|F(x_n)\|.$$

Therefore,

$$\begin{aligned} \|\bar{V}_n s + F(\bar{x}_n)\| &\le \|V(x_n)s + F(\bar{x}_n)\| + Ch\|s\| \\ &\le \|V(x_n)s + F(x_n)\| + \|F(x_n) - F(\bar{x}_n)\| + Ch\|s\| \\ &\le \eta_n\|F(x_n)\| + Lh + Ch(2M(1 + \eta_n)\|F(x_n)\|), \end{aligned} \tag{3.10}$$

where L is the Lipschitz constant of F. Since

$$\|F(x_n)\| \le \|F(\bar{x}_n)\| + Lh,$$

we may set

$$K_0 = 1 + 2L + 2MC(1 + \eta_n) \le 1 + 2L + 4MC$$

and obtain

$$\|\bar{V}_n s + F(\bar{x}_n)\| \leq K_0((\eta_n + h)\|F(\bar{x}_n)\| + h). \tag{3.11}$$

The inexact Newton condition (3.6) and (3.11) imply that

$$x_{n+1} = x_n + s = \bar{x}_n - \bar{V}_n^{-1}(F(\bar{x}_n) + r_n),$$

where

$$r_n = \bar{V}_n^{-1}(\bar{x}_n - x_n) - (\bar{V}_n s + F(\bar{x}_n)).$$

Hence

$$\|\bar{V}_n^{-1} r_n\| \leq MK_0((\eta_n + h)\|F(\bar{x}_n)\| + h) + h,$$

and

$$\begin{aligned}\|e_{n+1}\| &= \|\bar{e}_n - \bar{V}_n^{-1}F(\bar{x}_n) + \bar{V}_n^{-1} r_n\| \\ &\leq MK_0((\eta_n + h)\|F(\bar{x}_n)\| + h) + h + o(\|\bar{e}_n\|).\end{aligned}$$

If F is semismooth of order p at x^*, then (2.7) implies that there is $K_1 > 0$ such that

$$\begin{aligned}\|e_{n+1}\| &= \|\bar{e}_n - \bar{V}_n^{-1}F(\bar{x}_n) + \bar{V}_n^{-1} r_n\| \\ &\leq K_1\|\bar{e}_n\|^{1+p} + MK_0((\eta_n + h)\|F(\bar{x}_n)\| + h) + h.\end{aligned}$$

Since $\|F(\bar{x}_n)\| \leq L\|\bar{e}_n\|$ and $\|\bar{e}_n\| \leq \|e_n\| + h$, we obtain (3.8) with

$$K = 2K_1 + MK_0(1 + L) + 1$$

and complete the proof. □

3.3. Optimal choice of *h*. If $\bar{x}_n \neq x_n$, then the estimates (3.7) and (3.8) do not imply convergence but stagnation once the error is $O(h)$. This is analogous to convergence results [18] for Newton's method when there are errors, such as floating point roundoff, in the evaluation of F. In this case, however, h is larger than floating point roundoff, and we can combine Theorems 3.1 and 3.2 to estimate the optimal choice of h.

Suppose that F is piecewise C^1 (and hence semismooth of order 1 [24]) and can be evaluated up to an absolute error of ϵ_F. If we incorporate the error in F into the result of Theorem 3.1 in the standard way [18], we obtain

$$\partial_h^F F(x) \in \mathcal{D}(G, x, C', h'),$$

where $h' = O(h + \epsilon_F/h)$. Then the estimate (3.8) becomes

$$\|e_{n+1}\| \leq K((\eta_n + h + \epsilon_F/h)\|e_n\| + \|e_n\|^2 + h). \tag{3.12}$$

If we solve the equation for the step exactly, then $\eta_n = 0$. In that case, if $\|e_n\| = O(h^{1/2})$, then

$$\|e_{n+1}\| = O\left(\frac{\epsilon_F}{h^{1/2}} + h\right). \tag{3.13}$$

The term on the right-hand side of (3.13) is minimized when

$$h = O(\epsilon_F^{2/3}). \tag{3.14}$$

If, for example, $\epsilon_F \approx 10^{-15}$ is double precision floating point roundoff, (3.14) would say that the best results would be obtained if $h \approx 10^{-10}$, rather than 10^{-8} as a conventional analysis would predict. We provide numerical evidence for this in section 5.3.

4. Convergence of Ψtc.

The analysis of Ψtc in this paper follows the pattern of [8, 19], considering the iteration in two phases. For phase one, the initial or global phase, we show that Ψtc is a consistent convergent scheme for integration of the DAE. The scheme will be first order if F is semismooth of order 1, order p if F is semismooth of order $p < 1$, and convergent but with no order if F is merely semismooth.

From the analysis of the global phase we will conclude that, for sufficiently small δ_0, the iteration will approach x^*. For the second local phase of the iteration, we show that if x is near x^* and $\{\delta_n\}$ is bounded away from zero, then $\delta_n \to \delta_{max}$, and the terminal phase of convergence can be described by the results in section 3.

The analysis of the local phase does not depend on the dynamics, and we will defer the detailed assumptions on the DAE until section 4.2.

4.1. Local phase.

We consider the local phase first, as we did in [8, 19], in order to establish targets for the integration in the global phase. We seek to find ϵ_L so that if $x_0 \in \mathcal{B}(x^*, \epsilon_L)$ and $\{\delta_n\}$ remains bounded away from zero, then $\{x_n\}$ and $\{\delta_n\}$ in (2.12) satisfy $x_n \to x^*$ and $\delta_n \to \delta_{max}$.

The local convergence rates in the terminal phase depend on the following assumption.

ASSUMPTION 4.1. *F is semismooth at x^*. There are $C, h, \beta, \epsilon_L > 0$ such that for all $x \in \mathcal{B}(x^*, \epsilon_L)$ and all $\delta > 0$*

$$\|(D + \delta V(x))^{-1} D\| \leq 1/(1 + \beta\delta) \tag{4.1}$$

and

$$V(x) \in \mathcal{D}(F, x, C, h).$$

Note that Assumption 4.1 does not imply that every element in $\partial F(x^*)$ is nonsingular. It is an assumption on the particular element $V(x) \in \mathcal{D}(F, x, C, h)$, and the possibility of a singular matrix in $\partial F(x^*)$ is left open.

THEOREM 4.1. *Let the assumptions of Theorem 3.2 and Assumption 4.1 hold. Let $\{\delta_n\}$ be given by (2.10). Then there are C_T and ϵ_T such that if $\{\eta_n\}$ is sufficiently small and $x_0 \in \mathcal{B}(x^*, \epsilon_T)$, then either $\inf_n \delta_n = 0$ or $\delta_n \to \delta_{max}$, the Ψtc iteration converges, and, for n sufficiently large,*

$$\|e_{n+1}\| \leq C_T((\eta_n + \delta_n^{-1} + h)\|e_n\| + h) + o(\|e_n\|) \tag{4.2}$$

or, if F is semismooth of order p,

$$\|e_{n+1}\| \leq C_T(\|e_n\|^{1+p} + (\eta_n + \delta_n^{-1} + h)\|e_n\| + h). \tag{4.3}$$

Proof. We assume that x_0 is near enough to x^* so that the conclusions of Theorem 3.2 hold. If $x_n \in \mathcal{B}(\epsilon_T)$, then, following the proof of Theorem 3.2,

$$e_{n+1} = e_n - (\delta_n^{-1} D + \bar{V}_n)^{-1} F(\bar{x}_n) + r_n,$$

where

$$\|r_n\| = O((\eta_n + h)\|F(\bar{x}_n)\| + h).$$

Semismoothness and our assumptions imply that

$$F(\bar{x}_n) - \bar{V}_n e_n = O(h) + o(\|e_n\|)$$

and hence

$$\begin{aligned} e_{n+1} &= e_n - (\delta_n^{-1} D + \bar{V}_n)^{-1} \bar{V}_n e_n + R_n \\ &= (\delta_n^{-1} D + \bar{V}_n)^{-1} \delta_n^{-1} D e_n + R_n, \end{aligned}$$

where

$$R_n = O((\eta_n + h)\|F(\bar{x}_n)\| + h) + o(\|e_n\|).$$

If $\delta_n > \delta^*$ for all n, then Assumption 4.1 implies that

$$\|(\delta_n^{-1} D + \bar{V}_n)^{-1} \delta_n^{-1} D\| < 1/(1 + \beta\delta^*).$$

This implies that the iteration is q-linearly convergent, and hence $\delta_n \to \delta_{max}$ and $x_n \to x^*$.

The completion of the proof for large δ_n is a direct consequence of Theorem 3.2, since the inexact Newton condition

$$\|(\delta_n^{-1} D + V(x_n))s + F(x_n)\| \leq \eta_n \|F(x_n)\|$$

implies that there is C_h such that

$$\|V(x_n)s + F(x_n)\| \leq (\eta_n + C_h h)\|F(x_n)\| + \delta_n^{-1}\|D_{11}\|\|s\|,$$

and then C and η_n in (3.10) can be replaced by $C + \|D_{11}\|$ and $\eta_n + \delta_n^{-1} + C_h h$. This implies convergence if δ_{max} is sufficiently large. □

4.2. Global phase. In the analysis of the global phase we must assume that the Ψtc iteration is, for small δ, a stable explicit method for the DAE (2.8). To do this we must assume that the DAE is consistent and has index one. In the smooth case, one can express this in terms of the nonsingularity of g_v, the Jacobian of g with respect to the algebraic variables. In the nonsmooth case, however, one must take the limit in (2.1) in all components together. This means that the index assumption is more technical, using the nonsingularity of the matrix pencil $\delta^{-1}D + V(x)$ in part 7 of Assumption 4.2.

We assume that $V(x) \in \mathcal{D}(F, x, C, h)$ for a sufficiently small h. We decompose operators $V \in \partial F$ into blocks

$$V(x) = \begin{pmatrix} V_{uu} & V_{uv} \\ V_{vu} & V_{vv} \end{pmatrix}, \tag{4.4}$$

where $V_{uu} \in \partial_u f$, …, $V_{vv} \in \partial_v g$.

With this in mind we can formulate our assumptions on the dynamics. Define a neighborhood of the trajectory from x_0 as

$$S(\epsilon) = \{z \mid \inf_{t \geq 0} \|z - x(t)\| \leq \epsilon\}. \tag{4.5}$$

ASSUMPTION 4.2. $g(u_0, v_0) = 0$*; i.e., the initial values* (u_0, v_0) *are consistent.*

There are $\epsilon_G \in (0, \epsilon_T/2)$*, where* ϵ_T *is the radius from Theorem* 4.1*, such that the following hold:*

1. F *is semismooth in* $S(\epsilon_G)$*.*
2. *For all* $z_0 \in S(\epsilon_S)$*, the solution of* $Dz' = -F(z), z(0) = z_0$ *exists,* $z(t) \in S(\epsilon_G)$ *for all* t*, and* $\lim_{t\to\infty} z(t) = x^*$*.*

3. *$V(x) \in \mathcal{D}(F, x, C, h)$ for all $x \in S(\epsilon_G)$.*
4. *$V_{vv}(x)$ is nonsingular for all $x \in S(\epsilon_G)$, and there is M_V such that $\|V_{vv}(x)^{-1}\| \leq M_V$ for all $x \in S(\epsilon_S)$.*
5. *There is M_I such that for all h sufficiently small, $\delta > 0$, $x \in S(\epsilon_G)$, and $V \in \mathcal{D}(F, x, C, h)$, $D_{11} + \delta V_{uu}$ is nonsingular and*
$$\|(D_{11} + \delta V_{uu})^{-1}\| \leq M_I.$$
6. *There is $M_D > 0$ such that for all $x \in S(\epsilon_G)$*
$$\|V(x)\| \leq M_D.$$
7. *$(\delta^{-1} D + V(x))$ is nonsingular for all $x \in S(\epsilon_G)$, and $\delta > 0$.*

We analyze the global phase by showing that the global truncation error of the scheme

$$x_{n+1} = x_n - (\delta_n^{-1} D + V(x_n))^{-1} F(x_n)$$

is of order p, i.e.,

$$\|x_n - x(t_n)\| = O(\delta_{max}^p),$$

where $\delta_{max} = \max_{0 \leq m \leq n} \delta_m$. Since the SER formula implies that $\delta_{max} = O(\delta_0)$, this will imply, similarly to [8, 19], that if δ_0 is sufficiently small, then the Ψtc iteration will correctly track the solution until x_n is in the ball of local convergence required by Theorem 4.1.

We will use a simple consequence of pth order semismoothness.

LEMMA 4.2. *Let Assumption* 4.2 *hold. Let $x(t)$ be the solution to* (2.8). *Let $\delta > 0$ and let*

$$\sigma = \begin{pmatrix} \sigma^u \\ \sigma^v \end{pmatrix} = x(t + \delta) - x(t).$$

Then, for δ, h sufficiently small,

$$(D + \delta V(x(t)))\sigma = -\delta F(x(t)) + O(\delta h) + o(\delta^2), \tag{4.6}$$

and if F is semismooth of order p,

$$(D + \delta V(x(t)))\sigma = -\delta F(x(t)) + O(\delta^{2+p} + \delta h), \tag{4.7}$$

uniformly in t.

Proof. In the interest of brevity, we will give the proof for $h = 0$ and F semismooth of order p. The analysis for $h > 0$ and semismooth F follows the outlines of the proofs of Theorems 3.2 and 4.1.

Write $x(t) = (u(t), v(t))^T$. By integrating the DAE (2.8), we see that u is a Lipschitz continuous function of t. The semismoothness of F and the nonsingularity of V_{vv} (from parts 1 and 4 of Assumption 4.2) imply that v is a Lipschitz continuous function of u and hence also a Lipschitz continuous function of t. This Lipschitz continuity implies that

$$\|\sigma\| = O(\delta),$$

uniformly in t. Integrate (2.8) over the interval $[t, t+\delta]$ and use the Lipschitz continuity of F to obtain

$$D\sigma = -\int_t^{t+\delta} F(x(\tau))\,d\tau = -\delta F(x(t+\delta)) + O(\delta^2), \tag{4.8}$$

uniformly in t.

By the definition of semismoothness with $x = x(t+\delta)$ and $w = -\sigma$, we have, for δ sufficiently small,

$$\begin{aligned} F(x(t+\delta)) &= F(x(t)) + V(x(t))\sigma + O(\|\sigma\|^{1+p}) \\ &= F(x(t)) + V(x(t))\sigma + O(\delta^{1+p}). \end{aligned} \tag{4.9}$$

The estimate (4.9) is uniform in t because the set $\{x(t)\,|\,t \geq 0\}$ is compact.

Hence, multiplying (4.9) by δ and substituting into (4.8),

$$(D + \delta V(x(t)))\sigma = -\delta F(x(t)) + O(\delta^{2+p}),$$

as asserted. □

Lemma 4.2 will imply convergence of Ψtc in the same way as in the smooth case [8]. The objective is to show that for δ_0 sufficiently small, the Ψtc iteration remains in the tube $S(\epsilon_G)$. We will give a proof that is more general than the one in [8] in that it uses Assumption 4.2 rather than the stronger one in [8].

In the analysis we will let $\|\cdot\|$ denote the Euclidean norm on any of $R^N = R^{N_1+N_2}$, R^{N_1}, or R^{N_2}. The dimension will be clear from the context.

THEOREM 4.3. *Let Assumptions* 4.1 *and* 4.2 *hold. Then if* δ_0, $\{\eta_n\}$, *and* h *are sufficiently small and* $\{\delta_n\}$ *is bounded from below, then* $x_n \to x^*$ *and* (4.2) *holds. If* F *is semismooth of order* p, *then* (4.3) *holds.*

Proof. We begin with the special case in which F is semismooth of order p and $\eta_n = 0$ for all n. Part 2 of Assumption 4.2 implies that there is T such that $x(t) \in \mathcal{B}(x^*, \epsilon_T/2)$, the ball of local convergence from Theorem 4.1, for all $t \geq T$.

Let

$$t_n = \sum_{k=0}^{n} \delta_k.$$

We will use Lemma 4.2 to show that if δ_0 is sufficiently small and the sequence $\{\delta_n\}$ is bounded from below, then the Ψtc iteration is an accurate integrator for (2.8) in the sense that

$$\|x_n - x(t_n)\| = O(\delta^p + h), \tag{4.10}$$

where $\delta = \max_{0\leq k\leq n} \delta_k$. Hence, we can select δ_0 and h such that $x_n \in S(\epsilon_G)$ until $t_n > T$.

We begin by dividing (2.9) and (4.7) into the u and v components. We set

$$e_n = x_n - x(t_n) = \begin{pmatrix} e_n^u \\ e_n^v \end{pmatrix}, \; s_n = x_{n+1} - x_n = \begin{pmatrix} s_n^u \\ s_n^v \end{pmatrix}$$

and

$$\sigma_n = x(t_{n+1}) - x(t_n) = \begin{pmatrix} \sigma_n^u \\ \sigma_n^v \end{pmatrix}.$$

First note that if $V \in \mathcal{D}(F, x_n, C, h)$, then $V \in \mathcal{D}(F, x(t_n), C, h + \|e_n\|)$. Hence if $\|e_n\|$ is sufficiently small, Lemma 4.2 implies that

$$(D + \delta_n V)\sigma_n = -\delta F(x(t_n)) + O(\delta_n^{2+p} + \delta_n(h + \|e_n\|)). \tag{4.11}$$

We write the u component of (2.9) as

$$(D_{11} + \delta_n V_{uu})s_n^u + V_{uv}s_n^v = f(u_n, v_n), \tag{4.12}$$

and the u component of (4.11) as

$$(D_{11} + \delta_n V_{uu})\sigma_n^u + \delta_n V_{uv}\sigma_n^v = -f(u(t_n), v(t_n)) + O(\delta_n^{2+p} + \delta_n(h + \|e_n\|)). \tag{4.13}$$

Subtract (4.13) from (4.12) and use part 6 of Assumption 4.2 to obtain

$$(D_{11} + \delta_n V_{uu})(e_{n+1}^u - e_n^u) = O(\delta_n(\|e_n\| + \|e_{n+1}^v\|) + \delta_n^{2+p} + \delta_n h). \tag{4.14}$$

Hence, we may use part 5 of Assumption 4.2 to obtain

$$\|e_{n+1}^u\| \leq \|e_n^u\| + O(\delta_n(\|e_n\| + \|e_{n+1}^v\|) + \delta_n^{2+p} + \delta_n h). \tag{4.15}$$

The v component of (2.9) is

$$V_{vu}s_n^u + V_{vv}s_n^v = -g(u_n, v_n) = -V_{uv}e_n^u - V_{vv}e_n^v + O(\delta_n^{1+p} + h\delta_n + \|e_n\|\delta_n), \tag{4.16}$$

where we use the facts that $g(u(t), v(t)) = 0$ and $V \in \mathcal{D}(F, x(t_n), C, h + \|e_n\|)$ in the last equality. Adding $V_{uv}e_n^u + V_{vv}e_n^v$ to both sides of (4.16) and noting that $s_n = e_{n+1} - e_n$ yields

$$V_{vv}e_n^v = -V_{vu}e_{n+1}^u + O(\delta_n^{1+p} + \delta_n h + \delta_n\|e_n\|). \tag{4.17}$$

Parts 4 and 6 from Assumption 4.2 and (4.17) imply that

$$\|e_{n+1}^v\| = O(\|e_{n+1}^u\| + \delta_n\|e_n\| + \delta_n^{1+p} + \delta_n h). \tag{4.18}$$

Equations (4.15) and (4.18) together imply that there is $L > 0$ such that

$$\begin{gathered} \|e_{n+1}^u\| \leq (1 + \delta_n L)\|e_n\| + \delta_n L\|e_{n+1}^v\| + O(\delta_n^{1+p} + \delta_n h) \\ \text{and} \\ \|e_{n+1}^v\| \leq L\|e_{n+1}^u\| + \delta_n\|e_n\| + O(\delta_n^{1+p} + \delta_n h). \end{gathered} \tag{4.19}$$

So, if we define a new norm on R^N by

$$|||(u, v)^T||| = L\|u\| + \|v\|,$$

then

$$|||e_{n+1}||| = |||e_n||| + O(\delta_n|||e_n||| + \delta_n^{1+p} + \delta_n h). \tag{4.20}$$

Our assumption that $\{\delta_n\}$ is bounded from below implies that there is n^* such that $t_{n^*} = T^* \geq T$. Then, as is standard in the analysis of numerical methods for initial value problems [1], we may let $\delta_* = \max_{0 \le n \le n^*} \delta_n$ and use (4.20) to conclude that, for $0 \le n \le n^*$,

$$|||e_n||| = O(\delta_*^p + h),$$

which proves (4.10), because $||| \cdot |||$ is equivalent to the Euclidean norm on R^N.

If F is semismooth and $\{\eta_n\}$ is nonzero, then (4.10) becomes (see [19])

$$\|x_n - x(t_n)\| = O\left(h + \sum_{n=0}^{n^*} \eta_n \delta_n \right) + o(1) \text{ as } \delta_0 \to 0, \tag{4.21}$$

and the convergence result still holds if, say, $\eta_n = O(\delta_*)$ for all $0 \le n \le n^*$. □

5. Numerical example.

We illustrate the results with a simple one-dimensional example taken from [2,3,5]. This example is sufficient to illustrate the convergence results in this paper and allows us to refine the grids to a degree that was not possible in the two- and three-dimensional results that motivated this paper [11,15,16,25,32, 33].

We use direct methods to compute the Newton step in this section, so $\eta_n = 0$. In all but section 5.3, we compute $V \in \partial F(x)$ analytically, so $h = 0$ in those computations.

This example, taken from [5], is a Lipschitz reformulation of the boundary value problem [2,3]

$$-u_{zz} + \lambda \max(0, u)^p = 0, \quad z \in (0,1),$$

with boundary data

$$u(0) = u(1) = 0$$

and $p \in (0,1)$.

The reformulation adds a new variable

$$v = \begin{cases} u^p & \text{if } u \ge 0, \\ u & \text{if } u < 0 \end{cases}$$

to obtain a Lipschitz continuous elliptic-algebraic system, $F(x) = 0$, where $x = (u,v)^T$ and

$$F(x) = \begin{pmatrix} f(u,v) \\ g(u,v) \end{pmatrix} = \begin{pmatrix} -u_{zz} + \lambda \max(0,v) \\ u - \omega(v) \end{pmatrix} = 0, \tag{5.1}$$

where

$$\omega(v) = \begin{cases} v^{1/p} & \text{if } v \ge 0, \\ v & \text{if } v < 0. \end{cases}$$

We use SER (2.10) to control the sequence $\{\delta_n\}$ and use

$$D = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$$

in (2.9).

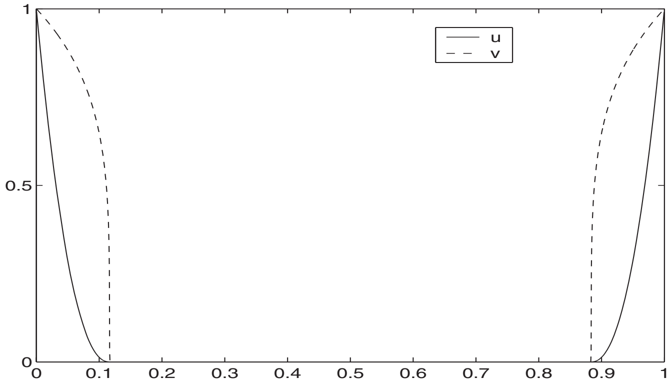


FIG. 5.1. *Solution.*

The reason we formulate the problem with DAE (rather than ODE) dynamics is that the pseudo-time variable should not be added to both equations in (5.1) but to only the first. The reason for this is that the true time-dependent system is

$$u_t = u_{zz} - \lambda \max(0, u)^p$$

and that the auxiliary variable v is used only to make the nonlinearity Lipschitz continuous. One might think that an ODE formulation would work equally well, but, in fact, the ODE formulation, which does not model the physics, failed to converge in our testing.

We discretize the problem with central differences, using a difference increment of δ_z. The nonsmooth nonlinearity is a substitution operator, and its generalized Jacobian is a set of diagonal matrices.

We report on several computations with $p = 0.1$ and $\lambda = 200$. This choice leads to a large "dead core" [2, 3], a region in which the solution vanishes. We plot the solution in Figure 5.1.

$\delta_0 = 1$ and $\delta_{max} = 10^6$ for all the computations. We terminate the nonlinear iteration when either

$$\|F(x_n)\|/\|F(x_0)\| < 10^{-13} \quad \text{or} \quad \|s_n\| < 10^{-10}, \tag{5.2}$$

where $s_n = x_{n+1} - x_n$. In the tables we see the superlinear convergence clearly in the reduction in the norms of the steps; this is consistent with the estimate $s_n = -e_n + o(\|e_n\|)$ which follows from local superlinear convergence. The superlinear convergence is less visible in the residual norms, because the generalized Jacobians become more ill-conditioned as the mesh is refined. The residual norms begin to stagnate after a reduction of 10^{12}.

5.1. Exact computation of the generalized Jacobian. For the results in this section we compute the generalized Jacobian analytically. If we let L_{δ_z} be the discretized Laplacian, we can write

$$F(x) = \begin{pmatrix} f(u,v) \\ g(u,v) \end{pmatrix} = \begin{pmatrix} -L_{\delta_z} u \\ u - v - \max(0,v)^{1/p} \end{pmatrix} + \begin{pmatrix} \lambda \\ 1 \end{pmatrix} \max(0, v). \tag{5.3}$$

In the above and in the discussion that follows, functions of vectors, $\max(0, v)^{1/p}$, for example, are understood to mean componentwise evaluation.

We use the known result for the scalar function $\max(0,v)$,

$$\partial \max(0,v) = \begin{cases} 0 & \text{if } v < 0, \\ [0,1] & \text{if } v = 0, \\ 1 & \text{if } v > 0, \end{cases}$$

to obtain

$$\partial F = \begin{pmatrix} -L_{\delta_z} & 0 \\ 1 & -1-(1/p)\max(0,v)^{(1-p)/p} \end{pmatrix} + \begin{pmatrix} 0 & \lambda \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & \partial \max(0,v). \end{pmatrix} \tag{5.4}$$

The notation should be clear. The 2, 2 blocks in the matrices denote multiplication operators in the continuous case, and diagonal matrices in the discrete case. We use this notation because of its compactness and close connection to the original differential equation.

The calculations in this section use $V(x_n) \in \partial F(x_n)$. We may use any choice from the set-valued map $\partial \max(0,v)$ and we choose $V \in \partial F$ using

$$\chi(v) = \left\{ \begin{array}{ll} 0 & \text{if } v \leq 0, \\ 1 & \text{if } v > 0 \end{array} \right\} \in \partial \max(0,v).$$

Had we used $\chi(v) = 1$ when $v = 0$, then $\mu(v)$ would vanish when $\mu = 0$, leading to singularity of V_{vv}. Note that the global convergence result does not require that all elements of V_{vv} be nonsingular for all choices of $V \in \partial F$, only that we choose one for which it is. The local convergence result does require that all elements of ∂F be nonsingular, as they are.

In Figure 5.2 we plot the norms of the steps and nonlinear residuals together with the growth of δ for a mesh of width $\delta_z = 1/2048$. δ grows smoothly in the early phase of the iteration and reaches its maximum rapidly. The superlinear convergence is clearly visible in the curve for the norms of the steps. The Jacobian of the nonlinear residual has a condition number of $O(1/h^2)$, and hence the residual norm reflects the error less accurately.

5.2. Verification of the assumptions. We will now explore verification of Assumptions 4.1 and 4.2 for the case $h = 0$. The case $h \neq 0$ is similar, requiring only the addition of an $O(h)$ perturbation to ∂F. We have not verified part 2 of Assumption 4.2, though we have done numerical experiments that indicate its validity. F is clearly semismooth, and we are using $V \in \partial F$ for all x. So parts 1 and 3 in Assumption 4.2 hold trivially.

Let

$$V(x) = \begin{pmatrix} -L_{\delta_z} & 0 \\ 1 & -1-(1/p)\max(0,v)^{(1-p)/p} \end{pmatrix} + \begin{pmatrix} 0 & \lambda \\ 0 & 1 \end{pmatrix} \chi(v).$$

Since the set $S(\epsilon_G)$ is compact, $\|v\|$ is bounded on $S(\epsilon_G)$, which implies part 6 of Assumption 4.2.

Now V_{vv} is the operator of componentwise multiplication by

$$-\mu(v) = -1-(1/p)\max(0,v)^{(1-p)/p} + \chi(v).$$

This operator is negative semidefinite. To see this, note that

$$\mu(v) = 1+(1/p)\max(0,v)^{(1-p)/p} - \chi(v) = \left\{ \begin{array}{ll} (1/p)\max(0,v)^{(1-p)/p} & \text{if } v > 0, \\ 1 & \text{if } v \leq 0 \end{array} \right\} > 0.$$

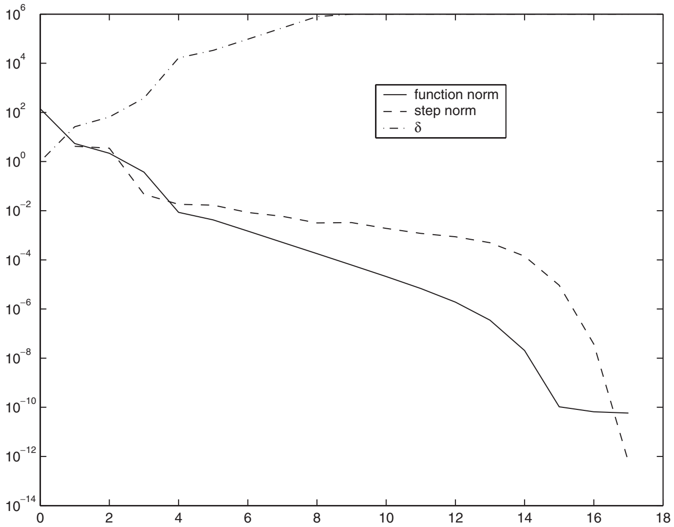


FIG. 5.2. *Analytic generalized Jacobian.*

To verify part 4 of Assumption 4.2, we observe that the smallest nonzero component of $v(t)$ is bounded away from zero. This follows from the boundary conditions on the differential equation and the compactness of $S(\epsilon_G)$. Hence

$$V_{vv}\eta = -\mu(v)\eta$$

has a uniformly bounded inverse, as assumed. We will now use part 4 of Assumption 4.2 in the verification of the remaining assumptions.

We now verify the bounds on the inverse, (4.1) in Assumption 4.1 and part 5 of Assumption 4.2. As part of that process, we will also verify the nonsingularity assumption, part 7 of Assumption 4.2.

We seek to solve

$$(D + \delta V)(\xi, \eta)^T = (\phi, \psi)^T,$$

which we write as two equations:

$$\begin{aligned} &(I - \delta L_{\delta_z})\xi + \lambda\chi(v)\eta = \phi, \\ &\delta\xi + \delta V_{vv}\eta = \psi. \end{aligned} \tag{5.5}$$

We may eliminate η by using the second equation in (5.5),

$$\eta = -V_{vv}^{-1}(\xi - \delta^{-1}\psi), \tag{5.6}$$

and hence

$$(I - \delta L_{\delta_z} - \lambda\chi(v)V_{vv}^{-1})\xi = \phi - \delta^{-1}\lambda\chi(v)V_{vv}^{-1}\psi. \tag{5.7}$$

Since $-\delta L_{\delta_z}$ is positive definite and $\delta, \lambda, \mu, \chi \geq 0$, we have

$$\|(I - \delta L_{\delta_z} - \lambda\chi(v)V_{vv}^{-1})^{-1}\| \leq \|(I - \delta L_{\delta_z})^{-1}\| \leq \|I\| = 1. \tag{5.8}$$

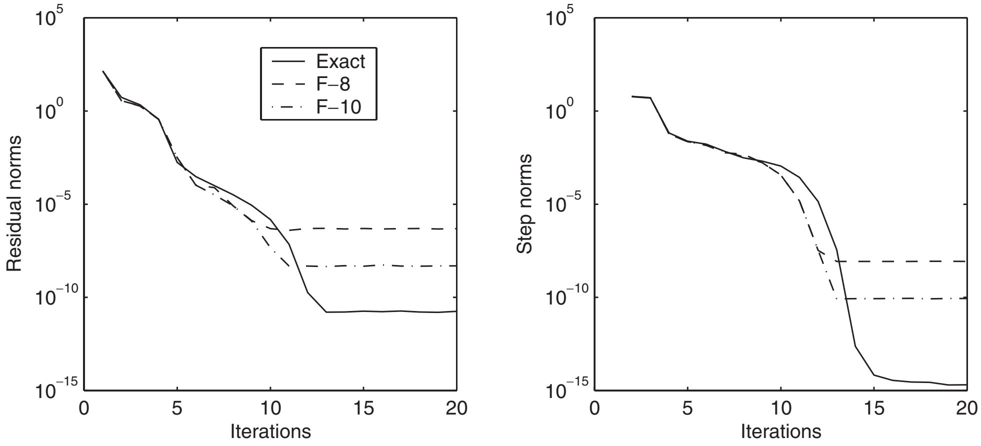


FIG. 5.3. *Norms of the steps and residuals.*

Hence we may solve for ξ, proving nonsingularity of $(D + \delta V)$.

The bound on V_{vv} and (5.8) imply

$$\|\xi\| \leq (\|\phi\| + \delta^{-1}\lambda M_V \|\psi\|). \tag{5.9}$$

Combining (5.9) with (5.6) completes the verification of part 5 of Assumption 4.2. To verify (4.1), we set $\psi = 0$ and use the second inequality in (5.8).

5.3. Computation of the generalized Jacobian by differences. For the results in this section we compute the generalized Jacobian with several choices of differences. The results were similar for all the meshes. In Figure 5.3 we plot residual and step norm histories for

- analytic generalized Jacobian (Exact),
- forward differences, increment 10^{-8} (F-8), and
- forward differences, increment 10^{-10} (F-10)

for a mesh of width $\delta_z = 1/2048$ and 20 iterations. In this way we can clearly see the point at which the iteration stagnates. As we predicted in section 3.3, the iteration is more accurate when the difference increment is $10^{-10} \approx \epsilon^{2/3}$ than it is with the standard choice of $10^{-8} \approx \epsilon^{1/2}$.

5.4. Mesh dependence and nested iteration. We used the analytic ∂F (5.4) in the computations reported in this section.

In Figure 5.4 we plot the progress of the iteration for mesh sizes of $1/128, 1/512$, and $1/2048$, terminating the iteration when $\|s\| < 10^{-13}$. In this way we can examine the dependence of the convergence on the mesh width. While the convergence in the early phase is identical for the three meshes and superlinear in the terminal phase, the global convergence becomes slower as the mesh is refined.

Nested iteration or grid sequencing means solving the problem to high precision on a coarse mesh, interpolating to a finer mesh in such a way that the interpolation error can be corrected with a few (e.g., one) iterations and continuing this until one has a solution on a target, finest mesh. We set $\delta = 10^6$ for the finer meshes, under the assumption that we are in the locally convergent phase of the iteration.

For this example, one would hope not only to eliminate the mesh-dependency in the iteration history that one sees in Figure 5.4, but also to approximate the solution up to truncation error at each level.

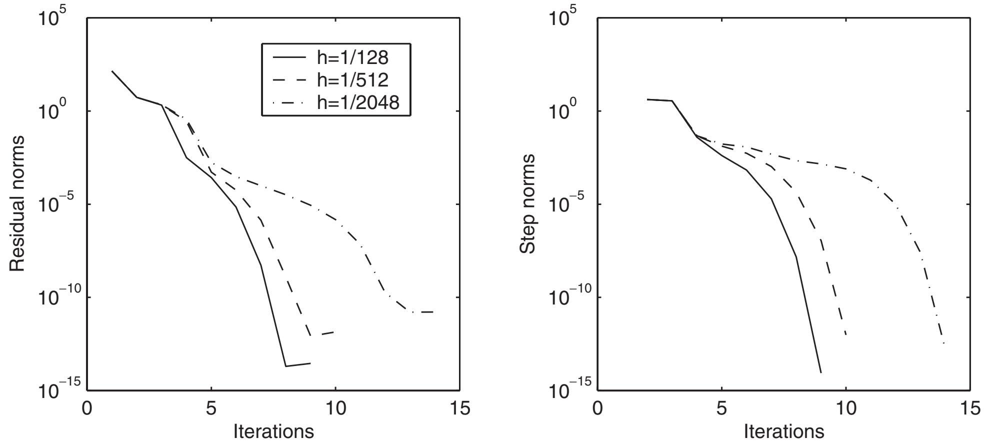


FIG. 5.4. *Mesh-dependence of convergence.*

TABLE 5.1
Step norms: $p = 0.1$, nested iteration.

$n\backslash\delta_z$	1/64	1/128	1/256	1/512	1/1024	1/2048
0	4.20e+00	2.02e-02	1.02e-02	5.72e-03	3.45e-03	3.61e-03
1	3.53e+00	1.13e-02	1.23e-03	1.13e-03	2.14e-03	6.16e-04
2	3.91e-02	8.95e-04	1.56e-04	2.15e-04	1.58e-04	1.37e-05
3	4.11e-03	6.44e-05	2.19e-06	6.18e-06	7.24e-05	7.14e-08
4	6.89e-04	3.26e-07				2.23e-12
5	1.94e-05					
6	1.47e-08					

TABLE 5.2
Step norms: $p = 0.5$, nested iteration.

$n\backslash\delta_z$	1/64	1/128	1/256	1/512	1/1024	1/2048
0	1.32e+00	1.52e-03	3.87e-04	9.74e-05	2.44e-05	6.13e-06
1	5.29e-01	4.37e-05	7.89e-06	1.39e-06	2.47e-07	4.25e-08
2	5.20e-03	1.10e-06	9.73e-08	4.75e-08		4.21e-09
3	6.59e-05		3.83e-09			1.76e-11
4	2.73e-05					
5	9.74e-08					

This was a successful strategy. However, the results must be interpreted in light of the continuity properties of the solution. $u^* \in C^2[0,1]$, and hence $u^* \in H^2[0,1]$. However, if $p < 1/2$, $v^* = (u^*)^p \notin H^2[0,1]$. This means that linear interpolation will not approximate v^* to second order if $p < 1/2$. To partially address this, we interpolate u from the coarse to fine mesh with linear interpolation and then compute v as

$$v = \max(0, u)^p.$$

This give us a better initial approximation of v on the finer mesh than directly interpolating v, but not, as Table 5.1 shows, a second order accurate one.

In Tables 5.1 and 5.2 we report the residual and step norms on a sequence of meshes $\{2^{-n}\}_{n=6}^{11}$ for $p = 0.1$ and $p = 0.5$. The initial steps at each mesh reflect both the error in the initial iterate and the truncation error in the interpolation.

The iterations for both values of p show that we have recovered mesh independence in the sense that the iteration requires a roughly constant number of steps to terminate at each level. The table for $p = 0.5$ clearly shows second order convergence. The interpolation error for $p = 0.1$ is visible in the sizes of the initial steps.

Acknowledgments. The authors thank Xiaojun Chen for several useful conversations. We also appreciate the efforts of Liqun Qi and two anonymous referees in making us aware of [29] and clearing up several of our misconceptions about semismooth functions.

REFERENCES

[1] U. M. Ascher and L. R. Petzold, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM, Philadelphia, 1998.

[2] A. K. Aziz, A. B. Stephens, and M. Suri, *Numerical methods for reaction-diffusion problems with non-differentiable kinetics*, Numer. Math., 53 (1988), pp. 1–11.

[3] J. W. Barrett and R. M. Shanahan, *Finite element approximation of a model reaction-diffusion problem with a non-Lipschitz nonlinearity*, Numer. Math., 59 (1991), pp. 217–242.

[4] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics Appl. Math. 14, SIAM, Philadelphia, 1995.

[5] X. Chen, *A superlinearly and globally convergent method for reaction and diffusion problems with a non-Lipschitzian operator*, in Topics in Numerical Analysis, Comput. Suppl. 15 Springer-Verlag, Vienna, 2001, pp. 79–90.

[6] X. Chen, Z. Nashed, and L. Qi, *Smoothing methods and semismooth methods for nondifferentiable operator equations*, SIAM J. Numer. Anal., 38 (2000), pp. 1200–1216.

[7] F. H. Clarke, *Optimization and Nonsmooth Analysis*, Classics Appl. Math. 5, SIAM, Philadelphia, 1990.

[8] T. S. Coffey, C. T. Kelley, and D. E. Keyes, *Pseudotransient continuation and differential-algebraic equations*, SIAM J. Sci. Comput., 25 (2003), pp. 553–569.

[9] T. S. Coffey, R. J. McMullan, C. T. Kelley, and D. S. McRae, *Globally convergent algorithms for nonsmooth nonlinear equations in computational fluid dynamics*, J. Comput. Appl. Math., 152 (2003), pp. 69–81.

[10] F. Facchinei, A. Fischer, and C. Kanzow, *Inexact Newton methods for semismooth equations with applications to variational inequality problems*, in Nonlinear Optimization and Applications, G. D. Pillo and F. Giannessi, eds., Plenum Press, New York, 1996, pp. 125–139.

[11] M. W. Farthing, C. E. Kees, T. Coffey, C. T. Kelley, and C. T. Miller, *Efficient steady-state solution techniques for variably saturated groundwater flow*, Advances in Water Resources, 26 (2003), pp. 833–849.

[12] M. Heinkenschloss, C. T. Kelley, and H. T. Tran, *Fast algorithms for nonsmooth compact fixed-point problems*, SIAM J. Numer. Anal., 29 (1992), pp. 1769–1792.

[13] M. Hintermüller, K. Ito, and K. Kunisch, *The primal-dual active set strategy as a semismooth Newton method*, SIAM J. Optim., 13 (2003), pp. 865–888.

[14] K. R. Kavanagh, *Nonsmooth Nonlinearities in Applications from Hydrology*, Ph.D. thesis, North Carolina State University, Raleigh, NC, 2003.

[15] K. R. Kavanagh, C. T. Kelley, R. C. Berger, J. P. Hallberg, and S. E. Howington, *Nonsmooth nonlinearities and temporal integration of Richards' equation*, in Computational Methods in Water Resources XIV, Vol. 2, S. M. Hassanizadeh, R. J. Schotting, W. G. Gray, and G. F. Pinder, eds., Elsevier, Amsterdam, 2002, pp. 947–954.

[16] C. E. Kees and C. T. Miller, *Higher order time integration methods for two-phase flow*, Advances in Water Resources, 25 (2002), pp. 159–177.

[17] C. T. Kelley, *Identification of the support of nonsmoothness*, in Large Scale Optimization: State of the Art, W. W. Hager, D. W. Hearn, and P. Pardalos, eds., Kluwer Academic Publishers, Boston, 1994, pp. 192–205.

[18] C. T. Kelley, *Iterative Methods for Solving Linear and Nonlinear Equations*, Frontiers Appl. Math. 16, SIAM, Philadelphia, 1995.

[19] C. T. Kelley and D. E. Keyes, *Convergence analysis of pseudo-transient continuation*, SIAM J. Numer. Anal., 35 (1998), pp. 508–523.

[20] C. T. Kelley and E. W. Sachs, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Sci. Comput., 15 (1994), pp. 645–667.

[21] C. T. Kelley and E. W. Sachs, *A trust region method for parabolic boundary control problems*, SIAM J. Optim., 9 (1999), pp. 1064–1081.

[22] J. Martinez and L. Qi, *Inexact Newton methods for solving nonsmooth equations*, J. Comput. Appl. Math., 60 (1995), pp. 127–145.

[23] J. M. Martinez, *Quasi-Newton methods for solving underdetermined nonlinear simultaneous equations*, J. Comput. Appl. Math., 34 (1991), pp. 171–190.

[24] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SiAM J. Control Optim., 15 (1977), pp. 959–972.

[25] C. T. Miller, G. A. Williams, C. T. Kelley, and M. D. Tocci, *Robust solution of Richards' equation for non-uniform porous media*, Water Resources Research, 34 (1998), pp. 2599–2610.

[26] Y. Mualem, *A new model for predicting the hydraulic conductivity of unsaturated porous media*, Water Resources Research, 12 (1976), pp. 513–522.

[27] W. Mulder and B. V. Leer, *Experiments with implicit upwind methods for the Euler equations*, J. Comput. Phys., 59 (1985), pp. 232–246.

[28] J.-S. Pang and L. Qi, *Nonsmooth equations: Motivation and algorithms*, SIAM J. Optim., 3 (1993), pp. 443–465.

[29] F. A. Potra, L. Qi, and D. Sun, *Secant methods for semismooth equations*, Numer. Math., (1998), pp. 305–324.

[30] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Programming, 58 (1993), pp. 353–367.

[31] L. Qi and X. Chen, *A globally convergent successive approximation method for severely nonsmooth equations*, SIAM J. Control Optim., 33 (1995), pp. 402–418.

[32] K. Staheli, J. H. Schmidt, and S. Swift, *Guidelines for Solving Groundwater Problems with ADH*, US Army Waterways Experiment Station, Vicksburg, MS, 1998.

[33] M. D. Tocci, C. T. Kelley, C. T. Miller, and C. E. Kees, *Inexact Newton methods and the method of lines for solving Richards' equation in two space dimensions*, Comput. Geosci., 2 (1998), pp. 291–310.

[34] M. Ulbrich, *Semismooth Newton methods for operator equations in function spaces*, SIAM J. Optim., 13 (2003), pp. 805–841.

[35] M. T. van Genuchten, *Predicting the hydraulic conductivity of unsaturated soils*, Soil Sci. Soc. Am. J., 44 (1980), pp. 892–898.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1407–1435

NUMERICAL PERIODIC NORMALIZATION FOR CODIM 1 BIFURCATIONS OF LIMIT CYCLES*

YU. A. KUZNETSOV[†], W. GOVAERTS[‡], E. J. DOEDEL[§], AND A. DHOOGE[‡]

Abstract. Explicit computational formulas for the coefficients of the periodic normal forms for all codim 1 bifurcations of limit cycles in generic autonomous ODEs are derived. They include second-order coefficients for the fold (limit point) bifurcation, as well as third-order coefficients for the flip (period-doubling) and Neimark–Sacker (torus) bifurcations. The formulas are independent of the dimension of the phase space and involve solutions of certain boundary-value problems on the interval $[0,T]$, where T is the period of the critical cycle, as well as multilinear functions from the Taylor expansion of the right-hand sides near the cycle. The formulas allow us to distinguish between sub- and supercritical bifurcations, in agreement with earlier asymptotic expansions of the bifurcating solutions. Our formulation makes it possible to use robust numerical boundary-value algorithms based on orthogonal collocation, rather than shooting techniques, which greatly expands its applicability. The actual implementation is described in detail. We include three numerical examples, in which codim 2 singularities are detected along branches of codim 1 bifurcations of limit cycles as zeros of the periodic normal form coefficients.

Key words. normal forms, limit cycles, bifurcations

AMS subject classifications. 34C20, 37G15, 37M20, 65L07

DOI. 10.1137/040611306

1. Introduction.

Isolated periodic orbits (limit cycles) of smooth differential equations

$$\dot{u} = f(u,\alpha), \quad u \in \mathbb{R}^n, \ \alpha \in \mathbb{R}^m, \tag{1.1}$$

play an important role in applications. In generic systems of the form (1.1) depending on one control parameter (i.e., with $m = 1$) a hyperbolic limit cycle exists for an open interval of parameter values α. At a boundary of such an interval, the limit cycle may not exist, degenerating into an equilibrium or an orbit homoclinic to an equilibrium or another nonhyperbolic limit cycle (see, for example, [3]). We do not consider such cases here, instead focusing on those where the cycle does exist at the boundary parameter values but loses its hyperbolicity due to the presence of a nontrivial multiplier μ, with $|\mu| = 1$.

The codim 1 bifurcations of limit cycles in generic systems (1.1) are well understood (see, for example, [8, 10, 3]). Let $u_0(t)$ be a periodic solution (with minimal period) corresponding to a limit cycle Γ of (1.1). The standard approach to the theoretical and numerical analysis of local bifurcations of limit cycles is based on *Poincaré maps*: Given a transversal section Σ to Γ at $u_0(0)$, such a map assigns to each point y of Σ close to $u_0(0)$ another point $\mathcal{P}(y,\alpha)$, where the orbit of (1.1) starting

*Received by the editors July 8, 2004; accepted for publication (in revised form) January 25, 2005; published electronically October 19, 2005.

http://www.siam.org/journals/sinum/43-4/61130.html

[†]Mathematical Institute, Utrecht University, Budapestlaan 6, P.O. Box 80010, 3508 TA Utrecht, The Netherlands, and Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia (Yu.Kuznetsov@math.uu.nl).

[‡]Department of Applied Mathematics and Computer Science, University of Gent, Krijgslaan 281-S9, B-9000, Gent, Belgium (Willy.Govaerts@UGent.be, Annick.Dhooge@UGent.be).

[§]Department of Computer Science, Concordia University, 1455 de Maisonneuve Blvd. West, Montreal, QC, H3G 1M8, Canada (Doedel@cs.concordia.ca).

at y intersects Σ again close to $u_0(0)$. In local coordinates in Σ, the Poincaré map will be represented by a smooth map $\mathcal{P} : \mathbb{R}^{n-1} \times \mathbb{R}^m \to \mathbb{R}^{n-1}$. The cycle corresponds to a fixed point y_0 of $\mathcal{P}(\cdot, \alpha)$; the eigenvalues of its linearization at the fixed point are the nontrivial multipliers of the periodic solution. Once the Poincaré map is introduced, the theory of local bifurcations of maps can be applied.

It is well known (see, for example, [8, 10, 1] for the general theory and [3] for computational formulas) that in generic smooth one-parameter families of maps

$$y \mapsto \mathcal{P}(y, \alpha), \quad y \in \mathbb{R}^{n-1},\ \alpha \in \mathbb{R}^1, \tag{1.2}$$

only the following three bifurcations of fixed points occur:

(1) *The fold.* The fixed point y_0 has a simple eigenvalue $\lambda_1 = 1$ and no other eigenvalues on the unit circle, while the restriction of (1.2) to a one-dimensional center manifold $\mathcal{W}^c(y_0)$ at the critical parameter value has the form $\xi \mapsto \xi + \tilde{b}\xi^2 + O(\xi^3)$, where $\tilde{b} \neq 0$. At the critical parameter value, two fixed points coalesce. This bifurcation is often called a *saddle-node bifurcation*, a *fold*, or a *limit point (LP)*, since two periodic solutions of (1.1) collide and disappear when the parameter passes the critical value. If $\mathcal{A}v = \mathcal{P}_y v$ and $\mathcal{B}(u, v) = \mathcal{P}_{yy}[u, v]$ are evaluated at the critical fixed point y_0, then

$$\tilde{b} = \frac{1}{2}\langle q^*, \mathcal{B}(q, q)\rangle, \tag{1.3}$$

where $\mathcal{A}q = q$, $\mathcal{A}^{\mathrm{T}}q^* = q^*$, and $\langle q^*, q\rangle = 1$. Here and in what follows, $\langle u, v\rangle = u^{\mathrm{H}}v = \bar{u}^{\mathrm{T}}v$ is the standard scalar product in an appropriate complex (or real) finite-dimensional vector space; here, $\mathbb{R}^{n-1}$. It should also be noted that the coefficient $\tilde{b}$ is not uniquely defined but depends on the normalization of q. A similar remark holds for all other normal form coefficients.

(2) *The flip.* The fixed point y_0 has a simple eigenvalue $\lambda_1 = -1$ and no other eigenvalues on the unit circle, while the restriction of (1.2) to a one-dimensional center manifold $\mathcal{W}^c(y_0)$ at the critical parameter value can be transformed to the normal form $\xi \mapsto -\xi + \tilde{c}\xi^3 + O(\xi^4)$, where $\tilde{c} \neq 0$. For nearby parameter values, a cycle of period 2 bifurcates from the fixed point. This is a *period-doubling (PD)* of the periodic solution of (1.1); i.e., there are nearby periodic solutions of approximately double (minimal) period. If $\mathcal{C}(u, v, w) = \mathcal{P}_{yyy}[u, v, w]$ is evaluated at y_0, then

$$\tilde{c} = \frac{1}{6}\langle p^*, \mathcal{C}(p, p, p) + 3\mathcal{B}(p, (I_{n-1} - \mathcal{A})^{-1}\mathcal{B}(p, p))\rangle, \tag{1.4}$$

where I_{n-1} is the $(n-1) \times (n-1)$ identity matrix, $\mathcal{A}p = -p$, $\mathcal{A}^{\mathrm{T}}p^* = -p^*$, and $\langle p^*, p\rangle = 1$.

(3) *The Neimark–Sacker (NS) bifurcation.* The fixed point y_0 has simple critical eigenvalues $\lambda_{1,2} = e^{\pm i\theta}$ and no other eigenvalues on the unit circle. Assume that

$$e^{iq\theta} - 1 \neq 0, \quad q = 1, 2, 3, 4 \ \text{ (no strong resonances)}.$$

Then the restriction of (1.2) to a two-dimensional center manifold $\mathcal{W}^c(y_0)$ at the critical parameter value can be transformed to the normal form $\eta \mapsto \eta e^{i\theta}(1 + \tilde{d}|\eta|^2) + O(|\eta|^4)$, where η is a complex variable and $\tilde{d}$ is a complex number. Further assume that Re $\tilde{d} \neq 0$. Under the above assumptions, a unique *closed invariant curve* around the fixed point appears when the parameter crosses the critical value. This curve corresponds to an *invariant torus*, on which the flow of (1.1) contains periodic or quasi-periodic motions. One has the following expression for $\tilde{d}$:

$$\tilde{d} = \frac{1}{2}e^{-i\theta}\langle v^*, \mathcal{C}(v,v,\bar{v}) + 2\mathcal{B}(v,(I_{n-1}-\mathcal{A})^{-1}\mathcal{B}(v,\bar{v})) + \mathcal{B}(\bar{v},(e^{2i\theta}I_{n-1}-\mathcal{A})^{-1}\mathcal{B}(v,v))\rangle, \tag{1.5}$$

where $\mathcal{A}v = e^{i\theta}v,\ \mathcal{A}^{\mathrm{T}}v^* = e^{-i\theta}v^*$, and $\langle v^*, v\rangle = 1$.

Although existing software (e.g., CONTENT [11]) can compute the normal form coefficients at codim 1 bifurcations of fixed points of general maps using (1.3)–(1.5), application of these capabilities to limit cycle analysis has been limited. One reason for this is the necessity to compute the Poincaré map (1.2) and its derivatives $(\mathcal{A}, \mathcal{B}, \mathcal{C}, \mathcal{D}, \dots)$ numerically. Finite differences work for low-dimensional nonstiff systems (1.1), where they allow the approximation of the Jacobian matrix $\mathcal{A} = \mathcal{P}_y$ with reasonable accuracy. However, this approach fails when the cycle has multipliers with a very big or very small modulus, i.e., when (1.1) is stiff. In general, finite-difference approximations of higher-order partial derivatives (i.e., $\mathcal{B}, \mathcal{C}, \mathcal{D}, \dots$) have very low accuracy due to loss of significant digits and are therefore unreliable. Better results for computing $\mathcal{B}$ and higher-order derivative tensors can be achieved by simultaneous numerical integration of the periodic solution and the corresponding variational equations over the period.

This method has been successfully used in [6] to compute normal form coefficients of the Poincaré map at the fold-flip bifurcation in a four-dimensional atmosphere circulation model. An interesting alternative to numerical integration of the variational equations is to compute the higher-order derivatives of the Poincaré map $\mathcal{P}$ by *automatic differentiation* [12] of the (for example, C-) code used to compute the Poincaré map; see [13, 14]. Both approaches, however, rely on the possibility of accurately finding the periodic solution by shooting, which is not always the case.

There are at least two approaches to the analysis of the limit cycle bifurcations that are not directly based on the Poincaré map and its derivatives. Since it is known which periodic solutions can bifurcate at generic codim 1 bifurcations of limit cycles, one can compute the Taylor series for the period $T(\varepsilon)$, for the corresponding parameter $\alpha(\varepsilon)$, and for the bifurcating solution itself, as functions of the solution amplitude ε. The solvability of the linear systems is guaranteed by the *Fredholm alternative*. This approach, which is conceptually similar to the *Lyapunov–Schmidt method*, has been successfully applied to all codim 1 bifurcations of limit cycles in [5, Chapter XI]. The resulting asymptotic expressions use the derivatives of the right-hand side of (1.1) with respect to u and α and involve solutions to linear boundary-value problems (BVPs) (on the interval $[0,T]$ in the LPC (limit point of cycles) and NS cases and on the interval $[0,2T]$ for the PD bifurcation). They allow one to distinguish between sub- and supercritical bifurcations. However, these formulas are rather involved—in particular, for the NS case, where one has to distinguish between various subharmonic and quasi-periodic solutions—and to our knowledge they have not been implemented in bifurcation software.

There is another theoretical approach [8] for the analysis of limit cycle bifurcations in (1.1), which avoids the Poincaré map reduction. First, in a neighborhood of Γ in $W^c(\Gamma)$, *normal coordinates* can be chosen so that the restricted system (1.1) becomes a nonautonomous T-*periodic* system in $\mathbb{R}^{n-1}$. This periodic system can be considered as an autonomous system with one cyclic variable (mod T). Near the bifurcation, this system can be restricted to an (n_c+1)-dimensional invariant *center manifold*

$W^c(\Gamma)$, thus giving a periodic n_c-dimensional system of ODEs. One can then apply (in general, $2T$-) periodic coordinate transformations to this system and write it as the sum of an autonomous n_c-dimensional normal form and higher-order periodic terms. The autonomous part of this *periodic normal form* allows one to study local and global bifurcations of (1.1) near the critical cycle. This approach is very useful for the theoretical analysis of limit cycle bifurcations (see [8] and [4] for normal forms for some codim 2 cases).

Since the late 1980s, an improvement of the latter approach is known [15], which combines the computation of the center manifold with the normalization of the ODEs restricted to this manifold. This technique leads to simple formulas for the computation of normal form coefficients in two codim 1 cases of equilibrium bifurcations in ODEs (derived earlier with the Lyapunov–Schmidt method), as well as in all five codim 2 cases (see [16]). Although a similar normalization technique was introduced in [17] for time-periodic systems and in [18, 7] for limit cycle bifurcations, it has remained mainly a theoretical tool up to now. There are no numerical algorithms for the computation of the coefficients of the normal forms on $W^c(\Gamma)$ that are based on this approach and that have been implemented in available bifurcation software.

Below we derive a powerful numerical normalization tool based on this technique. In a sense, we combine the periodic normal forms derived in [18] with the Fredholm alternative used in [5]. It should be noted that the idea to apply Fredholm's solvability condition to compute the normal form coefficients for time-periodic systems can be traced back to [17]. The main difference between our approach and that of [17] and [18] is that we avoid Fourier series solutions of the linear BVPs, instead solving them numerically using orthogonal collocation for discretization as in AUTO [19]. This leads to simple and explicit algorithms for the normal form coefficients. A further simplification occurs because we consider only the critical normal forms, and therefore we do not need derivatives of $f(u, \alpha)$ with respect to α. Our results fully agree with the asymptotic expansions for the bifurcating solutions derived in [5].

This paper is organized as follows. In section 2 we fix notation and formulate the periodic normalization on the center manifold. Then we apply this technique to derive explicit formulas to compute the critical normal form coefficients for fold, PD, and torus bifurcations of limit cycles. The formulas are independent of the dimension of the phase space and involve solutions to certain BVPs on the interval $[0, T]$, where T is the period of the critical cycle, as well as multilinear functions from the Taylor expansion of the right-hand sides of (1.1) near the cycle. In section 3 we show that our algorithms fit very well into the BVP-collocation framework of existing continuation software such as AUTO [19], CONTENT [11], and MATCONT [20]. Three numerical examples are given in section 4. Future work is discussed in section 5.

2. Periodic normalization on the center manifold.

Write (1.1) at the critical parameter values as

$$\dot{u} = F(u), \tag{2.1}$$

and suppose that it has a periodic solution $u_0(t) = u_0(t + T)$, where $T > 0$ is its (minimal) period. Develop $F(u_0(t) + v)$ into the Taylor series

$$F(u_0(t) + v) = F(u_0(t)) + A(t)v + \frac{1}{2}B(t; v, v) + \frac{1}{6}C(t; v, v, v) + O(\|v\|^4), \tag{2.2}$$

where

$$A(t)v = F_u(u_0(t))v, \quad B(t, v, v) = F_{uu}(u_0(t))[v, v], \quad C(t; v, v, v) = F_{uuu}(u_0(t))[v, v, v].$$

The multilinear forms A, B, and C are periodic in t with period T.

Consider the initial-value problem for the fundamental matrix solution $Y(t)$, namely,

$$\frac{dY}{dt} = A(t)Y, \quad Y(0) = I_n, \tag{2.3}$$

where I_n is the $n \times n$ identity matrix. The monodromy matrix $M = Y(T)$ always has a "trivial" eigenvalue $\mu_n = 1$. The cycle is hyperbolic if there are no other eigenvalues with $|\mu| = 1$ and is nonhyperbolic otherwise.

The cycle has a fold bifurcation if the eigenvalue $\mu_1 = 1$ of $Y(T)$ corresponds to a two-dimensional Jordan block and if there are no other critical eigenvalues of the monodromy matrix. The cycle has a period-doubling (flip) bifurcation if $\mu_1 = -1$ is simple and is the only nontrivial critical eigenvalue of $Y(T)$. Finally, at an NS (torus) bifurcation, there is a simple pair of nonreal eigenvalues $\mu_{1,2} = e^{\pm i\theta}$, such that $e^{iq\theta} \neq 1$ for $q = 1, 2, 3, 4$ (no strong resonances), and $Y(T)$ has no further critical multipliers other than 1. We will refer to these conditions as the *spectral assumptions*.

To describe the periodic normal forms for the three critical cases mentioned above, we parametrize the corresponding (n_c+1)-dimensional center manifold $W^c(\Gamma)$ near Γ by (τ, ξ), where $\tau \in [0, T]$ or $[0, 2T]$, and ξ is a real or complex coordinate, depending on the bifurcation. It follows from [18] that it is possible to select the ξ-coordinates so that the restriction of (2.1) to the corresponding critical center manifold $W^c(\Gamma)$ will take one of the following *periodic normal forms*.

The periodic normal form at the LPC bifurcation is

$$\left\{ \begin{array}{rcl} \dfrac{d\tau}{dt} & = & 1 - \xi + a\xi^2 + \cdots, \\ \dfrac{d\xi}{dt} & = & b\xi^2 + \cdots, \end{array} \right. \tag{2.4}$$

where $\tau \in [0, T]$, ξ is a real coordinate on $W^c(\Gamma)$ that is transverse to Γ, $a, b \in \mathbb{R}$, and the dots denote nonautonomous T-periodic $O(\xi^3)$-terms. One can show that b and $\tilde{b}$ vanish together, where $\tilde{b}$ is obtained via the Poincaré map reduction and given by (1.3).

The periodic normal form at the PD bifurcation is

$$\left\{ \begin{array}{rcl} \dfrac{d\tau}{dt} & = & 1 + a\xi^2 + \cdots, \\ \dfrac{d\xi}{dt} & = & c\xi^3 + \cdots, \end{array} \right. \tag{2.5}$$

where $\tau \in [0, 2T]$, ξ is a real coordinate on $W^c(\Gamma)$ that is transverse to Γ, $a, c \in \mathbb{R}$, and the dots denote nonautonomous $2T$-periodic $O(\xi^4)$-terms. The coefficient c determines the stability of the critical cycle; if $c \neq 0$, then sign c = sign $\tilde{c}$, where $\tilde{c}$ is obtained via the Poincaré map reduction and given by (1.4).

In the absence of strong resonances, the periodic normal form at the NS bifurcation is

$$\left\{ \begin{array}{rcl} \dfrac{d\tau}{dt} & = & 1 + a|\xi|^2 + \cdots, \\ \dfrac{d\xi}{dt} & = & \dfrac{i\theta}{T}\xi + d\xi|\xi|^2 + \cdots, \end{array} \right. \tag{2.6}$$

where $\tau \in [0,T]$, ξ is a complex coordinate on $W^c(\Gamma)$ that is complementary to τ, $a \in \mathbb{R}, d \in \mathbb{C}$, and the dots denote nonautonomous T-periodic $O(|\xi|^4)$-terms. If Re $d \neq 0$, then sign(Re d) = sign(Re $\tilde{d}$), where $\tilde{d}$ is given by (1.5), obtained via the Poincaré map reduction.

In view of the above, we can assume that a parametrization of the center manifold $W^c(\Gamma)$ has been selected so that the restriction of (2.1) to this manifold has one of the normal forms (2.4), (2.5), or (2.6). The Taylor expansions of T- or $2T$-periodic unknown functions involved in these parametrizations can be found by solving appropriate BVPs on $[0,T]$ or $[0,2T]$, respectively, so that (2.1) restricted to $W^c(\Gamma)$ will have the corresponding periodic normal form. The coefficients a, b, and c arise from the solvability conditions for the BVPs as integrals of scalar products over $[0,T]$, involving quadratic and cubic terms of (2.1) near the periodic solution u_0, as well as the critical eigenfunctions.

The following (or similar) construction will often be used below. Denote by $\mathcal{C}^k([a,b],\mathbb{R}^n)$ the space of k times continuously differentiable functions on $[a,b]$, with values in $\mathbb{R}^n$. Let $\varphi \in \mathcal{C}^1([0,T],\mathbb{R}^n)$ be the only solution of the BVP

$$\left\{\begin{array}{rcl} \dot{\varphi}(\tau) - A(\tau)\varphi(\tau) & = & 0,\ \tau \in [0,T], \\ \varphi(T) - \varphi(0) & = & 0, \\ \int_0^T \langle \varphi(\tau), \varphi(\tau)\rangle d\tau - 1 & = & 0, \end{array}\right.$$

and let $\varphi^* \in \mathcal{C}^1([0,T],\mathbb{R}^n)$ be a nontrivial solution of the adjoint BVP

$$\left\{\begin{array}{rcl} \dot{\varphi}^*(\tau) + A^{\mathrm{T}}(\tau)\varphi^*(\tau) & = & 0,\ \tau \in [0,T], \\ \varphi^*(T) - \varphi^*(0) & = & 0. \end{array}\right. \tag{2.7}$$

If $h \in \mathcal{C}^1([0,T],\mathbb{R}^n)$ is a solution of the singular BVP

$$\left\{\begin{array}{rcl} \dot{h}(\tau) - A(\tau)h(\tau) & = & g(\tau),\ \tau \in [0,T], \\ h(T) - h(0) & = & 0, \end{array}\right. \tag{2.8}$$

then $g \in \mathcal{C}^1([0,T],\mathbb{R}^n)$ satisfies

$$\int_0^T \langle \varphi^*(\tau), g(\tau)\rangle \, d\tau = 0. \tag{2.9}$$

Indeed, taking into account (2.7), we see that this integral equals

$$\int_0^T \langle \varphi^*(\tau), \dot{h}(\tau) - A(\tau)h(\tau)\rangle \, d\tau = -\int_0^T \langle \dot{\varphi}^*(\tau) + A^{\mathrm{T}}(\tau)\varphi^*(\tau), h(\tau)\rangle \, d\tau = 0.$$

We will refer to (2.9) as the *Fredholm solvability condition*. If (2.9) holds, then the problem (2.8) has a unique solution h, satisfying

$$\int_0^T \langle \varphi^*(\tau), h(\tau)\rangle \, d\tau = 0.$$

2.1. The fold bifurcation.

The two-dimensional critical center manifold $W^c(\Gamma)$ at the LPC bifurcation can be parametrized locally by (τ, ξ) as

$$u = u_0(\tau) + \xi v(\tau) + H(\tau, \xi), \quad \tau \in [0,T],\ \xi \in \mathbb{R}, \tag{2.10}$$

where H satisfies $H(T,\xi)=H(0,\xi)$ and has the Taylor expansion

$$H(\tau,\xi)=\frac{1}{2}h_2(\tau)\xi^2+O(\xi^3), \tag{2.11}$$

with $h_2(T)=h_2(0)$, while

$$\left\{\begin{array}{rcl} \dot{v}(\tau)-A(\tau)v(\tau)-F(u_0(\tau)) &=& 0,\ \tau\in[0,T],\\ v(T)-v(0) &=& 0,\\ \int_0^T\langle v(\tau),F(u_0(\tau))\rangle d\tau &=& 0. \end{array}\right. \tag{2.12}$$

The function v exists due to Lemma 2 of [18]. Note that (2.12) implies

$$\int_0^T\langle\varphi^*(\tau),F(u_0(\tau))\rangle\, d\tau=0 \tag{2.13}$$

for any φ^* satisfying (2.7). Moreover, due to the spectral assumptions at the LPC-point, we can also assume that

$$\int_0^T\langle\varphi^*(\tau),v(\tau)\rangle d\tau=1. \tag{2.14}$$

Therefore, φ^* is the unique solution of the BVP

$$\left\{\begin{array}{rcl} \dot{\varphi}^*(\tau)+A^{\mathrm{T}}(\tau)\varphi^*(\tau) &=& 0,\ \tau\in[0,T],\\ \varphi^*(T)-\varphi^*(0) &=& 0,\\ \int_0^T\langle\varphi^*(\tau),v(\tau)\rangle d\tau-1 &=& 0. \end{array}\right. \tag{2.15}$$

The function $h_2(\tau)$ can be found by solving an appropriate BVP, assuming that (2.1) restricted to $W^c(\Gamma)$ has the periodic normal form (2.4). The coefficient b arises from the solvability condition for the BVP as an integral over the interval $[0,T]$ of scalar products. Specifically, these scalar products involve the quadratic terms of (1.1) near the periodic solution u_0, the (generalized) eigenfunction v, and the adjoint eigenfunction φ^* defined by (2.15).

Substitute (2.10) into (2.1), using (2.2), (2.4), and (2.11), as well as

$$\frac{du}{dt}=\frac{\partial u}{\partial\xi}\frac{d\xi}{dt}+\frac{\partial u}{\partial\tau}\frac{d\tau}{dt}.$$

Collecting the ξ^0-terms in the resulting equation, we get the identity

$$\dot{u}_0=F(u_0),$$

where u_0 is the periodic solution of (2.1).

The ξ^1-terms provide another identity, namely,

$$\dot{v}-A(\tau)v-\dot{u}_0=0,$$

due to (2.12).

Finally, collecting the ξ^2-terms, we obtain the equation for h_2,

$$\dot{h}_2-A(\tau)h_2=B(\tau;v,v)-2a\dot{u}_0+2\dot{v}-2bv, \tag{2.16}$$

to be solved in the space of vector-functions on $[0,T]$ satisfying $h_2(T) = h_2(0)$. The differential operator $\frac{d}{d\tau} - A(\tau)$ is singular in this space, with $\dot{u}_0$ as the eigenfunction corresponding to zero eigenvalue. The null-eigenfunction φ^* of the adjoint operator $-\frac{d}{d\tau} - A^{\mathrm{T}}(\tau)$ is defined by (2.15). Thus, the Fredholm solvability condition implies that

$$\int_0^T \langle \varphi^*(\tau), B(\tau; v(\tau), v(\tau)) - 2a\dot{u}_0(\tau) + 2\dot{v}(\tau) - 2bv(\tau)\rangle \, d\tau = 0.$$

Using (2.13) and (2.14), we get the expression

$$b = \frac{1}{2}\int_0^T \langle \varphi^*(\tau), B(\tau; v(\tau), v(\tau)) + 2A(\tau)v(\tau)\rangle \, d\tau. \tag{2.17}$$

Here v and φ^* are defined by (2.12) and (2.15), respectively. Therefore, the critical coefficient b in the periodic normal form for the LPC bifurcation has been computed. The bifurcation is nondegenerate if $b \neq 0$. Note that the coefficient a does not enter into (2.17) due to (2.13).

2.2. The period-doubling bifurcation. The two-dimensional critical center manifold $W^c(\Gamma)$ at the PD bifurcation can be parametrized locally by (τ, ξ) as

$$u = u_0(\tau) + \xi w(\tau) + H(\tau, \xi), \quad \tau \in [0, 2T], \ \xi \in \mathbb{R}, \tag{2.18}$$

where the function H satisfies $H(2T, \xi) = H(0, \xi)$. It has the Taylor expansion

$$H(\tau, \xi) = \frac{1}{2}h_2(\tau)\xi^2 + \frac{1}{6}h_3(\tau)\xi^3 + O(\xi^4), \tag{2.19}$$

with $h_j(2T) = h_j(0)$, while

$$w(\tau) = \begin{cases} v(\tau), & \tau \in [0,T], \\ -v(\tau - T), & \tau \in [T, 2T], \end{cases} \tag{2.20}$$

with

$$\left\{ \begin{array}{rcl} \dot{v}(\tau) - A(\tau)v(\tau) & = & 0, \ \tau \in [0,T], \\ v(T) + v(0) & = & 0, \\ \int_0^T \langle v(\tau), v(\tau)\rangle d\tau - 1 & = & 0. \end{array} \right. \tag{2.21}$$

The function v exists due to Lemma 5 of [18].

The parametrization (2.18) provides a two-cover of $W^c(\Gamma)$ that is locally diffeomorphic to the Möbius band (see Figure 2.1).

The functions $h_2(\tau)$ and $h_3(\tau)$ can be found by solving appropriate BVPs, assuming that (2.1) restricted to $W^c(\Gamma)$ has the periodic normal form (2.5). The coefficients a and c arise from the solvability conditions for the BVPs as integrals of scalar products over the interval $[0,T]$. Specifically, these scalar products involve the quadratic and cubic terms of (1.1) near the periodic solution u_0, the eigenfunction v, and a similar adjoint eigenfunction v^* satisfying

$$\left\{ \begin{array}{rcl} \dot{v}^*(\tau) + A^{\mathrm{T}}(\tau)v^*(\tau) & = & 0, \ \tau \in [0,T], \\ v^*(T) + v^*(0) & = & 0, \\ \int_0^T \langle v^*(\tau), v(\tau)\rangle d\tau - 1/2 & = & 0. \end{array} \right. \tag{2.22}$$

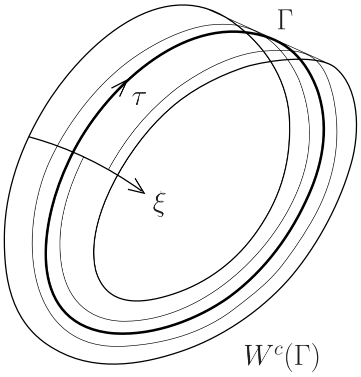


FIG. 2.1. *Center manifold $W^c(\Gamma)$ at the PD bifurcation.*

Similarly to (2.20), define

$$w^*(\tau) = \begin{cases} v^*(\tau), & \tau \in [0,T], \\ -v^*(\tau - T), & \tau \in [T, 2T]. \end{cases} \tag{2.23}$$

Note that

$$\int_0^{2T} \langle w^*(\tau), w(\tau)\rangle d\tau = 1. \tag{2.24}$$

To derive the normal form coefficients, we proceed as in section 2.1, namely, we substitute (2.18) into (2.1) and use (2.2), as well as (2.5) and (2.19).

Collecting the ξ^0-terms in the resulting equation, we get the identity

$$\dot{u}_0 = F(u_0),$$

where u_0 is the T-periodic solution of (2.1).

The ξ^1-terms provide the identity

$$\dot{w} = A(\tau)w,$$

due to (2.20) and (2.21).

Collecting the ξ^2-terms, we obtain the equation for h_2,

$$\dot{h}_2 - A(\tau)h_2 = B(\tau; w, w) - 2a\dot{u}_0, \tag{2.25}$$

to be solved in the space of functions on $[0, 2T]$ satisfying $h_2(2T) = h_2(0)$. In this space, the differential operator $\frac{d}{d\tau} - A(\tau)$ is singular with two linearly independent null-functions: $\psi = \dot{u}_0$ and w.

Thus, two Fredholm solvability conditions are involved, namely,

$$\int_0^{2T} \langle w^*(\tau), B(\tau; w(\tau), w(\tau)) - 2a\dot{u}_0(\tau)\rangle \, d\tau = 0,$$

which holds automatically for any a, due to (2.23) and the T-periodicity of the right-hand side of (2.25), and

$$\int_0^{2T} \langle \psi^*(\tau), B(\tau; w(\tau), w(\tau)) - 2a\dot{u}_0(\tau)\rangle \, d\tau = 0,$$

where ψ^* satisfies

$$\left\{ \begin{array}{rcl} \dot{\psi}^*(\tau) + A^{\mathrm{T}}(\tau)\psi^*(\tau) & = & 0, \ \tau \in [0,T], \\ \psi^*(T) - \psi^*(0) & = & 0, \\ \int_0^T \langle \psi^*(\tau), F(u_0(\tau)) \rangle \, d\tau - 1/2 & = & 0, \end{array} \right. \tag{2.26}$$

and is extended to $[T, 2T]$ by periodicity. Note that $\int_0^{2T} \langle \psi^*(\tau), F(u_0(\tau)) \rangle \, d\tau \neq 0$, since 0 is a semisimple eigenvalue of the differential operator $\frac{d}{d\tau} - A(\tau)$. This leads to the expression

$$a = \frac{1}{2} \int_0^{2T} \langle \psi^*(\tau), B(\tau; w(\tau), w(\tau)) \rangle \, d\tau$$

or, equivalently,

$$a = \int_0^{T} \langle \psi^*(\tau), B(\tau; v(\tau), v(\tau)) \rangle \, d\tau, \tag{2.27}$$

where v and ψ^* are defined by (2.21) and (2.26), respectively.

With a defined in this way, let h_2 be the unique solution of (2.25) in the space of functions on $[0, 2T]$ satisfying $h_2(0) = h_2(2T)$, as well as two orthogonality conditions:

$$\int_0^{2T} \langle w^*(\tau), h_2(\tau) \rangle \, d\tau = 0,$$

which holds automatically, due to the T-periodicity of h_2 ($h_2(0) = h_2(T)$), and

$$\int_0^{2T} \langle \psi^*(\tau), h_2(\tau) \rangle \, d\tau = 0,$$

which is equivalent to

$$\int_0^{T} \langle \psi^*(\tau), h_2(\tau) \rangle \, d\tau = 0.$$

Thus h_2 is the unique solution of the BVP

$$\left\{ \begin{array}{rcl} \dot{h}_2(\tau) - A(\tau)h_2(\tau) - B(\tau; v(\tau), v(\tau)) + 2aF(u_0(\tau)) & = & 0, \ \tau \in [0,T], \\ h_2(T) - h_2(0) & = & 0, \\ \int_0^T \langle \psi^*(\tau), h_2(\tau) \rangle \, d\tau & = & 0, \end{array} \right. \tag{2.28}$$

extended by periodicity to $[T, 2T]$. Collecting the ξ^3-terms, we get the equation for h_3,

$$\dot{h}_3 - A(\tau)h_3 = C(\tau; w, w, w) + 3B(\tau; w, h_2) - 6a\dot{w} - 6cw, \tag{2.29}$$

which again must be solved in the space of functions on $[0, 2T]$ satisfying $h_3(2T) = h_3(0)$. Its solvability implies

$$\int_0^{2T} \langle w^*(\tau), C(\tau; w(\tau), w(\tau), w(\tau)) + 3B(\tau; w(\tau), h_2(\tau)) - 6a\dot{w}(\tau) - 6cw(\tau) \rangle \, d\tau = 0.$$

Taking into account (2.24), we obtain

$$c = \frac{1}{6}\int_0^{2T} \langle w^*(\tau), C(\tau; w(\tau), w(\tau), w(\tau)) + 3B(\tau; w(\tau), h_2(\tau)) - 6aA(\tau)w(\tau)\rangle \, d\tau$$

and finally,

$$c = \frac{1}{3}\int_0^{T} \langle v^*(\tau), C(\tau; v(\tau), v(\tau), v(\tau)) + 3B(\tau; v(\tau), h_2(\tau)) - 6aA(\tau)v(\tau)\rangle \, d\tau, \tag{2.30}$$

where a is defined by (2.27), h_2 is the solution of (2.28), and v and v^* are defined by (2.21) and (2.22), respectively. Thus, the critical coefficient c in the periodic normal form for the PD bifurcation has been computed. The critical cycle is stable within the center manifold if $c < 0$ and is unstable if $c > 0$. In the former case, the bifurcation is supercritical, while in the latter case it is subcritical.

2.3. The torus bifurcation. The three-dimensional critical center manifold $W^c(\Gamma)$ at the NS bifurcation can be parametrized locally by (τ, ξ) as

$$u = u_0(\tau) + \xi v(\tau) + \bar{\xi}\bar{v}(\tau) + H(\tau, \xi, \bar{\xi}), \quad \tau \in [0, T], \ \xi \in \mathbb{C}, \tag{2.31}$$

where the real function H satisfies $H(T, \xi, \bar{\xi}) = H(0, \xi, \bar{\xi})$, and has the Taylor expansion

$$\begin{aligned} H(\tau, \xi, \bar{\xi}) = {} & \frac{1}{2}h_{20}(\tau)\xi^2 + h_{11}(\tau)\xi\bar{\xi} + \frac{1}{2}h_{02}(\tau)\bar{\xi}^2 \\ & + \frac{1}{6}h_{30}(\tau)\xi^3 + \frac{1}{2}h_{21}(\tau)\xi^2\bar{\xi} + \frac{1}{2}h_{12}(\tau)\xi\bar{\xi}^2 + \frac{1}{6}h_{03}(\tau)\bar{\xi}^3 \\ & + O(|\xi|^4), \end{aligned} \tag{2.32}$$

with $h_{ij}(T) = h_{ij}(0)$ and $h_{ij} = \bar{h}_{ji}$ so that h_{11} is real, while

$$\left\{ \begin{array}{rcl} \dot{v}(\tau) - A(\tau)v(\tau) + \dfrac{i\theta}{T}v(\tau) & = & 0, \ \tau \in [0, T], \\ v(T) - v(0) & = & 0, \\ \int_0^T \langle v(\tau), v(\tau)\rangle d\tau - 1 & = & 0. \end{array} \right. \tag{2.33}$$

The function v exists due to Lemma 2 of [18]. Recall that $\langle u, v\rangle = u^{\mathrm{H}}v = \bar{u}^{\mathrm{T}}v$.

Note that

$$w(\tau) = \exp\left(\frac{i\theta\tau}{T}\right) v(\tau)$$

satisfies

$$\left\{ \begin{array}{rcl} \dot{w}(\tau) - A(\tau)w(\tau) & = & 0, \ \tau \in [0, T], \\ w(T) - e^{i\theta}w(0) & = & 0, \\ \int_0^T \langle w(\tau), w(\tau)\rangle d\tau - 1 & = & 0, \end{array} \right.$$

which is often used in the defining system for the NS bifurcation.

As in the previous cases, the functions $h_{ij}(\tau)$ can be found by solving appropriate BVPs, assuming that (2.1) restricted to $W^c(\Gamma)$ has the periodic normal form (2.6).

Also introduce the adjoint eigenfunction v^* that satisfies

$$(2.34)\qquad \left\{\begin{array}{rcl} \dot{v}^*(\tau) + A^{\mathrm{T}}(\tau)v^*(\tau) - \frac{i\theta}{T}v^*(\tau) & = & 0,\ \tau \in [0,T], \\ v^*(T) - v^*(0) & = & 0, \\ \int_0^T \langle v^*(\tau), v(\tau)\rangle d\tau - 1 & = & 0. \end{array}\right.$$

Substitute (2.31) into (2.1), using (2.2), (2.6), and (2.32), as well as

$$\frac{du}{dt} = \frac{\partial u}{\partial \xi}\frac{d\xi}{dt} + \frac{\partial u}{\partial \bar{\xi}}\frac{d\bar{\xi}}{dt} + \frac{\partial u}{\partial \tau}\frac{d\tau}{dt}.$$

The ξ-independent terms in the resulting equation give the usual identity

$$\dot{u}_0 = F(u_0).$$

The ξ-terms give another identity, namely,

$$\dot{v} - A(\tau)v + \frac{i\theta}{T}v = 0,$$

while the $\bar{\xi}$-terms lead to the corresponding complex-conjugate identity.

Collecting the coefficients of the ξ^2- or $\bar{\xi}^2$-terms leads to the equation

$$(2.35)\qquad \dot{h}_{20} - A(\tau)h_{20} + \frac{2i\theta}{T}h_{20} = B(\tau; v, v)$$

or its complex-conjugate. This equation has a unique solution $h_{20}(\tau)$ satisfying $h_{20}(T) = h_{20}(0)$, since $e^{2i\theta}$ is not a multiplier of the critical cycle by the spectral assumptions. Thus, h_{20} can be found from the BVP

$$(2.36)\qquad \left\{\begin{array}{rcl} \dot{h}_{20}(\tau) - A(\tau)h_{20}(\tau) + \frac{2i\theta}{T}h_{20}(\tau) - B(\tau; v(\tau), v(\tau)) & = & 0,\ \tau \in [0,T], \\ h_{20}(T) - h_{20}(0) & = & 0. \end{array}\right.$$

The $|\xi|^2$-terms give

$$(2.37)\qquad \dot{h}_{11} - A(\tau)h_{11} = B(\tau; v, \bar{v}) - a\dot{u}_0,$$

where $h_{11}(T) = h_{11}(0)$. The differential operator $\frac{d}{d\tau} - A(\tau)$ has a nontrivial kernel spanned by $\dot{u}_0$. The null-eigenfunction of the adjoint operator $-\frac{d}{d\tau} - A^{\mathrm{T}}(\tau)$ is φ^*, given by the equation

$$(2.38)\qquad \left\{\begin{array}{rcl} \dot{\varphi}^*(\tau) + A^{\mathrm{T}}(\tau)\varphi^*(\tau) & = & 0,\ \tau \in [0,T], \\ \varphi^*(T) - \varphi^*(0) & = & 0, \\ \int_0^T \langle \varphi^*(\tau), F(u_0(\tau))\rangle d\tau - 1 & = & 0. \end{array}\right.$$

Note that $\int_0^T \langle \varphi^*(\tau), F(u_0(\tau))\rangle d\tau \neq 0$, since the trivial multiplier 1 is simple, due to the spectral assumptions. The Fredholm solvability condition implies

$$(2.39)\qquad a = \int_0^T \langle \varphi^*(\tau), B(\tau; v(\tau), \bar{v}(\tau))\rangle\ d\tau.$$

With a defined in this way, let h_{11} be the unique solution of (2.37) satisfying $h_{11}(T) = h_{11}(0)$ and

$$\int_0^T \langle \varphi^*(\tau), h_{11}(\tau) \rangle \, d\tau = 0,$$

i.e.,

$$\tag{2.40} \left\{ \begin{array}{rcl} \dot{h}_{11}(\tau) - A(\tau)h_{11}(\tau) - B(\tau; v(\tau), \bar{v}(\tau)) + aF(u_0(\tau)) & = & 0, \ \tau \in [0,T], \\ h_{11}(T) - h_{11}(0) & = & 0, \\ \int_0^T \langle \varphi^*(\tau), h_{11}(\tau) \rangle \, d\tau & = & 0. \end{array} \right.$$

Other spectral conditions assure the solvability of linear equations for h_{30}, h_{12}, and h_{03}. Since these coefficients are not used below, we do not write the corresponding equations explicitly.

Finally, the coefficients of the $\xi^2\bar{\xi}$-terms give the singular equation

$$\dot{h}_{21} - Ah_{21} + \frac{i\theta}{T} h_{21} = 2B(\tau; h_{11}, v) + B(\tau; h_{20}, \bar{v}) + C(\tau; v, v, \bar{v}) - 2a\dot{v} - 2dv.$$

If one takes into account (2.34), the Fredholm solvability condition implies

$$\begin{aligned} d = & \frac{1}{2} \int_0^T \langle v^*(\tau), B(\tau; h_{11}(\tau), v(\tau)) + B(\tau; h_{20}(\tau), \bar{v}(\tau)) + C(\tau; v(\tau), v(\tau), \bar{v}(\tau)) \rangle \, d\tau \\ & - a \int_0^T \langle v^*(\tau), A(\tau)v(\tau) \rangle \, d\tau + \frac{ia\theta}{T}, \end{aligned} \tag{2.41}$$

where a is defined by (2.39), h_{11} and h_{20} by (2.40) and (2.36), respectively, and v and v^* satisfy (2.33) and (2.34), respectively. Thus, the critical coefficient d in the periodic normal form for the NS bifurcation has been computed. The critical cycle is stable within the center manifold if Re $d < 0$ and is unstable if Re $d > 0$. In the former case, the NS bifurcation is supercritical, while in the latter case it is subcritical.

3. Implementation issues.

Numerical implementation of the formulas derived in the preceding sections requires the evaluation of integrals of scalar functions over $[0,T]$ and the solution of nonsingular linear BVPs with integral constraints. Such tasks can be carried out with continuation software such as AUTO [19], CONTENT [11], and MATCONT [20]. In these software packages, periodic solutions to (1.1) are computed with the method of *orthogonal collocation* with piecewise polynomials applied to properly formulated BVPs. The standard BVP for the periodic solutions is formulated on the unit interval $[0,1]$, so that the period T becomes a parameter, and it involves an integral phase condition:

$$\tag{3.1} \left\{ \begin{array}{rcl} \dot{x}(\tau) - Tf(x(\tau), \alpha) & = & 0, \ \tau \in [0,1], \\ x(0) - x(1) & = & 0, \\ \int_0^1 \langle x(\tau), \dot{\xi}(\tau) \rangle \, d\tau & = & 0, \end{array} \right.$$

where ξ is a previously calculated periodic solution, rescaled to $[0,1]$.

In the orthogonal collocation method [21], the problem (3.1) is replaced by the following discretization:

$$
\tag{3.2}
\begin{cases}
\displaystyle\sum_{j=0}^{m} x_{i,j}\dot{\ell}_{i,j}(\zeta_{i,k}) - Tf\left(\sum_{j=0}^{m} x_{i,j}\ell_{i,j}(\zeta_{i,k}),\alpha\right) & = \quad 0,\\
x_{0,0} - x_{N-1,m} & = \quad 0,\\
\displaystyle\sum_{i=0}^{N-1}\sum_{j=0}^{m-1}\sigma_{i,j}\langle x_{i,j},\dot{\xi}_{i,j}\rangle + \sigma_{N,0}\langle x_{N,0},\dot{\xi}_{N,0}\rangle & = \quad 0.
\end{cases}
$$

Here $x_{i,j}$ is the approximation of $x(\tau)$ at $m+1$ equidistant mesh points

$$\tau_{i,j} = \tau_i + \frac{j}{m}(\tau_{i+1} - \tau_i), \quad j = 0, 1, \ldots, m,$$

on each of N intervals $[\tau_i, \tau_{i+1}]$, where

$$0 = \tau_0 < \tau_1 < \cdots < \tau_N = 1.$$

The $\ell_{i,j}(\tau)$'s are the Lagrange basis polynomials, while $\zeta_{i,j}$ $(j = 1, \ldots, m)$ are Gauss points [2], i.e., the roots of the Legendre polynomial of degree m, all relative to the interval $[\tau_i, \tau_{i+1}]$.

With this choice of collocation points $\zeta_{i,j}$, the approximation error at the mesh points has order of accuracy m,

$$\|x(\tau_{i,j}) - x_{i,j}\| = \mathcal{O}(h^m),$$

where $h = \max_{i=1,2,\ldots,N}\{t_i\},\ t_i = \tau_i - \tau_{i-1}$ $(i = 1, \ldots, N)$, while for the main mesh points τ_i it has order of accuracy $2m$,

$$\|x(\tau_i) - x_{i,0}\| = \mathcal{O}(h^{2m})$$

("superconvergence").

The integration weight $\sigma_{i,j}$ of $\tau_{i,j}$ is given by $w_{j+1}t_{i+1}$ for $0 \le i \le N-1$ and $0 < j < m$. For $i = 0, \ldots, N-2$, the integration weight of $\tau_{i,m}$ $(\tau_{i,m} = \tau_{i+1,0})$ is given by $\sigma_{i,m} = w_{m+1}t_{i+1} + w_1 t_{i+2}$, and the integration weights of τ_0 and τ_N are given by $w_1 t_1$ and $w_{m+1}t_N$, respectively. In the above expressions, w_{j+1} is the Lagrange quadrature coefficient.

The numerical continuation of the solutions of (3.2) leads to structured, sparse linear systems, which in AUTO [19] and CONTENT [11] are solved by an efficient, specially adapted elimination algorithm that computes the multipliers as a by-product, without explicitly using the Poincaré map. To detect codim 1 bifurcations, one can specify test functions that are based on computing multipliers [22, 19] or on solving appropriate bordered linear BVPs [23].

Once a codim 1 bifurcation has been detected, one can compute the normal form coefficients using the formulas derived in the previous sections. All BVPs are reformulated on the unit interval $[0, 1]$, and all integrals are scaled accordingly. Moreover, if the bordering methods from [23] are used to continue LPC, PD, and NS bifurcations of limit cycles, then the computation of the normal form coefficients requires little extra effort, since all necessary eigenfunctions have already been computed, either while evaluating the test functionals or their gradients. These coefficients then serve as test functions for detecting codim 2 singularities of limit cycles due to *nonlinear degeneracies* of LPC, PD, or NS bifurcations, i.e., the cusp (CPC), the degenerate period-doubling (DP), and the degenerate NS or Chenciner (CH) bifurcation. Byproducts of these computations are test functions for detecting certain codim 2 singularities of limit cycles due to *linear degeneracies*, namely, the strong 1:1 resonance (R1), the strong 1:2 resonance (R2), the fold-Neimark–Sacker bifurcation (FN), and the fold-flip bifurcation (FF).

3.1. Discretization symbols. It is convenient to discretize all computed functions using the *same mesh* as in (3.2). For a given vector function $\eta \in \mathcal{C}^1([0,1], \mathbb{R}^n)$ we consider three different discretizations:

- $\eta_M \in \mathbb{R}^{(Nm+1)n}$, the vector of the function values at the mesh points;
- $\eta_C \in \mathbb{R}^{Nmn}$, the vector of the function values at the collocation points;
- $\eta_W = \left[\begin{smallmatrix}\eta_{W_1}\\ \eta_{W_2}\end{smallmatrix}\right] \in \mathbb{R}^{Nmn} \times \mathbb{R}^n$, where η_{W_1} is the vector of the function values at the collocation points multiplied by the Gauss–Legendre weights and the lengths of the corresponding mesh intervals, and $\eta_{W_2} = \eta(0)$.

Formally we also introduce the structured sparse matrix $L_{C\times M}$ that converts a vector η_M of function values at the mesh points into a vector η_C of its values at the collocation points, namely, $\eta_C = L_{C\times M}\eta_M$. This matrix is never formed explicitly; its entries are approximated by the $\ell_{i,j}(\zeta_{i,k})$-coefficients in (3.2). We also need a matrix $A_{C\times M}$ such that $A_{C\times M}\eta_M = (A(t)\eta(t))_C$. Again this matrix need not be formed explicitly. On the other hand, we do need the matrix $(D - TA(t))_{C\times M}$ explicitly; it is defined by $(D - TA(t))_{C\times M}\eta_M = (\dot{\eta}(t) - TA(t)\eta(t))_C$. Finally, let the tensors $B_{C\times M\times M}$ and $C_{C\times M\times M\times M}$ be defined by $B_{C\times M\times M}\eta_{1M}\eta_{2M} = (B(t;\eta_1(t), \eta_2(t)))_C$ and

$$C_{C\times M\times M\times M}\eta_{1M}\eta_{2M}\eta_{3M} = (C(t;\eta_1(t), \eta_2(t), \eta_3(t)))_C$$

for all $\eta_i \in \mathcal{C}^1([0,1], \mathbb{R}^n)$. (These tensors are not formed explicitly.)

Let $f(t), g(t) \in \mathcal{C}^0([0,1], \mathbb{R})$ be two scalar functions. Then the integral $\int_0^1 f(t)dt$ is represented by $\sum_{i=0}^{N-1}\sum_{j=1}^{m}\omega_j (f_C)_{i,j} t_{i+1} = \sum_{i=0}^{N-1}\sum_{j=1}^{m}(f_{W_1})_{i,j}$, where $(f_C)_{i,j} = f(\zeta_{i,j})$ and ω_j is the Gauss–Legendre quadrature coefficient. The integral $\int_0^1 f(t)g(t)dt$ is approximated with Gauss–Legendre by $f_{W_1}^{\mathrm{T}} g_C \approx f_{W_1}^{\mathrm{T}} L_{C\times M} g_M$, where equality holds if $g(t)$ is a piecewise polynomial of degree m or less on the given mesh. For vector functions $f(t), g(t) \in \mathcal{C}^0([0,1], \mathbb{R}^n)$, the integral $\int_0^1 \langle f(t), g(t)\rangle \, dt$ is formally approximated by the same expression: $f_{W_1}^{\mathrm{T}} g_C \approx f_{W_1}^{\mathrm{T}} L_{C\times M} g_M$, where again we have equality if $g(t)$ is a piecewise polynomial of degree m or less on the given mesh. Concerning the accuracy of the quadrature formulas, we first note that accuracy is not an important issue for the phase integral in (3.1), as this equation only selects a specific solution from the continuum of solutions obtained by phase shifts. Similarly, the discretization of the normalization integrals, for example, in (2.15), does not affect the inherent accuracy, including superconvergence at the main mesh points τ_i of the solution of the discretized BVP. Discretization of integrals, as specified above, follows the standard Gauss quadrature error, which has order of accuracy $2m$ if, as mentioned, the function $g(t)$ is a piecewise polynomial of degree m or less on the given mesh and if $f(t)$ is sufficiently smooth (in a piecewise sense). Otherwise, still assuming sufficient piecewise smoothness, the order of accuracy of the numerical integrals is $m + 1$ if m is odd, and $m + 2$ if m is even. In particular, for the often used choice $m = 4$, the integrals would then have order of accuracy 6.

We now consider the LPC, PD, and NS cases separately.

3.2. LPC bifurcation. The first task is to rescale the computed functions to the interval $[0,1]$. We start by defining $u_1(t) = u_0(Tt)$ for $t \in [0,1]$. The linear BVPs (2.12) and (2.15) are replaced by

$$\left\{\begin{array}{rcl} \dot{v}_1(t) - TA(t)v_1(t) - TF(u_1(t)) & = & 0, \ t \in [0,1], \\ v_1(0) - v_1(1) & = & 0, \\ \int_0^1 \langle v_1(t), F(u_1(t))\rangle dt & = & 0, \end{array}\right. \tag{3.3}$$

where $v(\tau) = v_1(\tau/T)$, and

$$\left\{\begin{array}{rcl} \dot{\varphi}_1^*(t) + TA^{\mathrm{T}}(t)\varphi_1^*(t) & = & 0, \ t \in [0,1], \\ \varphi_1^*(0) - \varphi_1^*(1) & = & 0, \\ \int_0^1 \langle \varphi_1^*(t), \varphi_1^*(t)\rangle dt - 1 & = & 0, \end{array}\right. \tag{3.4}$$

respectively. We then compute $I = \int_0^1 \langle \varphi_1^*(t), v_1(t)\rangle dt$. If $I = 0$, then we have a strong 1:1 resonance (a limit cycle with two nontrivial multipliers equal to 1). If not, then we rescale φ^* so that $I = 1$. It then follows that $\varphi^*(\tau) = \varphi_1^*(\tau/T)/T$. Thus, we obtain

$$b = \frac{1}{2}\int_0^1 \langle \varphi_1^*(t), B(t; v_1(t), v_1(t)) + 2A(t)v_1(t)\rangle \, dt. \tag{3.5}$$

We compute v_{1M} by solving the discretization of (3.3)

$$\left[\begin{array}{c} (D - TA(t))_{C\times M} \\ \delta_0 - \delta_1 \\ (g_{W_1})^{\mathrm{T}} L_{C\times M} \end{array}\right] v_{1M} = \left[\begin{array}{c} Tg_C \\ 0 \\ 0 \end{array}\right], \tag{3.6}$$

where $g(t) = F(u_1(t))$.

It is more efficient to compute φ_{1W}^* than φ_{1M}^*, since φ_1^* will be used only to compute integrals of the form $\int_0^1 \langle \varphi_1^*(t), \zeta(t)\rangle dt$. Moreover, φ_{1W}^* can be computed with the same matrix used in (3.6), thus saving factorization costs. Formally, the computation of φ_{1W}^* is based on Proposition A.1 from the appendix. Instead of approximating φ_1^* by solving

$$\left[\begin{array}{c} (D + TA^{\mathrm{T}}(t))_{C\times M} \\ \delta_0 - \delta_1 \end{array}\right] \varphi_{1M}^* = 0,$$

we remark that $\left[\begin{smallmatrix} \varphi_1^* \\ \varphi_1^*(0) \end{smallmatrix}\right]$ is orthogonal to the range of $\left[\begin{smallmatrix} D-TA(t) \\ \delta_0-\delta_1 \end{smallmatrix}\right]$. By discretization we obtain

$$(\varphi_{1W}^*)^{\mathrm{T}} \left[\begin{array}{c} (D - TA(t))_{C\times M} \\ \delta_0 - \delta_1 \end{array}\right] = 0.$$

To normalize $\varphi_{1W_1}^*$, we require

$$\sum_{i=0}^{N-1}\sum_{j=1}^{m} \left|(\varphi_{1W_1}^*)_{i,j}\right|_1 = 1. \tag{3.7}$$

Here $|.|_1$ denotes the 1-norm (sum of absolute values) of a vector. This choice is convenient for computational reasons. Then $\int_0^1 \langle \varphi_1^*(t), v_1(t)\rangle dt$ is approximated by $(\varphi_{1W_1}^*)^{\mathrm{T}} L_{C\times M} v_{1M}$, and if this quantity is nonzero, φ_{1W}^* is rescaled to ensure $\int_0^1 \langle \varphi_1^*(t), v_1(t)\rangle dt = 1$.

The integral (3.5) is finally approximated by

$$b = \frac{1}{2}(\varphi_{1W_1}^*)^{\mathrm{T}}(B_{C\times M\times M} v_{1M} v_{1M} + 2A_{C\times M} v_{1M}). \tag{3.8}$$

3.3. PD bifurcation. Again, we rescale the computed quantities to the interval $[0,1]$. The linear BVPs (2.21) and (2.22) are replaced by

$$\left\{\begin{array}{rcl} \dot{v_1}(t) - TA(t)v_1(t) &=& 0,\ t\in[0,1],\\ v_1(0)+v_1(1) &=& 0,\\ \int_0^1 \langle v_1(t), v_1(t)\rangle dt - 1 &=& 0, \end{array}\right. \tag{3.9}$$

where $v(\tau) = v_1(\tau/T)/\sqrt{T}$, and

$$\left\{\begin{array}{rcl} \dot{v_1}^*(t) + TA^{\mathrm{T}}(t)v_1^*(t) &=& 0,\ t\in[0,1],\\ v_1^*(0)+v_1^*(1) &=& 0,\\ \int_0^1 \langle v_1^*(t), v_1^*(t)\rangle dt - 1/2 &=& 0, \end{array}\right. \tag{3.10}$$

respectively. We note that the last equation in (3.10) differs from the last equation in (2.22). We then compute $I = \int_0^1 \langle v_1^*(t), v_1(t)\rangle dt$. If $I=0$, then we have a strong 1:2 resonance (a limit cycle with two multipliers equal to -1). If not, then we rescale v_1^* so that $I = 1/2$, which corresponds to the normalization condition used in (3.10). It then follows that $v^*(\tau) = v_1^*(\tau/T)/\sqrt{T}$.

We also replace (2.26) by

$$\left\{\begin{array}{rcl} \dot{\psi}_1^*(t) + TA^{\mathrm{T}}(t)\psi_1^*(t) &=& 0,\ t\in[0,1],\\ \psi_1^*(0)-\psi_1^*(1) &=& 0,\\ \int_0^1 \langle \psi_1^*(t), \psi_1^*(t)\rangle\, dt - 1 &=& 0. \end{array}\right. \tag{3.11}$$

Again, the last equation in (3.11) differs from the last equation in (2.26). We then compute $I = \int_0^1 \langle \psi_1^*(t), F(u_1(t))\rangle dt$. If $I=0$, then we have a fold-flip bifurcation. If not, then we rescale ψ_1^* so that $I=1$. It then follows that $\psi^*(\tau) = \psi_1^*(\tau/T)/T$.

This leads to the expression

$$a_1 = \int_0^1 \langle \psi_1^*(t), B(t; v_1(t), v_1(t))\rangle\, dt, \tag{3.12}$$

where $a_1 = aT$.

With a_1 defined in this way, let $h_{2,1}$ be the unique solution of the BVP

$$\left\{\begin{array}{rcl} \dot{h}_{2,1}(t) - TA(t)h_{2,1}(t) - B(t; v_1(t), v_1(t)) + 2a_1F(u_1(t)) &=& 0,\ t\in[0,1],\\ h_{2,1}(0)-h_{2,1}(1) &=& 0,\\ \int_0^1 \langle \psi_1^*(t), h_{2,1}(t)\rangle\, dt &=& 0, \end{array}\right. \tag{3.13}$$

where $h_2(\tau) = h_{2,1}(\tau/T)$.

Therefore we obtain

$$\begin{aligned} c = \frac{1}{3}\int_0^1 &\left\langle v_1^*(t), \frac{1}{T}C(t; v_1(t), v_1(t), v_1(t)) + 3B(t; v_1(t), h_{2,1}(t))\right\rangle\, dt \\ &- \frac{2a_1}{T}\int_0^1 \langle v_1^*(t), A(t)v_1(t)\rangle\, dt. \end{aligned} \tag{3.14}$$

We compute v_{1M} by solving

$$\begin{bmatrix} (D - TA(t))_{C\times M} \\ \delta_0 + \delta_1 \end{bmatrix} v_{1M} = 0. \tag{3.15}$$

We normalize v_{1M} by requiring $\sum_{i=0}^{N-1}\sum_{j=0}^{m}\sigma_j\langle (v_{1M})_{i,j},(v_{1M})_{i,j}\rangle = 1$, where σ_j is the Lagrange quadrature coefficient.

As in the LPC case, it is more efficient to compute v_{1W}^*, rather than v_M^*, since v_1^* will be used only to compute integrals of the form $\int_0^1\langle v_1^*(t),\zeta(t)\rangle dt$. Moreover, v_{1W}^* can be computed with the same matrix in (3.15), thus saving factorization costs. Formally, the computation of v_{1W}^* is based on Proposition A.2 (see the appendix). Instead of approximating v_1^* by solving

$$\begin{bmatrix} (D+TA^{\mathrm{T}}(t))_{C\times M} \\ \delta_0+\delta_1 \end{bmatrix} v_{1M}^* = 0,$$

we observe that $\left[\begin{smallmatrix} v_1^* \\ v_1^*(0)\end{smallmatrix}\right]$ is orthogonal to the range of $\left[\begin{smallmatrix} D-TA(t) \\ \delta_0+\delta_1\end{smallmatrix}\right]$. By discretization we obtain

$$(v_{1W}^*)^{\mathrm{T}}\begin{bmatrix} (D-TA(t))_{C\times M} \\ \delta_0+\delta_1 \end{bmatrix} = 0.$$

To normalize $v_{1W_1}^*$, we require $\sum_{i=0}^{N-1}\sum_{j=1}^{m}\left|(v_{1W_1}^*)_{i,j}\right|_1 = 1$. Then $\int_0^1\langle v_1^*(t),v_1(t)\rangle dt$ is approximated by $(v_{1W_1}^*)^{\mathrm{T}}L_{C\times M}v_{1M}$. If this quantity is nonzero, then v_{1W}^* is rescaled so that $\int_0^1\langle v_1^*(t),v_1(t)\rangle dt = 1/2$.

From Proposition A.1 it follows that we can approximate ψ_1^* like v_1^*, namely, we compute ψ_{1W}^* by solving

$$(\psi_{1W}^*)^{\mathrm{T}}\begin{bmatrix} (D-TA(t))_{C\times M} \\ \delta_0-\delta_1 \end{bmatrix} = 0,$$

and normalize $\psi_{1W_1}^*$ by requiring

$$\sum_{i=0}^{N-1}\sum_{j=1}^{m}\left|(\psi_{1W_1}^*)_{i,j}\right|_1 = 1.$$

Then $\int_0^1\langle\psi_1^*(t),F(u_1(t))\rangle dt$ is approximated by $(\psi_{1W_1}^*)^{\mathrm{T}}(F(u_1(t)))_C$ and if this quantity is nonzero, ψ_{1W}^* is rescaled so that $\int_0^1\langle\psi_1^*(t),F(u_1(t))\rangle dt = 1$.

Having found v_{1M} and ψ_{1W}^*, a_1 can be computed using (3.12) as

$$a_1 = (\psi_{1W_1}^*)^{\mathrm{T}}B_{C\times M\times M}v_{1M}v_{1M}.$$

Next, $(h_{2,1})_M$ is found by solving the discretization of (3.13), namely,

$$\begin{bmatrix} (D-TA(t))_{C\times M} \\ \delta_0-\delta_1 \\ (\psi_{W_1}^*)^{\mathrm{T}}L_{C\times M} \end{bmatrix}(h_{2,1})_M = \begin{bmatrix} B_{C\times M\times M}v_{1M}v_{1M}+2a_1g_C \\ 0 \\ 0 \end{bmatrix},$$

where $g_C = (F(u_1(t)))_C$.

Finally, (3.14) is approximated by

$$\begin{aligned} c = {}& \frac{1}{3T}(v_{1W_1}^*)^{\mathrm{T}}\left(C_{C\times M\times M\times M}v_{1M}v_{1M}v_{1M}+3TB_{C\times M\times M}v_{1M}(h_{2,1})_M\right) \\ & -\frac{2a_1}{T}(v_{1W_1}^*)^{\mathrm{T}}A_{C\times M}v_{1M}. \end{aligned} \tag{3.16}$$

3.4. Torus bifurcation. As before, we first rescale the time variable to the unit time interval. The linear BVPs (2.33) and (2.34) are replaced by

$$\left\{\begin{array}{rcl} \dot{v_1}(t) - TA(t)v_1(t) + i\theta v_1(t) & = & 0,\ t \in [0,1], \\ v_1(0) - v_1(1) & = & 0, \\ \int_0^1 \langle v_1(t), v_1(t)\rangle dt - 1 & = & 0, \end{array}\right. \tag{3.17}$$

where $v(\tau) = v_1(\tau/T)/\sqrt{T}$, and

$$\left\{\begin{array}{rcl} \dot{v_1}^*(t) + TA^{\mathrm{T}}(t)v_1^*(t) - i\theta v_1^*(t) & = & 0,\ t \in [0,1], \\ v_1^*(0) - v_1^*(1) & = & 0, \\ \int_0^1 \langle v_1^*(t), v_1^*(t)\rangle dt - 1 & = & 0, \end{array}\right. \tag{3.18}$$

respectively. Note that the last equation in (3.18) differs from the last equation in (2.34). To rescale v_1^* we first compute $I = \int_0^1 \langle v_1^*(t), v_1(t)\rangle dt$. If $I \neq 0$, then we rescale v_1^* so that $I = 1$. (The case $I = 0$ corresponds to a bifurcation of codimension three or higher.) It then follows that $v^*(\tau) = v_1^*(\tau/T)/\sqrt{T}$.

We also replace (2.38) by

$$\left\{\begin{array}{rcl} \dot{\varphi_1}^*(t) + TA^{\mathrm{T}}(t)\varphi_1^*(\tau) & = & 0,\ t \in [0,1], \\ \varphi_1^*(0) - \varphi_1^*(1) & = & 0, \\ \int_0^1 \langle \varphi_1^*(t), \varphi_1^*(t)\rangle dt - 1 & = & 0. \end{array}\right. \tag{3.19}$$

Again, note that the last equation in (3.19) differs from the last equation in (2.38). Now compute $I = \int_0^1 \langle \varphi_1^*(t), F(u_1(t))\rangle dt$. If $I = 0$, then we have a fold-Neimark–Sacker bifurcation. If $I \neq 0$, then we rescale φ_1^* so that $I = 1$. It follows that $\varphi^*(\tau) = \varphi_1^*(\tau/T)/T$. (2.36) is replaced by

$$\left\{\begin{array}{rcl} \dot{h}_{20,1}(t) - A(t)h_{20,1}(t) + 2i\theta h_{20,1}(t) - B(t; v_1(t), v_1(t)) & = & 0,\ t \in [0,1], \\ h_{20,1}(0) - h_{20,1}(1) & = & 0, \end{array}\right. \tag{3.20}$$

where $h_{20}(\tau) = h_{20,1}(\tau/T)$. This leads to the expression

$$a_1 = \int_0^1 \langle \varphi_1^*(\tau), B(t; v_1(t), \bar{v}_1(t))\rangle\ dt, \tag{3.21}$$

where $a = a_1/T$.

With a_1 defined in this way, let $h_{11,1}$ be the unique solution of the BVP

$$\left\{\begin{array}{rcl} \dot{h}_{11,1}(t) - A(t)h_{11,1}(t) - B(t; v_1(t), \bar{v}_1(t)) + a_1F(u_1(t)) & = & 0,\ t \in [0,1], \\ h_{11,1}(0) - h_{11,1}(1) & = & 0, \\ \int_0^1 \langle \varphi_1^*(t), h_{11,1}(t)\rangle\ dt & = & 0, \end{array}\right. \tag{3.22}$$

where $h_{11}(\tau) = h_{11,1}(\tau/T)$.

Finally we obtain

$$\begin{aligned} d = {} & \frac{1}{2}\int_0^1 \langle v_1^*(t), B(t; h_{11,1}(t), v_1(t)) + B(t; h_{20,1}(t), \bar{v}_1(t))\rangle\ dt \\ & + \frac{1}{2T}\int_0^1 \langle v_1^*(t), C(t; v_1(t), v_1(t), \bar{v}_1(t))\rangle\ dt - \frac{a_1}{T}\int_0^1 \langle v_1^*(t), A(t)v_1(t)\rangle\ dt + \frac{ia_1\theta}{T^2}. \end{aligned} \tag{3.23}$$

We compute v_{1M} by solving

$$\left[\begin{array}{c}(D-TA(t)+i\theta I_n)_{C\times M}\\ \delta_0-\delta_1\end{array}\right] v_{1M}=0, \tag{3.24}$$

where $(D-TA(t)+i\theta I_n)_{C\times M}$ is defined like $(D-TA(t))_{C\times M}$. We normalize v_{1M} by requiring that $\sum_{i=0}^{N-1}\sum_{j=0}^{m}\sigma_j\langle (v_{1M})_{i,j},(v_{1M})_{i,j}\rangle=1$, where σ_j is the Lagrange quadrature coefficient. Again, it is more efficient to compute v_{1W}^* than v_M^*, since v_1^* will be used only to compute integrals of the form $\int_0^1\langle v_1^*(t),\zeta(t)\rangle dt$. Moreover, v_{1W}^* can be computed with the same matrix in (3.23). Formally, the computation of v_{1W}^* is based on Proposition A.3 from the appendix. Instead of approximating v_1^* by solving

$$\left[\begin{array}{c}(D+TA^{\mathrm{T}}(t)-i\theta I_n)_{C\times M}\\ \delta_0+\delta_1\end{array}\right] v_{1M}^*=0,$$

we remark that $\left[\begin{smallmatrix}v_1^*\\ v_1^*(0)\end{smallmatrix}\right]$ is orthogonal to the range of $\left[\begin{smallmatrix}D-TA(t)+i\theta\\ \delta_0+\delta_1\end{smallmatrix}\right]$. By discretization we obtain

$$(v_{1W}^*)^{\mathrm{H}}\left[\begin{array}{c}(D-TA(t)+i\theta I_n)_{C\times M}\\ \delta_0+\delta_1\end{array}\right]=0.$$

To normalize $v_{1W_1}^*$ we require that $\sum_{i=0}^{N-1}\sum_{j=1}^{m}\left|(v_{1W_1}^*)_{i,j}\right|_1=1$. Then $\int_0^1\langle v_1^*(t),v_1(t)\rangle dt$ is approximated by $(v_{1W_1}^*)^{\mathrm{T}}L_{C\times M}v_{1M}$. If this quantity is nonzero, then v_{1W}^* is rescaled so that $\int_0^1\langle v_1^*(t),v_1(t)\rangle dt=1$.

From Proposition A.1 it follows that we can approximate φ_1^* like v_1^*. To be precise, we compute φ_{1W}^* by solving

$$(\varphi_{1W}^*)^{\mathrm{T}}\left[\begin{array}{c}(D-TA(t))_{C\times M}\\ \delta_0-\delta_1\end{array}\right]=0,$$

and we normalize $\varphi_{1W_1}^*$ by requiring that $\sum_{i=0}^{N-1}\sum_{j=1}^{m}\left|(\varphi_{1W_1}^*)_{i,j}\right|_1=1$. Then the integral $\int_0^1\langle\varphi_1^*(t),F(u_1(t))\rangle dt$ is approximated by $(\varphi_{1W_1}^*)^{\mathrm{T}}(F(u_1(t)))_C$. If this quantity is nonzero, then φ_{1W}^* is rescaled, so that $\int_0^1\langle\varphi_1^*(t),F(u_1(t))\rangle dt=1$. We compute $(h_{20,1})_M$ by solving

$$\left[\begin{array}{c}(D-TA(t)+2i\theta I_n)_{C\times M}\\ \delta_0-\delta_1\end{array}\right](h_{20,1})_M=\left[\begin{array}{c}B_{C\times M\times M}v_{1M}v_{1M}\\ 0\end{array}\right].$$

The coefficient a_1 can be approximated using (3.21) as

$$a_1=(\varphi_{W_1}^*)^{\mathrm{T}}B_{C\times M\times M}v_{1M}\bar{v}_{1M},$$

while $(h_{11,1})_M$ is found by solving the discretization of (3.22),

$$\left[\begin{array}{c}(D-TA(t))_{C\times M}\\ \delta_0-\delta_1\\ (\varphi_{W_1}^*)^{\mathrm{T}}L_{C\times M}\end{array}\right](h_{11,1})_M=\left[\begin{array}{c}B_{C\times M\times M}v_{1M}\bar{v}_{1M}-a_1(F(u_1(t)))_C\\ 0\\ 0\end{array}\right].$$

The normal form coefficient d defined by (3.23) is approximated by

(3.25)

$$\begin{aligned} d\ &=\ \frac{1}{2}(v_{1W_1}^*)^{\mathrm{T}}(B_{C\times M\times M}(h_{11,1})_M v_{1M}+B_{C\times M\times M}(h_{20,1})_M\bar{v}_{1M})\\ &\quad+\frac{1}{2T}(v_{1W_1}^*)^{\mathrm{T}}C_{C\times M\times M\times M}v_{1M}v_{1M}\bar{v}_{1M}-\frac{a_1}{T}(v_{1W_1}^*)^{\mathrm{T}}A_{C\times M}v_{1M}+\frac{ia_1\theta}{T^2}.\end{aligned}$$

4. Examples. The computations in this section are done with MATCONT [20]. In particular, the bordering methods from [23] are used to continue the codim 1 bifurcations of limit cycles in two parameters. The algorithms described above for computing the normal form coefficients are also implemented in the current version of MATCONT.

4.1. The LPC normal form coefficient in the ABC-reaction. We have computed the normal form coefficient b of (2.4) in a model of a continuously stirred tank reactor, with consecutive A → B → C reactions, as studied by Doedel and Heinemann [24]. It has three state variables, u_1, u_2, and u_3, and five parameters, p_1, p_2, p_3, p_4, and p_5:

$$\left\{ \begin{array}{lcl} \dot{u_1} & = & -u_1 + p_1(1-u_1)e^{u_3}, \\ \dot{u_2} & = & -u_2 + p_1(1-u_1-p_5u_2)e^{u_3}, \\ \dot{u_3} & = & -u_3 - p_3u_3 + p_1p_4(1-u_1+p_2p_5u_2)e^{u_3}. \end{array} \right. \tag{4.1}$$

This model is used as a demo in the AUTO manual [19]. In the notation of [24], we have $u_1 = y$, where $1-y$ is the concentration of reactant A; $u_2 = z$, the concentration of reactant B; $u_3 = \theta$, the temperature; $p_1 = D$, the Damkohler number; $p_2 = \alpha$, the ratio of reaction heats; $p_3 = \beta$, the heat transfer coefficient; $p_4 = B$, the adiabatic temperature rise; and $p_5 = \sigma$, the selectivity ratio.

In Figure 4.1 the equilibrium curve computed with MATCONT is represented. The parameter values are $p_2 = 1$, $p_3 = 1.5$, $p_4 = 8$, $p_5 = 0.04$, with free parameter p_1, starting from the equilibrium at $p_1 = 0.1$, for which $u_1 \approx 0.13304$, $u_2 \approx 0.13223$, and $u_3 \approx 0.42833$. The curve of equilibria contains four Hopf points, labeled by H, which we call, from left to right, $\mathsf{H}_1, \mathsf{H}_2, \mathsf{H}_3, \mathsf{H}_4$, respectively. As shown in [24], in the case $p_2 = 1$, the Hopf points H_1 and H_4 are connected by a family of periodic solutions, and H_2 and H_3 are similarly connected. The family of solutions that connects H_1 to H_4 contains three fold bifurcations of periodic solutions, as also observed in [24]. We continue the first fold bifurcation of periodic solutions numerically in two parameters p_1 and p_2, with the discretization parameters $N = 30$ (mesh intervals) and $m = 4$ (collocation points). This family contains a cusp point of periodic orbits (CPC) detected in MATCONT as a zero of the coefficient b computed with (3.8). In Figure 4.2(a)–(c) we present the normal form coefficient b, the first component u_1 of the state variables vector, and p_2, respectively, as functions of p_1.

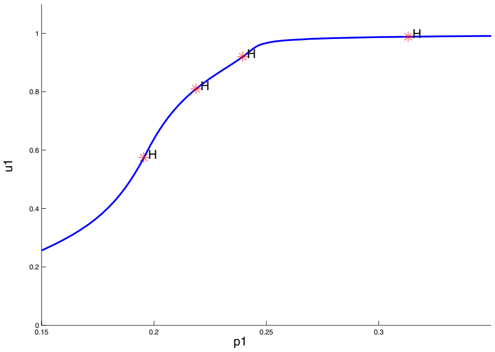
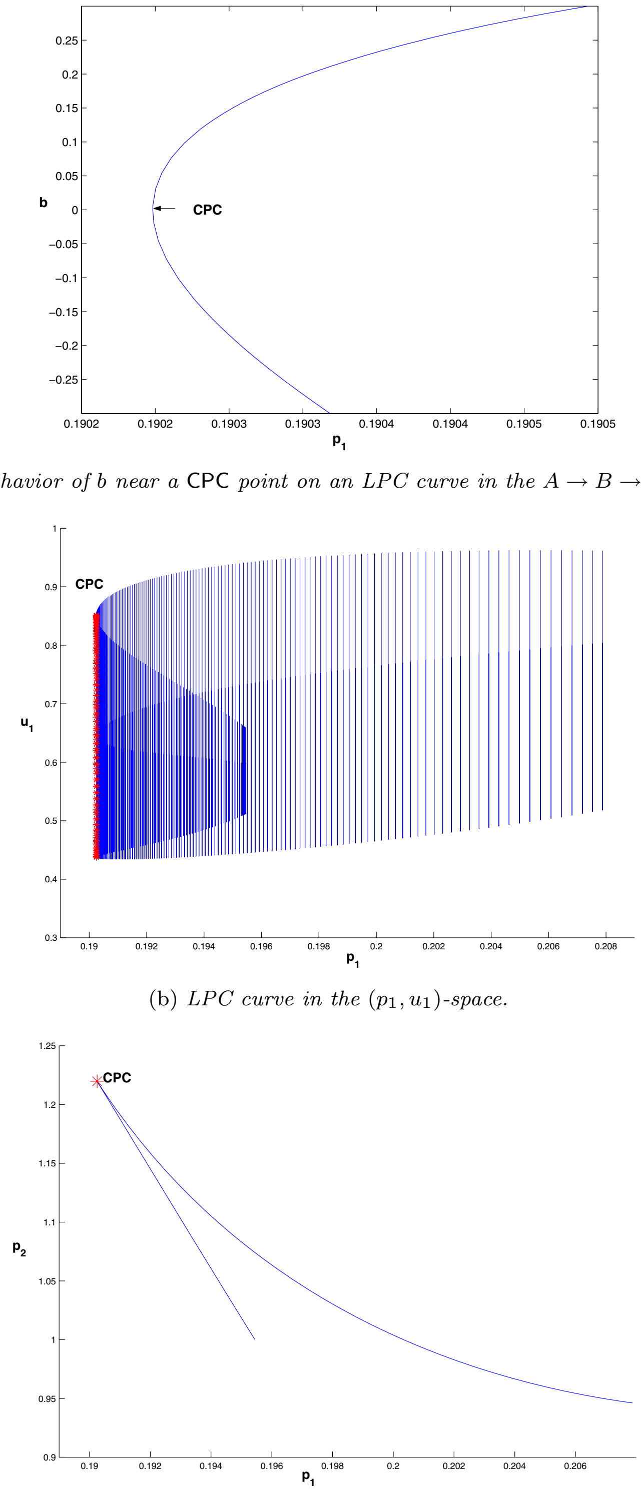


FIG. 4.1. *Equilibrium curve of the* $A \to B \to C$ *reaction for* $p_2 = 1$.

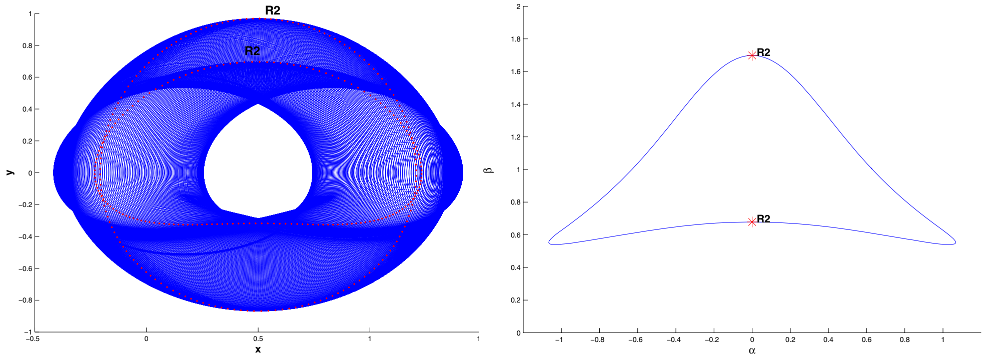


(a) *The behavior of b near a* CPC *point on an LPC curve in the $A \to B \to C$ reaction.*

(b) *LPC curve in the (p_1, u_1)-space.*

(c) *LPC curve in the (p_1, p_2)-space.*

FIG. 4.2.



(a) *PD curve with two* $1:2$ *points at* $\alpha = 0$.

(b) *The family in the* (α, β)*-space.*

Fig. 4.3.

4.2. The PD normal form coefficient in a feedback control system. We have used (3.16) to compute the PD normal form coefficient c of (2.5) in a feedback control system, described in [25, 9] and further used in [3, Example 5.4, p. 178]:

$$\left\{\begin{array}{rcl} \dot{x} & = & y, \\ \dot{y} & = & z, \\ \dot{z} & = & -\alpha z - \beta y - x + x^2. \end{array}\right. \tag{4.2}$$

Due to the special structure of this system, a good approximation to the PD curve can be found by the harmonic balance method; cf. [26, 27].

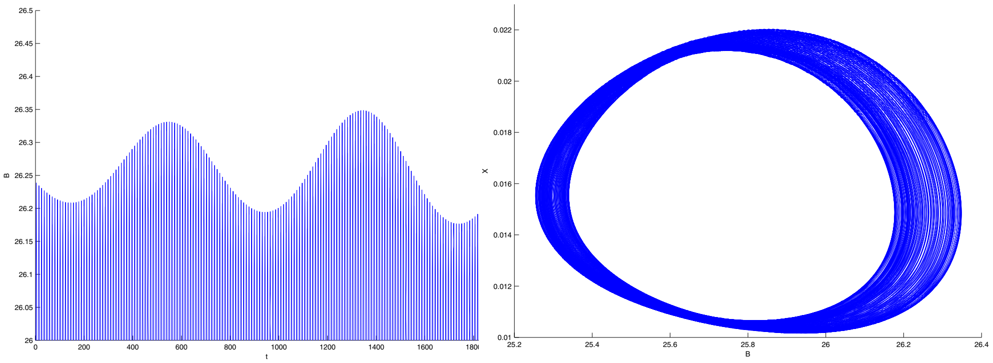
We have computed a family of periodic orbits of (4.2) numerically, as described in the CL_MATCONT manual and in [28], starting from the Hopf point for $\alpha = 1$ and $\beta = 1$ at $(0, 0, 0)$. We used $N = 20$ (mesh intervals) and $m = 4$ (collocation points) for the discretization. We detected two PD points with period $6.36407\dots$ at $\alpha \approx \pm 0.6303020$, respectively. The noncritical multipliers at the first PD point are inside the unit circle, so the periodic orbit could be stable. At the second PD point there is one multiplier outside the unit circle, and therefore the orbit is unstable. At the PD points the normal form coefficients c were computed. At the first PD point we find that $c \approx -0.04267737 < 0$. Therefore, the critical periodic orbit at the first PD point is stable, and a stable limit cycle with approximately double period exists for nearby parameter values. This was confirmed by computing the bifurcating periodic orbit and its multipliers. At the second PD point the normal form coefficient is $c \approx 0.04268605 > 0$. Hence the periodic orbit with double period is unstable in the center manifold. By computing the orbit with doubled period and monitoring the multipliers near this second PD point, we found that it has indeed two multipliers outside the unit circle. From the first PD point we computed the branch of PD cycles (see Figure 4.3). The normal form coefficient c is used as a test function. We also use $I = \int_0^1 \langle v_1^*(t), v_1(t) \rangle dt$ as another test function. We detected (at $\alpha = 0$) two strong 1:2 resonances R2 on this curve, where there are two multipliers equal to -1.

4.3. The NS normal form coefficient in a chemical model. The following model of the peroxidase-oxidase reaction was studied by Steinmetz and Larter [29]:

$$\left\{\begin{array}{rcl} \dot{A} & = & -k_1 ABX - k_3 ABY + k_7 - k_{-7} A, \\ \dot{B} & = & -k_1 ABX - k_3 ABY + k_8, \\ \dot{X} & = & k_1 ABX - 2k_2 X^2 + 2k_3 ABY - k_4 X + k_6, \\ \dot{Y} & = & -k_3 ABY + 2k_2 X^2 - k_5 Y, \end{array}\right. \tag{4.3}$$

where A, B, X, Y are state variables and k_1, k_2, k_3, k_4, k_5, k_6, k_7, k_8, and k_{-7} are parameters. The following (approximate) values correspond to an unstable equilibrium in (4.3):

Variable	Value	Parameter	Value
A	31.78997	k_1	0.1631021
B	1.45468	k_2	1250
X	0.01524586	k_3	0.046875
Y	0.1776113	k_4	20
		k_5	1.104
		k_6	0.001
		k_7	4.235322
		k_8	0.5
		k_{-7}	0.1175

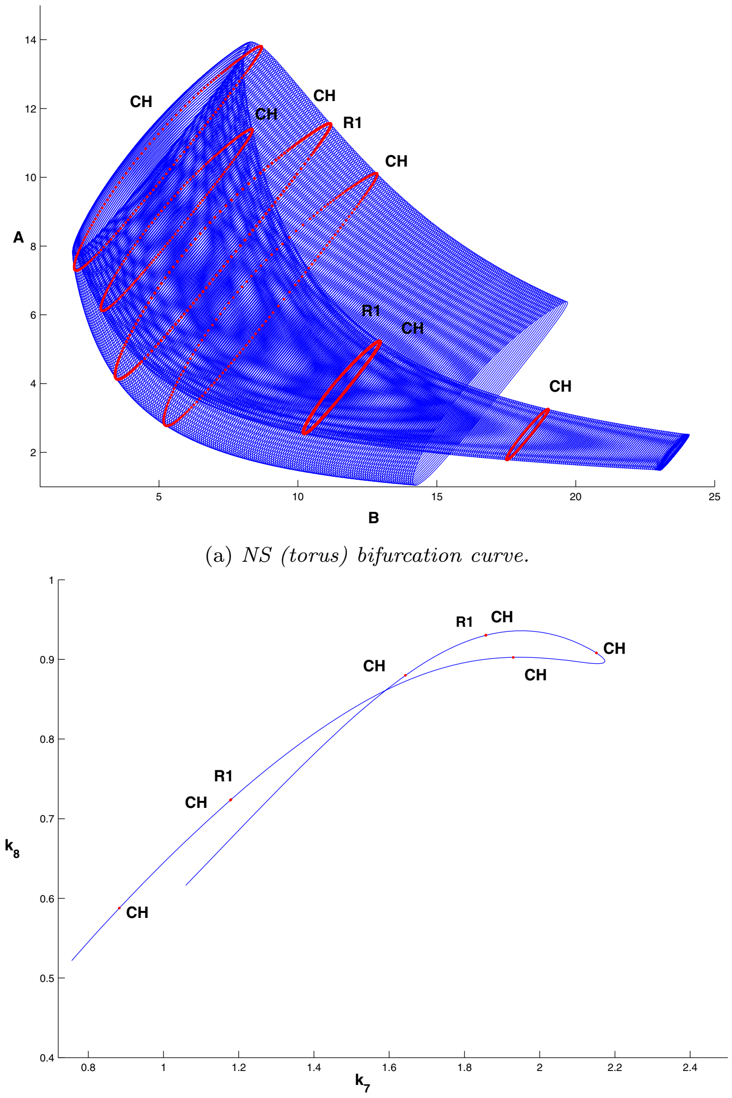


(a) *Modulated oscillations.*

(b) *Orbits on a stable 2-torus.*

FIG. 4.4.

We continued this equilibrium with decreasing k_7, keeping all other parameters fixed. We found a Hopf point at $k_7 \approx 0.712475$, where the first Lyapunov coefficient is negative. We then computed the family of stable limit cycles that bifurcates from the Hopf point. At $k_7 \approx 0.716434$ a *torus* (NS) bifurcation occurs. The real part of normal form coefficient d of (2.6) at this point is Re $d \approx -1.405999 \cdot 10^{-6}$, and therefore the emanating tori would be stable, locally. If we start a time integration from a point on the critical limit cycle, with a slightly increased parameter value, namely, $k_7 = 0.7167$, then after a transient period the orbit exhibits modulated oscillations with two frequencies near the limit cycle (see Figure 4.4). This is a motion on a stable two-dimensional torus that arises from the NS bifurcation. The NS point can be used as a starting point for the two-parameter continuation of the corresponding codim 1 bifurcation, using k_7 and k_8 as control parameters and $N = 50$, $m = 4$ as discretization parameters. We monitored Re d of the normal form coefficient d, computed with (3.25), during this continuation; it vanishes in a Chenciner bifurcation point (CH). The computed bifurcation curve is presented in Figures 4.5(a) and 4.5(b), in the (A, B)-plane and in the (k_7, k_8)-plane, respectively. The NS curve contains two additional codim 2 points, where a triple multiplier $\mu = 1$ is present (also counting the trivial multiplier). These are 1:1 strong resonance points [3] denoted by R1 in the figure. Between the 1:1 points, the *NS* curve is a *neutral saddle cycle* curve. Near such codim 2 points complicated homoclinic structures exist.



(a) *NS (torus) bifurcation curve.*

(b) *Codim 2 bifurcation points on the NS curve.*

FIG. 4.5.

It should be noted that the algorithm for the NS continuation, as implemented in MATCONT, is sufficiently robust to pass through the 1:1 resonance points (within a 10^{-3} parameter-range).

5. Discussion. The formulas for the normal form coefficients derived in this paper allow numerical verification of the nondegeneracy conditions (see [3]) for all codim 1 limit cycle bifurcations. In particular, the coefficients for the PD and torus bifurcations allow one to distinguish between sub- and supercritical cases.

These coefficients serve as test functions for detecting codim 2 bifurcations of limit cycles. One may try to use them to set up defining equations for the corresponding codim 2 bifurcations in three control parameters. However, any Newton-like continuation scheme would require the derivatives of the coefficients with respect to parameters, period, and the discretization variables. This problem requires further analysis.

The new algorithms fit very well into the BVP-framework (see [30, 23]) of AUTO [19], CONTENT [11], and, particularly, MATCONT [20], which contains our current prototype implementation.

The underlying technique can also be used to derive the coefficients of the periodic normal forms for codim 2 singularities of limit cycles. Although periodic normal forms are known in most of these codim 2 cases (see [8, 4]), substantial work remains to be done on the derivation and implementation of formulas for their coefficients. When implemented, such formulas will allow one to verify the nondegeneracy conditions for the codim 2 bifurcations.

A comparison of the numerical periodic normalization proposed in the current paper with the computation of normal form coefficients of the Poincaré map via automatic differentiation is also a task in future work.

Appendix A. Kernels of some differential-difference operators. In section 3 we used the orthogonality with respect to the following inner product: If $\zeta_1, \zeta_2 \in \mathcal{C}^0([0,1], \mathbb{C}^n)$ and $\eta_1, \eta_2 \in \mathbb{C}^n$, then

$$\left\langle \begin{bmatrix} \zeta_1 \\ \eta_1 \end{bmatrix}, \begin{bmatrix} \zeta_2 \\ \eta_2 \end{bmatrix} \right\rangle = \int_0^1 \langle \zeta_1(t), \zeta_2(t) \rangle \, dt + \langle \eta_1, \eta_2 \rangle = \int_0^1 \zeta_1^{\mathrm{H}}(t)\zeta_2(t)dt + \eta_1^{\mathrm{H}}\eta_2.$$

If this inner product vanishes, then we say that the corresponding vectors are orthogonal and write

$$\begin{bmatrix} \zeta_1 \\ \eta_1 \end{bmatrix} \perp \begin{bmatrix} \zeta_2 \\ \eta_2 \end{bmatrix}.$$

In section 3 we also used the following propositions.

PROPOSITION A.1. *Consider two differential-difference operators*

$$\phi_{1,2} : \mathcal{C}^1([0,1], \mathbb{R}^n) \to \mathcal{C}^0([0,1], \mathbb{R}^n) \times \mathbb{R}^n,$$

where

$$\phi_1(\zeta) = \begin{bmatrix} \dot{\zeta} - TA\zeta \\ \zeta(0) - \zeta(1) \end{bmatrix}, \quad \phi_2(\zeta) = \begin{bmatrix} \dot{\zeta} + TA^{\mathrm{T}}\zeta \\ \zeta(0) - \zeta(1) \end{bmatrix}.$$

If $\zeta \in \mathcal{C}^1([0,1], \mathbb{R}^n)$, *then* $\zeta \in \mathrm{Ker}(\phi_1)$ *if and only if*

$$\begin{bmatrix} \zeta \\ \zeta(0) \end{bmatrix} \perp \phi_2(\mathcal{C}^1([0,1], \mathbb{R}^n)),$$

and $\zeta \in \mathrm{Ker}(\phi_2)$ *if and only if*

$$\begin{bmatrix} \zeta \\ \zeta(0) \end{bmatrix} \perp \phi_1(\mathcal{C}^1([0,1], \mathbb{R}^n)).$$

Proof. If ζ is in the kernel of ϕ_1, then $\dot{\zeta} - TA(t)\zeta = 0$ and $\zeta(0) - \zeta(1) = 0$. For all $g \in \mathcal{C}^1([0,1], \mathbb{R}^n)$ we have

$$\begin{aligned}
&\textstyle\int_0^1 g(t)^{\mathrm{T}}\dot{\zeta}(t)dt - \int_0^1 Tg(t)^{\mathrm{T}}A(t)\zeta(t)dt = 0\\
\Rightarrow&\\
&\textstyle g(t)^{\mathrm{T}}\zeta(t)|_0^1 - \int_0^1 \dot{g}(t)^{\mathrm{T}}\zeta(t)dt - \int_0^1 Tg(t)^{\mathrm{T}}A(t)\zeta(t)dt = 0\\
\Rightarrow&\\
&\textstyle g(1)^{\mathrm{T}}\zeta(1) - g(0)^{\mathrm{T}}\zeta(0) - \int_0^1 (\dot{g}(t) + TA(t)^{\mathrm{T}}g(t))^{\mathrm{T}}\zeta(t)dt = 0\\
\Rightarrow&\\
&\textstyle -(g(0) - g(1))^{\mathrm{T}}\zeta(0) - \int_0^1 (\dot{g}(t) + TA(t)^{\mathrm{T}}g(t))^{\mathrm{T}}\zeta(t)dt = 0\\
\Rightarrow&\\
&\left\langle \begin{bmatrix} \dot{g} + TA^{\mathrm{T}}g \\ g(0) - g(1) \end{bmatrix}, \begin{bmatrix} \zeta \\ \zeta(0) \end{bmatrix} \right\rangle = 0.
\end{aligned}$$

Conversely, assume that $\langle [\begin{smallmatrix} \zeta \\ \zeta(0) \end{smallmatrix}], [\begin{smallmatrix} \dot{g} + TA^{\mathrm{T}}g \\ g(0) - g(1) \end{smallmatrix}] \rangle = 0$ for all $g \in \mathcal{C}^1([0,1], \mathbb{R}^n)$. Then

$$\begin{aligned}
&\textstyle\int_0^1 \zeta^{\mathrm{T}}(t)(\dot{g}(t) + TA(t)^{\mathrm{T}}g(t))dt + \zeta^{\mathrm{T}}(0)(g(0) - g(1)) = 0\\
\Rightarrow&\\
&\textstyle \zeta(1)^{\mathrm{T}}g(1) - \zeta(0)^{\mathrm{T}}g(0) + \zeta(0)^{\mathrm{T}}(g(0) - g(1)) - \int_0^1 (\dot{\zeta}(t) - TA(t)\zeta(t))^{\mathrm{T}}g(t)dt = 0\\
\Rightarrow&\\
&\textstyle -(\zeta(0) - \zeta(1))^{\mathrm{T}}g(1) - \int_0^1 (\dot{\zeta}(t) - TA(t)\zeta(t))^{\mathrm{T}}g(t)dt = 0.
\end{aligned}$$

If $\dot{\zeta}(t) - TA(t)\zeta(t) \neq 0$, then there exists a $g(t)$ with $g(1) = 0$ such that

$$\int_0^1 (\dot{\zeta}(t) - TA(t)\zeta(t))^{\mathrm{T}}g(t)dt \neq 0.$$

This is impossible, so $\dot{\zeta}(t) + TA(t)^{\mathrm{T}}\zeta(t) = 0$. Hence $(\zeta(0) - \zeta(1))^{\mathrm{T}}g(1) = 0$ for all g; hence $\zeta(0) - \zeta(1) = 0$. From $\dot{\zeta}(t) - TA(t)\zeta(t) = 0$ and $\zeta(0) = \zeta(1)$ it follows that $\zeta \in \mathrm{Ker}(\phi_1)$.

The proof of the second assertion is similar. □

PROPOSITION A.2. *Consider* $\phi_{1,2} : \mathcal{C}^1([0,1], \mathbb{R}^n) \to \mathcal{C}^0([0,1], \mathbb{R}^n) \times \mathbb{R}^n$, *where*

$$\phi_1(\zeta) = \begin{bmatrix} \dot{\zeta} - TA\zeta \\ \zeta(0) + \zeta(1) \end{bmatrix}, \phi_2(\zeta) = \begin{bmatrix} \dot{\zeta} + TA^{\mathrm{T}}\zeta \\ \zeta(0) + \zeta(1) \end{bmatrix}.$$

If $\zeta \in \mathcal{C}^1([0,1], \mathbb{R}^n)$, *then* $\zeta \in \mathrm{Ker}(\phi_1)$ *if and only if*

$$\begin{bmatrix} \zeta \\ \zeta(0) \end{bmatrix} \perp \phi_2(\mathcal{C}^1([0,1], \mathbb{R}^n)),$$

and $\zeta \in Ker(\phi_2)$ *if and only if*

$$\begin{bmatrix} \zeta \\ \zeta(0) \end{bmatrix} \perp \phi_1(\mathcal{C}^1([0,1], \mathbb{R}^n)).$$

Proof. The proof is as in Proposition A.1. □

PROPOSITION A.3. *Consider* $\phi_{1,2} : \mathcal{C}^1([0,1], \mathbb{C}^n) \to \mathcal{C}^0([0,1], \mathbb{C}^n) \times \mathbb{C}^n$, *where*

$$\phi_1(\zeta) = \begin{bmatrix} \dot{\zeta} - TA\zeta + i\theta I_n \\ \zeta(0) - \zeta(1) \end{bmatrix}, \phi_2(\zeta) = \begin{bmatrix} \dot{\zeta} + TA^{\mathrm{T}}\zeta - i\theta I_n \\ \zeta(0) - \zeta(1) \end{bmatrix}.$$

If $\zeta \in \mathcal{C}^1([0,1],\mathbb{C}^n)$*, then* $\zeta \in \mathrm{Ker}(\phi_1)$ *if and only if*

$$\begin{bmatrix} \zeta \\ \zeta(0) \end{bmatrix} \perp \phi_2(\mathcal{C}^1([0,1],\mathbb{C}^n)),$$

and $\zeta \in \mathrm{Ker}(\phi_2)$ *if and only if*

$$\begin{bmatrix} \zeta \\ \zeta(0) \end{bmatrix} \perp \phi_1(\mathcal{C}^1([0,1],\mathbb{C}^n)).$$

Proof. The proof is as in Proposition A.1. □

REFERENCES

[1] G. Iooss, *Bifurcations of Maps and Applications*, North–Holland, Amsterdam, 1979.

[2] C. De Boor and B. Swartz, *Collocation at Gaussian points*, SIAM J. Numer. Anal., 10 (1973), pp. 582–606.

[3] Yu. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, 2nd ed., Springer-Verlag, New York, 1998.

[4] S.-N. Chow and D. Wang, *Normal forms of bifurcating periodic orbits*, in Multiparameter Bifurcation Theory, Contemp. Math. 56, AMS, Providence, RI, 1986, pp. 9–18.

[5] G. Iooss and D. Joseph, *Elementary Stability and Bifurcation Theory*, Springer-Verlag, New York, 1980.

[6] Yu. A. Kuznetsov, H. G. E. Meijer, and L. van Veen, *The fold-flip bifurcation*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 14 (2004), pp. 2253–2282.

[7] G. Iooss and M. Adelmeyer, *Topics in Bifurcation Theory and Applications*, Adv. Ser. Nonlinear Dynam. 3, World Scientific, River Edge, NJ, 1992.

[8] V. I. Arnold, *Geometrical Methods in the Theory of Ordinary Differential Equations*, Springer-Verlag, New York, Heidelberg, Berlin, 1983.

[9] R. Genesio, A. Tesi, and F. Villoresi, *Models of complex dynamics in nonlinear systems*, Systems Control Lett., 25 (1995), pp. 185–192.

[10] J. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*, Springer-Verlag, New York, 1983.

[11] Yu. A. Kuznetsov and V. V. Levitin, *CONTENT: A Multiplatform Environment for Analyzing Dynamical Systems*, Dynamical Systems Laboratory, Centrum voor Wiskunde en Informatica, Amsterdam, ftp.cwi.nl/pub/CONTENT (1997).

[12] A. Griewank, D. Juedes, and J. Utke, ADOL-C*: A Package for the Automatic Differentiation of Algorithms Written in* C/C++*, Version* 1.7, Argonne National Laboratory, Argonne, IL, 1996.

[13] J. Guckenheimer and B. Meloon, *Computing periodic orbits and their bifurcations via automatic differentiation*, SIAM J. Sci. Comput., 22 (2000), pp. 951–985.

[14] Yu. A. Kuznetsov and H. G. E. Meijer, *Numerical normal forms for codim* 2 *bifurcations of fixed points with at most two critical eigenvalues*, SIAM J. Sci. Comput., 26 (2005), pp. 1932–1954.

[15] C. Elphick, E. Tirapegui, M. E. Brachet, P. H. Coullet, and G. Iooss, *A simple global characterization for normal forms of singular vector fields*, Phys. D, 32 (1987), pp. 95–127.

[16] Yu. A. Kuznetsov, *Numerical normalization techniques for all codim* 2 *bifurcations of equilibria in ODEs*, SIAM J. Numer. Anal., 36 (1999), pp. 1104–1124.

[17] C. Elphick, G. Iooss, and E. Tirapegui, *Normal form reduction for time-periodically driven differential equations*, Phys. Lett. A, 120 (1987), pp. 459–463.

[18] G. Iooss, *Global characterization of the normal form for a vector field near a closed orbit*, J. Differential Equations, 76 (1988), pp. 47–76.

[19] E. J. Doedel, A. R. Champneys, T. F. Fairgrieve, Yu. A. Kuznetsov, B. Sandstede, and X. J. Wang, auto97*: Continuation and Bifurcation Software for Ordinary Differential Equations (with HomCont)*, Concordia University, Montreal, Canada, ftp.cs.concordia.ca/pub/doedel/auto (1997).

[20] A. Dhooge, W. Govaerts, and Yu. A. Kuznetsov, matcont*: A* matlab *package for numerical bifurcation analysis of ODEs*, ACM Trans. Math. Software, 29 (2003), pp. 141–164.

[21] U. M. Ascher, R. M. M. Mattheij, and R. D. Russell, *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations*, Classics Appl. Math. 13, SIAM, Philadelphia, 1995.

[22] E. Doedel, H. B. Keller, and J.-P. Kernévez, *Numerical analysis and control of bifurcation problems.* II. *Bifurcation in infinite dimensions*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 1 (1991), pp. 745–772.

[23] E. J. Doedel, W. Govaerts, and Yu. A. Kuznetsov, *Computation of periodic solution bifurcations in ODEs using bordered systems*, SIAM J. Numer. Anal., 41 (2003), pp. 401–435.

[24] E. J. Doedel and R. F. Heinemann, *Numerical computation of periodic solution branches and oscillatory dynamics of the stirred tank reactor with $A \rightarrow B \rightarrow C$ reactions*, Chem. Eng. Sci., 38 (1983), pp. 1493–1499.

[25] R. Genesio and A. Tesi, *Harmonic balance methods for the analysis of chaotic dynamics in nonlinear systems*, Automatica, 28 (1992), pp. 531–548.

[26] A. Tesi, E. H. Abed, R. Genesio, and H. O. Wang, *Harmonic balance analysis of period-doubling bifurcations with implications for control of nonlinear dynamics*, Automatica, 32 (1996), pp. 1255–1271.

[27] G. Torrini, R. Genesio, and A. Tesi, *On the computation of characteristic multipliers for predicting limit cycle bifurcations*, Chaos Solitons Fractals, 9 (1998), pp. 121–133.

[28] A. Dhooge, W. Govaerts, Yu. A. Kuznetsov, W. Mestrom, and A. M. Riet, cl_matcont*: A continuation toolbox in* matlab, in Proceedings of the 2003 ACM Symposium on Applied Computing, Melbourne, FL, 2003, pp. 161–166.

[29] C. G. Steinmetz and R. Larter, *The quasiperiodic route to chaos in a model of the peroxidase-oxidase reaction*, J. Chem. Phys., 74 (1991), pp. 1388–1396.

[30] W.-J. Beyn, A. Champneys, E. Doedel, W. Govaerts, Yu. A. Kuznetsov, and B. Sandstede, *Numerical continuation, and computation of normal forms*, in Handbook of Dynamical Systems, Vol. 2, B. Fiedler, ed., North–Holland, Amsterdam, 2002, pp. 149–219.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1436–1454

WEAK BV CONVERGENCE OF A MOVING FINITE-ELEMENT METHOD FOR SINGULAR AXISYMMETRIC HARMONIC MAPS[*]

MORGAN PIERRE[†]

Abstract. We prove the convergence of a moving finite-element method for the minimization of a relaxed Dirichlet energy among axisymmetric maps from the disc into the sphere. The optimal mesh and the discrete minimizer exist for a discrete relaxed energy. Similarly to the true solution, the discrete minimizer has a boundary layer. Because of the consistency error introduced by the discretization of the energy, this discrete minimizer is nonconforming. We show that it converges to the solution of the continuous problem in an appropriate BV space. The proof is done for three different piecewise linear finite-element discretizations.

Key words. moving mesh, finite-element methods, harmonic maps, BV functions

AMS subject classifications. 26A45, 53C43, 65N12, 65N50

DOI. 10.1137/040605035

1. Introduction.

Moving mesh methods have been developed in the past three decades for problems whose solutions have sharp layers (see [3, 11, 17] and the references therein). They consist in moving automatically the nodes of a given initial mesh to new locations, the number and connectivity of nodes being fixed. For steady-state problems in which the true solution minimizes an energy functional, there exists a natural optimality criterion for the positioning of the nodes, which has been successfully used by several authors [5, 9, 10, 15]. The idea is to look for the triangulation that minimizes the discrete energy obtained by a finite-element discretization (moving finite elements). In other words, we deal with a constrained optimization problem: the unknowns are, as usual, the values of the discrete solution at the nodes (for a piecewise linear approximation), but also the position of the nodes; the constraints are given by the topology of the triangulation.

From a computational point of view, one major difficulty is the strong nonlinearity of the problem with respect to the nodes; another difficulty is the possibility of tangling. From a mathematical point of view, the existence of an optimal triangulation is a consequence of the continuity of the energy, if one allows degenerate triangles [4]; the optimal triangulation may not be unique. Furthermore, the discrete solution computed together with the optimal mesh converges to the true solution as the number of nodes tends to infinity. This assertion is an immediate consequence of a density result, but it holds only when the elements are conforming and when the energy is computed exactly. If the energy is computed with a quadrature formula, as this happens most often, it may be false because of the consistency error. For instance, if a discrete solution has a discrete energy level lower than the energy level of the true solution, then for every optimal mesh, the discrete energy level will be even lower. To our knowledge, there is no study of the convergence of an optimal mesh method that takes into consideration the consistency error.

[*]Received by the editors March 11, 2004; accepted for publication (in revised form) March 18, 2005; published electronically October 19, 2005.
http://www.siam.org/journals/sinum/43-4/60503.html

[†]CMLA, Ecole Normale Supérieure de Cachan, 61 av. du Président Wilson, 94235 Cachan Cedex, France. Current address: Laboratoire de Mathématiques, Université de Poitiers, Bd Marie et Pierre Curie BP 30179, 86962 Futuroscope Cedex, France (pierre@math.univ-poitiers.fr).

The aim of this paper is precisely to prove the convergence of an optimal mesh method in the presence of a consistency error. The solution of the continuous problem that we consider minimizes a relaxed Dirichlet energy among axisymmetric maps from the disc into the sphere. It is defined on $[0, 1]$ and has a "boundary layer of zero thickness" at 0 or, equivalently, a discontinuity. Because of the consistency error introduced by the discretization of the energy, the discrete minimizer is nonconforming; the main difficulty is to find appropriate error estimators.

The continuous problem has been studied by Alouges and the author [2]. Harmonic maps into the unit sphere are a simplified model for liquid crystals (see, for instance, [1]). The singularities arising in the context of harmonic maps can be roughly depicted as "boundary layers with zero thickness." In order to describe them, Giaquinta, Modica, and Souček [7, 8] have proposed to consider the graph of the unknown as a Cartesian current and to understand the convergence of minimizing currents. In our problem, we were able to simplify this approach by using weak *BV* convergence (which, in codimension 1, is equivalent to the convergence of Cartesian currents [7]).

In [2], we also computed the optimal mesh and we showed numerical evidence of the convergence of the solution as the number of nodes tends to infinity. The energy of the piecewise linear approximation could not be computed exactly because of a cosine term, so we used two different quadrature formulas: the midpoint formula (see Figures 4.1 and 4.2) and the Gaussian formula with two nodes. The optimal mesh and the discrete minimizer exist for a discrete relaxed energy. Similarly to the true solution, the graph of the discrete solution has a vertical part at 0. In order to compute this, we used an appropriate change of variable for the mesh variables and we implemented a projected conjugate gradient algorithm.

The paper is organized as follows. For the reader's convenience, we sum up in section 2 the main results obtained in [2] for the continuous problem. In section 3, we prove the convergence of the moving finite-element method, under the assumption that the discrete energy is computed exactly: the elements here are conforming. In section 4, the elements are nonconforming because of the midpoint formula; we introduce an external approximation [14] in the *BV* space, which allows us to prove the convergence. In section 5, we deal with the midpoint formula for the S^1 formulation of the problem. In section 6, we sum up the main results and we discuss the extension of the method in higher dimension.

2. The continuous problem. A map $u : B^2 \to S^2$ from the unit disc into the unit sphere is called *axially symmetric* if there exists an angle function $\theta : [0, 1] \to \mathbb{R}$ (θ is the latitude) which depends only on $r := \sqrt{x^2 + y^2}$ such that

$$u(x, y) = \left(\cos\theta(r)\frac{x}{r}, \cos\theta(r)\frac{y}{r}, \sin\theta(r)\right) \quad \forall (x, y) \in B^2.$$

If u is axisymmetric with angle function $\theta(r)$, its Dirichlet energy is

$$E(\theta) := \pi \int_0^1 \frac{\cos^2\theta}{r} + r\theta'^2 dr = \frac{1}{2}\int_{B^2} |\nabla u|^2 dxdy =: E(u). \tag{2.1}$$

Define the space of bounded energy functions with boundary condition α

$$\mathcal{E}_\alpha := \left\{\theta \in \mathcal{C}^0(]0, 1]),\ \theta(1) = \alpha,\ \frac{\cos\theta}{\sqrt{r}} \in L^2(0, 1),\ \text{and}\ \sqrt{r}\theta' \in L^2(0, 1)\right\}. \tag{2.2}$$

For every $\theta \in \mathcal{E}_\alpha$ there exists $k \in \mathbb{Z}$ such that $\lim_{r\to 0}\theta(r) = -\pi/2 + k\pi$. The integer k is called the $1d$ degree of θ. Thus $\mathcal{E}_\alpha = \bigcup_{k\in\mathbb{Z}} \mathcal{E}_{\alpha,k}$, where

$$\mathcal{E}_{\alpha,k} := \{\theta \in \mathcal{E}_\alpha,\ \theta(0) = -\pi/2 + k\pi\}. \tag{2.3}$$

We consider the following minimization problem:

$$\text{Minimize } E(\theta) \text{ in } \mathcal{E}_{\alpha,k}. \tag{2.4}$$

In general, the class $\mathcal{E}_{\alpha,k}$ does not have a minimizer. The theory for the two-dimensional Dirichlet problem (see [16] for instance) explains this by an energy concentration at $r = 0$ for minimizing sequences (*bubbling*). We were able to describe in a BV space this phenomenon of loss of compactness by concentration.

Recall that if $I =]a, b[$ is an open bounded interval,

$$BV(I) := \{y \in L^1(I),\ |y'|_{BV(I)} < \infty\},$$

where for all $y \in L^1(I)$, the total mass of y' is defined by

$$|y'|_{BV(I)} := \sup\left\{\int_I y(r)\varphi'(r)dr \mid \varphi \in \mathcal{C}_c^\infty(I,\mathbb{R}),\ ||\varphi||_\infty \le 1\right\}.$$

The space $BV(I)$ is a Banach space for the norm $\|y_n\|_{BV(I)} = |y'|_{BV(I)} + \|y\|_{L^1(I)}$. We will use the following.

Definition 2.1 (weak BV convergence). *Let $(y_n)_n$ be a sequence of functions in $BV(I)$ and let $y \in BV(I)$. Then $(y_n)_n$ converges weakly to y in $BV(I)$ if $\sup_n |y_n'|_{BV(I)} < \infty$ and $y_n \xrightarrow[n\to\infty]{} y$ strongly in $L^1(I)$. In this case, we have the following bound: $|y'|_{BV(I)} \le \liminf_n |y_n'|_{BV(I)}$.*

The injection $BV(I) \subset L^1(I)$ is compact [7], so every sequence $(y_n)_n$ in $BV(I)$ such that $\sup_n \|y_n\|_{BV(I)} < \infty$ converges, up to a subsequence, weakly in $BV(I)$ to some $y \in BV(I)$.

Turning back to (2.4), let $F \in \mathcal{C}^1(\mathbb{R},\mathbb{R})$ be the change of variable defined by

$$F(0) = 0 \text{ and } F'(\varphi) = |\cos\varphi| \quad \forall\varphi \in \mathbb{R}. \tag{2.5}$$

We recall that $F(\varphi + l\pi) = (\sin\varphi) + 2l$ for all $\varphi \in [-\pi/2, \pi/2]$ and for all $l \in \mathbb{Z}$. Then a fundamental inequality is the lower bound of the energy by the degree.

Lemma 2.2 (Alouges and Pierre). *For every $\theta \in \mathcal{E}_{\alpha,k}$,*

$$E(\theta) \ge 2\pi \int_0^1 |\theta' \cos\theta| dr \ge 2\pi \left|F(-\pi/2 + k\pi) - F(\alpha)\right|. \tag{2.6}$$

Moreover, $\Lambda_{\alpha,k} := \inf_{\theta\in\mathcal{E}_{\alpha,k}} E(\theta) = 2\pi\,|F(-\pi/2 + k\pi) - F(\alpha)|$.

From (2.6) it is clear that $F(\theta)$ belongs to $W^{1,1}(]0,1[) \subset BV(]0,1[)$ for every $\theta \in \mathcal{E}_{\alpha,k}$. In order to describe the concentration at $r = 0$ as the jump of a BV function, we included $[0,1]$ in $[-1,1]$ by symmetrization. For all $y := F \circ \theta$ with $\theta \in \mathcal{E}_\alpha$, its extension of degree k on $[-1,1]$ is the BV function built as

$$P_k(y)(r) := \begin{cases} y(r) & \text{if} \quad r \in [0,1], \\ -y(-r) + 4k - 2 & \text{if} \quad r \in [-1,0[. \end{cases}$$

Notice that if $\theta \in \mathcal{E}_{\alpha,k}$, then $y := P_k \circ F(\theta)$ is continuous on $[-1,1]$. Now we define as subsets of $BV(]-1,1[)$ the class $\mathcal{E}_{\alpha,k}$ and its "closure":

$$\mathcal{Y}_{\alpha,k} := P_k \circ F(\mathcal{E}_{\alpha,k}) \quad \text{and} \quad \overline{\mathcal{Y}}_{\alpha,k} := P_k \circ F(\mathcal{E}_\alpha). \tag{2.7}$$

The *BV* frame allowed us to introduce the following *relaxed energy*:

$$\overline{E}(y) := \inf\left\{\liminf_n E(y_n) \mid y_n \in \mathcal{Y}_{\alpha,k},\ y_n \rightharpoonup y \text{ weakly in } BV(]-1,1[)\right\} \quad \forall y \in \overline{\mathcal{Y}}_{\alpha,k}. \tag{2.8}$$

The relaxed energy is a lower semicontinuous (l.s.c.) extension of E called the *Lebesgue extension of* E [8]. In the following we use the notation $y = y_{\theta,k}$ for every $y \in \overline{\mathcal{Y}}_{\alpha,k}$ such that $y = P_k \circ F(\theta)$. In [2], we proved that

$$\overline{E}(y_{\theta,k}) = E(\theta) + 4|\theta(0) - (-\pi/2 + k\pi)| \qquad (y_{\theta,k} \in \overline{\mathcal{Y}}_{\alpha,k}). \tag{2.9}$$

With these definitions we obtained the existence and uniqueness.

THEOREM 2.3 (Alouges and Pierre). *Let* $(\alpha,k) \in [-\pi/2, 0] \times \mathbb{Z}$. *Then* $\inf_{y\in\overline{\mathcal{Y}}_{\alpha,k}} \overline{E}(y) = \inf_{\theta\in\mathcal{E}_{\alpha,k}} E(\theta)$, *and there exists a unique minimizer* $y_{\underline{\theta},k}$ *for* $\overline{E}$ *in* $\overline{\mathcal{Y}}_{\alpha,k}$. *Moreover, if* $(y_{\theta_n,k})_n$ *is a minimizing sequence for* $\overline{E}$ *in* $\overline{\mathcal{Y}}_{\alpha,k}$, *then* $y_{\theta_n,k} \underset{n\to\infty}{\rightharpoonup} y_{\underline{\theta},k}$ *weakly in* $BV(]-1,1[)$ *and* $\sqrt{r}\theta'_n \underset{n\to\infty}{\rightharpoonup} \sqrt{r}\underline{\theta}'$ *weakly in* $L^2(0,1)$.

In Theorem 2.3 the solution $y_{\underline{\theta},k}$ is regular (i.e., continuous on $[-1,1]$) if and only if $(\alpha,k) = (-\pi/2, 0)$ or $(\alpha,k) \in]-\pi/2, 0] \times \{0,1\}$. This corresponds exactly to the cases where problem (2.4) has a solution. In this case any minimizing sequence $(\theta_n)_n$ in $\mathcal{E}_{\alpha,k}$ converges strongly in the sense that $E(\theta_n) \xrightarrow[n\to\infty]{} E(\underline{\theta})$. We point out for the latter purpose that the minimizer $y_{\underline{\theta},k}$ is known explicitly. In particular, for $\alpha = -\pi/2$ then $\underline{\theta} \equiv -\pi/2$. If $\alpha \in]-\pi/2, 0]$, then $\underline{\theta} = \theta_{\alpha,1}$ for $k \geq 1$ and $\underline{\theta} = \theta_{\alpha,0}$ for $k \leq 0$, where $\theta_{\alpha,1}$ and $\theta_{\alpha,0}$ can be computed from

$$\theta' = \varepsilon \cos\theta \quad (0 < r \leq 1), \quad \theta(1) = \alpha, \tag{2.10}$$

with $\varepsilon = +1$ for $\theta_{\alpha,1}$ and $\varepsilon = -1$ for $\theta_{\alpha,0}$, respectively.

3. Conforming moving finite elements.

3.1. Moving finite elements and discrete relaxed energy. We use moving continuous P_1 elements to discretize the space $\mathcal{E}_{\alpha,k}$. Let

$$\mathcal{E}^N_{\alpha,k} := \Big\{\theta^N \in \mathcal{C}^0([0,1]),\ \theta^N(0) = -\frac{\pi}{2} + k\pi,\ \theta^N(1) = \alpha, \quad \text{and there exist}$$
$$r_0 = 0 < \cdots < r_N = 1 \text{ such that } \theta^N \text{ is affine on every } [r_i, r_{i+1}]\Big\}.$$

Since $\mathcal{E}^N_{\alpha,k} \subset \mathcal{E}_{\alpha,k}$, the elements here are *conforming* [14]. A function $\theta^N \in \mathcal{E}^N_{\alpha,k}$ is uniquely defined by the values $r_0 = 0 < \cdots < r_N = 1$ and $\theta_i := \theta^N(r_i)$ for $i \in \{0, \ldots, N\}$. Throughout section 3 we use the identification

$$\theta^N \simeq (r_0, r_1, \ldots, r_N, \theta_0, \ldots, \theta_N). \tag{3.1}$$

In particular, $\mathcal{E}^N_{\alpha,k} \simeq D^N$, where $D^N \subset \mathbb{R}^{2N+2}$ is defined by

$$D^N := \left\{r_0 = 0 < r_1 < \cdots < r_N = 1,\ \theta_0 = -\frac{\pi}{2} + k\pi, \theta_1, \ldots, \theta_N = \alpha\right\}. \tag{3.2}$$

In the remainder of this paper, we define various discrete energies E^N, which are consistent approximations of the exact energy E. A numerical solution to problem (2.4) is a minimizer of $E^N(r_0, r_1, \dots, r_N, \theta_0, \theta_1, \dots, \theta_N)$ in D^N under the constraints $r_0 = 0 < r_1 < \cdots < r_N = 1$. In order to guarantee the existence of a minimizer, we introduce an l.s.c. extension of E^N on the closure

$$\overline{D^N} = \left\{ r_0 = 0 \leq r_1 \leq \cdots \leq r_N = 1,\ \theta_0 = -\frac{\pi}{2} + k\pi, \theta_1, \dots, \theta_N = \alpha \right\}. \tag{3.3}$$

The following well-known lemma [6] gives such an extension (compare with (2.8)).

Lemma 3.1. *Let $D \subset \mathbb{R}^m$ and suppose $E : D \to [0, \infty)$ is continuous on D. Then $\overline{E} : \overline{D} \to [0, \infty]$, defined on the closure of D in $\mathbb{R}^m$ by*

$$\overline{E}(x) := \inf \left\{ \liminf_n E(x_n) \mid x_n \in D,\ x_n \to x \right\} \qquad (x \in \overline{D}),$$

is an extension of E which is l.s.c. on $\overline{D}$.

In the following we will extend the identification (3.1), in order to associate to every element $(r_0, \dots, \theta_N) \in \overline{D^N}$ its "P_1 interpolate" y^N, which is a BV function as in the continuous case.

3.2. Exact discretization. In the remainder of section 3 we suppose that E^N is exact. For every $\theta^N \in \mathcal{E}^N_{\alpha,k}$ that corresponds to $(r_0, \dots, r_N, \theta_0, \dots, \theta_N) \in D^N$, let

$$E^N(\theta^N) := \pi \sum_{i=0}^{N-1} \left(\int_{r_i}^{r_{i+1}} \frac{\cos^2 \theta^N}{r} dr \right) + \frac{(\theta_{i+1} - \theta_i)^2}{2} \left(\frac{r_{i+1} + r_i}{r_{i+1} - r_i} \right). \tag{3.4}$$

It is clear from (3.4) that E^N is continuous on D^N. Lemma 3.1 gives an l.s.c. extension $\overline{E^N}$ of E^N on $\overline{D^N}$. Clearly,

$$E^N(r_0, \dots, \theta_N) \geq \frac{\pi}{2} \sum_{i=0}^{N-1} (\theta_{i+1} - \theta_i)^2, \tag{3.5}$$

and by continuity of the right-hand side, the same estimate holds for $\overline{E^N}$. Thus, every minimizing sequence for $\overline{E^N}$ in $\overline{D^N}$ is bounded, and converges up to a subsequence to a minimizer. In particular, there exists a minimizer for $\overline{E^N}$ in $\overline{D^N}$.

Extending the identification $\mathcal{E}^N_{\alpha,k} \simeq D^N$, we want to see the minimizer for $\overline{E^N}$ as an element of $BV(]-1, 1[)$ as in the continuous case (2.7). Let

$$\mathcal{Y}^N_{\alpha,k} := P_k \circ F\big(\mathcal{E}^N_{\alpha,k}\big) \subset W^{1,1}(]-1, 1[) \tag{3.6}$$

denote the class $\mathcal{E}^N_{\alpha,k}$ in coordinates y. For every $y^N \in \mathcal{Y}^N_{\alpha,k}$ we denote $E^N(y^N) := E^N(\theta^N)$ where θ^N is the unique element in $\mathcal{E}^N_{\alpha,k}$ such that $y^N = P_k \circ F(\theta^N)$. The closure of $\mathcal{Y}^N_{\alpha,k}$ is defined by

$$\overline{\mathcal{Y}^N_{\alpha,k}} := \Big\{ y^N \in \overline{\mathcal{Y}}_{\alpha,k} \mid \exists\, y_n \in \mathcal{Y}^N_{\alpha,k},\ \sup_n E^N(y_n) < \infty \text{ and } y_n \underset{n\to\infty}{\rightharpoonup} y^N \text{ weakly in } BV(]-1, 1[) \Big\}. \tag{3.7}$$

We can now see the discrete relaxed energy as an analogue of (2.8).

Definition 3.2 (discrete relaxed energy). *For every $y^N \in \overline{\mathcal{Y}^N_{\alpha,k}}$,*

$$\overline{\overline{E^N}}(y^N) := \inf \left\{ \liminf_n E^N(y_n) \mid y_n \in \mathcal{Y}^N_{\alpha,k},\ y_n \underset{n\to\infty}{\rightharpoonup} y^N \text{ weakly in } BV \right\}. \tag{3.8}$$

Since $E^N(\theta^N) = E(\theta^N)$ for all $\theta^N \in \mathcal{E}^N_{\alpha,k}$, it is clear that $\overline{\mathcal{Y}^N_{\alpha,k}} \subset \overline{\mathcal{Y}}_{\alpha,k}$ and $\overline{E}(y^N) \leq \overline{\overline{E^N}}(y^N)$ for all $y^N \in \overline{\mathcal{Y}^N_{\alpha,k}}$. Moreover, by extending the identification $\mathcal{E}^N_{\alpha,k} \simeq D^N$ it is easy to see that

$$\overline{\mathcal{Y}^N_{\alpha,k}} \simeq \left\{ (r_0, \dots, \theta_N) \in \overline{D^N} \mid \overline{E^N}(r_0, \dots, \theta_N) < \infty \right\}$$

and that both definitions for the relaxed energy are equivalent, i.e., $\overline{\overline{E^N}}(y^N) = \overline{E^N}(r_0, \dots, \theta_N)$ for all $y^N \in \overline{\mathcal{Y}^N_{\alpha,k}}$, which corresponds to $(r_0, \dots, \theta_N) \in \overline{D^N}$. In the following we denote $\overline{\overline{E^N}} = \overline{E^N}$.

In order to prove that the discrete minimizer for the exact energy tends to the continuous minimizer when $N \to \infty$, we only need, by Theorem 2.3, to build a minimizing sequence made of P_1 elements. This is an easy density.

Lemma 3.3. *There exists a minimizing sequence for E in $\mathcal{E}_{\alpha,k}$ made of continuous P_1 finite elements.*

Proof. By P_1 interpolation [14], the energy of any regular function in $\mathcal{E}_{\alpha,k}$ can be approximated as closely as desired by a P_1 function in $\mathcal{E}_{\alpha,k}$. Clearly, $\inf_{\mathcal{E}_{\alpha,k}} E = \inf_{\mathcal{C}^2([0,1]) \cap \mathcal{E}_{\alpha,k}} E$, and this concludes the proof. □

As a consequence, we have the following proposition.

Proposition 3.4. *For every integer N let y^N be a minimizer for $\overline{E^N}$ in $\overline{\mathcal{Y}^N_{\alpha,k}}$. Then $(y^N)_N$ is a minimizing sequence for $\overline{E}$ in $\overline{\mathcal{Y}}_{\alpha,k}$. In particular, $y^N \underset{N\to\infty}{\rightharpoonup} \underline{y}$ weakly in $BV(]-1,1[)$, where $\underline{y}$ is the unique minimizer for $\overline{E}$ in $\overline{\mathcal{Y}}_{\alpha,k}$.*

4. Midpoint formula.

We use the same notation as in section 3.1. The difficulty now is that the integral in formula (3.4) cannot be computed exactly. It is necessary to use a quadrature formula, which introduces a *consistency error.* In section 4 we choose the midpoint formula. For every $\theta^N \in \mathcal{E}^N_{\alpha,k}$, which corresponds to $(r_0, \dots, \theta_N) \in D^N$,

$$E^N(\theta^N) := 2\pi \sum_{i=0}^{N-1} E_i(\theta^N), \tag{4.1}$$

where for all $i \in \{0, \dots, N-1\}$

$$E_i(\theta^N) := \left[\cos^2\left(\frac{\theta_i + \theta_{i+1}}{2}\right) \left(\frac{r_{i+1} - r_i}{r_{i+1} + r_i}\right) + \left(\frac{\theta_{i+1} - \theta_i}{2}\right)^2 \left(\frac{r_{i+1} + r_i}{r_{i+1} - r_i}\right) \right]. \tag{4.2}$$

The function E_i is defined on $\mathcal{E}^N_{\alpha,k} \simeq D^N$ and depends only on θ_i, θ_{i+1} and the ratio $t_i = r_i / r_{i+1}$.

It is clear from (4.2) that E^N is continuous on D^N. Lemma 3.1 gives an l.s.c. extension $\overline{E^N}$ of E^N on $\overline{D^N}$. By the same argument as in section 3.2, there exists a minimizer for $\overline{E^N}$ in $\overline{D^N}$. Figures 4.1 and 4.2 show the corresponding minimizer obtained in [2] for $\alpha = -\pi/4$, $k = 3$, $N = 23$ and $\alpha = -\pi/2$, $k = 1$, $N = 10$, respectively. Notice that a vertical part in the graph of the minimizer can only occur at $r = 0$, since if $r_i = r_{i+1} > 0$ and $\overline{E^N}(\theta^N) < \infty$, then $\theta_i = \theta_{i+1}$.

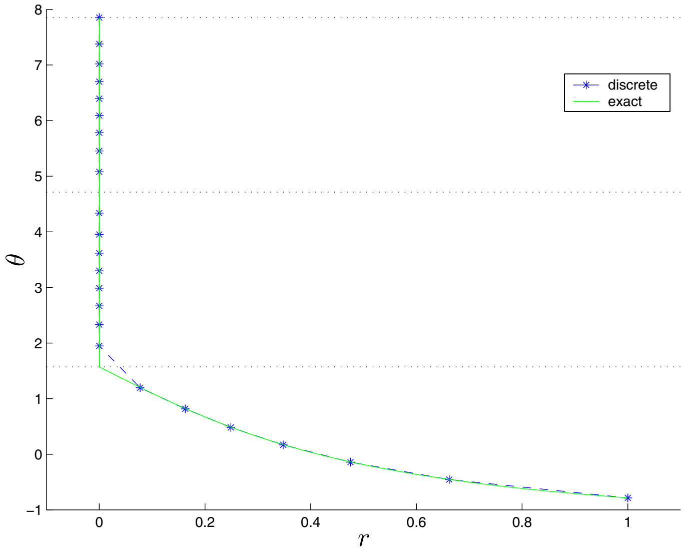


FIG. 4.1. *Midpoint for* $\alpha = -\pi/4$, $k = 3$, $N = 23$.

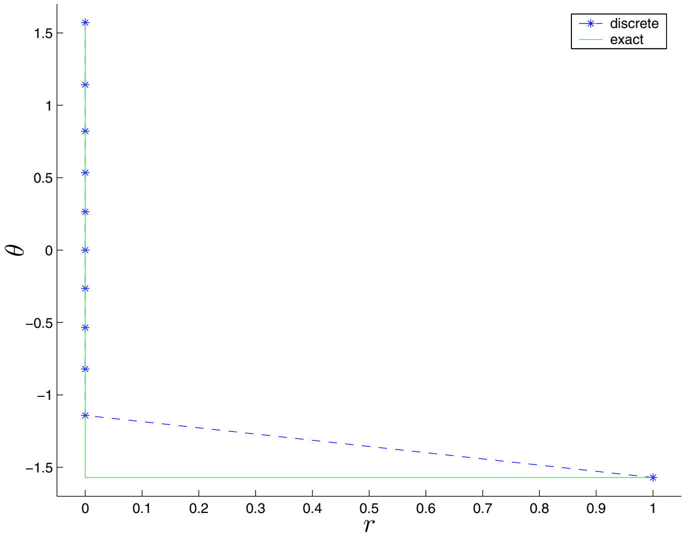


FIG. 4.2. *Midpoint for* $\alpha = -\pi/2$, $k = 1$, $N = 10$.

4.1. Discrete Euler–Lagrange equation. The first fundamental idea in the rest of section 4 is the change of variable

$$X_i := \frac{r_{i+1} - r_i}{r_{i+1} + r_i} = \frac{1 - t_i}{1 + t_i}. \tag{4.3}$$

With this formulation,

$$E_i(X_i, \theta_i, \theta_{i+1}) = \cos^2\left(\frac{\theta_i + \theta_{i+1}}{2}\right) X_i + \left(\frac{\theta_{i+1} - \theta_i}{2}\right)^2 \frac{1}{X_i}. \tag{4.4}$$

We recall that the values $r_0 = 0$, $r_N = 1$, $\theta_0 = -\pi/2 + k\pi$, and $\theta_N = \alpha$ are fixed, so $t_0 = 0$ and $X_0 = 1$. Thus E^N is a function of $(X_1, \ldots, X_{N-1}, \theta_1, \ldots, \theta_{N-1})$ denoted E_X^N and defined on $]0,1[^{N-1} \times \mathbb{R}^{N-1}$. It is clear that the map

$$\begin{aligned} D^N \cap \mathbb{R}^{N-1} &\to]0,1[^{N-1}, \\ (r_1, \ldots, r_{N-1}) &\to \left(t_1, \ldots, t_i = \frac{r_i}{r_{i+1}}, \ldots, t_{N-1}\right) \end{aligned} \tag{4.5}$$

is a smooth diffeomorphism: for all $(t_1, \ldots, t_{N-1}) \in]0,1[^{N-1}$ the inverse in $D^N \cap \mathbb{R}^{N-1}$ is computed by $r_i = t_i r_{i+1}$ for $i = N-1, N-2, \ldots, 1$. Thus

$$\inf_{D^N} E^N(r_0, \ldots, \theta_N) = \inf_{]0,1[^{N-1} \times \mathbb{R}^{N-1}} E_X^N(X_1, \ldots, X_{N-1}, \theta_1, \ldots, \theta_{N-1}). \tag{4.6}$$

The second fundamental idea is a well-known inequality, but we state it as a lemma because of its importance. The proof is immediate.

LEMMA 4.1. *Let $(a,b) \in \mathbb{R}^2$. Then*

$$\inf_{X \in]0,1]} \left(a^2 X + \frac{b^2}{X}\right) = \begin{cases} 2|ab| & \textit{if} \quad |a| > |b|, \ \textit{obtained for } X = \dfrac{|b|}{|a|}, \\ a^2 + b^2 & \textit{if} \quad |a| \leq |b|, \ \textit{obtained for } X = 1. \end{cases}$$

In particular we have $\inf_{X_i \in]0,1[} E_i(X_i, \theta_i, \theta_{i+1}) = J(\theta_i, \theta_{i+1})$ for all (θ_i, θ_{i+1}), where $J : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$J(\theta, \tilde{\theta}) := \begin{cases} 2\dfrac{|\tilde{\theta} - \theta|}{2}\left|\cos\left(\dfrac{\theta + \tilde{\theta}}{2}\right)\right| & \text{if } \dfrac{|\tilde{\theta} - \theta|}{2} \leq \left|\cos\left(\dfrac{\theta + \tilde{\theta}}{2}\right)\right|, \\ \left(\dfrac{\tilde{\theta} - \theta}{2}\right)^2 + \cos^2\left(\dfrac{\theta + \tilde{\theta}}{2}\right) & \text{if } \dfrac{|\tilde{\theta} - \theta|}{2} \geq \left|\cos\left(\dfrac{\theta + \tilde{\theta}}{2}\right)\right|. \end{cases} \tag{4.7}$$

The following result is a discrete version of the Euler–Lagrange equation (2.10).

PROPOSITION 4.2. *Let $(r_i^N, \theta_i^N)_{0 \leq i \leq N}$ be a minimizer for $\overline{E^N}$ in $\overline{D^N}$, with $\theta_N = \alpha$. Then for all $i \in \{0, \ldots, N-1\}$ such that $r_{i+1} > 0$ and*

$$(\theta_i \neq \theta_{i+1}) \ \textit{or} \ \left(\theta_i = \theta_{i+1} \not\equiv \frac{\pi}{2} \pmod{\pi}\right), \tag{4.8}$$

$$\frac{r_{i+1} - r_i}{r_{i+1} + r_i} = \begin{cases} \dfrac{|\theta_{i+1} - \theta_i|}{2\left|\cos\left(\frac{\theta_i + \theta_{i+1}}{2}\right)\right|} & \textit{if } \dfrac{|\theta_{i+1} - \theta_i|}{2} < \left|\cos\left(\dfrac{\theta_i + \theta_{i+1}}{2}\right)\right|, \\ 1 & \textit{if } \dfrac{|\theta_{i+1} - \theta_i|}{2} \geq \left|\cos\left(\dfrac{\theta_i + \theta_{i+1}}{2}\right)\right|. \end{cases} \tag{4.9}$$

Proof. Lemma 4.1 and equality (4.6) imply

$$\overline{E^N}(r_0, \ldots, \theta_N) = 2\pi \sum_{i=0}^{N-1} J(\theta_i, \theta_{i+1}). \tag{4.10}$$

For all $i \in \{0,\dots,N\}$ let $\overline{E_i}$ be the l.s.c. extension of E_i on $\overline{D^N}$ from Lemma 3.1. It is clear from the definition that $\overline{E^N}(r_0,\dots,\theta_N) \geq 2\pi\sum_{i=0}^{N-1}\overline{E_i}(r_0,\dots,\theta_N)$, and this together with Lemma 4.1 implies $\overline{E_i}(r_0,\dots,\theta_N) = J(\theta_i,\theta_{i+1})$ for all i.

Now assume there exists $i \in \{0,\dots,N-1\}$ with $r_{i+1} > 0$ such that (4.8) is satisfied and (4.9) is not satisfied. Necessarily, $r_i < r_{i+1}$, otherwise $r_i = r_{i+1} > 0$ would imply $\theta_i = \theta_{i+1}$, with $\cos((\theta_i+\theta_{i+1})/2) \neq 0$ by (4.8), and (4.9) would be satisfied. So X_i belongs to $]0,1]$. We apply Lemma 4.1 by noticing that the minimizer $X \in]0,1]$ given in the lemma is unique except for $a = b = 0$. This is not possible here by assumption (4.8) and X_i is not the minimizer since (4.9) is not satisfied, so $E_i(X_i,\theta_i,\theta_{i+1}) > J(\theta_i,\theta_{i+1})$. It is clear by continuity that $E_i(X_i,\theta_i,\theta_{i+1}) = \overline{E_i}(r_i,r_{i+1},\theta_i,\theta_{i+1})$, hence a contradiction, and this concludes the proof. □

In agreement with Figure 4.2 we have the following.

Corollary 4.3. *If $\alpha = -\pi/2$ and $(r_i^N,\theta_i^N)_{0\leq i\leq N}$ is a minimizer for $\overline{E^N}$ in $\overline{D^N}$, then $r_0 = r_1 = \cdots = r_{i_N} = 0$, where $i_N := \max\{i \mid \theta_i \neq -\pi/2\}$.*

Proof. If $r_{i_N+1} = 0$, then the assertion is proved. Otherwise, $r_{i_N+1} > 0$, and since $\theta_{i_N} \neq -\pi/2$, $\theta_{i_N+1} = -\pi/2$, we can apply Proposition 4.2 with $i = i_N$. We find $(r_{i_N+1} - r_{i_N})/(r_{i_N+1} + r_{i_N}) = 1$, so $r_{i_N} = 0$ as expected. □

Remark. The numerical computations show that $(\theta_i^N)_{0\leq i\leq N}$ is monotone, but we have not been able to prove it.

4.2. Upper bound for the degree. Similarly to the continuous case (2.6), the energy E^N defined by the midpoint formula (4.1) is greater than the degree. The proof of this assertion is based on the following (stronger) lemma. We recall that J is defined by (4.7) and F by (2.5).

Lemma 4.4. *Let $(\theta,\tilde{\theta}) \in \mathbb{R}^2$. If $\theta \neq \tilde{\theta}$, then $J(\theta,\tilde{\theta}) > |F(\theta) - F(\tilde{\theta})|$.*

Proof. First assume there exists $l \in \mathbb{Z}$ such that $[\theta,\tilde{\theta}] \subset [\pi/2 + l\pi, \pi/2 + (l+1)\pi]$. Then $|\tilde{\theta} - \theta|/2 > |\sin((\tilde{\theta}-\theta)/2)|$ and $\cos((\theta+\tilde{\theta})/2) \neq 0$, so by Lemma 4.1

$$J(\theta,\tilde{\theta}) > 2\left|\sin\left(\frac{\tilde{\theta}-\theta}{2}\right)\right|\left|\cos\left(\frac{\theta+\tilde{\theta}}{2}\right)\right|. \tag{4.11}$$

The right-hand side of (4.11) being equal to $|\sin\tilde{\theta} - \sin\theta| = |F(\theta) - F(\tilde{\theta})|$, Lemma 4.4 is proved.

If the previous assumption is not satisfied, there exists $l \in \mathbb{Z}$ such that

$$\frac{\pi}{2} + l\pi \in]\theta,\tilde{\theta}[. \tag{4.12}$$

Let $m \geq 1$ denote the number of $l \in \mathbb{Z}$ that satisfy (4.12). Assume $\theta < \tilde{\theta}$ (the proof for $\theta > \tilde{\theta}$ is similar). Then we have a monotone sequence

$$\theta < \frac{\pi}{2} + l\pi < \frac{\pi}{2} + (l+1)\pi < \cdots < \frac{\pi}{2} + (l+m-1)\pi < \tilde{\theta},$$

where l is the smallest integer that satisfies (4.12). We define $h := \pi/2 + l\pi - \theta$ and $k := \tilde{\theta} - (\pi/2 + (l+m-1)\pi)$. Notice that $h \in]-0,\pi]$ and $k \in]0,\pi]$. Then,

$$\begin{aligned}\left|F(\theta) - F(\tilde{\theta})\right| &= |1 - \cos h| + 2(m-1) + |1 - \cos k|,\\ &= 2(m-1) + 2\sin^2\left(\frac{h}{2}\right) + 2\sin^2\left(\frac{k}{2}\right).\end{aligned} \tag{4.13}$$

On the other hand,

$$J(\theta,\tilde{\theta}) = \inf_{X\in]0,1]} \cos^2\left(m\frac{\pi}{2}+\frac{h-k}{2}\right)X + \left(\frac{h+k+(m-1)\pi}{2}\right)^2\frac{1}{X}.$$

Case $m=1$. Since $\sin^2(\frac{h-k}{2}) \le |\frac{h-k}{2}|^2 \le (\frac{h+k}{2})^2$, from Lemma 4.1

$$J(\theta,\tilde{\theta}) = \sin^2\left(\frac{h-k}{2}\right) + \left(\frac{h+k}{2}\right)^2. \tag{4.14}$$

With (4.13) and (4.14), Lemma 4.5, which follows, concludes the proof for $m=1$.

Case $m\ge 2$. Since

$$\cos^2\left(\frac{m\pi}{2}+\frac{h-k}{2}\right) \le 1 \le \left(\frac{\pi}{2}\right)^2 \le \left(\frac{h+k+(m-1)\pi}{2}\right)^2,$$

with Lemma 4.1 we get

$$J(\theta,\tilde{\theta}) = \cos^2\left(\frac{m\pi}{2}+\frac{h-k}{2}\right) + \left(\frac{h+k+(m-1)\pi}{2}\right)^2. \tag{4.15}$$

If $m=2$, then from (4.13) and (4.15), Lemma 4.6 which follows, concludes the proof. If $m\ge 3$, then by (4.15),

$$J(\theta,\tilde{\theta}) \ge \left(\frac{(m-1)\pi}{2}\right)^2 > 2(m+1) \ge 2(m-1) + 2\sin^2\left(\frac{h}{2}\right) + 2\sin^2\left(\frac{k}{2}\right).$$

The relation (4.13) concludes the proof. □

LEMMA 4.5. *For all* $(h,k)\in[0,\pi]^2$ *with* $(h,k)\ne(0,0)$,

$$\sin^2\left(\frac{h-k}{2}\right) + \left(\frac{h+k}{2}\right)^2 > 2\sin^2\left(\frac{h}{2}\right) + 2\sin^2\left(\frac{k}{2}\right).$$

Proof. Let $G:\mathbb{R}^2\to\mathbb{R}$ be defined by

$$G(h,k) := \sin^2\left(\frac{h-k}{2}\right) + \left(\frac{h+k}{2}\right)^2 - 2\sin^2\left(\frac{h}{2}\right) - 2\sin^2\left(\frac{k}{2}\right).$$

Let $(\underline{h},\underline{k})$ be a minimizer for G in $[0,\pi]\times[0,\pi]$ and assume that $(\underline{h},\underline{k})\in]0,\pi[^2$. Then $(\partial_h G,\partial_k G)(\underline{h},\underline{k}) = (0,0)$. A computation gives

$$\partial_h G(h,k) = \frac{\sin(h-k)}{2} + \frac{k-h}{2} + (h-\sin h) \quad \forall (h,k)\in[0,\pi]^2.$$

If $\pi\ge k\ge h>0$, then $|\sin(h-k)|\le(k-h)$ and $h\ge\sin h$, so $\partial_h G(h,k)\ge k/2>0$: thus $\underline{k}<\underline{h}$. Using the symmetry $G(h,k)=G(k,h)$ the condition $\partial_k G(\underline{h},\underline{k})=0$ implies this time $\underline{k}>\underline{h}$, hence a contradiction. Therefore $(\underline{h},\underline{k})\in\partial\left([0,\pi]^2\right)$. Now from the preceding, we have $\partial_h G(h,\pi)\ge 0$ for all $h\in[0,\pi]$, so $G(h,\pi)\ge G(0,\pi)>0$ for all $h\in[0\pi]$. Moreover, $G(h,0)=(h/2)^2-\sin^2(h/2)>0$ for all $h\in]0,\pi]$. By symmetry $G(h,k)>0$ on $\partial\left([0,\pi]^2\right)-\{(0,0)\}$ and this concludes the proof. □

LEMMA 4.6. *For all* $(h,k)\in[0,\pi]^2$,

$$\cos^2\left(\frac{h-k}{2}\right) + \left(\frac{h+k+\pi}{2}\right)^2 > 2 + 2\sin^2\left(\frac{h}{2}\right) + 2\sin^2\left(\frac{k}{2}\right).$$

Proof. The proof is similar to the previous one. Let $G : \mathbb{R}^2 \to \mathbb{R}$ be defined by

$$G(h,k) := \cos^2\left(\frac{h-k}{2}\right) + \left(\frac{h+k+\pi}{2}\right)^2 - 2\sin^2\left(\frac{h}{2}\right) - 2\sin^2\left(\frac{k}{2}\right) - 2.$$

Let $(\underline{h},\underline{k})$ be a minimizer for G in $[0,\pi]\times[0,\pi]$. Since for all $(h,k)\in[0,\pi]^2$,

$$\begin{aligned}\partial_h G(h,k) &= -\frac{\sin(h-k)}{2} + \frac{h+k+\pi}{2} - \sin h,\\ &\geq \frac{\pi}{2} - \frac{\sin(h-k)}{2} - \sin h > 0,\end{aligned} \tag{4.16}$$

necessarily $(\underline{h},\underline{k}) \in \partial\left([0,\pi]^2\right)$. Moreover, from (4.16), $G(h,0) \geq G(0,0) > 0$ and $G(h,\pi) \geq G(0,\pi) > 0$ for all $h\in[0,\pi]$. The symmetry $G(h,k) = G(k,h)$ concludes the proof. □

Lemma 4.4 implies the upper bound on the degree (compare with Lemma 2.2).

PROPOSITION 4.7. *For all $\theta^N \in \mathcal{E}^N_{\alpha,k}$ which corresponds to $(r_0,\dots,\theta_N)\in D^N$,*

$$E^N(r_0,\dots,\theta_N) \geq 2\pi\int_0^1 |(\theta^N)'\cos\theta^N|dr \geq \Lambda_{\alpha,k}. \tag{4.17}$$

Proof. Let $\theta^N \in \mathcal{E}^N_{\alpha,k}$. Recall that $F(\theta)$ is a primitive of $|\cos\theta|$ on $\mathbb{R}$. On $[r_i, r_{i+1}]$, $(\theta^N)'$ is constant, so $\int_{r_i}^{r_{i+1}} |(\theta^N)'\cos\theta^N|dr = |F(\theta_{i+1}) - F(\theta_i)|$. On the other hand, by Lemma 4.1, for all $i\in\{0,\dots,N-1\}$, $E_i(\theta^N) \geq J(\theta_i,\theta_{i+1})$. Together with Lemma 4.4, a summation on all $i\in\{0,\dots,N-1\}$ concludes the proof (the value of the infimum $\Lambda_{\alpha,k} := \inf_{\mathcal{E}_{\alpha,k}} E$ is given in Lemma 2.2). □

As a consequence, we have the following lemma.

LEMMA 4.8. *For E^N given by the midpoint formula, $\inf_{D^N} E^N \xrightarrow[N\to\infty]{} \Lambda_{\alpha,k}$.*

Proof. By Proposition 4.7, it is sufficient to find a sequence $(\theta^N)_N$, with $\theta^N \in \mathcal{E}^N_{\alpha,k}$ for every N, such that $E^N(\theta^N) \xrightarrow[N\to\infty]{} \Lambda_{\alpha,k}$. Let $\epsilon > 0$. By Lemma 3.3, for N large enough, there exists $\theta^N \in \mathcal{E}^N_{\alpha,k}$, which corresponds to $(r_0,\dots,\theta_N)\in D^N$, such that $E(\theta^N) < \Lambda_{\alpha,k} + \epsilon$. Define $I^M\theta^N$ as the P_1 interpolate of θ^N with respect to the uniform subdivision of each $[r_i,r_{i+1}]$ into M segments. Then $I^M\theta^N$ belongs to $\mathcal{E}^{NM}_{\alpha,k}$ and for M large enough, $E^{NM}(I^M\theta^N) \leq E(\theta^N) + \epsilon \leq \Lambda_{\alpha,k} + 2\epsilon$, and this concludes the proof (recall that $(\inf_{D^N} E^N)_N$ is nonincreasing). □

4.3. Nonconforming moving elements.

In order to see the minimizer for $\overline{E^N}$ as a BV function, we use the y coordinates as in section 3.2. Recall that $\mathcal{Y}^N_{\alpha,k}$ (3.6) denotes the class $\mathcal{E}^N_{\alpha,k}$ in the y coordinates. For every $y^N \in \mathcal{Y}^N_{\alpha,k}$ we denote $E^N(y^N) := E^N(\theta^N)$, where $\theta^N \in \mathcal{E}^N_{\alpha,k}$ satisfies $y^N = P_k \circ F(\theta^N)$. We point out that, as a consequence of Proposition 4.7,

$$2\pi\int_0^1 |(y^N)'|dr \leq E^N(y^N) \quad \forall y^N \in \mathcal{Y}^N_{\alpha,k}. \tag{4.18}$$

The closure of $\mathcal{Y}^N_{\alpha,k}$ for the midpoint formula is defined by

$$\begin{aligned}\overline{\overline{\mathcal{Y}^N_{\alpha,k}}} := \Big\{y^N \in BV(]-1,1[) \mid \exists\, y_n \in \mathcal{Y}^N_{\alpha,k},\ \sup_n E^N(y_n) < \infty\\ \text{and } y_n \underset{n\to\infty}{\rightharpoonup} y^N \text{ weakly in } BV(]-1,1[)\Big\}.\end{aligned}$$

In order to describe $\overline{\overline{\mathcal{Y}^N_{\alpha,k}}}$ we introduce the following Hilbert space:

$$V_\alpha := \left\{\theta \in \mathcal{C}^0(]0,1]) \mid \theta(1) = \alpha,\ \sqrt{r}\theta' \in L^2(0,1)\right\}. \tag{4.19}$$

The space V_α is an affine Hilbert space for the norm $(\theta^2(1) + \int_0^1 r\theta'^2 dr)^{1/2}$, which is isomorphic to the space of radial functions in $H^1(B^2)$ with Dirichlet boundary condition α. The upper bound $E^N(\theta^N) \geq 2\pi \int_0^1 r(\theta^N)'^2 dr$ which is true for all $\theta^N \in \mathcal{E}^N_{\alpha,k}$ implies that for all $y^N \in \overline{\overline{\mathcal{Y}^N_{\alpha,k}}}$ there exists a unique $\theta \in V_\alpha$ such that $y^N = F \circ P_k(\theta)$. Moreover, θ is a continuous P_1 element with respect to a subdivision consisting of M segments for some $M \leq N$; θ is called the *regular part* of y^N. Since $E^N \neq E$, $\overline{\overline{\mathcal{Y}^N_{\alpha,k}}} \not\subset \overline{\mathcal{Y}}_{\alpha,k}$. The elements in $\overline{\overline{\mathcal{Y}^N_{\alpha,k}}}$ are *nonconforming.* However, $\overline{\overline{\mathcal{Y}^N_{\alpha,k}}} \subset F \circ P_k\,(V_\alpha)$ for all N and $\mathcal{E}_\alpha \subset V_\alpha$, so $\overline{\mathcal{Y}}_{\alpha,k} \subset F \circ P_k\,(V_\alpha)$: we have an *external approximation* of $\overline{\mathcal{Y}}_{\alpha,k}$ [14].

We define the discrete relaxed energy as in the continuous case (2.8).

Definition 4.9 (discrete relaxed energy). *For every $y^N \in \overline{\overline{\mathcal{Y}^N_{\alpha,k}}}$*

$$\overline{\overline{E^N}}(y^N) := \inf\left\{\liminf_n E^N(y_n) \mid y_n \in \mathcal{Y}^N_{\alpha,k},\ y_n \underset{n\to\infty}{\rightharpoonup} y^N \textit{ weakly in } BV\right\}. \tag{4.20}$$

By extending the identification $\mathcal{E}^N_{\alpha,k} \simeq D^N$ it is easy to show that

$$\overline{\overline{\mathcal{Y}^N_{\alpha,k}}} \simeq \left\{(r_0, \dots, \theta_N) \in \overline{D^N} \mid \overline{E^N}(r_0, \dots, \theta_N) < \infty\right\},$$

and that Definition 4.9 is equivalent to the one given by Lemma 3.1, i.e., $\overline{\overline{E^N}}(y^N) = \overline{E^N}(r_0, \dots, \theta_N)$ for all $y^N \in \overline{\overline{\mathcal{Y}^N_{\alpha,k}}}$, which corresponds to $(r_0, \dots, \theta_N) \in \overline{D^N}$. In the following we denote $\overline{\overline{E^N}} = \overline{E^N}$.

4.4. Convergence analysis. We begin by proving an intermediate convergence result.

Lemma 4.10. *Let $(r_i^N, \theta_i^N)_{0\leq i\leq N}$ be a minimizer for $\overline{E^N}$ in $\overline{D^N}$. Then*

$$\max_{0\leq i\leq N}\left\{|\theta^N_{i+1} - \theta^N_i|\right\} \xrightarrow[N\to\infty]{} 0. \tag{4.21}$$

Proof. Assume (4.21) is false. Then there exist a subsequence $\left(\theta_i^{N'}\right)_{0\leq i\leq N'}$ and $\beta > 0$ such that $\min_{N'} \max_{0\leq i<N'}\{|\theta^{N'}_{i+1} - \theta^{N'}_i|\} \geq \beta$. By definition (4.7), J is continuous on $\mathbb{R}^2$. Thus from Lemma 4.4 for every $M > \beta > 0$,

$$\epsilon(\beta, M) := \min_{\substack{|\tilde\theta-\theta|\geq\beta \\ |\theta|\leq M,\ |\tilde\theta|\leq M}} \left\{J(\theta, \tilde\theta) - |F(\tilde\theta) - F(\theta)|\right\} > 0. \tag{4.22}$$

Now let $M := \sup_{N'} \max_{0\leq i\leq N'} \left|\theta^{N'}_i\right|$. Since the corresponding sequence $(y^{N'})_{N'}$ is uniformly bounded in $BV(]-1,1[)$ by (4.18), we have $M < \infty$. Using successively (4.10), Lemma 4.4, and (4.22), we obtain

$$\overline{E^{N'}}\left(r_0^{N'}, \dots, \theta^{N'}_{N'}\right) \geq \Lambda_{\alpha,k} + 2\pi\epsilon(\beta, M).$$

On the other hand, from Lemma 4.8 we get for N' large enough

$$\overline{E^{N'}}\left(r_0^{N'}, \dots, \theta^{N'}_{N'}\right) < \Lambda_{\alpha,k} + 2\pi\epsilon(\beta, M),$$

a contradiction, and this concludes the proof of Lemma 4.10. □

We can state the main convergence result as follows.

THEOREM 4.11. *For every integer N, let y^N be a minimizer for $\overline{E^N}$ in $\overline{\overline{\mathcal{Y}^N_{\alpha,k}}}$ and let $\theta^N \in V_\alpha$ be the regular part of y^N defined in section 4.3. Then,*

$$\overline{E^N}(y^N) \xrightarrow[N\to\infty]{} \Lambda_{\alpha,k}, \tag{4.23}$$

$$y^N \quad\rightharpoonup\quad y \text{ weakly in } BV(]-1,1[), \tag{4.24}$$

$$\sqrt{r}(\theta^N)' \quad\rightharpoonup\quad \sqrt{r}\underline{\theta}' \text{ weakly in } L^2(0,1), \tag{4.25}$$

where $y = y_{\underline{\theta},k}$ is the unique minimizer for $\overline{E}$ in $\overline{\mathcal{Y}}_{\alpha,k}$.

Proof. Equation (4.23) is given by Lemma 4.8. Estimate (4.18) shows that $(y^N)_N$ is uniformly bounded in BV. Moreover, $\sqrt{r}(\theta^N)'$ is uniformly bounded in L^2. So there exist $\theta \in V_\alpha$ and $y = P_k \circ F(\theta) \in BV$ such that, up to a subsequence,

$$\begin{aligned} y^N &\underset{N\to\infty}{\rightharpoonup} y \text{ weakly in } BV(]-1,1[),\\ \sqrt{r}(\theta^N)' &\underset{N\to\infty}{\rightharpoonup} \sqrt{r}\theta' \text{ weakly in } L^2(]0,1[). \end{aligned} \tag{4.26}$$

Case $\alpha = -\pi/2$. Lemma 4.10 and Corollary 4.3 imply that θ^N converges uniformly to $-\pi/2$ on $[0,1]$, so $\theta \equiv -\pi/2$, and the uniqueness concludes the proof.

Case $\alpha \in]-\pi/2, \pi/2[$. We assume $k \geq 1$, the proof for $k \leq 0$ being similar. The idea is to pass at the limit in the discrete version of the Euler–Lagrange equation (4.9) in order to prove the uniqueness of the limit. Define

$$\underline{r} := \max\{r \in]0,1[, \theta(r) = \pm\pi/2\} < 1 \tag{4.27}$$

(if $\theta(r) \in]-\pi/2, \pi/2[$ for all $r \in]0,1]$, then we set $\underline{r} := 0$).

We first show that $\underline{r} := 0$. For all N let $\underline{r}^N := \max\{r \in]0,1[, \theta^N(r) = \pm\pi/2\} < 1$. Let $(r_0^N, \dots, r_N^N, \theta_0^N, \dots, \theta_N^N) \in \overline{D^N}$ be the minimizer in $\overline{D^N}$ corresponding to $y^N \in \overline{\overline{\mathcal{Y}^N_{\alpha,k}}}$. We define the important index

$$i_N := \min\left\{i \in \{0,\dots,N\} \mid r_i > \underline{r}^N\right\}. \tag{4.28}$$

The interest of i_N is the following: from the proof of Lemma 4.4, i_N is the unique index such that $r_{i_N}^N > 0$ and $r_{i_N-1}^N = 0$ (see Figure 4.1). Moreover, by Proposition 4.2

$$\left(\frac{r_{i+1}^N - r_i^N}{r_{i+1}^N + r_i^N}\right)\left|\cos\left(\frac{\theta_i^N + \theta_{i+1}^N}{2}\right)\right| = \frac{|\theta_{i+1}^N - \theta_i^N|}{2} \quad \forall i \in \{i_N, \dots, N-1\}. \tag{4.29}$$

Lemma 4.12, which follows the proof, gives a natural result concerning i_N.

Let $s \in]\underline{r}, 1[$ and define

$$\beta(s) := \min\left\{\left(\min_{r\in[s,1]} \theta(r) + \frac{\pi}{2}\right), \left(\frac{\pi}{2} - \max_{[s,1]} \theta(r)\right)\right\} > 0.$$

The compact inclusion $H^1([s,1]) \subset \mathcal{C}^0([s,1])$ implies that θ^N converges uniformly to θ on $[s,1]$. Thus for N large enough, $\theta^N(r) \in [-\pi/2 + \beta(s)/2, \pi/2 - \beta(s)/2]$ for all $r \in [s,1]$. In particular, $\min_{[s,1]} \cos\theta^N \geq \sin(\beta(s)/2) > 0$ for all N large enough. Using (4.29) and Lemma 4.10 we obtain

$$\max_{\substack{i\in\{0,\dots,N-1\}\\ r_i^N \geq s}} (r_{i+1}^N - r_i^N) \xrightarrow[N\to\infty]{} 0. \tag{4.30}$$

This means that we have a family of subdivisions whose maximal mesh-size tends to 0 as $N \to \infty$. More precisely, let $i_{s,N} := \min\left\{i \mid r_i^N > s\right\} \geq i_N$. It is clear by the same proof as for Lemma 4.12 that $r_{i_{s,N}}^N \xrightarrow[N\to\infty]{} s$. Similarly, setting $t \in \,]s,1[$ and letting $i_{t,N} := \min\{i \mid r_i^N > t\} \geq i_{s,N}$, we have $r_{i_{t,N}}^N \xrightarrow[N\to\infty]{} t$. Thus $s < r_{i_{s,N}}^N < \cdots < r_{i_{t,N}-1}^N \leq t$ is a family of subdivisions of $[s,t]$ whose maximum step-size tends to 0.

Now let $i \in \{i_{s,N}-1,\dots,N-1\}$ such that $r_{i+1}^N > r_i^N$ (the case $r_{i+1}^N = r_i^N$ is not a problem since $\theta_i^N = \theta_{i+1}^N$). On $[r_i^N, r_{i+1}^N]$, (4.29) can be written

$$\left(\frac{2}{r_i^N + r_{i+1}^N}\right)\cos^2\left(\frac{\theta_i^N + \theta_{i+1}^N}{2}\right) = \left(\frac{r_i^N + r_{i+1}^N}{2}\right)(\theta^N)'^2. \tag{4.31}$$

Multiplying by $(r_{i+1}^N - r_i^N)$ and summing on all $i_{s,N} \leq i < i_{t,N}$, we obtain

$$\int_s^t g^N(r)dr = \int_s^t r(\theta^N)'^2 dr + \gamma(s,t,N), \tag{4.32}$$

where g^N is the piecewise constant function given by the left-hand side of (4.31) on every $[r_i, r_{i+1}]$. Here $\gamma(s,t,N)$ corresponds to the integration on $[s, r_{i_{s,N}}^N] \cup [r_{i_{t,N}-1}^N, t]$ and $\gamma(s,t,N) \xrightarrow[N\to\infty]{} 0$ (s and t are fixed). The uniform convergence of θ^N on $[s,1]$, Lemma 4.12, and (4.30) imply that $g^N \xrightarrow[N\to\infty]{} \cos^2\theta/r$ uniformly on $[s,t]$. On the other hand, $\sqrt{r}(\theta^N)' \rightharpoonup \sqrt{r}\theta'$ weakly in $L^2(0,1)$, so by passing to the limit in (4.32) and using l.s.c. of the L^2 norm we get $\int_s^t r\theta'^2 dr \leq \int_s^t (\cos^2\theta/r)dr$. This is true for every $\underline{r} < s < t < 1$, so $r\theta'^2(r) \leq \cos^2\theta(r)/r$ for a.e. $r \in [\underline{r}, 1]$.

This is equivalent to

$$-\frac{\cos\theta}{r} \leq \theta' \leq \frac{\cos\theta}{r} \quad \text{for a.e. } r \in [\underline{r}, 1]. \tag{4.33}$$

The two solutions of the Cauchy problem corresponding to the two equality cases in (4.33) with initial condition $\theta(1) = \alpha$ are $\theta_{\alpha,0}$ and $\theta_{\alpha,1}$ (see (2.10)). The regularity of $(r,\Theta) \to \cos\Theta/r$ on $]0,\infty[\,\times\,\mathbb{R}$ implies $\theta_{\alpha,0}(r) \leq \theta(r) \leq \theta_{\alpha,1}(r)$ for all $r \in [\underline{r}, 1]$. Since $-\pi/2 < \theta_{\alpha,0}(r) < \theta_{\alpha,1}(r) < \pi/2$ for all $r \in \,]0,1]$ we obtain $\underline{r} = 0$ as expected.

Now set $\epsilon > 0$ and let $i_{\epsilon,N} := \min\{i \mid r_i^N > \epsilon\}$ again. Equality (4.10) yields

$$\frac{\overline{E^N}(y^N)}{2\pi} = \sum_{i<i_{\epsilon,N}} J(\theta_i^N, \theta_{i+1}^N) + \sum_{i\geq i_{\epsilon,N}} J(\theta_i^N, \theta_{i+1}^N). \tag{4.34}$$

Denote $S_{\epsilon,N}^+$ the previous sum on the indexes $i \geq i_{\epsilon,N}$. Then, using the previous function g^N, (4.31), and the definition (4.7) of J

$$\begin{aligned}\int_\epsilon^1 g^N(r)dr &= 2\sum_{i\geq i_{\epsilon,N}}\left(\frac{r_{i+1}^N - r_i^N}{r_{i+1}^N + r_i^N}\right)\cos^2\left(\frac{\theta_i^N + \theta_{i+1}^N}{2}\right) + \gamma(\epsilon,N)\\ &= S_{\epsilon,N}^+ + \gamma(\epsilon,N)\end{aligned}$$

Here $\gamma(\epsilon,N)$ corresponds to the integration on $[\epsilon, r_{i_{\epsilon,N}}^N]$ so $\gamma(\epsilon,N) \xrightarrow[N\to\infty]{} 0$ (ϵ is fixed). We deduce $S_{\epsilon,N}^+ \leq \overline{E^N}(y^N)/(2\pi) - \left|F(-\pi/2 + k\pi) - F(\theta^N(r_{i_{\epsilon,N}}^N)\right|$ from (4.34) and Lemma 4.4. Letting $N \to \infty$ and using the uniform convergence on $[\epsilon,1]$,

$$\int_\epsilon^1 \frac{\cos^2\theta}{r}dr \leq \frac{\Lambda_{\alpha,k}}{2\pi} - \left|F\left(-\frac{\pi}{2} + k\pi\right) - F(\theta(\epsilon))\right|.$$

In particular with (4.33), $E(\theta) \leq 2\pi \int_0^1 \cos^2\theta/r dr \leq \Lambda_{\alpha,k} < \infty$, so $\theta \in \mathcal{E}_{\alpha,l}$ with necessarily $l = 1$ or $l = 0$ since $\underline{r} = 0$. Moreover, using the continuity of θ at $r = 0$,

$$E(\theta) \leq \Lambda_{\alpha,k} - 2\pi|F(-\pi/2 + k\pi) - F(\theta(0))|.$$

The case $l = 0$ is impossible since $\Lambda_{\alpha,k} - 2\pi|F(-\pi/2 + k\pi) - F(-\pi/2)| < 0$ (see the remark following (2.6)). Thus $l = 1$, and since $\Lambda_{\alpha,k} - |F(-\pi/2+k\pi) - F(\pi/2)| = \Lambda_{\alpha,1}$, θ is a minimizer in $\mathcal{E}_{\alpha,1}$. In other words $\theta = \theta_{\alpha,1}$. The uniqueness of the limit θ concludes the proof. □

Lemma 4.12. *Let i_N be defined by* (4.28). *Then* $r^N_{i_N} \xrightarrow[N\to\infty]{} \underline{r}$.

Proof. First notice that the uniform convergence of θ^N on $[\underline{r}+\epsilon, 1]$ for every $\epsilon > 0$ implies that $\underline{r}^N \xrightarrow[N\to\infty]{} \underline{r}$. Assume by contradiction that $r^N_{i_N}$ does not converge to $\underline{r}$ as N tends to infinity. Since $\underline{r}^N < r^N_{i_N} \leq 1$, there exists a subsequence of indexes $(i_{N'})_{N'}$ such that $r^{N'}_{i_{N'}} \xrightarrow[N'\to\infty]{} b$ and $\theta^{N'}_{i_{N'}-1} \xrightarrow[N'\to\infty]{} \varphi$, with $\underline{r} < b \leq 1$ and $\varphi \notin]-\pi/2, \pi/2[$. Now $\theta^{N'}_{i_{N'}} = \theta^{N'}(r^N_{i_{N'}})$ tends to $\theta(b) \in]-\pi/2, \pi/2[$, so $|\theta^{N'}_{i_{N'}-1} - \theta^{N'}_{i_{N'}}| \xrightarrow[N\to\infty]{} |\varphi - b| \neq 0$. This contradicts Lemma 4.10 and concludes the proof. □

5. The S^1 formulation.

In this section, we apply the moving finite-element method to the S^1 formulation of problem (2.4): the energy is a quadratic functional of S^1-valued maps. This shows how the method could apply to the Dirichlet problem for S^2-valued maps in dimension 2 or 3 of domain, as considered in [8].

5.1. The continuous formulation.

Every function $\theta \in \mathcal{E}_{\alpha,k}$ can be seen as the S^1-valued map $u : [0,1] \to S^1 \subset \mathbb{R}^2$:

$$u(r) := (\cos\theta(r), \sin\theta(r)) =: (c(r), s(r)) \quad \forall r \in [0,1].$$

The Dirichlet energy with this formulation is given by

$$E(u) := \pi \int_0^1 \frac{c^2(r)}{r} + r|u'(r)|^2 dr. \tag{5.1}$$

This is a quadratic energy: it allows us to define the space of maps with bounded energy

$$H^1_{axi} := \left\{ u = (c,s) \in \mathcal{C}^0(]0,1], \mathbb{R}^2) \mid u(1) = (\cos\alpha, \sin\alpha),\ E(u) < \infty \right\},$$

which is an affine Hilbert space, isomorphic to the subspace of $H^1(B^2, \mathbb{R}^3)$ made of axisymmetric vector fields with boundary condition α. The identification $\theta \simeq u$ allows us to see $\mathcal{E}_{\alpha,k}$ as a subset of H^1_{axi}. We denote this subset by $\mathcal{U}_{\alpha,k}$.

One point of interest regarding this formulation is that $\mathcal{U}_{\alpha,k}$ has the metric induced by the norm in H^1_{axi} (whereas $\mathcal{E}_{\alpha,k}$ does not have a canonical metric). In fact, $\mathcal{U}_{\alpha,k}$ is connected and closed for the strong H^1_{axi} topology, but it is not closed for the weak H^1_{axi} topology. By Theorem 2.3, E has a minimizer in the class $\mathcal{U}_{\alpha,k}$ if and only if $(\alpha,k) = (-\pi/2, 0)$ or $(\alpha,k) \in]-\pi/2, 0] \times \{0,1\}$. If (α,k) does not satisfy this assumption, then a minimizing sequence in $\mathcal{U}_{\alpha,k}$ converges weakly in H^1_{axi}, but not strongly: a part of the energy concentrates at 0.

For the extension to higher dimension and for numerical applications, the S^1 formulation is interesting. However, it is not convenient for the mathematical description of the limit of minimizing sequences (singular minimizers). Since the codimension is 2, the boundary layer is not completely described by a discontinuity. The weak BV

convergence would have to be replaced by the convergence in the sense of currents, as in [8]. But even in this context, it is not clear how one could introduce the change of variable F.

5.2. Midpoint formula, conforming element, and discrete degree. We use a discretization based on P_1 elements in H^1_{axi}. The integer N is fixed. Let $r_0 = 0 < \cdots < r_N = 1$ be a subdivision of $[0,1]$ and let $(u_i)_{0\le i\le N} \in (S^1)^{N+1}$ represent a P_1 element in H^1_{axi}. For all $i \in \{0,\dots,N\}$ we denote $u_i = (c_i, s_i) \in \mathbb{R}^2$, where $c_i^2 + s_i^2 = 1$. We define $E^N(r_0,\dots,u_N) := 2\pi \sum_{i=0}^{N-1} E_i(r_0,\dots,u_N)$, where for all i

$$E_i(r_0,\dots,u_N) := \frac{(c_i + c_{i+1})^2}{|u_i + u_{i+1}|^2}\left(\frac{r_{i+1} - r_i}{r_{i+1} + r_i}\right) + \frac{|u_i - u_{i+1}|^2}{|u_i + u_{i+1}|^2}\left(\frac{r_{i+1} + r_i}{r_{i+1} - r_i}\right). \tag{5.2}$$

Here $|\cdot|$ is the euclidean norm in $\mathbb{R}^2$. The function E_i is defined on

$$D_u^N := \{r_0 = 0 < \cdots < r_N = 1\} \times \left\{(u_i)_{0\le i\le N} \in (S^1)^{N+1},\ u_{i+1} \neq -u_i\right\}. \tag{5.3}$$

Let $u^N = (c^N, s^N)$ be the unique P_1 element in H^1_{axi} defined by the values $u^N(r_i) = u_i$ for all $i \in \{0,\dots,N\}$, i.e., $u^N \in \mathcal{C}^0([0,1],\mathbb{R}^2)$ and u^N is affine on every $[r_i, r_{i+1}]$. Assume $u_{i+1} \neq -u_i$ for all i. Since $|u^N| \ge |u_i + u_{i+1}|/2 > 0$ on $[r_i, r_{i+1}]$, the map $u^N/|u^N|$ belongs to $\mathcal{C}^0([0,1], S^1)$. The discrete energy (5.2) is the approximation of $E(u^N/|u^N|)$ by the midpoint formula. Notice that the quadratic energy $E(u^N)$ does not have good properties for the optimal mesh method [13].

Now set $\alpha \in [-\pi/2, \pi/2]$ and assume $u^N(1) = (\cos\alpha, \sin\alpha)$ is given. Then, the map $u^N/|u^N|$ admits a unique lift $\theta^N \in \mathcal{C}^0([0,1],\mathbb{R})$ such that $\theta^N(1) = \alpha$ and $u^N(r) = (\cos\theta^N(r), \sin\theta^N(r))$. Since $E(\theta^N) = E(u^N) < \infty$, $\theta^N \in \mathcal{E}_{\alpha,k}$ for a unique $k \in \mathbb{Z}$. The integer k such that $\theta^N(0) = -\pi/2 + k\pi$ is the degree of u.

Now, the discrete formulation of problem (2.4) is

$$\text{Minimize } E^N \text{ on } D^N_{u,\alpha,k} = \left\{(r_0,\dots,u_N) \in D^N_u \,\middle|\, u^N(1) = \alpha \text{ and } \deg(u) = k\right\}. \tag{5.4}$$

The conditions $u_{i+1} \neq -u_i$ and u^N affine on $[r_i, r_{i+1}]$ imply $|\theta^N(r_{i+1}) - \theta^N(r_i)| < \pi$, so $|\theta(0)| < (N+1)\pi$. Conversely, it is possible to construct for all $\epsilon > 0$ a function $\theta^N \in \mathcal{C}^0([0,1])$ such that $\pi > \theta^N(r_{i+1}) - \theta^N(r_i) \ge \pi - \epsilon$, and for ϵ small enough $\theta(0) \ge \alpha + N\pi - N\epsilon \ge -\pi/2 + (N-1)\pi$. Then $k \ge N-1$ (and similarly we can have $k \le -N+2$). So for any $k \in \mathbb{Z}$, if N is large enough, $D^N_{u,\alpha,k}$ is not empty. Notice also that by the identification $u^N \simeq \theta^N$, every $D^N_{u,\alpha,k}$ is connected (as in the continuous case).

Figure 5.1 shows the numerical solution to (5.4) obtained by a projected gradient algorithm with the change of variable X_i (4.3) for $\alpha = -\pi/4$, $k = 2$, and $N = 17$; the circles of radius 1 represent the cylinder $[0,1] \times S^1$ in which the true solution lives. The results are comparable to the computations with the midpoint formula in codimension 1 (Figures 4.1 and 4.2).

5.3. Upper bound for the degree. The following lemma is the equivalent of Lemma 4.4, which is itself the key of the proof of the convergence Theorem 4.11. Therefore, seeing the discrete S^1 energy E^N (5.2) as a function of $(r_0,\dots,\theta_N)$, it is possible to obtain a convergence result similar to Theorem 4.11 for any sequence of minimizing lifts $(\theta^N)_N$.

LEMMA 5.1. *Let $0 \le r_i < r_{i+1} \le 1$ and let $(\theta_i, \theta_{i+1}) \in \mathbb{R}^2$ such that $0 < |\theta_{i+1} - \theta_i| < \pi$. Define $u_j := (\cos\theta_j, \sin\theta_j)$ for $j = i, i+1$.*

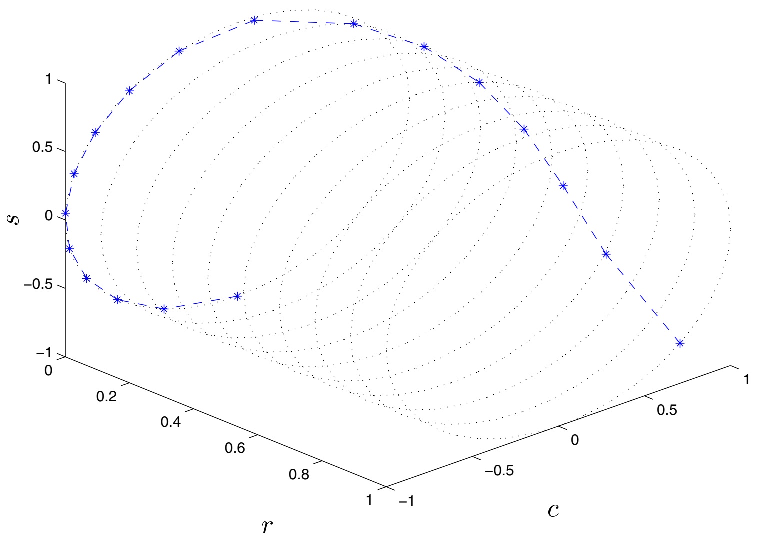


FIG. 5.1. *Midpoint for* S^1 *with* $\alpha = -\pi/4$, $k = 2$, $N = 17$.

Then $E_i(r_i, r_{i+1}, u_i, u_{i+1}) > |F(\theta_{i+1}) - F(\theta_i)|$.

In particular, $E^N(r_0, \dots, u_N) \geq \Lambda_{\alpha,k}$ for all $(r_0, \dots, u_N) \in D^N_{u,\alpha,k}$.

Proof. Assume $[\theta_i, \theta_{i+1}] \subset [\pi/2 + l\pi, \pi/2 + (l+1)\pi]$ for some $l \in \mathbb{Z}$. By Lemma 4.1

$$\begin{aligned} E_i(r_i, r_{i+1}, u_i, u_{i+1}) &\geq 2\frac{|\cos\theta_i + \cos\theta_{i+1}| \cdot |u_{i+1} - u_i|}{|u_i + u_{i+1}|^2}| \\ &\geq 2\frac{\left|\cos\left(\frac{\theta_{i+1}-\theta_i}{2}\right)\right| \cdot \left|\sin\left(\frac{\theta_{i+1}-\theta_i}{2}\right)\right|}{\left|\cos\left(\frac{\theta_i+\theta_{i+1}}{2}\right)\right|} \\ &> |\sin\theta_{i+1} - \sin\theta_i| = |F(\theta_{i+1}) - F(\theta_i)|. \end{aligned}$$

If the previous assumption is not satisfied, we assume without loss of generality $\theta_i < \theta_{i+1}$. There exists $l \in \mathbb{Z}$ such that $\pi/2 + (l-1)\pi < \theta_i < \pi/2 + l\pi < \theta_{i+1} < \pi/2 + (l+1)\pi$. We define $h := \pi/2 + l\pi - \theta_i$ and $k := \theta_{i+1} - (\pi/2 + l\pi)$. Notice that $h \in]0, \pi[$, $k \in]0, \pi[$, and $h + k < \pi$. Then with $X_i = (r_{i+1} - r_i)/(r_{i+1} + r_i) \in]0, 1[$

$$E_i(r_i, r_{i+1}, u_i, u_{i+1}) = \frac{|\sin h - \sin k|^2}{4\cos^2\left(\frac{h+k}{2}\right)} X_i + \frac{\left|\sin^2\left(\frac{h+k}{2}\right)\right|}{\cos^2\left(\frac{h+k}{2}\right)} \frac{1}{X_i}.$$

Using Lemmas 5.2 and 4.1

$$E_i(r_i, r_{i+1}, u_i, u_{i+1}) \geq \frac{|\sin h - \sin k|^2}{4\cos^2\left(\frac{k+h}{2}\right)} + \frac{\sin^2\left(\frac{k+h}{2}\right)}{\cos^2\left(\frac{k+h}{2}\right)}.$$

On the other hand (4.13) yields $|F(\theta_{i+1}) - F(\theta_i)| = 2\sin^2(h/2) + 2\sin^2(k/2)$. The following lemma concludes the proof. □

LEMMA 5.2. *For all* $(h, k) \in [0, \pi]^2$ *such that* $h + k < \pi$,

$$\left|\frac{\sin h - \sin k}{2}\right|^2 \leq \sin^2\left(\frac{h+k}{2}\right).$$

Proof. The proof is similar to the proof of Lemma 4.5. □

LEMMA 5.3. *For all* $(h,k) \in [0,\pi]^2$ *such that* $0 < h+k < \pi$,

$$\frac{|\sin h - \sin k|^2}{4\cos^2\left(\frac{k+h}{2}\right)} + \frac{\sin^2\left(\frac{k+h}{2}\right)}{\cos^2\left(\frac{k+h}{2}\right)} > 2\sin^2\left(\frac{h}{2}\right) + 2\sin^2\left(\frac{k}{2}\right).$$

Proof. Define $K := \{(h,k) \in [0,\pi]^2 \mid h+k \leq \pi \text{ and } k \leq h\}$ and $G : K \to \mathbb{R}$ by

$$G(h,k) = \left(\frac{\sin h - \sin k}{2}\right)^2 + \sin^2\left(\frac{h+k}{2}\right) - 2\left(\sin^2\left(\frac{h}{2}\right) + \sin^2\left(\frac{k}{2}\right)\right)\cos^2\left(\frac{h+k}{2}\right).$$

Then G admits a minimizer $(\underline{h},\underline{k})$ on K. We have

$$\partial_h G(h,k) = \cos\left(\frac{h+k}{2}\right)\left[A(h,k) + \left(2\sin^2\left(\frac{h}{2}\right) + 2\sin^2\left(\frac{k}{2}\right)\right)\sin\left(\frac{h+k}{2}\right)\right],$$

with

$$A(h,k) := \cos(h)\sin\left(\frac{h-k}{2}\right) + \sin\left(\frac{h+k}{2}\right) - \sin(h)\cos\left(\frac{h+k}{2}\right).$$

Now

$$A(h,k) = \sin\left(\frac{h-k}{2} - h\right) + \sin(h)\left[\cos\left(\frac{h-k}{2}\right) - \cos\left(\frac{h+k}{2}\right)\right] + \sin\left(\frac{h+k}{2}\right),$$

so for all $(h,k) \in \overset{\circ}{K}$, $0 < (h-k)/2 < (h+k)/2 < \pi/2$ and $\partial_h G(h,k) > 0$. Thus $(\underline{h},\underline{k}) \in \partial K$. We have $\partial_h G(h,k) > 0$ for all $(h,k) \in \partial K$ such that $0 < h+k < \pi$ and this concludes the proof. □

6. Conclusion.

The main results in this paper are the convergence Theorem 4.11 for the midpoint formula and the external approximation by *BV* functions introduced in section 4.3. They are based on the *BV* bound given by the discrete energy in Proposition 4.7, which is the discrete equivalent of Lemma 2.2. Lemma 5.1 provides the same *BV* bound for the S^1 formulation discretized by the midpoint formula. For the quadratic discretized energy obtained by suppressing the term $|u_i + u_{i+1}|^2$ in formula (5.2), such a bound does not hold and the optimal mesh method does not converge [13]. Thus, the discretized problem should mimic the properties of the continuous problem. For the Gaussian quadrature with two nodes instead of the midpoint formula, the numerical computations [2] indicate that $\min_{\mathcal{E}_{\alpha,k}} E^N > \Lambda_{\alpha,k}$, but we have not been able to prove this.

As pointed out in the introduction, the results of section 3 apply directly in every dimension to every steady-state problem in which the solution minimizes an energy functional. Indeed, Lemma 3.1 provides an l.s.c. extension of the discrete energy, and this guarantees the existence of a discrete minimizer in the closure of its domain of definition, provided the discrete energy is coercive in the sense of (3.5). Assuming that the elements are conforming and that the energy is exact, a density result is enough to show the convergence of the method.

For a triangulation, closing the constraints means that we allow degenerate triangles. Concerning the Poisson equation, the optimal mesh might have such degenerate triangles, but the discrete minimizer is continuous [4]; for regular elliptic problems, one expects the same behavior. For the two- or three-dimensional formulation of problem (2.4) as in [8], or the Plateau problem, which is similar, the graph of the true solution may have a vertical part; one expects a similar vertical part for the discrete solution; in particular, the optimal mesh would have degenerate triangles. Using conforming elements as in section 5.2, we would obtain in two or three dimensions a convergence result similar to Proposition 3.4: the minimizing sequence would converge in the sense of currents (rather than in the weak BV sense), but only up to a subsequence because the limit is not unique.

For practical reasons, we have not yet computed numerical solutions for the two- or three-dimensional formulation of problem (2.4). However, we can refer the reader to [9, 12] for numerical simulations concerning the Plateau problem and elliptic problems with large gradients.

REFERENCES

[1] F. Alouges, *A new algorithm for computing liquid crystal stable configurations: The harmonic mapping case*, SIAM J. Numer. Anal., 34 (1997), pp. 1708–1726.

[2] F. Alouges and M. Pierre, *Mesh optimization for singular axisymmetric harmonic maps from the disc into the sphere*, Numer. Math., to appear.

[3] N. Carlson and K. Miller, *Design and application of a gradient-weighted moving finite element code* I: *In one dimension*, SIAM J. Sci. Comput., 19 (1998), pp. 728–765.

[4] P. de Oliveira, *Existence de maillages optimaux dans les méthodes d'éléments finis*, RAIRO Anal. Numér., 14 (1980), pp. 279–290.

[5] M. Delfour, G. Payre, and J. P. Zolesio, *An optimal triangulation for second-order elliptic problems*, Comput. Methods Appl. Mech. Engrg., 50 (1985), pp. 231–261.

[6] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North-Holland, Amsterdam, 1976.

[7] M. Giaquinta, G. Modica, and J. Souček, *Cartesian Currents in the Calculus of Variations* I, Springer-Verlag, Berlin, 1998.

[8] M. Giaquinta, G. Modica, and J. Souček, *Cartesian Currents in the Calculus of Variations* II, Springer-Verlag, Berlin, 1998.

[9] F. Hülsemann and Y. Tourigny, *A new moving mesh algorithm for the finite element solution of variational problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1416–1438.

[10] R. Li, W.-B. Liu, and T. Tang, *Adaptive finite element approximation for distributed elliptic optimal control problems*, SIAM J. Control Optim., 41 (2002), pp. 1321–1349.

[11] R. Li, T. Tang, and P.-W. Zhang, *Moving mesh methods in multiple dimensions based on harmonic maps*, J. Comput. Phys., 170 (2001), pp. 562–588.

[12] R. Li, T. Tang, and P.-W. Zhang, *A moving mesh finite element algorithm for singular problems in two and three space dimensions*, J. Comput. Phys., 177 (2002), pp. 365–393.

[13] M. Pierre, *Graphes et Maillages Adaptés Pour le Calcul d'Applications Harmoniques Minimisantes*, Ph.D. thesis, ENS Cachan, France, 2002.

[14] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics*, Springer-Verlag, New-York, 2000.

[15] S. Rippa and B. Schiff, *Minimum energy triangulations for elliptic problems*, Comput. Methods Appl. Mech. Engrg., 84 (1990), pp. 257–274.

[16] A. Soyeur, *The Dirichlet problem for harmonic maps from the disc into the* 2*-sphere*, Proc. Roy. Soc. Edinburgh Sect. A., 113 (1989), pp. 229–234.

[17] T. Tang, *Moving mesh methods for computational fluid dynamics*, in Recent Advances in Adaptive Computations, Z.-C. Shi, Z. Chen, T. Tang, and D. Yu, eds., Contemp. Math., AMS, Providence, RI, 2005, to appear.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1455–1480

AN EULER–BERNOULLI BEAM WITH DYNAMIC CONTACT: DISCRETIZATION, CONVERGENCE, AND NUMERICAL RESULTS*

JEONGHO AHN[†] AND DAVID E. STEWART[†]

Abstract. In this paper, we formulate a time-discretization using the implicit Euler method for contact conditions and the midpoint rule for the elastic part of the equations. The energy functional is defined, and convergence for the time-discretization is investigated. Our time-discretization leads to energy dissipation. Using this time discretization and the finite element method with B-spline basis functions, we compute numerical solutions. We show that there is a converging subsequence, and the limit of any such converging subsequence is a solution of the dynamic impact problem. In order to solve the linear complementarity problem that arises in the numerical method, we use a smoothed guarded Newton method. We also investigate numerically the question of whether the numerical solutions converge strongly to their limit and if energy is conserved for the limit. Our numerical results give some evidence that this is so.

Key words. impact, Euler–Bernoulli beam, frictionless contact

AMS subject classifications. Primary, 74M20; Secondary, 74M15, 74K10, 35L85

DOI. 10.1137/S0036142903432619

1. Introduction.

The Euler–Bernoulli equation is an approximate equation for a thin beam such as a rod. Combining this with Signorini contact conditions will give our formulation of the Euler–Bernoulli beam with frictionless contact. The solution $u(x,t)$ of our formulation represents the vertical displacement of the rod from an initially horizontal position for time t and position x along the beam. For small displacements, we have a linear relationship between stress and strain. The Euler–Bernoulli equation is

$$\rho A \frac{\partial^2 u}{\partial t^2} = -EI\frac{\partial^4 u}{\partial x^4} + f(x,t) \quad \text{in } (0,L)\times(0,T],$$

where L is the length of the rod, A is the area of the cross-section of the rod, ρ is the density of the rod, E is the Young modulus for the rod, and I is the second moment of inertia. Note that I is given by $I = \int_{\mathcal{A}}(y-\overline{y})^2 dx\,dy$, where $\mathcal{A}$ is the cross-section of the rod as a subset of the plane and $\overline{y}$ is the vertical center of area, which is $\overline{y} = \int_{\mathcal{A}} y\,dx\,dy / \int_{\mathcal{A}} dx\,dy$. The function $f(x,t)$ is the body force applied to the rod, and time t is between the initial time $t=0$ and some fixed time $t=T$. We denote vectors and matrices by bold letters.

This problem that we are considering is close to Schatzman's model [13], which has been studied in the literature. Her analysis is based on the use of characteristics, and conservation of energy is shown. We note that Euler–Bernoulli beam's boundary conditions are different from those of Schatzman's model.

In the purely elastic case, Euler–Bernoulli beams are not dissipative, and there are no characteristics. While the question of whether there can be conservation of

*Received by the editors July 30, 2003; accepted for publication (in revised form) October 15, 2004; published electronically October 19, 2005. This research was supported in part by NSF grants DMS-98-04316 and DMS-01-39708.

http://www.siam.org/journals/sinum/43-4/43261.html

[†]Department of Mathematics, University of Iowa, Iowa City, IA 52242 (jahn@math.uiowa.edu, dstewart@math.uiowa.edu).

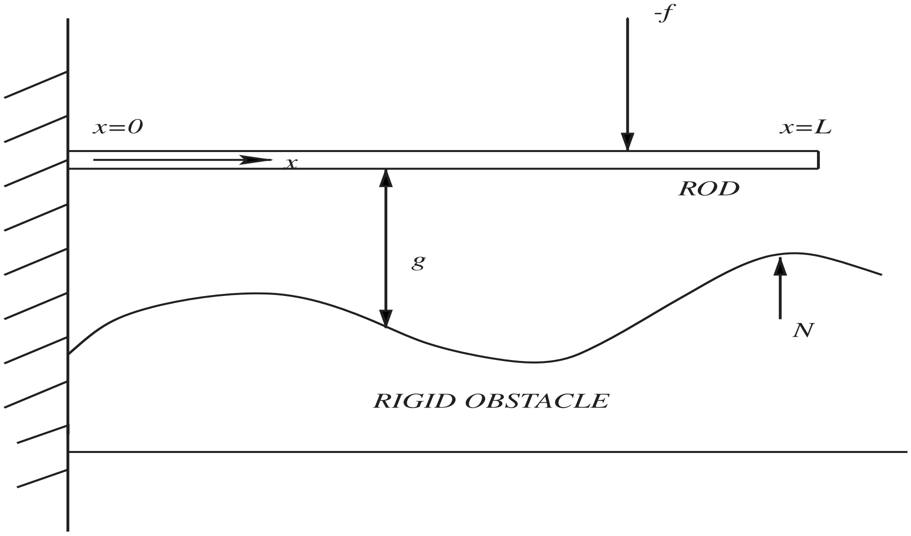


FIG. 1. *Euler–Bernoulli beam with frictionless contact.*

energy for Signorini contact conditions is an open one, this paper gives some evidence that energy is conserved generically.

The Euler–Bernoulli beam with Signorini contact conditions comes from the physical situation illustrated in Figure 1.

The boundary conditions can be identified from the figure. Suppose that the end of the rod at $x = 0$ is clamped horizontally. Then the boundary condition of $x = 0$ has the homogeneous *essential* boundary conditions, i.e., $u(0, t) = u_x(0, t) = 0$. If the end of rod is free, then we have the *natural* boundary conditions $u_{xx}(L, t) = u_{xxx}(L, t) = 0$. Note that subscripts denote derivatives with respect to the subscripted variables. These last two boundary conditions can be obtained from the usual variational techniques.

If we impose frictionless Signorini contact conditions along the length of the rod, we represent the equation of motion as

$$\rho A \frac{\partial^2 u}{\partial t^2} = -EI \frac{\partial^4 u}{\partial x^4} + f(x, t) + N(x, t), \tag{1.1}$$

where from contact criterion, the magnitude of the vertical contact forces (pressures), $N(x, t)$ satisfy the linear complementary problem (LCP) with the complementarity conditions

$$0 \le N(x, t) \quad \perp \quad u(x, t) + g(x) \ge 0 \qquad \text{for all } x \in (0, L),\ t > 0. \tag{1.2}$$

Note that ρA, EI, f, and g are given, while the unknowns are the functions u and N. Note that $g(x)$, called the *gap function*, displays a measure of the initial normalized "gap" between the rod and the rigid foundation, where the position of the rod is on the same as its clamped point of the rod horizontally. Note that for vectors $\mathbf{a}$, $\mathbf{b}$, $0 \le \mathbf{a} \perp \mathbf{b} \ge 0$ means that $\mathbf{a}$, $\mathbf{b} \ge 0$ componentwise and $\mathbf{a}^T\mathbf{b} = 0$. If a and b are scalars, both are nonnegative and either a or b is zero. From the physical point of view, the LCP condition can be interpreted in the following way: When there is a gap between the rod and rigid foundation, i.e., the rod does not reach to rigid foundation, the contact force N must be zero; when there is a contact force, the rod touches the rigid foundation, i.e., there is no gap between them. We assume that applied body force $f(x,t) = f(x)$. So the body force f and gap function g do not depend on time t. We also assume that the gap function $g(x) \geq 0$. Thus we are led to consider solving the following PDE:

$$\rho A u_{tt} = -EI u_{xxxx} + f(x) + N(x,t) \quad \text{in } (0,L) \times (0,T], \tag{1.3}$$

$$0 \leq N(x,t) \perp u(x,t) + g(x) \geq 0 \quad \text{in } (0,L) \times (0,T], \tag{1.4}$$

$$u(0,t) = u_x(0,t) = 0 \quad \text{on } (0,T], \tag{1.5}$$

$$u_{xx}(L,t) = u_{xxx}(L,t) = 0 \quad \text{on } (0,T], \tag{1.6}$$

$$u(x,0) = u^0(x) \quad \text{in } (0,L), \tag{1.7}$$

$$u_t(x,0) = v^0(x) \quad \text{in } (0,L). \tag{1.8}$$

We will assume that $f \in L^2(0,L)$, $u^0 \in H^2_{cf}(0,L)$, $v^0 \in L^2(0,L)$, $g \in C[0,L]$, and that $g(0) > 0$. Note that $H^2_{cf}(0,L)$ is the subset of $H^2(0,L)$, which satisfies the clamped end condition at $x = 0$ ("c" denotes "clamped," while "f" denotes "free"). In this paper, we focus on a numerical approach to the PDE, since the existence of solutions to the PDE has been already shown in [1]. Note that (1.7), (1.8) are the initial conditions for displacement and velocity, respectively.

2. Convergence of the time discretization.

In this section we set up a time-discretization. Also, the convergence for our semidiscretization is investigated. The following section will consider discretization in both time and space. Note that throughout this paper, C refers to a quantity that depends only on the data of the problem and not on the parameters of the approximations used but that may be different in each occurrence.

2.1. Formulation of the discrete-time problem.

In order to obtain a numerical formulation, we will employ a hybrid of two numerical schemes in time:

- Elasticity (u_{xxxx})—Midpoint rule is used,
- Contact condition—Implicit Euler is used.

First we consider a partition of time: $0 = t_0 < t_1 < t_2 < \cdots < t_l < t_{l+1} < \cdots < T$.

We denote by $u^l(x)$ numerical solution of displacement $u(x,t_l)$ and by $v^l(x)$ numerical solution of velocity $v(x,t_l)$ and $N^l(x)$ numerical solution of magnitude of contact force, $N(x,t_l)$, respectively, at each discretized time $t_l = lh$. Then the time step size is $h = t_{l+1} - t_l$ for $l \geq 0$. From (1.3), we take $\rho A = EI = 1$ by proper scaling.

Using our numerical scheme, we establish a numerical formulation:

$$\frac{v^{l+1} - v^l}{h} = -\left(\frac{u^{l+1}_{xxxx} + u^l_{xxxx}}{2}\right) + f(x) + N^l, \tag{2.1}$$

$$\frac{u^{l+1} - u^l}{h} = \frac{v^{l+1} + v^l}{2}, \tag{2.2}$$

$$0 \leq N^l \perp u^{l+1} + g \geq 0, \tag{2.3}$$

where $u^l = u^l(x)$, $v^l = v^l(x)$, $N^l = N^l(x)$ for each $l \geq 0$.

2.2. Energy dissipation in the semidiscrete case.

In this subsection, we will see that numerical formulations (2.1–2.3) cause energy dissipation.

Now we define the energy functional that is dependent on displacement u and velocity v:

$$E(u,v)=\frac{1}{2}\int_0^L\left(|v|^2+|u_{xx}|^2\right)\,dx-\int_0^L f\cdot u\,dx. \tag{2.4}$$

The first term $\frac{1}{2}\int_0^L |v|^2dx$ is the kinetic energy, the second term $\frac{1}{2}\int_0^L |u_{xx}|^2dx$ is the elastic energy, and the last term $-\int_0^L f\cdot u\,dx$ is the external potential energy.

Lemma 2.1. *In the semidiscrete case, energy is dissipated.*

Proof. We want to show that $E(u^{l+1},v^{l+1})\le E(u^l,v^l)$. Using (2.1)–(2.2), we have

$$\begin{aligned}\int_0^L \frac{|v^{l+1}|^2-|v^l|^2}{2h}\,dx &= -\int_0^L \frac{(u^{l+1}_{xxxx}+u^l_{xxxx})(u^{l+1}-u^l)}{2h}\,dx\\ &\quad+\int_0^L \frac{f\cdot(u^{l+1}-u^l)}{h}\,dx+\int_0^L \frac{N^l\cdot(u^{l+1}-u^l)}{h}\,dx. \end{aligned}\tag{2.5}$$

Multiplying by h on both sides of (2.5) and using integration by parts and the boundary conditions, we obtain

$$\begin{aligned}\int_0^L \frac{|v^{l+1}|^2-|v^l|^2}{2}\,dx &= -\int_0^L \frac{|u^{l+1}_{xx}|^2-|u^l_{xx}|^2}{2}\,dx\\ &\quad+\int_0^L f\cdot(u^{l+1}-u^l)\,dx+\int_0^L N^l\cdot(u^{l+1}-u^l)\,dx.\end{aligned}$$

Thus from the LCP condition (2.3),

$$\begin{aligned}\frac{1}{2}\int_0^L\left(|v^{l+1}|^2-|v^l|^2\right)dx &= -\frac{1}{2}\int_0^L\left(|u^{l+1}_{xx}|^2-|u^l_{xx}|^2\right)dx+\int_0^L f\cdot(u^{l+1}-u^l)\,dx\\ &\quad+\int_0^L N^l\cdot(u^{l+1}+g)\,dx-\int_0^L N^l\cdot(u^l+g)\,dx\\ &\le -\frac{1}{2}\int_0^L\left(|u^{l+1}_{xx}|^2-|u^l_{xx}|^2\right)dx+\int_0^L f\cdot(u^{l+1}-u^l)\,dx\end{aligned}\tag{2.6}$$

as $\int_0^L N^l\cdot(u^{l+1}+g)\,dx=0$ by (2.3), but N^l and $u^l+g\ge 0$ so $\int_0^L N^l\cdot(u^l+g)\,dx\ge 0$. Therefore we have

$$\begin{aligned}E(u^{l+1},v^{l+1}) &= \left(\frac{1}{2}\int_0^L\left(|v^{l+1}|^2+|u^{l+1}_{xx}|^2\right)dx\right)-\int_0^L f\cdot u^{l+1}\,dx\\ &\le\left(\frac{1}{2}\int_0^L |v^l|^2+|u^l_{xx}|^2\,dx\right)-\int_0^L f\cdot u^l\,dx=E(u^l,v^l)\end{aligned}$$

as desired. □

From (2.6), we observe that the energy E is conserved if $N^l=0$ and energy is dissipated by the LCP condition (2.3) if $N^l(x)>0$ for some $x\in(0,L)$. Assume that the initial energy is finite. Then Lemma 2.1 shows that $v^l\in L^2(0,L)$ and $u^l\in H^2_{cf}(0,L)$ for all l and $h>0$ and that they are bounded in these spaces independently of l and $h>0$.

2.3. Convergence of the semidiscrete scheme. Since the fourth-order differential operator $K = \partial^4/\partial x^4$ is an elliptic self-adjoint with our boundary condition, we have orthonormal basis ϕ_i with $\partial^4\phi_i/\partial x^4 = \lambda_i\phi_i$ satisfying the homogeneous boundary conditions $\phi_i(0) = \phi_i'(0) = \phi_i''(L) = \phi_i'''(L) = 0$. We order the eigenvalues λ_i so that $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_i \leq \cdots$ and $\lim_{i\to\infty}\lambda_i = \infty$. Properties of these eigenfunctions are discussed in [1]. For the PDE system (1.3–1.8), we can write the discrete-time solution quantities as $u^l(x) = \sum_{i=1}^{\infty} u_i^l\phi_i(x)$, $v^l(x) = \sum_{i=1}^{\infty} v_i^l\phi_i(x)$, and $N^l(x) = \sum_{i=1}^{\infty} N_i^l\phi_i(x)$.

So using the above numerical solution expressions and the numerical formulation (2.1–2.2), for any $i \geq 1$ we have

$$\frac{v_i^{l+1} - v_i^l}{h} = -\lambda_i\left(\frac{u_i^{l+1} + u_i^l}{2}\right) + N_i^l, \tag{2.7}$$

$$\frac{u_i^{l+1} - u_i^l}{h} = \frac{v_i^{l+1} + v_i^l}{2}. \tag{2.8}$$

Note that when we investigate the convergence of our numerical scheme, we will not consider the external body force $f(x)$.

LEMMA 2.2. *From* (2.7) *and* (2.8), u_i^{l+1} *and* v_i^{l+1} *are expressed in terms of* u_i^l *and* v_i^l *for each* $i \geq 1$ *and each* $l \geq 0$ *in the following way:*

$$\begin{aligned}\begin{bmatrix} u_i^{l+1} \\ v_i^{l+1} \end{bmatrix} &= \begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{1/2} \end{bmatrix}\begin{bmatrix} \cos\chi_i & \sin\chi_i \\ -\sin\chi_i & \cos\chi_i \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{-1/2} \end{bmatrix}\begin{bmatrix} u_i^l \\ v_i^l \end{bmatrix} \\ &\quad + \frac{hN_i^l}{1 + h^2\lambda_i/4}\begin{bmatrix} \frac{h}{2} \\ 1 \end{bmatrix},\end{aligned} \tag{2.9}$$

where $\chi_i = \chi(h\lambda_i^{1/2})$*, i.e., function* χ_i *depends only on* $h\lambda_i^{1/2}$*.*

Proof. From (2.8), we have

$$v_i^{l+1} = \frac{2}{h}(u_i^{l+1} - u_i^l) - v_i^l. \tag{2.10}$$

Multiplying by h on (2.7) and plugging (2.10) into (2.7), we obtain

$$\frac{2}{h}(u_i^{l+1} - u_i^l) - 2v_i^l = -\frac{h}{2}(\lambda_i u_i^{l+1} + \lambda_i u_i^l) + hN_i^l. \tag{2.11}$$

Thus multiplying by $h/2$ on (2.11), we have the discrete-time solution at the next step:

$$u_i^{l+1} = (1 + h^2\lambda_i/4)^{-1}\left[(1 - h^2\lambda_i/4)u_i^l + hv_i^l + h^2N_i^l/2\right].$$

Using (2.10), we obtain the next step's velocity:

$$v_i^{l+1} = \frac{1}{1 + h^2\lambda_i/4}\left[-h\lambda_i u_i^l + (1 - h^2\lambda_i/4)v_i^l + hN_i^l\right].$$

Therefore, solving the equations for u^{l+1} and v^{l+1} in terms of u^l and v^l gives

$$\begin{bmatrix} u_i^{l+1} \\ v_i^{l+1} \end{bmatrix} = \frac{1}{1 + h^2\lambda_i/4}\begin{bmatrix} 1 - h^2\lambda_i/4 & h \\ -h\lambda_i & 1 - h^2\lambda_i/4 \end{bmatrix}\begin{bmatrix} u_i^l \\ v_i^l \end{bmatrix} + \frac{hN_i^l}{1 + h^2\lambda_i/4}\begin{bmatrix} \frac{h}{2} \\ 1 \end{bmatrix}.$$

The above system can be written as

$$\begin{bmatrix} u_i^{l+1} \\ v_i^{l+1} \end{bmatrix} = \frac{1}{1+h^2\lambda_i/4} \begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 1-\frac{h^2}{4}\lambda_i & h\lambda_i^{\frac{1}{2}} \\ -h\lambda_i^{\frac{1}{2}} & 1-\frac{h^2}{4}\lambda_i \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{-\frac{1}{2}} \end{bmatrix} \begin{bmatrix} u_i^l \\ v_i^l \end{bmatrix} + \frac{hN_i^l}{1+h^2\lambda_i/4} \begin{bmatrix} \frac{h}{2} \\ 1 \end{bmatrix}.$$

Note that we have

$$\left(\frac{1-h^2\lambda_i/4}{1+h^2\lambda_i/4}\right)^2 + \left(\frac{h\lambda_i^{1/2}}{1+h^2\lambda_i/4}\right)^2 = 1.$$

So we can write

$$\sin\chi_i = \frac{h\lambda_i^{1/2}}{1+h^2\lambda_i/4}, \qquad \cos\chi_i = \frac{1-h^2\lambda_i/4}{1+h^2\lambda_i/4},$$

where $\chi_i = \chi(h\lambda_i^{1/2})$. Hence the result follows. □

Indeed, we can require that χ_i be restricted to $[0,\pi]$.

Remark 2.1. Consider a sequence of vectors $\mathbf{z}_{l+1} = \mathbf{C}\mathbf{z}_l + \mathbf{b}_l$, for $l \in \mathbf{N}$. Then we have

$$\mathbf{z}_l = \mathbf{C}^l \mathbf{z}_0 + \sum_{j=0}^{l-1} \mathbf{C}^{l-1-j}\mathbf{b}_j.$$

It is easy to prove this formula using mathematical induction.

Lemma 2.3. *From* (2.7) *and* (2.8), u_i^l *for each* $l \geq 1$ *can be expressed as*

$$u_i^l = u_i^0 \cos(l\chi_i) + v_i^0 \sin(l\chi_i)/\lambda_i^{1/2} + \frac{h}{1+h^2\lambda_i/4} \sum_{j=0}^{l-1} \left(\frac{h\cos((l-1-j)\chi_i)}{2} + \frac{\sin((l-1-j)\chi_i)}{\lambda_i^{1/2}} \right) N_i^j, \tag{2.12}$$

where u_i^0 *and* v_i^0 *are coefficients for the initial displacement and velocity, respectively.*

Proof. In order to apply Remark 2.1, we set

$$\mathbf{z}_l = \begin{bmatrix} u_i^l \\ v_i^l \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{1/2} \end{bmatrix} \begin{bmatrix} \cos\chi_i & \sin\chi_i \\ -\sin\chi_i & \cos\chi_i \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{-1/2} \end{bmatrix}, \mathbf{b}_l = \begin{bmatrix} \frac{h}{2}N_i^l \\ N_i^l \end{bmatrix}.$$

So from Lemma 2.2, we have

$$\begin{bmatrix} u_i^l \\ v_i^l \end{bmatrix} = \left(\begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{1/2} \end{bmatrix} \begin{bmatrix} \cos\chi_i & \sin\chi_i \\ -\sin\chi_i & \cos\chi_i \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{-1/2} \end{bmatrix} \right)^l \begin{bmatrix} u_i^0 \\ v_i^0 \end{bmatrix} + \frac{h}{1+h^2\lambda_i/4} \sum_{j=0}^{l-1} \left(\begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{1/2} \end{bmatrix} \begin{bmatrix} \cos\chi_i & \sin\chi_i \\ -\sin\chi_i & \cos\chi_i \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \lambda_i^{-1/2} \end{bmatrix} \right)^{l-1-j} \begin{bmatrix} \frac{h}{2}N_i^j \\ N_i^j \end{bmatrix}.$$

By mathematical induction,

$$\begin{bmatrix} u_i^l \\ v_i^l \end{bmatrix} = \begin{bmatrix} \cos(l\chi_i) & \lambda_i^{-1/2}\sin(l\chi_i) \\ -\lambda_i^{1/2}\sin(l\chi_i) & \cos(l\chi_i) \end{bmatrix} \begin{bmatrix} u_i^0 \\ v_i^0 \end{bmatrix} + \frac{h}{1+h^2\lambda_i/4} \sum_{j=0}^{l-1} \begin{bmatrix} \cos((l-1-j)\chi_i) & \lambda_i^{-1/2}\sin((l-1-j)\chi_i) \\ -\lambda_i^{1/2}\sin((l-1-j)\chi_i) & \cos((l-1-j)\chi_i) \end{bmatrix} \begin{bmatrix} \frac{h}{2}N_i^j \\ N_i^j \end{bmatrix}.$$

Multiplying by row vector [1, 0] on both sides of the above system, the coefficient u_i^l of $u^l(x)$ is obtained as desired. □

We define the impulse response function (or fundamental solution of the time-discretization) for fixed $x^* \in (0, L)$ to be

$$w^l(x) = \sum_{i=1}^{\infty} w_i^l \phi_i(x),$$

where $w_i^l = (h\cos(l\chi_i)/2 + \sin(l\chi_i)/\lambda_i^{1/2})/(1 + h^2\lambda_i/4)$. Similar to the fundamental solution of the PDE system, we extend $w_i^l = 0$ for $l < 0$. Thus using this form of impulse response function with Lemma 2.3, we have

$$u_i^l = u_i^0 \cos(l\chi_i) + v_i^0 \frac{\sin(l\chi_i)}{\lambda_i^{1/2}} + h\sum_{j=0}^{l-1} w_i^{l-j-1} N_i^j. \tag{2.13}$$

Recalling the fundamental solution of the PDE system, we define the impulse response function for fixed $x^* \in (0, L)$ as

$$w^l(\cdot, x^*) = \sum_{i=1}^{\infty} w_i^l \phi_i(x^*)\phi_i(\cdot). \tag{2.14}$$

Lemma 2.4. *Using the impulse response function $w^l(\cdot)$, the discrete-time solution $u^l(\cdot)$ can be expressed as*

$$u^l(\cdot) = \sum_{i=1}^{\infty} u_i^0 \cos(l\chi_i)\cdot\phi_i(\cdot) + \sum_{i=1}^{\infty} v_i^0 \frac{\sin(l\chi_i)}{\lambda_i^{1/2}}\phi_i(\cdot) + h\sum_{j=0}^{l-1}\int_0^L w^{l-j-1}(\cdot, x^*)N^j(x^*)dx^*.$$

Proof. Employing (2.13) for fixed $x^* \in (0, L)$, we have

$$\sum_{i=1}^{\infty} u_i^l\phi_i(\cdot) = \sum_{i=1}^{\infty} u_i^0 \cos(l\chi_i)\cdot\phi_i(\cdot) + \sum_{i=1}^{\infty} v_i^0 \frac{\sin(l\chi_i)}{\lambda_i^{1/2}}\phi_i(\cdot) + h\sum_{j=0}^{l-1}\sum_{i=1}^{\infty} w_i^{l-j-1}N_i^j\phi_i(\cdot).$$

Since $N^j(\cdot) = \sum_{r=1}^{\infty} N_r^j\phi_r(\cdot)$ and ϕ_i is orthonormal basis in $L^2(0, L)$, we have

$$N_i^j = \int_0^L \sum_{r=1}^{\infty} N_r^j \phi_r(x^*)\phi_i(x^*)dx^* = \int_0^L N^j(x^*)\phi_i(x^*)dx^*.$$

Thus we obtain

$$\begin{aligned} &u^l(\cdot) \\ &= \sum_{i=1}^{\infty} u_i^0 \cos(l\chi_i)\cdot\phi_i(\cdot) + \sum_{i=1}^{\infty} v_i^0 \frac{\sin(l\chi_i)}{\lambda_i^{1/2}}\phi_i(\cdot) + h\sum_{j=0}^{l-1}\sum_{i=1}^{\infty} w_i^{l-j-1}\phi_i(\cdot)\int_0^L N^j(x^*)\phi_i(x^*)dx^* \\ &= \sum_{i=1}^{\infty} u_i^0 \cos(l\chi_i)\cdot\phi_i(\cdot) + \sum_{i=1}^{\infty} v_i^0 \frac{\sin(l\chi_i)}{\lambda_i^{1/2}}\phi_i(\cdot) + h\sum_{j=0}^{l-1}\int_0^L w^{l-j-1}(\cdot, x^*)N^j(x^*)dx^*, \end{aligned}$$

as required. □

We now need a lemma giving some basic bounds on the function $\chi(s)$. These basic bounds will be used to establish a uniform Hölder continuity result for the discrete fundamental solution w and then for the solution u_h of the discrete-time problem.

LEMMA 2.5. *If* $\cos\chi(s) = (1 - s^2/4)/(1 + s^2/4)$ *and* $\sin\chi(s) = s/(1+s^2/4)$, *then* $\chi(s) \leq s$ *for* $0 \leq s \leq 2$.

Proof. Taking a derivative $\sin\chi(s)$ with respect to s, we have

$$\frac{d\,\sin\chi(s)}{d\chi}\frac{d\chi}{ds} = \frac{d}{ds}\left(\frac{s}{1+s^2/4}\right) = \frac{1}{1+s^2/4}\cdot\frac{1-s^2/4}{1+s^2/4} < \cos\chi(s) \quad \text{for } s \neq 2.$$

So if $s \neq 2$, $d\chi/ds < 1$. Since $\chi(0) = 0$, we have $\chi(s) \leq s$ for $s \geq 0$, using Proposition 3.7 in [1]. If $s = 2$, $\cos\chi(2) = 0$. Therefore $\chi(2) = \pi/2 < 2$, and the result follows. ▯

LEMMA 2.6. *The following uniform Hölder continuity property holds for* $p = 2\gamma$, $0 < p \leq 1$*:*

$$\left|\frac{\sin\left((l+r)\chi(h\lambda^{1/2})\right) - \sin\left(l\chi(h\lambda^{1/2})\right)}{\lambda^\gamma}\right| \leq C_p \cdot (rh)^p,$$

where C_p *is independent of* h *and* λ.

Proof. Suppose $r \geq 1$. Then we have

$$\left|\frac{\sin\left((l+r)\chi(h\lambda^{1/2})\right) - \sin\left(l\chi(h\lambda^{1/2})\right)}{\lambda^\gamma}\right| \leq \frac{2}{\lambda^\gamma}\left|\sin\left(\frac{r\chi}{2}\right)\right|.$$

Since $r\chi - \sin(r\chi/2) \geq 0$ for $r\chi \geq 0$, we have

$$\lambda^{-\gamma}\left|\sin\left((l+r)\chi\right) - \sin(l\chi)\right| \leq 2\lambda^{-\gamma}\min(r\chi, 1).$$

Applying Lemma 2.5 for $h\lambda^{1/2} \leq 2$,

$$\lambda^{-\gamma}\left|\sin\left((l+r)\chi\right) - \sin(l\chi)\right| \leq 2\lambda^{-\gamma}\min(rh\lambda^{1/2}, 1), \tag{2.15}$$

and for $h\lambda^{1/2} \geq 2$, (2.15) also holds by inspection as $rh\lambda^{1/2} > 1$. Dividing by $(rh)^p$ on both sides of (2.15),

$$\frac{\lambda^{-\gamma}\left|\sin\left((l+r)\chi\right) - \sin\left(l\chi\right)\right|}{(rh)^p} \leq 2\lambda^{-\gamma}(rh)^{-p}\min(rh\lambda^{1/2}, 1). \tag{2.16}$$

Thus, from (2.16), if $rh\lambda^{1/2} \leq 1$,

$$\begin{aligned}\lambda^{-\gamma}\left|\sin\left((l+r)\chi\right) - \sin\left(l\chi\right)\right|(rh)^{-p} &\leq 2\lambda^{-\gamma}(rh)^{-p}(rh\lambda^{1/2})\\ &= 2\lambda^{-\gamma+1/2}(rh)^{1-p}\\ &\leq 2\lambda^{-\gamma+1/2}\lambda^{p/2-1/2} = 2\lambda^{p/2-\gamma}.\end{aligned}$$

If $rh\lambda^{1/2} \geq 1$, $\left|\lambda^{-\gamma}\sin\left((l+r)\chi\right) - \lambda^{-\gamma}\sin\left(l\chi\right)\right|(rh)^{-p} \leq 2\lambda^{-\gamma}(rh)^{-p} \leq 2\lambda^{p/2-\gamma}$. Thus putting $p = 2\gamma$, we have

$$\lambda^{-\gamma}\left|\sin\left((l+r)\chi(h\lambda^{1/2})\right) - \sin\left(l\chi(h\lambda^{1/2})\right)\right| \leq 2(rh)^p,$$

as required. ▯

Let the value $u_h(\cdot,t)$ be a continuous piecewise linear interpolant of $u_h(\cdot,lh)=u^l$ and $u_h(\cdot,(l+1)h)=u^{l+1}$ for $t\in[lh,(l+1)h]$. Then recalling Lemma 2.4, the value of $u_h(\cdot,lh)$ computed at step l is expressed as

$$\begin{aligned} u_h(\cdot,lh) &= \sum_{i=1}^{\infty} u_i^0\cos(l\chi_i)\cdot\phi_i(\cdot)+\sum_{i=1}^{\infty} v_i^0\frac{\sin(l\chi_i)}{\lambda_i^{1/2}}\phi_i(\cdot)\\ &\quad + h\sum_{j=0}^{l-1}\int_0^L w_h(\cdot,(l-j-1)h,x^*)N^j(x^*)dx^*, \end{aligned} \tag{2.17}$$

where $w_h(\cdot,lh,x^*)=\sum_{i=1}^{\infty}(h\cos(l\chi_i)/2+\sin(l\chi_i)/\lambda_i^{1/2})\phi_i(\cdot)\phi_i(x^*)/(1+h^2\lambda_i/4)$. Now we define the discrete-time contact force $N_h(x,t)$ as

$$N_h(x,t)=h\sum_{j=0}^{\lfloor T/h\rfloor-1}\delta(t-(j+1)h)N^j(x),$$

where δ is the Dirac-δ function and $\lfloor T/h\rfloor$ is the number of time steps. We also identify N_h with a nonnegative Borel measure on $[0,L]\times[0,T]$ by

$$N_h(B)=\int_B N_h(x,t)\,dx\,dt,$$

where B is any Borel set in $[0,L]\times[0,T]$. It can be shown easily that the Borel measures N_h can be expressed in another way:

$$\int_0^T\int_0^L N_h(x,t)\,dx\,dt=h\sum_{l=0}^{\lfloor T/h\rfloor-1}\int_0^L N^l(x)\,dx. \tag{2.18}$$

Lemma 2.7. *The Borel measures N_h are uniformly bounded as measures on $[0,L]\times[0,T]$ as $h\downarrow 0$ for $v^0\in L^2(0,L)$ and $u^0\in H^2_{cf}(0,L)$.*

Proof. Multiplying h and $x^2/2$ on both sides in (2.1) and taking integrals on both sides in (2.1),

$$\int_0^L\frac{x^2}{2}(v^{l+1}-v^l)dx=-\frac{h}{2}\int_0^L\frac{x^2}{2}(u^{l+1}_{xxxx}+u^l_{xxxx})dx+h\int_0^L\frac{x^2}{2}N^l dx.$$

Note that we do not consider the body force $f(x)$. Thus, taking the sum over $l\geq 0$ and using an integration by parts,

$$\begin{aligned} h&\sum_{l=0}^{\lfloor T/h\rfloor-1}\int_0^L\frac{x^2}{2}N^l(x)\,dx\\ &=\sum_{l=0}^{\lfloor T/h\rfloor-1}\int_0^L\frac{x^2}{2}(v^{l+1}-v^l)dx+\frac{h}{2}\sum_{l=0}^{\lfloor T/h\rfloor-1}\int_0^L\frac{x^2}{2}(u^{l+1}_{xxxx}+u^l_{xxxx})dx\\ &=\sum_{l=0}^{\lfloor T/h\rfloor-1}\int_0^L\frac{x^2}{2}(v^{l+1}-v^l)dx+\frac{h}{2}\sum_{l=0}^{\lfloor T/h\rfloor-1}\int_0^L(u^{l+1}_{xx}+u^l_{xx})dx\\ &\leq\frac{L^2}{2}(\|v^{\lfloor T/h\rfloor}\|_{L^2(0,L)}+\|v^0\|_{L^2(0,L)})+TL^{1/2}\cdot\max_{0\leq l\leq\lfloor T/h\rfloor}\|u^l_{xx}\|_{L^2(0,L)}. \end{aligned} \tag{2.19}$$

We want to show that $\int_0^L N^l dx$ is bounded by $\int_0^L \frac{x^2}{2} N^l dx$ for all $l \geq 0$. Since $g(0) > 0$ and $u(0, t_{l+1}) = \partial u(0, t_{l+1})/\partial x = 0$ and $u^{l+1} \in H^2_{cf}(0, L)$, there is an $\eta > 0$ such that $u^{l+1}(x) > -g(x)$ for all $x \in [0, \eta]$. So by the LCP condition of the numerical formulation, $N^l = 0$ for $0 \leq x \leq \eta$. Since $N^l \geq 0$ and $x^2/2 \geq \eta^2/2 > 0$ for $[\eta, L]$, we have

$$\int_0^L \frac{x^2}{2} N^l dx = \int_\eta^L \frac{x^2}{2} N^l dx \geq \frac{\eta^2}{2} \int_\eta^L N^l dx = \frac{\eta^2}{2} \int_0^L N^l dx. \tag{2.20}$$

So by (2.18), the Borel measure N_h is bounded, independent of h as $h \downarrow 0$. The proof is complete. □

LEMMA 2.8. *The discrete-time solution $t \mapsto u_h(\cdot, t)$ is uniformly Hölder continuous into $H^\beta(0, L)$ as $h \downarrow 0$ with an exponent $0 < p \leq 1$ and $\beta > 0$ in the following sense:*

$$\| u_h(\cdot, (l + r)h) - u_h(\cdot, lh) \|_{H^\beta(0,L)} \leq C_p (rh)^p$$

for integers l and r, where $\beta/2 + p < 3/4$ and C_p is independent of h.

Proof. Applying (2.17), the last term of $u_h(\cdot, (l + r)h)$ becomes

$$h \sum_{j=0}^{l+r-1} \int_0^L w_h(\cdot, (l + r - j - 1)h, x^*) N^j(x^*) dx^*. \tag{2.21}$$

Similarly, the last term of $u_h(\cdot, lh)$ becomes

$$\sum_{j=0}^{l-1} \int_0^L w_h(\cdot, (l - j - 1)h, x^*) N^j(x^*) dx^*. \tag{2.22}$$

We denote (2.21) by (I) and (2.22) by (II). Thus, using Lemma 2.4, we have

$$\begin{aligned} &\| u_h(\cdot, (l + r)h) - u_h(\cdot, lh) \|_{H^\beta(0,L)} \\ &\quad \leq \left\| \sum_{i=1}^\infty u_i^0 \left(\cos\left((l + r)\chi_i\right) - \cos(l\chi_i) \right) \phi_i(x) \right\|_{H^\beta(0,L)} \\ &\qquad + \left\| \sum_{i=1}^\infty v_i^0 \frac{\sin\left((l + r)\chi_i\right) - \sin(l\chi_i)}{\lambda_i^{1/2}} \phi_i(x) \right\|_{H^\beta(0,L)} \\ &\qquad + h \left\| \int_0^L ((\mathrm{I}) - (\mathrm{II})) dx^* \right\|_{H^\beta(0,L)}. \end{aligned} \tag{2.23}$$

Using Lemma 2.6 and Proposition 3.3 in [1], in the first term of (2.23) we have

$$\begin{aligned} &\left\| \sum_{i=1}^\infty u_i^0 \left(\cos\left((l + r)\chi_i\right) - \cos(l\chi_i) \right) \phi_i(x) \right\|^2_{H^\beta(0,L)} \\ &\quad = \sum_{i=1}^\infty \left(\cos\left((l + r)\chi_i\right) - \cos(l\chi_i) \right)^2 \lambda_i^{\beta/2} (u_i^0)^2 + \sum_{i=1}^\infty \left(\cos\left((l + r)\chi_i\right) - \cos(l\chi_i) \right)^2 (u_i^0)^2 \\ &\quad = \sum_{i=1}^\infty \left[\frac{\cos\left((l + r)\chi_i\right) - \cos(l\chi_i)}{\lambda_i^{p/2}} \right]^2 \left(\lambda_i^{\beta/2+p-1} \cdot \lambda_i (u_i^0)^2 + \lambda_i^{p-1} \cdot \lambda_i (u_i^0)^2 \right) \\ &\quad = (rh)^{2p} \sum_{i=1}^\infty \lambda_i (u_i^0)^2 \left[\lambda_i^{\beta/2+p-1} + \lambda_i^{p-1} \right]. \end{aligned}$$

Similarly, in the second term of (2.23), we have

$$\left\|\sum_{i=1}^{\infty} v_i^0 \frac{\sin\left((l+r)\chi_i\right)-\sin(l\chi_i)}{\lambda_i^{1/2}}\phi_i(x)\right\|_{H^\beta(0,L)}^2 = (rh)^{2p}\sum_{i=1}^{\infty}\lambda_i(v_i^0)^2\left[\lambda_i^{\beta/2+p-1}+\lambda_i^{p-1}\right].$$

Note that since $v^0 \in L^2(0,L)$ and $u^0 \in H^2_{cf}(0,L)$, $\|v^0\|^2_{L^2(0.L)} = \sum_i (v_i^0)^2 < \infty$ and $|u^0|^2_{H^2(0,L)} = \sum_i \lambda_i(u_i^0)^2 < \infty$. Similarly in the third term of (2.23),

$$\begin{aligned}
&\left\|\int_0^L ((\mathrm{I})-(\mathrm{II}))dx^*\right\|_{H^\beta(0,L)} \\
&\quad\le h\int_0^L \|(\mathrm{I})-(\mathrm{II})\|_{H^\beta(0,L)}dx^* \\
&\quad\le h\int_0^L \sum_{j=0}^{l-1}\left\|\left(w_h(\cdot,(l+r-j-1)h,x^*)-w_h(\cdot,(l-j-1)h,x^*)\right)N^j(x^*)\right\|_{H^\beta(0,L)}dx^* \\
&\quad\quad+ h\int_0^L \sum_{j=l}^{l+r-1}\left\|w_h(\cdot,(l+r-j-1)h,x^*)N^j(x^*)\right\|_{H^\beta(0,L)}dx^*.
\end{aligned} \tag{2.24}$$

Recall that $w_h(\cdot,lh,x^*) = \sum_{i=1}^{\infty}(h\cos(l\chi_i)/2+\sin(l\chi_i)/\lambda_i^{1/2})\phi_i(\cdot)\phi_i(x^*)/(1+h^2\lambda_i/4)$ and $\max_{0\le x\le L}|\phi_i(x)| \le M$ as proved in Lemma 3.2 in [1]. Note that $\lambda_i^{1/2}h/(1+\lambda_ih^2/4) \le 1$ and $1/(1+\lambda_ih^2/4) \le 1$ for $\lambda_i^{1/2}h \ge 0$. In the first term of (2.24) for $0 \le j \le l-1$, it can be shown that we have

$$\begin{aligned}
&\|\left(w_h(\cdot,(l+r-j-1)h,x^*)-w_h(\cdot,(l-j-1)h,x^*)\right)N^j(x^*)\|^2_{H^\beta(0,L)} \\
&\quad\le \left(\frac{5}{4}+\sqrt{2}\right)M^2|N^j(x^*)|^2(rh)^{2p}\sum_{i=1}^{\infty}\left(\lambda_i^{\beta/2+p-1}+\lambda_i^{p-1}\right).
\end{aligned} \tag{2.25}$$

Similar to the second term of (2.24), for $l \le j \le l+r-1$ we have

$$\begin{aligned}
&\|w_h(\cdot,(l+r-j-1)h,x^*)N^j(x^*)\|^2_{H^\beta(0,L)} \\
&\quad\le C|N^j(x^*)|^2(rh)^{2p}\sum_{i=1}^{\infty}\left(\lambda_i^{\beta/2+p-1}+\lambda_i^{p-1}\right).
\end{aligned} \tag{2.26}$$

Note that for sufficiently large i, there exist $C > 0$ such that $\lambda_i \le Ci^4$. This was shown in Lemma 3.2 of [1]. Therefore by (2.18) and Lemma 2.7 and integral test, the result follows, provided that $\beta/2+p < 3/4$. □

Note that the condition of Lemma 2.8 is the same case as for the penalty method.

Lemma 2.9. *In a certain subsequence with $h \downarrow 0$, the time-discretized functions u_h, v_h, and N_h converge to a solution, u_h uniformly in $C([0,L]\times[0,T])$, v_h weak* in $L^\infty(0,T;L^2(0,L))$, and N_h weak* in the space of measures on $[0,L]\times[0,T]$. Furthermore, the solution (u, N) converged by (u_h, N_h) satisfies the complementarity condition $0 \le u+g \perp N \ge 0$ in the weak sense.*

Proof. By Lemma 2.7 and the Riesz representation theorem [8, Thm. 4.2, p. 268] and Alaoglu's theorem [12, Thm. 6.62, p. 203], a subsequence converges $N_h \rightharpoonup^* N$ as measures. Since $N_h \ge 0$, $N \ge 0$. Then since $C^p(0,T;H^\beta(0,L))$ is compactly embedded in $C([0,L]\times[0,T])$, by the Arzela–Ascoli theorem [8, pp. 57–59] there exists a suitable subsequence of u_h corresponding to subsequence of N_h such that $u_h \to u$ in $C([0,L]\times[0,T])$. We also denote this subsequence by u_h and restrict our attention to this subsequence. Since $u_h + g \geq 0$ for each $h > 0$, it follows that $u + g \geq 0$.

Since v_h is uniformly bounded in $L^\infty(0,T;L^2(0,L))$ and $L^\infty(0,T;L^2(0,L)) \simeq L^1(0,T;L^2(0,L))^*$, by Alaoglu's theorem there is a weak* converging subsequence, also denoted v_h, and we restrict our attention to this subsequence.

Since $u_h(\cdot,t)$ is an interpolant of $u_h(\cdot,lh)$ and $u_h(\cdot,(l+1)h)$, and $N_h(x,t) = h\sum_{j=0}^{\lfloor T/h\rfloor-1}\delta(t-(j+1)h)N^j(x)$ for $t\in[lh,\,(l+1)h]$, we have

$$\begin{aligned}
&\int_0^T\int_0^L N_h(x,t)(u_h(x,t)+g(x))\,dx\,dt\\
&\quad= h\int_0^T\int_0^L\left[\sum_{j=0}^{\lfloor T/h\rfloor-1}\delta(t-(j+1)h)N^j(x)\right](u_h(x,t)+g(x))\,dxdt\\
&\quad= h\int_0^L\left[\sum_{j=0}^{\lfloor T/h\rfloor-1}\int_0^T\delta(t-(j+1)h)(u_h(x,t)+g(x))\,dt\right]N^j(x)\,dx\\
&\quad= h\int_0^L\sum_{j=0}^{\lfloor T/h\rfloor-1}N^j(x)(u^{j+1}(x)+g(x))\,dx = 0.
\end{aligned}$$

So taking limits in the subsequence gives

$$0 = \int_0^T\int_0^L N_h(x,t)(u_h(x,t)+g(x))\,dxdt \to \int_0^T\int_0^L N(x,t)(u(x,t)+g(x))\,dxdt = 0$$

as desired. □

2.4. Do the discrete-time solutions converge strongly? While we cannot fully answer this question at this time, we will lay the groundwork in this section for the numerical evidence to be presented later for strong convergence.

We recall the expression of numerical solutions (discrete time solutions) $u^l(x)$ at each discretized time t_l such as $u^l(x) = \sum_{i=1}^\infty u_i^l\phi_i(x)$. Note that we write $u_i^{l;h}$ and $v_i^{l;h}$ instead of u_i^l and v_i^l, respectively, in order to show the dependence on $h > 0$ more explicitly. Then we consider numerical trajectories $u_h(x,t)$ by piecewise continuous linear interpolation of $u_h(x,t_l) = u^{l;h}(x)$ and $v_h(x,t)$ by the piecewise constant interpolation of $v_h(x,t_l) = v^{l;h}(x)$ for each $l \geq 0$. So we express these as $u_h(x,t) = \sum_{i=1}^\infty u_i^h(t)\phi_i(x)$ and $v_h(x,t) = \sum_{i=1}^\infty v_i^h(t)\phi_i(x)$. Then the value of $u_i^h(t)$ is the linear interpolant of $u_i^h(lh) = u_i^{l;h}$ and $u_i^h((l+1)h) = u_i^{l+1;h}$ for $t\in[lh,(l+1)h]$.

Let $\mathbf{u}^{l;h} = (u_1^{l;h}, u_2^{l;h}, u_3^{l;h},\ldots)$, $\mathbf{v}^{l;h} = (v_1^{l;h}, v_2^{l;h}, v_3^{l;h},\ldots)$, and $\boldsymbol{\omega}^{l;h} = (\omega_1^{l;h},\omega_2^{l;h},\omega_3^{l;h},\ldots)$, where $\omega_i^{l;h} = \lambda_i^{1/2}u_i^{l;h}$ for $i \geq 1$. We use notation ℓ^2 as the Hilbert space of sequences $\mathbf{x} = (x_1,x_2,x_3,\ldots)$, where $\|\mathbf{x}\|_{\ell^2} = \sqrt{\sum_i^\infty |x_i|^2} < \infty$. Using the energy functional, it can be shown easily that

$$E(u^l,v^l) = \sum_{i=1}^\infty\left(\left(v_i^{l;h}\right)^2 + \lambda_i\left(u_i^{l;h}\right)^2\right)$$

and $\boldsymbol{\omega}^{l;h}$, $\mathbf{v}^{l;h}$ are uniformly bounded in ℓ^2.

Now suppose that we do not consider body force f in the energy function. Thus by the energy boundedness we have

$$\sum_{i=1}^{\infty} \lambda_i \left(u_i^h(t)\right)^2 \text{ and } \sum_{i=1}^{\infty} \left(v_i^h(t)\right)^2 < \infty.$$

So $\boldsymbol{\omega}^h$, $\mathbf{v}^h \in \ell^2$ and are uniformly bounded in ℓ^2, where $\boldsymbol{\omega}^h = (\omega_1^h(t), \omega_2^h(t), \omega_3^h(t), \ldots)$ and $\mathbf{v}^h = (v_1^h(t), v_2^h(t), v_3^h(t), \ldots)$. Thus there are a subsequence of $\mathbf{v}^h$ and a subsequence of $\boldsymbol{\omega}^h$ that are convergent to $\mathbf{v}(t)$ and $\boldsymbol{\omega}(t)$, respectively, in ℓ^2, as $h \downarrow 0$. These facts induce Lemma 2.10.

By inspection of the eigenfunctions, the frequency of oscillation is proportional on $\lambda^{1/4}$. So high frequency modes correspond to large eigenvalues, and low frequency modes correspond to small eigenvalues. Also, only the elastic energy defines the modes, since they are eigenfunctions of the fourth-order operator $K = \partial^4/\partial x^4$ in the continuous case, or eigenvectors of $\mathbf{M}^{-1}\mathbf{K}$ in the fully discretized case, which will be considered in the following section. In the next lemma, it is shown that the amount of energy in the high frequency modes is negligible under the assumption of the strong convergence. In the physical point of view, energy in the high frequency modes is equivalent to heat. In section 3, the fact that $\boldsymbol{\omega}^{l;h}$, $\mathbf{v}^{l;h}$ are uniformly bounded in ℓ^2 will be supported by numerical evidence. The detailed argument will be presented in subsections 3.5 and 3.6.

Lemma 2.10. *Let $t \in [lh, (l+1)h]$ for any $l \geq 1$. Suppose that $\boldsymbol{\omega}^{l;h} \to \boldsymbol{\omega}(t)$ and $\mathbf{v}^{l;h} \to \mathbf{v}(t)$ (strongly) in ℓ^2, as $h \downarrow 0$, $lh \to t$. Then we have*

$$\lim_{c\to\infty} \limsup_{h\downarrow 0} \frac{1}{2} \sum_{i;i\geq c} \left(|v_i^{l;h}|^2 + \lambda_i |u_i^{l;h}|^2\right) = 0.$$

Proof. For the fixed $l \geq 1$ and any $c \geq 1$, we obtain

$$\left(\sum_{i=c}^{\infty} |\omega_i^{l;h}|^2\right)^{1/2} \leq \left(\sum_{i=c}^{\infty} |\omega_i^{l;h} - \omega_i(t)|^2\right)^{1/2} + \left(\sum_{i=c}^{\infty} |\omega_i(t)|^2\right)^{1/2}.$$

Since $\|\boldsymbol{\omega}^{l;h} - \boldsymbol{\omega}(t)\|_{\ell^2} \to 0$ as $h \downarrow 0$, $lh \to t$, we obtain

$$\limsup_{h\downarrow 0} \left(\sum_{i=c}^{\infty} |\omega_i^{l;h}|^2\right)^{1/2} \leq \limsup_{h\downarrow 0} \left[\left(\sum_{i=c}^{\infty} |\omega_i^{l;h} - \omega_i(t)|^2\right)^{1/2} + \left(\sum_{i=c}^{\infty} |\omega_i(t)|^2\right)^{1/2}\right]$$

$$= \limsup_{h\downarrow 0} \left(\sum_{i=c}^{\infty} |\omega_i(t)|^2\right)^{1/2}. \tag{2.27}$$

Since $\sum_{i=c}^{\infty} |\omega_i(t)|^2 = \|\boldsymbol{\omega}(t)\|_{\ell^2}^2 - \sum_{i=1}^{c-1} |\omega_i(t)|^2$, we have

$$\lim_{c\to\infty} \sum_{i=c}^{\infty} |\omega_i(t)|^2 = \|\boldsymbol{\omega}(t)\|_{\ell^2}^2 - \lim_{c\to\infty} \sum_{i=1}^{c-1} |\omega_i(t)|^2 = \|\boldsymbol{\omega}(t)\|_{\ell^{2-}}^2 - \|\boldsymbol{\omega}(t)\|_{\ell^2}^2 = 0. \tag{2.28}$$

Thus, combining (2.27) with (2.28),

$$\lim_{c\to\infty} \limsup_{h\downarrow 0} \left(\sum_{i=c}^{\infty} |\omega_i^{l;h}|^2\right)^{1/2} \leq 0.$$

Since $|\omega_i^{l;h}|^2 = \lambda_i |u_i^{l;h}|^2 \geq 0$ for each $i \geq 1$, we have for elastic energy

$$\lim_{c\to\infty} \limsup_{h\downarrow 0} \sum_{i;i\geq c} |\omega_i^{l;h}|^2 = \lim_{c\to\infty} \limsup_{h\downarrow 0} \sum_{i;i\geq c} \lambda_i |u_i^{l;h}|^2 = 0.$$

Similar to the above argument, we have for kinetic energy

$$\lim_{c\to\infty} \limsup_{h\downarrow 0} \sum_{i;i\geq c} |v_i^{l;h}|^2 = 0.$$

Therefore the result follows. □

We note that in general, $\mathbf{u}^l \rightharpoonup \mathbf{u}$ in ℓ^p with $1 < p < \infty$ if and only if $\lim_{l\to\infty} u_i^l = u_i$ for $i \geq 1$ and $\sup_{1\leq l<\infty} \|\mathbf{u}^l\|_{\ell^p} < \infty$.

3. Discretization in time and space.

3.1. Finite element method and B-splines. The finite element method is one of the most popular numerical methods for solving static elliptic boundary value problems. So we will approximate the solution in the spatial domain $[0, L]$, using the finite element method [4, 6]. We partition the domain $[0, L]$ into

$$0 = x_0 < x_1 < x_2 < x_3 < x_4 < \cdots < x_{m+1} = L.$$

We denote $k = x_{i+1} - x_i$ as size of subinterval $[x_{i+1}, x_i]$ for $i \geq 1$. Let

$$V = H_{cf}(0, L) = \{w \in H^2(0, L) \mid w(0) = w'(0) = 0\},$$

where $H_{cf}(0, L)$ is a subset of Sobolev space $H^2(0, L)$, using the same norm. We choose B-spline functions $\psi_i(x)$, $1 \leq i \leq m+1$ for the basis functions. The B-spline will be a cubic spline [2, pp. 166–176] with nodes x_i, $i = 1, 2, 3, \ldots, m+1$. Note that unlike the usual piecewise continuous linear basis function, we need $m+1$ basis functions from the construction of a B-spline. Thus the finite element space becomes

$$V_k = \operatorname{span}\{\, \psi_i \mid 1 \leq i \leq m+1 \,\}.$$

These basis function will need to be in H_{cf}^2. Thus we can construct the standard B-spline function $B(s)$, according to the property of B-splines and the condition that $B(0) = 1$, $B(s) = B(-s)$, and $B'(0) = 0$:

$$B(s) = \frac{2}{3} \begin{cases} 1 + \frac{3}{4}|s|^3 - \frac{3}{2}|s|^2 & \text{if } |s| \leq 1, \\ \frac{1}{4}(2 - |s|)^3 & \text{if } 1 \leq |s| \leq 2, \\ 0 & \text{if } |s| \geq 2. \end{cases}$$

Thus $B(s)$ is piecewise cubic on interval $[i, i+1]$ for $i \in \mathbb{Z}$. For most basis functions $i = 2, 3, \ldots, m+1$, we use shifted B-splines:

$$\psi_i(x) = B\left(\frac{x}{k} - i\right) = B\left(\frac{x - x_i}{k}\right),$$

where $x_i = i\,k$, $1 \leq i \leq m+1$. In order to satisfy essential boundary condition, we need to change the first basis function:

$$\psi_1(x) = 2\left(B\left(\frac{x}{k} + 1\right) + B\left(\frac{x}{k} - 1\right)\right) - B\left(\frac{x}{k}\right).$$

We write the approximate solution u^l, v^l, N^l as

$$u^l(x) = \sum_{i=1}^{m+1} \widehat{u}_i^l \psi_i(x),\ v^l(x) = \sum_{i=1}^{m+1} \widehat{v}_i^l \psi_i(x),\ \text{and } N^l = \sum_{i=1}^{m+1} \widehat{N}_i^l \psi_i(x). \tag{3.1}$$

In this section we want to use $\widehat{u}_i^l$ to indicate coefficients of the basis functions in the finite elements, in contrast to u_i^l, which indicate coefficients of the eigenfunctions. Using (2.2), we have discrete-time equations of motion

$$\frac{2}{h^2}u^{l+1} + \frac{1}{2}u_{xxxx}^{l+1} = \frac{2}{h^2}u^l - \frac{1}{2}u_{xxxx}^l + \frac{2}{h}v^l + f(x) + N^l. \tag{3.2}$$

First setting $u^l(x) = \sum_{i=1}^{m+1} \widehat{u}_i^l \psi_i(x)$ and multiplying by basis function $\psi_i(x)$ on both sides of (3.2) and by integrating by parts, we obtain the Galerkin approximation for one time step:

$$\left(\mathbf{M} + \frac{h^2}{4}\mathbf{K}\right)\mathbf{u}^{l+1} = \left(\mathbf{M} - \frac{h^2}{4}\mathbf{K}\right)\mathbf{u}^l + h\mathbf{M}\mathbf{v}^l + \frac{h^2}{2}\left(\mathbf{f} + \mathbf{M}\mathbf{N}^l\right), \tag{3.3}$$

$$\mathbf{v}^{l+1} = \frac{2}{h}(\mathbf{u}^{l+1} - \mathbf{u}^l) - \mathbf{v}^l, \tag{3.4}$$

where the mass ($\mathbf{M}$) and stiffness matrices ($\mathbf{K}$) have the following forms, respectively:

$$M_{ij} = \int_0^L \psi_i \psi_j \, dx \text{ and } K_{ij} = \int_0^L \psi_i'' \psi_j'' \, dx.$$

From (3.3) and (3.4) we will obtain numerical solutions $\mathbf{u}^l = (\widehat{u}_1^l, \widehat{u}_2^l, \dots, \widehat{u}_{m+1}^l)^T$, $\mathbf{v}^l = (\widehat{v}_1^l, \widehat{v}_2^l, \dots, \widehat{v}_{m+1}^l)^T$, and $\mathbf{N}^l = (\widehat{N}_1^l, \widehat{N}_2^l, \dots, \widehat{N}_{m+1}^l)^T$ at each discretized time t_l. Also note that $\mathbf{f} = (f_1, f_2, \dots, f_{m+1})$ is the load vector, where $f_i = \int_0^l f(x)\psi_i(x)\,dx$. Recalling that each basis function is $\psi_i(x) = B((x - x_i)/k)$, the mass and stiffness matrices are banded matrix with three subdiagonals and three superdiagonals. Note that these matrices $\mathbf{M}$ and $\mathbf{K}$ are symmetric positive definite.

3.2. Energy dissipation in the fully discrete case. If the fully discrete scheme has the same linear complementary condition as the semidiscrete case, energy dissipation can fail to hold. This was indeed observed in some preliminary computations. So in the fully discrete case, we need to modify the complementarity condition in order to guarantee energy dissipation. Following the definition of energy functional (2.4) and (3.1), we can define energy functional in the fully discrete case:

$$E(\mathbf{u}^l, \mathbf{v}^l) = \frac{1}{2}\left((\mathbf{v}^l)^T \mathbf{M}\mathbf{v}^l + (\mathbf{u}^l)^T \mathbf{K}\mathbf{u}^l\right) - \mathbf{f}\cdot\mathbf{u}^l. \tag{3.5}$$

LEMMA 3.1. *If we have complementarity condition*

$$\mathbf{0} \le \mathbf{M}\mathbf{N}^l \quad \perp \quad \mathbf{u}^{l+1} + \mathbf{g} \ge \mathbf{0}, \tag{3.6}$$

where $\mathbf{g} = (g_1, g_2, \dots, g_{m+1})^T$ *and* $g_i = g(x_i)$ *and with the discrete equations of motion* (3.3), (3.4)*, then energy is dissipated.*

Proof. Using numerical formulation (2.1)–(2.2) and (3.1), we have

$$\frac{1}{2h}\left(\sum_{i=1}^{m+1} \widehat{v}_i^{\,l+1}\psi_i(x) - \sum_{i=1}^{m+1} \widehat{v}_i^{\,l}\psi_i(x)\right)\left(\sum_{j=1}^{m+1} \widehat{v}_j^{\,l+1}\psi_j(x) + \sum_{j=1}^{m+1} \widehat{v}_j^{\,l}\psi_j(x)\right)$$

$$= -\frac{1}{2h}\left(\sum_{i=1}^{m+1}\widehat{u}_i^{\,l+1}\psi_i^{''''}(x) + \sum_{i=1}^{m+1}\widehat{u}_i^{\,l}\psi_i^{''''}(x)\right)\left(\sum_{j=1}^{m+1}\widehat{u}_j^{\,l+1}\psi_j(x) - \sum_{j=1}^{m+1}\widehat{u}_j^{\,l}\psi_j(x)\right)$$
$$+ \frac{1}{h}f\left(\sum_{j=1}^{m+1}\widehat{u}_j^{\,l+1}\psi_j(x) - \sum_{j=1}^{m+1}\widehat{u}_j^{\,l}\psi_j(x)\right)$$
$$+ \frac{1}{h}\sum_{i=1}^{m+1}\widehat{N}_i^{\,l}\psi_i(x)\left(\sum_{j=1}^{m+1}\widehat{u}_j^{\,l+1}\psi_j(x) - \sum_{j=1}^{m+1}\widehat{u}_j^{\,l}\psi_j(x)\right).$$

Then taking the integral with respect to x and using integration by parts and using mass and stiffness matrix,

$$\frac{1}{2}\left((\mathbf{v}^{l+1})^T\mathbf{M}\mathbf{v}^l - (\mathbf{v}^l)^T\mathbf{M}\mathbf{v}^l\right)$$
$$= -\frac{1}{2}\left((\mathbf{u}^{l+1})^T\mathbf{K}\mathbf{u}^{l+1} - (\mathbf{u}^l)^T\mathbf{K}\mathbf{u}^l\right) + \mathbf{f}^T(\mathbf{u}^{l+1} - \mathbf{u}^l) + (\mathbf{N}^l)^T\mathbf{M}(\mathbf{u}^{l+1} - \mathbf{u}^l)$$
$$= -\frac{1}{2}((\mathbf{u}^{l+1})^T\mathbf{K}\mathbf{u}^{l+1} - (\mathbf{u}^l)^T\mathbf{K}\mathbf{u}^l) + \mathbf{f}^T(\mathbf{u}^{l+1} - \mathbf{u}^l) + (\mathbf{N}^l)^T\mathbf{M}(\mathbf{u}^{l+1} + \mathbf{g} - \mathbf{u}^l - \mathbf{g}).$$

Thus by the complementarity condition (3.6), we have

$$E(\mathbf{u}^{l+1}, \mathbf{v}^{l+1}) = \frac{1}{2}\left((\mathbf{v}^{l+1})^T\mathbf{M}\mathbf{v}^l + (\mathbf{u}^{l+1})^T\mathbf{K}\mathbf{u}^{l+1}\right) - \mathbf{f}\cdot\mathbf{u}^{l+1}$$
$$\leq \frac{1}{2}\left((\mathbf{v}^l)^T\mathbf{M}\mathbf{v}^{l+1} + (\mathbf{u}^l)^T\mathbf{K}\mathbf{u}^l\right) - \mathbf{f}\cdot\mathbf{u}^l = E(\mathbf{u}^l, \mathbf{v}^l),$$

as required. □

Notice that we apply the complementarity condition in Lemma 3.1, when we compute numerical solutions.

3.3. Convergence of the numerical scheme in time and space. In this subsection, we investigate the convergence of contact force N in the fully discretization. As we constructed the mass matrix $\mathbf{M}$ in subsection 3.1, we obtain the mass matrix $\mathbf{M} = \frac{4}{9}k \cdot \mathbf{B}$, where $\mathbf{B}$ is a 3-banded matrix. In Lemma 3.2, we will see that the inverse of $\mathbf{B}$ is bounded in the matrix 1-norm, based on the boundedness of its spectrum [2, p. 588] and results on inverses of banded matrices [5]. Let $[r_1, r_2]$, $r_1 > 0$ be an interval containing the spectrum of $\mathbf{B}$.

LEMMA 3.2. *Set* $r = r_2/r_1$, $q := q(r) := (\sqrt{r}-1)\ (\sqrt{r}+1)$, $C_0 = (1+r^{1/2})/(2r_1 r)$, *and* $\varpi = q^{2/3}$. *For any subinterval* $k > 0$, *we have*

$$\left\|\mathbf{B}^{-1}\right\|_1 \leq C\left(1 + 2\frac{\varpi}{1-\varpi}\right),$$

where $C := C(r_1, r) := \max\{r_1^{-1}, C_0\}$.

Proof. Using Propositions 2.1 and 2.2 in [5, p. 492], we have

$$\left\|\mathbf{B}^{-1}\right\|_1 = \max_{1\leq j\leq m+1}\sum_{i=1}^{m+1}\left|B_{ij}^{-1}\right| \leq C \max_{1\leq j\leq m+1}\sum_{i=1}^{m+1}\varpi^{|i-j|}. \tag{3.7}$$

To bound (3.7) simply, take $m \to \infty$. Then since $0 < \varpi < 1$,

$$\left\|\mathbf{B}^{-1}\right\|_1 \leq C\max_j\sum_{i=1}^{\infty}\varpi^{|i-j|} \leq C\sum_{i=-\infty}^{\infty}\varpi^i = C\left(1 + 2\frac{\varpi}{1-\varpi}\right),$$

as required. □

Recalling (2.18), let Borel measure $N_{h,k}$ be

$$\int_0^T \int_0^L N_{h,k}(x,t)\,dx\,dt = h \sum_{l=0}^{\lfloor T/h \rfloor -1} \int_0^L \sum_{i=1}^{m+1} \widehat{N}_i{}^l \psi_i(x) dx.$$

LEMMA 3.3. *Suppose that* (3.6) *is satisfied. Then Borel measures $N_{h,k}$ are uniformly bounded as measures on $[0,L]\times[0,T]$ as $h\downarrow 0$ and $k\downarrow 0$, i.e.,*

$$\int_0^T \int_0^L |N_{h,k}(x,t)|\,dx\,dt \le C,$$

where C does not depend on h, k.

Proof. Since $x \mapsto x^2/2$ is a quadratic function that satisfies the clamped end conditions at $x=0$, we can take $x^2/2 = \sum_{i=1}^{m+1} \alpha_i \psi_i(x)$ in the finite element space V_k as it is a spanned by a set of B-splines. The coefficients α_i can be exactly computed to be

$$\alpha_i = \begin{cases} k^2/6, & i=1, \\ \left(x_i^2 - k^2/3\right)/2, & i>1. \end{cases}$$

As we can see, $\alpha_i - x_i^2/2 \to 0$ as $k \to 0$. Using an argument similar to Lemma 2.7, we have

$$\int_0^L \frac{x^2}{2} N^l dx = \int_\eta^L \frac{x^2}{2} N^l dx = \int_\eta^L \sum_{i=1}^{m+1} \alpha_i \psi_i(x) \sum_{j=1}^{m+1} \widehat{N}_j^l \psi_j(x) dx.$$

Since $x^2/2 \ge \eta^2/2$ for $x \in [\eta, L]$ and $\mathbf{M}\mathbf{N}^l \ge \mathbf{0}$, we obtain

$$\begin{aligned}\int_0^L \frac{x^2}{2} N^l dx &\ge \int_0^L \sum_{i=1}^{m+1} \alpha_i \psi_i(x) \sum_{j=1}^{m+1} \widehat{N}_j^l \psi_j(x) = \sum_{j=1}^{m+1} \alpha_i \int_0^L \psi_i \psi_j dx\, \widehat{N}_j^l \\ &= \sum_{j=1}^{m+1} \alpha_i \int_0^L \psi_i \psi_j dx\, \widehat{N}_j^l = \sum_{i=1}^{m+1} \alpha_i \left(\mathbf{M}\mathbf{N}^l\right)_i.\end{aligned}$$

Since there is an $\eta > 0$ where $\widehat{u}_i^{l+1} + g(x_i) > 0$ for all $0 \le x_i \le \eta$ and all l, and $0 \le \mathbf{u}^l + \mathbf{g} \perp \mathbf{M}\,\mathbf{N}^l \ge 0$, we see that $(\mathbf{M}\,\mathbf{N}^l)_i = 0$ whenever $0 \le x_i \le \eta$. Thus

$$\sum_{i=1}^{m+1} \alpha_i \left(\mathbf{M}\mathbf{N}^l\right)_i \ge \min_{k\,i\ge\eta} \alpha_i \left\|\mathbf{M}\mathbf{N}^l\right\|_1 \ge \left(\frac{\eta^2}{2} - \frac{k^2}{6}\right) \left\|\mathbf{M}\mathbf{N}^l\right\|_1 \ge \frac{\eta^2}{4} \left\|\mathbf{M}\mathbf{N}^l\right\|_1$$

for sufficiently small $k>0$. From (2.19), $h\eta^2 \sum_{l=0}^{\lfloor T/h\rfloor -1} \left\|\mathbf{M}\mathbf{N}^l\right\|_1 \le C$, where C does not depend on h, k.

Therefore by Lemma 3.2,

$$\begin{aligned} h \sum_{l=0}^{\lfloor T/h\rfloor -1} \int_0^L \left| \sum_{i=1}^{m+1} \widehat{N}_i^l \psi_i(x) \right| dx &\le h \sum_{l=0}^{\lfloor T/h\rfloor -1} \sum_{i=1}^{m+1} \left|\widehat{N}_i^l\right| \|\psi_i\|_{L^1(0,L)} \\ &\le O(k)\, h \sum_{l=0}^{\lfloor T/h\rfloor -1} \left\|\mathbf{N}^l\right\|_1 \le O(k)\, h \sum_{l=0}^{\lfloor T/h\rfloor -1} \left\|\mathbf{M^{-1}}\mathbf{M}\mathbf{N}^l\right\|_1 \end{aligned}$$

$$\begin{aligned}&\le O(k)\left\|\mathbf{M}^{-1}\right\|_1 h\sum_{l=0}^{\lfloor T/h\rfloor-1}\left\|\mathbf{M}\mathbf{N}^l\right\|_1\\&=\frac{1}{\eta^2}O(k)O\left(\frac{1}{k}\right)=O(1),\end{aligned}$$

as required. □

We now need to show that any weak* limiting measure N is necessarily nonnegative. This is needed as the numerical method requires $\mathbf{M}\,\mathbf{N}^l\ge\mathbf{0}$ rather than $\mathbf{N}^l\ge\mathbf{0}$.

Lemma 3.4. *If N is a weak* limit of a subsequence $N_{h,k}\rightharpoonup^* N$ in the space of measures on $[0,T]\times[0,L]$, then $N\ge 0$.*

Proof. Suppose that $N_{h,k}\rightharpoonup^* N$ as measures on $[0,T]\times[0,L]$. Then for any continuous $\varphi\colon[0,T]\times[0,L]\to\mathbb{R}$, $\int_0^T\int_0^L\varphi(x,t)N_{h,k}(x,t)\,dx\,dt\to\int_0^T\int_0^L\varphi(x,t)N(x,t)\,dx\,dt$ in the subsequence with h, $k\to 0$. To show that $N\ge 0$ consider any $\varphi\colon[0,T]\times[0,L]\to\mathbb{R}_+$. Let $\varphi_k(x,t)=\sum_{i=1}^{m+1}\varphi(x_i,t)\psi_i(x)$, which is a pseudo-interpolant. Since φ is continuous on a compact set, it is uniformly continuous, and so for any $\epsilon>0$ there is a $\delta>0$ such that if $|x-x'|<\delta$ and $|t-t'|<\delta$ then $|\varphi(x,t)-\varphi(x',t')|<\epsilon$. If $x_i\le x\le x_{i+1}$ then $\psi_j(x)=0$ unless $i-1\le j\le i+2$. Choosing $0<k<\delta/2$, we see that

$$|\varphi(x,t)-\varphi_k(x,t)|\le\sum_{j=i-1}^{i+2}|\varphi(x,t)-\varphi(x_j,t)|\,\psi_j(x)\le\sum_{j=i-1}^{i+2}\epsilon\,\psi_j(x)=\epsilon.$$

Since this is true for all (x,t), $\varphi_k\to\varphi$ uniformly in $C([0,T]\times[0,L])$. Thus $\iint\varphi_k N_{h,k}\to\iint\varphi N$ as $h,k\to 0$ in a suitable subsequence. Now

$$\begin{aligned}\int_0^T\int_0^L\varphi_k(x,t)N_{h,k}(x,t)\,dx\,dt&=h\sum_{l=0}^{\lfloor T/h\rfloor-1}\int_0^L\sum_{i=1}^{m+1}\varphi(x_i,t_l)\psi_i(x)\sum_{j=1}^{m+1}N_j^l\psi_j(x)\,dx\\&=h\sum_{l=0}^{\lfloor T/h\rfloor-1}\sum_{i=1}^{m+1}\varphi(x_i,t_l)(\mathbf{M}\,\mathbf{N}^l)_i\ge 0,\end{aligned}$$

since $\mathbf{M}\,\mathbf{N}^l\ge\mathbf{0}$ and $\varphi(x_i,t_l)\ge 0$ for all i and l. Taking the limit in the subsequence gives

$$\int_0^T\int_0^L\varphi(x,t)N(x,t)\,dx\,dt\ge 0.$$

Since this holds for all continuous nonnegative φ, it follows that $N\ge 0$ in the sense of measures. □

From Lemmas 3.3 and 3.4 we see that there is a subsequence $N_{h,k}$ that converges weak* to a nonnegative measure N. It can also be shown that there are subsequences in which $u_{h,k}\rightharpoonup^* u$ in $L^\infty(0,T;H^2_{cf}(0,L))$ and $v_{h,k}\rightharpoonup^* v$ in $L^\infty(0,T;L^2(0,L))$ with $u_t=v$ and $v_t=-u_{xxxx}+N(x,t)$ by integrating against sufficiently smooth functions over $[0,T]\times[0,L]$. It is also clear that the numerical solutions $u_{h,k}$ are in a compact subset of $L^2([0,T]\times[0,L])$, and so $u_{h,k}+g\ge 0$ implies in the limit that $u+g\ge 0$ (continuity of the limit ensures that this holds everywhere, not just almost everywhere). To show that $u+g\perp N$ in the limit, we need a compactness result.

Noting that $du_{h,k}/dt(t) = \frac{1}{2}(v_{h,k}(t) + v_{h,k}(t-h))$ (with $v_{h,k}(t) = v^0$ for $-h \leq t < 0$), we see that $du_{h,k}/dt$ is uniformly bounded in $L^\infty(0,T;L^2(0,L))$ while $u_{h,k}$ is uniformly bounded in $C(0,T;H^2_{cf}(0,L))$. If we pick $0 < \theta < 1$ we can obtain the following bounds in the interpolation space $H^{2\theta}_{cf}(0,L) := [L^2(0,L),\, H^2_{cf}(0,L)]_\theta \subseteq H^{2\theta}(0,L)$ (see Kuttler [7, section 22.6, equation (62)], Triebel [16, Thm. 1.3.3(g)], and Bramble and Zhang [3, Appendix A, Thms. A.1 and A.2]):

$$\begin{aligned}\|u_{h,k}(t) - u_{h,k}(s)\|_{H^{2\theta}(0,L)} &\leq C_\theta\ \|u_{h,k}(t) - u_{h,k}(s)\|^\theta_{H^2(0,L)}\ \|u_{h,k}(t) - u_{h,k}(s)\|^{1-\theta}_{L^2(0,L)}\\ &\leq C_\theta\ \left(\|u_{h,k}(t)\|_{H^2(0,L)} + \|u_{h,k}(s)\|_{H^2(0,L)}\right)^\theta\\ &\qquad\times\left(\int_s^t \frac{1}{2}\,\|v_{h,k}(\tau) + v_{h,k}(\tau-h)\|_{L^2(0,L)}\,d\tau\right)^{1-\theta}\\ &\leq C_\theta\ \left(2\sqrt{2E^0}\right)^\theta\left(\sqrt{2E^0}\right)^{1-\theta}|t-s|^{1-\theta}\end{aligned}$$

so that the numerical trajectories $u_{h,k}$ are uniformly Hölder continuous as functions $[0,T] \to H^{2\theta}(0,L)$. Note that $E^0 := E(u^0,v^0)$ is the initial energy. For $2\theta > 1/2$, $H^{2\theta}(0,L)$ is compactly embedded into $C[0,L]$ [15, Props. 4.3 and 4.4], so by the Arzela–Ascoli theorem [8, Thm. III.3.1, p. 57] there is a convergent subsequence in $C([0,T]\times[0,L])$. Taking such a subsequence, we note that if $g_k(x) = \sum_{i=1}^{m+1} g_i\psi_i(x)$, $g_k \to g$ uniformly and so

$$\begin{aligned}\int_0^T\int_0^L (u_{h,k}(x,t) + g_k(x))\,N_{h,k}(x,t)\,dx\,dt &= \sum_{l=1}^{\lfloor T/h\rfloor-1}\sum_{i,j=1}^{m+1}\left(\widehat{u}_i^{l+1} + g_i\right)\widehat{N}_j^l\int_0^L \psi_i(x)\psi_j(x)\,dxh\\ &= \sum_{l=1}^{\lfloor T/h\rfloor-1}(\mathbf{u}^{l+1} + \mathbf{g})^T\mathbf{M}\,\mathbf{N}^l h = 0\end{aligned}$$

since $\mathbf{0} \leq \mathbf{u}^{l+1} + \mathbf{g} \perp \mathbf{M}\,\mathbf{N}^l \geq \mathbf{0}$. Taking the limit in the subsequence then gives

$$\int_0^T\int_0^L (u(x,t) + g(x))\,N(x,t)\,dx\,dt = 0$$

as desired. Thus there is a subsequence in which we get the desired convergence, and the limit for any converging subsequence satisfies the conditions for a solution.

3.4. Nonsmooth Newton method. To solve the linear system (3.3) for one time step with the linear complementarity condition (3.6), we consider using the nonsmooth Newton method (see Qi and Sun [11] for details). In order to find the next step solution $\mathbf{u}^{l+1}$ from the linear system (3.3) and the complementarity condition (3.6), we consider the mapping $\mathbf{F} : \mathbf{R}^{m+1} \to \mathbf{R}^{m+1}$:

$$\mathbf{F} : \mathbf{u}^{l+1} \mapsto \min(\mathbf{M}\mathbf{N}^l, \mathbf{u}^{l+1} + \mathbf{g}). \tag{3.8}$$

Note that $\min(\mathbf{a},\mathbf{b})$ is meant componentwise for vectors $\mathbf{a}$ and $\mathbf{b}$, and so $\min(\mathbf{a},\mathbf{b}) = \mathbf{0}$ is equivalent to $0 \leq \mathbf{a} \perp \mathbf{b} \geq 0$. Thus the complementarity condition (3.6) is equivalent to $\mathbf{F}(\mathbf{u}^{l+1}) = \mathbf{0}$. Since $\mathbf{M}\mathbf{N}^l$ is implicitly a function of $\mathbf{u}^{l+1}$ via the linear system (3.3), we can express $\mathbf{M}\mathbf{N}^l$ as

$$\mathbf{M}\mathbf{N}^l = \frac{2}{h^2}\left[\left(\mathbf{M} + \frac{h^2}{4}\mathbf{K}\right)\mathbf{u}^{l+1} - \left(\mathbf{M} - \frac{h^2}{4}\mathbf{K}\right)\mathbf{u}^l - h\mathbf{M}\mathbf{v}^l\right] - \mathbf{f}. \tag{3.9}$$

Table 3.1
Average number of linear systems solved per time step.

	$k = 1/5$	$k = 1/25$	$k = 1/50$	$k = 1/500$
$h = 1/10$	20.48	28.08	29.36	35.79
$h = 1/20$	19.68	25.95	26.99	27.23
$h = 1/50$	19.20	23.36	23.96	18.28
$h = 1/100$	19.19	22.79	23.01	19.14
$h = 1/1000$	19.28	22.87	19.87	14.60

We can find the next approximate solution $\mathbf{u}^{l+1}$ by using the nonsmooth Newton method:

$$\mathbf{u}_{n+1}^{l+1} = \mathbf{u}_n^{l+1} - \nabla\mathbf{F}(\mathbf{u}_n^{l+1})^{-1}F(\mathbf{u}_n^{l+1}) \text{ for } n \geq 0.$$

Even though $\mathbf{F}$ is a nonsmooth function, Newton's method can still converge superlinearly since $\mathbf{F}$ is a *semismooth* function [10, 9]. This is because max and min are semismooth functions, and Newton's method for semismooth function still converges locally at a superlinear rate provided $\mathbf{F}$ is "BD regular" [11]. That is, it converges superlinearly provided $\partial_B\mathbf{F}(\mathbf{u}) := \{\lim_{j\to\infty}\nabla\mathbf{F}(\mathbf{u}_j) \mid \lim_{j\to\infty}\mathbf{u}_j = \mathbf{u}\}$ does not contain any singular matrices.

In practice, in order to obtain computation, we use a smooth approximation $\theta_\alpha(a,b)$ to $\min(a,b)$

$$\theta_\alpha(a,b) = \frac{1}{2}\left((a+b) - h_\alpha(a-b) + \alpha\right),$$

where $h_\alpha(y) = \sqrt{y^2+\alpha^2} - \alpha$ is an approximation to $|y|$. $\alpha > 0$ is chosen by an adaptive strategy reducing by a factor of 10 or increasing by a factor of 2 in order to succeed the guarded Newton method. The number α is called a smoothing parameter. Clearly, as $\alpha \to 0$, we have

$$\theta_\alpha(a,b) \to \min(a,b). \tag{3.10}$$

At each stage we solve $\mathbf{F}_\alpha(\mathbf{u}) := \theta_\alpha(\mathbf{MN}^l, \mathbf{u}^{l+1} + \mathbf{g}) = 0$ using a guarded Newton method and then reduce α, usually by a factor of 10, and repeat the procedure until α is sufficiently small. We use the function $\|\mathbf{F}_\alpha(\mathbf{u})\|_2$ as the merit function for the guarded Newton method.

This method has proven to be quite efficient, typically requiring only 20 to 30 linear solves per time step, as can be seen in Table 3.1. This means that the time for the computations per time step grows linearly with the size of the problem as shown in Table 3.3.

3.5. Numerical evidence for strong convergence. In this subsection, we present the numerical evidence that our numerical solutions converge strongly (via Lemma 2.10), and our method of obtaining and assessing this evidence. Let $\boldsymbol{\phi}_i$ be the ith eigenvector with eigenvalue λ_i of the generalized eigenproblem (3.11). Then we have

$$\boldsymbol{\phi}_i^T\mathbf{M}\boldsymbol{\phi}_i = 1 \text{ and } \mathbf{K}\boldsymbol{\phi}_i = \lambda_i\mathbf{M}\boldsymbol{\phi}_i, \tag{3.11}$$

where $\boldsymbol{\phi}_i = ((\phi_i)_1, (\phi_i)_2, (\phi_i)_3, \ldots, (\phi_i)_{m+1})$. Note that this is the Galerkin discretization of the eigenfunction problem $\partial^4\phi_i(x)/\partial x^4 = \lambda_i\phi_i(x)$ and $\int_0^L \phi_i(x)^2 = 1$ with our boundary conditions. Also note that $\mathbf{M}^{-1}\mathbf{K}$ is self-adjoint with respect to the inner products $\langle \mathbf{z}, \mathbf{w}\rangle_{\mathbf{M}} = \mathbf{z}^T\mathbf{M}\mathbf{w}$ and $\langle \mathbf{z}, \mathbf{w}\rangle_{\mathbf{K}} = \mathbf{z}^T\mathbf{K}\mathbf{w}$. So for any given function $\beta : \mathbf{R} \to \mathbf{R}$, we can define $\beta(\mathbf{M}^{-1}\mathbf{K})$ via $\beta(\mathbf{M}^{-1}\mathbf{K})\boldsymbol{\phi}_i = \beta(\lambda_i)\boldsymbol{\phi}_i$. In particular, let $\kappa^*(\lambda) = 1$ if $\lambda \le \lambda_c$ and $\kappa^*(\lambda) = 0$ otherwise. The $\kappa^*(\mathbf{M}^{-1}\mathbf{K})\mathbf{z}$ is the projection onto $\operatorname{span}\{\boldsymbol{\phi}_i \mid i = 1, 2, \dots,$ and $\lambda_i \le \lambda_c\}$ that is orthogonal with respect to both $\langle \cdot, \cdot\rangle_{\mathbf{M}}$ and $\langle \cdot, \cdot\rangle_{\mathbf{K}}$. The elastic energy in the modes i with $\lambda_i \le \lambda_c$ is therefore $\frac{1}{2}(\kappa^*(\mathbf{M}^{-1}\mathbf{K})\mathbf{u})^T\mathbf{K}\,\kappa^*(\mathbf{M}^{-1}\mathbf{K})\mathbf{u}$ and the kinetic energy is $\frac{1}{2}(\kappa^*(\mathbf{M}^{-1}\mathbf{K})\mathbf{v})^T\mathbf{M}\,\kappa^*(\mathbf{M}^{-1}\mathbf{K})\mathbf{v}$. Since $\kappa^*(\mathbf{M}^{-1}\mathbf{K})$ is not easily computable without performing a complete (and expensive) eigenvalue/eigenvector decomposition of $\mathbf{M}^{-1}\mathbf{K}$, we will instead construct a rational approximation to it.

Choosing $\lambda_c > 0$ for any cut-off $c \ge 1$, we have $(1/\lambda_c)\mathbf{M}^{-1}\mathbf{K}\boldsymbol{\phi}_i = (\lambda_i/\lambda_c)\boldsymbol{\phi}_i$. Thus for any large integer $p > 0$,

$$\left(\mathbf{I} + \left(\frac{1}{\lambda_c}\mathbf{M}^{-1}\mathbf{K}\right)^{2p}\right)^{-1}\boldsymbol{\phi}_i = \frac{1}{1 + (\lambda_i/\lambda_c)^{2p}}\boldsymbol{\phi}_i.$$

Then we fix a continuous map κ of λ, which approximates the step function $\kappa^*(\lambda)$:

$$\kappa(\lambda) = \frac{1}{(1 + (\lambda/\lambda_c)^{2p})} \text{ and then } \kappa(\mathbf{M}^{-1}\mathbf{K}) = \left(\mathbf{I} + \left(\frac{1}{\lambda_c}\mathbf{M}^{-1}\mathbf{K}\right)^{2p}\right)^{-1}. \tag{3.12}$$

Employing (3.11), it can be proved that at each time step $l \ge 1$, the energy in the fully discrete case with no body force is

$$\frac{1}{2}\left((\mathbf{v}^l)^T\mathbf{M}\mathbf{v}^l + (\mathbf{u}^l)^T\mathbf{K}\mathbf{u}^l\right) = \frac{1}{2}\sum_{i=1}^{m+1}\left(|\widehat{v}_i^l|^2 + \lambda_i|\widehat{u}_i^l|^2\right). \tag{3.13}$$

Using (3.13), we can demonstrate numerical evidence using Lemma 2.10 that the convergence is strong. The ratio

$$\frac{(\kappa(\mathbf{M}^{-1}\mathbf{K})\mathbf{u}^l)^T\mathbf{K}\,\kappa(\mathbf{M}^{-1}\mathbf{K})\mathbf{u}^l + (\kappa(\mathbf{M}^{-1}\mathbf{K})\mathbf{v}^l)^T\mathbf{M}\,\kappa(\mathbf{M}^{-1}\mathbf{K})\mathbf{v}^l}{(\mathbf{u}^l)^T\mathbf{K}\,\mathbf{u}^l + (\mathbf{v}^l)^T\mathbf{M}\,\mathbf{v}^l}$$

is the ratio of the elastic and kinetic energy in the modes with $\lambda_i \le \lambda_c$ to the total elastic and kinetic energy for the numerical solution at time step l. Following Lemma 2.10, this should go to one as $\lambda_c \uparrow \infty$, uniformly in the numerical parameters $h > 0$, l and $k > 0$. Of course, for *fixed* $k > 0$, this will happen as $\lambda_c \uparrow \infty$ anyway. So we need to first fix λ_c and then compute these ratios for k and h becoming small; from the apparent limits of the energy ratios for several fixed λ_c, we observe the overall trend as $\lambda_c \uparrow \infty$. This will be done in subsection 3.6.

3.6. Computing $\boldsymbol{\kappa}(\mathbf{M}^{-1}\mathbf{K})\mathbf{z}$.

In this subsection, we discuss how to efficiently compute $\kappa(\mathbf{M}^{-1}\mathbf{K})\mathbf{z}$. Note that we do not compute $\kappa^*(\mathbf{M}^{-1}\mathbf{K})$ directly using an eigendecomposition of $\mathbf{M}^{-1}\mathbf{K}$, as this is computationally expensive. So we use a rational function $\kappa(\lambda)$ to approximate the step function $\kappa^*(\lambda)$. We can then compute $\kappa(\mathbf{M}^{-1}\mathbf{K})\mathbf{z}$ for any vector $\mathbf{z}$. For simplicity we choose

$$\kappa(\lambda) = \frac{1}{1 + (\lambda/\lambda_c)^{2p}} \text{ for } p \text{ moderately large.}$$

Table 3.2

The ratio of energy $\boldsymbol{E_c}$ in low frequency modes to total energy $\boldsymbol{E}$ in actual computation.

The number of nodes	c	$h=1/10$	$h=1/50$	$h=1/100$
500	10	0.650153	0.407755	0.380260
	30	0.910487	0.812236	0.777214
	100	0.997099	0.986641	0.972011
	300	0.999846	0.999166	0.997870
1000	10	0.653481	0.412869	0.378693
	30	0.917148	0.855211	0.755536
	100	0.997575	0.981944	0.968371
	300	0.999846	0.998196	0.997980

In fact, we implement this function for $p = 5$. The key to efficient computation of $\kappa(\mathbf{M}^{-1}\mathbf{K})\mathbf{z}$ is the factorization of $\kappa(\lambda)$. The zeros of the denominator (3.12) are solutions of $(\lambda_j/\lambda_c)^{2p} = -1$. The solutions of this equation are

$$\lambda_j/\lambda_c = \zeta_j := \exp((2j+1)\pi i/2p), \quad j = 0, 1, 2, \ldots, 2p-1, \text{ where } i = \sqrt{-1}.$$

Thus since $\kappa(\lambda) = \lambda_c^{2p}(\lambda - \lambda_c\zeta_0)^{-1}(\lambda - \lambda_c\zeta_1)^{-1}\cdots(\lambda - \lambda_c\zeta_{2p-1})^{-1}$, we have

$$\kappa(\mathbf{M}^{-1}\mathbf{K}) = \lambda_c^{2p}(\mathbf{K} - \lambda_c\zeta_0\mathbf{M})^{-1}\mathbf{M}(\mathbf{K} - \lambda_c\zeta_1\mathbf{M})^{-1}\cdots\mathbf{M}(\mathbf{K} - \lambda_c\zeta_{2p-1}\mathbf{M})^{-1}\mathbf{M}.$$

The right side of the linear system $(\mathbf{K} - \lambda_c\zeta_j\mathbf{M})\mathbf{z} = \mathbf{w}$ has matrices over the complex numbers $\mathbf{C}$. We can change the complex matrix into two real ones so that we have an equivalent real banded matrix with double the bandwidth. Thus the linear system $(\mathbf{K} - \lambda_c\zeta_j\mathbf{M})\mathbf{z} = \mathbf{w}$ can be solved as a banded system with an upper and lower bandwidth of six, which can be done in $O(m+1)$ time. The matrix-vector products $\mathbf{Mz}$ can also be computed in $O(m+1)$ time. Thus $\kappa(\mathbf{M}^{-1}\mathbf{K})\mathbf{z}$ can be computed in just $O(p(m+1))$ time.

The ratios contained in Table 3.2 are obtained as follows: Let $E(\mathbf{u}^l, \mathbf{v}^l)$ be the total energy in actual computation and let $E_c(\mathbf{u}^l, \mathbf{v}^l)$ be the energy in the low frequency modes. Then the ratio that we use is

$$\tau = \frac{\sum_{l=0}^{\lfloor T/h \rfloor} E_c(\mathbf{u}^l, \mathbf{v}^l)}{\sum_{l=0}^{\lfloor T/h \rfloor} E(\mathbf{u}^l, \mathbf{v}^l)}.$$

Looking across the rows of Table 3.2, we note that there does seem to be some slow convergence of the ratios as h goes to zero, and this ratio increases as m (the number of grid nodes) increases; this limit seems to be very close to one for large λ_c. By picking $c = 100$, i.e., the lowest 100 out of 500 or 1000 possible modes, we can account for about 97 percent of the total kinetic and elastic energy. This implies that we can account for almost all the energy in the bottom 100 frequency modes and account for about 75 percent of the total energy in the bottom 30 modes. So Table 3.2 presents substantial numerical evidence of the applicability of Lemma 2.10 and, therefore, of strong convergence of the numerical solutions.

3.7. Numerical experiments and discussion of the results. The package that we used for handling the matrices and vectors is Meschach [14], which uses the C programming language. We took particular advantage of the banded matrix routines in that package. Our numerical experiments were performed on a Hewlett-Packard Visualize B2000.

In this subsection, we show our numerical simulation results. In our computation, we take the length of the rod to be $L = 20$ and the initial displacement $u^0(x) = x^2/4$,

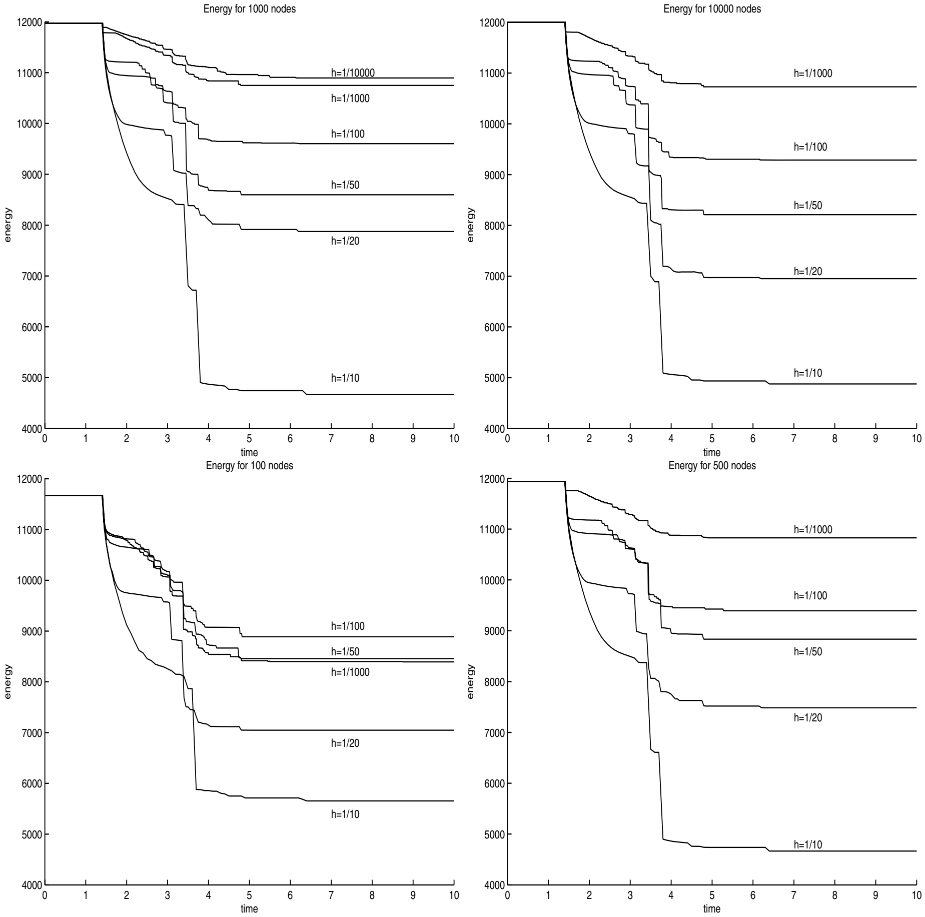


Fig. 2. *Energy function.*

which is consistent with the essential boundary condition and the initial velocity $v^0(x) = -2 \cdot x$ and gap function $g(x) = (x - 12)^2$, and the end time $T = 10$. We assume that the rod is moving downward, in a negative direction in simulation. The gap function g indicates the distance between the rigid foundation and the initial position where the rod is located vertically. Note that the potential energy is not included in our computation.

From the energy functional in (3.5) in the fully-discrete case, we obtain four graphs for the total energy in Figure 2. According to those graphs, our numerical implementation supports the energy dissipation that we anticipated theoretically. The first graph shows that the energy function using 100 nodes is erratic. Indeed, we anticipated that the smaller time step size h we used, the higher the energy. This appears to be true for all the cases except for $k = 0.2$ and for $h = 0.01$ and $h = 0.001$. We would conjecture that the reason is that the approximations are not sufficiently refined for this value of k. On the other hand, other graphs show that energy conservation is expected as step size h becomes smaller and smaller.

In Figure 3, the motion of the rod is presented. Each curve is the profile of the rod at a given time. In this simulation, we used $k = 1/1000$ in space and a time step of $h = 1/100$. According to our numerical experiments, that case brings the most

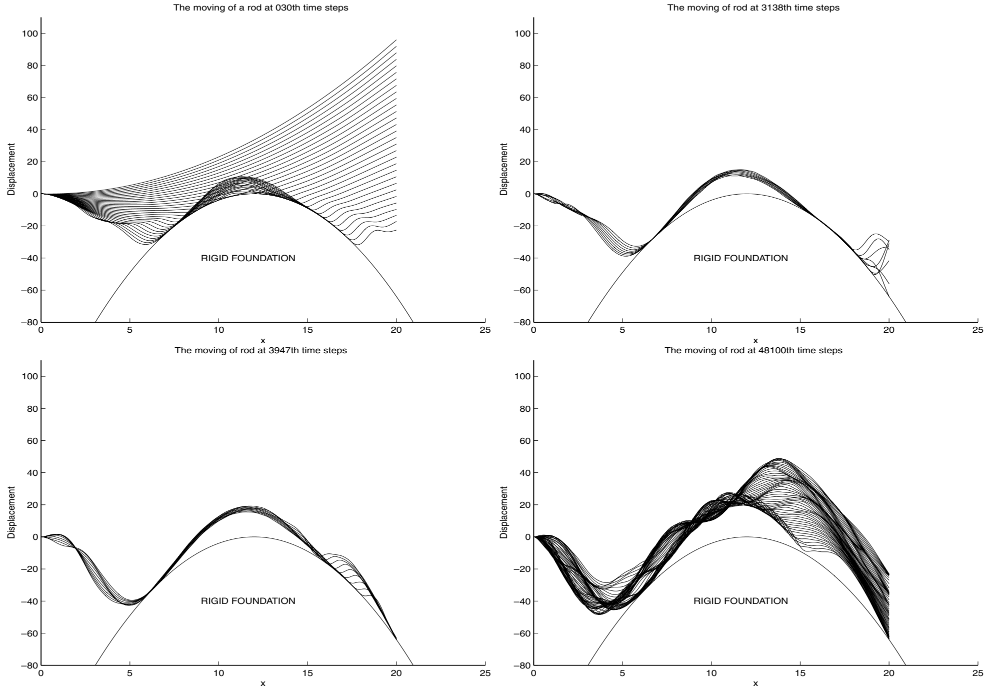


FIG. 3. *Flow of solution.*

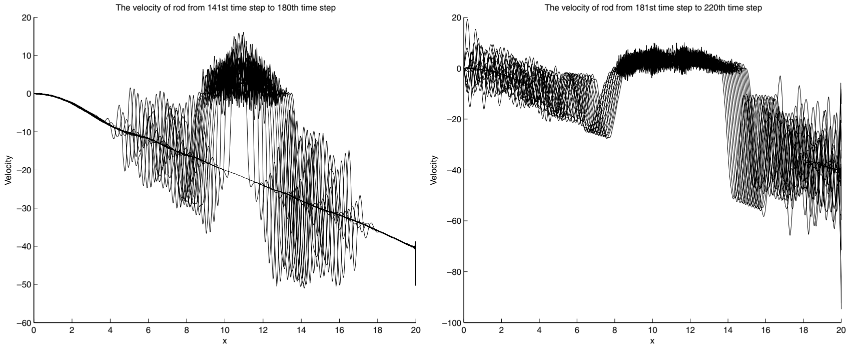


FIG. 4. *The velocity of the rod at each time step* 141–220.

comfortable and solid result. An interesting point is that the end of the rod touches the rigid foundation at some time step, and it oscillates very rapidly. See the pictures at the top right and bottom left of Figure 3.

Figures 4 and 5 present the velocity of the rods. Particularly in Figure 4, high frequencies appear when the rod touches the rigid foundation. So we would guess by the phenomenon that the rate of deformation of the rod is very fast in the time steps. Figure 5 shows the velocity after the rod bounces away from the rigid foundation.

Finally, we have 3-dimensional pictures showing the contact force in Figure 6. When the end of the rod touches the rigid foundation, the contact force is largest there. Even though the number of nodes in the two pictures are different and they show different magnitudes for the contact force, the graphs have a similar shape.

Table 3.3 shows the speed of the computations. Note that in the case of $k = 1/500$

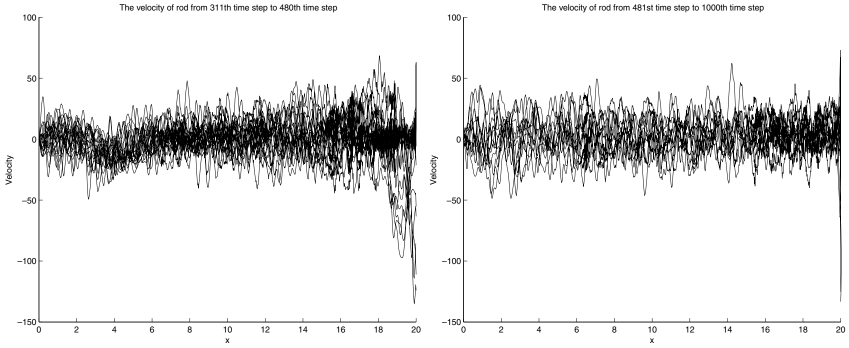


FIG. 5. *The velocity of the rod at each time step* 311–1000.

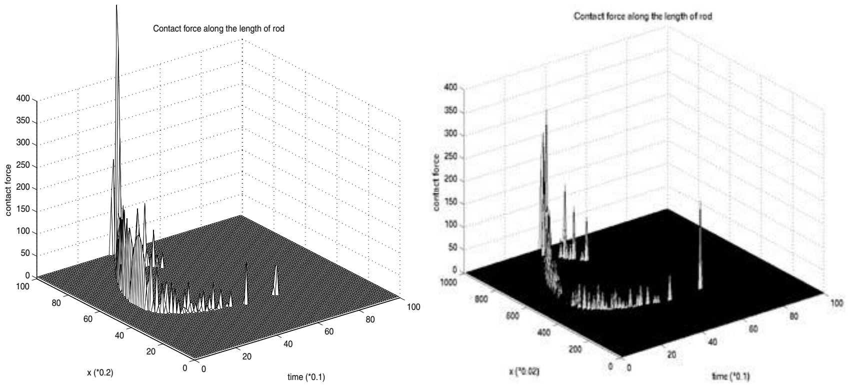


FIG. 6. *Contact force at each time step.*

TABLE 3.3
Computation time (u:user time, s:system time).

$h \ \backslash \ k$	2×10^{-1}	4×10^{-2}	1×10^{-2}	2×10^{-3}
1×10^{-1}	0.783u 0.007s	6.460u 0.041s	19.892u 0.035s	1083.158u 2.931s
5×10^{-2}	1.503u 0.003s	11.082u 0.044s	37.642u 0.113s	1295.474u 3.453s
2×10^{-2}	3.632u 0.011s	25.968u 0.033s	81.783u 0.158s	1633.937u 3.597s
1×10^{-2}	7.283u 0.039s	47.621u 0.179s	151.851u 0.255s	3852.925u 7.367s
1×10^{-3}	73.242u 0.390s	477.255u 1.234s	1220.408u 2.054s	29675.837u 65.416s

we use the different convergence criterion $\|\mathbf{F}(\mathbf{u})\|_2 < \epsilon$ for stopping. This was necessary because of difficulties with roundoff and ill-conditioning in the stiffness matrix $\mathbf{K}$. So instead we used $\|\nabla\mathbf{F}(\mathbf{u})^{-1}\mathbf{F}(\mathbf{u})\|_2 < \epsilon$ to avoid these numerical difficulties. So in Table 3.3, we can see that the ratio of times differs from the other cases.

4. Conclusion.

In this paper we consider semidiscrete and fully discrete approximations to the motion of an Euler–Bernoulli beam with frictionless contact. For both the semidiscrete and fully discrete approximations, we are able to show that there is a subsequence of the discrete approximations that converges (albeit in a sufficiently weak sense) to a (weak) solution of the PDE and the Signorini contact conditions. The fully discrete approximation was developed using the finite element method using B-splines to construct the basis functions. This scheme was implemented, and the linear complementarity problems (LCPs) that arise at each time step were solved using a smoothed guarded Newton method applied to a reformulation of the LCP as a nonsmooth equation. These methods turn out to be quite efficient, especially since the one-dimensional structure of the problem results in banded matrices when handled properly. Furthermore, the number of linear systems solved per time-step seems not to grow as the discretization parameters (h in time and k in space) go to zero.

Open questions of particular interest to the authors are the question of strong convergence of the solutions and the related question of whether the limiting solution conserves energy or not. A numerical scheme is devised in this paper to test the question of strong convergence in a computationally efficient manner. The results from the computation give evidence that the numerical solutions for our problem do indeed converge strongly, and, even though the time-discretization is dissipative, the limit solution also conserves energy. No analytical demonstration of energy conservation is given; it can be demonstrated to be false in general, but it may be true generically.

Acknowledgments. We would like to acknowledge the comments of an anonymous referee, which improved the presentation of this paper and prompted us to complete the convergence proof for the discretization in both time and space.

REFERENCES

[1] J. Ahn and D. E. Stewart, *An Euler–Bernoulli Beam with Dynamic Contact: Penalty Approximation and Existence*, Computational Mathematics Technical Report 163, University of Iowa, http://www.www.math.uiowa.edu/comp-math/comp-math-2005.htm.

[2] K. E. Atkinson, *An Introduction to Numerical Analysis*, 2nd ed., J. Wiley and Sons, 1988.

[3] J. H. Bramble and X. Zhang, *The analysis of multigrid methods*, in Handbook of Numerical Analysis, Vol. VII, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 2000, pp. 391–402.

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

[5] S. Demko, W. F. Moss, and P. W. Smith, *Decay rates for inverse of band matrices*, Math. Comp., 43 (1984), pp. 491–499.

[6] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.

[7] K. Kuttler, *Modern Analysis*, CRC Press, Boca Raton, FL, 1998.

[8] S. Lang, *Real and Functional Analysis*, 2nd ed., Grad. Texts in Math. 142, Springer-Verlag, Berlin, Heidelberg, New York, 1993.

[9] R. Mifflin, *An algorithm for constrained optimization of semi-smooth functions*, Math. Oper. Res., 2 (1977), pp. 191–207.

[10] R. Mifflin, *Semismooth and semiconvex functions in constrained optimization*, SIAM J. Control Optim., 15 (1977), pp. 959–972.

[11] L. Q. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program., 58 (1993), pp. 353–367.

[12] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Texts Appl. Math. 13, Springer-Verlag, New York, Berlin, Heidelberg, 1993.

[13] M. Schatzman, *A hyperbolic problem of second order with unilateral constraints: The vibrating string with a concave obstacle*, J. Math. Anal. Appl., 73 (1980), pp. 138–191.

[14] D. E. Stewart and Z. Leyk, *Meschach: Matrix Computations in C*, Proc. Centre Math. Appl. Austral. Nat. Univ. 32, Australian National University, Canberra, 1994.

[15] M. E. Taylor, *Partial Differential Equations* I*: Basic Theory*, Appl. Math. Sci. 115, Springer-Verlga, New York, 1996.

[16] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North-Holland, Amsterdam, New York, 1978.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1481–1503

A POSTERIORI ERROR ESTIMATIONS OF SOME CELL-CENTERED FINITE VOLUME METHODS[*]

SERGE NICAISE[†]

Abstract. This paper presents the natural framework to residual based a posteriori error estimation of some cell-centered finite volume methods for the Laplace equation in $\mathbb{R}^d, d = 2$ or 3. For that purpose we associate with the finite volume solution a reconstructed approximation, which is a kind of Morley interpolant. The error is then the difference between the exact solution and this Morley interpolant. The residual error estimator is based on the jump of normal and tangential derivatives of the Morley interpolant. We then prove the equivalence between the discrete H^1 seminorm of the error and the residual error estimator. Numerical tests confirm our theoretical results.

Key words. finite volume method, cell-centered method, a posteriori error estimates

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/S0036142903437787

1. Introduction.

The finite volume method is a well-adapted method for the discretization of various partial differential equations and is very popular in the engineering community [24]. The mathematical community recently started to analyze it in detail. Presently, existence and uniqueness results as well as a priori error estimates are available for a quite large class of problems; we refer to [10] and the references cited there. Contrary to the finite element methods [26], a posteriori error estimates for finite volume methods are less developed, and until now only a few results have been obtained in that direction. See [14, 22, 1, 12, 13] for cell-centered finite volume methods, [17, 19, 25, 23] for vertex-centered methods, and [2, 3, 15, 16] for finite volume element methods. Since finite volume methods have some similarities with the finite element methods, we may hope that this gap will be filled soon.

The goal of our paper is to present the natural framework to residual based a posteriori (efficient and reliable) error estimation of some cell-centered finite volume methods for linear elliptic equations. In a first attempt we restrict ourselves to the Laplace equation in $\mathbb{R}^d, d = 2$ or 3. The case of diffusion–convection–reaction equations will be only sketched; for details, we refer to a forthcoming paper [20]. The key idea is the reconstruction of a piecewise polynomial approximation of the finite volume solution, its principal property being that the mean of its flux through any edge/face of the mesh is equal to the numerical flux through that edge/face (this interpolant is consequently smoother than the approximated solution). This reconstructed approximation is then a kind of Morley interpolant of the finite volume solution. In general a Morley interpolant is not in H^1, and therefore the Morley interpolant may be considered as a nonconforming approximation of the exact solution. The second key idea is to use the Helmholtz decomposition of the error, the difference between the exact solution and this Morley interpolant, as was done in [7] for the a posteriori error analysis of a nonconforming finite element approximation of the Laplace equation. As in

[*] Received by the editors November 17, 2003; accepted for publication (in revised form) March 24, 2005; published electronically October 19, 2005.

http://www.siam.org/journals/sinum/43-4/43778.html

[†] Université de Valenciennes et du Hainaut Cambrésis, MACS, ISTV, F-59313 Valenciennes Cedex 9, France (snicaise@univ-valenciennes.fr).

[7] the residual error estimator is then naturally based on the jump of normal and tangential derivatives of the Morley interpolant. We finally show the equivalence between the discrete H^1 seminorm of the error and the residual error estimator. The proof of the upper error bound uses the Helmholtz decomposition of the error and some quasi-orthogonality relations obtained using the above-mentioned property of the Morley interpolant. The proof of the lower error bound is more standard and simply uses some Green's formulas and inverse inequalities as for finite element methods [26].

Note that our purposes also require the introduction of new finite elements of Morley type on rectangles and tetrahedra.

We further give explicitly the size of the constants appearing in the error estimates by estimating the constants involved in the interpolation error estimates (using some related eigenvalue problems and extension techniques) and in some inverse inequalities. In particular, we obtain constants in the upper error bound that are quite close to unity.

The idea to interpolate the finite volume solution by a smoother function having the above-mentioned property on the flux was presented in [12] in an L^1 framework for time-dependent nonlinear convection–diffusion equations in $\mathbb{R}^d \times \mathbb{R}^+$. In that paper the authors obtain a reliable estimator in an L^1-norm, instead of the energy norm. Furthermore, their interpolant is a piecewise linear Lagrange interpolant on a dual mesh. As a consequence, to guarantee the property on the flux, the (primal) mesh has to be admissible in the sense of [10, Def. 9.1], a deep obstacle for adaptivity. To avoid this admissibility condition and use the energy norm framework, we need to use the natural degrees of freedom on the mesh, namely, the mean of the flux on the edges/faces, and consequently use higher order polynomials.

The outline of the paper is as follows: In section 2 we describe the so-called cell-centered method for the Laplace equation on a mesh made of triangles, rectangles, or tetrahedra. Some standard inverse inequalities and interpolation error estimates are recalled in section 3, where some constants are specified as explicitly as possible. Section 4 is devoted to the introduction of some finite elements of Morley type. In section 5 we introduce the Morley interpolant of the approximated solution and prove its main properties. The upper and lower error bounds are then deduced in section 6. The upper error bound is based on the properties of the Morley interpolant and the use of the Helmholtz decomposition of the error, while the lower error bound is proved in a quite standard way. In section 7 we briefly describe how to extend our results to diffusion–convection–reaction equations. Finally, section 8 is devoted to numerical experiments that confirm our theoretical considerations.

2. Discretization of the Laplace equation. Let Ω be an open subset of $\mathbb{R}^d$, $d = 2$ or 3, with a polygonal ($d = 2$) or polyhedral ($d = 3$) boundary Γ.

As usual, we denote by $L^2(\cdot)$ the Lebesgue spaces and by $H^s(\cdot)$, $s \geq 0$, the standard Sobolev spaces. If D is an open subset of $\mathbb{R}^d$, $d = 2$ or 3, the usual norm and seminorm of $H^s(D)$ are denoted by $\| \cdot \|_{s,D}$ and $| \cdot |_{s,D}$. For brevity the $L^2(D)$-norm will be denoted by $\| \cdot \|_D$ and in the case $D = \Omega$, we will drop the index Ω. The space $H_0^1(\Omega)$ is defined, as usual, by $H_0^1(\Omega) := \{v \in H^1(\Omega)/v = 0 \text{ on } \Gamma\}$. In what follows the symbol $| \cdot |$ will denote either the Euclidean norm in $\mathbb{R}^d$, $d = 2$ or 3, or the length of a line segment, or the area of a plane face, or finally the measure of a domain of $\mathbb{R}^d$.

We consider the standard elliptic problem: for $f \in L^2(\Omega)$ let $u \in H_0^1(\Omega)$ be the variational solution of

$$-\Delta u = f \text{ in } \Omega, \tag{1}$$

which means that u satisfies

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall v \in H_0^1(\Omega). \tag{2}$$

To approximate this problem by a finite volume scheme we fix a family of meshes $T_h, h > 0$, regular in Ciarlet's sense [4, p. 124]. In two dimensions we assume that all elements of T_h are either triangles or rectangles, while in three dimensions the mesh consists only of tetrahedra. For $K \in T_h$ we recall that h_K is the diameter of K and $h = \max_{K \in T_h} h_K$.

For any edge/face E of K, we denote by $h_{E,K}$ its height in K, namely, $h_{E,K} = \frac{d|K|}{|E|}$ if K is a triangle or a tetrahedron and $h_{E,K} = \frac{|K|}{|E|}$ if K is a rectangle. For an edge/face E, its mean height is $h_E = \frac{1}{2}(h_{E,K} + h_{E,L})$, when E is the edge/face of K and L. The regularity of the mesh implies in particular that for any edge/face E of K one has

$$\sigma_1 h_{E,K} \le h_K \le \sigma_2 h_{E,K}, \tag{3}$$

$$\sigma_3 h_{E,K} \le h_E \le \sigma_4 h_{E,K} \tag{4}$$

for some positive constants $\sigma_i, i = 1, \dots, 4$, depending on the aspect ratio of T_h.

Let us define E_h as the set of edges ($d = 2$) or faces ($d = 3$) of the triangulation and set $E_h^{int} = \{E \in E_h / E \subset \Omega\}$ the set of interior edges/faces of T_h, while $E_h^{ext} = E_h \setminus E_h^{int}$ is the set of exterior edges/faces of T_h.

For an edge E of a two-dimensional (2D) element K, introduce $n_{K,E} = (n_x, n_y)$ the unit outward normal vector to K along E. Similarly for a face E of a tetrahedron K, set $n_{K,E} = (n_x, n_y, n_z)$ the unit outward normal vector to K on E. Furthermore, for each edge/face E, we fix one of the two normal vectors and denote it by n_E. In two dimensions additionally introduce the tangent vector $t_{K,E} = n_{K,E}^\perp := (-n_y, n_x)^\top$ such that it is oriented positively (with respect to K); similarly set $t_E := n_E^\perp$.

The jump of some function v across an edge/face E at a point $y \in E$ is defined as

$$\begin{aligned} [\![v(y)]\!]_E &:= \lim_{\alpha \to +0} v(y + \alpha n_E) - v(y - \alpha n_E) \quad \forall E \in E_h^{int}, \\ [\![v(y)]\!]_E &:= v(y) \quad \forall E \in E_h^{ext}. \end{aligned}$$

For any $K \in T_h$ or $E \in E_h$, we denote by $\mathcal{M}_K \chi$ and $\mathcal{M}_E \chi$ the mean of χ on K and E, respectively, i.e.,

$$\mathcal{M}_K \chi = \frac{1}{|K|} \int_K \chi(x) \, dx \quad \forall K \in T_h, \qquad \mathcal{M}_E \chi = \frac{1}{|E|} \int_E \chi(x) \, ds(x) \quad \forall E \in E_h.$$

Finally, we will need local subdomains (also called patches). As usual, let ω_K be the union of all elements having a common edge/face with K. Similarly let ω_E be the union of both elements having E as edge/face.

The finite volume approximation u_h of u is piecewise constant on T_h, i.e., $u_h := (u_K)_{K \in T_h}$ (u_K being the approximation of $u(x_K)$ for $K \in T_h$, x_K being the "center" of the box K). To deduce the approximated equation satisfied by u_h, we first integrate (1) on a control volume K and use the divergence formula to obtain

$$- \sum_{E \in E_K} \int_E \nabla u \cdot n_{K,E} \, ds = \int_K f(x) \, dx \quad \forall K \in T_h, \tag{5}$$

where E_K is the set of edges/faces of K. The diffusion flux $\int_E \nabla u \cdot n_{K,E}$ is approximated by a numerical diffusion flux $F_{K,E}(u_h)$ obtained using quadrature rules and finite differences (see, e.g., [10]) and is consequently a linear combination of some values of u_h around E [10, 5, 6]. For our further uses we do not need its exact form but the principle of conservation of flux is required: $F_{K,E}(u_h) = -F_{L,E}(u_h)$ for $E = \bar{K} \cap \bar{L}$. These approximations lead to the following system. Find a solution $u_h := (u_K)_{K \in T_h}$ of

$$- \sum_{E \in E_K} F_{K,E}(u_h) = \int_K f(x)\, dx \quad \forall K \in T_h. \tag{6}$$

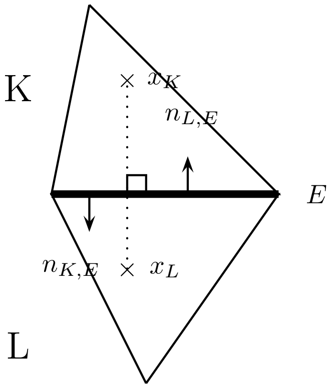


FIG. 1. *The standard orthogonality condition.*

If the mesh T_h is admissible in the sense of [10, Def. 9.1], i.e., satisfies standard orthogonality conditions (see Figure 1), then the numerical diffusion flux is defined by

$$\begin{aligned} F_{K,E}(u_h) &:= \frac{|E|(u_L - u_K)}{d(x_K, x_L)} \quad \text{if } E = \overline{K} \cap \overline{L}, \\ F_{K,E}(u_h) &:= -\frac{|E| u_K}{d(x_K, \Gamma)} \quad \text{if } E \subset \overline{K} \cap \partial\Omega. \end{aligned} \tag{7}$$

For general meshes, a possible choice for $F_{K,E}(u_h)$ is proposed in [5, 6] using the diamond cell method.

From now on we suppose that system (6) is well defined. This is the case if the mesh T_h is admissible in the sense of [10] and if $F_{K,E}(u_h)$ is given by (7) (see, for instance, [10]); while for an arbitrary mesh and the choice of $F_{K,E}(u_h)$ from [5, 6], system (6) is well defined under some geometrical conditions on the mesh [5, 6].

3. Some analytic tools.

3.1. Bubble functions, extension operator, and inverse inequalities.

For our further analysis we require standard bubble functions and extension operators that satisfy certain properties recalled here for the sake of completeness.

We need two types of bubble functions, namely, b_K and b_E associated with an element K and an edge/face E, respectively. For a triangle or a tetrahedron K, denoting by $\lambda_{a_i^K}$, $i = 1, \ldots, d+1$, the barycentric coordinates of K and by $a_i^E, i = 1, \ldots, d$, the vertices of the edge/face $E \subset \partial K$, we recall that $b_K = (d+1)^{d+1}\prod_{i=1}^{d+1} \lambda_{a_i^K}$ and $b_E = d^d \prod_{i=1}^{d} \lambda_{a_i^E}$.

For a rectangle K we here enumerate its vertices in a clockwise sense. Denoting by $\lambda_{a_i^K}$, $i = 1, \dots, 4$, the "barycentric" coordinates of K, namely, $\lambda_{a_i^K}$ is the unique element in $\mathbb{Q}_1(K)$ such that $\lambda_{a_i^K}(a_j^K) = \delta_{i,j}$, then we recall that $b_K = 8\lambda_{a_1^K}\lambda_{a_3^K}$ and $b_E = 4\lambda_{a_1^K}(\lambda_{a_2^K} + \lambda_{a_3^K})$ if the endpoints of the edge E are a_1^K and a_2^K.

One recalls that $b_K = 0$ on ∂K, $b_E = 0$ on $\partial\omega_E$, $\|b_K\|_{\infty,K} = \|b_E\|_{\infty,\omega_E} = 1$.

In two dimensions for an edge $E \subset \partial K$ using temporarily the local coordinates system (x, y) such that E is included into the x-axis, then the extension $F_{\text{ext}}(v_E)$ of $v_E \in C(E)$ to K is defined by $F_{\text{ext}}(v_E)(x, y) = v_E(x)$. We proceed similarly in three dimensions.

Now we may recall the so-called inverse inequalities, whose proof uses classical scaling techniques and the fact that all norms are equivalent in a finite-dimensional space [26].

LEMMA 3.1 (inverse inequalities). *Let $K \in T_h$, $E \in E_K$, $v_K \in \mathbb{P}_{k_0}(K)$, and $v_E \in \mathbb{P}_{k_1}(E)$ for some nonnegative integers k_0 and k_1. Then there exist positive constants β_0, β_1 (resp., α_0, α_1, and α_2) depending on the form of K (triangle, rectangle, or tetrahedron), on the aspect ratio of the mesh T_h, and on the polynomial degree k_0 (resp., k_1) such that*

$$\|v_K b_K^{1/2}\|_K^2 \le \|v_K\|_K^2 \le \beta_0 \|v_K b_K^{1/2}\|_K^2, \tag{8}$$

$$\|\nabla(v_K b_K)\|_K^2 \le \beta_1 h_K^{-2} \|v_K\|_K^2, \tag{9}$$

$$\|v_E b_E^{1/2}\|_E^2 \le \|v_E\|_E^2 \le \alpha_0 \|v_E b_E^{1/2}\|_E^2, \tag{10}$$

$$\|F_{\text{ext}}(v_E) b_E\|_K^2 \le \alpha_1 h_K \|v_E\|_E^2, \tag{11}$$

$$\|\nabla(F_{\text{ext}}(v_E) b_E)\|_K^2 \le \alpha_2 h_K^{-1} \|v_E\|_E^2. \tag{12}$$

Remark 3.2. In the above lemma, if K is a square and $k_1 = 2$, then $\alpha_0 = \frac{10+\sqrt{30}}{4} \approx 1.967$, $\alpha_1 = \frac{8(6+\sqrt{21})}{315} \approx 0.269$, $\alpha_2 = \frac{8(56+\sqrt{881})}{105} \approx 6.528$. These numbers are obtained by reducing estimates (10)–(12) to an eigenvalue problem. Namely using the standard basis of $\mathbb{P}_2$, estimate (12) is equivalent to $(AX, X) \le \alpha_2(BX, X)$ for all $X \in \mathbb{R}^3$, where A and B are two explicit 3×3 matrices. Therefore, α_2 is the largest eigenvalue of the matrix $B^{-1/2} \cdot A \cdot B^{1/2}$, or equivalently the largest eigenvalue of the matrix $B^{-1} \cdot A$, since B is invertible. A direct calculation yields the value of α_2. The other estimates are proved in the same manner. □

3.2. Interpolation error estimates. Here we collect some standard interpolation error estimates but we specify as explicitly as possible the involved constants. As usual we start with the reference elements, which are the unit triangle $\hat{K}$ of vertices $(0, 0), (1, 0), (0, 1)$, the unit square $\hat{K}$ of vertices $(0, 0), (1, 0), (0, 1), (1, 1)$, or the unit tetrahedron $\hat{K}$ of vertices $(0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1)$.

LEMMA 3.3. *Let $\hat{E}$ be the edge/face of $\hat{K}$ included in the axis/plane $x_d = 0$. Then there exist two positive constants μ and α such that for all $v \in H^1(\hat{K})$, the following estimates hold:*

$$\|v - \mathcal{M}_{\hat{K}} v\|_{\hat{K}} \le \mu \|\nabla v\|_{\hat{K}}, \tag{13}$$

$$\|v - \mathcal{M}_{\hat{E}} v\|_{\hat{E}} \le \alpha \|\nabla v\|_{\hat{K}}. \tag{14}$$

If $\hat{K}$ is the reference square, then $\mu = \frac{1}{\pi}$ and $\alpha = \frac{1}{\sqrt{\pi \tanh \pi}} \approx 0.565244$. If $\hat{K}$ is the reference triangle, then $\mu = \frac{1}{\pi}$ and $\alpha = \frac{1}{\mu_1}$, where μ_1^2 is the first positive root of *the transcendental equation*

$$\sinh x + \tan x = 0. \tag{15}$$

This means that $\alpha \approx 0.730276$. *If* $\hat{K}$ *is the reference tetrahedron, then* $\mu \leq \frac{(2(11+4\sqrt{6}))^{1/4}}{\sqrt{3}\pi} \approx 0.466715$ *and* $\alpha \leq \frac{(2(11+4\sqrt{6}))^{1/4}}{\sqrt{\pi \tanh \pi}} \approx 1.43549$.

Proof. The two estimates are Poincaré-like inequalities and follow from the Bramble–Hilbert lemma. But this argument does not give an estimate for μ and α. Therefore, we argue as follows. For the first estimate, denote by λ_1^2 the first positive eigenvalue of the Laplace operator on $\hat{K}$ with Neumann boundary conditions. Then by the min-max principle, we know that $\lambda_1^2 = \min_{\substack{v \in H^1(\hat{K}) \\ v \neq 0, \mathcal{M}_{\hat{K}} v = 0}} \frac{\|\nabla v\|_{\hat{K}}^2}{\|v\|_{\hat{K}}^2}$. This identity is equivalent to

$$\lambda_1^2 \|v\|_{\hat{K}}^2 \leq \|\nabla v\|_{\hat{K}}^2 \quad \forall v \in H^1(\hat{K}) : \mathcal{M}_{\hat{K}} v = 0.$$

We then obtain (13) with $\mu = \lambda_1^{-1}$.

If $\hat{K}$ is the unit square, it is well known that $\lambda_1^2 = \pi^2$ and consequently $\mu = \frac{1}{\pi}$. If $\hat{K}$ is the reference triangle, we use the following extension operator from $\hat{K}$ to the unit square $(0,1)^2$, temporarily denoted by $\hat{S}$. Namely, for $v \in H^1(\hat{K})$, we define its extension Ev to $\hat{S}$ by

$$\begin{aligned} &Ev(y_1, y_2) = v(y_1, y_2) \quad \text{if } (y_1, y_2) \in \hat{K}, \\ &Ev(y_1, y_2) = v(1 - y_2, 1 - y_1) \quad \text{if } (y_1, y_2) \in \hat{S} \setminus \hat{K}. \end{aligned}$$

Note that $Ev \in H^1(\hat{S})$ and from $\|v\|_{\hat{K}}^2 = \|v\|_{\hat{S}}^2$, $\|\nabla v\|_{\hat{K}}^2 = \|\nabla v\|_{\hat{S}}^2$, we easily get

$$\min_{\substack{v \in H^1(\hat{K}) \\ v \neq 0, \mathcal{M}_{\hat{K}} v = 0}} \frac{\|\nabla v\|_{\hat{K}}^2}{\|v\|_{\hat{K}}^2} \geq \min_{\substack{v \in H^1(\hat{S}) \\ v \neq 0, \mathcal{M}_{\hat{S}} v = 0}} \frac{\|\nabla v\|_{\hat{S}}^2}{\|v\|_{\hat{S}}^2} = \pi^2.$$

On the other hand, one readily checks that the function $\psi(x, y) = \sqrt{2}(\cos(\pi x) + \cos(\pi(1 - y))$ is an eigenvector of the eigenvalue π^2 of the Laplace operator with Neumann boundary conditions in $\hat{K}$. Therefore, we actually have $\lambda_1^2 = \pi^2$.

We use a similar argument for the reference tetrahedron. Namely we use an extension operator from $\hat{K}$ to the standard reference pentahedron $\hat{P} = \hat{E} \times (0, 1)$. For that purpose denote by $\hat{K}_2$ and $\hat{K}_3$ the tetrahedra of vertices $(1,0,0), (0,1,0), (0,0,1), (1,0,1)$ and $(0,1,0), (0,0,1), (1,0,1), (0,1,1)$, respectively. We remark that $\hat{P} = \hat{K} \cup \hat{K}_2 \cup \hat{K}_3$, that $\hat{K}$ and $\hat{K}_2$ have a common face, and similarly that $\hat{K}_2$ and $\hat{K}_3$ have a common face. Note further that $|\hat{K}| = |\hat{K}_2| = |\hat{K}_3| = \frac{1}{6}$. Therefore, as before there exists an affine transformation F_1 which maps $\hat{K}$ onto $\hat{K}_2$ and let their common face be invariant. Similarly denote by F_2 the affine transformation which maps $\hat{K}_2$ onto $\hat{K}_3$ and let their common face be invariant. Denote by $A_i, i = 1, 2$, the 3×3 matrices and by $b_i \in \mathbb{R}^3, i = 1, 2$, such that $F_i(x) = A_i x + b_i$ for all $x \in \mathbb{R}^3$. Now we are able to define the extension operator E: for $v \in H^1(\hat{K})$, we define

$$\begin{aligned} &Ev(y) = v(y), \quad \text{if } y \in \hat{K}, \qquad Ev(y) = v(F_1^{-1}(y)) \quad \text{if } y \in \hat{K}_2, \\ &Ev(y) = v(F_1^{-1}(F_2^{-1}(y))) \quad \text{if } y \in \hat{K}_3. \end{aligned}$$

Using the above properties between the tetrahedra $\hat{K}, \hat{K}_2$, and $\hat{K}_3$ and some changes of variables, we readily check that $Ev \in H^1(\hat{P})$ and satisfies $\int_{\hat{P}} Ev(y)\, dy = 3 \int_{\hat{K}} v(x)\, dx$, and

$$\int_{\hat{P}} |Ev(y)|^2\, dy = 3 \int_{\hat{K}} |v(x)|^2\, dx, \int_{\hat{P}} |\nabla Ev(y)|^2\, dy = \int_{\hat{K}} \nabla v(x)^\top \cdot T \cdot \nabla v(x)\, dx,$$

where the matrix T is given by

$$T = Id + (A_1^\top A_1)^{-1} + A_1^{-1}(A_2^\top A_2)^{-1}A_1^{-\top} = \begin{pmatrix} 5 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 5 \end{pmatrix}.$$

An easy calculation yields $\|T\|_2 = \sqrt{2(11+4\sqrt{6})} \approx 6.44949$.

These identities directly lead to

$$\min_{\substack{v\in H^1(\hat{K})\\ v\neq 0, \mathcal{M}_{\hat{K}}v=0}} \frac{\|\nabla v\|^2_{\hat{K}}}{\|v\|^2_{\hat{K}}} \geq \frac{3}{\|T\|_2} \min_{\substack{v\in H^1(\hat{P})\\ v\neq 0, \mathcal{M}_{\hat{P}}v=0}} \frac{\|\nabla v\|^2_{\hat{P}}}{\|v\|^2_{\hat{P}}} = \frac{3\pi^2}{\|T\|_2}.$$

We then conclude that $\mu \leq \sqrt{\frac{\|T\|_2}{3\pi^2}}$.

For the second estimate, we start with the following nonstandard eigenvalue problem in the unit square $\hat{K} = (0,1)^2$. Find λ^2 and $v \in H^1(\hat{K})$ solution of

$$\int_{\hat{K}} \nabla v \cdot \nabla w = \lambda^2 \int_{\hat{E}} vw \quad \forall w \in H^1(\hat{K}). \tag{16}$$

For this eigenvalue problem, let us show that the min-max principle holds at least for the first positive eigenvalue $\tilde{\lambda}_1^2$. Namely $\tilde{\lambda}_1^2$ is characterized by

$$\tilde{\lambda}_1^2 = \min_{\substack{v\in H^1(\hat{K})\\ \|v\|_{\hat{E}}\neq 0, \mathcal{M}_{\hat{E}}v=0}} \frac{\|\nabla v\|^2_{\hat{K}}}{\|v\|^2_{\hat{E}}}. \tag{17}$$

Denote by m the above right-hand side. Consider a minimizing sequence v_n of the above minimum, namely, for all $n \in \mathbb{N}$, let $v_n \in H^1(\hat{K})$ be such that

$$\mathcal{M}_{\hat{E}}v_n = 0, \qquad \|v_n\|_{\hat{E}} = 1, \qquad \|\nabla v_n\|^2_{\hat{K}} \to m \text{ as } n \to \infty.$$

Since $\|\nabla v\|_{\hat{K}} + \|v\|_{\hat{E}}$ is a norm on $H^1(\hat{K})$ equivalent to the standard norm, the sequence $(v_n)_n$ is bounded in $H^1(\hat{K})$. Therefore, there exists a subsequence, still denoted by $(v_n)_n$, such that

$$v_n \to v \text{ in } H^1(\hat{K}) \text{ as } n \to \infty.$$

From the above properties of the sequence $(v_n)_n$, we deduce that v satisfies

$$\mathcal{M}_{\hat{E}}v = 0, \qquad \|v\|_{\hat{E}} = 1, \qquad \|\nabla v\|^2_{\hat{K}} = m.$$

It remains to show that v is an eigenvector of problem (16) corresponding to the eigenvalue m. For that purpose let us fix $z \in H^1(\hat{K})$ such that

$$\mathcal{M}_{\hat{E}}z = 0 \quad \text{and} \quad \int_{\hat{E}} vz = 0. \tag{18}$$

Consider the mapping

$$\Phi : \mathbb{R} \to \mathbb{R} : \alpha \to \frac{\|\nabla(v+\alpha z)\|^2_{\hat{K}}}{1+\alpha^2\|z\|^2_{\hat{K}}}.$$

From the above minimization problem and the properties of v, the mapping Φ hits its minimum at $\alpha = 0$. Since Φ is smooth, we deduce that $\Phi'(0) = 0$, or

$$\int_{\hat{K}} \nabla v \cdot \nabla z = 0. \tag{19}$$

Since any $w \in H^1(\hat{K})$ such that $\mathcal{M}_{\hat{E}} w = 0$ may be written in the form $w = \beta v + z$, with $\beta \in \mathbb{R}$ and $z \in H^1(\hat{K})$ satisfying (18), we deduce that

$$\int_{\hat{K}} \nabla v \cdot \nabla w = \beta \int_{\hat{K}} \nabla v \cdot \nabla v + \int_{\hat{K}} \nabla v \cdot \nabla z.$$

By the properties of v and identity (19), we conclude that

$$\int_{\hat{K}} \nabla v \cdot \nabla w = \beta m = m \int_{\hat{E}} vw;$$

this last identity follows from (18).

To find the first eigenvalue of problem (16), we remark that its strong form is

$$\begin{cases} -\Delta v = 0 \text{ in } \hat{K}, \\ \frac{\partial v}{\partial n} \;=\; \lambda^2 v \text{ on } \hat{E}, \\ \frac{\partial v}{\partial n} \;=\; 0 \text{ on } \partial\hat{K} \setminus \hat{E}. \end{cases}$$

Using the standard argument of separation of variables, one finds a family of eigenvalues, its smallest one being $\pi \tanh \pi \approx 3.12988$. In order to be sure that this value is the smallest eigenvalue of problem (16), we penalize it by an integral term in $\hat{K}$. Namely, for any $\epsilon > 0$, we consider the problem

$$\int_{\hat{K}} \nabla v \cdot \nabla w = \lambda_\epsilon^2 \left(\int_{\hat{E}} vw + \epsilon \int_{\hat{K}} vw \right) \quad \forall w \in H^1(\hat{K}). \tag{20}$$

This problem is an eigenvalue problem related to a selfadjoint nonnegative operator. For that problem one can find all the eigenvalues by separation of variables. Since the eigenvalues of (20) depend continously on ϵ, the first positive eigenvalue $\tilde{\lambda}_{1,\epsilon}^2$ tends to the first eigenvalue of problem (16). By direct calculations one shows that

$$\tilde{\lambda}_{1,\epsilon}^2 \to \pi \tanh \pi.$$

We therefore conclude that $\tilde{\lambda}_1^2 = \pi \tanh \pi$. By the above "min-max" principle (17), we deduce that $\alpha = 1/\tilde{\lambda}_1$.

For the unit triangle, we start with the minization problem (17) (as before $\tilde{\lambda}_1^2$ is the first positive eigenvalue of problem (16)). Using the extension operator E from $\hat{K}$ to $\hat{S}$, we deduce that

$$\tilde{\lambda}_1^2 = \min_{\substack{v \in H^1(\hat{K}) \\ \|v\|_{\hat{E}} \neq 0, \mathcal{M}_{\hat{E}} v = 0}} \frac{\|\nabla v\|_{\hat{K}}^2}{\|v\|_{\hat{E}}^2} \geq \min_{\substack{v \in H^1(\hat{S}) \\ \|v\|_{\hat{E}} + \|v\|_{\hat{F}} \neq 0, \mathcal{M}_{\hat{E} \cup \hat{F}} v = 0}} \frac{\|\nabla v\|_{\hat{S}}^2}{\|v\|_{\hat{E}}^2 + \|v\|_{\hat{F}}^2},$$

where $\hat{F}$ is the edge of $\hat{S}$ included into the line $x_1 = 1$. The right-hand side is related to the eigenvalue problem: find λ^2 and $v \in H^1(\hat{S})$ solution of

$$\int_{\hat{S}} \nabla v \cdot \nabla w = \lambda^2 \int_{\hat{E} \cup \hat{F}} vw \quad \forall w \in H^1(\hat{S}). \tag{21}$$

The same arguments as before give, as first positive eigenvalue μ_1^2, the first positive root of the transcendental equation (15).

As for the first estimate we deduce that μ_1^2 is the first positive eigenvalue of problem (16). Indeed, if $w(x_1, x_2)$ is the eigenvector of problem (21) associated with the eigenvalue μ_1^2, then one readily checks that $v(x_1, x_2) = w(x_1, x_2) - w(1-x_2, 1-x_1)$ is an eigenvector of problem (16) associated with the eigenvalue μ_1^2.

As before the situation is not so convenient for the unit tetrahedron Therefore, we first state the following estimate on the reference prism $\hat{P}$:

$$\|v - \mathcal{M}_{\hat{E}} v\|_{\hat{E}} \le \frac{1}{\sqrt{\pi \tanh \pi}} \|\nabla v\|_{\hat{P}} \quad \forall v \in H^1(\hat{P}),$$

obtained as for the unit square. Now, using the extension operator E and this estimate, we may write

$$\|v - \mathcal{M}_{\hat{E}} v\|_{\hat{E}} = \|Ev - \mathcal{M}_{\hat{E}} Ev\|_{\hat{E}} \le \frac{1}{\sqrt{\pi \tanh \pi}} \|\nabla Ev\|_{\hat{P}} \le \frac{1}{\sqrt{\pi \tanh \pi}} \sqrt{\|T\|_2} \|\nabla v\|_{\hat{K}};$$

this last estimate follows from the above properties of Ev. □

Remark 3.4. To our knowledge, the exact value of μ is not explicitly known for the unit tetrahedron. Numerical tests give for λ_1^2 the approximated value $\lambda_1^2 \approx 14.444208445$. This gives for μ the approximated value $\mu \approx 0.26312$, which is relatively smaller than our theoretical upper bound. Similarly the exact value of α is not explicitly known for the unit tetrahedron; an approximated value is 0.340355, and therefore our theoretical upper bound is far from being optimal.

In the above arguments, the main difference between the unit triangle and the unit tetrahedron concerns the extension operator. For the triangle, the extension operator uses an orthogonal transformation, which is impossible for the unit tetrahedron. That last case still requires more investigations. □

The above lemma and scaling arguments lead to the following lemma.

Lemma 3.5. *There exist two positive constants μ and α depending on $\hat{K}$ such that for all $K \in T_h$ and $v \in H^1(K)$, the following estimates hold:*

$$\|v - \mathcal{M}_K v\|_K \le \mu \hat{\rho}^{-1} h_K \|\nabla v\|_K, \tag{22}$$

$$\|v - \mathcal{M}_E v\|_E \le \alpha \hat{\rho}^{-1} h_{E,K}^{-1/2} h_K \|\nabla v\|_K, \tag{23}$$

where E is an edge/face of K, and $\hat{\rho}$ is the diameter of the inscribed ball of $\hat{K}$.

4. Some finite elements of Morley type. As already mentioned the main idea of our a posteriori error analysis is to use an interpolant p satisfying

$$\int_E \frac{\partial p}{\partial n_{K,E}} \, ds = F_{K,E}(u_h) \quad \forall E \in E_K.$$

This means that we need to use finite elements having as degrees of freedom the mean of the normal derivative of p on each edge/face. The simplest element is the so-called Morley triangle [18, 4] usually used for the approximation of the plate problem. For our further uses we extend this kind of elements to rectangles and tetrahedra. We start by recalling the Morley triangle as well as a recent extension due to Nilssen, Tai, and Winther [21] and then introduce our new elements.

4.1. Triangles. Here K is a (nondegenerate) triangle with vertices a_i^K, $i = 1, 2, N_f := 3$.

The standard Morley triangle is defined by the triple (K, P_K, Σ_K) [18, 4], where $P_K = \mathbb{P}_2(K)$ and

$$\Sigma_K = \{p(a_i)\}_{i=1,\ldots,N_f} \cup \left\{ \int_E \frac{\partial p}{\partial n_{K,E}}\, ds \right\}_{E \in E_K}. \tag{24}$$

Note that this element is not a C^0-element; an extension which has this property was recently built in [21, sect. 4], where they take

$$\begin{aligned} P_K &= \mathbb{P}_2(K) \oplus \mathbb{P}_1(K) b_K = \{q + p b_K : q \in \mathbb{P}_2(K), p \in \mathbb{P}_1(K)\}, \\ \Sigma_K &= \{p(a_i^K)\}_{i=1,2,3} \cup \{p(m_E)\}_{E \in E_K} \cup \left\{ \int_E \frac{\partial p}{\partial n_{K,E}}\, ds \right\}_{E \in E_K}. \end{aligned}$$

4.2. Rectangles. Here K is a (nondegenerate) rectangle of vertices a_i^K, $i = 1, \ldots, N_f := 4$.

The first element is defined by $P_K = \mathbb{P}_2(K) \oplus \operatorname{Span}\{x^3 - 3xy^2, y^3 - 3yx^2\}$ with degrees of freedom Σ_K defined by (24). We readily check that the triple (K, P_K, Σ_K) is a finite element. The above choice is motivated by the fact that $\Delta q \in \mathbb{R}\ \forall q \in P_K$, since $x^3 - 3xy^2$ and $y^3 - 3yx^2$ are the unique homogeneous polynomials of degree 3 which are harmonic.

The second example is to take $P_K = \mathbb{Q}_2(K)$ and $\Sigma_K := \{p(a_i^K)\}_{i=1,\ldots,5} \cup \{\int_E \frac{\partial p}{\partial n_{K,E}}\, ds\}_{E \in E_K}$, where a_5^K is the center of gravity of K.

4.3. Tetrahedra. Here K is a (nondegenerate) tetrahedron with vertices a_i^K, $i = 1, 2, 3, N_f := 4$.

Inspired from the second triangular example from [21] we choose $P_K = \mathbb{P}_1(K) \oplus \mathbb{P}_1(K) b_K = \{q + p b_K : p, q \in \mathbb{P}_1(K)\}$, and Σ_K defined by (24).

Similar to Lemma 4.1 of [21] (adapted to our setting) we can prove the following lemma.

Lemma 4.1. *The above triple (K, P_K, Σ_K) is a C^0-finite element.*

5. The Morley interpolant.

5.1. Definition. For any vertex a of the triangulation we fix $(w_K(a))_{K \in T_h : a \in K}$ suitable weights of interpolation around K. Since our analysis below is independent of their choice, we do not describe them. They may be obtained using a discrete projection of piecewise constant functions over affine functions on ω_a [5, 6], a standard technique to get a recovered gradient at the vertex a, leading further to the $\mathbb{P}_1$-exactness. Namely for any vertex a the weights $w_K(a)$ may be fixed such that $w(a) = \sum_{K \subset \omega_a} w_K(a) u_K$, where $w \in \mathbb{P}_1(\omega_a)$ is the discrete projection of u_h on $\mathbb{P}_1(\omega_a)$, i.e., $w \in \mathbb{P}_1(\omega_a)$ is the unique minimizer of

$$\sum_{K \subset \omega_a} |q(x_K) - u_K|^2, \quad q \in \mathbb{P}_1(\omega_a).$$

This choice implies that if u_h were $\mathbb{P}_1(\omega_a)$, then we would have $w = u_h$ in ω_a. For instance, if ω_a is made of four squares, then this choice yields $w_K(a) = 1/4$.

We now introduce the Morley finite element space

$$
\begin{aligned}
V_h := \Big\{ & v_h \in L^2(\Omega) : v_{h|K} \in P_K \ \forall K \in T_h, \\
& v_{h|K}(a_i^K) = v_{h|L}(a_j^L) \ \forall K, L \in T_h, i, j \in \{1, \dots, N_f\} : a_i^K = a_j^L, \\
& v_{h|K}(a_i^K) = 0 \ \forall K, L \in T_h, i \in \{1, \dots, N_f\} : a_i^K \in \Gamma, \\
& \int_E \frac{\partial v_{h|K}}{\partial n_E}\, ds = \int_E \frac{\partial v_{h|L}}{\partial n_E}\, ds \ \forall E \in E_h, K, L \in T_h : E = K \cap L \Big\}.
\end{aligned}
$$

Since V_h is not necessarily included into $H_0^1(\Omega)$, the space V_h is equipped with the norm $|\cdot|_{1,h} := (\sum_{K \in T_h} |\cdot|_{1,K}^2)^{1/2}$. Notice that V_h is indeed included into $H_0^1(\Omega)$ for the second-triangular example and for our three-dimensional (3D) example.

Definition 5.1. *For $u_h = (u_K)_{K \in T_h}$, we define its Morley interpolant $I_M u_h$ as the unique element v_h in V_h satisfying*

$$
v_{h|K}(a_i^K) = \sum_{L \in T_h : a_i^K \in L} w_L(a_i^K) u_L \quad \forall K \in T_h, \quad i \in \{1, \dots, N_f\} : a_i^K \in \Omega, \tag{25}
$$

$$
v_{h|K}(a_i^K) = 0 \quad \forall K \in T_h, \quad i \in \{1, \dots, N_f\} : a_i^K \in \Gamma, \tag{26}
$$

$$
\int_E \frac{\partial v_{h|K}}{\partial n_{K,E}}\, ds = F_{K,E}(u_h) \quad \forall E \in E_K, \quad K \in T_h. \tag{27}
$$

For the second triangular element we have to add the conditions

$$
\begin{aligned}
& v_{h|K}(m_E) = v_{h|L}(m_E) = \tfrac{1}{2}(u_K + u_L) \quad \forall E \in E_h, \quad K, L \in T_h : E = K \cap L, \\
& v_{h|K}(m_E) = 0 \quad \forall E \in E_h, \quad K \in T_h : E \subset K \cap \Gamma.
\end{aligned}
$$

Similarly for the first-rectangular element we must add $v_{h|K}(a_5^K) = u_K$ for all $K \in T_h$.

5.2. Some useful properties. We first prove a basic property of the Morley interpolant.

Lemma 5.2. *If u_h is solution of* (6)*, then $I_M u_h$ satisfies*

$$
\int_K \Delta(I_M u_h)\, dx = -\int_K f(x)\, dx \quad \forall K \in T_h. \tag{28}
$$

Proof. By Green's formula and property (27) satisfied by $I_M u_h$, we have

$$
\int_K \Delta(I_M u_h)\, dx = \sum_{E \in E_K} \int_E \frac{\partial(I_M u_h)}{\partial n_{K,E}}\, ds = \sum_{E \in E_K} F_{K,E}(u_h),
$$

and we conclude by (6). □

Now we prove some quasi-orthogonality relations that will be used for the upper error bound. We first define the gradient jump of $I_M u_h$ in the normal and tangential direction by

$$
\begin{aligned}
& J_{E,n}(u_h) = \left[\!\left[\tfrac{\partial}{\partial n_E}(I_M u_h) \right]\!\right]_E \quad \forall E \in E_h^{int}, \\
& J_{E,t}(u_h) = \begin{cases} \left[\!\left[\tfrac{\partial}{\partial t_E}(I_M u_h) \right]\!\right]_E \quad \forall E \in E_h & \text{for nonconforming 2D cases,} \\ 0 \quad \forall E \in E_h & \text{for conforming cases.} \end{cases}
\end{aligned}
$$

Lemma 5.3. *If u is a solution of* (2) *and u_h is a solution of* (6)*, then*

$$
\begin{aligned}
\sum_{K \in T_h} \int_K \nabla(u - I_M u_h) \cdot \nabla \chi\, dx = & \sum_{K \in T_h} \int_K (f + \Delta I_M u_h)(\chi - \mathcal{M}_K \chi)\, dx \\
& - \sum_{E \in E_h^{int}} \int_E J_{E,n}(u_h)(\chi - \mathcal{M}_E \chi)\, ds \quad \forall \chi \in H_0^1(\Omega).
\end{aligned} \tag{29}
$$

Proof. For brevity denote the left-hand side of (29) by $I_1(\chi)$. By (2) and Green's formula on each triangle K, and recalling that $\chi \in H_0^1(\Omega)$, we may write

$$\begin{aligned} I_1(\chi) &= \int_\Omega f\chi\,dx + \sum_{K\in T_h}\int_K \Delta(I_M u_h)\chi\,dx - \sum_{K\in T_h}\int_{\partial K}\frac{\partial(I_M u_h)}{\partial n_K}\chi\,ds \\ &= \sum_{K\in T_h}\int_K (f+\Delta(I_M u_h))\chi\,dx - \sum_{E\in E_h^{int}}\int_E J_{E,n}(u_h)\chi\,ds. \end{aligned}$$

Using identity (28), we arrive at

$$I_1(\chi) = \sum_{K\in T_h}\int_K (f+\Delta(I_M u_h))(\chi - \mathcal{M}_K\chi)\,dx - \sum_{E\in E_h^{int}}\int_E J_{E,n}(u_h)\chi\,ds.$$

The conclusion now follows from the fact that $\int_E J_{E,n}(u_h)\,ds = 0$, for all $E\in E_h^{int}$, due to (27) and the principle of conservation of flux, $F_{K,E}(u_h) = -F_{L,E}(u_h)$, if $E = K\cap L$, $K, L\in T_h$. □

COROLLARY 5.4. *Under the assumptions of Lemma* 5.3 *the next estimate holds*

$$\begin{aligned} |I_1(\chi)| \le \sqrt{2}\Big\{&\frac{\mu^2}{\hat\rho^2}\sum_{K\in T_h} h_K^2\|f+\Delta(I_M u_h)\|_K^2 \\ &+ \frac{\alpha^2 N_f}{4\hat\rho^2}\sum_{K\in T_h}\sum_{E\in E_h^{int}\cap E_K} h_{E,K}^{-1}h_K^2\|J_{E,n}(u_h)\|_E^2\Big\}^{1/2}|\chi|_{1,\Omega}. \end{aligned} \tag{30}$$

Proof. Identity (29) and Cauchy–Schwarz's inequality yield

$$\begin{aligned} |I_1(\chi)| &\le \sum_{K\in T_h}\|f+\Delta(I_M u_h)\|_K\|\chi-\mathcal{M}_K\chi\|_K + \sum_{E\in E_h^{int}}\|J_{E,n}(u_h)\|_E\|\chi-\mathcal{M}_E\chi\|_E \\ &\le \sum_{K\in T_h}\|f+\Delta(I_M u_h)\|_K\|\chi-\mathcal{M}_K\chi\|_K \\ &\quad + \frac{1}{2}\sum_{K\in T_h}\sum_{E\in E_h^{int}\cap E_K}\|J_{E,n}(u_h)\|_E\|\chi-\mathcal{M}_E\chi\|_E. \end{aligned}$$

By the interpolation error estimates (22) and (23), we obtain

$$|I_1(\chi)| \le \sum_{K\in T_h} h_K\left(\frac{\mu}{\hat\rho}\|f+\Delta(I_M u_h)\|_K + \frac{\alpha}{2\hat\rho}\sum_{E\in E_h^{int}\cap E_K} h_{E,K}^{-1/2}\|J_{E,n}(u_h)\|_E\right)|\chi|_{1,K}.$$

We conclude by the discrete Cauchy–Schwarz's inequality and the well-known estimate $(\sum_{i=1}^l a_i)^2 \le l\sum_{i=1}^l a_i^2$, valid for $l = 2,3,4$ and all real numbers a_i. □

LEMMA 5.5. *Assume that* $d = 2$. *If* u *is the solution of* (2) *and* u_h *is the solution of* (6), *then*

$$\sum_{K\in T_h}\int_K \nabla(u - I_M u_h)\cdot \operatorname{curl} g\,dx = \sum_{E\in E_h}\int_E J_{E,t}(u_h)(g - \mathcal{M}_E g)\,ds \quad \forall g\in H^1(\Omega), \tag{31}$$

where $\operatorname{curl} g = (\partial_2 g, -\partial_1 g)^\top$ *is the vectorial curl of* g.

Proof. Denote the left-hand side of (31) by $I_2(g)$. Green's formula on each element K leads to (see Theorem I.2.11 of [11])

$$I_2(g) = -\sum_{K\in T_h}\int_{\partial K}\frac{\partial}{\partial t}(u - I_M u_h)g\,ds = \sum_{E\in E_h}\int_E J_{E,t}(u_h)g\,ds,$$

since $u \in H_0^1(\Omega)$ and $g \in H^1(\Omega)$. The conclusion follows from the property

$$\int_E J_{E,t}(u_h)\,ds = 0. \tag{32}$$

Indeed, if $a_E^i, i = 1, 2$, are the two extremities of E, we have $\int_E J_{E,t}(u_h)\,ds = [\![u_h]\!]_E(a_E^1) - [\![u_h]\!]_E(a_E^2)$. Using properties (25) and (26), we have $[\![u_h]\!]_E(a_E^i) = 0$, $i = 1, 2$, and therefore (32) holds. □

COROLLARY 5.6. *Under the assumptions of Lemma* 5.5 *the following estimate holds:*

$$|I_2(g)| \leq \frac{\alpha}{\hat\rho}\sqrt{N_f}\left\{\sum_{K\in T_h}\sum_{E\in E_K} h_{E,K}^{-1}h_K^2\|J_{E,t}(u_h)\|_E^2\right\}^{1/2}|g|_{1,\Omega}. \tag{33}$$

Remark 5.7. The above fundamental properties are only based on the definition of the scheme (6), the continuity of the interpolant at the interior nodes, the property (26), and the interpolation property (27). Therefore, our further analysis works for any finite element (K, P_K, Σ_K) such that the associated interpolant satisfies these properties. But the finite element and the definition of the interpolant should be well chosen in order to guarantee the convergence of $I_M u_h$ to the exact solution u. That is the reason of the introduction of the weights $w_K(a)$ in (25) since it was shown in [5, 6] that for a triangulation made of rectangles, the choice of the weights described at the beginning of section 5.1 guarantees the convergence of u_h to u. Convergence analysis for arbitrary triangulations and appropriate weights is still to be done, but it is outside the scope of this paper. □

6. Error estimators.

6.1. Residual error estimators.

The exact element residual is defined by $R_K := f + \Delta I_M u_h$ on K. As usual this exact residual is replaced by some finite-dimensional approximation called approximate element residual $r_K \in \mathbb{P}_k(K)$. A realistic choice is to take $r_K = \mathcal{M}_K f + \Delta I_M u_h$ since in the case $\Delta I_M u_h \in \mathbb{R}$ we have (thanks to Lemma 5.2) $r_K = 0$.

DEFINITION 6.1 (residual error estimator). *The local and global residual error estimators and approximation terms are defined by*

$$\eta_K^2 := h_K^2\left(\|r_K\|_K^2 + \sum_{E\in E_K\cap E_h^{int}} h_{E,K}^{-1}\|J_{E,n}(u_h)\|_E^2 + \sum_{E\in E_K} h_{E,K}^{-1}\|J_{E,t}(u_h)\|_E^2\right),$$

$$\eta^2 := \sum_{K\in T_h}\eta_K^2,$$

$$\zeta_K^2 := \sum_{K'\subset\omega_K} h_{K'}^2\|R_{K'} - r_{K'}\|_{K'}^2, \qquad \zeta^2 := \sum_{K\in T_h}\zeta_K^2.$$

6.2. Upper error bound.

THEOREM 6.2. *Let u be a solution of* (2) *and let u_h be a solution of* (6). *Then the error $e := u - I_M u_h$ is bounded by*

$$|e|_{1,h} \leq c_{up}(\eta + \zeta), \tag{34}$$

where $c_{up} = \frac{1}{\hat{\rho}} \max\{2\mu, \alpha N_f^{1/2}\}$.

Proof. Denote by $\nabla_h e$ the brokent gradient of e, namely,

$$(\nabla_h e)_{|K} = \nabla e_{|K} \text{ on } K \quad \forall K \in T_h.$$

As in Theorem 3.1 of [7] we use its Helmholtz decomposition

$$\nabla_h e = \nabla \chi + \operatorname{curl} \psi, \tag{35}$$

with $\chi \in H_0^1(\Omega)$ and $\psi \in H^1(\Omega)$ if $d = 2$ and $\chi = e$ and $\psi = 0$ if $d = 3$ such that

$$|\chi|_{1,\Omega}^2 + |\psi|_{1,\Omega}^2 \leq |e|_{1,h}^2. \tag{36}$$

This estimate is direct in three dimensions, while in two dimensions it directly follows from the identity $\int_\Omega \nabla \chi \cdot \operatorname{curl} \psi = 0$, a consequence of Green's formula (recalling that $\chi = 0$ on the boundary).

Owing to identity (35) we may write (with the notation from Lemmas 5.3 and 5.5)

$$|e|_{1,h}^2 = \int_\Omega |\nabla_h e|^2 \, dx = \int_\Omega \nabla_h e \cdot (\nabla \chi + \operatorname{curl} \psi) \, dx = I_1(\chi) + I_2(\psi).$$

Using estimates (30) and (33), we obtain

$$|e|_{1,h}^2 \leq \sqrt{2} \left(\frac{\mu^2}{\hat{\rho}^2} \Xi^2 + \frac{\alpha^2 N_f}{4\hat{\rho}^2} \eta_n^2 \right)^{1/2} |\chi|_{1,\Omega} + \frac{\alpha N_f^{1/2}}{\hat{\rho}} \eta_t |\psi|_{1,\Omega},$$

where for brevity we set $\Xi^2 = \sum_{K \in T_h} h_K^2 \|f + \Delta(I_M u_h)\|_K^2$ and

$$\eta_n^2 = \sum_{K \in T_h} \sum_{E \in E_h^{int} \cap E_K} h_{E,K}^{-1} h_K^2 \|J_{E,n}(u_h)\|_E^2, \eta_t^2 = \sum_{K \in T_h} \sum_{E \in E_K} h_{E,K}^{-1} h_K^2 \|J_{E,t}(u_h)\|_E^2.$$

By the discrete Cauchy–Schwarz inequality and estimate (36), we obtain

$$\begin{aligned} |e|_{1,h}^2 \leq \frac{2\mu^2}{\hat{\rho}^2} \Xi^2 + \frac{\alpha^2 N_f}{2\hat{\rho}^2} \eta_n^2 + \frac{\alpha^2 N_f}{\hat{\rho}^2} \eta_t^2 \leq \frac{4\mu^2}{\hat{\rho}^2} \zeta^2 \\ + \frac{4\mu^2}{\hat{\rho}^2} \sum_{K \in T_h} h_K^2 \|r_K\|_K^2 + \frac{\alpha^2 N_f}{\hat{\rho}^2} (\eta_n^2 + \eta_t^2); \end{aligned}$$

this last estimate follows from the well-known estimate $(a+b)^2 \leq 2a^2 + 2b^2$, valid for all real numbers a, b. By the definition of c_{up} and η the above estimate implies that $|e|_{1,h} \leq c_{up}(\xi^2 + \eta^2)^{1/2} \leq c_{up}(\eta + \xi)$. □

Remark 6.3. Thanks to Lemma 3.3, we can estimate the constant c_{up} appearing in the above upper bound. For a triangulation made of rectangles, then $c_{up} = 2 \max\{\frac{1}{\pi}, \alpha\} = 2\alpha \approx 1.13049$. For a triangulation made of triangles, then $c_{up} = \frac{1}{\hat{\rho}} \max\{\frac{2}{\pi}, \sqrt{3}\alpha\} = \frac{\sqrt{3}\alpha}{\hat{\rho}} \approx 2.15928$. Finally for a mesh made of tetrahedra, one has $c_{up} \leq 2 \frac{(2(11+4\sqrt{6}))^{1/4}}{\sqrt{\pi \tanh \pi}} \approx 9.27912$. For that last case, by Remark 3.4, a numerical upper bound for c_{up} is 2.20009. In both cases, the exact value, or the numerical upper bound for the tetrahedral case, of c_{up} is quite close to unity. □

6.3. Lower error bound.

Theorem 6.4. *For all elements K, the following local lower error bound holds*

$$\eta_K \leq c_{low}(\|\nabla_h e\|_{\omega_K} + \zeta_K), \tag{37}$$

where $c_{low}^2 = \max\{2\beta_0^2\beta_1 + 2N_f\alpha_0^2\sigma_1^{-1}\sigma_2^2\sigma_3(3\alpha_2\sigma_4^{-1} + 8\alpha_1\beta_0^2\beta_1\sigma_1^{-1}\sigma_4), 2\beta_0^2 + 8N_f\alpha_0^2\alpha_1(1 + 2\beta_0^2)\sigma_1^{-1}\sigma_2^2\sigma_3\sigma_4\}$.

Proof.

Element residual. For a fixed element K denote $w_K = r_K b_K$ which belongs to $H_0^1(K)$. From the definition of R_K and integration by parts, we may write

$$\begin{aligned}\int_K r_K w_K &= \int_K (r_K - R_K)w_K - \int_K \Delta(u - I_M u_h)w_K \\ &= \int_K (r_K - R_K)w_K + \int_K \nabla e \cdot \nabla w_K.\end{aligned}$$

By Cauchy–Schwarz's inequality and the inverse inequalities (8) and (9), we conclude that

$$h_K\|r_K\|_K \;\leq\; \beta_0(\beta_1^{1/2}\|\nabla e\|_K + h_K\|r_K - R_K\|_K). \tag{38}$$

Normal jump. Fix an arbitrary $E \in E_h^{int}$. Recall that $J_{E,n}(u_h) \in \mathbb{P}_k(E)$ for some $k \in \mathbb{N}$ and set $w_E := F_{\text{ext}}(J_{E,n}(u_h))b_E \in H_0^1(\omega_E)$. By elementwise partial integration, we get

$$\begin{aligned}\int_E J_{E,n}(u_h)w_E = -\sum_{K\subset\omega_E}\int_{\partial K}\frac{\partial e}{\partial n_K}w_E &= -\sum_{K\subset\omega_E}\int_K(\nabla e\cdot\nabla w_E + \Delta e w_E)\,dx \\ &\leq \|\nabla_h e\|_{\omega_E}\,\|\nabla w_E\|_{\omega_E} \;+ \sum_{K\subset\omega_E}\|\Delta e\|_K\|w_E\|_K.\end{aligned}$$

Inequalities (10)–(12) and properties (3) and (4) in the previous estimate lead to

$$h_E\|J_{E,n}(u_h)\|_E^2 \;\leq\; 2\alpha_0^2\sigma_1^{-1}\left(2\alpha_2\sigma_4^{-1}\|\nabla_h e\|_{\omega_E}^2 + 2\alpha_1\sigma_4\sum_{K\subset\omega_E}h_K^2\|R_K\|_K^2\right).$$

By estimate (38) we arrive at

$$\begin{aligned}h_E\|J_{E,n}(u_h)\|_E^2 \leq{}& 4\alpha_0^2\sigma_1^{-1}(\alpha_2\sigma_4^{-1} + 4\alpha_1\sigma_4\beta_0^2\beta_1)\|\nabla_h e\|_{\omega_E}^2 \\ &+ 8\alpha_0^2\alpha_1\sigma_1^{-1}\sigma_4(1+2\beta_0^2)\sum_{K\subset\omega_E}h_K^2\|R_K - r_K\|_K^2.\end{aligned} \tag{39}$$

Tangential jump (in two dimensions). For a fixed edge E set $w_E := F_{\text{ext}}(J_{E,t}(u_h))$ $b_E \in H_0^1(\omega_E)$. For $u \in H^1(\omega_E)$ and $w_E \in H_0^1(\omega_E)$, partial integration leads to

$$0 = \int_{\partial\omega_E}\frac{\partial u}{\partial t}w_E \;=\; -\int_{\omega_E}\nabla u\cdot\operatorname{curl} w_E.$$

For $I_M u_h$ we integrate elementwise and obtain using the above identity

$$\begin{aligned}\int_E J_{E,t}(u_h)w_E = -\sum_{K\subset\omega_E}\int_{\partial K}\frac{\partial(I_M u_h)}{\partial t}w_E &= \sum_{K\subset\omega_E}\int_K\nabla(I_M u_h)\cdot\operatorname{curl} w_E \\ &= -\sum_{K\subset\omega_E}\int_K\nabla(u - I_M u_h)\cdot\operatorname{curl} w_E \leq \|\nabla_h e\|_{\omega_E}\,\|\nabla w_E\|_{\omega_E}.\end{aligned}$$

The inverse inequalities (10) and (12), as well as (3) and (4), lead to

$$h_E \|J_{E,t}(u_h)\|_E^2 \ \leq \ 2\alpha_0^2 \alpha_2 \sigma_1^{-1} \sigma_4^{-1} \|\nabla_h e\|_{\omega_E}^2. \tag{40}$$

The conclusion follows from estimates (38)–(40) and properties (3) and (4). ∎

Remark 6.5. In the above proof, we see that if $r_K = 0$, then the constant c_{low} reduces to $c_{low}^2 = 2N_f \alpha_0^2 \sigma_1^{-1} \sigma_2^2 \sigma_3 \max\{3\alpha_2 \sigma_4^{-1}, 2\alpha_1 \sigma_4\}$. Let us illustrate this constant c_{low} in the particular case considered in the next section. Take a triangulation made of squares, build the Morley interpolant with the help of the first rectangular element from section 4, and choose $r_K = 0$. Then by Remark 3.2, we have $c_{low} = \sqrt{24\alpha_0^2 \alpha_2} \approx 24.622$. ∎

7. Diffusion–convection–reaction equations. In this section we describe how to extend the above results to diffusion–convection–reaction equations; for details we refer the reader to [20].

Consider the linearized diffusion–convection–reaction problem: for $f \in L^2(\Omega)$ let $u \in H_0^1(\Omega)$ be the unique solution of

$$Au := \operatorname{div}(-\epsilon \nabla u + \mathbf{v} u) + bu = f \text{ in } \Omega, \tag{41}$$

where ϵ is a fixed positive constant, $\mathbf{v} \in \mathbb{R}^d$, and b is a nonnegative real number.

Integrating (41) on a control volume K and using the divergence formula, we obtain

$$\sum_{E \in E_K} \int_E (-\epsilon \nabla u + \mathbf{v} u) \cdot n_{K,E} \, ds + \int_K bu \, dx = \int_K f(x) \, dx \quad \forall K \in T_h.$$

The continuous diffusion flux $-\epsilon \nabla u \cdot n_{K,E}$ is approximated as before, the convection flux $\mathbf{v} u \cdot n_{K,E}$ by a first order upwind scheme, and the reaction term $\int_K u$ by a simple quadrature formula (see [10]). These approximations lead to the following system. Find $u_h := (u_K)_{K \in T_h}$, the solution of

$$\sum_{E \in E_K} \left(-\epsilon F_{K,E}(u_h) + v_{K,E} F_E^C(u_h)\right) + bF_K^R(u_h) = \int_K f(x) \, dx \quad \forall K \in T_h, \tag{42}$$

where $v_{K,E} = \mathbf{v} \cdot n_{K,E}$, the quantity $F_{K,E}(u_h)$ is supposed to satisfy the principle of conservation of flux, while the quantities $F_E^C(u_h)$ and $F_K^R(u_h)$ are, respectively, defined by

$$F_E^C(u_h) := |E| u_{E,+}, \tag{43}$$

where for $E \in E_h^{int}$, $u_{E,+} = u_{K_{E,+}}$, $K_{E,+}$ being the upstream control volume, i.e., $v_{K_{E,+},E} \geq 0$; while for $E \in \bar{K} \cap \Gamma$, $u_{E,+} = u_K$ if $v_{K,E} \geq 0$, and $u_{E,+} = 0$ else. $F_K^R(u_h) = |K| u_K$.

For a restricted admissible mesh in the sense of [10, Def. 9.4], if the numerical diffusion fluxes $F_{K,E}(u_h)$ are given by (7), then system (42) is well defined as proved in [9]. For a general mesh as here, we simply assume that system (42) has a unique solution.

As for the Laplace operator, we associate with the finite volume solution u_h its Morley interpolant $I_M u_h$. This interpolant is related to the quantities involved in (42), namely, the diffusion and convection fluxes, and the reaction term. For that purpose, for each element K, we build a C^0-finite element (K, P_K, Σ_K) having as degrees of freedom the mean of the normal derivative and of the function on each edge, as well as the mean on K. For instance, if K is a triangle, we may take

$$P_K = \{q + (p + \alpha b_K)b_K : q \in \mathbb{P}_2(K), p \in \mathbb{P}_1(K), \alpha \in \mathbb{R}\},$$
$$\Sigma_K = \left\{p(a_i^K)\right\}_{i=1,2,3} \cup \left\{\int_K p\right\} \cup \left\{\int_E p\,ds\right\}_{E\in E_K} \cup \left\{\int_E \frac{\partial p}{\partial n_{K,E}}\,ds\right\}_{E\in E_K}.$$

Then for $u_h = (u_K)_{K\in T_h}$, we define its interpolant $I_M u_h$ as the unique element v_h in $V_h \cap H_0^1(\Omega)$ satisfying (25)–(27) and

$$\int_E v_h\,ds = F_E^C(u_h) \quad \forall E \in E_h^{int}, \qquad \int_K v_h\,dx = F_K^R(u_h) \quad \forall K \in T_h.$$

The key point of our a posteriori analysis is the following basic property of the Morley interpolant, obtained using Green's formula and the above properties of $I_M u_h$.

LEMMA 7.1. *If u_h is a solution of* (42), *then $I_M u_h$ satisfies*

$$\int_K (A(I_M u_h) - f)\,dx = 0 \quad \forall K \in T_h : \text{meas}_{d-1}(K \cap \Gamma) = 0.$$

This property and similar arguments to those used before allow us to prove the following error bounds.

THEOREM 7.2. *Let u be a solution of* (41), *and let u_h be a solution of* (42). *Then the error is bounded by*

$$(44)\qquad \int_\Omega \left(\epsilon|\nabla e|^2 + \left(\frac{1}{2}\text{div}\,\mathbf{v} + b\right)|e|^2\right)^{1/2} \le c_1(\eta + \zeta),$$

where c_1 is a positive constant depending on the aspect ratio of the mesh and of the size of ϵ.

For all elements K, the following local lower error bound holds:

$$(45)\qquad \eta_K \le c_2\left(\left(\int_{\omega_K} \left(\epsilon|\nabla e|^2 + \left(\frac{1}{2}\text{div}\,\mathbf{v} + b\right)|e|^2\right)\right)^{1/2} + \zeta_K\right),$$

where c_2 is a positive constant depending on the aspect ratio of the mesh and of the size of ϵ.

Remark 7.3. 1. By modifying appropriately the estimator η_K, we may skip the dependence of c_1 with respect to ϵ and give explicitly the dependence of c_2 on this parameter; see [20].

2. In the case of a large Peclet number $Pe \equiv \epsilon^{-1}|\mathbf{v}|$ and/or large number $\Gamma \equiv \epsilon^{-1}b$, problem (41) is singularly perturbed and the solution may generate sharp boundary or interior layers, where the solution of the limit problem (corresponding to $\epsilon = 0$) is not smooth or does not satisfy the Dirichlet boundary condition. In that case, the use of anisotropic meshes is recommended. This will be addressed in [20]. □

8. Numerical results. In this section we present two numerical tests that illustrate the efficiency and reliability of our estimator. The second example further indicates that our estimator is appropriate for adaptivity. Additionally, for both examples we provide the order of convergence of the error $|e|_{1,h}$; both cases confirm that $I_M u_h$ is a good approximation of u.

8.1. A smooth solution. The first example is for a smooth solution in the unit square $]0,1[^2$ and quasi-uniform meshes made of squares. Namely, we consider problem (1) in $]0,1[^2$ with the following prescribed exact solution $u(x,y) = xy(1-x)(1-y)$. The meshes are uniform ones made of squares of size $h = 1/n$ for $n = 4, 8, \dots, 256$ obtained by dividing each segment in n subintervals. Since the meshes are made of squares we use the scheme (6) with the numerical diffusion flux given by (7); furthermore, the Morley interpolant is built using the first-rectangular element from subsection 4.2 and the weights $w_K(a) = 1/4$.

We first investigate the order of convergence of the approximated solution u_h as well as its interpolant $I_M u_h$ to the exact solution u in different norms. Namely, we present in Figure 2 the following norms: $\|\bar{u} - u_h\|$ (where $\bar{u}$ is piecewise constant on T_h and is equal to $u(x_K)$ on each K); $|u - u_h|_{1,T_h}$ (where the mesh-depending norm $|\cdot|_{1,T_h}$ is defined in [10, Def. 9.3]); $\|u - I_M u_h\|$; and $|u - I_M u_h|_{1,h}$. These quantities are illustrated in Figure 2 by lines 1–4, respectively, in a double logarithmic scale so that the slope of the curve corresponds to the order of convergence. Theorem 9.3 of [10] yields the order of convergence 1 for $|u - u_h|_{1,T_h}$. Figure 2 even reveals a better order of convergence of about 1.5, probably due to the smoothness of u. For the L^2-norms, we remark a quadratic order of convergence, a usual phenomenon. On the other hand, for the discrete H^1-norm of the reconstructed approximation, we also see a quadratic order of convergence. This seems to be a superconvergence effect, probably due to the smoothness of the solution and of the use of structured meshes.

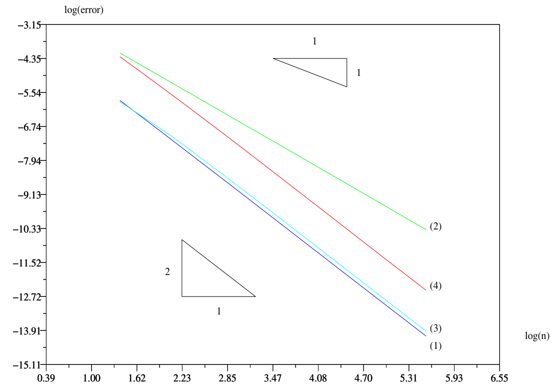


FIG. 2. *Illustration of different norms for test* 1.

Now we investigate the main theoretical results which are the upper and lower error bounds (34) and (37). In order to present them appropriately, we consider the ratios

$$q_{up} := \frac{|u - I_M u_h|_{1,h}}{\eta + \xi} \text{ as a function of } |\log n|,$$

$$q_{low} := \max_{K \in T_h} \frac{\eta_K}{\|\nabla(u - I_M u_h)\|_{\omega_K} + \xi_K} \text{ as a function of } |\log n|.$$

The first ratio q_{up} is frequently referred to as the effectivity index. It measures the reliability of the estimator and is related to the global upper error bound. The second ratio is related to the local lower error bound and measures the efficiency of the estimator. From our theoretical considerations, both ratios should be bounded from above which is confirmed experimentally as shown in Figure 3. Hence our estimator is reliable and efficient.

In Figure 4 we compare the discrete H^1 seminorm $|u - I_M u_h|_{1,h}$ and the global error estimator η with respect to n. We remark that the orders of convergence are the same (namely, 2). This figure further confirms the equivalence between $|u - I_M u_h|_{1,h}$ and η. All related quantities are summarized in Table 1.

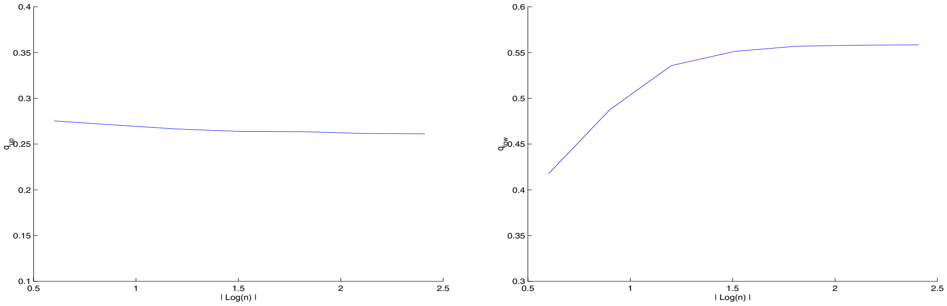


FIG. 3. q_{up} *(left) and* q_{low} *(right) in dependence of* $|\log n|$ *for test* 1.

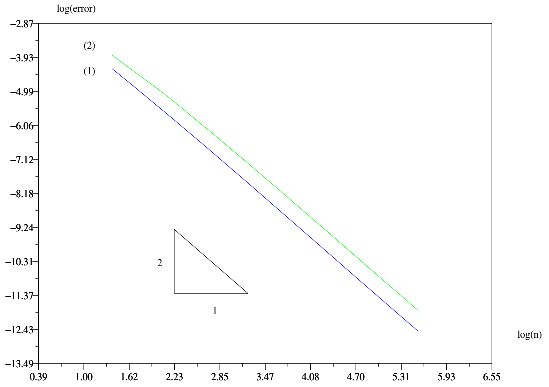


FIG. 4. *Comparison between* $-\ln |u - I_M u_h|_{1,h}$ *(line* (1)*) and* $-\ln \eta$ *(line* (2)*) with respect to* $\ln n$ *for test* 1.

8.2. A nonsmooth solution. For the second example we use the L-shaped domain $\Omega :=]-1,1[^2 \setminus]0,1[\times]-1,0[$ with the exact singular solution given in polar coordinates by $u = r^{2/3} \sin\left(\frac{2\theta}{3}\right)$ considered as a solution of the Dirichlet problem with nonhomogeneous Dirichlet boundary conditions. As before the domain is discretized using uniform meshes made of squares of size $h = 1/n$ for $n = 4, 8, \ldots, 256$. Since u presents singular behavior near the origin and uniform meshes are used, we have a reduction of the order of convergence from 1 to 2/3 for the norm $|u - u_h|_{1,h}$ (see Theorem 2.4 of [8] and Figure 5). From Figure 5 we notice the same phenomenon of reduction of the order of convergence for the other norms.

TABLE 1
Data for test 1.

n	$\Vert\bar{u}-u_h\Vert_{0,\Omega}$	$\vert u-u_h\vert_{1,T_h}$	$\Vert u-I_M u_h\Vert_{0,\Omega}$	$\vert u-I_M u_h\vert_{1,h}$	η	q_{low}	q_{up}
4	2.96e−03	1.57e−02	2.84e−03	1.37e−02	2.09e−02	0.4174	0.2753
8	7.57e−04	5.58e−03	8.58e−04	3.72e−03	6.43e−03	0.4875	0.2708
16	1.91e−04	1.99e−03	2.25e−04	9.53e−04	1.74e−03	0.5357	0.2665
32	4.77e−05	7.09e−04	5.68e−05	2.40e−04	4.48e−04	0.5513	0.2638
64	1.19e−05	2.51e−04	1.42e−05	6.00e−05	1.14e−04	0.5569	0.2636
128	2.99e−06	8.90e−05	3.56e−06	1.50e−05	2.86e−05	0.5580	0.2616
256	7.47e−07	3.15e−05	8.90e−07	3.75e−06	7.17e−06	0.5584	0.2612

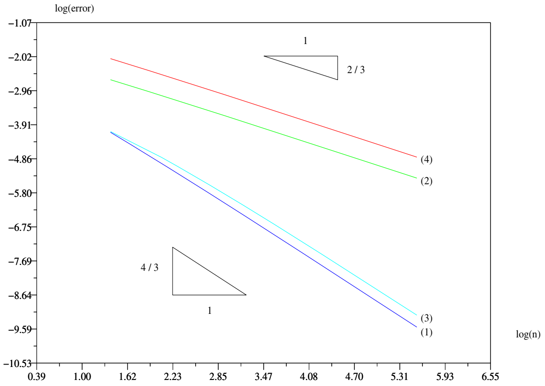


FIG. 5. *Illustration of different norms for test* 2.

TABLE 2
Data for tests 2.

n	$\Vert\bar{u}-u_h\Vert_{0,\Omega}$	$\vert u-u_h\vert_{1,T_h}$	$\Vert u-I_M u_h\Vert_{0,\Omega}$	$\vert u-I_M u_h\vert_{1,h}$	η	q_{low}	q_{up}
4	1.63e−02	7.02e−02	1.66e−02	1.26e−01	4.26e−01	2.4949	0.2958
8	6.86e−03	4.52e−02	8.11e−03	8.11e−02	2.70e−01	2.4661	0.3003
16	2.81e−03	2.87e−02	3.59e−03	5.15e−02	1.71e−01	2.4527	0.3022
32	1.14e−03	1.82e−02	1.51e−03	3.26e−02	1.07e−01	2.4471	0.3030
64	4.56e−04	1.15e−02	6.21e−04	2.06e−02	6.78e−02	2.4448	0.3033
128	1.82e−04	7.23e−03	2.52e−04	1.30e−02	4.27e−02	2.4438	0.3035
256	7.25e−05	4.55e−03	1.01e−04	8.16e−03	2.69e−02	2.4435	0.3034

Again we have tested the rate of convergence of $\Vert\bar{u}-u_h\Vert$, $|u-u_h|_{1,T_h}$, $\Vert u-I_M u_h\Vert$, and $|u-I_M u_h|_{1,h}$. Here we notice that both discrete H^1-norms have a rate of convergence of $2/3$ and that the rate of convergence of the L^2-norms is twice, namely, $4/3$.

As before we further check the boundedness of the ratios q_{up} and q_{low}. These quantities are given in Table 2 and illustrated in Figures 6 and 7. From these figures we can draw the same conclusion as before, namely, the efficiency and reliability of our estimator.

From Tables 1 and 2, we see that the experimental bounds for q_{up} and q_{low} are quite smaller than the theoretical ones. This is quite realistic since the experimental

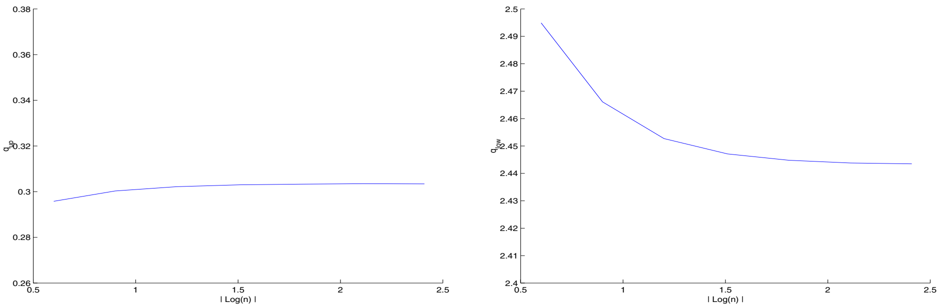


FIG. 6. *q_{up} (left) and q_{low} (right) in dependence of $|\log n|$ for test* 2.

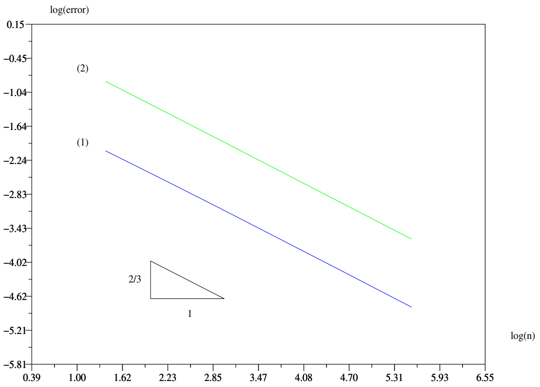


FIG. 7. *Comparison between $-\ln|u - I_M u_h|_{1,h}$ (line* (1)*) and $-\ln\eta$ (line* (2)*) with respect to $\ln n$ for test* 2.

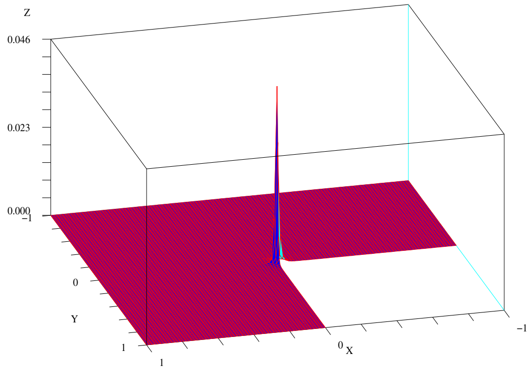


FIG. 8. *Distribution of the local estimator for test* 2 *and $n = 64$.*

values depend on the chosen solution, while the theoretical analysis always considers the worse case.

Finally, in Figure 8 we give the distribution of the local residual error estimators for our second example with the mesh corresponding to $n = 64$. From this figure we may conclude that our estimator is appropriate for adaptivity, since it detects the region of large errors, namely, the neighborhood of the origin.

9. Conclusions.

We have proposed and rigorously analyzed a new a posteriori error estimator for a cell-centered finite volume method that is reliable and efficient. This estimator is based on the construction of an appropriate interpolant of Morley type and the use of a Helmholtz decomposition of the error. The size of the constants appearing in the error estimates has been given as explicitly as possible, as a function of the aspect ratio of the mesh and of the form of the elements (triangles, rectangles, or tetrahedra). Some numerical experiments confirm our theoretical predictions and show that our estimator is appropriate for adaptivity.

The extension of our method to diffusion–convection–reaction equations is briefly described; the details are postponed to a forthcoming paper.

Adaptive algorithms are not considered here since they require more investigations. They will be considered in forthcoming works.

Acknowledgments. I am very grateful to Karim Djadel (University of Lille, France) who conducted the numerical experiments presented in section 8. I further thank Joachim Schöberl (University of Linz, Austria) and Philipp Frauenfelder (ETH Zürich, Switzerland) who gave me the numerical values of μ and α presented in Remark 3.4 and obtained using their *hp*-FEM code.

REFERENCES

[1] A. Agouzal and F. Oudin, *A posteriori error estimator for finite volume methods*, Appl. Math. Comput., 110 (2000), pp. 239–250.

[2] A. Bergam and Z. Mghazli, *Estimateurs a posteriori d'un schéma de volumes finis pour un problème non linéaire*, C. R. Math Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 475–478.

[3] A. Bergam, Z. Mghazli, and R. Verfürth, *A posteriori estimators for the finite volume discretization of an elliptic problem*, Numer. Math., 95 (2003), pp. 599–624.

[4] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[5] Y. Coudière, J.-P. Villa, and P. Villedieu, *Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.

[6] Y. Coudière and P. Villedieu, *Convergence rate of a finite volume scheme for the linear convection–diffusion equation on locally refined meshes*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1123–1149.

[7] E. Dari, R. Durán, C. Padra, and V. Vampa, *A posteriori error estimators for nonconforming finite element methods*, M2AN Math. Model. Numer. Anal., 30 (1996), pp. 385–400.

[8] K. Djadel, S. Nicaise, and J. Tabka, *Some refined finite volume methods for elliptic problems with corner singularities*, Int. J. Finite Volumes, (2003), http://averoes.math.univ-paris13.fr/IJFVDB, PAPERS/RevFv_Nicaise.pdf.

[9] J. Droniou, *Error estimates for the convergence of a finite volume discretization of convection–diffusion equations*, J. Numer. Math., 11 (2003), pp. 1–32.

[10] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. 7, P. Ciarlet and J.-L. Lions, eds., North Holland, Amsterdam, 2000, pp. 723–1020.

[11] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations, Theory and Algorithms*, Springer Series in Comput. Math. 5, Springer, New York, 1986.

[12] R. Herbin and M. Ohlberger, *A posteriori error estimate for finite volume approximation of convection–diffusion problems*, in Finite Volume for Complex Applications, R. Herbin and D. Kröner, eds., Hermès, London, 2002, pp. 753–760.

[13] N. Jullian, *An error indicator for cell-centered finite volumes for linear convection–iffusion problems*, in Finite Volume for Complex Applications, R. Herbin and D. Kröner, eds., Hermès, London, 2002, pp. 777–784.

[14] D. Kröner and M. Ohlberger, *A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multi dimensions*, Math. Comp., 69 (1999), pp. 25–39.

[15] R. Lazarov and S. Tomov, *Adaptive finite volume element method for convection-diffusion-reaction problems in* 3*-d*, in Scientific Computing and Application, Y. W. P. Minev and Y. Lin, eds., Nova Science Publishing House, Huntington, NY, 2001, pp. 91–106.

[16] R. Lazarov and S. Tomov, *A posteriori error estimates for finite volume approximations of convection-diffusion-reaction equations*, Comput. Geosci., 6 (2002), pp. 483–503.

[17] J. Mackenzie, T. Sonar, and G. Warnecke, *A posteriori error estimates for the cell-vertex finite volume method*, in Adaptive Methods: Algorithms, Theory and Applications, W. Hackbusch and G. Wittum, eds., Vieweg, Berlin, 1994, pp. 221–235.

[18] L. Morley, *The triangular equilibrium element in the solution of plate bending problems*, Aero. Quarterly, 19 (1968), pp. 149–169.

[19] K. W. Morton and E. Süli, *A posteriori and a priori error analysis of finite volume methods*, in The Mathematics of Finite Elements and Applications, J. R. Whiteman, ed., John Wiley and Sons, New York, 1994, pp. 267–288.

[20] S. Nicaise, *A posteriori error estimations of some cell-centered finite volume methods for diffusion-convection-reaction problems*, SIAM J. Numer. Anal., submitted.

[21] T. K. Nilssen, X.-C. Tai, and R. Winther, *A robust nonconforming H^2-element*, Math. Comp., 70 (2000), pp. 489–505.

[22] M. Ohlberger, *A posteriori error estimate for finite volume approximations to singularly perturbed nonlinear convection-diffusion equations*, Numer. Math., 87 (2001), pp. 737–761.

[23] M. Ohlberger, *A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 355–387.

[24] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*, Ser. Comput. Methods Mech. Thermal Sci., McGraw Hill, New York, 1980.

[25] T. Sonar and E. Süli, *A dual graph-norm refinement indicator for finite volume approximations of the Euler equations*, Numer. Math., 78 (1998), pp. 619–658.

[26] R. Verfürth, *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, Chichester, 1996.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1504–1535

NUMERICAL RECONSTRUCTION OF HEAT FLUXES[*]

JIANLI XIE[†] AND JUN ZOU[‡]

Abstract. This paper studies the reconstruction of heat fluxes on an inner boundary of a heat conductive system when the measurement of temperature in a small subregion near the outer boundary of the physical domain is available. We will first consider two different regularization formulations for this severely ill-posed inverse problem and justify their well-posedness; then we will propose two fully discrete finite element methods to approximate the resultant nonlinear minimization problems. The existence and uniqueness of the discrete minimizers and convergence of the finite element solution are rigorously demonstrated. A conjugate gradient method is formulated to solve the nonlinear finite element optimization problems. Numerical experiments are given to demonstrate the stability and effectiveness of the proposed reconstruction methods.

Key words. numerical reconstruction, heat flux, regularization, optimization, finite elements

AMS subject classifications. 65M30, 35R25, 65M60

DOI. 10.1137/030602551

1. Introduction.

Consider a heat conductive system which occupies an open bounded domain Ω with an outer boundary Γ_o and an inner boundary Γ_i; see Figure 1. We are interested in a heat conductive system which can be modeled by the parabolic equation

$$\frac{\partial u}{\partial t} = \nabla \cdot (\alpha(x,t)\nabla u) \qquad \text{in} \quad \Omega \times (0,T), \tag{1.1}$$

assuming the initial condition

$$u(x,0) = u_0(x) \qquad \text{in} \quad \Omega \tag{1.2}$$

and the heat flux exchanges through the outer and inner boundaries Γ_o and Γ_i as follows:

$$-\alpha(x,t)\frac{\partial u}{\partial n} = c(x,t)(u(x,t) - u_a(x,t)) \qquad \text{on} \quad \Gamma_o \times (0,T), \tag{1.3}$$

$$-\alpha(x,t)\frac{\partial u}{\partial n} = q(x,t) \qquad \text{on} \quad \Gamma_i \times (0,T). \tag{1.4}$$

Here $\alpha(x,t)$ is the heat conductivity, $c(x,t)$ and $u_a(x,t)$ are specified functions, and $q(x,t)$ is the heat flux on the inner boundary Γ_i.

The forward initial-boundary value problem (1.1)–(1.4) has been well studied. The focus of this paper is on a physically more interesting and challenging inverse problem: Is it possible to effectively reconstruct the heat flux $q(x,t)$ on the inner boundary Γ_i for all time $t \in [0,T]$ when Γ_i is inaccessible?

[*] Received by the editors December 18, 2003; accepted for publication (in revised form) March 4, 2005; published electronically October 31, 2005.
http://www.siam.org/journals/sinum/43-4/60255.html

[†] Department of Mathematics, Shanghai Jiao Tong University, Shanghai 200240, People's Republic of China (jlxie@online.sh.cn).

[‡] Department of Mathematics, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong (zou@math.cuhk.edu.hk). The work of this author was fully supported by Hong Kong RGC Grants (Projects CUHK4244/01P and 403403).

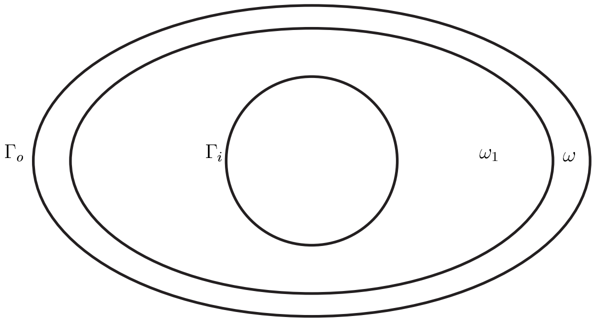


FIG. 1. *Physical domain* $\Omega = \omega_1 \cup (\bar{\omega} \setminus \Gamma_o)$.

In order to possibly reconstruct the heat flux $q(x,t)$, some extra information on the temperature $u(x,t)$ is needed. One choice is to assume the temperature data available in a small subregion ω near the outer boundary Γ_o (see Figure 1). Some high furnaces in steel companies are such examples, where special small devices are installed inside the furnaces but near the outer boundary to measure temperature.

This reconstruction problem is known to be a severely ill-posed inverse problem. One of the main difficulties in the reconstruction comes from both the space and time dependence of the heat flux $q(x,t)$ and the fact that the inner boundary is away from the small measurement subregion. The most severe instability of an inverse problem is triggered when the reconstruction involves some profile at the initial time and on some large boundary portion of a physical domain [7], [17], [19], [20], as is the case encountered here. As far as ill-posed inverse problems are concerned, not much work is found in the literature addressing numerical reconstructions of some physical profiles of both space and time; even less work can be found on convergence and stability analysis for numerical reconstruction methods. We refer readers to [1], [2], [3], [8], [9], [18], and the references therein for numerical reconstructions of profiles of some time-independent parameters in parabolic and elliptic systems.

The aim of this paper is to justify both theoretically and numerically the validation and effectiveness of two regularization formulations for solving the aforementioned severely ill-posed inverse problem of heat flux reconstruction. Indeed, as will be seen from the theory, numerical analysis, and simulations developed in what follows, the regularization methods are very stable and effective in numerical reconstruction of heat fluxes, without any constraints enforced on the search space of heat fluxes if appropriate regularizations are selected. In particular, the resulting nonlinear finite element minimization systems can be efficiently solved by conjugate gradient method.

The rest of this paper is organized as follows. In section 2, we investigate the first formulation with an L^2-regularization of both space and time for the heat flux and validate the "true" well-posedness of the formulation under no constraints on the search space of heat fluxes. In section 3, we study the ill-posedness of heat flux reconstruction and the stability of the regularization. In section 4, we study an alternative formulation of the inverse problem, which uses an L^2-regularization in space and H^1-regularization in time. As will be seen, this formulation turns out to be able to demonstrate much more satisfactory reconstructions. Regarding the approximation of the regularized nonlinear minimization systems, it is very tricky and essential to decide how to effectively discretize in both time and space the nonlinear optimizations and the associated parabolic equation so that the resulting fully discrete schemes converge. For this purpose, two fully discrete finite element approximations are proposed in sections 5 and 6, and the unique existence of discrete minimizers and their convergence to the continuous minimizer are rigorously demonstrated. For solving the nonlinear finite element minimization systems involved in the formulations, a conjugate gradient method is formulated in section 7, and the numerical experiments are presented in section 8 to verify the effectiveness of the proposed reconstruction methods.

We end this section with some useful notation. We define

$$H^m(0,T;B) = \Big\{u(t) \in B \text{ for a.e. } t \in (0,T) \text{ and } \|u\|_{H^m(0,T;B)} < \infty\Big\}$$

for a Banach space B and $m \geq 0$, with its norm given by

$$\|u\|_{H^m(0,T;B)} = \left\{\sum_{k=0}^{m} \int_0^T \|u^{(k)}(t)\|_B^2 dt\right\}^{1/2}.$$

For a given domain $\mathcal{O}$, $H^m(\mathcal{O})$ stands for the standard Sobolev space of mth order for any $m \geq 0$. The norms and seminorms of $H^m(\mathcal{O})$ are denoted by $\|\cdot\|_{m,\mathcal{O}}$ and $|\cdot|_{m,\mathcal{O}}$, respectively. When $m = 0$, we write $L^2(\mathcal{O}) = H^0(\mathcal{O})$ with the norm $\|\cdot\|_{0,\mathcal{O}}$. The domain $\mathcal{O}$ in the subindex will be dropped if $\mathcal{O} = \Omega$.

Further, C is frequently used to denote a generic constant, which depends only on the given data such as domain Ω and coefficients in (1.1)–(1.4) and is independent of unknown functions involved and the discrete time step τ and mesh size h.

2. First regularization formulation.

Recall that the inverse problem of interest here is to reconstruct the heat flux $q(x,t)$ in (1.4) on the inner boundary Γ_i, given the temperature measurement $z(x,t) \approx u(x,t)$ in the small subdomain ω (cf. Figure 1). The first approach we will study for solving the inverse problem is to formulate it into the following constrained minimizing process with L^2-regularization in both space and time for possible heat fluxes:

$$\min J(q) = \frac{1}{2}\int_0^T\!\!\int_\omega (u(q)-z)^2 dxdt + \frac{\beta}{2}\int_0^T\!\!\int_{\Gamma_i} q^2 dsdt \tag{2.1}$$

subject to $q \in L^2(0,T;L^2(\Gamma_i))$ and $u(q) \equiv u(q)(\cdot,t) \in H^1(\Omega)$ satisfying

$$u(x,0) = u_0(x) \qquad \text{in} \quad \Omega, \tag{2.2}$$

$$\int_\Omega \frac{\partial u}{\partial t} v dx + \int_\Omega \alpha \nabla u \cdot \nabla v dx + \int_{\Gamma_o} c\, u\, v ds = \int_{\Gamma_o} c\, u_a v ds - \int_{\Gamma_i} q\, v ds \tag{2.3}$$

for all $v \in H^1(\Omega)$ and for a.e. $t \in (0,T)$.

In what follows, we will demonstrate that the inverse problem for reconstruction of heat flux is an ill-posed problem and that the formulation (2.1)–(2.3) is a true regularization of the inverse problem; that is, the minimizer q not only exists uniquely, but also depends on the observation data z continuously.

For the subsequent analysis, we often use the following compactness result (cf. [13]).

LEMMA 2.1. *Suppose that $B_0 \subset B \subset B_1$ are Banach spaces, B_0 and B_1 are reflexive, and B_0 is compactly embedded into B. Let*

$$W = \left\{v;\ v \in L^2(0,T;B_0),\ v' = \frac{dv}{dt} \in L^2(0,T;B_1)\right\},$$

with the norm $\|v\|_W = \|v\|_{L^2(0,T;B_0)} + \|v'\|_{L^2(0,T;B_1)}$. *Then* W *is compactly embedded into* $L^2(0,T;B)$.

Throughout this section, the parameter functions $\alpha(x,t)$, $c(x,t)$, and $u_a(x,t)$ in (1.1)–(1.4) are assumed to satisfy the following natural conditions:

$$\begin{gathered} \alpha(x,t) \ge \alpha_0 > 0 \quad \text{for a.e.} \quad (x,t) \in \Omega \times (0,T), \\ c(x,t) \ge c_0 > 0 \quad \text{for a.e.} \quad (x,t) \in \Gamma_o \times (0,T), \\ \alpha(x,t) \in L^2(0,T;L^2(\Omega)); \quad c(x,t),\ u_a(x,t) \in L^2(0,T;L^2(\Gamma_o)). \end{gathered} \tag{2.4}$$

We start with the following unique existence.

THEOREM 2.2. *There exists a unique minimizer to the optimization problem* (2.1)–(2.3).

Proof. Clearly $\min J(q)$ is finite over $L^2(0,T;L^2(\Gamma_i))$; thus there exists a minimizing sequence $\{q^n\}$ such that

$$\lim_{n\to\infty} J(q^n) = \inf J(q). \tag{2.5}$$

This implies the boundedness of $\{q^n\}$ in $L^2(0,T;L^2(\Gamma_i))$ and thus the existence of such a subsequence, still denoted[1] as q^n, and $\{q^n\}$ converges to q^* weakly in $L^2(0,T;L^2(\Gamma_i))$. We now prove that this q^* is the unique minimizer of (2.1)–(2.3). We divide the proof into four steps.

Step 1. Letting $u^n \equiv u(q^n)(x,t)$, we show that there exists a subsequence of $\{u^n\}$ such that

$$u^n \to u^* \text{ weakly in } L^2(0,T;H^1(\Omega)) \text{ and } L^2(0,T;L^2(\Gamma_o)). \tag{2.6}$$

By the definition of $u(q^n)$ in (2.2)–(2.3), $u^n \in H^1(\Omega)$ satisfies $u^n(x,0) = u_0(x)$, and

$$\int_\Omega \frac{\partial u^n}{\partial t} v dx + \int_\Omega \alpha \nabla u^n \cdot \nabla v dx + \int_{\Gamma_o} c\, u^n v ds = \int_{\Gamma_o} c\, u_a v ds - \int_{\Gamma_i} q^n v ds \tag{2.7}$$

holds for any $v \in H^1(\Omega)$ and a.e. $t \in (0,T)$. Taking $v = u^n$ in (2.7), we obtain

$$\frac{1}{2}\frac{d}{dt}\|u^n\|_0^2 + \int_\Omega \alpha |\nabla u^n|^2 dx + \int_{\Gamma_o} c\,|u^n|^2 ds = \int_{\Gamma_o} c\, u_a u^n ds - \int_{\Gamma_i} q^n u^n ds. \tag{2.8}$$

Integrating over $(0,t)$, we derive

$$\begin{aligned} &\frac{1}{2}\|u^n(\cdot,t)\|_0^2 + \int_0^t\!\!\int_\Omega \alpha |\nabla u^n(x,t)|^2 dxdt + \int_0^t\!\!\int_{\Gamma_o} c(x,t)|u^n(x,t)|^2 dsdt \\ &= \frac{1}{2}\|u_0\|_0^2 + \int_0^t\!\!\int_{\Gamma_o} c(x,t)u_a(x,t)u^n(x,t)dsdt - \int_0^t\!\!\int_{\Gamma_i} q^n(x,t)u^n(x,t)dsdt; \end{aligned}$$

then by the Cauchy–Schwarz inequality and assumptions in (2.4), we have

$$\begin{aligned} &\frac{1}{2}\|u^n(\cdot,t)\|_0^2 + \alpha_0 \|\nabla u^n\|^2_{L^2(0,t;L^2(\Omega))} + c_0 \|u^n\|^2_{L^2(0,t;L^2(\Gamma_o))} \\ &\le \frac{1}{2}\|u_0\|_0^2 + \|c\, u_a\|_{L^2(0,T;L^2(\Gamma_o))} \|u^n\|_{L^2(0,t;L^2(\Gamma_o))} + \|q^n\|_{L^2(0,T;L^2(\Gamma_i))} \|u^n\|_{L^2(0,t;L^2(\Gamma_i))} \\ &\le \frac{1}{2}\|u_0\|_0^2 + C\Big(\|u^n\|_{L^2(0,t;L^2(\Gamma_o))} + \|u^n\|_{L^2(0,t;L^2(\Gamma_i))}\Big). \end{aligned}$$

[1] Where no confusion exists, throughout this paper we shall always use the same notation to denote a subsequence taken from some sequence.

Using the Sobolev trace theorem, we can estimate the above last term as follows, a technique that will be frequently used in the subsequent analysis:

$$\begin{aligned}\|u^n\|^2_{L^2(0,t;L^2(\Gamma_i))} &= \int_0^t \|u^n(\cdot,s)\|^2_{L^2(\Gamma_i)}ds \le \int_0^t \|u^n(\cdot,s)\|^2_{H^{1/2}(\Gamma_i)}ds\\ &\le \int_0^t \|u^n(\cdot,s)\|^2_{H^1(\Omega)}ds\\ &= \int_0^t \|u^n(\cdot,s)\|^2_{L^2(\Omega)}ds + \int_0^t \|\nabla u^n(\cdot,s)\|^2_{L^2(\Omega)}ds\\ &\le \Big(\|u^n\|_{L^2(0,t;L^2(\Omega))} + \|\nabla u^n\|_{L^2(0,t;L^2(\Omega))}\Big)^2.\end{aligned}$$

Taking the square root on both sides, plugging the result into the previous estimate, and then using Young's inequality, we obtain

$$\begin{aligned}\|u^n(\cdot,t)\|_0^2 &\le \|u^n(\cdot,t)\|_0^2 + \alpha_0\|\nabla u^n\|^2_{L^2(0,t;L^2(\Omega))} + c_0\|u^n\|^2_{L^2(0,t;L^2(\Gamma_o))}\\ &\le \|u_0\|_0^2 + C + \int_0^t \|u^n(\cdot,s)\|^2_{L^2(\Omega)}ds.\end{aligned} \tag{2.9}$$

This gives the boundedness of $\{u^n\}$ in $L^\infty(0,T;L^2(\Omega))$ by applying Gronwall's inequality; then using this bound one can get the boundedness of $\{u^n\}$ in $L^2(0,T;H^1(\Omega))$ and $L^2(0,T;L^2(\Gamma_o))$ from the second inequality in (2.9). Now the convergence in (2.6) follows immediately from this boundedness.

Step 2. We prove $u^* = u(q^*)$. Taking any function $\Psi(t) \in C^1[0,T]$ with $\Psi(T) = 0$, multiplying both sides of (2.7) by Ψ, and then integrating over $t \in (0,T)$, we get

$$\begin{aligned}&\int_0^T\int_{\Gamma_o} c\,u_a v\Psi(t)dsdt - \int_0^T\int_{\Gamma_i} q^n v\Psi(t)dsdt\\ &= -\int_0^T\int_\Omega u^n v\Psi'(t)dxdt + \int_0^T\int_\Omega \alpha\nabla u^n\cdot\nabla v\Psi(t)dxdt\\ &\quad-\int_\Omega \Psi(0)u_0(x)vdx + \int_0^T\int_{\Gamma_o} c\,u^n v\Psi(t)dsdt.\end{aligned}$$

By the weak convergence of q^n and u^n, we deduce from above that

$$\begin{aligned}&\int_0^T\int_{\Gamma_o} c\,u_a v\Psi(t)dsdt - \int_0^T\int_{\Gamma_i} q^* v\Psi(t)dsdt\\ &= \int_0^T\int_\Omega \alpha\nabla u^*\cdot\nabla v\Psi(t)dxdt + \int_0^T\int_{\Gamma_o} c\,u^* v\Psi(t)dsdt\\ &\quad-\int_\Omega \Psi(0)u_0(x)vdx - \int_0^T\int_\Omega u^* v\Psi'(t)dxdt.\end{aligned} \tag{2.10}$$

Noting that (2.10) is also true for any $\Psi(t) \in C_0^\infty(0,T)$, by integration by parts over $t \in (0,T)$ for the last term we have

$$\int_\Omega \frac{\partial u^*}{\partial t}vdx + \int_\Omega \alpha\nabla u^*\cdot\nabla vdx + \int_{\Gamma_o} c\,u^* vds = \int_{\Gamma_o} c\,u_a vds - \int_{\Gamma_i} q^* vds \qquad \forall v \in H^1(\Omega)$$

for a.e. $t \in (0,T)$. Using this and integration by parts again for the last term in (2.10) shows that $u^*(x,0) = u_0(x)$. This verifies $u^* = u(q^*)$.

Step 3. We prove the strong convergence

$$\lim_{n\to\infty}\int_0^T\int_\omega |u^n - z|^2 dxdt = \int_0^T\int_\omega |u^* - z|^2 dxdt. \tag{2.11}$$

It suffices to prove the strong convergence of $\{u^n\}$ in $L^2(0,T;L^2(\Omega))$. By Lemma 2.1, we need only show the boundedness of $\{\frac{\partial u^n}{\partial t}\}$ in $L^2(0,T;(H^1(\Omega))')$.

It follows from (2.7) that for any $v \in L^2(0,T;H^1(\Omega))$,

$$\left|\left\langle \frac{\partial u^n}{\partial t}, v\right\rangle\right| \le C(\|u^n\|_{H^1(\Omega)} + \|u^n\|_{L^2(\Gamma_o)} + \|u_a\|_{L^2(\Gamma_o)} + \|q^n\|_{L^2(\Gamma_i)})\|v\|_{H^1(\Omega)}; \tag{2.12}$$

this, along with the boundedness of $\{u^n\}$ proved in Step 1, implies the boundedness of $\{\frac{\partial u^n}{\partial t}\}$ in $L^2(0,T;(H^1(\Omega))')$.

Step 4. We prove q^* is a unique minimizer to the system (2.1)–(2.3). Using the results in Step 3 and the lower semicontinuity of a norm, we have

$$\begin{aligned} J(q^*) &= \frac{1}{2}\int_0^T\int_\omega |u(q^*) - z|^2 dxdt + \frac{\beta}{2}\int_0^T\int_{\Gamma_i} |q^*|^2 dsdt \\ &\le \lim_{n\to\infty}\int_0^T\int_\omega |u(q^n) - z|^2 dxdt + \frac{\beta}{2}\lim_{n\to\infty}\inf\int_0^T\int_{\Gamma_i} |q^n|^2 dsdt \\ &\le \lim_{n\to\infty}\inf J(q^n) = \inf J(q), \end{aligned} \tag{2.13}$$

so q^* is indeed a minimizer. The uniqueness of minimizers is a consequence of the convexity of $u(q)$ and the strict convexity of $J(q)$. □

PROPOSITION 2.3. *Assume that $\{q^n\}$, with $q^n \in L^2(0,T;L^2(\Gamma_i))$, is a minimizing sequence of $J(q)$ in* (2.1)*; then $\{q^n\}$ converges to the unique minimizer of $J(q)$ strongly in $L^2(0,T;L^2(\Gamma_i))$.*

Proof. From the proof of Theorem 2.2, we know any subsequence of $\{q^n\}$ has a subsequence converging weakly to the unique minimizer of $J(q)$. Thus the whole sequence $\{q^n\}$ converges weakly to the unique minimizer of $J(q)$. Further, one notices from (2.5), (2.11), and (2.13) that

$$\lim_{n\to\infty}\int_0^T\int_{\Gamma_i} |q^n|^2 dsdt = \int_0^T\int_{\Gamma_i} |q^*|^2 dsdt;$$

thus the weak and norm convergences imply the strong convergence. □

3. Ill-posedness of heat flux reconstruction and stability of the regularization.

Next, we study the ill-posedness of heat flux reconstruction and stability of the regularization system (2.1)–(2.3). The following theorem confirms the ill-posedness of the heat flux reconstruction problem (1.1)–(1.4).

THEOREM 3.1. *Let $u(q)$ be a mapping from $L^2(0,T;L^2(\Gamma_i))$ to $L^2(0,T;L^2(\omega))$, defined by the system* (2.2)*–*(2.3) *associated with any given heat flux q in $L^2(0,T;L^2(\Gamma_i))$. Then there exists a sequence $\{q^n\}$ from $L^2(0,T;L^2(\Gamma_i))$ such that $u(q^n) \to 0$ but $\|q^n\|_{L^2(0,T;L^2(\Gamma_i))} \to \infty$, and the inverse of $u(\cdot)$ is unbounded.*

Proof. From the proof of Theorem 2.2, we know for any bounded sequence $\{q^n\}_{n=1}^\infty$ there exists a subsequence $\{q^{n_k}\}_{k=1}^\infty$ such that $\{u(q^{n_k})\}_{k=1}^\infty$ is strongly convergent in $L^2(0,T;L^2(\omega))$. Therefore, as an operator from $L^2(0,T;L^2(\Gamma_i))$ to $L^2(0,T;L^2(\omega))$,

$u(\cdot)$ is compact. On the other hand, one can directly verify that $u(\cdot)$ is a one-to-one mapping and can be decomposed into $u(q) = w(q) + u(0)$, where $w(q)(\cdot, t) \in H^1(\Omega)$ solves the parabolic system (1.1)–(1.4) with $w(q)(x, 0) = 0$ in Ω and $u_a \equiv 0$. The rest of the proof follows the routine procedure; for example, see [10, pp. 13–14]. □

The next theorem shows that the solution q to the regularization system (2.1)–(2.3) depends continuously on the observation data z, so system (2.1)–(2.3) is a "true" regularization to the original inverse problem $u(q) = z$. The detailed proof can be found in [14].

Theorem 3.2. *Let $\{z^n\}$ be a sequence such that*

$$z^n \to z \quad in \quad \mathrm{L}^2(0, \mathrm{T}; \mathrm{L}^2(\omega)) \quad as \quad \mathrm{n} \to \infty, \tag{3.1}$$

and let $\{q^n\}$ be the minimizers of problem (2.1)–(2.3) *with z replaced by z^n. Then the whole sequence $\{q^n\}$ converges in $L^2(0, T; L^2(\Gamma_i))$ to the unique minimizer of* (2.1)–(2.3).

4. An alternative formulation. In this section, we investigate an alternative formulation for reconstruction of heat fluxes in the heat conductive system (1.1)–(1.4), using an L^2-regularization in space and H^1-regularization in time for heat fluxes. As one can see from numerical results in section 8, this new formulation is able to generate more satisfactory reconstructions. This results in the following constrained minimization:

(4.1)
$$\min J(q) = \frac{1}{2}\int_0^T\int_\omega (u(q) - z)^2 dxdt + \frac{\beta}{2}\left(\int_{\Gamma_i} q^2(x, 0)ds + \int_0^T\int_{\Gamma_i} |q_t(x, t)|^2 dsdt\right)$$

subject to $q \in H^1(0, T; L^2(\Gamma_i))$ and $u(q) \equiv u(q)(\cdot, t) \in H^1(\Omega)$ satisfying

$$u(x, 0) = u_0(x) \qquad \text{in} \quad \Omega, \tag{4.2}$$

$$\int_\Omega \frac{\partial u}{\partial t} v dx + \int_\Omega \alpha \nabla u \cdot \nabla v dx + \int_{\Gamma_o} c\, u\, v ds = \int_{\Gamma_o} c\, u_a v ds - \int_{\Gamma_i} q\, v ds \tag{4.3}$$

for all $v \in H^1(\Omega)$ and a.e. $t \in (0, T)$.

The following theorem justifies the well-posedness of the system (4.1)–(4.3) and its stability with respect to the observation data.

Theorem 4.1. *There exists a unique minimizer to the optimization problem* (4.1)–(4.3), *and the minimizer depends on the observation data z continuously.*

Proof. It is clear that $\min J(q)$ is finite over $H^1(0, T; L^2(\Gamma_i))$; thus there exists a minimizing sequence $\{q^n\}$ such that

$$\lim_{n\to\infty} J(q^n) = \inf J(q).$$

This implies the boundedness of $\{q^n\}$ in $H^1(0, T; L^2(\Gamma_i))$ and the existence of a subsequence, still denoted as $\{q^n\}$, such that

$$\begin{aligned} q^n &\to q^* \text{ weakly in } L^2(0, T; L^2(\Gamma_i)), \\ \frac{\partial q^n}{\partial t} &\to p^* \text{ weakly in } L^2(0, T; L^2(\Gamma_i)), \\ q^n(x, 0) &\to q_0^* \text{ weakly in } L^2(\Gamma_i). \end{aligned}$$

We can show that $p^* = \partial q^*/\partial t$ and $q^*(x,0) = q_0^*$. In fact, taking any function $\varphi(x) \in L^2(\Gamma_i)$ and $\psi(t) \in C_0^\infty(0,T)$, we deduce

$$\int_0^T\int_{\Gamma_i} \frac{\partial}{\partial t} q^n(x,t)\varphi(x)\psi(t)dsdt = -\int_0^T\int_{\Gamma_i} q^n(x,t)\varphi(x)\psi'(t)dsdt.$$

Passing to the limit, we derive

$$\int_0^T\int_{\Gamma_i} p^*(x,t)\varphi(x)\psi(t)dsdt = -\int_0^T\int_{\Gamma_i} q^*(x,t)\varphi(x)\psi'(t)dsdt.$$

This shows $p^* = \partial q^*/\partial t$.

Then letting $\varphi(x) \in L^2(\Gamma_i)$ and $\psi(t) \in C^\infty(0,T)$ with $\psi(T) = 0$ and $\psi(0) = 1$, we obtain

$$\int_0^T\int_{\Gamma_i} \frac{\partial}{\partial t} q^n(x,t)\varphi(x)\psi(t)dsdt = \int_{\Gamma_i} q^n(x,0)\varphi(x)ds - \int_0^T\int_{\Gamma_i} q^n(x,t)\varphi(x)\psi'(t)dsdt.$$

By the weak convergence of $\partial q^n/\partial t$, $q^n(x,0)$, and q^n, we deduce

$$\int_0^T\int_{\Gamma_i} \frac{\partial}{\partial t} q^*(x,t)\varphi(x)\psi(t)dsdt = \int_{\Gamma_i} q_0^*(x)\varphi(x)ds - \int_0^T\int_{\Gamma_i} q^*(x,t)\varphi(x)\psi'(t)dsdt.$$

Integrating by parts the left-hand side, we obtain for any $\varphi(x) \in L^2(\Gamma_i)$ that

$$\int_{\Gamma_i} q_0^*(x)\varphi(x)ds = \int_{\Gamma_i} q^*(x,0)\varphi(x)ds,$$

which implies $q^*(x,0) = q_0^*$. The rest of the proof is similar to those of Theorems 2.2 and 3.2. □

Similarly to Proposition 2.3, we have the following strong convergence (cf. [14]).

PROPOSITION 4.2. *Any minimizing sequence $\{q^n\}$ of $J(q)$ in (4.1) over $H^1(0,T;$ $L^2(\Gamma_i))$ converges to the unique minimizer of $J(q)$ strongly in $H^1(0,T;L^2(\Gamma_i))$.*

5. Finite element approximation of system (2.1)–(2.3) and its convergence. We now propose a fully discrete finite element method for solving the continuous minimization problem (2.1)–(2.3). For the sake of exposition, we study in detail the case where the outer and inner boundaries Γ_o and Γ_i are both circles centered at the origin; see Figure 2. The subsequent results can be extended to more general domains by combining the analysis used here and the finite element analysis for the case when the approximation of the physical domain is involved [4].

Let us start with a triangulation of the domain Ω. To do so, we generate a set of circles all centered at the origin, starting with Γ_i and ending with Γ_o. Next we choose a set of quasi-uniformly distributed points on Γ_o, which are then connected to the origin to yield a set of radial lines, and the intersections of these lines with all the previous generated circles also yield a partition of each circle; see Figure 2. Now the triangulation $\mathcal{T}^h$ of Ω is formed by these sectorial elements. The arc segments on Γ_o and Γ_i generate naturally two triangulations of Γ_o and Γ_i, respectively, denoted by Γ_o^h and Γ_i^h.

For each sectorial element K, say $K = \{(r\cos\theta, r\sin\theta);\ r_1 \le r \le r_2, \theta_1 \le \theta \le \theta_2\}$, there exists a one-to-one mapping $\hat{F}_K : \hat{K} \to K$ such that $K = \hat{F}_K(\hat{K})$, where $\hat{K}$

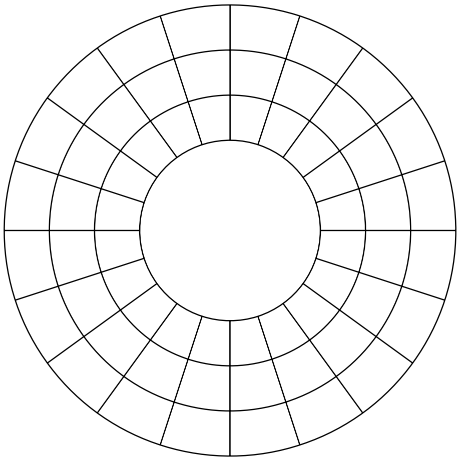


FIG. 2. *Circular partition of* Ω *and partition of each circle.*

is a rectangular reference element. For example, if $\hat{K} = [0,1] \times [0,1]$, we can take $\hat{F}_K$ as

$$\begin{cases} x = (r\,r_2 + (1-r)r_1)\cos(\theta\,\theta_2 + (1-\theta)\theta_1), \\ y = (r\,r_2 + (1-r)r_1)\sin(\theta\,\theta_2 + (1-\theta)\theta_1). \end{cases} \tag{5.1}$$

Now we can define the finite element space V^h to be

$$V^h = \Big\{ v_h \in C(\bar{\Omega});\ v_h(x)\big|_K = \hat{v} \circ \hat{F}_K^{-1}(x)\ \forall \hat{v} \in \mathcal{Q}_1(\hat{K}) \Big\},$$

where $\mathcal{Q}_1(\hat{K})$ is the space of bilinear functions on the reference element $\hat{K}$, and $V^h_{\Gamma_o}$, $V^h_{\Gamma_i}$ are the restrictions of V^h on Γ_o and Γ_i, respectively.

To fully discretize the system (2.1)–(2.3), we also need the time discretization. For this, we divide the time interval $[0,T]$ into M equally spaced subintervals using nodal points

$$\Delta: \quad 0 = t_0 < t_1 < \cdots < t_M = T \tag{5.2}$$

with $t_n = n\tau$, $\tau = T/M$. For a continuous mapping $u : [0,T] \to L^2(\Omega)$, we define $u^n = u(\cdot, t_n)$ for $0 \le n \le M$. For a given sequence $\{u^n\}_{n=0}^M \subset L^2(\Omega)$, we define its difference quotient and the averaging $\bar{u}^n$ of a function $u(\cdot,t)$ as follows:

$$\partial_\tau u^n = \frac{u^n - u^{n-1}}{\tau}, \quad \bar{u}^n = \frac{1}{\tau}\int_{t_{n-1}}^{t_n} u(\cdot,t)dt, \tag{5.3}$$

where for $n = 0$, we let $\bar{u}^0 = u(\cdot, 0)$.

In our subsequent convergence analysis, we need a crucial projection operator Q_h from $L^2(\Omega)$ into V^h defined on sectorial elements, which should possess the following L^2- and H^1-stability and optimal L^2-norm error estimate:

$$\lim_{h\to 0} \|v - Q_h v\|_1 = 0 \quad \forall v \in H^1(\Omega), \tag{5.4}$$

$$\|Q_h v\|_0 \le C\|v\|_0, \quad \|Q_h v\|_1 \le C\|v\|_1 \quad \forall v \in H^1(\Omega), \tag{5.5}$$

$$\|v - Q_h v\|_0 \le Ch\|v\|_1 \quad \forall v \in H^1(\Omega). \tag{5.6}$$

Noting that the transform $\hat{F}_K$: $\hat{K} \to K$ is not of polynomial type and that the functions in V^h may not be piecewise polynomials, the standard L^2-projection operator from $L^2(\Omega)$ into V^h (cf. [16]) does not satisfy these properties. Instead, we introduce a novel weighted L^2-projection operator Q_h from $L^2(\Omega)$ into V^h as follows:

$$\sum_{K\in\mathcal{T}^h}\int_K (Q_h w)\, v\, |J_K^{-1}(x,y)|\, dxdy = \sum_{K\in\mathcal{T}^h}\int_K w\, v\, |J_K^{-1}(x,y)|\, dxdy \quad \forall\, w \in L^2(\Omega), v \in V^h,$$

where $J_K(x,y) = \hat{J}_K(r,\theta)$ for all x, y and r, θ defined by (5.1) and $\hat{J}_K(r,\theta)$ is the Jacobian determinant of the transform $\hat{F}_K$. One can show that this weighted operator Q_h is well-defined, and it possesses all the properties (5.4), (5.5), and (5.6). The detailed proof was given in Xie [14].

Now we are ready to formulate the finite element approximation of the minimization (2.1)–(2.3). We approximate the heat flux $q(x,t)$ by a piecewise constant function $q_{h,\tau}(x,t)$ over the time partition Δ in (5.2):

$$q_{h,\tau}(x,t) = \sum_{n=1}^{M} \chi_n(t) q_h^n(x), \tag{5.7}$$

where $q_h^n(x) \in V_{\Gamma_i}^h$ and $\chi_n(t)$ is the characteristic function on the interval (t_{n-1}, t_n).

Using the composite trapezoidal rule for the time discretization of the first integral in (2.1) and the exact time integration for the second term, the fully discrete finite element approximation to problem (2.1)–(2.3) can be formulated as follows:

$$\min J_{h,\tau}(q_{h,\tau}) = \frac{\tau}{2}\sum_{n=0}^{M} \alpha_n \int_\omega (u_h^n - z^n)^2 dx + \frac{\beta\tau}{2}\sum_{n=1}^{M}\int_{\Gamma_i} |q_h^n|^2 ds \tag{5.8}$$

over all $q_h^n \in V_{\Gamma_i}^h$ with $u_h^n \equiv u_h^n(q_{h,\tau}) \in V^h$ satisfying

$$u_h^0 = Q_h u_0(x), \tag{5.9}$$

$$\begin{aligned}&\int_\Omega \partial_\tau u_h^n \phi_h dx + \int_\Omega \bar{a}^n \nabla u_h^n \cdot \nabla \phi_h dx + \int_{\Gamma_o} \bar{c}^n u_h^n \phi_h ds\\ &= \int_{\Gamma_o} \bar{c}^n \bar{u}_a^n \phi_h ds - \int_{\Gamma_i} q_h^n \phi_h ds \quad \forall \phi_h \in V^h\end{aligned} \tag{5.10}$$

for $n = 1, 2, \ldots, M$. Here $\{\alpha_n\}$ are the coefficients of the composite trapezoidal rule, i.e., $\alpha_0 = \alpha_M = \frac{1}{2}$ and $\alpha_n = 1$ for all $n \neq 0, M$.

For convenience, the minimization of $J_{h,\tau}$ also shall be regarded as the minimization over the product space $\prod_{n=1}^M V_{\Gamma_i}^h$, and we will often write (5.8) as

$$\min J_{h,\tau}(\{q_h^1, q_h^2, \ldots, q_h^M\}) = \frac{\tau}{2}\sum_{n=0}^{M} \alpha_n \int_\omega (u_h^n - z^n)^2 dx + \frac{\beta\tau}{2}\sum_{n=1}^{M}\int_{\Gamma_i} |q_h^n|^2 ds. \tag{5.11}$$

Before verifying the existence of a unique minimizer to the finite element minimization (5.8)–(5.10), we first derive some useful a priori estimates on the discrete solutions u_h^n to the system (5.9)–(5.10).

In the rest of this section, we assume on the functions $\alpha(x,t)$ and $c(x,t)$ in (1.1)–(1.4) that

$$\alpha \in H^1(0,T;L^\infty(\Omega)) \quad \text{and} \quad c \in H^1(0,T;L^\infty(\Gamma_o))$$

and introduce two related constants

$$C_1 = \|\alpha\|_{H^1(0,T;L^\infty(\Omega))}, \quad C_2 = \|c\|_{H^1(0,T;L^\infty(\Gamma_o))}.$$

The following auxiliary lemma (cf. [14]) will be needed in the subsequent analysis.

Lemma 5.1. *For any*

$$f \in H^1(0,T;L^\infty(\Omega)) \quad \textit{and} \quad g \in L^2(0,T;L^\infty(\Omega)),$$

we have the estimates

$$\|\bar{f}^n - \bar{f}^{n-1}\|_{L^\infty(\Omega)} \le \sqrt{\tau}\|f_t\|_{L^2(t_{n-2},t_n;L^\infty(\Omega))}, \tag{5.12}$$

$$\|\bar{f}^n\bar{g}^n - \overline{fg}^n\|_{L^2(\Omega)} \le \frac{2}{3}\|f_t\|_{L^2(t_{n-1},t_n;L^\infty(\Omega))}\, \|g\|_{L^2(t_{n-1},t_n;L^2(\Omega))}. \tag{5.13}$$

Lemma 5.2. *Assume that u_h^n is the solution of the finite element system* (5.9)–(5.10) *corresponding to $q_{h,\tau}$. Then we have the following stability estimates:*

$$\max_{1\le n\le M}\|u_h^n\|_0^2 + \tau\sum_{n=1}^M \|\nabla u_h^n\|_0^2 + \tau\sum_{n=1}^M \|u_h^n\|_{0,\Gamma_o}^2$$
$$\le C\,(\|u_0\|_0^2 + C_2^2\|u_a\|_{L^2(0,T;L^2(\Gamma_o))}^2 + \|q_{h,\tau}\|_{L^2(0,T;L^2(\Gamma_i))}^2), \tag{5.14}$$

$$\max_{1\le n\le M}\|\nabla u_h^n\|_0^2 + \max_{1\le n\le M}\|u_h^n\|_{0,\Gamma_o}^2 + \tau\sum_{n=1}^M \|\partial_\tau u_h^n\|_0^2$$
$$\le C\,\tau^{-1}(\|u_0\|_1^2 + C_2^2\|u_a\|_{L^2(0,T;L^2(\Gamma_o))}^2 + \|q_{h,\tau}\|_{L^2(0,T;L^2(\Gamma_i))}^2), \tag{5.15}$$

$$\tau\sum_{n=1}^M \|\partial_\tau u_h^n\|_{(H^1(\Omega))'}^2$$
$$\le C\,(C_1^2 + C_2^2 + \tau^{-1}h^2)\,(\|u_0\|_0^2 + C_2^2\|u_a\|_{L^2(0,T;L^2(\Gamma_o))}^2 + \|q_{h,\tau}\|_{L^2(0,T;L^2(\Gamma_i))}^2). \tag{5.16}$$

Proof. The proof of (5.14) follows directly by taking $\phi_h = \tau u_h^n$ in (5.10) and then applying the Sobolev trace theorem and Young's and Gronwall's inequalities.

Next, we show (5.15). Taking $\phi_h = \tau\partial_\tau u_h^n = u_h^n - u_h^{n-1}$ in (5.10), we obtain

$$\tau\|\partial_\tau u_h^n\|_0^2 + \int_\Omega \bar{\alpha}^n \nabla u_h^n \cdot \nabla(u_h^n - u_h^{n-1})dx + \int_{\Gamma_o} \bar{c}^n u_h^n(u_h^n - u_h^{n-1})ds$$
$$= \int_{\Gamma_o} \bar{c}^n \bar{u}_a^n(u_h^n - u_h^{n-1})ds - \int_{\Gamma_i} q_h^n(u_h^n - u_h^{n-1})ds.$$

Summing up the above equation over $n = 1, 2, \ldots, k \le M$, we obtain

$$\begin{aligned}
&\tau \sum_{n=1}^{k} \|\partial_\tau u_h^n\|_0^2 + \frac{1}{2}\sum_{n=1}^{k}\int_\Omega \bar{\alpha}^n(|\nabla u_h^n|^2 - |\nabla u_h^{n-1}|^2)dx \\
&\quad + \frac{1}{2}\sum_{n=1}^{k}\int_{\Gamma_o} \bar{c}^n(|u_h^n|^2 - |u_h^{n-1}|^2)ds \\
&\le \sum_{n=1}^{k}\int_{\Gamma_o} \bar{c}^n \bar{u}_a^n(u_h^n - u_h^{n-1})ds - \sum_{n=1}^{k}\int_{\Gamma_i} q_h^n(u_h^n - u_h^{n-1})ds.
\end{aligned}$$

Then using the discrete integration by parts formula

$$\sum_{n=1}^{k}(a_n - a_{n-1})b_n = a_k b_k - a_0 b_0 - \sum_{n=1}^{k} a_{n-1}(b_n - b_{n-1}), \tag{5.17}$$

where b_0 appearing on the right-hand side can be any real number, we derive

$$\begin{aligned}
&\tau \sum_{n=1}^{k} \|\partial_\tau u_h^n\|_0^2 + \frac{1}{2}\alpha_0\|\nabla u_h^k\|_0^2 + \frac{1}{2}c_0\|u_h^k\|_{0,\Gamma_o}^2 \\
&\le \frac{1}{2}\int_\Omega \bar{\alpha}^0|\nabla u_h^0|^2dx + \frac{1}{2}\sum_{n=1}^{k}\int_\Omega(\bar{\alpha}^n - \bar{\alpha}^{n-1})|\nabla u_h^{n-1}|^2dx \\
&\quad + \frac{1}{2}\int_{\Gamma_o}\bar{c}^0|u_h^0|^2ds + \frac{1}{2}\sum_{n=1}^{k}\int_{\Gamma_o}(\bar{c}^n - \bar{c}^{n-1})|u_h^{n-1}|^2ds \\
&\quad + \int_{\Gamma_o}\bar{c}^k\bar{u}_a^k u_h^k ds - \sum_{n=1}^{k}\int_{\Gamma_o}(\bar{c}^n\bar{u}_a^n - \bar{c}^{n-1}\bar{u}_a^{n-1})u_h^{n-1}ds \\
&\quad - \int_{\Gamma_i} q_h^k u_h^k ds + \sum_{n=1}^{k}\int_{\Gamma_i}(q_h^n - q_h^{n-1})u_h^{n-1}ds,
\end{aligned}$$

where $\bar{u}_a^0$ and q_h^0 are taken to be 0. We now estimate the terms on the right-hand side of the above inequality. First, for those terms without summation, we can deduce by using the properties of Q_h and the Sobolev trace theorem that

$$\frac{1}{2}\int_\Omega \bar{\alpha}^0|\nabla u_h^0|^2dx + \frac{1}{2}\int_{\Gamma_o}\bar{c}^0|u_h^0|^2ds \le C\,(C_1 + C_2)\|u_0\|_1^2,$$

$$\int_{\Gamma_o}\bar{c}^k\bar{u}_a^k u_h^k ds \le \frac{1}{2}\|\bar{c}^k\bar{u}_a^k\|_{0,\Gamma_o}^2 + \frac{1}{2}\|u_h^k\|_{0,\Gamma_o}^2 \le \tau^{-1}\left(\tau\sum_{n=1}^{k}\|\bar{c}^n\bar{u}_a^n\|_{0,\Gamma_o}^2 + \tau\sum_{n=1}^{M}\|u_h^n\|_{0,\Gamma_o}^2\right),$$

$$\int_{\Gamma_i} q_h^k u_h^k ds \le \frac{1}{2}\|q_h^k\|_{0,\Gamma_i}^2 + \frac{1}{2}\|u_h^k\|_1^2 \le \tau^{-1}\left(\|q_{h,\tau}\|_{L^2(0,T;L^2(\Gamma_i))}^2 + \tau\sum_{n=1}^{M}\|u_h^n\|_1^2\right).$$

Using (5.12) we have the following estimates:

$$\sum_{n=1}^{k}\int_\Omega(\bar{\alpha}^n - \bar{\alpha}^{n-1})|\nabla u_h^{n-1}|^2dx \le \frac{4}{3}C_1\sqrt{\tau}\sum_{n=1}^{k}\|\nabla u_h^{n-1}\|_0^2,$$

$$\sum_{n=1}^{k}\int_{\Gamma_o}(\bar{c}^n - \bar{c}^{n-1})|u_h^{n-1}|^2ds \le \frac{4}{3}C_2\sqrt{\tau}\sum_{n=1}^{k}\|u_h^{n-1}\|_{0,\Gamma_o}^2.$$

Applying the Cauchy–Schwarz inequality and the Sobolev trace theorem, we have

$$\sum_{n=1}^{k}\int_{\Gamma_o}(\bar{c}^n\bar{u}_a^n-\bar{c}^{n-1}\bar{u}_a^{n-1})u_h^{n-1}ds\le\sum_{n=1}^{k}\|\bar{c}^n\bar{u}_a^n\|_{0,\Gamma_o}^2+\sum_{n=1}^{k}\|u_h^{n-1}\|_{0,\Gamma_o}^2,$$
$$\sum_{n=1}^{k}\int_{\Gamma_i}(q_h^n-q_h^{n-1})u_h^{n-1}ds\le\sum_{n=1}^{k}\|q_h^n\|_{0,\Gamma_i}^2+\sum_{n=1}^{k}\|u_h^{n-1}\|_1^2.$$

Combining all these estimates with (5.14), we obtain (5.15).

It remains to show (5.16). For any $\phi\in H^1(\Omega)$, taking $\phi_h=Q_h\phi$ in (5.10), we have

$$\int_\Omega\partial_\tau u_h^nQ_h\phi dx+\int_\Omega\bar{\alpha}^n\nabla u_h^n\nabla Q_h\phi dx+\int_{\Gamma_o}\bar{c}^nu_h^nQ_h\phi ds=\int_{\Gamma_o}\bar{c}^n\bar{u}_a^nQ_h\phi ds-\int_{\Gamma_i}q_h^nQ_h\phi ds.$$

Using the property of Q_h in (5.5) and the Cauchy–Schwarz inequality, we derive

$$\left|\int_\Omega\partial_\tau u_h^nQ_h\phi dx\right|\le C\left(C_1\|\nabla u_h^n\|_0+C_2\|u_h^n\|_{0,\Gamma_o}+C_2\|\bar{u}_a^n\|_{0,\Gamma_o}+\|q_h^n\|_{0,\Gamma_i}\right)\|\phi\|_1.$$

On the other hand, applying the Cauchy–Schwarz inequality and the property of Q_h in (5.6), we obtain

$$\left|\int_\Omega\partial_\tau u_h^n(\phi-Q_h\phi)dx\right|\le Ch\|\partial_\tau u_h^n\|_0\|\phi\|_1.$$

It follows from the above two inequalities that for any $\phi\in H^1(\Omega)$,

$$\left|\int_\Omega\partial_\tau u_h^n\phi dx\right|\le C\left(C_1\|\nabla u_h^n\|_0+C_2\|u_h^n\|_{0,\Gamma_o}+C_2\|\bar{u}_a^n\|_{0,\Gamma_o}\right.$$
$$\left.+\|q_h^n\|_{0,\Gamma_i}+h\|\partial_\tau u_h^n\|_0\right)\|\phi\|_1,$$

which implies

$$\|\partial_\tau u_h^n\|_{(H^1(\Omega))'}\le C\left(C_1\|\nabla u_h^n\|_0+C_2\|u_h^n\|_{0,\Gamma_o}+C_2\|\bar{u}_a^n\|_{0,\Gamma_o}+\|q_h^n\|_{0,\Gamma_i}+h\|\partial_\tau u_h^n\|_0\right).$$

Taking squares on both sides and adding up over $n=1,\dots,M$, (5.16) then follows from (5.14) and (5.15). □

Remark 5.3. Fortunately, the unbounded factor τ^{-1} in the estimate (5.15) can be cancelled in the subsequent convergence analysis; see (5.28) and the last estimate in the proof of Lemma 5.5.

Based on the stability estimates (5.14)–(5.16), we are now ready to show the existence and uniqueness of minimizers to the finite element system (5.9)–(5.11).

THEOREM 5.4. *There exists a unique minimizer to the finite element system* (5.9)–(5.11).

Proof. By the stability estimates of Lemma 5.2 and the same argument as in Theorem 2.2, we know there exists a minimizing sequence $\{q_h^{1,k},q_h^{2,k},\dots,q_h^{M,k}\}_{k=1}^\infty$ such that

$$\lim_{k\to\infty}J_{h,\tau}(\{q_h^{1,k},q_h^{2,k},\dots,q_h^{M,k}\})=\inf_{\{q_h^n\}_{n=1}^M\in V_\Gamma^h}J_{h,\tau}(\{q_h^1,q_h^2,\dots,q_h^M\}),$$

and

$$q_h^{n,k} \to q_h^{n,*} \quad \text{in any norm for} \quad n = 1, 2, \ldots, M \quad \text{as} \quad k \to \infty.$$

Next, we prove $\{q_h^{1,*}, q_h^{2,*}, \ldots, q_h^{M,*}\}$ is the unique minimizer of (5.9)–(5.11). Let $q_{h,\tau}^k$ and $q_{h,\tau}^*$ be the functions defined in (5.7) by $\{q_h^{n,k}\}_{n=1}^M$ and $\{q_h^{n,*}\}_{n=1}^M$, respectively; then $u_h^n(q_{h,\tau}^k)$ and $u_h^n(q_{h,\tau}^*)$ are the finite element solutions to (5.9)–(5.10) corresponding to $q_{h,\tau}^k$ and $q_{h,\tau}^*$, respectively.

Let $w_h^{n,k} = u_h^n(q_{h,\tau}^k) - u_h^n(q_{h,\tau}^*)$; then $w_h^{0,k} = 0$ and for $n = 1, 2, \ldots, M$, $w_h^{n,k}$ solves

$$\int_\Omega \partial_\tau w_h^{n,k} \phi_h dx + \int_\Omega \bar{\alpha}^n \nabla w_h^{n,k} \cdot \nabla \phi_h dx + \int_{\Gamma_o} \bar{c}^n w_h^{n,k} \phi_h ds$$
$$= \int_{\Gamma_i} (q_h^{n,*} - q_h^{n,k}) \phi_h ds \quad \forall \phi_h \in V^h.$$

Taking $\phi_h = \tau w_h^{n,k}$ in the above equation, one can directly show by Gronwall's inequality that

$$\max_{1 \le n \le M} \|w_h^{n,k}\|_0^2 \le C\tau \sum_{n=1}^M \|q_h^{n,*} - q_h^{n,k}\|_{0,\Gamma_i}^2. \tag{5.18}$$

This proves $w_h^{n,k} \to 0$, and so we have $u_h^n(q_{h,\tau}^k) \to u_h^n(q_{h,\tau}^*)$ as $k \to \infty$.

The rest of the proof is basically the same as that of Theorem 2.2. □

The remaining part of this section is devoted to one of the central issues of our interest: Will the discrete minimizer of the system (5.8)–(5.10) converge to the global minimizer of the continuous problem (2.1)–(2.3)? If yes, is the convergence only weak or can it be strong in some norm? To answer this question, we need some preparations.

For a given function $f \in C([0, T]; X)$, with X being a Banach space, we define a step function approximation, based on the time partition (5.2):

$$S_\Delta f(x, t) = \sum_{n=1}^M \chi_n(t) f(x, t_n). \tag{5.19}$$

We know (cf. [21]) that

$$\lim_{\tau \to 0} \int_0^T \|S_\Delta f(\cdot, t) - f(\cdot, t)\|_X^2 dt = 0. \tag{5.20}$$

Next, we shall demonstrate a most important and technical result in the paper: for any weakly convergent sequence $q_{h,\tau}$ in $L^2(0, T; L^2(\Gamma_i))$ with respect to h and τ, the corresponding finite element solution $u_h^n(q_{h,\tau})$ defined in (5.9)–(5.10) will converge strongly in $L^2(0, T; L^2(\omega))$. More accurately, we have the following lemma.

LEMMA 5.5. *If $q_{h,\tau}$ converges to some q weakly in $L^2(0, T; L^2(\Gamma_i))$ as h and τ tend to* 0, *then*

$$\tau \sum_{n=0}^M \alpha_n \int_\omega (u_h^n(q_{h,\tau}) - z^n)^2 dx \to \int_0^T \int_\omega (u(q) - z)^2 dx dt.$$

Proof. For $1 \le n \le M$, we shall use the following notation:

$$u_h^n = u_h^n(q_{h,\tau}), \quad u^n = u(q)(\cdot, t_n).$$

By (5.20), we can directly verify

$$\lim_{\tau\to 0} \tau \sum_{n=0}^{M} \alpha_n \int_\omega (u^n - z^n)^2 dx = \int_0^T \int_\omega (u(q) - z)^2 dx dt.$$

Therefore it suffices to show

$$\lim_{\substack{h\to 0\\ \tau\to 0}} \tau \sum_{n=0}^{M} \alpha_n \int_\omega (u_h^n - z^n)^2 dx = \lim_{\substack{h\to 0\\ \tau\to 0}} \tau \sum_{n=0}^{M} \alpha_n \int_\omega (u^n - z^n)^2 dx$$

or, equivalently,

$$\lim_{\substack{h\to 0\\ \tau\to 0}} \tau \sum_{n=0}^{M} \int_\omega (u_h^n - u^n)^2 dx = 0. \tag{5.21}$$

For this, we construct two interpolations based on $\{u_h^n\}$: the first one is the piecewise linear interpolation over the time partition (5.2),

$$u_{h,\tau}(x,t) = \frac{t - t_{n-1}}{\tau} u_h^n + \frac{t_n - t}{\tau} u_h^{n-1}, \qquad t \in (t_{n-1}, t_n),$$

while the second one is the piecewise constant interpolation

$$\tilde{u}_{h,\tau}(x,t) = \sum_{n=1}^{M} \chi_n(t) u_h^n(x).$$

By straightforward computations, we have

$$\|\tilde{u}_{h,\tau}\|^2_{L^2(0,T;H^1(\Omega))} = \tau \sum_{n=1}^{M} \|u_h^n\|_1^2, \quad \left\| \frac{\partial}{\partial t} u_{h,\tau} \right\|^2_{L^2(0,T;(H^1(\Omega))')} = \tau \sum_{n=1}^{M} \|\partial_\tau u_h^n\|^2_{(H^1(\Omega))'}$$

and

$$\begin{aligned} &\|u_{h,\tau}\|^2_{L^2(0,T;H^1(\Omega))} \\ &= \frac{\tau}{3} \sum_{n=1}^{M} \int_\Omega (|u_h^n|^2 + |u_h^{n-1}|^2 + u_h^n u_h^{n-1} + |\nabla u_h^n|^2 + |\nabla u_h^{n-1}|^2 + \nabla u_h^n \cdot \nabla u_h^{n-1}) dx \\ &\le \tau \sum_{n=0}^{M} \|u_h^n\|_1^2. \end{aligned}$$

These, together with the stability estimates (5.14)–(5.16), indicate that both $\{u_{h,\tau}\}$ and $\{\tilde{u}_{h,\tau}\}$ are bounded in $L^2(0,T;H^1(\Omega))$ and $\{\frac{\partial}{\partial t} u_{h,\tau}\}$ is bounded in $L^2(0,T;(H^1(\Omega))')$. So by Lemma 2.1 there exist a subsequence of $\{u_{h,\tau}\}$ such that

$$u_{h,\tau} \to u^* \text{ weakly in } L^2(0,T;H^1(\Omega)) \text{ and strongly in } L^2(0,T;L^2(\Omega)), \tag{5.22}$$

$$\frac{\partial}{\partial t} u_{h,\tau} \to v^* \text{ weakly in } L^2(0,T;(H^1(\Omega))'), \tag{5.23}$$

and a subsequence of $\{\tilde{u}_{h,\tau}\}$ such that

$$\tilde{u}_{h,\tau} \to \tilde{u}^* \text{ weakly in } L^2(0,T;H^1(\Omega)) \tag{5.24}$$

for some $u^*, \tilde{u}^* \in L^2(0,T;H^1(\Omega))$ and $v^* \in L^2(0,T;(H^1(\Omega))')$.

From (5.23), we know for any $\varphi(x) \in H^1(\Omega)$ and $\psi(t) \in C_0^\infty(0,T)$,

$$\lim_{\substack{h\to 0\\ \tau\to 0}} \int_0^T\int_\Omega \frac{\partial u_{h,\tau}(x,t)}{\partial t}\varphi(x)\psi(t)dxdt = \int_0^T\int_\Omega v^*(x,t)\varphi(x)\psi(t)dxdt. \tag{5.25}$$

Integrating by parts the left-hand side and using (5.22), we obtain

$$-\int_0^T\int_\Omega u^*(x,t)\varphi(x)\psi'(t)dxdt = \int_0^T\int_\Omega v^*(x,t)\varphi(x)\psi(t)dxdt,$$

which gives

$$v^*(x,t) = \frac{\partial u^*(x,t)}{\partial t}. \tag{5.26}$$

Next, taking any $\varphi(x) \in H^1(\Omega)$ and $\psi(t) \in C^1[0,T]$ with $\psi(T) = 0$, integrating by parts to both sides of (5.25), and noting (5.26), we get

$$\begin{aligned}&\lim_{\substack{h\to 0\\ \tau\to 0}} \left\{-\int_\Omega Q_h u_0(x)\varphi(x)\psi(0)dx - \int_0^T\int_\Omega u_{h,\tau}(x,t)\varphi(x)\psi'(t)dxdt\right\}\\ &= -\int_\Omega u^*(x,0)\varphi(x)\psi(0)dx - \int_0^T\int_\Omega u^*(x,t)\varphi(x)\psi'(t)dxdt.\end{aligned}$$

By the convergence property of Q_h and (5.22) we obtain

$$u^*(x,0) = u_0(x). \tag{5.27}$$

Next, we show $u^*(x,t) = \tilde{u}^*(x,t)$. In fact, by direct computing and (5.15), we obtain

$$\int_0^T \|u_{h,\tau}(\cdot,t) - \tilde{u}_{h,\tau}(\cdot,t)\|_0^2 dt = \frac{\tau^3}{3}\sum_{n=1}^M \|\partial_\tau u_h^n\|_0^2 \le C\tau; \tag{5.28}$$

this with (5.22) proves that $\tilde{u}_{h,\tau}$ converges to $\tilde{u}^*$ strongly in $L^2(0,T;L^2(\Omega))$ and $u^*(x,t) = \tilde{u}^*(x,t)$.

Below we will show $u^* = u(q)$. For any $\varphi(x) \in H^1(\Omega)$ and $\psi(t) \in C_0^\infty(0,T)$, let $\phi(x,t) = \varphi(x)\psi(t)$ and $\phi_{h,\tau}(x,t) = \sum_{n=1}^M \chi_n(t)Q_h\phi(x,t_n)$. Then we have

$$\begin{aligned}&\int_0^T \|\phi(\cdot,t) - \phi_{h,\tau}(\cdot,t)\|_1^2 dt\\ &\le 2\int_0^T \|\phi(\cdot,t) - S_\Delta\phi(\cdot,t)\|_1^2 dt + 2\int_0^T \|S_\Delta\phi(\cdot,t) - \phi_{h,\tau}(\cdot,t)\|_1^2 dt\\ &\le 2\int_0^T \|\phi(\cdot,t) - S_\Delta\phi(\cdot,t)\|_1^2 dt + 2\,T\max_{0\le t\le T}|\psi(t)|^2\,\|Q_h\varphi(\cdot) - \varphi(\cdot)\|_1^2.\end{aligned}$$

Therefore by (5.20) and the convergence property of Q_h, we deduce

$$\phi_{h,\tau} \text{ converges to } \phi \text{ strongly in } L^2(0,T;H^1(\Omega)). \tag{5.29}$$

By direct computations we have the following equalities:

$$\int_0^T\int_\Omega \frac{\partial}{\partial t}u_{h,\tau}(x,t)\phi_{h,\tau}(x,t)dxdt = \tau\sum_{n=1}^M\int_\Omega \partial_\tau u_h^n Q_h\phi(x,t_n)dx,$$

$$\int_0^T\int_\Omega \alpha(x,t)\nabla\tilde{u}_{h,\tau}(x,t)\nabla\phi_{h,\tau}(x,t)dxdt = \tau\sum_{n=1}^M\int_\Omega \bar{\alpha}^n\nabla u_h^n\nabla Q_h\phi(x,t_n)dx,$$

$$\int_0^T\int_{\Gamma_o} c(x,t)\tilde{u}_{h,\tau}(x,t)\phi_{h,\tau}(x,t)dsdt = \tau\sum_{n=1}^M\int_{\Gamma_o} \bar{c}^n u_h^n Q_h\phi(x,t_n)ds,$$

$$-\int_0^T\int_{\Gamma_o} c(x,t)u_a(x,t)\phi_{h,\tau}(x,t)dsdt = -\tau\sum_{n=1}^M\int_{\Gamma_o} \overline{cu}_a^n Q_h\phi(x,t_n)ds,$$

$$\int_0^T\int_{\Gamma_i} q_{h,\tau}(x,t)\phi_{h,\tau}(x,t)dsdt = \tau\sum_{n=1}^M\int_{\Gamma_i} q_h^n Q_h\phi(x,t_n)ds;$$

adding them together and using the discrete parabolic equation (5.10), we obtain

$$\begin{aligned}&\int_0^T\int_\Omega \frac{\partial}{\partial t}u_{h,\tau}(x,t)\phi_{h,\tau}(x,t)dxdt + \int_0^T\int_\Omega \alpha(x,t)\nabla\tilde{u}_{h,\tau}(x,t)\nabla\phi_{h,\tau}(x,t)dxdt\\ &+\int_0^T\int_{\Gamma_o} c(x,t)\tilde{u}_{h,\tau}(x,t)\phi_{h,\tau}(x,t)dsdt - \int_0^T\int_{\Gamma_o} c(x,t)u_a(x,t)\phi_{h,\tau}(x,t)dsdt\\ &= -\int_0^T\int_{\Gamma_i} q_{h,\tau}(x,t)\phi_{h,\tau}(x,t)dsdt + \tau\sum_{n=1}^M\int_{\Gamma_o}(\bar{c}^n\bar{u}_a^n - \overline{cu}_a^n)Q_h\phi(x,t_n)ds.\end{aligned}\tag{5.30}$$

Taking the limit as h and τ tend to 0 and using the convergence (5.22)–(5.24) and (5.29), we derive that for any $\varphi(x)\in H^1(\Omega)$ and $\psi(t)\in C_0^\infty(0,T)$

$$\begin{aligned}&\int_0^T\int_\Omega \frac{\partial u^*}{\partial t}\varphi(x)\psi(t)dxdt + \int_0^T\int_\Omega \alpha\nabla u^*\cdot\nabla\varphi(x)\psi(t)dxdt + \int_0^T\int_{\Gamma_o} c\,u^*\varphi(x)\psi(t)dsdt\\ &= \int_0^T\int_{\Gamma_o} c\,u_a\varphi(x)\psi(t)dsdt - \int_0^T\int_{\Gamma_i} q\,\varphi(x)\psi(t)dsdt,\end{aligned}\tag{5.31}$$

where we have used the limit

$$\lim_{\substack{h\to 0\\ \tau\to 0}}\tau\sum_{n=1}^M\int_{\Gamma_o}(\bar{c}^n\bar{u}_a^n - \overline{cu}_a^n)Q_h\phi(x,t_n)ds = 0.\tag{5.32}$$

To see this, it follows from (5.13), the trace theorem, and the Cauchy–Schwarz inequality that

$$\begin{aligned}&\tau\sum_{n=1}^M\int_{\Gamma_o}(\bar{c}^n\bar{u}_a^n - \overline{cu}_a^n)Q_h\phi(x,t_n)ds\\ &\le C\,\tau\max_{0\le t\le T}|\psi(t)|\,\|Q_h\varphi\|_1\sum_{n=1}^M\left\|\bar{c}^n\bar{u}_a^n - \overline{cu}_a^n\right\|_{0,\Gamma_o}\\ &\le C\,\tau\max_{0\le t\le T}|\psi(t)|\,\|\varphi\|_1\left(\sum_{n=1}^M\|c_t\|^2_{L^2(t_{n-1},t_n;L^\infty(\Gamma_o))}\right)^{\frac12}\left(\sum_{n=1}^M\|u_a\|^2_{L^2(t_{n-1},t_n;L^2(\Gamma_o))}\right)^{\frac12}\\ &\le C\,\tau\max_{0\le t\le T}|\psi(t)|\,\|\varphi\|_1\,\|c_t\|_{L^2(0,T;L^\infty(\Gamma_o))}\,\|u_a\|_{L^2(0,T;L^2(\Gamma_o))}.\end{aligned}$$

Clearly, the fact that $u^* = u(q)$ follows then from (5.31).

Now we can show the desired relation (5.21). For this, setting $f(x,t) = u_{h,\tau}(x,t) - u(x,t)$, we can write and estimate using Lemma 5.2 as follows:

$$\begin{aligned}
&\tau \sum_{n=1}^{M} \int_\Omega (u_h^n - u^n)^2 dx - \int_0^T \|u_{h,\tau}(\cdot,t) - u(\cdot,t)\|_0^2 dt \\
&= \sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \int_\Omega \Big(|f(x,t_n)|^2 - |f(x,t)|^2 \Big) dx dt \\
&\le \left\{ \sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \|f(\cdot,t_n) + f(\cdot,t)\|_0^2 dt \right\}^{\frac{1}{2}} \left\{ \sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \|f(\cdot,t_n) - f(\cdot,t)\|_0^2 dt \right\}^{\frac{1}{2}} \\
&\le C \left\{ \sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \|f(\cdot,t_n) - f(\cdot,t)\|_0^2 dt \right\}^{\frac{1}{2}}.
\end{aligned}$$

By (5.22), the second term at the left-hand side of the above inequality tends to 0 as $h, \tau \to 0$. But the last term can be estimated as follows:

$$\begin{aligned}
&\sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \|f(\cdot,t_n) - f(\cdot,t)\|_0^2 dt \\
&= \sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \|u - u^n + (t_n - t)\partial_\tau u_h^n\|_0^2 dt \\
&\le 2 \sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \|u - u^n\|_0^2 dt + 2 \sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \int_\Omega (t_n - t)^2 |\partial_\tau u_h^n|^2 dx dt \\
&= 2 \sum_{n=1}^{M} \int_{t_{n-1}}^{t_n} \|u - u^n\|_0^2 dt + \frac{2}{3}\tau^3 \sum_{n=1}^{M} \|\partial_\tau u_h^n\|_0^2.
\end{aligned}$$

From (5.20) and (5.15), the last two terms both tend to 0, and (5.21) follows. □

Finally, we are ready to show the main convergence results of this section.

THEOREM 5.6. *Let $\{q_{h,\tau}^*\}$ be a sequence of minimizers to the finite element minimization problem* (5.8)–(5.10)*; then as h and τ tend to* 0*, the whole sequence $\{q_{h,\tau}^*\}$ converges strongly in $L^2(0,T;L^2(\Gamma_i))$ to the unique minimizer of the continuous problem* (2.1)–(2.3).

Proof. Using the stability estimate (5.14), it is easy to know that $J_{h,\tau}(q_{h,\tau}^*) \le C$ for some constant C independent of h and τ. This implies that $\{q_{h,\tau}^*\}$ is bounded in $L^2(0,T;L^2(\Gamma_i))$ and there exists a subsequence of $\{q_{h,\tau}^*\}$, still denoted as $\{q_{h,\tau}^*\}$, such that $q_{h,\tau}^* \to q^*$ weakly in $L^2(0,T;L^2(\Gamma_i))$ as $h, \tau \to 0$.

Now for any $q \in L^2(0,T;L^2(\Gamma_i))$ and any fixed $\varepsilon > 0$, by the density results there exists a $q_\varepsilon \in H^1(0,T;H^{1/2}(\Gamma_i))$ such that

$$\|q - q_\varepsilon\|_{L^2(0,T;L^2(\Gamma_i))} \le \varepsilon.$$

Then we define an extension $\tilde{q}_\varepsilon$ of q_ε as follows: $\tilde{q}_\varepsilon \in H^1(\Omega)$ solves

$$-\Delta \tilde{q}_\varepsilon = 0 \;\text{ in }\; \Omega, \quad \tilde{q}_\varepsilon = q_\varepsilon \;\text{ on }\; \Gamma_i, \quad \tilde{q}_\varepsilon = 0 \;\text{ on }\; \Gamma_o.$$

One can verify that $\tilde{q}_\varepsilon \in H^1(0,T;H^1(\Omega))$ and $\|\tilde{q}_\varepsilon\|_{H^1(0,T;H^1(\Omega))} \le C\|q_\varepsilon\|_{H^1(0,T;H^{1/2}(\Gamma_i))}$. Define

$$\tilde{q}_\varepsilon^{h,\tau}(x,t) = \sum_{n=1}^{M} \chi_n(t) Q_h \tilde{q}_\varepsilon(x,t_n).$$

Let $q_\varepsilon^{h,\tau}$ be the restriction of $\tilde{q}_\varepsilon^{h,\tau}$ on Γ_i; then $q_\varepsilon^{h,\tau} \in V_{\Gamma_i}^h$ and for any $\varepsilon > 0$,

$$\begin{aligned}
&\|q_\varepsilon^{h,\tau} - q_\varepsilon\|^2_{L^2(0,T;L^2(\Gamma_i))} \le \|q_\varepsilon^{h,\tau} - q_\varepsilon\|^2_{L^2(0,T;H^{1/2}(\Gamma_i))} \le C\|\tilde{q}_\varepsilon^{h,\tau} - \tilde{q}_\varepsilon\|^2_{L^2(0,T;H^1(\Omega))} \\
&= C\sum_{n=1}^{M}\int_{t_{n-1}}^{t_n} \|Q_h\tilde{q}_\varepsilon(\cdot,t_n) - \tilde{q}_\varepsilon(\cdot,t)\|_1^2 dt \\
&\le C\sum_{n=1}^{M}\int_{t_{n-1}}^{t_n} \|Q_h\tilde{q}_\varepsilon(\cdot,t_n) - Q_h\tilde{q}_\varepsilon(\cdot,t) + Q_h\tilde{q}_\varepsilon(\cdot,t) - \tilde{q}_\varepsilon(\cdot,t)\|_1^2 dt \\
&\le C\int_0^T \|S_\Delta\tilde{q}_\varepsilon(\cdot,t) - \tilde{q}_\varepsilon(\cdot,t)\|_1^2 dt + C\int_0^T \|Q_h\tilde{q}_\varepsilon(\cdot,t) - \tilde{q}_\varepsilon(\cdot,t)\|_1^2 dt.
\end{aligned}$$

Thus $q_\varepsilon^{h,\tau} \to q_\varepsilon$ in $L^2(0,T;L^2(\Gamma_i))$ as $h,\tau \to 0$. Using this and Lemma 5.5, we can derive

$$\begin{aligned}
J(q^*) &\le \lim_{\substack{h\to 0\\ \tau\to 0}} \frac{\tau}{2}\sum_{n=0}^{M}\alpha_n \int_\omega (u_h^n(q_{h,\tau}^*) - z^n)^2 dx + \frac{\beta}{2}\liminf_{\substack{h\to 0\\ \tau\to 0}} \int_0^T\int_{\Gamma_i} |q_{h,\tau}^*|^2 ds dt \\
&\le \liminf_{\substack{h\to 0\\ \tau\to 0}} J_{h,\tau}(q_{h,\tau}^*) \le \liminf_{\substack{h\to 0\\ \tau\to 0}} J_{h,\tau}(q_\varepsilon^{h,\tau}) \\
&= \frac{1}{2}\int_0^T\int_\omega (u(q_\varepsilon) - z)^2 dx dt + \frac{\beta}{2}\int_0^T\int_{\Gamma_i} q_\varepsilon^2 ds dt \\
&= J(q_\varepsilon).
\end{aligned}$$

Letting $\varepsilon \to 0$, we deduce

$$J(q^*) \le J(q) \qquad \forall\ q \in L^2(0,T;L^2(\Gamma_i)), \tag{5.33}$$

which indicates that q^* is the unique minimizer of the continuous problem (2.1)–(2.3).

The strong convergence follows by the same trick as used in Proposition 2.3. □

Remark 5.7. All the results obtained in this paper can be naturally extended to a three-dimensional domain Ω with every two-dimensional cross-section being the domain as in Figure 2.

6. Finite element approximation of system (4.1)–(4.3) and its convergence. Next, we shall discuss the discretization of system (4.1)–(4.3). As we did for system (2.1)–(2.3), we use the composite trapezoidal rule for the time discretization of the first integral in (4.1) and the exact time integration for the second term. But as the time derivative of the identifying parameter $q(x,t)$ is involved in the regularization term now, we cannot ensure the convergence of the resultant fully discrete scheme for the entire system (4.1)–(4.3) if the backward Euler scheme is still used for approximating the parabolic problem (4.3). Instead we shall adopt the Crank–Nicolson scheme. This results in the following finite element approximation of (4.1)–(4.3):

(6.1)

$$\min J_{h,\tau}(q_{h,\tau}) = \frac{\tau}{2}\sum_{n=0}^{M}\alpha_n\int_\omega (u_h^n - z^n)^2 dx + \frac{\beta}{2}\left(\int_{\Gamma_i} |q_h^0|^2 ds + \tau\sum_{n=1}^{M}\int_{\Gamma_i} |\partial_\tau q_h^n|^2 ds\right)$$

over all $q_h^n \in V_{\Gamma_i}^h$ with $u_h^n \equiv u_h^n(q_{h,\tau}) \in V^h$ satisfying $u_h^0 = Q_h u_0$ in Ω and

$$\int_\Omega \partial_\tau u_h^n \phi_h dx + \int_\Omega \bar{\alpha}^n \nabla \frac{u_h^n + u_h^{n-1}}{2} \cdot \nabla \phi_h dx + \int_{\Gamma_o} \bar{c}^n \frac{u_h^n + u_h^{n-1}}{2} \phi_h ds$$
$$= \int_{\Gamma_o} \bar{c}^n \bar{u}_a^n \phi_h ds - \int_{\Gamma_i} \frac{q_h^n + q_h^{n-1}}{2} \phi_h ds \quad \forall \phi_h \in V^h \tag{6.2}$$

for $n = 1, 2, \ldots, M$. Here $\{\alpha_n\}$ are the coefficients of the composite trapezoidal rule: $\alpha_0 = \alpha_M = \frac{1}{2}$ and $\alpha_n = 1$ for all $n \neq 0, M$. The heat flux q is approximated by $q_{h,\tau}$, a piecewise linear interpolation based on $\{q_h^n\}$ over the time partition Δ in (5.2):

$$q_{h,\tau}(x,t) = \frac{t - t_{n-1}}{\tau} q_h^n + \frac{t_n - t}{\tau} q_h^{n-1}, \qquad t \in (t_{n-1}, t_n). \tag{6.3}$$

For the fully discrete finite element scheme (6.1)–(6.2), we can show (cf. [14]) the following theorem.

Theorem 6.1. *There exists a unique minimizer to the finite element problem* (6.1)*–*(6.2)*.*

In the rest of this section, we study the convergence of the discrete minimizer of (6.1)–(6.2) to the global minimizer of the continuous problem (4.1)–(4.3). For this purpose, we assume on functions $\alpha(x,t)$, $c(x,t)$, and $u_a(x,t)$ in (1.1)–(1.4) that

$$\alpha \in W^{1,\infty}(0,T; L^\infty(\Omega)) \quad \text{and} \quad c,\ u_a \in W^{1,\infty}(0,T; L^\infty(\Gamma_o)) \tag{6.4}$$

and introduce three related constants:

$$C_1 = \|\alpha\|_{W^{1,\infty}(0,T;L^\infty(\Omega))}, \quad C_2 = \|c\|_{W^{1,\infty}(0,T;L^\infty(\Gamma_o))}, \quad C_3 = \|u_a\|_{W^{1,\infty}(0,T;L^\infty(\Gamma_o))}. \tag{6.5}$$

Using these constants, we can derive the following estimates (cf. [14]):

$$\|\bar{\alpha}^n - \bar{\alpha}^{n-1}\|_{L^\infty(\Omega)} \le C_1\, \tau, \quad \|\bar{c}^n \bar{u}_a^n - \bar{c}^{n-1} \bar{u}_a^{n-1}\|_{L^\infty(\Gamma_o)} \le \frac{5}{3} C_2 C_3\, \tau. \tag{6.6}$$

For the convergence analysis, we first establish some stability estimates of the finite element solution to (6.2).

Lemma 6.2. *Let u_h^n be the finite element solution of system* (6.2) *corresponding to the given heat flux $\{q_h^n\}_{n=0}^M$; then we have the following stability estimates:*

$$\max_{1\le n\le M} \|u_h^n\|_0^2 + \tau \sum_{n=1}^M \left\| \nabla \frac{u_h^n + u_h^{n-1}}{2} \right\|_0^2 + \tau \sum_{n=1}^M \left\| \frac{u_h^n + u_h^{n-1}}{2} \right\|_{0,\Gamma_o}^2$$
$$\le C \left(\|u_0\|_0^2 + C_2^2 C_3^2 + \tau \sum_{n=0}^M \|q_h^n\|_{0,\Gamma_i}^2 \right), \tag{6.7}$$

$$\tau \sum_{n=1}^M \|\partial_\tau u_h^n\|_0^2 + \max_{1\le n\le M} \|\nabla u_h^n\|_0^2 + \max_{1\le n\le M} \|u_h^n\|_{0,\Gamma_o}^2$$
$$\le C \left(\|u_0\|_1^2 + C_2^2 C_3^2 + \max_{1\le n\le k} \|q_h^n\|_{0,\Gamma_i}^2 + \tau \sum_{n=1}^M \|\partial_\tau q_h^n\|_{0,\Gamma_i}^2 + \tau \sum_{n=0}^M \|q_h^n\|_{0,\Gamma_i}^2 \right). \tag{6.8}$$

Proof. Taking $\phi_h = \tau \frac{u_h^n + u_h^{n-1}}{2}$ in (6.2), we have

$$\frac{1}{2}\|u_h^n\|_0^2 - \frac{1}{2}\|u_h^{n-1}\|_0^2 + \alpha_0 \tau \left\| \nabla \frac{u_h^n + u_h^{n-1}}{2} \right\|_0^2 + c_0 \tau \left\| \frac{u_h^n + u_h^{n-1}}{2} \right\|_{0,\Gamma_o}^2$$
$$\leq \tau \int_{\Gamma_o} \bar{c}^n \bar{u}_a^n \frac{u_h^n + u_h^{n-1}}{2} ds - \tau \int_{\Gamma_i} \frac{q_h^n + q_h^{n-1}}{2} \frac{u_h^n + u_h^{n-1}}{2} ds.$$

Summing up the above equation over $n = 1, 2, \ldots, k \leq M$, we derive

$$\frac{1}{2}\|u_h^k\|_0^2 - \frac{1}{2}\|u_h^0\|_0^2 + \alpha_0 \tau \sum_{n=1}^k \left\| \nabla \frac{u_h^n + u_h^{n-1}}{2} \right\|_0^2 + c_0 \tau \sum_{n=1}^k \left\| \frac{u_h^n + u_h^{n-1}}{2} \right\|_{0,\Gamma_o}^2$$
$$\leq \tau \sum_{n=1}^k \int_{\Gamma_o} \bar{c}^n \bar{u}_a^n \frac{u_h^n + u_h^{n-1}}{2} ds - \tau \sum_{n=1}^k \int_{\Gamma_i} \frac{q_h^n + q_h^{n-1}}{2} \frac{u_h^n + u_h^{n-1}}{2} ds;$$

then (6.7) follows by applying the trace theorem and Young's and Gronwall's inequalities.

Next, taking $\phi_h = \tau \partial_\tau u_h^n$ in (6.2), we have

$$\tau \|\partial_\tau u_h^n\|_0^2 + \frac{1}{2} \int_\Omega \bar{\alpha}^n (|\nabla u_h^n|^2 - |\nabla u_h^{n-1}|^2) dx + \frac{1}{2} \int_{\Gamma_o} \bar{c}^n (|u_h^n|^2 - |u_h^{n-1}|^2) ds$$
$$= \int_{\Gamma_o} \bar{c}^n \bar{u}_a^n (u_h^n - u_h^{n-1}) ds - \int_{\Gamma_i} \frac{q_h^n + q_h^{n-1}}{2} (u_h^n - u_h^{n-1}) ds.$$

Summing up the above equation over $n = 1, 2, \ldots, k \leq M$ and using the formula (5.17), we deduce

$$\tau \sum_{n=1}^k \|\partial_\tau u_h^n\|_0^2 + \frac{1}{2} \alpha_0 \|\nabla u_h^k\|_0^2 + \frac{1}{2} c_0 \|u_h^k\|_{0,\Gamma_o}^2$$
$$\leq \frac{1}{2} \int_\Omega \bar{\alpha}^0 |\nabla u_h^0|^2 dx + \frac{1}{2} \sum_{n=1}^k \int_\Omega (\bar{\alpha}^n - \bar{\alpha}^{n-1}) |\nabla u_h^{n-1}|^2 dx$$
$$+ \frac{1}{2} \int_{\Gamma_o} \bar{c}^0 |u_h^0|^2 ds + \frac{1}{2} \sum_{n=1}^k \int_{\Gamma_o} (\bar{c}^n - \bar{c}^{n-1}) |u_h^{n-1}|^2 ds$$
$$+ \int_{\Gamma_o} \bar{c}^k \bar{u}_a^k u_h^k ds - \int_{\Gamma_o} \bar{c}^0 \bar{u}_a^0 u_h^0 ds - \sum_{n=1}^k \int_{\Gamma_o} (\bar{c}^n \bar{u}_a^n - \bar{c}^{n-1} \bar{u}_a^{n-1}) u_h^{n-1} ds$$
$$- \frac{1}{2} \int_{\Gamma_i} (q_h^k + q_h^{k-1}) u_h^k ds + \frac{1}{2} \int_{\Gamma_i} q_h^0 u_h^0 ds$$
$$+ \sum_{n=1}^k \int_{\Gamma_i} \left(\frac{q_h^n + q_h^{n-1}}{2} - \frac{q_h^{n-1} + q_h^{n-2}}{2} \right) u_h^{n-1} ds.$$

We now estimate all the terms on the right-hand side above. First, for those terms without summation, we can easily deduce

$$\int_\Omega \bar{\alpha}^0 |\nabla u_h^0|^2 dx + \int_{\Gamma_o} \bar{c}^0 |u_h^0|^2 ds \le C\,(C_1 + C_2)\|u_0\|_1^2,$$

$$\int_{\Gamma_o} \bar{c}^0 \bar{u}_a^0 u_h^0 ds + \frac{1}{2}\int_{\Gamma_i} q_h^0 u_h^0 ds \le C\,(C_2 C_3 + \|q_h^0\|_{0,\Gamma_i})\|u_0\|_1,$$

$$\int_{\Gamma_o} \bar{c}^k \bar{u}_a^k u_h^k ds \le \frac{1}{4} c_0 \|u_h^k\|_{0,\Gamma_o}^2 + C\,C_2^2 C_3^2,$$

$$-\frac{1}{2}\int_{\Gamma_i} (q_h^k + q_h^{k-1}) u_h^k ds \le \frac{1}{4}\alpha_0 \|\nabla u_h^k\|_0^2 + C\Big(\max_{1\le n\le M} \|u_h^n\|_0^2 + \max_{1\le n\le M} \|q_h^n\|_{0,\Gamma_i}^2\Big).$$

Using (6.6), we obtain the following estimates:

$$\frac{1}{2}\sum_{n=1}^k \int_\Omega (\bar{\alpha}^n - \bar{\alpha}^{n-1}) |\nabla u_h^{n-1}|^2 dx \le \frac{1}{2} C_1\,\tau \sum_{n=1}^k \|\nabla u_h^{n-1}\|_0^2,$$

$$\frac{1}{2}\sum_{n=1}^k \int_{\Gamma_o} (\bar{c}^n - \bar{c}^{n-1}) |u_h^{n-1}|^2 ds \le \frac{1}{2} C_2\,\tau \sum_{n=1}^k \|u_h^{n-1}\|_{0,\Gamma_o}^2,$$

$$-\sum_{n=1}^k \int_{\Gamma_o} (\bar{c}^n \bar{u}_a^n - \bar{c}^{n-1}\bar{u}_a^{n-1}) u_h^{n-1} ds \le C\,\tau \sum_{n=1}^k \|u_h^{n-1}\|_{0,\Gamma_o}^2 + C.$$

For the last term, we use the Cauchy–Schwarz inequality to obtain

$$\sum_{n=1}^k \int_{\Gamma_i} \left(\frac{q_h^n + q_h^{n-1}}{2} - \frac{q_h^{n-1} + q_h^{n-2}}{2} \right) u_h^{n-1} ds$$

$$\le \frac{1}{2}\tau \sum_{n=1}^k \|u_h^{n-1}\|_{0,\Gamma_i}^2 + \frac{1}{8\tau}\sum_{n=1}^k \left\| (q_h^n + q_h^{n-1}) - (q_h^{n-1} + q_h^{n-2}) \right\|_{0,\Gamma_i}^2$$

$$\le C\,\tau \sum_{n=1}^k \Big(\|\nabla u_h^{n-1}\|_0^2 + \|u_h^{n-1}\|_0^2 + \|\partial_\tau q_h^n\|_{0,\Gamma_i}^2 \Big).$$

Now (6.8) follows by combining all of the above estimates and using Gronwall's inequality. □

As we did for the finite element system (5.8)–(5.10), we need the following crucial technical result for the convergence of the finite element approximation (6.1)–(6.2).

LEMMA 6.3. *If $q_{h,\tau}$ converges to some q weakly in $H^1(0,T;L^2(\Gamma_i))$ as h and τ tend to* 0*, then*

$$\lim_{\substack{h\to 0\\ \tau\to 0}} \tau \sum_{n=0}^M \alpha_n \int_\omega (u_h^n(q_h^n) - z^n)^2 dx = \int_0^T\!\!\int_\omega (u(q) - z)^2 dx dt.$$

Proof. As in the proof of Lemma 5.5, it suffices to show (5.21).

We first construct two interpolations based on $\{u_h^n\}$: one is a piecewise linear interpolation over the time partition Δ,

$$u_{h,\tau}(x,t) = \frac{t - t_{n-1}}{\tau} u_h^n + \frac{t_n - t}{\tau} u_h^{n-1}, \quad t \in (t_{n-1}, t_n) \quad \text{for} \;\; n = 1, 2, \ldots, M,$$

and the other is a piecewise constant interpolation,

$$\tilde{u}_{h,\tau}(x,t) = \frac{1}{2}(u_h^n + u_h^{n-1}), \quad t \in (t_{n-1}, t_n) \quad \text{for} \quad n = 1, 2, \dots, M.$$

Using the definition of $q_{h,\tau}$ in (6.3) and the simple identity

$$q_h^n = \tau \sum_{k=1}^{n} \partial_\tau q_h^k + q_h^0,$$

we can directly see that

$$\tau \sum_{n=1}^{M} \|\partial_\tau q_h^n\|_{0,\Gamma_i}^2 = \tau \sum_{n=1}^{M} \left\| \frac{\partial}{\partial t} q_{h,\tau} \right\|_{0,\Gamma_i}^2 = \left\| \frac{\partial}{\partial t} q_{h,\tau} \right\|_{L^2(0,T;L^2(\Gamma_i))}^2,$$

$$\|q_h^n\|_{0,\Gamma_i}^2 \le 2\tau T \sum_{k=1}^{n} \|\partial_\tau q_h^k\|_{0,\Gamma_i}^2 + 2\|q_h^0\|_{0,\Gamma_i}^2.$$

With these relations, the assumption on $q_{h,\tau}$, and the stability estimates (6.7)–(6.8), we can easily check that both $\{u_{h,\tau}\}$ and $\{\tilde{u}_{h,\tau}\}$ are bounded in $L^2(0,T;H^1(\Omega))$ and that $\{\frac{\partial}{\partial t} u_{h,\tau}\}$ is bounded in $L^2(0,T;L^2(\Omega))$. So there exist a subsequence $\{u_{h,\tau}\}$ such that

$$u_{h,\tau} \to u^* \text{ weakly in } L^2(0,T;H^1(\Omega)) \text{ and strongly in } L^2(0,T;L^2(\Omega)),$$

$$\frac{\partial}{\partial t} u_{h,\tau} \to \frac{\partial u^*}{\partial t} \text{ weakly in } L^2(0,T;L^2(\Omega)),$$

and a subsequence $\{\tilde{u}_{h,\tau}\}$ such that

$$\tilde{u}_{h,\tau} \to \tilde{u}^* \text{ weakly in } L^2(0,T;H^1(\Omega)) \tag{6.9}$$

for some $u^* \in H^1(0,T;L^2(\Omega))$ and $\tilde{u}^* \in L^2(0,T;H^1(\Omega))$. We can further show that $u^*(x,0) = u_0(x)$ and $u^* = \tilde{u}^*$ using the fact that

$$\int_0^T \|u_{h,\tau}(\cdot,t) - \tilde{u}_{h,\tau}(\cdot,t)\|_0^2 dt = \frac{\tau^3}{12} \sum_{n=1}^{M} \|\partial_\tau u_h^n\|_0^2 \to 0. \tag{6.10}$$

Next, we show $u^* = u(q)$. Let $\phi_{h,\tau}(x,t)$ be defined as in the proof of Lemma 5.5. By simple computations we have the following equalities:

$$\int_0^T \int_\Omega \frac{\partial}{\partial t} u_{h,\tau}(x,t) \phi_{h,\tau}(x,t) dx dt = \tau \sum_{n=1}^{M} \int_\Omega \partial_\tau u_h^n Q_h \phi(x,t_n) dx,$$

$$\int_0^T \int_\Omega \alpha(x,t) \nabla \tilde{u}_{h,\tau}(x,t) \nabla \phi_{h,\tau}(x,t) dx dt = \tau \sum_{n=1}^{M} \int_\Omega \bar{\alpha}^n \nabla \frac{u_h^n + u_h^{n-1}}{2} \nabla Q_h \phi(x,t_n) dx,$$

$$\int_0^T \int_{\Gamma_o} c(x,t) \tilde{u}_{h,\tau}(x,t) \phi_{h,\tau}(x,t) ds dt = \tau \sum_{n=1}^{M} \int_{\Gamma_o} \bar{c}^n \frac{u_h^n + u_h^{n-1}}{2} Q_h \phi(x,t_n) ds,$$

$$\int_0^T \int_{\Gamma_o} c(x,t) u_a(x,t) \phi_{h,\tau}(x,t) ds dt = \tau \sum_{n=1}^{M} \int_{\Gamma_o} \overline{cu}_a^n Q_h \phi(x,t_n) ds,$$

$$\int_0^T \int_{\Gamma_i} q_{h,\tau}(x,t) \phi_{h,\tau}(x,t) ds dt = \tau \sum_{n=1}^{M} \int_{\Gamma_i} \frac{q_h^n + q_h^{n-1}}{2} Q_h \phi(x,t_n) ds.$$

Adding them together and using (6.2), we obtain a similar equation to (5.30). The rest of the proof is basically the same as that in Lemma 5.5. □

By virtue of Lemma 6.3, we can show the convergence of the finite element approximation (6.1)–(6.2), following the same lines as for Theorem 5.6; see Xie [14].

THEOREM 6.4. *Let* $\{q^*_{h,\tau}\}$ *be a sequence of minimizers to the discrete minimization problem* (6.1)–(6.2)*; then as* h *and* τ *tend to* 0*, the whole sequence* $\{q^*_{h,\tau}\}$ *converges strongly in* $H^1(0,T;L^2(\Gamma_i))$ *to the unique minimizer of the continuous problem* (4.1)–(4.3).

7. Solutions of finite element minimizaton problems.

In this section, we shall formulate a conjugate gradient algorithm to solve the nonlinear finite element minimization problems (5.8)–(5.10) and (6.1)–(6.2). We present details only for system (6.1)–(6.2), while the algorithm for system (5.8)–(5.10) can be formulated similarly; for details of the latter system, we refer to Xie [14].

We first derive the Gateaux derivative of the cost functional $J_{h,\tau}(q_{h,\tau})$ in (6.1), or the form $J_{h,\tau}(\{q_h^0,\dots,q_h^M\})$. Let $N=\dim(V^h_{\Gamma_i})$, and let $\{\psi_i\}_{i=1}^N$ be the basis of $V^h_{\Gamma_i}$. For any element from space $V^h_{\Gamma_i}\times\cdots\times V^h_{\Gamma_i}$, say $\{q_h^0,\dots,q_h^M\}$, let $\mathcal{U}_h^n\equiv u_h^n(q_{h,\tau})'p_{h,\tau}$ be the Gateaux derivative of solution $u_h^n(q_{h,\tau})$ to (6.1)–(6.2) in the direction $p_{h,\tau}$, or $\{p_h^0,\dots,p_h^M\}$. We easily see that $\mathcal{U}_h^0=0$, and for $n=1,2,\dots,M$ and any $\phi_h\in V^h$, the derivative $\mathcal{U}_h^n\in V^h$ satisfies

$$\int_\Omega \partial_\tau\mathcal{U}_h^n\phi_h dx+\int_\Omega \bar\alpha^n\nabla\frac{\mathcal{U}_h^n+\mathcal{U}_h^{n-1}}{2}\cdot\nabla\phi_h dx+\int_{\Gamma_o}\bar c^n\frac{\mathcal{U}_h^n+\mathcal{U}_h^{n-1}}{2}\phi_h ds$$
$$=-\int_{\Gamma_i}\frac{p_h^n+p_h^{n-1}}{2}\phi_h ds.$$

This enables us to derive the first and second derivatives of the cost functional $J_{h,\tau}$ in (6.1):

$$J_{h,\tau}(q_{h,\tau})'p_{h,\tau}=\tau\sum_{n=1}^M\alpha_n\int_\omega(u_h^n-z^n)\mathcal{U}_h^n dx+\beta\left(\int_{\Gamma_i}q_h^0p_h^0ds+\tau\sum_{n=1}^M\int_{\Gamma_i}\partial_\tau q_h^n\,\partial_\tau p_h^n ds\right),\tag{7.1}$$

$$J_{h,\tau}(q_{h,\tau})''p_{h,\tau}r_{h,\tau}=\tau\sum_{n=1}^M\alpha_n\int_\omega\Big(u_h^n(q_{h,\tau})'p_{h,\tau}\Big)\Big(u_h^n(q_{h,\tau})'r_{h,\tau}\Big)dx+\beta\left(\int_{\Gamma_i}p_h^0r_h^0ds+\tau\sum_{n=1}^M\int_{\Gamma_i}\partial_\tau p_h^n\,\partial_\tau r_h^n ds\right).\tag{7.2}$$

Clearly, evaluating the derivatives of $J_{h,\tau}$ at a given point $q_{h,\tau}$ using formula (7.1) is extremely expensive. To reduce the cost, we introduce an adjoint equation for the Crank–Nicolson scheme (6.2), which seems to have not been studied in the literature before. This needs to be done carefully in order to meet our final goal. A discrete sequence $\{w_h^n\}_{n=0}^M$ is defined in such a way that $w_h^M=0$ and $w_h^n\in V^h$ for $n\neq M$ solves

$$-\int_\Omega\frac{w_h^n-w_h^{n-1}}{\tau}\phi_h dx+\int_\Omega\frac{\bar\alpha^{n+1}\nabla w_h^n+\bar\alpha^n\nabla w_h^{n-1}}{2}\cdot\nabla\phi_h dx$$
$$+\int_{\Gamma_o}\frac{\bar c^{n+1}w_h^n+\bar c^n w_h^{n-1}}{2}\phi_h ds=\alpha_n\int_\omega(u_h^n-z^n)\phi_h dx\quad\forall\phi_h\in V^h.\tag{7.3}$$

Now taking $\phi_h = \mathcal{U}_h^n$ in (7.3), we can rewrite the first term in $J_{h,\tau}(q_{h,\tau})'p_{h,\tau}$ as

$$\begin{aligned} J_{h,\tau}^1 = &-\sum_{n=1}^{M}\int_\Omega \frac{w_h^n - w_h^{n-1}}{\tau}\mathcal{U}_h^n dx + \sum_{n=1}^{M}\int_\Omega \frac{\bar{\alpha}^{n+1}\nabla w_h^n + \bar{\alpha}^n\nabla w_h^{n-1}}{2}\cdot\nabla\mathcal{U}_h^n dx \\ &+\sum_{n=1}^{M}\int_{\Gamma_o}\frac{\bar{c}^{n+1}w_h^n + \bar{c}^n w_h^{n-1}}{2}\mathcal{U}_h^n ds. \end{aligned}$$

Then using formula (5.17), the identity[2]

$$\sum_{n=1}^{k}(a_n + a_{n-1})b_n = a_k b_k - a_0 b_0 + \sum_{n=1}^{k} a_{n-1}(b_n + b_{n-1}),$$

and the equation for $\mathcal{U}_h^n$, we obtain

$$\begin{aligned} J_{h,\tau}^1 = &\sum_{n=1}^{M}\int_\Omega \frac{\mathcal{U}_h^n - \mathcal{U}_h^{n-1}}{\tau} w_h^{n-1}dx + \sum_{n=1}^{M}\int_\Omega \bar{\alpha}^n\nabla\frac{\mathcal{U}_h^n + \mathcal{U}_h^{n-1}}{2}\cdot\nabla w_h^{n-1}dx \\ &+\sum_{n=1}^{M}\int_{\Gamma_o}\bar{c}^n\frac{\mathcal{U}_h^n + \mathcal{U}_h^{n-1}}{2}w_h^{n-1}ds \\ = &-\sum_{n=1}^{M}\int_{\Gamma_i}\frac{p_h^n + p_h^{n-1}}{2}w_h^{n-1}ds. \end{aligned}$$

This, along with (7.1), leads to a very simple formula for evaluating the derivative of $J_{h,\tau}$:

$$\begin{aligned} J_{h,\tau}(q_{h,\tau})'p_{h,\tau} = \int_{\Gamma_i}\Bigg(\beta q_h^0 p_h^0 + \sum_{n=1}^{M}\Bigg\{&\frac{\beta(q_h^n - q_h^{n-1})(p_h^n - p_h^{n-1})}{\tau} \\ &- \frac{\tau(p_h^n + p_h^{n-1})w_h^{n-1}}{2}\Bigg\}\Bigg)ds. \end{aligned} \tag{7.4}$$

Next, we are going to formulate the conjugate gradient method for the nonlinear minimization (6.1). Let us first establish one-to-one correspondences between finite element functions and their coefficient vectors. For any $q_h^j \in V_{\Gamma_i}^h$, we write its representation in terms of the basis $\{\psi_i\}_{i=1}^N$ as

$$q_h^j = \sum_{i=1}^{N} q_i^j \psi_i.$$

Then each finite element function $q_{h,\tau}$ or $\{q_h^0, q_h^1, \dots, q_h^M\}$ corresponds uniquely to an $(M+1)N$-dimensional vector

$$\mathbf{q} = (q_1^0, \dots, q_N^0, q_1^1, \dots, q_N^1, q_1^2, \dots, q_N^2, \dots, q_1^M, \dots, q_N^M)^T.$$

[2] This crucial identity has not been seen in the literature before and has no continuous counterpart, unlike the widely used identity (5.17) that is known as the discrete integration by parts formula.

Letting $f(\mathbf{q}) = J_{h,\tau}(q_{h,\tau})$, one can directly verify the relation for the first derivatives of $f(\mathbf{q})$,

$$\frac{\partial f(\mathbf{q})}{\partial q_i^j} = J_{h,\tau}(\{q_h^0, q_h^1, \ldots, q_h^M\})'(\{0, \ldots, \psi_i, \ldots, 0\}),$$

and the relation for the Hessian $H = (h_{ij})$,

$$h_{ik} :\equiv \frac{\partial^2 f(\mathbf{q})}{\partial q_i^j \partial q_k^l} = J_{h,\tau}(\{q_h^0, q_h^1, \ldots, q_h^M\})''(\{0, \ldots, \psi_i, \ldots, 0\})(\{0, \ldots, \psi_k, \ldots, 0\}).$$

This leads to the following expression:

$$f(\mathbf{q}) = \frac{1}{2}\mathbf{q}^T H \mathbf{q} + \nabla f(\mathbf{0})^T \mathbf{q} + f(\mathbf{0}).$$

We see that the evaluation of the Hessian H is extremely expensive. Fortunately, only its products with vectors are needed in the conjugate gradient method, and such products can be done with much less cost by noting the identity that $H\mathbf{q} = \nabla f(\mathbf{q}) - \nabla f(\mathbf{0})$ and the simple formula (7.4).

We are now ready to state the conjugate gradient algorithm for solving the discrete minimization problem (6.1)–(6.2). We shall use $(J_{h,\tau}(q_{h,\tau}))'$ for $\nabla f(\mathbf{q})$ to emphasize the dependence of the first order derivatives on mesh size h and time step τ.

Conjugate Gradient Algorithm

Step 1. Given a tolerance ε, compute $\tilde{\mathbf{g}}_0 = (J_{h,\tau}(0))'$.

Step 2. Given an initial guess $q_{h,\tau}^{(0)}$, solve the direct problem (6.2) and the adjoint equation (7.3), then compute $\mathbf{g}_0 = (J_{h,\tau}(q_{h,\tau}^{(0)}))'$ by using (7.4). Set $\mathbf{d}_0 := -\mathbf{g}_0$ and $k := 0$.

Step 3. Solve the one-dimensional problem

$$J_{h,\tau}(q_{h,\tau}^{(k)} + \alpha_k d_{h,\tau}^{(k)}) = \min_{\alpha} J_{h,\tau}(q_{h,\tau}^{(k)} + \alpha d_{h,\tau}^{(k)})$$

by computing

$$\alpha_k = -\frac{\mathbf{g}_k^T \mathbf{d}_k}{\mathbf{d}_k^T((J_{h,\tau}(d_{h,\tau}^{(k)}))' - \tilde{\mathbf{g}}_0)}.$$

Set $q_{h,\tau}^{(k+1)} := q_{h,\tau}^{(k)} + \alpha_k d_{h,\tau}^{(k)}$ and $k := k+1$. Compute

$$\mathbf{g}_k = (J_{h,\tau}(q_{h,\tau}^{(k)}))',$$
$$\beta_k = \frac{\mathbf{g}_k^T \mathbf{g}_k}{\mathbf{g}_{k-1}^T \mathbf{g}_{k-1}},$$
$$\mathbf{d}_k = -\mathbf{g}_k + \beta_k \mathbf{d}_{k-1}.$$

Step 4. If $\|\mathbf{g}_k\| \leq \varepsilon \|\mathbf{g}_0\|$, stop; otherwise goto Step 3.

8. Numerical experiments. In this section we show some numerical experiments on heat flux reconstructions using the two regularization methods (5.8)–(5.10) and (6.1)–(6.2). The physical domain Ω is taken to be $\Omega = \{(x, y);\ (\frac{1}{2})^2 \leq x^2 + y^2 \leq 1\}$. The domain Ω is triangulated as in Figure 2 using sectorial elements, with each

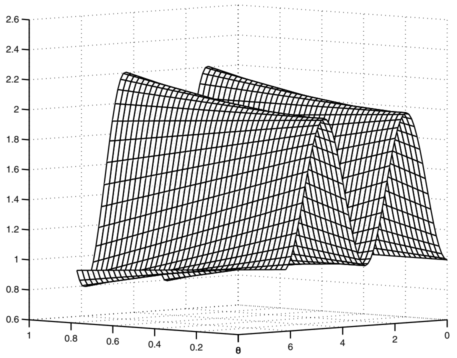


FIG. 3. *Exact q.*

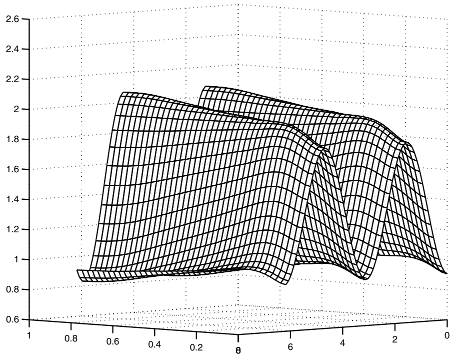


FIG. 4. *Method* (6.1) *used,* $\beta = 7 \times 10^{-6}$*, iter* $= 21$*, error* $= 4.19 \times 10^{-2}$.

circle divided into 60 arcs of equal length. The time interval $[0, 1]$ is divided into 40 equally spaced subintervals. For the conjugate gradient method, we take the tolerance $\epsilon = 10^{-4}$, and the initial guess $q^{(0)}_{h,\tau}$ of the heat flux is taken to be a constant zero everywhere in the whole space-time domain. In all three-dimensional figures shown below, the x-axis stands for the time interval varying from 0 to 1 and the y-axis stands for the inner boundary $\Gamma_i = \{(x, y);\ x^2 + y^2 = (\frac{1}{2})^2\}$ represented by the polar coordinate θ varying from 0 to 2π, while the z-axis shows the magnitude of the heat flux at each point (t, θ). The errors listed under each figure are the relative L^2-norm errors between the exact and numerically reconstructed heat fluxes.

In our simulations, the coefficients α, c, and u_a in (1.1) and (1.3) are taken to be $\alpha(x, t) = 1, c(x, t) = 1$, and $u_a(x, t) = 0$. In order to select more general and difficult profiles of heat fluxes for our tests, we add a source term $f(x, t)$ in (1.1). As our first example, we try the exact solution $u(x, y, t)$ and the heat flux $q(x, y, t)$ to be reconstructed as follows:

$$u(x, y, t) = x^2 + 2y^2 + t + \sin(xyt), \quad q(x, y, t) = 4x^2 + 8y^2 + 4xyt\cos(xyt).$$

Instead of the exact data $u(x, y, t)$, we use the perturbed data of the form $z(x, y, t) = u(x, y, t) + \delta\, u(x, y, t)$ as the measurement data, with the noise level $\delta = 1\%$ (1% relative noise pointwise). We first test the case when the measurement region is taken to be $\omega = \{(x, y);\ (\frac{3}{4})^2 \leq x^2 + y^2 \leq 1\}$. Figure 3 plots the exact heat flux, while Figure 4 shows the numerically reconstructed heat flux using the finite element method

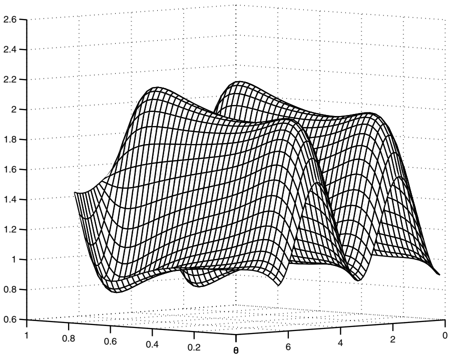


FIG. 5. *Method* (5.8) *used,* $\beta = 2 \times 10^{-6}$, *iter* $= 5$, *error* $= 7.73 \times 10^{-2}$.

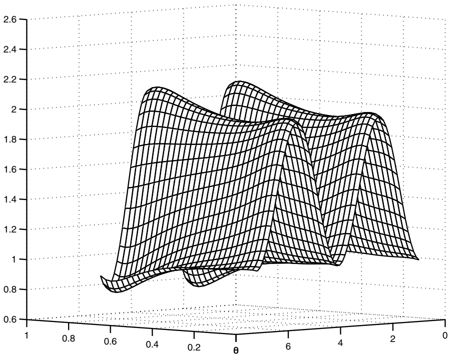


FIG. 6. *Method* (5.8) *used, plots from the initial* 4 *and last* 5 *time points removed, error* $= 3.09 \times 10^{-2}$.

(6.1) with L^2-regularization in space but H^1-regularization in time for heat fluxes. From Figure 4 we see that the numerical reconstruction works very well, considering the difficult oscillation of the heat flux in space. Also the conjugate gradient iteration performs very stably for such an oscillating heat flux, starting with a very bad initial guess of constant zero everywhere in the space-time domain. Figure 5 presents the numerical reconstruction using the finite element method (5.8) with L^2-regularization in both space and time for heat fluxes. One finds that the quality of reconstruction is far from satisfactory compared to the result we have seen in Figure 4 using the finite element method (6.1); the reconstruction is especially bad near the initial and terminal time. But interestingly, when we remove the bad reconstruction at a few initial and terminal time points, the remaining reconstruction seems very satisfactory again; see Figure 6.

We have also tried to see the effects of the measurement region. When the measurement region is reduced to a smaller subdomain $\omega = \{(x, y);\, (\frac{4}{5})^2 \le x^2 + y^2 \le 1\}$, the numerical reconstructions are not affected too much; see Figures 7 and 8.

As our second example, we take the exact solutions $u(x, y, t)$ and $q(x, y, t)$ in (1.1) and (1.4) as the following functions:

$$u(x, y, t) = \sin \pi t(x \cos \pi y + y \sin \pi x),$$
$$q(x, y, t) = 2 \sin \pi t(\pi xy(\cos \pi x - \sin \pi y) + x \cos \pi y + y \sin \pi x).$$

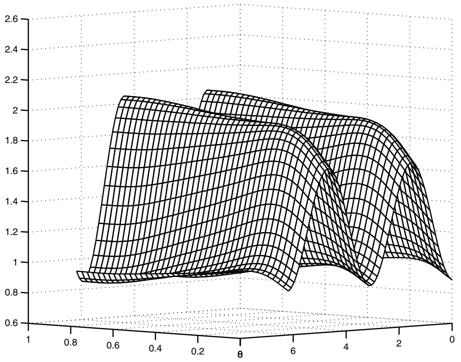


FIG. 7. *Method* (6.1) *used,* $\beta = 10^{-5}$*, iter* $= 23$*, error* $= 5.25 \times 10^{-2}$.

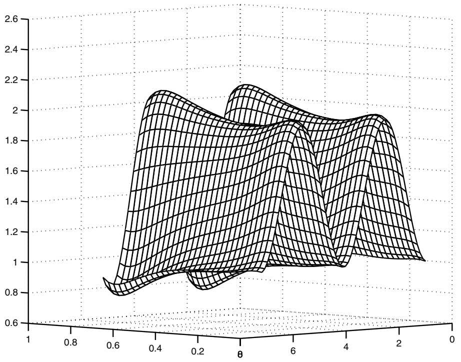


FIG. 8. *Method* (5.8) *used,* $\beta = 6 \times 10^{-6}$*, iter* $= 13$*, plots from the initial* 4 *and last* 5 *time points removed, error* $= 3.57 \times 10^{-2}$.

Again, the perturbed data $z(x, y, t) = u(x, y, t) + \delta\, u(x, y, t)$, with 1% noise pointwise, is taken to be the measurement data in ω. We first test the case when the measurement region is taken to be $\omega = \{(x, y);\ (\frac{3}{4})^2 \le x^2 + y^2 \le 1\}$. Figure 9 plots the exact heat flux q, which appears to be very challenging for numerical reconstruction as it oscillates widely in both time and space direction. Figure 10 shows the numerically reconstructed heat flux using the finite element method (6.1) with L^2-regularization in space but H^1-regularization in time for heat fluxes. This demonstrates very satisfactory performance of the numerical reconstruction algorithm, especially the stability and effectiveness of the conjugate gradient iteration, considering that it is such an oscillating heat flux and that it starts with a very bad initial guess of constant zero everywhere in the space-time domain. Figure 11 presents the numerical reconstruction using the finite element method (5.8) with L^2-regularization in both space and time for heat fluxes. Again its quality of reconstruction is not as good as the one obtained using the finite element method (6.1), and the accuracy is much worse.

When the measurement subregion is reduced to a smaller subdomain $\omega = \{(x, y);\ (\frac{4}{5})^2 \le x^2 + y^2 \le 1\}$, again the numerical reconstructions have not been affected much, as we have seen in the first example.

9. Concluding remarks. The inverse problem of reconstructing profiles of both time- and space-dependent heat fluxes on an inner boundary of a heat conductive

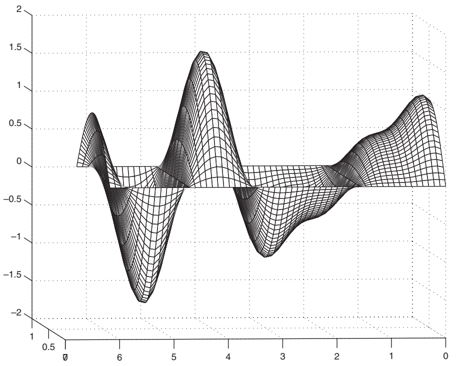


Fig. 9. *The exact q.*

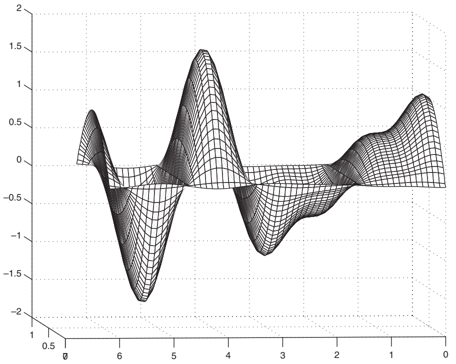


Fig. 10. *Method* (6.1) *used*, $\beta = 10^{-9}$, *iter* $= 30$, *error* $= 3.68 \times 10^{-2}$.

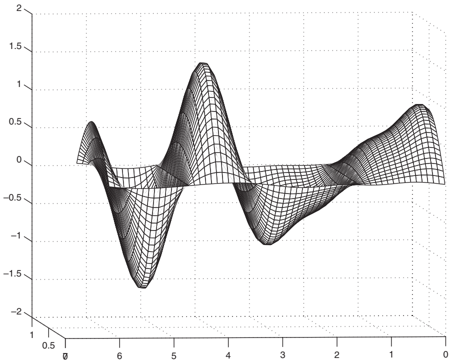


Fig. 11. *Method* (5.8) *used*, $\beta = 4 \times 10^{-8}$, *iter* $= 29$, *error* $= 11.59 \times 10^{-2}$.

system is investigated. The reconstruction problem is severely ill-posed as it involves the heat flux profile at the initial time and on the inner boundary. Validation and effectiveness of two regularization formulations are justified both theoretically and numerically for the reconstruction, without any constraints enforced on the search spaces of heat fluxes when appropriate regularizations are selected. Regarding the approximation of the regularized nonlinear minimization systems, it is very tricky and essential to decide how to effectively discretize in both time and space the nonlinear optimizations and the associated parabolic equation and its adjoint so that the resulting fully discrete schemes converge. Two such discrete approaches are proposed to approximate two nonlinear minimization formulations: the first uses the backward Euler scheme in time, while the second requires the Crank–Nicolson scheme, with both adopting piecewise linear finite elements for space approximation and the trapezoidal and midpoint rules for discretization of the cost functionals. A novel weighted discrete projection operator Q_h is introduced which possesses both L^2- and H^1-stability and L^2-optimal error estimate, crucial to the success of convergence analysis of two fully discrete schemes. The resulting nonlinear finite element minimization systems are shown to be well suited for the solutions by conjugate gradient method. Numerical experiments have demonstrated the stability and effectiveness of the reconstruction algorithms.

There exists little work on numerical reconstruction of both time- and space-dependent physical profiles, and even less on convergence analysis of numerical reconstruction methods. As we have seen, the convergence analyses of the fully discrete schemes are much more difficult and trickier than the cases with only space-dependent profiles. This paper provides a relatively complete study on reconstruction of both time- and space-dependent heat fluxes, including well-posedness of the regularized systems, convergence of fully discrete approximations, numerical algorithms for solving discrete nonlinear minimizations, and numerical experiments. Most technical tools should be useful in theoretical and numerical analysis of regularization methods for other inverse problems.

Acknowledgment. The authors wish to thank an anonymous referee for his constructive comments.

REFERENCES

[1] T. F. Chan and X.-C. Tai, *Level set and total variation regularization for elliptic inverse problems with discontinuous coefficients*, J. Comput. Phys., 193 (2003), pp. 40–66.

[2] T. F. Chan and X.-C. Tai, *Identification of discontinuous coefficients in elliptic problems using total variation regularization*, SIAM J. Sci. Comput., 25 (2003), pp. 881–904.

[3] Z. Chen and J. Zou, *An augmented Lagrangian method for identifying discontinuous parameters in elliptic systems*, SIAM J. Control Optim., 37 (1999), pp. 892–910.

[4] Z. Chen and J. Zou, *Finite element methods and their convergence for elliptic and parabolic interface problems*, Numer. Math., 79 (1998), pp. 175–202.

[5] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[6] L. C. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1999.

[7] H. W. Engl and J. Zou, *A new approach to convergence rate analysis of Tikhonov regularization for parameter identification in heat conduction*, Inverse Problems, 16 (2000), pp. 1907–1923.

[8] K. Ito and K. Kunisch, *The augmented Lagrangian method for parameter estimation in elliptic systems*, SIAM J. Control Optim., 28 (1990), pp. 113–136.

[9] Y. L. Keung and J. Zou, *An efficient linear solver for nonlinear parameter identification problems*, SIAM J. Sci. Comput., 22 (2000), pp. 1511–1526.

[10] A. Kirsch, *An Introduction to the Mathematical Theory of Inverse Problems*, Appl. Math. Sci. 120, Springer-Verlag, New York, 1996.

[11] K. Kunisch and J. Zou, *Iterative choice of regularization parameters in linear inverse problems*, Inverse Problems, 14 (1998), pp. 1247–1264.

[12] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.

[13] R. Temam, *Navier-Stokes Equations*, North–Holland, Amsterdam, 1977.

[14] J. L. Xie, *Numerical Reconstruction of Heat Fluxes*, Ph.D. thesis, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong, 2003.

[15] J. L. Xie and J. Zou, *An improved model function method for choosing regularization parameters in linear inverse problems*, Inverse Problems, 18 (2002), pp. 631–643.

[16] J. Xu, *Theory of Multilevel Methods*, Ph.D. thesis, Cornell University, Ithaca, NY, 1989.

[17] M. Yamamoto, *Lipschitz stability in inverse parabolic problems by the Carleman estimate*, Inverse Problems, 14 (1998), pp. 1229–1245.

[18] M. Yamamoto and J. Zou, *Simultaneous reconstruction of the initial temperature and heat radiative coefficient*, Inverse Problems, 17 (2001), pp. 1181–1202.

[19] H.-M. Yin, *The Lipschitz continuity of the interface in the heat equation with strong absorption*, Nonlinear Anal., 20 (1993), pp. 413–416.

[20] H.-M. Yin, *Solvability of a class of parabolic inverse problems*, Adv. Differential Equations, 1 (1996), pp. 1005–1024.

[21] E. Zeidler, *Nonlinear Functional Analysis and Its Applications*, Vol. 2, Springer-Verlag, New York, 1985.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1536–1571

CONVERGENCE OF THE LAGRANGE–GALERKIN METHOD FOR THE EQUATIONS MODELLING THE MOTION OF A FLUID-RIGID SYSTEM[*]

JORGE SAN MARTÍN[†], JEAN-FRANÇOIS SCHEID[‡], TAKÉO TAKAHASHI[‡], AND MARIUS TUCSNAK[‡]

Abstract. In this paper, we consider a Lagrange–Galerkin scheme to approximate a two-dimensional fluid-rigid body problem. The equations of the system are the Navier–Stokes equations in the fluid part, coupled with ordinary differential equations for the dynamics of the rigid body. In this problem, the equations of the fluid are written in a domain whose variation is one of the unknowns. We introduce a numerical method based on the use of characteristics and on finite elements with a fixed mesh. Our main result asserts the convergence of this scheme.

Key words. fluid-structure interaction, incompressible Navier–Stokes equations, finite element method, Lagrange–Galerkin method

AMS subject classifications. 35Q30, 76D05, 65M12, 76M10

DOI. 10.1137/S0036142903438161

1. Introduction.

The aim of this paper is to analyze a Lagrange–Galerkin approximation of the equations modelling the motion of a two-dimensional rigid body immersed in a fluid. We first briefly describe the equations modelling this system. Assume that the system fluid-rigid body occupies a bounded domain $\mathcal{O}$ in $\mathbb{R}^2$ with a regular boundary $\partial\mathcal{O}$. The solid is supposed to occupy at each instant t a closed connected subset $B(t) \subset \mathcal{O}$ which is surrounded by a viscous homogeneous fluid filling the domain $\Omega(t) = \mathcal{O}\backslash B(t)$.

The motion of the fluid is described by the classical Navier–Stokes equations, whereas the motion of the rigid body is governed by the balance equations for linear and angular momentum (Newton's laws). More precisely, we consider the following system coupling partial differential and ordinary differential equations:

$$\rho_f \frac{\partial \mathbf{u}}{\partial t} - \nu\Delta\mathbf{u} + \rho_f(\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \rho_f \mathbf{f}, \quad \mathbf{x} \in \Omega(t), \quad t \in [0,T], \tag{1.1}$$

$$\operatorname{div}\mathbf{u} = 0, \quad \mathbf{x} \in \Omega(t), \quad t \in [0,T], \tag{1.2}$$

$$\mathbf{u} = 0, \quad \mathbf{x} \in \partial\mathcal{O}, \quad t \in [0,T], \tag{1.3}$$

$$\mathbf{u} = \boldsymbol{\zeta}'(t) + \omega(t)(\mathbf{x} - \boldsymbol{\zeta}(t))^{\perp}, \quad \mathbf{x} \in \partial B(t), \quad t \in [0,T], \tag{1.4}$$

$$M\boldsymbol{\zeta}''(t) = -\int_{\partial B(t)} \boldsymbol{\sigma}\mathbf{n}\,\mathrm{d}\Gamma + \rho_s \int_{B(t)} \mathbf{f}(\mathbf{x},t)\,\mathrm{d}\mathbf{x}, \quad t \in [0,T], \tag{1.5}$$

[*] Received by the editors November 29, 2003; accepted for publication (in revised form) January 19, 2005; published electronically October 31, 2005.
http://www.siam.org/journals/sinum/43-4/43816.html

[†] Departamento de Ingeniería Matemática, Universidad de Chile, Casilla 170/3-Correo 3, Santiago, Chile (jorge@dim.uchile.cl). The work of this author was partially supported by Conicyt under grant Fondecyt 1010402 and by the Center for Mathematical Modelling from Chile.

[‡] Institut Élie Cartan, UMR 7502 UHP/CNRS/INRIA, B.P. 239, F-54506 Vandoeuvre-lès-Nancy Cedex, France (Jean-Francois.Scheid@iecn.u-nancy.fr, Takeo.Takahashi@iecn.u-nancy.fr, Marius.Tucsnak@iecn.u-nancy.fr). The work of these authors was supported by the CORIDA research team of INRIA. The work of the fourth author was partially supported by Conicyt under grant Fondecyt 1010402.

(1.6)
$$J\omega'(t) = -\int_{\partial B(t)} (\mathbf{x} - \boldsymbol{\zeta}(t))^{\perp} \cdot \boldsymbol{\sigma}\mathbf{n} \, \mathrm{d}\Gamma + \rho_s \int_{B(t)} (\mathbf{x} - \boldsymbol{\zeta}(t))^{\perp} \cdot \mathbf{f}(\mathbf{x}, t) \, \mathrm{d}\mathbf{x}, \quad t \in [0, T],$$

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}), \quad \mathbf{x} \in \Omega(0), \tag{1.7}$$

$$\boldsymbol{\zeta}(0) = \boldsymbol{\zeta}_0 \in \mathbb{R}^2, \quad \boldsymbol{\zeta}'(0) = \boldsymbol{\zeta}_1 \in \mathbb{R}^2, \quad \omega(0) = \omega_0 \in \mathbb{R}. \tag{1.8}$$

In the above equations the unknowns are $\mathbf{u}(\mathbf{x}, t)$ (the Eulerian velocity field of the fluid), $p(\mathbf{x}, t)$ (the pressure of the fluid), $\boldsymbol{\zeta}(t)$ (the position of the mass center of the rigid body), and $\omega(t)$ (the angular velocity of the rigid body). The domain $B(t)$ is defined by

$$B(t) = \{\mathbf{R}_{-\theta(t)}\mathbf{y} + \boldsymbol{\zeta}(t), \ \mathbf{y} \in B\},$$

where

$$\theta(t) = \int_0^t \omega(s) \, \mathrm{d}s, \tag{1.9}$$

$B = B(0)$, and $\mathbf{R}_\theta$ is the rotation matrix of angle θ. Moreover, we have denoted by $\partial B(t)$ the boundary of the rigid body at instant t and by $\mathbf{n}(\mathbf{x}, t)$ the unit normal to $\partial B(t)$ at the point $\mathbf{x}$ directed to the interior of the rigid body.

The constants ρ_f and ρ_s are, respectively, the density of the fluid and the density of the rigid body. In what follows, we assume that the densities of the fluid and of the solid are equal, that is

$$\rho_f = \rho_s = 1, \tag{1.10}$$

and that the rigid body is a ball in $\mathbb{R}^2$. Assumption (1.10) is clearly restrictive but it is important for the forthcoming analysis (see Remarks 2.1 and 2.4 below). On the contrary, the assumption that the rigid body is a ball is not essential but avoids some technicalities.

The constants M and J are the mass and the moment of inertia of the rigid body, and the positive constant ν is the viscosity of the fluid. Moreover, $\mathbf{f}(\mathbf{x}, t)$ is the applied force (per unit mass).

For all $\mathbf{x} = \binom{x_1}{x_2}$, we denote by $\mathbf{x}^\perp$ the vector $\mathbf{x}^\perp = \binom{x_2}{-x_1}$. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$, then $\mathbf{x} \cdot \mathbf{y}$ stands for the inner product of $\mathbf{x}$ and $\mathbf{y}$ and $|\mathbf{x}|$ stands for the corresponding norm. We have also denoted by w' and w'' the derivatives of a function w depending only on the time t.

Finally, the stress tensor (also called the Cauchy stress) is defined by

$$\boldsymbol{\sigma}(\mathbf{x}, t) = -p(\mathbf{x}, t)\mathbf{Id} + 2\nu\mathbf{D}(\mathbf{u}), \tag{1.11}$$

where $\mathbf{Id}$ is the identity matrix and $\mathbf{D}(\mathbf{u})$ is the tensor field defined by

$$D(u)_{k,l} = \frac{1}{2}\left(\frac{\partial u_k}{\partial x_l} + \frac{\partial u_l}{\partial x_k}\right).$$

The main difficulties of this problem are that

- the equations of the structure are coupled with those of the fluid,
- the domain of the fluid is variable, and it is one of the unknowns of the problem (we thus have a free boundary problem).

The well-posedness of this type of system has been recently studied in a large number of papers (see, for instance, Desjardins and Esteban [6], Gunzburger, Lee, and Seregin [17], San Martín, Starovoitov, and Tucsnak [25], Grandmont and Maday [16], Takahashi [27], and the references therein).

The literature on the numerical approximation of the solution of (1.1)–(1.8) also contains a large number of recent papers. Some of these papers are based on an arbitrary Lagrangian Eulerian (ALE) formulation; see, for example, Grandmont, Guimet, and Maday [15], Nobile [22], Maury and Glowinski [21], Maury [19], [20], Formaggia and Nobile [9], and Farhat, Geuzaine, and Grandmont [8]. In the ALE method, at each time step, the mesh is moved with an arbitrary velocity in the fluid to follow the motion of the rigid body. The stability of the ALE method is studied in [9] (in the case of the finite element context) and in [8] (in the case of the finite volume context). We also mention the work of Gastaldi [11], where, in the case of an advection-diffusion equation in a moving two-dimensional domain, a priori error estimates that are optimal both in space and time have been obtained.

Another approach, developed by Glowinski et al. [14], [13] is based on a fictitious domain formulation: the rigid bodies are filled by the surrounding fluid, and the constraint of rigid body motion is relaxed by introducing a distributed Lagrange multiplier.

As far as we know, the only proof of the convergence of one of these methods is given in [15] for a simplified problem in one space dimension. The main novelty brought in by our paper consists of the fact that we construct a new approximation method using a fixed mesh and that we prove a convergence result. This method is inspired by the Galerkin–Lagrange approximation, which is commonly used for Navier–Stokes equations (see Pironneau [23] and Süli [26]).

The remaining part of this paper is organized as follows. In section 2 we introduce some function spaces and semidiscretize our problem with respect to the time variable. In section 3 we give the full discretization of the problem and state the main result. Section 4 is devoted to the study of the finite element spaces that were introduced in the previous section. Section 5 is devoted to the study of a change of variables that plays an important role in the proof of the main result. In section 6 we prove the consistency of our scheme. Finally, in section 7 we give the proof of the main result.

2. Notation and preliminaries.

2.1. Notation and function spaces. Throughout this paper, we shall use the classical Sobolev spaces $H^s(\Omega)$, $H^s_0(\Omega)$, $H^{-s}(\Omega)$, $s \geqslant 0$, and the space of Lipschitz continuous functions $C^{0,1}(\overline{\Omega})$ on the closure of Ω. We also define

$$L^2_0(\Omega) = \left\{ f \in L^2(\Omega) \;\middle|\; \int_\Omega f \, \mathrm{d}\mathbf{x} = 0 \right\}$$

and denote by $\mathcal{L}^2_0(\Omega)$, $\mathcal{H}^s(\Omega)$, $\mathcal{H}^s_0(\Omega)$, $\mathcal{H}^{-s}(\Omega)$, $s \geqslant 0$, the spaces $\left[L^2_0(\Omega)\right]^2$, $[H^s(\Omega)]^2$, $[H^s_0(\Omega)]^2$, $[H^{-s}(\Omega)]^2$. The usual inner product in $\mathcal{L}^2(\mathcal{O})$ will be denoted by

$$(\mathbf{u}, \mathbf{v}) = \int_{\mathcal{O}} \mathbf{u} \cdot \mathbf{v} \, \mathrm{d}\mathbf{x} \quad \forall \mathbf{u}, \mathbf{v} \in \mathcal{L}^2(\mathcal{O}). \tag{2.1}$$

If $\mathbf{A}$ is a matrix, we denote by $\mathbf{A}^*$ its transpose. For any 2×2 matrices $\mathbf{A}, \mathbf{B} \in \mathcal{M}_{2\times 2}$, we denote by $\mathbf{A} : \mathbf{B}$ their inner product $\mathbf{A} : \mathbf{B} = \mathrm{Trace}(\mathbf{A}^*\mathbf{B})$, and by $|\mathbf{A}|$ the corresponding norm. For convenience, we use the same notation as in (2.1) for the inner product in $L^2(\mathcal{O}, \mathcal{M}_{2\times 2})$, that is,

$$(\mathbf{A},\mathbf{B}) = \int_{\mathcal{O}} \mathbf{A} : \mathbf{B} \ \mathrm{d}\mathbf{x} \quad \forall \mathbf{A}, \mathbf{B} \in L^2(\mathcal{O}, \mathcal{M}_{2\times 2}). \tag{2.2}$$

We also define the spaces

$$\mathcal{K}(\boldsymbol{\zeta}) = \{\mathbf{u} \in \mathcal{H}_0^1(\mathcal{O}) \mid \mathbf{D}(\mathbf{u}) = 0 \quad \text{in } B(\boldsymbol{\zeta})\} \tag{2.3}$$

and

$$\widehat{\mathcal{K}}(\boldsymbol{\zeta}) = \{\mathbf{u} \in \mathcal{H}_0^1(\mathcal{O}) \mid \operatorname{div} \mathbf{u} = 0 \text{ in } \mathcal{O}, \quad \mathbf{D}(\mathbf{u}) = 0 \text{ in } B(\boldsymbol{\zeta})\}, \tag{2.4}$$

where $\boldsymbol{\zeta} \in \mathbb{R}^2$ and $B(\boldsymbol{\zeta}) = \{\mathbf{x} \in \mathbb{R}^2, |\mathbf{x} - \boldsymbol{\zeta}| \leqslant 1\}$. According to Lemma 1.1 of [29, p. 18], for any $\mathbf{u} \in \mathcal{K}(\boldsymbol{\zeta})$, there exist $\mathbf{l}_{\mathbf{u}} \in \mathbb{R}^2$ and $\omega_{\mathbf{u}} \in \mathbb{R}$ such that

$$\mathbf{u}(\mathbf{y}) = \mathbf{l}_{\mathbf{u}} + \omega_{\mathbf{u}} (\mathbf{y} - \boldsymbol{\zeta})^{\perp} \quad \forall \mathbf{y} \in B(\boldsymbol{\zeta}).$$

These spaces are specific to our problem. In fact, if the solution $\mathbf{u}$ of (1.1)–(1.8) is extended by

$$\mathbf{u}(\mathbf{x},t) = \boldsymbol{\zeta}'(t) + \omega(t)(\mathbf{x} - \boldsymbol{\zeta}(t))^{\perp} \quad \forall \mathbf{x} \in B(\boldsymbol{\zeta}(t)),$$

then we easily see that $\mathbf{u}(t) \in \widehat{\mathcal{K}}(\boldsymbol{\zeta}(t))$.

In what follows, the solution $\mathbf{u}$ *of* (1.1)–(1.8) *will be extended as above.*

We also notice that, by using (1.10), for any $\mathbf{u}, \mathbf{v} \in \mathcal{K}(\boldsymbol{\zeta})$ we have

$$(\mathbf{u},\mathbf{v}) = \int_{\mathcal{O}\setminus B(\boldsymbol{\zeta})} \mathbf{u} \cdot \mathbf{v} \ \mathrm{d}\mathbf{x} + M \mathbf{l}_{\mathbf{u}} \cdot \mathbf{l}_{\mathbf{v}} + J \omega_{\mathbf{u}} \, \omega_{\mathbf{v}}. \tag{2.5}$$

Remark 2.1. In the case of different densities $\rho_F \neq \rho_S$, the natural inner product to be used seems to be

$$\langle \mathbf{u}, \mathbf{v} \rangle_{\zeta} = \rho_F \int_{\mathcal{O}\setminus B(\boldsymbol{\zeta})} \mathbf{u} \cdot \mathbf{v} \ \mathrm{d}\mathbf{x} + M \mathbf{l}_{\mathbf{u}} \cdot \mathbf{l}_{\mathbf{v}} + J \omega_{\mathbf{u}} \, \omega_{\mathbf{v}},$$

which clearly depends on the position of the ball. This fact would considerably complicate the further analysis.

An important ingredient of the numerical method we use is given by the characteristic functions whose level lines are the integral curves of the velocity field. More precisely (see, for instance, [23], [26]), the characteristic function $\widetilde{\boldsymbol{\psi}} : [0,T]^2 \times \mathcal{O} \to \mathcal{O}$ is defined as the solution of the initial value problem

$$\begin{cases} \dfrac{d}{dt} \widetilde{\boldsymbol{\psi}}(t; s, \mathbf{x}) = \mathbf{u}(\widetilde{\boldsymbol{\psi}}(t; s, \mathbf{x}), t), \\ \widetilde{\boldsymbol{\psi}}(s; s, \mathbf{x}) = \mathbf{x}. \end{cases} \tag{2.6}$$

It is well known that the material derivative $D_t \mathbf{u} = \partial \mathbf{u}/\partial t + (\mathbf{u} \cdot \nabla)\mathbf{u}$ of $\mathbf{u}$ at instant t_0 satisfies

$$D_t \mathbf{u}(\mathbf{x}, t_0) = \frac{d}{dt} \Big[\mathbf{u}(\widetilde{\boldsymbol{\psi}}(t; t_0, \mathbf{x}), t) \Big]_{|t=t_0}. \tag{2.7}$$

Remark 2.2. By using a classical result of Liouville (see, for instance, Arnold [1, p. 251]), if

$$\boldsymbol{\zeta} \in \mathcal{H}^2(0,T), \quad \omega \in H^1(0,T), \quad \mathbf{u} \in C([0,T];\widehat{\mathcal{K}}(\boldsymbol{\zeta}(t))),$$

then we have that

$$\det \mathbf{J}_{\widetilde{\boldsymbol{\psi}}} = 1, \tag{2.8}$$

where we have denoted by

$$\mathbf{J}_{\widetilde{\boldsymbol{\psi}}} = \left(\frac{\partial \widetilde{\psi}_i}{\partial y_j}\right)_{i,j}$$

the jacobian matrix of the transformation $\mathbf{y} \mapsto \widetilde{\boldsymbol{\psi}}(\mathbf{y})$.

2.2. Weak form and semidiscretization scheme. In this subsection we give a weak form of (1.1)–(1.8) which is then used to discretize the problem with respect to time.

The fact that (2.9) is called a "weak formulation" of the system (1.1)–(1.8) is justified by the following result.

LEMMA 2.3. *Assume that*

$$\mathbf{u} \in L^2(0,T;\mathcal{H}^2(\Omega(t))) \cap H^1(0,T;\mathcal{L}^2(\Omega(t))) \cap C([0,T];\mathcal{H}^1(\Omega(t))),$$
$$p \in L^2(0,T;H^1(\Omega(t))),$$
$$\boldsymbol{\zeta} \in \mathcal{H}^2(0,T), \quad \omega \in H^1(0,T)$$

and that $\mathbf{u}$ *is extended by*

$$\mathbf{u}(\mathbf{x},t) = \boldsymbol{\zeta}'(t) + \omega(t)(\mathbf{x} - \boldsymbol{\zeta}(t))^{\perp} \quad \forall \mathbf{x} \in B(\boldsymbol{\zeta}(t)).$$

Then $(\mathbf{u},p,\boldsymbol{\zeta},\omega)$ *is the solution of* (1.1)–(1.8) *if and only if* $\mathbf{u}(t) \in \widehat{\mathcal{K}}(\boldsymbol{\zeta}(t))$ *for all* t *and* $(\mathbf{u},p)$ *satisfies*

$$\left(\frac{d}{dt}\left[\mathbf{u} \circ \widetilde{\boldsymbol{\psi}}\right](t), \boldsymbol{\varphi}\right) + 2\nu\left(\mathbf{D}(\mathbf{u}(t)), \mathbf{D}(\boldsymbol{\varphi})\right) - \int_{\Omega(t)} (\operatorname{div} \boldsymbol{\varphi}) p(t)\ \mathrm{d}\mathbf{x} \\ = (\mathbf{f}(t), \boldsymbol{\varphi}) \quad \forall \boldsymbol{\varphi} \in \mathcal{K}(\boldsymbol{\zeta}(t)). \tag{2.9}$$

We skip the proof of Lemma 2.3 since it is similar to the proof of the corresponding result for the classical Navier–Stokes system (see, for instance, [24, Chap. 12]).

Remark 2.4. In the case of different densities $\rho_F \neq \rho_S$, a similar weak statement can be obtained (see, for instance, [5]). In this case $\mathbf{u}$ in the first term of (2.9) should be replaced by $\rho\mathbf{u}$, where $\rho = \rho_F$ in the fluid and $\rho = \rho_S$ in the moving solid. Thus ρ would depend on the time and a transport equation for ρ should be added to the system.

By using the weak formulation given above we can derive a semidiscrete version of our system. For $N \in \mathbb{N}^*$ we denote $\Delta t = T/N$ and $t_k = k\Delta t$ for $k = 0,\dots,N$. Denote by $(\mathbf{u}^k, \boldsymbol{\zeta}^k) \in \widehat{\mathcal{K}}(\boldsymbol{\zeta}^k) \times \mathbb{R}^2$ the approximation of the solution of (1.1)–(1.8) at the time $t = t_k$. We approximate the position of the rigid ball at instant t_{k+1} by $\boldsymbol{\zeta}^{k+1}$, which is defined by the relation

$$\boldsymbol{\zeta}^{k+1} = \boldsymbol{\zeta}^k + \mathbf{u}^k(\boldsymbol{\zeta}^k)\Delta t. \tag{2.10}$$

We then define characteristic function $\overline{\boldsymbol{\psi}}$ associated to the semidiscretized velocity field as the solution of

$$\left\{\begin{array}{l}\dfrac{d}{dt}\overline{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{x})=\mathbf{u}^k(\overline{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{x})),\\ \overline{\boldsymbol{\psi}}(t_{k+1};t_{k+1},\mathbf{x})=\mathbf{x},\end{array}\right. \tag{2.11}$$

and we denote

$$\overline{\boldsymbol{X}}^{\boldsymbol{k}}(\mathbf{x})=\overline{\boldsymbol{\psi}}(t_k;t_{k+1},\mathbf{x})\quad\forall\mathbf{x}\in\mathcal{O}. \tag{2.12}$$

One can easily check that $\overline{\boldsymbol{X}}^{\boldsymbol{k}}(\mathcal{O})=\mathcal{O}$.

We next define $\mathbf{u}^{k+1}\in\widehat{\mathcal{K}}(\boldsymbol{\zeta}^{k+1})$ as the solution of the following Stokes type system:

$$\left(\frac{\mathbf{u}^{k+1}-\mathbf{u}^k\circ\overline{\boldsymbol{X}}^{\boldsymbol{k}}}{\Delta t},\boldsymbol{\varphi}\right)+2\nu\left(\mathbf{D}(\mathbf{u}^{k+1}),\mathbf{D}(\boldsymbol{\varphi})\right)=(\mathbf{f}^{k+1},\boldsymbol{\varphi})\quad\forall\boldsymbol{\varphi}\in\widehat{\mathcal{K}}(\boldsymbol{\zeta}^{k+1}), \tag{2.13}$$

where $\mathbf{f}^{k+1}=\mathbf{f}(t_{k+1})$.

The above equation can be rewritten by using a mixed formulation. To achieve this, we first define

$$M(\boldsymbol{\zeta})=\left\{p\in L^2_0(\mathcal{O})\mid p=0\text{ in }B(\boldsymbol{\zeta})\right\}, \tag{2.14}$$

$$a(\mathbf{u},\mathbf{v})=2\nu\int_{\mathcal{O}}\mathbf{D}(\mathbf{u}):\mathbf{D}(\mathbf{v})\,\mathrm{d}\mathbf{x}\quad\forall\mathbf{u},\mathbf{v}\in\mathcal{H}^1(\mathcal{O}), \tag{2.15}$$

$$b(\mathbf{u},p)=-\int_{\mathcal{O}}\operatorname{div}(\mathbf{u})p\,\mathrm{d}\mathbf{x}\quad\forall\mathbf{u}\in\mathcal{H}^1(\mathcal{O}),\quad\forall p\in L^2_0(\mathcal{O}). \tag{2.16}$$

With the above notation, it is clear that (2.13) is equivalent to the system

$$\left(\frac{\mathbf{u}^{k+1}-\mathbf{u}^k\circ\overline{\boldsymbol{X}}^{\boldsymbol{k}}}{\Delta t},\boldsymbol{\varphi}\right)+a(\mathbf{u}^{k+1},\boldsymbol{\varphi})+b(\boldsymbol{\varphi},p^{k+1})=(\mathbf{f}^{k+1},\boldsymbol{\varphi})\quad\forall\boldsymbol{\varphi}\in\mathcal{K}(\boldsymbol{\zeta}^{k+1}), \tag{2.17}$$

$$b(\mathbf{u}^{k+1},q)=0\quad\forall q\in M(\boldsymbol{\zeta}^{k+1}) \tag{2.18}$$

of unknowns $(\mathbf{u}^{k+1},p^{k+1})\in\mathcal{K}(\boldsymbol{\zeta}^{k+1})\times M(\boldsymbol{\zeta}^{k+1})$.

Remark 2.5. The requirement $p=0$ in $B(\boldsymbol{\zeta})$ for the definition of $M(\boldsymbol{\zeta})$ allows us to define the form b on the whole domain $\mathcal{O}$. This extension does not affect the form b since $\operatorname{div}(\mathbf{u})=0$ in $B(\boldsymbol{\zeta})$ for all $\mathbf{u}\in\mathcal{K}(\boldsymbol{\zeta})$.

It is well known (see, for example, [12, Corollary I.4.1., p. 61]) that the mixed formulation (2.17), (2.18) is a well-posed problem, provided that the spaces $\mathcal{K}(\boldsymbol{\zeta})$, $M(\boldsymbol{\zeta})$ and the bilinear form b satisfy an *inf-sup* condition. The fact that this *inf-sup* condition is satisfied in our case follows from the result below.

LEMMA 2.6. *Suppose that $\boldsymbol{\zeta}\in\mathcal{O}$ is such that $d(\boldsymbol{\zeta},\partial\mathcal{O})=1+\eta$, with $\eta>0$. Then there exists a constant $\beta>0$, depending only on η and on $\mathcal{O}$, such that for all $q\in M(\boldsymbol{\zeta})$ there exists $\mathbf{u}\in\mathcal{K}(\boldsymbol{\zeta})$ with*

$$\int_{\mathcal{O}}\operatorname{div}(\mathbf{u})\,q\,\mathrm{d}\mathbf{x}\geq\beta\|\mathbf{u}\|_{\mathcal{H}^1(\mathcal{O})}\|q\|_{L^2(\mathcal{O})}. \tag{2.19}$$

The proof of the above result can be obtained by slightly modifying the approach used for the mixed formulation of the standard Stokes system (see, for instance [12, p. 81]), so it is left to the reader.

3. Full discretization and statement of the main result. In order to discretize the problem (2.17), (2.18) with respect to the space variable we introduce two families of finite element spaces. We first define a family of finite element spaces that approximate the space $\mathcal{K}(\boldsymbol{\zeta})$ defined in (2.3). Let h denote a discretization parameter, $0 < h < 1$, and let P_1 be the space of all affine functions in $\mathbb{R}^2$.

Consider a quasi-uniform triangulation $\mathcal{T}_h$ of $\mathcal{O}$, as defined, for instance, in [2, p. 106] (this assumption will be accepted in the remainder of this paper and will allow us to make use of inverse estimates). If $T \in \mathcal{T}_h$ is a triangle of vertices $\mathbf{x}_1,\ \mathbf{x}_2$, and $\mathbf{x}_3$, we denote by $\varphi_1(\mathbf{x})$, $\varphi_2(\mathbf{x})$, and $\varphi_3(\mathbf{x})$ the corresponding barycentric coordinates of $\mathbf{x} \in \mathbb{R}^2$ with respect to the vertices $\mathbf{x}_1,\ \mathbf{x}_2$, and $\mathbf{x}_3$ (see, for instance, [4, p. 45] for the definition of barycentric coordinates). We associate to this triangulation two classical approximation spaces used in the mixed finite element methods for the Stokes system. The first space, classically used for the approximation of the velocity field in the mixed statement of the Stokes system, is denoted by $\mathcal{W}_h$ and is defined as the subspace of $\mathcal{H}_0^1(\mathcal{O})$ formed by the P_1-bubble finite elements associated to $\mathcal{T}_h$. More precisely, $\boldsymbol{\varphi} \in \mathcal{W}_h$ if and only if

$$\boldsymbol{\varphi}(\mathbf{x}) = \varphi_1(\mathbf{x})\boldsymbol{\alpha_1} + \varphi_2(\mathbf{x})\boldsymbol{\alpha_2} + \varphi_3(\mathbf{x})\boldsymbol{\alpha_3} + \frac{\varphi_1(\mathbf{x})\varphi_2(\mathbf{x})\varphi_3(\mathbf{x})}{\displaystyle\int_T \varphi_1\varphi_2\varphi_3 \, \mathrm{d}\mathbf{x}}\boldsymbol{\lambda} \qquad \forall\, \mathbf{x} \in T$$

for some constant vectors $\boldsymbol{\alpha_1},\ \boldsymbol{\alpha_2},\ \boldsymbol{\alpha_3},\ \boldsymbol{\lambda} \in \mathbb{R}^2$. We may notice that all functions in $\mathcal{W}_h$ are continuous.

The second space, classically used for the approximation of the pressure in mixed statements of the Stokes system, is denoted by E_h and is defined by

$$E_h = \left\{ q \in C(\overline{\mathcal{O}}) \ \middle|\ q_{|T} \in P_1(T) \right\}. \tag{3.1}$$

For our problem we use two spaces that are related to the presence of the rigid body. The first one, which is used for the approximation of the velocity field, is denoted by $\mathcal{K}_h(\boldsymbol{\zeta})$ and defined by

$$\mathcal{K}_h(\boldsymbol{\zeta}) = \mathcal{W}_h \cap \mathcal{K}(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \mathcal{O}.$$

The second one, which is used for the approximation of the pressure, is denoted by $M_h(\boldsymbol{\zeta})$ and defined by

$$M_h(\boldsymbol{\zeta}) = E_h \cap M(\boldsymbol{\zeta}) \quad \forall \boldsymbol{\zeta} \in \mathcal{O}.$$

We also define the finite element space (see [23])

$$\mathcal{R}_h = \{\mathbf{rot}\ \varphi_h, \quad \varphi_h \in E_h, \quad \varphi_h = 0 \text{ on } \partial\mathcal{O}\}.$$

We denote by $\mathbf{P}$ the orthogonal projection from $\mathcal{L}^2$ onto $\mathcal{R}_h$. More precisely, if $\mathbf{u} \in \mathcal{L}^2(\mathcal{O})$, then $\mathbf{Pu} \in \mathcal{R}_h$ satisfies

$$(\mathbf{u} - \mathbf{Pu}, \mathbf{r}_h) = 0 \quad \forall \mathbf{r}_h \in \mathcal{R}_h.$$

Let N be a positive integer. We denote $\Delta t = T/N$ and $t_k = k\Delta t$. Assume that the approximate solution $(\mathbf{u}_h^k, p_h^k, \boldsymbol{\zeta}_h^k)$ of (1.1)–(1.8) at $t = t_k$ is known. We describe below the numerical scheme allowing us to determinate the approximate solution $(\mathbf{u}_h^{k+1}, p_h^{k+1}, \boldsymbol{\zeta}_h^{k+1})$ at $t = t_{k+1}$. First, we compute $\boldsymbol{\zeta}_h^{k+1} \in \mathbb{R}^2$ by

$$\boldsymbol{\zeta}_h^{k+1} = \boldsymbol{\zeta}_h^k + \mathbf{u}_h^k(\boldsymbol{\zeta}_h^k)\Delta t. \tag{3.2}$$

We denote by $\mathbf{P}\mathbf{u}_h^k$ the projection of $\mathbf{u}_h^k$ onto $\mathcal{R}_h$. Then we define the characteristic function $\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}$ associated to the fully discretized velocity field as the solution of

$$\left\{\begin{aligned} &\frac{d}{dt}\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t;t_{k+1},\mathbf{x}) = \mathbf{P}\mathbf{u}_h^k(\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t;t_{k+1},\mathbf{x})),\\ &\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t_{k+1};t_{k+1},\mathbf{x}) = \mathbf{x}. \end{aligned}\right. \tag{3.3}$$

We also define

$$\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}(\mathbf{x}) = \overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t_k;t_{k+1},\mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{O}, \tag{3.4}$$

and as for the problem (2.11), one can check that $\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}(\mathcal{O}) = \mathcal{O}$ (see Remark 3.1 below).

Then we define $(\mathbf{u}_h^{k+1}, p_h^{k+1}) \in \mathcal{K}_h(\boldsymbol{\zeta}_h^{k+1}) \times M_h(\boldsymbol{\zeta}_h^{k+1})$ as the solution of the problem

$$\left(\frac{\mathbf{u}_h^{k+1} - \mathbf{u}_h^k \circ \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}}{\Delta t}, \boldsymbol{\varphi}\right) + a(\mathbf{u}_h^{k+1}, \boldsymbol{\varphi}) + b(\boldsymbol{\varphi}, p_h^{k+1}) = (\mathbf{f}_h^{k+1}, \boldsymbol{\varphi}) \quad \forall \boldsymbol{\varphi} \in \mathcal{K}_h(\boldsymbol{\zeta}_h^{k+1}), \tag{3.5}$$

$$b(\mathbf{u}_h^{k+1}, q) = 0 \quad \forall q \in M_h(\boldsymbol{\zeta}_h^{k+1}), \tag{3.6}$$

where $\mathbf{f}_h^{k+1}$ is the $\mathcal{L}^2$-projection of $\mathbf{f}^{k+1} = \mathbf{f}(t_{k+1})$ on $(E_h)^2$. We take $\boldsymbol{\zeta}_h^0 = \boldsymbol{\zeta}^0$, and the initial approximate velocity $\mathbf{u}_h^0$ is the $\mathcal{H}_0^1$-projection of $\mathbf{u}_0$ onto $\mathcal{K}_h(\boldsymbol{\zeta}_h^0)$.

Remark 3.1. In (3.3), we use the projection of $\mathbf{u}_h^k$ on $\mathcal{R}_h$ rather than the function $\mathbf{u}_h^k$ itself because $\operatorname{div}(\mathbf{P}\mathbf{u}_h^k) = 0$ in $\mathcal{O}$. By using a classical result of Liouville, this implies that $\det \mathbf{J}_{\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}} = 1$ and in particular that $\det \mathbf{J}_{\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}} = 1$. This property, combined with the fact that the velocity field $\mathbf{P}\mathbf{u}_h^k$ vanishes along the boundary $\partial\mathcal{O}$, entails the invariance property of the whole domain $\mathcal{O}$ through $\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}$, i.e., $\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}(\mathcal{O}) = \mathcal{O}$. Moreover, since $\mathbf{P}\mathbf{u}_h^k$ is constant in each triangle, the initial value problem (3.3) can be solved exactly.

In what follows, we suppose that

$$\begin{aligned} &\mathbf{f} \in C([0,T];\mathcal{H}^1(\mathcal{O})), \quad \mathbf{u}_0 \in \mathcal{H}^2(\Omega), \quad \operatorname{div}(\mathbf{u}_0) = 0 \quad \text{in } \Omega,\\ &\mathbf{u}_0 = 0 \quad \text{on } \partial\mathcal{O}, \quad \mathbf{u}_0(\mathbf{y}) = \boldsymbol{\zeta}_1 + \omega_0(\mathbf{y} - \boldsymbol{\zeta}_0)^\perp \quad \text{on } \partial B. \end{aligned} \tag{3.7}$$

The corresponding solution $(\mathbf{u}, p, \boldsymbol{\zeta}, \omega)$ of problem (1.1)–(1.8) will be assumed to satisfy the following regularity hypotheses:

$$\left\{\begin{aligned} &\mathbf{u} \in C([0,T];\mathcal{H}^2(\Omega(t))) \cap H^1(0,T;\mathcal{L}^2(\Omega(t))),\\ &D_t^2\mathbf{u} \in L^2(0,T;\mathcal{L}^2(\Omega(t))), \quad \mathbf{u} \in C([0,T];\mathcal{C}^{0,1}(\overline{\mathcal{O}}))\\ &p \in C([0,T];\mathcal{H}^1(\Omega(t))), \quad \boldsymbol{\zeta} \in \mathcal{H}^3(0,T), \quad \omega \in H^2(0,T). \end{aligned}\right. \tag{3.8}$$

Moreover, we assume that

$$\operatorname{dist}(B(t), \partial\mathcal{O}) > 0 \quad \forall t \in [0,T]. \tag{3.9}$$

The hypotheses (3.8) and (3.9) imply the existence of $\eta > 0$ such that

$$\operatorname{dist}(B(t), \partial\mathcal{O}) > 3\eta \quad \forall t \in [0,T]. \tag{3.10}$$

THEOREM 3.2. *Let $C_0 > 0$ be a fixed constant. Suppose that $\mathcal{O}$ is the interior of a convex polygon and that $(\mathbf{u}, p, \boldsymbol{\zeta}, \omega)$ is a solution of* (1.1)–(1.8) *satisfying* (3.8) *and* (3.9)*. Moreover, assume that $\mathbf{f}$ and $\mathbf{u}_0$ satisfy* (3.7)*. Consider the functions $\boldsymbol{\zeta}_h^k$, $\mathbf{u}_h^k$, and p_h^k defined in this section. Then there exist two positive constants C and τ^* not depending on h and Δt such that for all $0 < \Delta t \leqslant \tau^*$ and for all $h \leqslant C_0 (\Delta t)^2$ we have*

$$\sup_{1 \leqslant k \leqslant N} \left(|\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \|\mathbf{u}(t_k) - \mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})} \right) \leqslant C\Delta t.$$

Remark 3.3. For the Navier–Stokes system, the same type of result is obtained in [23] for $h \leqslant C_0\Delta t$ and in [26] for $h^2 \leqslant C_0 \Delta t \leqslant C_1 h^\sigma$ and $\sigma > 1/2$ (for h and Δt small enough).

Remark 3.4. It can be easily checked, by using the fact that $\det \mathbf{J}_{\overline{\boldsymbol{\psi}_h^k}} = 1$, that our method is unconditionally stable.

4. Some properties of the finite element spaces. Next we give some technical results on the finite element spaces introduced above. Throughout this section we consider $\boldsymbol{\zeta} \in \mathcal{O}$ such that $\operatorname{dist}(B(\boldsymbol{\zeta}), \partial\mathcal{O}) > 2\eta$ and we suppose that $h < \eta$. Therefore, we have that

$$\operatorname{dist}(B(\boldsymbol{\zeta}), \partial\mathcal{O}) > 2h. \tag{4.1}$$

Notice that, by definition, if $q \in M_h(\boldsymbol{\zeta})$, then $q = 0$ in $B(\boldsymbol{\zeta})$. Since q is a P_1 function in each triangle, it follows that $q_{|A_h} = 0$, where

$$A_h = \bigcup_{\substack{T \in \mathcal{T}_h \\ \overset{\circ}{T} \cap \overset{\circ}{B}(\boldsymbol{\zeta}) \neq \emptyset}} T.$$

Moreover, if we denote by Q_h the union of all triangles $T \in \mathcal{T}_h$ such that the three vertices of T are contained in $\overline{A_h}$, then, by using again the fact that q is a P_1 function in each triangle, it follows that

$$q_{|Q_h} = 0 \qquad \forall\, q \in M_h(\boldsymbol{\zeta}).$$

A similar argument shows that

$$\mathbf{D}(\mathbf{u})_{|A_h} = 0 \qquad \forall\, \mathbf{u} \in \mathcal{K}_h(\boldsymbol{\zeta}).$$

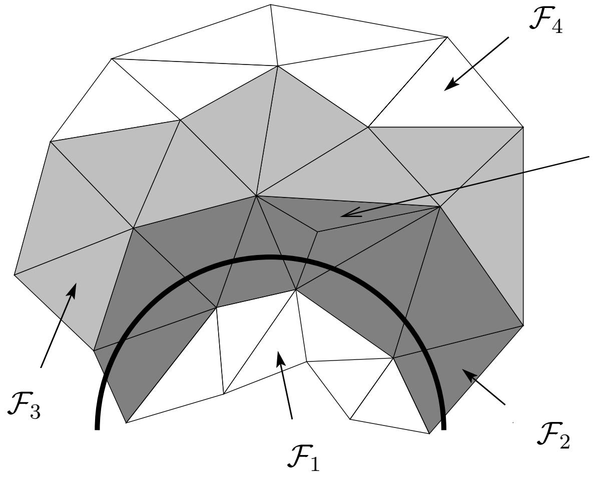
In order to study the properties of the spaces $\mathcal{K}_h(\boldsymbol{\zeta})$ and $M_h(\boldsymbol{\zeta})$ defined above we divide the triangles in $\mathcal{T}_h$ into four categories. These categories are defined as follows (see Figure 1):

- $\mathcal{F}_1$ is the subset of $\mathcal{T}_h$ formed by all triangles $T \in \mathcal{T}_h$ such that $\overline{T} \subset B(\boldsymbol{\zeta})$.
- $\mathcal{F}_2$ is the subset formed by all triangles $T \in \mathcal{T}_h \setminus \mathcal{F}_1$ such that $\overline{T} \subset \overline{Q}_h$.
- $\mathcal{F}_3$ is the subset formed by all triangles $T \in \mathcal{T}_h$ such that $\overline{T} \cap \overline{Q_h} \neq \emptyset$ and $T \not\subset \overline{Q_h}$.
- $\mathcal{F}_4 = \mathcal{T}_h \setminus (\mathcal{F}_1 \cup \mathcal{F}_2 \cup \mathcal{F}_3)$.

LEMMA 4.1. *There exists a positive constant C_1 (not depending on the position of $B(\boldsymbol{\zeta})$) such that*

$$\inf_{\mathbf{v}_h \in \mathcal{K}_h(\boldsymbol{\zeta})} \|\mathbf{v} - \mathbf{v}_h\|_{L^2(\mathcal{O})} \leqslant C_1 h^{\frac{3}{2}} \left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))} \right),$$
$$\inf_{\mathbf{v}_h \in \mathcal{K}_h(\boldsymbol{\zeta})} \|\mathbf{v} - \mathbf{v}_h\|_{\mathcal{H}^1(\mathcal{O})} \leqslant C_1 \sqrt{h} \left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))} \right),$$

for all $\mathbf{v} \in \mathcal{K}(\boldsymbol{\zeta}) \cap \mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}))$.



This kind of triangle is not included in A_h since $T \cap B = \emptyset$, but it is included in Q_h since its three vertices are in $\overline{A_h}$.

FIG. 1. *Splitting of the triangulation into four families of triangles.*

Proof. Let $\mathbf{v} \in \mathcal{K}(\boldsymbol{\zeta}) \cap \mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}))$. This means, in particular, that

$$\mathbf{v}(\mathbf{x}) = \mathbf{l} + \omega \mathbf{x}^{\perp} \qquad \forall\, \mathbf{x} \in B(\boldsymbol{\zeta}),$$

for some $\mathbf{l} \in \mathbb{R}^2$ and $\omega \in \mathbb{R}$. In the remaining part of this section we denote

$$\mathbf{R}(\mathbf{x}) = \mathbf{l} + \omega \mathbf{x}^{\perp} \qquad \forall\, \mathbf{x} \in \mathbb{R}^2.$$

We denote by $\mathbf{v}_{Ih}$ the unique function in $(E_h)^2$ which agrees with $\mathbf{v}$ at every node $\mathbf{x}_j$ of the triangulation $\mathcal{T}_h$ (recall the definition of E_h in (3.1)). Then we consider the function $\mathbf{v}_h \in (E_h)^2$ whose value in a node $\mathbf{x_j}$ of $\mathcal{T}_h$ is defined by

$$\mathbf{v_h}(\mathbf{x_j}) = \begin{cases} \mathbf{R}(\mathbf{x_j}) & \text{if} \quad \mathbf{x_j} \in \overline{A_h}, \\ \mathbf{v}_{Ih}(\mathbf{x_j}) & \text{if} \quad \mathbf{x_j} \notin \overline{A_h}. \end{cases}$$

Since $\mathbf{v_h}$ is affine in each triangle, it follows that

$$\mathbf{v_h}(\mathbf{x}) = \mathbf{R}(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \overline{Q_h}. \tag{4.2}$$

We will show that there exists a positive constant C_1 (not depending on the position of $B(\boldsymbol{\zeta})$) such that

$$\|\mathbf{v} - \mathbf{v}_h\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C_1 h^{\frac{3}{2}} \left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))} \right), \tag{4.3}$$

$$\|\mathbf{v} - \mathbf{v}_h\|_{\mathcal{H}^1(\mathcal{O})} \leqslant C_1 \sqrt{h} \left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))} \right). \tag{4.4}$$

In order to prove the above inequalities, we divide the domain $\mathcal{O}$ into four parts:

$$\mathcal{O} = B(\boldsymbol{\zeta}) \cup (Q_h \setminus B(\boldsymbol{\zeta})) \cup \left(\bigcup_{T \in \mathcal{F}_3} T \right) \cup \left(\bigcup_{T \in \mathcal{F}_4} T \right).$$

Let us first remark that

$$\mathbf{v} = \mathbf{R} \ \text{ in } B(\boldsymbol{\zeta}). \tag{4.5}$$

On the other hand it is clear that Q_h is contained in the closed ball of center $\boldsymbol{\zeta}$ and radius $1+h$, denoted by $B_h(\boldsymbol{\zeta})$. Let us remark that the ball $B_h(\boldsymbol{\zeta})$ is included in the domain $\mathcal{O}$ due to condition (4.1). According to a classical result (see, for instance, Lemma 5.11 in Fujita and Sauer [10]), there exists a universal constant $C > 0$, such that for all $\boldsymbol{\varphi} \in \mathcal{H}^1(\mathcal{O} \setminus B(\boldsymbol{\zeta}))$,

$$\|\boldsymbol{\varphi}\|_{\mathcal{L}^2(B_h(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))} \leq C\left(\sqrt{h}\|\boldsymbol{\varphi}\|_{\mathcal{L}^2(\partial B(\boldsymbol{\zeta}))} + h\|\nabla\boldsymbol{\varphi}\|_{[L^2(B_h(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))]^4}\right). \tag{4.6}$$

The above relation with $\boldsymbol{\varphi} = \mathbf{v} - \mathbf{R}$ and (4.5) imply that

$$\|\mathbf{v}-\mathbf{R}\|_{\mathcal{L}^2(B_h(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))} \leqslant Ch\,\|\nabla\left(\mathbf{v}-\mathbf{R}\right)\|_{[L^2(B_h(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))]^4}. \tag{4.7}$$

By again applying Lemma 5.11 in [10] (this time for the function $\nabla\left(\mathbf{v}-\mathbf{R}\right)$), we obtain that

$$\begin{aligned}\|\nabla\left(\mathbf{v}-\mathbf{R}\right)\|_{[L^2(B_h(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))]^4} \leqslant C\Big(\sqrt{h}\,\|\nabla\left(\mathbf{v}-\mathbf{R}\right)\|_{[L^2(\partial B(\boldsymbol{\zeta}))]^4} \\ + h\,\|\nabla\left(\mathbf{v}-\mathbf{R}\right)\|_{[H^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}))]^4}\Big).\end{aligned}$$

The above inequality, combined with the trace theorem in Sobolev spaces, gives that

$$\|\nabla\left(\mathbf{v}-\mathbf{R}\right)\|_{[L^2(B_h(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))]^4} \leq C\sqrt{h}\|\mathbf{v}-\mathbf{R}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}. \tag{4.8}$$

From (4.7) and (4.8) it follows that

$$\|\mathbf{v}-\mathbf{R}\|_{\mathcal{L}^2(B_h(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))} \leqslant Ch^{\frac{3}{2}}\|\mathbf{v}-\mathbf{R}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}. \tag{4.9}$$

The above relation implies, by using the fact that $Q_h \subset B_h(\boldsymbol{\zeta})$ and (4.2), that

$$\|\mathbf{v}-\mathbf{v_h}\|_{\mathcal{L}^2(Q_h\setminus B(\boldsymbol{\zeta}))} \leq Ch^{\frac{3}{2}}\|\mathbf{v}-\mathbf{R}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}. \tag{4.10}$$

Consequently, we have that

$$\|\mathbf{v}-\mathbf{v}_h\|_{\mathcal{L}^2(Q_h\setminus B(\boldsymbol{\zeta}))} \leqslant C_1 h^{\frac{3}{2}}\left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))}\right). \tag{4.11}$$

On the other hand, (4.8) and (4.9) imply that

$$\|\mathbf{v}-\mathbf{R}\|_{\mathcal{H}^1(B_h(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))} \leq Ch^{\frac{1}{2}}\|\mathbf{v}-\mathbf{R}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}.$$

The above relation implies, by using the fact that $Q_h \subset B_h(\boldsymbol{\zeta})$ and (4.2), that

$$\|\mathbf{v}-\mathbf{v_h}\|_{\mathcal{H}^1(Q_h\setminus B(\boldsymbol{\zeta}))} \leq Ch^{\frac{1}{2}}\|\mathbf{v}-\mathbf{R}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))},$$

which clearly implies

$$\|\mathbf{v}-\mathbf{v_h}\|_{\mathcal{H}^1(Q_h\setminus B(\boldsymbol{\zeta}))} \leq C\sqrt{h}\left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))}\right). \tag{4.12}$$

Let us now consider a triangle $T \in \mathcal{F}_3$. In order to estimate the restriction of $\mathbf{v}-\mathbf{v_h}$ to T we use the interpolating function $\mathbf{v}_{Ih}$. More precisely, we have

$$\|\mathbf{v}-\mathbf{v}_h\|_\alpha \leq \|\mathbf{v}-\mathbf{v}_{Ih}\|_\alpha + \|\mathbf{v}_{Ih}-\mathbf{v}_h\|_\alpha, \quad \alpha \in \{0,1\}, \tag{4.13}$$

where $\|\cdot\|_\alpha$ stands for the $\mathcal{L}^2$-norm or the $\mathcal{H}^1$-norm on T. We first estimate the second term in the right-hand side of (4.13). Since the function $\mathbf{v}_{Ih} - \mathbf{v}_h$ is affine in T, we have

$$\mathbf{v}_{Ih}(\mathbf{x}) - \mathbf{v}_h(\mathbf{x}) = \sum_{i=1}^{3} (\mathbf{v}_{Ih}(\mathbf{x}_i) - \mathbf{v}_h(\mathbf{x}_i))\, \varphi_i(x),$$

where $(\mathbf{x}_i)$ are the nodes of T and (φ_i) are the corresponding Lagrange barycentric functions. We have

$$\|\mathbf{v}_{Ih} - \mathbf{v}_h\|_\alpha \leq \sum_{i=1}^{3} |\mathbf{v}_{Ih}(\mathbf{x}_i) - \mathbf{v}_h(\mathbf{x}_i)|\, \|\varphi_i\|_\alpha. \tag{4.14}$$

A simple calculation shows that

$$\|\varphi_i\|_{L^2(T)} \leq Ch \tag{4.15}$$

and

$$\|\nabla\varphi_i\|_{L^2(T)} \leq C. \tag{4.16}$$

Since the mesh is quasi-uniform, the constant C can be chosen independent of the triangle. We now estimate $|\mathbf{v}_{Ih}(\mathbf{x}_i) - \mathbf{v}_h(\mathbf{x}_i)|$. Since $T \not\subset Q_h$, it follows that T has at most two nodes in Q_h and, consequently, at least one node such that $\mathbf{v}_{Ih}(\mathbf{x}_i) - \mathbf{v}_h(\mathbf{x}_i) = 0$. Therefore we tackle only the nodes in Q_h. If $\mathbf{x}_i$ is a node in Q_h, then

$$|\mathbf{v}_{Ih}(\mathbf{x}_i) - \mathbf{v}_h(\mathbf{x}_i)| = |\mathbf{v}(\mathbf{x}_i) - \mathbf{R}(\mathbf{x}_i)|\,. \tag{4.17}$$

Relations (4.14), (4.15), and (4.17) imply that

$$\begin{aligned}\|\mathbf{v}_{Ih} - \mathbf{v}_h\|_{\mathcal{L}^2(T)} &\leqslant Ch\, \|\mathbf{v} - \mathbf{R}\|_{\mathcal{L}^\infty(T)} \\ &\leqslant Ch \left(\|\mathbf{v} - \mathbf{v}_{Ih}\|_{\mathcal{L}^\infty(T)} + \|\mathbf{v}_{Ih} - \mathbf{R}\|_{\mathcal{L}^\infty(T)} \right).\end{aligned}$$

By using a classical interpolation error (see, for example, [2, Corollary 4.4.7]) and an inverse estimate (see, for example, [2, Lemma 4.5.3]), the above inequality yields

$$\|\mathbf{v}_{Ih} - \mathbf{v}_h\|_{\mathcal{L}^2(T)} \leqslant Ch \left(h\, \|\mathbf{v}\|_{\mathcal{H}^2(T)} + h^{-1}\, \|\mathbf{v}_{Ih} - \mathbf{R}\|_{\mathcal{L}^2(T)} \right),$$

which implies that

$$\begin{aligned}\|\mathbf{v}_{Ih} - \mathbf{v}_h\|_{\mathcal{L}^2(T)} &\leqslant C \left(h^2\, \|\mathbf{v}\|_{\mathcal{H}^2(T)} + \|\mathbf{v}_{Ih} - \mathbf{v}\|_{\mathcal{L}^2(T)} + \|\mathbf{v} - \mathbf{R}\|_{\mathcal{L}^2(T)} \right) \\ &\leqslant C \left(h^2\, \|\mathbf{v}\|_{\mathcal{H}^2(T)} + \|\mathbf{v} - \mathbf{R}\|_{\mathcal{L}^2(T)} \right).\end{aligned}$$

Above we have used again a classical result on the interpolation error (see, for example, [2, Theorem 4.4.4]).

Now, summing up the above relation for all triangles $T \in \mathcal{F}_3$ we obtain that

$$\|\mathbf{v}_{Ih} - \mathbf{v}_h\|_{\mathcal{L}^2\left(\bigcup_{T\in\mathcal{F}_3} T\right)} \leqslant C \left(h^2\, \|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v} - \mathbf{R}\|_{\mathcal{L}^2\left(\bigcup_{T\in\mathcal{F}_3} T\right)} \right). \tag{4.18}$$

In order to estimate the last term in the right-hand side of (4.18) we proceed as previously by introducing the closed ball $B_{2h}(\boldsymbol{\zeta})$ of center $\boldsymbol{\zeta}$ and radius $1+2h$. This ball is included in $\mathcal{O}$ thanks to (4.1). It is clear that all triangles of $\mathcal{F}_3$ are contained in $B_{2h}(\boldsymbol{\zeta}) \setminus B(\boldsymbol{\zeta})$. Then we can once again use Lemma 5.11 in Fujita and Sauer [10] and prove an estimate similar to (4.9), namely,

$$\|\mathbf{v}-\mathbf{R}\|_{\mathcal{L}^2(B_{2h}(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))} \leqslant Ch^{\frac{3}{2}} \|\mathbf{v}-\mathbf{R}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}. \tag{4.19}$$

From (4.18) and (4.19) we deduce that

$$\|\mathbf{v}_{Ih}-\mathbf{v}_h\|_{\mathcal{L}^2\left(\bigcup_{T\in\mathcal{F}_3} T\right)} \leqslant Ch^{\frac{3}{2}} \left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))}\right). \tag{4.20}$$

The above relation, combined with (4.13) and with an interpolation error estimate (see [2, Theorem 4.4.4]), implies that

$$\|\mathbf{v}-\mathbf{v}_h\|_{\mathcal{L}^2\left(\bigcup_{T\in\mathcal{F}_3} T\right)} \leqslant Ch^{\frac{3}{2}} \left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))}\right). \tag{4.21}$$

Now we turn to the H^1-estimate for the family $\mathcal{F}_3$ of triangles. From the usual inverse inequality (see [2, Lemma 4.5.3]) and the L^2-estimate (4.20) we obtain

$$\|\nabla(\mathbf{v}_{Ih}-\mathbf{v}_h)\|_{\left[L^2\left(\bigcup_{T\in\mathcal{F}_3} T\right)\right]^4} \leq C_1 h^{\frac{1}{2}} \left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))}\right), \tag{4.22}$$

which implies, together with (4.13) and an interpolation error estimate (see [2, Theorem 4.4.4]), that

$$\|\nabla(\mathbf{v}-\mathbf{v}_h)\|_{\left[L^2\left(\bigcup_{T\in\mathcal{F}_3} T\right)\right]^4} \leq C_1 h^{\frac{1}{2}} \left(\|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} + \|\mathbf{v}\|_{\mathcal{H}^2(B(\boldsymbol{\zeta}))}\right). \tag{4.23}$$

Finally, we consider the case of the triangle family $\mathcal{F}_4$. Interpolation error estimates lead to

$$\|\mathbf{v}-\mathbf{v}_h\|_{L^2\left(\bigcup_{T\in\mathcal{F}_4} T\right)} \leq C_1 h^2 \|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} \tag{4.24}$$

and

$$\|\nabla(\mathbf{v}-\mathbf{v}_h)\|_{L^2\left(\bigcup_{T\in\mathcal{F}_4} T\right)} \leq C_1 h \|\mathbf{v}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}. \tag{4.25}$$

Relations (4.11), (4.21), (4.24) and the fact that $\mathbf{v}=\mathbf{v}_h$ in $B(\boldsymbol{\zeta})$ imply (4.3). Moreover, (4.12), (4.23), (4.25) and the fact that $\mathbf{v}=\mathbf{v}_h$ in $B(\boldsymbol{\zeta})$ imply (4.4). □

LEMMA 4.2. *There exists a positive constant C_2 (independent of the position of $B(\boldsymbol{\zeta})$) such that*

$$\inf_{q_h\in M_h(\boldsymbol{\zeta})} \|q-q_h\|_{L^2(\mathcal{O})} \leqslant C_2 h^{\frac{1}{2}} \|q\|_{H^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} \tag{4.26}$$

for all $q \in M(\boldsymbol{\zeta}) \cap H^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}))$.

Proof. The proof of this lemma is similar to that of Lemma 4.1. Consider a function $q \in M(\boldsymbol{\zeta}) \cap H^1(\mathcal{O} \setminus B(\boldsymbol{\zeta}))$. According to a classical result (see, for example, [3, Theorem IX.7]), there exists $\widetilde{q} \in H^1(\mathcal{O})$ such that

$$\widetilde{q}_{|\mathcal{O}\setminus B(\boldsymbol{\zeta})} = q, \quad \|\widetilde{q}\|_{H^1(\mathcal{O})} \leq C\|q\|_{H^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}, \tag{4.27}$$

and it can be proved that we can choose the constant C independent of the position of $B(\boldsymbol{\zeta})$. Moreover, by a classical interpolation argument (see, for example, [2, Theorem 4.4.4]), there exists $\widetilde{q}_h \in E_h$ such that

$$\|\widetilde{q} - \widetilde{q}_h\|_{L^2(\mathcal{O})} \leq Ch\|\widetilde{q}\|_{H^1(\mathcal{O})}.$$

The above relation and (4.27) clearly imply that there exists a constant $C > 0$ such that

$$\|q - \widetilde{q}_h\|_{L^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))} \leq Ch\|q\|_{H^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}. \tag{4.28}$$

Denote by q_h the function in E_h satisfying the conditions

$$q_h(\mathbf{x}_i) = 0 \;\text{ if }\; \mathbf{x}_i \in \overline{A_h},$$

$$q_h(\mathbf{x}_i) = \widetilde{q}_h(\mathbf{x}_i) \;\text{ if }\; \mathbf{x}_i \in \mathcal{T}_h \setminus \overline{A_h}.$$

Then as in the proof of Lemma 4.1, we can show that

$$\|q - q_h\|_{L^2(\mathcal{O})} \leqslant C_2 h^{\frac{1}{2}}\|q\|_{H^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}. \qquad \square$$

We next show that the finite element spaces $\mathcal{K}_h(\boldsymbol{\zeta})$, $M_h(\boldsymbol{\zeta})$ and the bilinear form b satisfy a discrete *inf-sup* condition. This proves in particular that the approximate problem (3.5)–(3.6) is well-posed (see [12, Theorem II.1.1., p. 114]). More precisely, the following result holds.

LEMMA 4.3. *There exists a constant $\beta^* > 0$ such that for all $q_h \in M_h(\boldsymbol{\zeta})$ there exists $\mathbf{u}_h \in \mathcal{K}_h(\boldsymbol{\zeta})$ with*

$$\int_{\mathcal{O}} \operatorname{div}(\mathbf{u}_h) q_h \, \mathrm{d}\mathbf{x} \geqslant \beta^* \|\mathbf{u}_h\|_{\mathcal{H}^1(\mathcal{O})}\|q_h\|_{L^2(\mathcal{O})}. \tag{4.29}$$

Proof. Let $q_h \in M_h(\boldsymbol{\zeta})$. Since $M_h(\boldsymbol{\zeta}) \subset M(\boldsymbol{\zeta})$, Lemma 2.6 yields the existence of $\mathbf{u} \in \mathcal{K}(\boldsymbol{\zeta})$ such that

$$\int_{\mathcal{O}} \operatorname{div}(\mathbf{u}) q_h \, \mathrm{d}\mathbf{x} \geqslant \beta\|\mathbf{u}\|_{\mathcal{H}^1(\mathcal{O})}\|q_h\|_{L^2(\mathcal{O})},$$

with β independent of q_h. In order to prove the conclusion of the lemma it suffices to show the existence of $\mathbf{u}_h \in \mathcal{K}_h(\boldsymbol{\zeta})$ such that

$$\int_{\mathcal{O}} \operatorname{div}(\mathbf{u}_h) q_h \, \mathrm{d}\mathbf{x} = \int_{\mathcal{O}} \operatorname{div}(\mathbf{u}) q_h \, \mathrm{d}\mathbf{x}, \tag{4.30}$$

$$\|\mathbf{u}_h\|_{\mathcal{H}^1(\mathcal{O})} \leqslant C\|\mathbf{u}\|_{\mathcal{H}^1(\mathcal{O})}, \tag{4.31}$$

where C is a constant independent of q_h.

Note that (4.30) is equivalent to

$$\int_{\mathcal{O}} \mathbf{u}_h \cdot \nabla q_h \ \mathrm{d}\mathbf{x} = \int_{\mathcal{O}} \mathbf{u} \cdot \nabla q_h \ \mathrm{d}\mathbf{x}.$$

Since ∇q_h is constant in each triangle and vanishes in any triangle from $\mathcal{F}_1 \cup \mathcal{F}_2$, in order to check (4.30), it suffices to show that

$$\int_T \mathbf{u}_h \ \mathrm{d}\mathbf{x} = \int_T \mathbf{u} \ \mathrm{d}\mathbf{x} \qquad \forall\, T \in \mathcal{F}_3 \cup \mathcal{F}_4. \tag{4.32}$$

Note first that if $\mathbf{u}_h \in \mathcal{K}_h(\boldsymbol{\zeta})$, then for any triangle $T \in \mathcal{T}_h$ of vertices $\mathbf{x}_1,\ \mathbf{x}_2,\ \mathbf{x}_3$ and of corresponding barycentric functions $\varphi_1,\ \varphi_2,\ \varphi_3$, we have

$$\mathbf{u}_h(\mathbf{x}) = \overline{\mathbf{u}}_h(x) + \frac{\varphi_1(\mathbf{x})\varphi_2(\mathbf{x})\varphi_3(\mathbf{x})}{\displaystyle\int_T \varphi_1\varphi_2\varphi_3 \ \mathrm{d}\mathbf{x}}\boldsymbol{\lambda} \qquad \forall\, \mathbf{x} \in T, \tag{4.33}$$

where $\overline{\mathbf{u}}_h \in \mathcal{C}(\overline{\mathcal{O}})$ satisfies

$$\overline{\mathbf{u}}_h(x) = \varphi_1(\mathbf{x})\boldsymbol{\alpha_1} + \varphi_2(\mathbf{x})\boldsymbol{\alpha_2} + \varphi_3(\mathbf{x})\boldsymbol{\alpha_3} \qquad \forall\, \mathbf{x} \in T, \tag{4.34}$$

for some constant vectors $\boldsymbol{\alpha_1},\ \boldsymbol{\alpha_2},\ \boldsymbol{\alpha_3},\ \boldsymbol{\lambda} \in \mathbb{R}^2$ (these constants depend on the triangle T). Notice that, since the restriction of $\mathbf{u}_h$ to triangles in $\mathcal{F}_1 \cup \mathcal{F}_2$ is a rigid velocity field, the constant $\boldsymbol{\lambda}$ in (4.33) is equal to zero for all triangles in $\mathcal{F}_1 \cup \mathcal{F}_2$. If $\overline{\mathbf{u}}_h$ satisfies (4.34) and $T \in \mathcal{F}_3 \cup \mathcal{F}_4$, then condition (4.32) holds provided that

$$\lambda = \int_T (\mathbf{u} - \overline{\mathbf{u}}_h) \ \mathrm{d}\mathbf{x} \qquad \forall\, T \in \mathcal{F}_3 \cup \mathcal{F}_4. \tag{4.35}$$

Some simple calculations show that there exists a constant $C > 0$ (independent of the triangle) such that

$$\left\| \frac{\varphi_1\varphi_2\varphi_3}{\displaystyle\int_T \varphi_1\varphi_2\varphi_3 \ \mathrm{d}\mathbf{x}} \right\|_{H^1(T)} \leqslant \frac{C}{h^2}. \tag{4.36}$$

Moreover, (4.35) and the Cauchy–Schwarz inequality imply that

$$|\lambda| \leqslant Ch\|\mathbf{u} - \overline{\mathbf{u}}_h\|_{\mathcal{L}^2(T)} \qquad \forall\, T \in \mathcal{F}_3 \cup \mathcal{F}_4, \tag{4.37}$$

for some constant C. From (4.33), (4.36), and (4.37) it follows that

$$\|\mathbf{u}_h\|_{\mathcal{H}^1(T)} \leqslant \|\overline{\mathbf{u}}_h\|_{\mathcal{H}^1(T)} + \frac{C}{h}\|\mathbf{u} - \overline{\mathbf{u}}_h\|_{\mathcal{L}^2(T)} \qquad \forall\, T \in \mathcal{F}_3 \cup \mathcal{F}_4. \tag{4.38}$$

The remaining part of the proof is devoted to the construction of $\overline{\mathbf{u}}_h$ such that $\mathbf{u}_h$ satisfies (4.31). According to a classical result (see, for instance, [12, Theorem I.A.2., p. 101]), there exists a function $\overline{\mathbf{u}}_h^c \in C(\overline{\mathcal{O}})$ which is affine in each triangle $T \in \mathcal{T}_h$ such that

$$\|\mathbf{u} - \overline{\mathbf{u}}_h^c\|_{\mathcal{L}^2(T)} \leq Ch\|\mathbf{u}\|_{\mathcal{H}^1(T)}, \tag{4.39}$$

$$\|\overline{\mathbf{u}}_h^c\|_{\mathcal{H}^1(T)} \leq C\|\mathbf{u}\|_{\mathcal{H}^1(T)}, \tag{4.40}$$

with the constant C independent of h. We are now in a position to define $\overline{\mathbf{u}}_h$. This function is defined by

$$\overline{\mathbf{u}_h}(\mathbf{x}) = \begin{cases} \overline{\mathbf{u}}_h^c(\mathbf{x}) & \text{if} \quad \mathbf{x} \in \bigcup\limits_{T\in\mathcal{F}_4} T, \\ \mathbf{R}(\mathbf{x}) & \text{if} \quad \mathbf{x} \in \bigcup\limits_{T\in\mathcal{F}_1\cup\mathcal{F}_2} T, \end{cases} \tag{4.41}$$

where $\mathbf{R}$ is the extension of $\mathbf{u}_{|B(\boldsymbol{\zeta})}$ (which is a rigid velocity field) to $\mathbb{R}^2$. We remark that relation (4.41) also defines the values of $\overline{\mathbf{u}}_h$ in the triangles of $\mathcal{F}_3$. Indeed, the vertices of each triangle in $\mathcal{F}_3$ are also vertices of a triangle in either $\mathcal{F}_2$ or in $\mathcal{F}_4$. In order to prove (4.31) we estimate the terms in the right-hand side of (4.38). We first consider a triangle $T \in \mathcal{F}_4$. By using the fact that $\overline{\mathbf{u}}_h = \overline{\mathbf{u}}_h^c$ in T, (4.39), and (4.40), we obtain that

$$\|\overline{\mathbf{u}}_h\|_{\mathcal{H}^1(T)} + \frac{1}{h}\|\mathbf{u} - \overline{\mathbf{u}}_h\|_{\mathcal{L}^2(T)} \leq C\|\mathbf{u}\|_{\mathcal{H}^1(T)} \qquad \forall\, T \in \mathcal{F}_4, \tag{4.42}$$

with the constant C independent of $\mathbf{u}$. We next consider a triangle $T \in \mathcal{F}_3$. We first notice that

$$\begin{aligned} \|\overline{\mathbf{u}}_h\|_{\mathcal{H}^1(T)} + \frac{1}{h}\|\mathbf{u} - \overline{\mathbf{u}}_h\|_{\mathcal{L}^2(T)} &\leq \|\overline{\mathbf{u}}_h^c\|_{\mathcal{H}^1(T)} + \frac{1}{h}\|\mathbf{u} - \overline{\mathbf{u}}_h^c\|_{\mathcal{L}^2(T)} \\ &\quad + \|\overline{\mathbf{u}}_h^c - \overline{\mathbf{u}}_h\|_{\mathcal{H}^1(T)} + \frac{1}{h}\|\overline{\mathbf{u}}_h - \overline{\mathbf{u}}_h^c\|_{\mathcal{L}^2(T)} \qquad \forall\, T \in \mathcal{F}_3. \end{aligned} \tag{4.43}$$

The first two terms in the right-hand side of (4.43) can be directly estimated by using (4.39) and (4.40). Moreover, by using inverse estimates (see, for example, [2, Lemma 4.5.3]), there exists a positive constant C independent of h such that

$$\|\overline{\mathbf{u}}_h^c - \overline{\mathbf{u}}_h\|_{\mathcal{H}^1(T)} + \frac{1}{h}\|\overline{\mathbf{u}}_h - \overline{\mathbf{u}}_h^c\|_{\mathcal{L}^2(T)} \leqslant C\,\|\overline{\mathbf{u}}_h^c - \overline{\mathbf{u}}_h\|_{\mathcal{L}^\infty(T)} \qquad \forall\, T \in \mathcal{F}_3.$$

The above relation and the fact that $\overline{\mathbf{u}}_h$ is equal either to $\mathbf{R}$ or to $\overline{\mathbf{u}}_h^c$ in the vertices of a triangle $T \in \mathcal{F}_3$ imply that

$$\|\overline{\mathbf{u}}_h^c - \overline{\mathbf{u}}_h\|_{\mathcal{H}^1(T)} + \frac{1}{h}\|\mathbf{u}_h - \overline{\mathbf{u}}_h^c\|_{\mathcal{L}^2(T)} \leq C\,\|\overline{\mathbf{u}}_h^c - \mathbf{R}\|_{\mathcal{L}^\infty(T)} \qquad \forall\, T \in \mathcal{F}_3.$$

The above inequality, combined once again with an inverse inequality, implies that

$$\|\overline{\mathbf{u}}_h^c - \overline{\mathbf{u}}_h\|_{\mathcal{H}^1(T)} + \frac{1}{h}\|\mathbf{u}_h - \overline{\mathbf{u}}_h^c\|_{\mathcal{L}^2(T)} \leq \frac{C}{h}\|\overline{\mathbf{u}}_h^c - \mathbf{R}\|_{\mathcal{L}^2(T)} \qquad \forall\, T \in \mathcal{F}_3. \tag{4.44}$$

On the other hand,

$$\|\overline{\mathbf{u}}_h^c - \mathbf{R}\|_{\mathcal{L}^2(T)} \leq \|\overline{\mathbf{u}}_h^c - \mathbf{u}\|_{\mathcal{L}^2(T)} + \|\mathbf{u} - \mathbf{R}\|_{\mathcal{L}^2(T)} \qquad \forall\, T \in \mathcal{F}_3. \tag{4.45}$$

Combining (4.39), (4.45), (4.44), and (4.43), we obtain

$$\|\overline{\mathbf{u}}_h\|_{\mathcal{H}^1(T)} + \frac{1}{h}\|\mathbf{u} - \overline{\mathbf{u}}_h\|_{\mathcal{L}^2(T)} \leq C\|\mathbf{u}\|_{\mathcal{H}^1(T)} + \frac{C}{h}\|\mathbf{u} - \mathbf{R}\|_{\mathcal{L}^2(T)} \qquad \forall\, T \in \mathcal{F}_3. \tag{4.46}$$

We recall that all triangles of $\mathcal{F}_3$ are contained in $B_{2h}(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta})$. Therefore, by taking the sum of the above relations for all $T \in \mathcal{F}_3$ and by using (4.6), combined with the fact that $\mathbf{u} = \mathbf{R}$ on $\partial B(\boldsymbol{\zeta})$, we obtain

$$\|\overline{\mathbf{u}}_h\|_{\mathcal{H}^1\left(\bigcup_{T\in\mathcal{F}_3} T\right)} + \frac{1}{h}\|\mathbf{u}-\overline{\mathbf{u}}_h\|_{\mathcal{L}^2\left(\bigcup_{T\in\mathcal{F}_3} T\right)} \leq C\|\mathbf{u}\|_{\mathcal{H}^1(B_{2h}(\boldsymbol{\zeta})\setminus B(\boldsymbol{\zeta}))}. \tag{4.47}$$

Now by combining (4.42) and (4.47) in (4.38), we obtain

$$\|\mathbf{u}_h\|_{\mathcal{H}^1\left(\bigcup_{T\in\mathcal{F}_3\cup\mathcal{F}_4} T\right)} \leq C\|\mathbf{u}\|_{\mathcal{H}^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}))}. \tag{4.48}$$

We next consider the triangles $T \in \mathcal{F}_1 \cup \mathcal{F}_2$. By using the fact that $\mathbf{u}_h = \overline{\mathbf{u}}_h = \mathbf{R}$ in T, we obtain that

$$\|\mathbf{u}_h\|_{\mathcal{H}^1\left(\bigcup_{T\in\mathcal{F}_1\cup\mathcal{F}_2} T\right)} = \|\mathbf{R}\|_{\mathcal{H}^1\left(\bigcup_{T\in\mathcal{F}_1\cup\mathcal{F}_2} T\right)}.$$

A simple calculation shows that the right-hand side of the above relation is bounded by $C\|\mathbf{u}\|_{\mathcal{H}^1(B(\boldsymbol{\zeta}))}$, where C is a constant independent of h. We thus obtain

$$\|\mathbf{u}_h\|_{\mathcal{H}^1\left(\bigcup_{T\in\mathcal{F}_1\cup\mathcal{F}_2} T\right)} \leq C\|\mathbf{u}\|_{\mathcal{H}^1(B(\boldsymbol{\zeta}))}. \tag{4.49}$$

If we join (4.48) and (4.49), we see that the function $\mathbf{u}_h$ satisfies (4.31). This concludes the proof of the lemma. □

Now, we are in position to introduce a projector in $\mathcal{K}_h(\boldsymbol{\zeta}) \times M_h(\boldsymbol{\zeta})$ that will be a key ingredient in the proof of the convergence result.

LEMMA 4.4. *Suppose that* $\mathbf{V} \in \mathcal{K}(\boldsymbol{\zeta})$ *and that* $P \in M(\boldsymbol{\zeta})$. *Then there exists a unique couple* $(\mathbf{V}_h, P_h)$ *in* $\mathcal{K}_h(\boldsymbol{\zeta}) \times M_h(\boldsymbol{\zeta})$ *such that*

$$\left\{\begin{array}{rcll} a\left(\mathbf{V}-\mathbf{V}_h, \boldsymbol{\varphi}\right) + b\left(\boldsymbol{\varphi}, P-P_h\right) & = & 0 & \forall\, \boldsymbol{\varphi} \in \mathcal{K}_h(\boldsymbol{\zeta}), \\ b\left(\mathbf{V}-\mathbf{V}_h, q\right) & = & 0 & \forall\, q \in M_h(\boldsymbol{\zeta}). \end{array}\right. \tag{4.50}$$

Moreover, if we suppose in addition that $\mathbf{V}_{|\mathcal{O}\setminus B(\boldsymbol{\zeta})} \in \mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}))$ *and that* $P_{|\mathcal{O}\setminus B(\boldsymbol{\zeta})} \in H^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}))$, *then there exists a positive constant* C *such that*

$$\|\mathbf{V}-\mathbf{V}_h\|_{\mathcal{L}^2(\mathcal{O})} \leqslant Ch.$$

Proof. The result in Lemma 4.3 combined with Theorem 1.1 in [12, p. 114] implies the existence and uniqueness of $(\mathbf{V}_h, P_h)$ in $\mathcal{K}_h(\boldsymbol{\zeta}) \times M_h(\boldsymbol{\zeta})$, satisfying (4.50) together with

$$\|\mathbf{V}-\mathbf{V}_h\|_{\mathcal{H}^1(\mathcal{O})}+\|P-P_h\|_{L^2(\mathcal{O})} \leqslant C\left\{\inf_{\mathbf{v}\in\mathcal{K}_h(\boldsymbol{\zeta})}\|\mathbf{V}-\mathbf{v}\|_{\mathcal{H}^1(\mathcal{O})} + \inf_{q\in M_h(\boldsymbol{\zeta})}\|P-q\|_{L^2(\mathcal{O})}\right\}.$$

Using Lemmas 4.1 and 4.2, we obtain

$$\|\mathbf{V}-\mathbf{V}_h\|_{\mathcal{H}^1(\mathcal{O})} + \|P-P_h\|_{L^2(\mathcal{O})} \leqslant Ch^{1/2}\left\{\|\mathbf{V}\|_{\mathcal{H}^2(\mathcal{O}\setminus B)} + \|\mathbf{V}\|_{\mathcal{H}^2(B)} + \|P\|_{H^1(\mathcal{O})}\right\}.$$

Moreover, by applying the usual Aubin–Nitsche duality argument (see, for example, [12, p. 119]), one can easily prove

$$\|\mathbf{V}-\mathbf{V}_h\|_{\mathcal{L}^2(\mathcal{O})} \leqslant Ch\left\{\|\mathbf{V}\|_{\mathcal{H}^2(\mathcal{O}\setminus B)} + \|\mathbf{V}\|_{\mathcal{H}^2(B)} + \|P\|_{H^1(\mathcal{O})}\right\}. \qquad □$$

5. Definition and properties of the change of variables. In order to prove Theorem 3.2, we should be able to compare the exact solution, which is rigid in the ball $B(\boldsymbol{\zeta}(t_k))$, with the approximate solution, which is rigid in the ball $B(\boldsymbol{\zeta}_h^k)$. This will be achieved by the use of a change of variables that maps the exact ball onto the approximate one. This section is devoted to the description and main properties of this transformation.

5.1. Change of variables. In this section, we suppose that $\mathcal{O}$ is convex. In what follows, we need a change of variables, transforming a function in $\widehat{\mathcal{K}}(\boldsymbol{\zeta}_1)$ into a function in $\widehat{\mathcal{K}}(\boldsymbol{\zeta}_2)$, where $\boldsymbol{\zeta}_i \in \mathcal{O}$ are such that

$$\operatorname{dist}(\boldsymbol{\zeta}_i, \partial\mathcal{O}) > 1 + 2\eta, \quad i \in \{1,2\}, \quad \text{with } \eta > 0. \tag{5.1}$$

In this case, $B(\boldsymbol{\zeta}_i)$ is contained in $\mathcal{O}$ and the distance between $B(\boldsymbol{\zeta}_i)$ and $\partial\mathcal{O}$ is greater than 2η. Let $\xi \in C^\infty(\mathbb{R}^2, \mathbb{R})$ be a compactly supported function such that

- $\xi = 1$ if $\mathbf{x} \in \mathcal{O}$ and $\operatorname{dist}(\mathbf{x}, \partial\mathcal{O}) > 2\eta$,
- $\xi = 0$ if $\mathbf{x} \notin \mathcal{O}$ or $\operatorname{dist}(\mathbf{x}, \partial\mathcal{O}) \leqslant \eta$.

Let $\boldsymbol{\Lambda}$ be the mapping defined by

$$\boldsymbol{\Lambda}(\mathbf{x}) = \left[(\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2) \cdot \mathbf{x}^\perp\right] (\mathbf{rot}\ \xi) + \xi(\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2) \quad \forall \mathbf{x} \in \mathbb{R}^2. \tag{5.2}$$

We need several properties of the field $\boldsymbol{\Lambda}$ and of the associated flow. Since these properties are similar to those proved in [27] we state them here without proof.

Lemma 5.1. *Let $\boldsymbol{\Lambda}$ be the mapping defined by* (5.2)*. Then we have*

(i) $\boldsymbol{\Lambda} = 0$ *outside* $\mathcal{O}$,

(ii) $\operatorname{div} \boldsymbol{\Lambda} = 0$ *in* $\mathbb{R}^2$,

(iii) $\boldsymbol{\Lambda}(\mathbf{x}) = \boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2$ *if* $\mathbf{x} \in \mathcal{O}$ *and if* $\operatorname{dist}(\mathbf{x}, \partial\mathcal{O}) > 2\eta$.

In other words, the restriction of $\boldsymbol{\Lambda}$ to a neighborhood of $\partial\mathcal{O}$ is zero and $\boldsymbol{\Lambda}$ is a translation when restricted to points of $\mathcal{O}$ at distance to $\partial\mathcal{O}$ larger than 2η.

We consider next the initial value problem

$$\begin{cases} \dfrac{d}{d\lambda}\boldsymbol{\psi}(\lambda) = \boldsymbol{\Lambda}(\boldsymbol{\psi}(\lambda)), \quad \lambda > 0, \\ \boldsymbol{\psi}(0) = \mathbf{y}, \end{cases} \tag{5.3}$$

with $\boldsymbol{\Lambda}$ given by (5.2).

Lemma 5.2. *For all $\mathbf{y} \in \mathbb{R}^2$, the initial value problem* (5.3) *admits a unique solution $\boldsymbol{\psi}(\lambda, \mathbf{y})$ on* $[0,1]$*. Denote*

$$\mathbf{X}_{\boldsymbol{\zeta}_2,\boldsymbol{\zeta}_1}(\mathbf{y}) = \mathbf{X}(\mathbf{y}) = \boldsymbol{\psi}(1, \mathbf{y}). \tag{5.4}$$

Then $\mathbf{X}$ is a C^∞-diffeomorphism from $\mathcal{O}$ onto itself, and $\mathbf{X}(B(\boldsymbol{\zeta}_2)) = B(\boldsymbol{\zeta}_1)$. If we denote by

$$\mathbf{J_X} = \left(\frac{\partial X_i}{\partial y_j}\right)_{i,j}$$

the jacobian matrix of the transformation $\mathbf{y} \mapsto \mathbf{X}(\mathbf{y})$, then the above change of variables satisfies

$$\det \mathbf{J_X}(\mathbf{y}) = 1 \quad \forall \mathbf{y} \in \mathbb{R}^2. \tag{5.5}$$

We denote by

$$\mathbf{Y}_{\boldsymbol{\zeta}_2,\boldsymbol{\zeta}_1} = \mathbf{Y} = \mathbf{X}^{-1} \tag{5.6}$$

the inverse of $\mathbf{X}$ on $\mathcal{O}$.

5.2. Properties of the change of variables. In this subsection, we use the change of variables defined by the mapping $\mathbf{X}$ in Lemma 5.2 to transform functions in $\widehat{\mathcal{K}}(\boldsymbol{\zeta}_1)$ (resp., $\mathcal{K}(\boldsymbol{\zeta}_1)$, $M(\boldsymbol{\zeta}_1)$) into functions in $\widehat{\mathcal{K}}(\boldsymbol{\zeta}_2)$ (resp., $\mathcal{K}(\boldsymbol{\zeta}_2)$, $M(\boldsymbol{\zeta}_2)$). We also give the expressions of $\Delta\mathbf{u}$ and ∇p after the transformation.

Consider $(\mathbf{u},\ p) \in \mathcal{H}^1(\mathcal{O}) \times L^2(\mathcal{O})$ and define as in [18] the functions $(\mathbf{U},\ P) \in \mathcal{H}^1(\mathcal{O}) \times L^2(\mathcal{O})$ by

$$\mathbf{U}(\mathbf{y}) = \mathbf{J}_{\mathbf{Y}}(\mathbf{X}(\mathbf{y}))\mathbf{u}(\mathbf{X}(\mathbf{y})) \quad \forall \mathbf{y} \in \mathcal{O}, \tag{5.7}$$

$$P(\mathbf{y}) = p(\mathbf{X}(\mathbf{y})) \quad \forall \mathbf{y} \in \mathcal{O}. \tag{5.8}$$

We can easily check, by using the definition of $\boldsymbol{\Lambda}$, that

$$\mathbf{X}(\mathbf{y}) = \mathbf{y} + \boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2 \quad \forall \mathbf{y} \in B(\boldsymbol{\zeta}_2), \tag{5.9}$$

$$\mathbf{Y}(\mathbf{x}) = \mathbf{x} - \boldsymbol{\zeta}_1 + \boldsymbol{\zeta}_2 \quad \forall \mathbf{x} \in B(\boldsymbol{\zeta}_1), \tag{5.10}$$

Consequently, if $\mathbf{u} \in \mathcal{K}(\boldsymbol{\zeta}_1)$, then $\mathbf{U} \in \mathcal{K}(\boldsymbol{\zeta}_2)$ and if $p \in M(\boldsymbol{\zeta}_1)$, then $P \in M(\boldsymbol{\zeta}_2)$.

By using (5.5), we obtain the following result (see, for instance, [18, Proposition 2.4]).

LEMMA 5.3. *If* $\mathbf{X}$ *is defined by* (5.4)*, then for all* $\mathbf{u} \in \mathcal{H}^1(\mathcal{O})$*, the function* $\mathbf{U}$ *defined as above satisfies the relation*

$$\operatorname{div}\,[\mathbf{U}(\mathbf{y})] = \operatorname{div}\,[\mathbf{u}(\mathbf{X}(\mathbf{y}))] \quad \forall \mathbf{y} \in \mathcal{O}.$$

This lemma implies in particular that if $\mathbf{u} \in \widehat{\mathcal{K}}(\boldsymbol{\zeta}_1)$, then $\mathbf{U} \in \widehat{\mathcal{K}}(\boldsymbol{\zeta}_2)$.

In order to write down the expressions of $\Delta\mathbf{u}$ and ∇p after the change of variables, we define (see [18])

$$\begin{aligned}[\mathbf{L}\mathbf{U}]_i = \sum_{j,k} \frac{\partial}{\partial y_j}\left(g^{jk}\frac{\partial U_i}{\partial y_k}\right) + 2\sum_{j,k,l} g^{kl}\Gamma^i_{jk}\frac{\partial U_j}{\partial y_l} \\ + \sum_{j,k,l}\left\{\frac{\partial}{\partial y_k}(g^{kl}\Gamma^i_{jl}) + \sum_m g^{kl}\Gamma^m_{jl}\Gamma^i_{km}\right\}U_j,\end{aligned} \tag{5.11}$$

$$[\mathbf{G}P]_i = \sum_{j=1}^{2} g^{ij}\frac{\partial P}{\partial y_j}, \tag{5.12}$$

where we denote (see, for instance, [7])

$$g^{ij} = \sum_k \frac{\partial Y_i}{\partial x_k}\frac{\partial Y_j}{\partial x_k} \quad \text{(metric contravariant tensor)}, \tag{5.13}$$

$$g_{ij} = \sum_k \frac{\partial X_k}{\partial y_i}\frac{\partial X_k}{\partial y_j} \quad \text{(metric covariant tensor)}, \tag{5.14}$$

and

$$\Gamma^k_{ij} = \frac{1}{2}\sum_l g^{kl}\left\{\frac{\partial g_{il}}{\partial y_j} + \frac{\partial g_{jl}}{\partial y_i} - \frac{\partial g_{ij}}{\partial y_l}\right\} \quad \text{(Christoffel symbol)}. \tag{5.15}$$

We are now in position to write down the expressions of $\Delta\mathbf{u}$ and ∇p after the change of variables (see again [18] for details).

Proposition 5.4. *Suppose that*

$$(\mathbf{u}, p) \in \mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}_1)) \times H^1(\mathcal{O} \setminus B(\boldsymbol{\zeta}_1)).$$

Then, we have that

$$(\mathbf{U}, P) \in \mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}_2)) \times H^1(\mathcal{O} \setminus B(\boldsymbol{\zeta}_2)).$$

Moreover, for all $\mathbf{y} \in \mathcal{O} \setminus B(\boldsymbol{\zeta}_2)$, *we have that*

$$[\mathbf{L}\mathbf{U}](\mathbf{y}) = \mathbf{J}_{\mathbf{Y}}(\mathbf{X}(\mathbf{y}))\ [(\Delta \mathbf{u}) \circ \mathbf{X}]\,(\mathbf{y}), \quad [\mathbf{G}P](\mathbf{y}) = \mathbf{J}_{\mathbf{Y}}(\mathbf{X}(\mathbf{y}))\ [(\nabla p) \circ \mathbf{X}]\,(\mathbf{y}).$$

In the remaining part of this section, we denote by C a positive constant which may depend only on ξ and $\mathcal{O}$. We give below (without proofs) several estimates of the dependence of the change of variables defined in (5.4) on the points $\boldsymbol{\zeta}_1$ and $\boldsymbol{\zeta}_2$. For the proofs of these estimates, we refer to [27] and [28].

Lemma 5.5. *Let* $\boldsymbol{\Lambda}$ *be the function defined by* (5.2). *Then, for all* $\boldsymbol{\zeta}_1,\ \boldsymbol{\zeta}_2 \in \mathcal{O}$ *satisfying* (5.1) *we have*

$$\|\boldsymbol{\Lambda}\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,, \quad \|\nabla \boldsymbol{\Lambda}\|_{[L^\infty(\mathcal{O})]^4} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,,$$

$$\left\|\frac{\partial^2 \boldsymbol{\Lambda}}{\partial x_i \partial x_j}\right\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,, \quad \left\|\frac{\partial^3 \boldsymbol{\Lambda}}{\partial x_i \partial x_j \partial x_k}\right\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,.$$

Lemma 5.6. *Let* $\boldsymbol{\Lambda}$, $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$ *be as in Lemma* 5.5. *Then the functions* $\mathbf{X}$ *and* $\mathbf{Y}$ *defined by* (5.4) *and* (5.6) *satisfy the following inequalities:*

$$\|\mathbf{X}\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C, \quad \|\mathbf{Y}\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C,$$

$$\|\mathbf{J}_{\mathbf{X}} - \mathbf{Id}\|_{[L^\infty(\mathcal{O})]^4} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,, \quad \|\mathbf{J}_{\mathbf{Y}} - \mathbf{Id}\|_{[L^\infty(\mathcal{O})]^4} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,,$$

$$\left\|\frac{\partial^2 Y_i}{\partial x_j \partial x_k}\right\|_{L^\infty(\mathcal{O})} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,, \quad \left\|\frac{\partial^2 X_i}{\partial y_j \partial y_k}\right\|_{L^\infty(\mathcal{O})} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,,$$

$$\left\|\frac{\partial^3 Y_i}{\partial x_j \partial x_l \partial x_k}\right\|_{L^\infty(\mathcal{O})} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,, \quad \left\|\frac{\partial^3 X_i}{\partial y_j \partial y_l \partial y_k}\right\|_{L^\infty(\mathcal{O})} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\,.$$

Lemma 5.7. *Let* $\boldsymbol{\Lambda}$, $\boldsymbol{\zeta}_1$, $\boldsymbol{\zeta}_2$ *be as in Lemma* 5.5. *Moreover, suppose that*

$$(\mathbf{U}, P) \in \mathcal{H}^2(\mathcal{O} \setminus B(\boldsymbol{\zeta}_2)) \times H^1(\mathcal{O} \setminus B(\boldsymbol{\zeta}_2))$$

and that $\mathbf{L}$ *and* $\mathbf{G}$ *are given by* (5.11) *and* (5.12). *Then we have*

(i) $\|\nu[(\mathbf{L} - \Delta)\mathbf{U}]\|_{\mathcal{L}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}_2))} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\ \ \|\mathbf{U}\|_{\mathcal{H}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}_2))}$,

(ii) $\|[(\nabla - \mathbf{G})P]\|_{\mathcal{L}^2(\mathcal{O}\setminus B(\boldsymbol{\zeta}_2))} \leqslant C\,|\boldsymbol{\zeta}_1 - \boldsymbol{\zeta}_2|\ \ \|P\|_{H^1(\mathcal{O}\setminus B(\boldsymbol{\zeta}_2))}$.

6. Consistency of the fully discretized scheme. This section is devoted to the consistency of our fully discretized scheme. The main result in this section asserts that the solution $(\mathbf{u}, p, \boldsymbol{\zeta}, \omega)$ of (1.1)–(1.8) satisfies the scheme (3.2)–(3.6) with consistency errors that will be estimated. Since $(\mathbf{u}(t_k), p(t_k))$ belongs to $\mathcal{K}(\boldsymbol{\zeta}(t_k)) \times M(\boldsymbol{\zeta}(t_k))$ and not to $\mathcal{K}(\boldsymbol{\zeta}_h^k) \times M(\boldsymbol{\zeta}_h^k)$, we need the change of variables introduced in the previous section.

6.1. Consistency in time. In this subsection we show that the exact values at instants $t = t_k$ of a strong solution of (1.1)–(1.8) satisfy a perturbed version of the semidiscretized problem introduced in subsection 2.2 and we estimate these perturbations with respect to the time step. The precise statement is given in Lemma 6.1 below.

Consider the solution $(\mathbf{u}, p, \boldsymbol{\zeta}, \omega)$ of (1.1)–(1.8) and assume that (3.8) and (3.10) hold. In what follows, we will use the notation

$$\widetilde{\boldsymbol{X}}(\mathbf{x}) = \widetilde{\boldsymbol{\psi}}(t_k; t_{k+1}, \mathbf{x}) \quad \forall \mathbf{x} \in \mathcal{O}, \tag{6.1}$$

where $\widetilde{\boldsymbol{\psi}}$ is defined by relation (2.6). Note that $\widetilde{\boldsymbol{X}}(\mathcal{O}) = \mathcal{O}$.

Let $\boldsymbol{\varepsilon}_k$, $\boldsymbol{\delta}_k$, $\boldsymbol{\alpha}_k$, $\boldsymbol{\beta}_k$, γ_k be quantities defined by

$$\boldsymbol{\varepsilon}_k = \boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}'(t_k)\Delta t, \tag{6.2}$$

$$\boldsymbol{\delta}_k(t, \mathbf{x}) = \mathbf{u}(\widetilde{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{x}), t) - \mathbf{u}(\widetilde{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{x}), t_k), \tag{6.3}$$

$$\boldsymbol{\alpha}_k = \frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k) \circ \widetilde{\boldsymbol{X}}}{\Delta t} - \frac{d}{dt}\left[\mathbf{u} \circ \widetilde{\boldsymbol{\psi}}\right](t_{k+1}), \tag{6.4}$$

$$\boldsymbol{\beta}_k = \frac{\boldsymbol{\zeta}'(t_{k+1}) - \boldsymbol{\zeta}'(t_k)}{\Delta t} - \boldsymbol{\zeta}''(t_{k+1}), \tag{6.5}$$

$$\gamma_k = \frac{\omega'(t_{k+1}) - \omega'(t_k)}{\Delta t} - \omega''(t_{k+1}). \tag{6.6}$$

By using the fact that $\mathbf{u}\left(\boldsymbol{\zeta}(t_k), t_k\right) = \boldsymbol{\zeta}'(t_k)$ and relations (2.6), (1.1), (1.5), and (1.6) together with the above definitions, we infer that the exact solution $(\mathbf{u}, p, \boldsymbol{\zeta}, \omega)$ satisfies

$$\boldsymbol{\zeta}(t_{k+1}) = \boldsymbol{\zeta}(t_k) + \mathbf{u}\left(\boldsymbol{\zeta}(t_k), t_k\right)\Delta t + \boldsymbol{\varepsilon}_k, \tag{6.7}$$

$$\left\{\begin{aligned} &\frac{d}{dt}\widetilde{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{x}) = \mathbf{u}\left(\widetilde{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{x}), t_k\right) + \boldsymbol{\delta}_k(t, \mathbf{x}), \\ &\widetilde{\boldsymbol{\psi}}(t_{k+1}; t_{k+1}, \mathbf{x}) = \mathbf{x} \end{aligned}\right. \tag{6.8}$$

for all $\mathbf{x} \in \mathcal{O}$ and for all $t \in [t_k, t_{k+1}]$, together with

$$\frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k) \circ \widetilde{\boldsymbol{X}}}{\Delta t} - \nu\Delta\mathbf{u}(t_{k+1}) + \nabla p(t_{k+1}) = \mathbf{f}^{k+1} + \boldsymbol{\alpha}_k \quad \text{in } \mathcal{O} \setminus B(\boldsymbol{\zeta}(t_{k+1})), \tag{6.9}$$

$$M\frac{\boldsymbol{\zeta}'(t_{k+1}) - \boldsymbol{\zeta}'(t_k)}{\Delta t} = -\int_{\partial B(\boldsymbol{\zeta}(t_{k+1}))} \boldsymbol{\sigma}(t_{k+1})\mathbf{n}\ \mathrm{d}\Gamma + \int_{B(\boldsymbol{\zeta}(t_{k+1}))} \mathbf{f}^{k+1}\ \mathrm{d}\mathbf{x} + \boldsymbol{\beta}_k, \tag{6.10}$$

$$\begin{aligned} J\frac{\omega(t_{k+1}) - \omega(t_k)}{\Delta t} = &-\int_{\partial B(\boldsymbol{\zeta}(t_{k+1}))} (\mathbf{y} - \boldsymbol{\zeta}(t_{k+1}))^{\perp} \cdot \boldsymbol{\sigma}(t_{k+1})\mathbf{n}\ \mathrm{d}\Gamma \\ &+ \int_{B(\boldsymbol{\zeta}(t_{k+1}))} (\mathbf{y} - \boldsymbol{\zeta}(t_{k+1}))^{\perp} \cdot \mathbf{f}^{k+1}\ \mathrm{d}\mathbf{x} + \gamma_k. \end{aligned} \tag{6.11}$$

Moreover, if we denote

$$\theta(t) = \int_0^t \omega(s)\ \mathrm{d}s$$

and by $\mathbf{R}_\theta$ the rotation matrix of angle θ, then we also define the matrix $\mathbf{E}_k$ by

$$\mathbf{R}_{\theta(t_{k+1})-\theta(t_k)} = \mathbf{Id} - \Delta t\,\omega(t_{k+1})\mathbf{R}_{-\pi/2} + \mathbf{E}_k. \tag{6.12}$$

By using the Taylor–Lagrange inequality, we easily obtain the following consistency error estimates.

Lemma 6.1. *The elements $\boldsymbol{\alpha}_k$, $\boldsymbol{\beta}_k$, γ_k, $\boldsymbol{\delta}_k$, $\boldsymbol{\varepsilon}_k$, and $\mathbf{E}_k$ defined by* (6.2)–(6.6) *satisfy the following inequalities:*

$$\begin{gathered} |\boldsymbol{\varepsilon}_k| \leqslant C\,(\Delta t)^2, \quad \|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} \leqslant C\Delta t \left\|\frac{\partial \mathbf{u}}{\partial t}\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}, \\ \|\boldsymbol{\alpha}_k\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\sqrt{\Delta t}\left\|\frac{d^2}{dt^2}[\mathbf{u}\circ\widetilde{\boldsymbol{\psi}}]\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}, \\ |\boldsymbol{\beta}_k| \leqslant C\Delta t, \quad |\gamma_k| \leqslant C\Delta t, \quad |\mathbf{E}_k| \leqslant C\,(\Delta t)^2. \end{gathered} \tag{6.13}$$

6.2. Transformed system. We need to compare $\mathbf{u}(t_k) \in \mathcal{K}(\boldsymbol{\zeta}(t_k))$, which is a rigid velocity field in $B(\boldsymbol{\zeta}(t_k))$, with $\mathbf{u}_h^k \in \mathcal{K}(\boldsymbol{\zeta}_h^k)$, which is a rigid velocity field in $B(\boldsymbol{\zeta}_h^k)$. This will be done by using the change of variables introduced in section 5.1. To this end, we suppose that $|\boldsymbol{\zeta}_h^k - \boldsymbol{\zeta}(t_k)| < \eta$. This hypothesis and (3.10) imply that

$$\operatorname{dist}(B(\boldsymbol{\zeta}(t_k)), \partial\mathcal{O}) > 2\eta. \tag{6.14}$$

With this assumption, we can transform $\mathbf{u}(t_k)$ by using the change of variables introduced in section 5.1: we denote (see (5.4), (5.6))

$$\mathbf{X}^k = \mathbf{X}_{\boldsymbol{\zeta}_h^k,\boldsymbol{\zeta}(t_k)}, \quad \mathbf{Y}^k = \mathbf{Y}_{\boldsymbol{\zeta}_h^k,\boldsymbol{\zeta}(t_k)}. \tag{6.15}$$

We also define (see (5.7) and (5.8))

$$\begin{gathered} \mathbf{U}^k(\mathbf{y}) = \mathbf{J}_{\mathbf{Y}^k}(\mathbf{X}^k(\mathbf{y}))\mathbf{u}\left(\mathbf{X}^k(\mathbf{y}), t_k\right), \quad P^k(\mathbf{y}) = p^k(\mathbf{X}^k(\mathbf{y})), \\ \boldsymbol{S}^k = -P^k\,\mathbf{Id} + 2\nu\mathbf{D}(\mathbf{U}^k), \quad \mathbf{F}^k(\mathbf{y}) = \mathbf{J}_{\mathbf{Y}^k}(\mathbf{X}^k(\mathbf{y}))\mathbf{f}(\mathbf{X}^k(\mathbf{y}), t_k). \end{gathered} \tag{6.16}$$

We recall that, according to Lemma 5.3, $\mathbf{U}^k \in \widehat{\mathcal{K}}(\boldsymbol{\zeta}_h^k)$ and $P^k \in M(\boldsymbol{\zeta}_h^k)$. We introduce the following notation that will be useful in what follows:

$$\widehat{\boldsymbol{X}} = \mathbf{Y}^k \circ \widetilde{\boldsymbol{X}} \circ \mathbf{X}^{k+1} \tag{6.17}$$

and

$$\widehat{\mathbf{J}} = \left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)\left(\mathbf{J}_{\mathbf{X}^k} \circ \widehat{\boldsymbol{X}}\right). \tag{6.18}$$

Before stating the main result of this section, we give some properties on the characteristics. First note that, according to Lemma 5.2, we have

$$\mathbf{X}^{k+1}(B(\boldsymbol{\zeta}_h^{k+1})) = B(\boldsymbol{\zeta}(t_{k+1})), \tag{6.19}$$

$$\mathbf{Y}^k(B(\boldsymbol{\zeta}(t_k))) = B(\boldsymbol{\zeta}_h^k).$$

Moreover, we can easily check that the function $\widetilde{\boldsymbol{X}}$ defined by (6.1) satisfies

$$\widetilde{\boldsymbol{X}}(\mathbf{x}) = \boldsymbol{\zeta}(t_k) + \mathbf{R}_{\theta(t_{k+1})-\theta(t_k)}(\mathbf{x} - \boldsymbol{\zeta}(t_{k+1})) \quad \forall \mathbf{x} \in B(\boldsymbol{\zeta}(t_{k+1})). \tag{6.20}$$

Consequently, we have

$$\widetilde{\boldsymbol{X}}(B(\boldsymbol{\zeta}(t_{k+1}))) = B(\boldsymbol{\zeta}(t_k)),$$

and therefore, we obtain

$$\widehat{\boldsymbol{X}}(B(\boldsymbol{\zeta}_h^{k+1})) = B(\boldsymbol{\zeta}_h^k). \tag{6.21}$$

We summarize some of the above properties in the following diagram:

$$\begin{tikzcd}
B(\boldsymbol{\zeta}_h^{k+1}) \arrow[r, "\mathbf{X}^{k+1}"] \arrow[d, "\widehat{\boldsymbol{X}}"'] & B(\boldsymbol{\zeta}(t_{k+1})) \arrow[d, "\widetilde{\boldsymbol{X}}"] \\
B(\boldsymbol{\zeta}_h^k) & B(\boldsymbol{\zeta}(t_k)) \arrow[l, "\mathbf{Y}^k"]
\end{tikzcd}$$

Next, we turn to the main result of this subsection: we show that $\mathbf{U}^{k+1}$ and P^{k+1} satisfy a mixed weak formulation with test functions in $\mathcal{K}(\boldsymbol{\zeta}_h^{k+1})$ and $M(\boldsymbol{\zeta}_h^{k+1})$.

Proposition 6.2. *The functions* $(\mathbf{U}^{k+1}, P^{k+1})$ *defined by* (6.16) *satisfy*

$$\begin{aligned} \left(\frac{1}{\Delta t}\left[\mathbf{U}^{k+1} - \widehat{\mathbf{J}}\left(\mathbf{U}^k \circ \widehat{\boldsymbol{X}}\right)\right], \boldsymbol{\varphi}\right) + a(\mathbf{U}^{k+1}, \boldsymbol{\varphi}) + b(\boldsymbol{\varphi}, P^{k+1}) \\ = (\mathbf{f}_h^{k+1}, \boldsymbol{\varphi}) + (\mathbf{A}_k, \boldsymbol{\varphi}) \quad \forall \boldsymbol{\varphi} \in \mathcal{K}(\boldsymbol{\zeta}_h^{k+1}), \end{aligned} \tag{6.22}$$

$$b(\mathbf{U}^{k+1}, q) = 0 \quad \forall q \in M(\boldsymbol{\zeta}_h^{k+1}), \tag{6.23}$$

with

(6.24)

$$\|\mathbf{A}_k\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\left(|\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}| + h + \Delta t + C\sqrt{\Delta t}\left\|\frac{d^2}{dt^2}[\mathbf{u} \circ \widetilde{\boldsymbol{\psi}}]\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}\right).$$

Proof.

First step. We transform (6.9).

By using Proposition 5.4, we have that $\mathbf{U}^{k+1}$ and P^{k+1} satisfy

$$\begin{aligned} &\left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right) \frac{\mathbf{u}(t_{k+1}) - \mathbf{u}(t_k) \circ \widetilde{\boldsymbol{X}}}{\Delta t} \circ \mathbf{X}^{k+1} - \nu[\mathbf{L}^{k+1}\mathbf{U}^{k+1}] + [\mathbf{G}^{k+1}P^{k+1}] \\ &\quad = \left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)(\mathbf{f}(\mathbf{X}^{k+1}, t_{k+1})) + \left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)(\boldsymbol{\alpha}_{k+1} \circ \mathbf{X}^{k+1}), \\ &\qquad \text{in } \mathcal{O} \setminus B(\boldsymbol{\zeta}_h^{k+1}). \end{aligned}$$

The above relation and (6.16) imply

$$\begin{aligned} &\frac{1}{\Delta t}\left[\mathbf{U}^{k+1} - \left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)\left(\mathbf{J}_{\mathbf{X}^k} \circ \widehat{\boldsymbol{X}}\right)\left(\mathbf{U}^k \circ \widehat{\boldsymbol{X}}\right)\right] - \nu\Delta\mathbf{U}^{k+1} + \nabla P^{k+1} \\ &= \nu[(\mathbf{L}^{k+1} - \Delta)\mathbf{U}^{k+1}] + [(\nabla - \mathbf{G}^{k+1})P^{k+1}] + \mathbf{F}^{k+1} + \left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)(\boldsymbol{\alpha}_{k+1} \circ \mathbf{X}^{k+1}), \\ &\qquad \text{in } \quad \mathcal{O} \setminus B(\boldsymbol{\zeta}_h^{k+1}), \end{aligned} \tag{6.25}$$

where $\widehat{\boldsymbol{X}}$ is defined by (6.17).

By taking the inner product of the previous equation with $\boldsymbol{\varphi} \in \mathcal{K}(\boldsymbol{\zeta}_h^{k+1})$ and by using (6.18), we obtain

$$\text{(6.26)} \quad \int_{\mathcal{O}\setminus B(\boldsymbol{\zeta}_h^{k+1})} \left(\frac{1}{\Delta t} \left[\mathbf{U}^{k+1} - \widehat{\mathbf{J}} \left(\mathbf{U}^k \circ \widehat{\boldsymbol{X}} \right) \right] \cdot \boldsymbol{\varphi} \right) \, \mathrm{d}\mathbf{y} \\ - \int_{\mathcal{O}\setminus B(\boldsymbol{\zeta}_h^{k+1})} \left(\operatorname{div} \boldsymbol{S}^{k+1} \cdot \boldsymbol{\varphi} \right) \, \mathrm{d}\mathbf{y} = \int_{\mathcal{O}\setminus B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{F}^{k+1} \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y} + A_1$$

with

$$\text{(6.27)} \quad A_1 = \int_{\mathcal{O}\setminus B(\boldsymbol{\zeta}_h^{k+1})} \left(\nu[(\mathbf{L}^{k+1} - \Delta)\mathbf{U}^{k+1}] + [(\nabla - \mathbf{G}^{k+1})P^{k+1}] \right) \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y} \\ + \int_{\mathcal{O}\setminus B(\boldsymbol{\zeta}_h^{k+1})} \left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1} \right) \left(\boldsymbol{\alpha}_{k+1} \circ \mathbf{X}^{k+1} \right) \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y}.$$

Second step. We transform the integral

$$\int_{B(\boldsymbol{\zeta}_h^{k+1})} \frac{\mathbf{U}^{k+1} - \widehat{\mathbf{J}} \left(\mathbf{U}^k \circ \widehat{\boldsymbol{X}} \right)}{\Delta t} \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y}$$

by using (6.10)–(6.11). From (5.3) (with $\mathbf{Y}^{k+1}$ as in (6.15)), combined with (5.9) and with (5.10), we obtain that

$$\text{(6.28)} \qquad \mathbf{J}_{\mathbf{Y}^{k+1}}(\mathbf{x}) = \mathbf{Id} \quad \forall \mathbf{x} \in B(\boldsymbol{\zeta}(t_{k+1})).$$

The above relation, (6.16), and (5.9) imply that for all $\mathbf{y} \in B(\boldsymbol{\zeta}_h^{k+1})$,

$$\text{(6.29)} \qquad \mathbf{U}^{k+1}(\mathbf{y}) = \mathbf{u} \left(\mathbf{y} + \boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}, t_{k+1} \right).$$

In particular, we have that

$$\text{(6.30)} \qquad \mathbf{U}^{k+1}(\mathbf{y}) = \boldsymbol{\zeta}'(t_{k+1}) + \omega(t_{k+1})(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^{\perp} \quad \forall \mathbf{y} \in B(\boldsymbol{\zeta}_h^{k+1}).$$

Similarly, we have

$$\text{(6.31)} \qquad \mathbf{U}^{k}(\mathbf{y}) = \boldsymbol{\zeta}'(t_{k}) + \omega(t_{k})(\mathbf{y} - \boldsymbol{\zeta}_h^{k})^{\perp} \quad \forall \mathbf{y} \in B(\boldsymbol{\zeta}_h^{k}).$$

Relations (6.19) and (6.21) yield

$$\text{(6.32)} \qquad \left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1} \right) \left(\mathbf{J}_{\mathbf{X}^{k}} \circ \widehat{\boldsymbol{X}} \right) = \mathbf{Id} \quad \text{in} \quad B(\boldsymbol{\zeta}_h^{k+1}).$$

Simple calculations combined with relations (5.9) and (6.20) yield

$$\widehat{\boldsymbol{X}}(\mathbf{y}) = \mathbf{R}_{\theta(t_{k+1})-\theta(t_k)}(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1}) + \boldsymbol{\zeta}_h^{k} \quad \forall \mathbf{y} \in B(\boldsymbol{\zeta}_h^{k+1}).$$

The above relation, (6.32), and (6.31) imply that for all $\mathbf{y} \in B(\boldsymbol{\zeta}_h^{k+1})$, we have that

$$\widehat{\mathbf{J}}(\mathbf{U}^k \circ \widehat{\boldsymbol{X}})(\mathbf{y}) = \boldsymbol{\zeta}'(t_k) + \omega(t_k)\mathbf{R}_{\theta(t_{k+1})-\theta(t_k)}(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^{\perp}.$$

By using (6.12), the previous equality can be written as

$$\begin{aligned}\widehat{\mathbf{J}}(\mathbf{U}^k \circ \widehat{\boldsymbol{X}})(\mathbf{y}) &= \boldsymbol{\zeta}'(t_k) + \omega(t_k)(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^\perp \\ &\quad + \Delta t\ \omega(t_k)\omega(t_{k+1})(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1}) + \omega(t_k)\mathbf{E}_k(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^\perp \quad \forall \mathbf{y} \in B(\boldsymbol{\zeta}_h^{k+1}).\end{aligned}$$

By taking the inner product of the above relation with $\boldsymbol{\varphi} \in \mathcal{K}(\boldsymbol{\zeta}_h^{k+1})$ and by integrating on $B(\boldsymbol{\zeta}_h^{k+1})$, we obtain that

$$\begin{aligned}(6.33)\quad \int_{B(\boldsymbol{\zeta}_h^{k+1})} \widehat{\mathbf{J}}(\mathbf{U}^k \circ \widehat{\boldsymbol{X}})(\mathbf{y}) \cdot \boldsymbol{\varphi}\ \mathrm{d}\mathbf{y} &= M\mathbf{l}_{\boldsymbol{\varphi}} \cdot \boldsymbol{\zeta}'(t_k) + J\omega(t_k)\omega_{\boldsymbol{\varphi}} \\ &\quad + \omega(t_k) \int_{B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{E}_k(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^\perp \cdot \boldsymbol{\varphi}\ \mathrm{d}\mathbf{y}.\end{aligned}$$

Relation (6.30) implies that, for all $\boldsymbol{\varphi} \in \mathcal{K}(\boldsymbol{\zeta}_h^{k+1})$, we have

$$\int_{B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{U}^{k+1} \cdot \boldsymbol{\varphi}\ \mathrm{d}\mathbf{y} = M\mathbf{l}_{\boldsymbol{\varphi}} \cdot \boldsymbol{\zeta}'(t_{k+1}) + J\omega(t_{k+1})\omega_{\boldsymbol{\varphi}}.$$

The above equality and (6.33) yield that, for all $\boldsymbol{\varphi} \in \mathcal{K}(\boldsymbol{\zeta}_h^{k+1})$, we have

$$\begin{aligned}\int_{B(\boldsymbol{\zeta}_h^{k+1})} \frac{\mathbf{U}^{k+1} - \widehat{\mathbf{J}}(\mathbf{U}^k \circ \widehat{\boldsymbol{X}})}{\Delta t} \cdot \boldsymbol{\varphi}\ \mathrm{d}\mathbf{y} &= M\mathbf{l}_{\boldsymbol{\varphi}} \cdot \frac{\boldsymbol{\zeta}'(t_{k+1}) - \boldsymbol{\zeta}'(t_k)}{\Delta t} \\ &\quad + J\frac{\omega(t_{k+1}) - \omega(t_k)}{\Delta t}\omega_{\boldsymbol{\varphi}} - \frac{\omega(t_k)}{\Delta t} \int_{B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{E}_k(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^\perp \cdot \boldsymbol{\varphi}\ \mathrm{d}\mathbf{y}.\end{aligned}$$

The above relation and (6.10)–(6.11) imply that

$$\begin{aligned}(6.34)\quad \int_{B(\boldsymbol{\zeta}_h^{k+1})} \frac{\mathbf{U}^{k+1} - \widehat{\mathbf{J}}(\mathbf{U}^k \circ \widehat{\boldsymbol{X}})}{\Delta t} \cdot \boldsymbol{\varphi}\ \mathrm{d}\mathbf{y} &= -\mathbf{l}_{\boldsymbol{\varphi}} \cdot \int_{\partial B(\boldsymbol{\zeta}(t_{k+1}))} \boldsymbol{\sigma}^{k+1}\mathbf{n}\ \mathrm{d}\Gamma \\ - \omega_{\boldsymbol{\varphi}} \int_{\partial B(\boldsymbol{\zeta}(t_{k+1}))} (\mathbf{y} - \boldsymbol{\zeta}(t_{k+1}))^\perp \cdot \boldsymbol{\sigma}^{k+1}\mathbf{n}\ \mathrm{d}\Gamma &+ \mathbf{l}_{\boldsymbol{\varphi}} \cdot \int_{B(\boldsymbol{\zeta}(t_{k+1}))} \mathbf{f}^{k+1}\ \mathrm{d}\mathbf{x} \\ + \omega_{\boldsymbol{\varphi}} \int_{B(\boldsymbol{\zeta}(t_{k+1}))} &(\mathbf{x} - \boldsymbol{\zeta}(t_{k+1}))^\perp \cdot \mathbf{f}^{k+1}(x)\ \mathrm{d}\mathbf{x} \\ &+ \mathbf{l}_{\boldsymbol{\varphi}} \cdot \boldsymbol{\beta}_k + \omega_{\boldsymbol{\varphi}}\gamma_k - \frac{\omega(t_k)}{\Delta t} \int_{B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{E}_k(\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^\perp \cdot \boldsymbol{\varphi}\ \mathrm{d}\mathbf{y}.\end{aligned}$$

On the other hand, by using relations (5.9), (5.10), and (6.28), we easily obtain that

$$\int_{\partial B(\boldsymbol{\zeta}_h^{k+1})} \boldsymbol{S}^{k+1}\mathbf{n}\ \mathrm{d}\Gamma = \int_{\partial B(\boldsymbol{\zeta}(t_{k+1}))} \boldsymbol{\sigma}^{k+1}\mathbf{n}\ \mathrm{d}\Gamma$$

and that

$$\int_{\partial B(\boldsymbol{\zeta}_h^{k+1})} (\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^\perp \cdot \boldsymbol{S}^{k+1}\mathbf{n}\ \mathrm{d}\Gamma = \int_{\partial B(\boldsymbol{\zeta}(t_{k+1}))} (\mathbf{y} - \boldsymbol{\zeta}(t_{k+1}))^\perp \cdot \boldsymbol{\sigma}^{k+1}\mathbf{n}\ \mathrm{d}\Gamma.$$

The above relations and (6.34) yield that

$$(6.35)\quad \int_{B(\boldsymbol{\zeta}_h^{k+1})} \frac{\mathbf{U}^{k+1} - \widehat{\mathbf{J}}(\mathbf{U}^k \circ \widehat{\boldsymbol{X}})}{\Delta t} \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y} = - \int_{\partial B(\boldsymbol{\zeta}_h^{k+1})} \Big(\boldsymbol{S}^{k+1} \mathbf{n} \Big) \cdot \boldsymbol{\varphi} \, \mathrm{d}\Gamma$$
$$+ \int_{B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{F}^{k+1} \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y} + \mathbf{l}_{\boldsymbol{\varphi}} \cdot \boldsymbol{\beta}_k + \omega_{\boldsymbol{\varphi}} \gamma_k - \frac{\omega(t_k)}{\Delta t} \int_{B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{E}_k (\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^{\perp} \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y}.$$

Third step. By integrating by parts, we have that

$$(6.36)\quad 2\nu \int_{\mathcal{O} \setminus B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{D}(\mathbf{U}^{k+1}) : \mathbf{D}(\boldsymbol{\varphi}) \, \mathrm{d}\mathbf{y} - \int_{\mathcal{O} \setminus B(\boldsymbol{\zeta}_h^{k+1})} P^{k+1} \operatorname{div}(\boldsymbol{\varphi}) \, \mathrm{d}\mathbf{y}$$
$$= \int_{\partial B(\boldsymbol{\zeta}_h^{k+1})} \Big(\boldsymbol{S}^{k+1} \mathbf{n} \Big) \cdot \boldsymbol{\varphi} \, \mathrm{d}\Gamma - \int_{\mathcal{O} \setminus B(\boldsymbol{\zeta}_h^{k+1})} \operatorname{div}(\boldsymbol{S}^{k+1}) \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y}.$$

Summing (6.36), (6.35), and (6.26) yields (6.22) with

$$(\mathbf{A}_k, \boldsymbol{\varphi}) = (\mathbf{F}^{k+1} - \mathbf{f}_h^{k+1}, \boldsymbol{\varphi}) + \mathbf{l}_{\boldsymbol{\varphi}} \cdot \boldsymbol{\beta}_k + \omega_{\boldsymbol{\varphi}} \gamma_k - \frac{\omega(t_k)}{\Delta t} \int_{B(\boldsymbol{\zeta}_h^{k+1})} \mathbf{E}_k (\mathbf{y} - \boldsymbol{\zeta}_h^{k+1})^{\perp} \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y}$$
$$+ \int_{\mathcal{O} \setminus B(\boldsymbol{\zeta}_h^{k+1})} \big(\nu [(\mathbf{L}^{k+1} - \Delta) \mathbf{U}^{k+1}] + [(\nabla - \mathbf{G}^{k+1}) P^{k+1}] \big) \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y}$$
$$+ \int_{\mathcal{O} \setminus B(\boldsymbol{\zeta}_h^{k+1})} \big(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1} \big) \left(\boldsymbol{\alpha}_{k+1} \circ \mathbf{X}^{k+1} \right) \cdot \boldsymbol{\varphi} \, \mathrm{d}\mathbf{y}.$$

The above relation, combined with relation (3.7) and Lemmas 5.6, 5.7, and 6.1, implies the proposition. □

6.3. Some results on characteristics. In this subsection, we give some results on the functions $\mathbf{X}^k$, $\widehat{\boldsymbol{X}}$, and $\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}$ that will be used in the proof of the main result.

LEMMA 6.3. *There exists a positive constant C independent of h and k such that*

$$\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C \left(\|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} \Delta t + |\boldsymbol{\varepsilon}_k| \right).$$

Proof. We denote by $\boldsymbol{\Lambda}^k$ (resp., $\boldsymbol{\Lambda}^{k+1}$) the mapping defined by (5.2) with $\boldsymbol{\zeta}_1 = \boldsymbol{\zeta}(t_k)$ and $\boldsymbol{\zeta}_2 = \boldsymbol{\zeta}_h^k$ (resp., $\boldsymbol{\zeta}_1 = \boldsymbol{\zeta}(t_{k+1})$ and $\boldsymbol{\zeta}_2 = \boldsymbol{\zeta}_h^{k+1}$). Let $\boldsymbol{\psi}^k$ and $\boldsymbol{\psi}^{k+1}$ be the solution of (5.3) corresponding to the velocity fields $\boldsymbol{\Lambda}^k$ and $\boldsymbol{\Lambda}^{k+1}$, respectively.

By using (5.3), we have that

$$(\boldsymbol{\psi}^{k+1} - \boldsymbol{\psi}^k)(\lambda) = \int_0^\lambda \boldsymbol{\Lambda}^{k+1}(\boldsymbol{\psi}^{k+1}(\mu)) - \boldsymbol{\Lambda}^k(\boldsymbol{\psi}^k(\mu)) \, \mathrm{d}\mu.$$

Therefore, by Lemma 5.5, there exists a positive constant C such that for all $\lambda \in [0,1]$, we have that

$$\left| (\boldsymbol{\psi}^{k+1} - \boldsymbol{\psi}^k)(\lambda) \right| \leqslant \|\boldsymbol{\Lambda}^{k+1} - \boldsymbol{\Lambda}^k\|_{\mathcal{L}^\infty(\mathcal{O})} + C \int_0^\lambda \left| (\boldsymbol{\psi}^{k+1}(\mu) - \boldsymbol{\psi}^k(\mu)) \right| \, \mathrm{d}\mu.$$

The above inequality and Gronwall's lemma yield

$$\left| (\boldsymbol{\psi}^{k+1} - \boldsymbol{\psi}^k)(\lambda) \right| \leqslant C \|\boldsymbol{\Lambda}^{k+1} - \boldsymbol{\Lambda}^k\|_{\mathcal{L}^\infty(\mathcal{O})}$$

for all $\lambda \in [0,1]$. In particular, for $\lambda = 1$, we have that

$$\|\mathbf{X}^{k+1} - \mathbf{X}^k\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C\|\mathbf{\Lambda}^{k+1} - \mathbf{\Lambda}^k\|_{\mathcal{L}^\infty(\mathcal{O})}. \tag{6.37}$$

By using relation (5.2), there exists a positive constant C such that

$$\|\mathbf{\Lambda}^{k+1} - \mathbf{\Lambda}^k\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C|\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1} - \boldsymbol{\zeta}(t_k) + \boldsymbol{\zeta}_h^k|.$$

The above relation, combined with (3.2) and (6.7), yields

$$\|\mathbf{\Lambda}^{k+1} - \mathbf{\Lambda}^k\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C|\mathbf{u}_h^k(\boldsymbol{\zeta}_h^k) - \mathbf{u}\left(\boldsymbol{\zeta}(t_k), t_k\right)|\Delta t + C|\boldsymbol{\varepsilon}_k|. \tag{6.38}$$

On the other hand, by (6.29), we have $\mathbf{u}\left(\boldsymbol{\zeta}(t_k), t_k\right) = \mathbf{U}^k(\boldsymbol{\zeta}_h^k)$ and, moreover, $\mathbf{u}_h^k - \mathbf{U}^k \in \mathcal{K}(\boldsymbol{\zeta}_h^k)$. Then, owing to (2.5), we readily check that

$$|\mathbf{u}_h^k(\boldsymbol{\zeta}_h^k) - \mathbf{U}^k(\boldsymbol{\zeta}_h^k)| \leqslant \frac{1}{\sqrt{M}}\|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}. \tag{6.39}$$

Therefore, the above relation and (6.38) imply that

$$\|\mathbf{\Lambda}^{k+1} - \mathbf{\Lambda}^k\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C\|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}\Delta t + C|\boldsymbol{\varepsilon}_k|. \tag{6.40}$$

Relations (6.37) and (6.40) yield the conclusion of the lemma. □

A similar estimate holds for the jacobian matrices $\mathbf{J}_{\mathbf{X}^{k+1}}$ and $\mathbf{J}_{\mathbf{X}^k}$. Since the proof of this estimate is completely similar to the proof of Lemma 6.3, we give below only its statement and skip the proof.

Lemma 6.4. *There exists a positive constant C independent of k and h such that*

$$\|\mathbf{J}_{\mathbf{X}^{k+1}} - \mathbf{J}_{\mathbf{X}^k}\|_{\mathcal{L}^\infty(\mathcal{O})} \leqslant C\left(\|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}\Delta t + |\boldsymbol{\varepsilon}_k|\right).$$

The functions $\widehat{\boldsymbol{X}}$ and $\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}$ are close to the identity in the sense made precise below.

Lemma 6.5. *The functions $\widehat{\boldsymbol{X}}$ and $\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}$ defined by* (6.17) *and* (3.4) *satisfy the following estimates:*

$$\|\widehat{\boldsymbol{X}} - \mathbf{Id}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\left(|\boldsymbol{\varepsilon}_k| + \Delta t\|\mathbf{U}^k - \mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})} + \sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} + \Delta t\right), \tag{6.41}$$

$$\|\widehat{\boldsymbol{X}} - \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\left(|\boldsymbol{\varepsilon}_k| + \Delta t\|\mathbf{U}^k - \mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})} + \sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} + h\Delta t\right). \tag{6.42}$$

Proof. Let us define

$$\widehat{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{y}) = \mathbf{Y}^k(\widetilde{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{X}^{k+1}(\mathbf{y}))), \tag{6.43}$$

where $\widetilde{\boldsymbol{\psi}}$ is defined by (2.6). Note that $\widehat{\boldsymbol{\psi}}(t_k; t_{k+1}, \mathbf{y}) = \widehat{\boldsymbol{X}}(\mathbf{y})$ for all $\mathbf{y} \in \mathcal{O}$.

We have that

$$\frac{d}{dt}\widehat{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{y}) = \mathbf{J}_{\mathbf{Y}^k}(\widetilde{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{X}^{k+1}(\mathbf{y})))\frac{d}{dt}\widetilde{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{X}^{k+1}(\mathbf{y})).$$

By using (6.8) we obtain that

$$\begin{aligned}\frac{d}{dt}\widehat{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{y}) = {} & \left[\mathbf{J}_{\mathbf{Y}^k} \circ \mathbf{X}^k\right](\widehat{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{y}))\left[\mathbf{u}\left(\mathbf{X}^k\left(\widehat{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{y})\right), t_k\right)\right] \\ & + \left[\mathbf{J}_{\mathbf{Y}^k} \circ \mathbf{X}^k\right](\widehat{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{y}))\left[\boldsymbol{\delta}_k\left(t, \mathbf{X}^k\left(\widehat{\boldsymbol{\psi}}(t; t_{k+1}, \mathbf{y})\right)\right)\right].\end{aligned}$$

The above relation and (6.16) yield

$$\frac{d}{dt}\widehat{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{y}) = \mathbf{U}^k(\widehat{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{y})) + \left[\mathbf{J}_{\mathbf{Y}^k}\circ\mathbf{X}^k\right](\widehat{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{y}))\left[\boldsymbol{\delta}_k\left(t,\mathbf{X}^k\left(\widehat{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{y})\right)\right)\right]. \tag{6.44}$$

On the other hand, we have that

$$\widehat{\boldsymbol{\psi}}(t_{k+1};t_{k+1},\mathbf{y}) = \mathbf{Y}^k\circ\mathbf{X}^{k+1}(\mathbf{y}). \tag{6.45}$$

Therefore, by using (6.44) and (6.45), we get

$$\widehat{\boldsymbol{X}}(\mathbf{y})-\mathbf{y} = \mathbf{Y}^k\circ\mathbf{X}^{k+1}(\mathbf{y})-\mathbf{y}-\int_{t_k}^{t_{k+1}}\mathbf{U}^k(\widehat{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{y}))\,\mathrm{d}t - \int_{t_k}^{t_{k+1}}\left[\mathbf{J}_{\mathbf{Y}^k}\circ\mathbf{X}^k\right](\widehat{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{y}))\left[\boldsymbol{\delta}_k\left(t,\mathbf{X}^k\left(\widehat{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{y})\right)\right)\right]\,\mathrm{d}t,$$

which yields

$$\|\widehat{\boldsymbol{X}}-\mathbf{Id}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant \|\mathbf{Y}^k\circ\mathbf{X}^{k+1}-\mathbf{Id}\|_{\mathcal{L}^2(\mathcal{O})} + \int_{t_k}^{t_{k+1}}\left\|\mathbf{U}^k(\widehat{\boldsymbol{\psi}}(s))\right\|_{\mathcal{L}^2(\mathcal{O})}\,\mathrm{d}s + C\sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}. \tag{6.46}$$

By Lemma 5.6, there exists a positive constant C such that

$$\|\mathbf{Y}^k\circ\mathbf{X}^{k+1}-\mathbf{Id}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\|\mathbf{X}^{k+1}-\mathbf{X}^k\|_{\mathcal{L}^\infty(\mathcal{O})}.$$

The above relation and Lemma 6.3 yield

$$\|\mathbf{Y}^k\circ\mathbf{X}^{k+1}-\mathbf{Id}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\left(\Delta t\|\mathbf{u}_h^k-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}+|\boldsymbol{\varepsilon}_k|\right). \tag{6.47}$$

Relations (6.46) and (6.47), together with (3.8) and (6.16), imply

$$\|\widehat{\boldsymbol{X}}-\mathbf{Id}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\left(\Delta t\|\mathbf{u}_h^k-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}+|\boldsymbol{\varepsilon}_k|\right)+C\Delta t+\sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}.$$

Therefore, we deduce (6.41).

Now we turn to the proof of (6.42): by using (3.3), (6.44), and (6.45), we obtain

$$\widehat{\boldsymbol{\psi}}(t;t_{k+1},\mathbf{y})-\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t;t_{k+1},\mathbf{y}) = \mathbf{Y}^k\circ\mathbf{X}^{k+1}(\mathbf{y})-\mathbf{y} - \int_t^{t_{k+1}}\left(\mathbf{U}^k(\widehat{\boldsymbol{\psi}}(s;t_{k+1},\mathbf{y}))-\mathbf{P}\mathbf{u}_h^k(\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(s;t_{k+1},\mathbf{y}))\right)\,\mathrm{d}s - \int_t^{t_{k+1}}(\mathbf{J}_{\mathbf{Y}^k}\circ\mathbf{X}^k)(\widehat{\boldsymbol{\psi}}(s;t_{k+1},\mathbf{y}))\left[\boldsymbol{\delta}_k\left(s,\mathbf{X}^k\left(\widehat{\boldsymbol{\psi}}(s;t_{k+1},\mathbf{y})\right)\right)\right]\,\mathrm{d}s,$$

which yields

$$\|\widehat{\boldsymbol{\psi}}(t)-\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t)\|_{\mathcal{L}^2(\mathcal{O})} \leqslant \|\mathbf{Y}^k\circ\mathbf{X}^{k+1}-\mathbf{Id}\|_{\mathcal{L}^2(\mathcal{O})} + \int_t^{t_{k+1}}\left\|\mathbf{U}^k(\widehat{\boldsymbol{\psi}}(s))-\mathbf{P}\mathbf{u}_h^k(\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(s))\right\|_{\mathcal{L}^2(\mathcal{O})}\,\mathrm{d}s + C\sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}. \tag{6.48}$$

Relations (6.48) and (6.47) imply

$$\begin{aligned}\|\widehat{\boldsymbol{\psi}}(t)-\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t)\|_{\mathcal{L}^2(\mathcal{O})} &\leqslant C\left(\Delta t\|\mathbf{u}_h^k-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}+|\boldsymbol{\varepsilon}_k|\right)\\ &\quad+\int_t^{t_{k+1}}\left\|\mathbf{U}^k(\widehat{\boldsymbol{\psi}}(s))-\mathbf{P}\mathbf{u}_h^k(\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(s))\right\|_{\mathcal{L}^2(\mathcal{O})}\,\mathrm{d}s+C\sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}.\end{aligned}$$

By using (3.8) and Remark 3.1, we have that

$$\begin{aligned}\|\widehat{\boldsymbol{\psi}}(t)-\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t)\|_{\mathcal{L}^2(\mathcal{O})} &\leqslant C\left(\Delta t\|\mathbf{u}_h^k-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}+|\boldsymbol{\varepsilon}_k|+\Delta t\|\mathbf{U}^k-\mathbf{P}\mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})}\right)\\ &\quad+C\int_t^{t_{k+1}}\left\|\widehat{\boldsymbol{\psi}}(s)-\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(s)\right\|_{\mathcal{L}^2(\mathcal{O})}\,\mathrm{d}s+C\sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}.\end{aligned}$$

Therefore, by Gronwall's lemma, we get that

$$\begin{aligned}\|\widehat{\boldsymbol{\psi}}(t)-\overline{\boldsymbol{\psi}}_{\boldsymbol{h}}^{\boldsymbol{k}}(t)\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\Big(&\Delta t\|\mathbf{u}_h^k-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}+|\boldsymbol{\varepsilon}_k|+\Delta t\|\mathbf{U}^k-\mathbf{P}\mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})}\\ &+\sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}\Big).\end{aligned}$$

In particular for $t=t_k$, we obtain that

$$\begin{aligned}\|\widehat{\boldsymbol{X}}-\overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\Big(&\Delta t\|\mathbf{u}_h^k-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}+|\boldsymbol{\varepsilon}_k|+\Delta t\|\mathbf{U}^k-\mathbf{P}\mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})}\\ &+\sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}\Big).\end{aligned}\tag{6.49}$$

Since $\mathbf{P}$ is an orthogonal projection in $\mathcal{L}^2(\mathcal{O})$, we have that

$$\|\mathbf{U}^k-\mathbf{P}\mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})}\leqslant\|\mathbf{U}^k-\mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})}+\|\mathbf{P}\mathbf{U}^k-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}.\tag{6.50}$$

Now, since $\mathbf{U}^k\in\mathcal{H}_0^1(\mathcal{O})$ and $\operatorname{div}(\mathbf{U}^k)=0$, there exists a stream function $\psi\in H^2(\mathcal{O})\cap H_0^1(\mathcal{O})$ of $\mathbf{U}^k$, i.e., $\mathbf{U}^k=\mathbf{rot}\ \psi$. Let ψ_h be the Lagrange interpolated function of ψ on the triangulation $\mathcal{T}_h$. We denote $\widetilde{\mathbf{U}_h^k}=\mathbf{rot}\ \psi_h$. Since $\widetilde{\mathbf{U}_h^k}\in\mathcal{R}_h$, we have that

$$\begin{aligned}\|\mathbf{P}\mathbf{U}^k-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}&\leqslant\|\widetilde{\mathbf{U}_h^k}-\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}=\|\mathbf{rot}\ (\psi-\psi_h)\|_{\mathcal{L}^2(\mathcal{O})}\\ &\leqslant Ch\|\psi\|_{H^2(\mathcal{O})}\leqslant Ch\|\mathbf{U}^k\|_{\mathcal{H}^1(\mathcal{O})}.\end{aligned}$$

The above equation, (6.49), and (6.50) imply the result. □

7. Proof of the main result.

We can now prove Theorem 3.2.

First step. Assume that $h\leq C(\Delta t)^2$. We first show that if (3.10) holds and if

$$\operatorname{dist}(B(\boldsymbol{\zeta}_h^k),\partial\mathcal{O})>2\eta,\quad \operatorname{dist}(B(\boldsymbol{\zeta}_h^{k+1}),\partial\mathcal{O})>2\eta,\tag{7.1}$$

then there exist two positive constants C_0 and C_1 independent of Δt and h such that the error $e_h^k=\|\mathbf{U}^k-\mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})}+|\boldsymbol{\zeta}(t_k)-\boldsymbol{\zeta}_h^k|$ satisfies the following inequality:

$$e_h^{k+1}\leqslant e_h^k(1+C_0\Delta t)+C_0\Delta t\beta_h^k,\tag{7.2}$$

where

$$\sum_{k=0}^N\beta_h^k\leqslant C_1.$$

Let us remark that assumption (7.1) together with (3.10) allows us to perform the change of variables defined in section 5 and to define $\mathbf{U}^k$, $\mathbf{U}^{k+1}$, and P^{k+1} (see (6.16)).

By using (4.50), there exists $(\mathbf{U}_h^{k+1}, P_h^{k+1}) \in \mathcal{K}_h(\boldsymbol{\zeta}_h^{k+1}) \times M_h(\boldsymbol{\zeta}_h^{k+1})$ such that

$$\left\{ \begin{array}{rcl} a\left(\mathbf{U}^{k+1} - \mathbf{U}_h^{k+1}, \boldsymbol{\varphi}\right) + b\left(\boldsymbol{\varphi}, P^{k+1} - P_h^{k+1}\right) & = & 0 \quad \forall \boldsymbol{\varphi} \in \mathcal{K}_h(\boldsymbol{\zeta}_h^{k+1}) \\ b\left(\mathbf{U}^{k+1} - \mathbf{U}_h^{k+1}, q\right) & = & 0 \quad \forall q \in M_h(\boldsymbol{\zeta}_h^{k+1}). \end{array} \right. \tag{7.3}$$

Subtracting (7.3) and (3.5) from (6.22) yields

$$\begin{aligned} \frac{1}{\Delta t}\left(\mathbf{U}^{k+1} - \mathbf{u}_h^{k+1}, \boldsymbol{\varphi}\right) + a(\mathbf{U}_h^{k+1} - \mathbf{u}_h^{k+1}, \boldsymbol{\varphi}) + b(\boldsymbol{\varphi}, P_h^{k+1} - p_h^{k+1}) \\ = \frac{1}{\Delta t}\left(\widehat{\mathbf{J}}\left(\mathbf{U}^k \circ \widehat{\boldsymbol{X}}\right) - \mathbf{u}_h^k \circ \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}, \boldsymbol{\varphi}\right) + (\mathbf{A}_k, \boldsymbol{\varphi}) \quad \forall \boldsymbol{\varphi} \in \mathcal{K}(\boldsymbol{\zeta}_h^{k+1}), \\ b(\mathbf{U}_h^{k+1} - \mathbf{u}_h^{k+1}, q) = 0 \quad \forall q \in M_h(\boldsymbol{\zeta}_h^{k+1}). \end{aligned}$$

In particular, for $\boldsymbol{\varphi} = \mathbf{U}_h^{k+1} - \mathbf{u}_h^{k+1}$ and $q = P_h^{k+1} - p_h^{k+1}$, we easily obtain that

$$\begin{aligned} \left\|\mathbf{U}_h^{k+1} - \mathbf{u}_h^{k+1}\right\|_{\mathcal{L}^2(\mathcal{O})} \leqslant \left\|\widehat{\mathbf{J}}\left(\mathbf{U}^k \circ \widehat{\boldsymbol{X}}\right) - \mathbf{u}_h^k \circ \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}\right\|_{\mathcal{L}^2(\mathcal{O})} \\ + \Delta t\|\mathbf{A}_k\|_{\mathcal{L}^2(\mathcal{O})} + \left\|\mathbf{U}^{k+1} - \mathbf{U}_h^{k+1}\right\|_{\mathcal{L}^2(\mathcal{O})}. \end{aligned} \tag{7.4}$$

On the other hand, since

$$\widehat{\mathbf{J}} = \left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)\left(\mathbf{J}_{\mathbf{X}^k} \circ \widehat{\boldsymbol{X}}\right),$$

we have that

$$\begin{aligned} \left\|\widehat{\mathbf{J}}\left(\mathbf{U}^k \circ \widehat{\boldsymbol{X}}\right) - \mathbf{u}_h^k \circ \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}\right\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\left\|\left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)\left(\mathbf{J}_{\mathbf{X}^k} \circ \widehat{\boldsymbol{X}}\right) - \mathbf{Id}\right\|_{\mathcal{L}^2(\mathcal{O})} \\ + \left\|\mathbf{U}^k \circ \widehat{\boldsymbol{X}} - \mathbf{U}^k \circ \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}\right\|_{\mathcal{L}^2(\mathcal{O})} \\ + \left\|\mathbf{U}^k \circ \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}} - \mathbf{u}_h^k \circ \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}\right\|_{\mathcal{L}^2(\mathcal{O})}. \end{aligned} \tag{7.5}$$

Since $\left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)\mathbf{J}_{\mathbf{X}^{k+1}} = \mathbf{Id}$, we infer from Lemma 5.6 that

$$\begin{aligned} \left\|\left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)\left(\mathbf{J}_{\mathbf{X}^k} \circ \widehat{\boldsymbol{X}}\right) - \mathbf{Id}\right\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\|\widehat{\boldsymbol{X}} - \mathbf{Id}\|_{\mathcal{L}^2(\mathcal{O})}\, |\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| \\ + C\|\mathbf{J}_{\mathbf{X}^k} - \mathbf{J}_{\mathbf{X}^{k+1}}\|_{\mathcal{L}^2(\mathcal{O})}. \end{aligned}$$

By using Lemmas 6.4 and 6.5 and the above inequality, we obtain that

$$\begin{aligned} \left\|\left(\mathbf{J}_{\mathbf{Y}^{k+1}} \circ \mathbf{X}^{k+1}\right)\left(\mathbf{J}_{\mathbf{X}^k} \circ \widehat{\boldsymbol{X}}\right) - \mathbf{Id}\right\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\Big(\Delta t|\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| \\ + \Delta t\|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} + \sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} + |\boldsymbol{\varepsilon}_k|\Big). \end{aligned} \tag{7.6}$$

By using (3.8) and Lemma 5.6, we easily check that

$$\|\mathbf{U}^k \circ \widehat{\boldsymbol{X}} - \mathbf{U}^k \circ \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\|\widehat{\boldsymbol{X}} - \overline{\boldsymbol{X}}_{\boldsymbol{h}}^{\boldsymbol{k}}\|_{\mathcal{L}^2(\mathcal{O})}.$$

The above inequality, relations (7.4), (7.5), and (7.6), Lemma 6.5, and the fact that $\det \mathbf{J}_{\overline{\boldsymbol{X}}_h^k} = 1$ imply that

$$\begin{aligned}\|\mathbf{U}_h^{k+1} - \mathbf{u}_h^{k+1}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\Big(&\Delta t|\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \Delta t\|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} \\ &+ \sqrt{\Delta t}\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} + |\boldsymbol{\varepsilon}_k| + h\Delta t\Big) \\ &+ \|\mathbf{U}^k - \mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})} + \Delta t\|\mathbf{A}_k\|_{\mathcal{L}^2(\mathcal{O})} + \|\mathbf{U}^{k+1} - \mathbf{U}_h^{k+1}\|_{\mathcal{L}^2(\mathcal{O})}.\end{aligned} \tag{7.7}$$

By using Lemma 4.4, Proposition 6.2, and Lemma 6.1, we have the following inequalities:

$$\|\mathbf{U}^{k+1} - \mathbf{U}_h^{k+1}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant Ch,$$

$$\|\mathbf{A}_k\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\left(|\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}| + h + \Delta t + C\sqrt{\Delta t}\left\|\frac{d^2}{dt^2}[\mathbf{u}\circ\widetilde{\boldsymbol{\psi}}]\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}\right),$$

$$\|\boldsymbol{\delta}_k\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} \leqslant C\Delta t\left\|\frac{\partial\mathbf{u}}{\partial t}\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))},$$

$$|\boldsymbol{\varepsilon}_k| \leqslant C\,(\Delta t)^2.$$

The above inequalities and (7.7) yield that

$$\begin{aligned}\|\mathbf{U}^{k+1} - \mathbf{u}_h^{k+1}\|_{\mathcal{L}^2(\mathcal{O})} \leqslant \|\mathbf{U}^k - \mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})} + C\Bigg(&(\Delta t)^2 + h\Delta t + h \\ &+ \Delta t|\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}| + \Delta t\|\mathbf{U}^k - \mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})} \\ &+ (\Delta t)^{3/2}\left\|\frac{\partial\mathbf{u}}{\partial t}\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} + (\Delta t)^{3/2}\left\|\frac{d^2}{dt^2}[\mathbf{u}\circ\widetilde{\boldsymbol{\psi}}]\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}\Bigg).\end{aligned} \tag{7.8}$$

On the other hand, (3.2), (6.7), (6.31), and (6.39) imply that

$$\begin{aligned}|\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}| &\leqslant |\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \Delta t|\mathbf{u}_h^k(\boldsymbol{\zeta}_h^k) - \mathbf{u}\left(\boldsymbol{\zeta}(t_k), t_k\right)| + |\boldsymbol{\varepsilon}_k| \\ &\leqslant |\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + C\Delta t\|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} + |\boldsymbol{\varepsilon}_k|.\end{aligned} \tag{7.9}$$

Combining (7.8) and (7.9), we obtain that

$$\begin{aligned}&\|\mathbf{U}^{k+1} - \mathbf{u}_h^{k+1}\|_{\mathcal{L}^2(\mathcal{O})} + |\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}| \\ &\qquad \leqslant (1 + C\Delta t)\left(|\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}\right) \\ &\qquad + C\Bigg(h + (\Delta t)^2 + h\Delta t + (\Delta t)^{3/2}\left\|\frac{\partial\mathbf{u}}{\partial t}\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} \\ &\qquad\qquad + (\Delta t)^{3/2}\left\|\frac{d^2}{dt^2}[\mathbf{u}\circ\widetilde{\boldsymbol{\psi}}]\right\|_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}\Bigg).\end{aligned}$$

The above inequality and the hypothesis $h \leqslant C\,(\Delta t)^2$ imply the existence of a positive constant C_0 such that

$$\begin{aligned}&\|\mathbf{U}^{k+1} - \mathbf{u}_h^{k+1}\|_{\mathcal{L}^2(\mathcal{O})} + |\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}| \\ &\qquad \leqslant (1 + C_0\Delta t)\left(|\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})}\right) \\ &\qquad + C_0\Delta t\left(\Delta t + \left\|\frac{\partial\mathbf{u}}{\partial t}\right\|^2_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} + \left\|\frac{d^2}{dt^2}[\mathbf{u}\circ\widetilde{\boldsymbol{\psi}}]\right\|^2_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}\right),\end{aligned}$$

which is exactly (7.2).

Second step. We show that if Δt is small enough, then the error $e_h^k = \|\mathbf{U}^k - \mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})} + |\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k|$ satisfies $e_h^k \leqslant C_1 \Delta t$ with a constant C_1 independent of k, Δt, and h. This fact implies, in particular, that (7.1) holds.

Define

$$C_1 = C_0 \exp(C_0 T) \left(\left\| \frac{\partial \mathbf{u}}{\partial t} \right\|^2_{\mathcal{L}^2(\mathcal{O}\times(0,T))} + \left\| \frac{d^2}{dt^2}[\mathbf{u} \circ \widetilde{\boldsymbol{\psi}}] \right\|^2_{\mathcal{L}^2(\mathcal{O}\times(0,T))} \right) + \exp(C_0 T).$$

It can be easily checked that

$$(1 + C_0 \Delta t)^n \, C_0 \left(\left\| \frac{\partial \mathbf{u}}{\partial t} \right\|^2_{\mathcal{L}^2(\mathcal{O}\times(0,T))} + \left\| \frac{d^2}{dt^2}[\mathbf{u} \circ \widetilde{\boldsymbol{\psi}}] \right\|^2_{\mathcal{L}^2(\mathcal{O}\times(0,T))} \right) + (1 + C_0 \Delta t)^n - 1 \leqslant C_1 \quad \forall n \in \{0, \dots, N\}.$$

Moreover, there exists a positive constant C_2 such that

$$\|\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C_2.$$

Let $N_0 \in \mathbb{N}$ be such that $(2C_1 + C_2)\Delta t < \eta$ for all $N \geqslant N_0$. Next we prove by induction over k that for $N \geqslant N_0$ and for $k \in \{0, \dots, N\}$ we have

$$\tag{7.10} |\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} \leqslant \Bigg[(1 + C_0 \Delta t)^k - 1 + C_0 (1 + C_0 \Delta t)^k \left(\left\| \frac{\partial \mathbf{u}}{\partial t} \right\|^2_{\mathcal{L}^2(\mathcal{O}\times(0,t_k))} + \left\| \frac{d^2}{dt^2}[\mathbf{u} \circ \widetilde{\boldsymbol{\psi}}] \right\|^2_{\mathcal{L}^2(\mathcal{O}\times(0,t_k))} \right) \Bigg] \Delta t.$$

The relation (7.10) is true for $k = 0$. Suppose that we have shown (7.10) for a given $k \geqslant 0$. Then, we deduce that

$$\tag{7.11} |\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| \leqslant C_1 \Delta t < \eta$$

and therefore, by using (3.10), we have that $\operatorname{dist}(B(\boldsymbol{\zeta}_h^k), \partial\mathcal{O}) > 2\eta$.

By using (3.2) and (3.10), we also have that

$$\begin{aligned} |\boldsymbol{\zeta}_h^{k+1} - \boldsymbol{\zeta}_h^k| &\leqslant \frac{1}{\sqrt{\pi}} \left(\|\mathbf{U}^k - \mathbf{u}_h^k\|_{\mathcal{L}^2(\mathcal{O})} + \|\mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} \right) \Delta t \\ &\leqslant \frac{C_1 + C_2}{\sqrt{\pi}} \Delta t. \end{aligned}$$

The above relation, the fact that $(2C_1 + C_2)\Delta t < \eta$, and (7.11) imply that

$$\operatorname{dist}(B(\boldsymbol{\zeta}_h^{k+1}), \partial\mathcal{O}) > 2\eta.$$

Thus, we can apply the first step of the proof to obtain that

$$\begin{aligned} &|\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}| + \|\mathbf{u}_h^{k+1} - \mathbf{U}^{k+1}\|_{\mathcal{L}^2(\mathcal{O})} \\ &\quad \leqslant (1 + C_0 \Delta t) \Big(|\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} \Big) \\ &\qquad + C_0 \Delta t \left(\Delta t + \left\| \frac{\partial \mathbf{u}}{\partial t} \right\|^2_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} + \left\| \frac{d^2}{dt^2}[\mathbf{u} \circ \widetilde{\boldsymbol{\psi}}] \right\|^2_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} \right). \end{aligned}$$

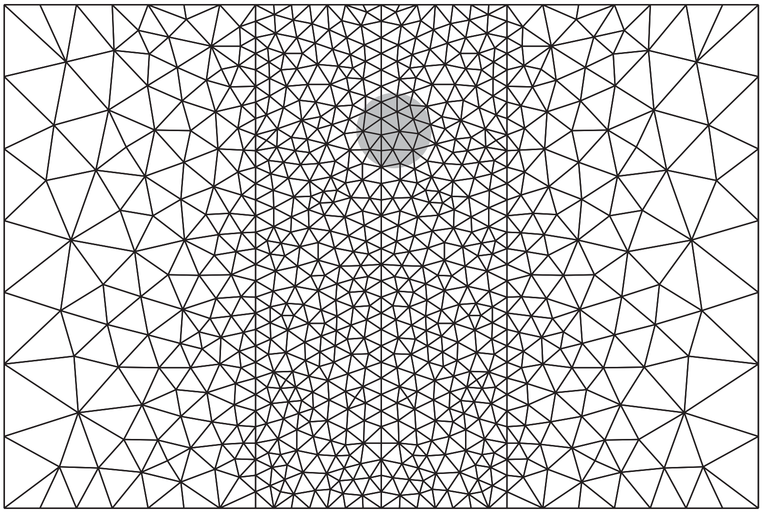


FIG. 2. *Initial position and mesh.*

The above relation and (7.10) imply that

$$\begin{aligned}|\boldsymbol{\zeta}(t_{k+1}) - \boldsymbol{\zeta}_h^{k+1}| + \|\mathbf{u}_h^{k+1} - \mathbf{U}^{k+1}\|_{\mathcal{L}^2(\mathcal{O})} &\leqslant (1+C_0\Delta t)[(1+C_0\Delta t)^k - 1]\Delta t \\ &+ C_0(1+C_0\Delta t)^{k+1}\left(\left\|\frac{\partial \mathbf{u}}{\partial t}\right\|^2_{\mathcal{L}^2(\mathcal{O}\times(0,t_k))} + \left\|\frac{d^2}{dt^2}[\mathbf{u}\circ\widetilde{\boldsymbol{\psi}}]\right\|^2_{\mathcal{L}^2(\mathcal{O}\times(0,t_k))}\right)\Delta t \\ &+ C_0\Delta t\left(\Delta t + \left\|\frac{\partial \mathbf{u}}{\partial t}\right\|^2_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))} + \left\|\frac{d^2}{dt^2}[\mathbf{u}\circ\widetilde{\boldsymbol{\psi}}]\right\|^2_{\mathcal{L}^2(\mathcal{O}\times(t_k,t_{k+1}))}\right),\end{aligned}$$

which implies (7.10) for $k+1$.

Third step. From the previous steps we conclude that if Δt is small enough and if $h \leqslant C(\Delta t)^2$, then

$$|\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \|\mathbf{u}_h^k - \mathbf{U}^k\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C_1\Delta t \qquad \forall\, k \in \{0,\dots,N\}.$$

The above relation, Lemma 5.6, (3.8), and Lemma 4.4 imply that if Δt is small enough and if $h \leqslant C(\Delta t)^2$, then

$$|\boldsymbol{\zeta}(t_k) - \boldsymbol{\zeta}_h^k| + \|\mathbf{u}_h^k - \mathbf{u}(t_k)\|_{\mathcal{L}^2(\mathcal{O})} \leqslant C\Delta t \qquad \forall\, k \in \{0,\dots,N\},$$

which is the conclusion of the theorem. □

8. Concluding remarks. We implemented the numerical method we proposed, and several numerical tests have been performed. Let us briefly describe the results obtained in the case of a rigid ball falling vertically under the action of a vertical force oriented downward. At instant $t = 0$ the velocity field in the fluids and in the solid is supposed to vanish.

We use a mesh with 1432 triangles and 752 vertices (see Figure 2).

Far from the ball the space discretization parameter is $h_1 \approx 0.57$, whereas in the neighborhood of the ball it is given by $h_2 \approx 0.12$. For the time discretization, we choose the time step $\Delta t = 0.1$. Moreover, we choose the radius of the ball equal

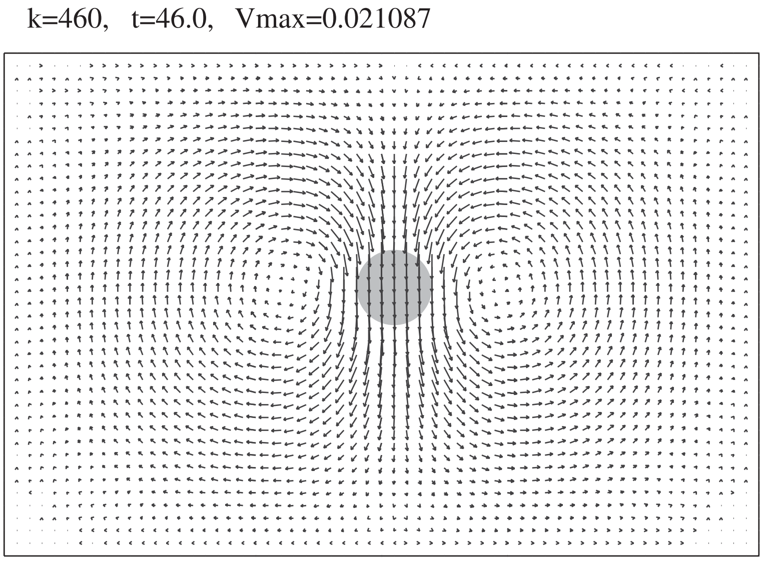


FIG. 3. *Position and velocity field at time $t = 46.0$.*

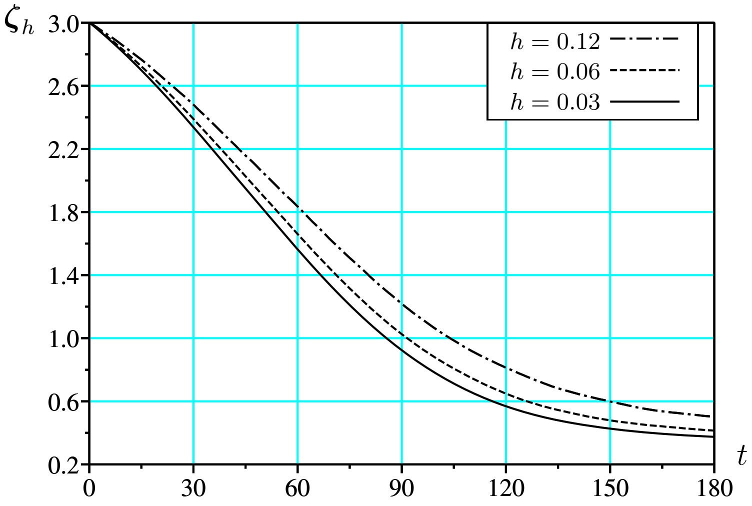


FIG. 4. *Position of the ball.*

to 0.3, the viscosity $\mu = 1$, and the downward force of intensity equal to one (all quantities are given in International System (IS) units). In Figure 3 we represent the configuration of the system for $k = 460$ (corresponding to $t = 46.0$).

We repeated the calculation twice by dividing each mesh size by two (this means that each triangle was each time divided into four smaller triangles). More precisely, we used the meshes described in the table below.

	h	Triangles	Vertices	CPU time
Mesh 1	0.12	1432	752	3 hours
Mesh 2	0.06	5728	2935	11 hours
Mesh 3	0.03	22912	11597	8 days

The last column represents the time used by a Pentium IV computer with a 2.4 GHz CPU clock to achieve the calculation.

In Figure 4 we represented the height of the center of the ball versus the time t for the different meshes.

REFERENCES

[1] V.I. Arnold, *Ordinary Differential Equations*, Springer-Verlag, Berlin, 1992. Translated from the third Russian edition by Roger Cooke.

[2] S.C. Brenner and L.R. Scott, *The Mathematical Theory of Finite Element Methods*, Texts Appl. Math. 15, Springer-Verlag, New York, 1994.

[3] H. Brezis, *Analyse fonctionnelle. Théorie et applications*, Collection Mathématiques Appliquées pour la Maîtrise, Masson, Paris, 1983.

[4] P.-G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.

[5] C. Conca, H.J. San Martín, and M. Tucsnak, *Existence of solutions for the equations modelling the motion of a rigid body in a viscous fluid*, Comm. Partial Differential Equations, 25 (2000), pp. 1019–1042.

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

[7] L.P. Eisenhart, *Riemannian Geometry*, Princeton University Press, Princeton, NJ, 1949.

[8] C. Farhat, P. Geuzaine, and C. Grandmont, *The discrete geometric conservation law and the nonlinear stability of ALE schemes for the solution of flow problems on moving grids*, J. Comput. Phys., 174 (2001), pp. 669–694.

[9] L. Formaggia and F. Nobile, *A stability analysis for the arbitrary Lagrangian Eulerian formulation with finite elements*, East-West J. Numer. Math., 7 (1999), pp. 105–131.

[10] H. Fujita and N. Sauer, *On existence of weak solutions of the Navier-Stokes equations in regions with moving boundaries*, J. Fac. Sci. Univ. Tokyo Sect. I, 17 (1970), pp. 403–420.

[11] L. Gastaldi, *A priori error estimates for the arbitrary Lagrangian Eulerian formulation with finite elements*, East-West J. Numer. Math., 9 (2001), pp. 123–156.

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

[13] R. Glowinski, T.-W. Pan, T.I. Hesla, D.D. Joseph, and J. Périaux, *A distributed Lagrange multiplier/fictitious domain method for the simulation of flow around moving rigid bodies: Application to particulate flow*, Comput. Methods Appl. Mech. Engrg., 184 (2000), pp. 241–267.

[14] R. Glowinski, T.-W. Pan, T.I. Hesla, D.D. Joseph, and J. Périaux, *A fictitious domain approach to the direct numerical simulation of incompressible viscous flow past moving rigid bodies: Application to particulate flow*, J. Comput. Phys., 169 (2001), pp. 363–426.

[15] C. Grandmont, V. Guimet, and Y. Maday, *Numerical analysis of some decoupling techniques for the approximation of the unsteady fluid structure interaction*, Math. Models Methods Appl. Sci., 11 (2001), pp. 1349–1377.

[16] C. Grandmont and Y. Maday, *Existence for an unsteady fluid-structure interaction problem*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 609–636.

[17] M.D. Gunzburger, H.-C. Lee, and G.A. Seregin, *Global existence of weak solutions for viscous incompressible flows around a moving rigid body in three dimensions*, J. Math. Fluid Mech., 2 (2000), pp. 219–266.

[18] A. Inoue and M. Wakimoto, *On existence of solutions of the Navier-Stokes equation in a time dependent domain*, J. Fac. Sci. Univ. Tokyo Sect. IA Math., 24 (1977), pp. 303–319.

[19] B. Maury, *A many-body lubrication model*, C. R. Acad. Sci. Paris Sér. I Math., 325 (1997), pp. 1053–1058.

[20] B. Maury, *Direct simulations of* 2*D fluid-particle flows in biperiodic domains*, J. Comput. Phys., 156 (1999), pp. 325–351.

[21] B. Maury and R. Glowinski, *Fluid-particle flow: A symmetric formulation*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 1079–1084.

[22] F. Nobile, *Numerical approximation of fluid-structure interaction problems with application to haemodynamics*, Thèse de doctorat de l'École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2001.

[23] O. Pironneau, *On the transport-diffusion algorithm and its applications to the Navier-Stokes equations*, Numer. Math., 38 (1982), pp. 309–332.

[24] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.

[25] J.A. San Martín, V. Starovoitov, and M. Tucsnak, *Global weak solutions for the two-dimensional motion of several rigid bodies in an incompressible viscous fluid*, Arch. Ration. Mech. Anal., 161 (2002), pp. 113–147.

[26] E. Süli, *Convergence and nonlinear stability of the Lagrange-Galerkin method for the Navier-Stokes equations*, Numer. Math., 53 (1988), pp. 459–483.

[27] T. Takahashi, *Analysis of strong solutions for the equations modeling the motion of a rigid-fluid system in a bounded domain*, Adv. Differential Equations, 8 (2003), pp. 1499–1532.

[28] T. Takahashi, *Analyse des équations modélisant le mouvement des systèmes couplant des solides rigides et des fluides visqueux*, Thèse de doctorat de l'Université Henri Poincaré - Nancy I, Nancy, France, 2002.

[29] R. Temam, *Problèmes mathématiques en plasticité*, Gauthier-Villars, Montrouge, France, 1983.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1572–1595

A DISCONTINUOUS SUBGRID EDDY VISCOSITY METHOD FOR THE TIME-DEPENDENT NAVIER–STOKES EQUATIONS*

SONGUL KAYA† AND BÉATRICE RIVIÈRE‡

Abstract. In this paper we provide an error analysis of a subgrid scale eddy viscosity method using discontinuous polynomial approximations for the numerical solution of the incompressible Navier–Stokes equations. Optimal continuous in time error estimates of the velocity are derived. The analysis is completed with some error estimates for two fully discrete schemes, which are first and second order in time, respectively.

Key words. error analysis, Navier–Stokes, discontinuous Galerkin, fully discrete scheme, high order method

AMS subject classifications. 76F65, 74S05

DOI. 10.1137/S0036142903434862

1. Introduction.

The goal of this paper is to formulate and analyze a subgrid eddy viscosity method for solving the incompressible time-dependent Navier–Stokes equations. If the separation point between large and small scales is held fixed, the model can be viewed as a large eddy simulation (LES) model. On the other hand, if the separation point is decreased as the mesh size tends to zero, the model can be viewed (and analyzed, as herein) as a numerical regularization of the Navier–Stokes equations.

For many flows in nature, capturing all the scales in a numerical simulation is an impossible task, since the scale separation may span several orders of magnitude. Global diffusion is the traditional phenomenology to model the dispersive effects of unresolved scales on resolved scales. The traditional approach for incorporating the effects of unresolved scales on the resolved ones for the Navier–Stokes equations utilizes eddy viscosity models. These models, first formulated by Boussinesq [5] and developed by Taylor and Prandlt [10], introduce a dissipation mechanism (Smagorinsky [29]). Standard eddy viscosity models act on all scales of motion, and their effects can be too diffusive on the coarse scales (Lewandowski [26] and Iliescu and Layton [19]). The idea of applying the eddy viscosity models on only the small scales results in the subgrid eddy viscosity method, introduced and analyzed by Guermond [14], Layton [24], and John and Kaya [20]. This subgrid eddy viscosity method can also be thought of as an extension to general domains and boundary conditions of the spectral vanishing viscosity idea of Maday and Tadmor [27]. Recently, Hughes, Mazzei, and Jansen [17] proposed a variational multiscale method (VMM) in which the diffusion acts only at the finest resolved scales. VMM is a promising approach in multiscale turbulence modelling. There are different choices on how to define coarse and small scales within the VMM framework. One approach is to define fluctuations via bubble functions and means via L^2 projection (Guermond [14] and Hughes [16]). Another possibility is to

*Received by the editors September 15, 2003; accepted for publication (in revised form) February 25, 2005; published electronically October 31, 2005.

http://www.siam.org/journals/sinum/43-4/43486.html

†Department of Applied Mathematics, Illinois Institute of Technology, Chicago, IL 60616 (kaya@iit.edu).

‡Department of Mathematics, 301 Thackeray Hall, University of Pittsburgh, Pittsburgh, PA 15260 (riviere@math.pitt.edu).

define fluctuations via the finest resolved scales in a hierarchy of finite element spaces, and means via elliptic or Stokes projection (Layton [24], Kaya and Layton [22], and Hughes [18]).

For any numerical method, the error equation arising from the Navier–Stokes equations contains a convection-like term and a reaction (or stretching) term. Discontinuous Galerkin (DG) methods, first introduced in the work of Reed and Hill [28] and Lesaint and Raviart [25], are particularly efficient in controlling convective error terms. On the other hand, (generally nonlinear) eddy viscosity models are, in a sense, intended to give some control of the error's reaction-like terms. Indeed, the exponential sensitivity of trajectories of the Navier–Stokes equations (arising from reaction-like terms) is widely believed to be limited to the small scales. It is thus conjectured that by modelling their action on the large scales, the exponential sensitivity introduced by the reaction-like terms will be contained.

DG methods have recently become more popular in the science and engineering community. They use piecewise polynomial functions with no continuity constraint across element interfaces. As a result, variational formulations must include jump terms across interfaces [31]. The DG methods offers several advantages, including (i) flexibility in the design of the meshes and in the construction of trial and test spaces, (ii) local conservation of mass, (iii) h-p adaptivity, and (iv) higher order local approximations. DG methods have become widely used for solving computational fluid problems, especially diffusion and pure convection problems [3]. The reader should refer to Cockburn, Karniadakis, and Shu [6] for a historical review of DG methods. For the steady-state Navier–Stokes equations, a totally discontinuous finite element method is formulated in [12], while in [21], the velocity is approximated by discontinuous polynomials that are pointwise divergence-free, and the pressure by continuous polynomials.

Combining DG and eddy viscosity techniques is clearly advantageous. While convective effects are accurately modelled by DG, the dispersive effects of small scales on the large scales are correctly taken into account with the eddy viscosity model. Besides, due to the absence of continuity constraints, one can select various basis functions (such as hierarchical basis functions) for the coarse and refined scales. As an appropriate first step, we consider in this paper the combination of DG methods with a linear eddy viscosity model. We show that the errors are optimal with respect to the mesh size and depend on the Reynolds number in a reasonable fashion. The particular eddy viscosity model considered here was introduced in [24], and complete numerical analysis for Navier–Stokes equations was performed in [20] where it was combined with the classical finite element method.

The outline of the paper is as follows. The model problem and notation are presented in section 2. In section 3, a variational formulation and scheme are introduced. Section 4 contains the continuous in time algorithm, some stability results, and some error estimates. In section 5 , two fully discrete schemes are formulated and analyzed. Conclusions are given in the last section.

2. Notation and preliminaries.

We consider the time-dependent Navier–Stokes equations for incompressible flow as follows:

$$\boldsymbol{u}_t - \nu \Delta \boldsymbol{u} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \nabla p = \boldsymbol{f} \quad \text{in } \Omega \text{ for } 0 < t \leq T, \tag{2.1}$$

$$\nabla \cdot \boldsymbol{u} = 0 \quad \text{in } \Omega \text{ for } 0 < t \leq T, \tag{2.2}$$

$$\boldsymbol{u} = \boldsymbol{u}_0 \quad \text{in } \Omega \text{ for } t = 0, \tag{2.3}$$

$$\boldsymbol{u} = 0 \quad \text{on } \partial\Omega \text{ for } 0 < t \leq T, \tag{2.4}$$

where $\boldsymbol{u}$ is the fluid velocity, p the pressure, $\boldsymbol{f}$ the external force, $\nu > 0$ the kinematic viscosity, and $\Omega \subset \mathbb{R}^2$ a bounded, simply connected domain with polygonal boundary $\partial\Omega$. We also impose the usual normalization condition on the pressure, namely, that $\int_\Omega p = 0$.

Let $\mathcal{K}_h = \{E_j, j = 1, \dots, N_h\}$ denote a nondegenerate triangulation of the domain Ω. Let h denote the maximum diameter of the elements E_j in $\mathcal{K}_h$. We denote the edges of $\mathcal{K}_h$ by $\{e_1, e_2, \dots, e_{P_h}, e_{P_h+1}, \dots, e_{M_h}\}$, where $e_k \subset \Omega$ for $1 \le k \le P_h$ and $e_k \subset \partial\Omega$ for $P_{h+1} \le k \le M_h$. With each edge we associate a normal unit vector $\mathbf{n}_k$. For $k > P_h$, the unit vector $\mathbf{n}_k$ is taken to be outward normal to $\partial\Omega$. Let e_k be an edge shared by elements E_i and E_j with $\mathbf{n}_k$ exterior to E_i. We define the jump $[\phi]$ and average $\{\phi\}$ of a function ϕ by

$$[\phi] = (\phi|_{E_i})|_{e_k} - (\phi|_{E_j})|_{e_k}, \quad \{\phi\} = \frac{1}{2}(\phi|_{E_i})|_{e_k} + \frac{1}{2}(\phi|_{E_j})|_{e_k}.$$

If e belongs to the boundary $\partial\Omega$, the jump and average of ϕ coincide with its trace on e. We shall use standard notation for Sobolev spaces [1]. For any nonnegative integer s and $r \ge 1$, the classical Sobolev space on a domain $E \subset \mathbb{R}^2$ is

$$W^{s,r}(E) = \{v \in L^r(E) : \forall\, |m| \le s,\, \partial^m v \in L^r(E)\},$$

where $\partial^m v$ are the partial derivatives of v of order $|m|$. The usual norm in $W^{s,r}(E)$ is denoted by $\|\cdot\|_{s,r,E}$ and the seminorm by $|\cdot|_{s,r,E}$. The L^2 inner-product is denoted by $(\cdot,\cdot)_E$ and by $(\cdot,\cdot)$ if $E = \Omega$. For the Hilbert space $H^s(E) = W^{s,2}(E)$, the norm is denoted by $\|\cdot\|_{s,E}$. By $H_0^1(E)$ we shall understand the subspace of $H^1(E)$ functions that vanish on ∂E. Throughout the paper, boldface characters denote vector quantities. Define

$$\boldsymbol{V} = \{\boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega) : \quad \nabla \cdot \boldsymbol{v} = 0\}, \quad \boldsymbol{H} = \{\boldsymbol{v} \in L^2(\Omega)^2 : \quad \nabla \cdot \boldsymbol{v} = 0, \boldsymbol{v} = \boldsymbol{0}\}.$$

For any function $\boldsymbol{\phi}$ that depends on time t and space $\boldsymbol{x}$, denote

$$\boldsymbol{\phi}(t)(\boldsymbol{x}) = \boldsymbol{\phi}(t, \boldsymbol{x}) \quad \forall t \in [0, T], \forall \boldsymbol{x} \in \Omega.$$

If Y denotes a functional space in the space variable with the norm $\|\cdot\|_Y$ and if $\boldsymbol{\phi} = \boldsymbol{\phi}(t, \boldsymbol{x})$, then for $s > 0$

$$\|\boldsymbol{\phi}\|_{L^s(0,T;Y)} = \left[\int_0^T \|\boldsymbol{\phi}(t)\|_Y^s dt\right]^{1/s}, \quad \|\boldsymbol{\phi}\|_{L^\infty(0,T;Y)} = \max_{0 \le t \le T} \|\boldsymbol{\phi}(t)\|_Y.$$

Recall that for a vector function $\boldsymbol{\phi}$, the tensor $\nabla\boldsymbol{\phi}$ is defined as $(\nabla\boldsymbol{\phi})_{i,j} = \frac{\partial \phi_i}{\partial x_j}$ and the tensor product of two tensors $\boldsymbol{T}$ and $\boldsymbol{S}$ is defined as $\boldsymbol{T} : \boldsymbol{S} = \sum_{i,j} T_{ij} S_{ij}$. We define the following *broken* norm for positive s:

$$|\!|\!|\cdot|\!|\!|_s = \left[\sum_{j=1}^{N_h} \|\cdot\|_{s,E_j}^2\right]^{1/2}.$$

From [30], if $\boldsymbol{f} \in L^2(0, T; \boldsymbol{V}')$ and $\boldsymbol{u}_0 \in \boldsymbol{H}$, there exists a solution $(\boldsymbol{u}, p)$ of (2.1)–(2.4) such that $\boldsymbol{u} \in L^\infty(0, T; L^2(\Omega)^2) \cap L^2(0, T; \boldsymbol{V})$. In addition, we will assume that $\boldsymbol{u} \in L^\infty(0, T; \boldsymbol{W}^{2,4/3}(\Omega))$ and $p \in L^\infty(0, T; W^{1,4/3}(\Omega))$ for the DG formulation to be well defined. For the analysis obtained in sections 4 and 5, we require extra regularity on the solution: $\boldsymbol{u} \in L^\infty(0,T;\boldsymbol{H}^2(\Omega)), p \in L^2(0,T;H^1(\Omega))$. This assumption is valid if the data is more regular [30]: $\boldsymbol{f} \in L^\infty(0,T;\boldsymbol{H}), \boldsymbol{f}_t \in L^2(0,T;\boldsymbol{V}'), \boldsymbol{f}(0) \in \boldsymbol{H}$, $\boldsymbol{u}_0 \in \boldsymbol{H}^2(\Omega) \cap \boldsymbol{V}$. The following functional spaces are defined:

$$\begin{aligned}\boldsymbol{X} &= \{\boldsymbol{v} \in (L^2(\Omega))^2 : \boldsymbol{v}|_{E_j} \in \boldsymbol{W}^{2,4/3}(E_j) \quad \forall E_j \in \mathcal{K}_h\},\\ Q &= \{q \in L_0^2(\Omega) : q|_{E_j} \in W^{1,4/3}(E_j) \quad \forall E_j \in \mathcal{K}_h\},\end{aligned}$$

where $L_0^2(\Omega)$ is given by

$$L_0^2(\Omega) = \left\{ q \in L^2(\Omega) : \int_\Omega q = 0 \right\}.$$

We associate to $(\boldsymbol{X}, Q)$ the following norms:

$$\|\boldsymbol{v}\|_X = (\|\!|\nabla \boldsymbol{v}\|\!|_0^2 + J(\boldsymbol{v},\boldsymbol{v}))^{\frac{1}{2}} \quad \forall \boldsymbol{v} \in \boldsymbol{X}, \quad \|q\|_Q = \|q\|_{0,\Omega} \quad \forall q \in Q,$$

where the jump term J is defined as

$$J(\boldsymbol{u},\boldsymbol{v}) = \sum_{k=1}^{M_h} \frac{\sigma}{|e|} \int_{e_k} [\boldsymbol{u}] \cdot [\boldsymbol{v}]. \tag{2.5}$$

In this jump term, $|e|$ denotes the measure of the edge e and σ is a constant parameter that will be specified later.

Recall the following property of norm $\|\cdot\|_X$ [12]: for each real number $p \in [2,\infty)$ there exists a constant $C(p)$ such that

$$\|\boldsymbol{v}\|_{L^p(\Omega)} \leq C(p)\|\boldsymbol{v}\|_X \quad \forall \boldsymbol{v} \in \boldsymbol{X}. \tag{2.6}$$

For any positive integer r, the finite-dimensional subspaces are

$$\begin{aligned}\boldsymbol{X}^h &= \{\boldsymbol{v}^h \in \mathbf{X} : \boldsymbol{v}^h \in (\mathbb{P}_r(E_j))^2 \quad \forall E_j \in \mathcal{K}_h\},\\ Q^h &= \{q^h \in Q : q^h \in \mathbb{P}_{r-1}(E_j) \quad \forall E_j \in \mathcal{K}_h\}.\end{aligned}$$

We assume that for each integer $r \geq 1$, there exists an operator $R_h \in \mathcal{L}(\boldsymbol{H}^1(\Omega); \boldsymbol{X}^h)$ such that

$$\|R_h(\boldsymbol{v}) - \boldsymbol{v}\|_X \leq Ch^r |\boldsymbol{v}|_{r+1,\Omega} \quad \forall \boldsymbol{v} \in \boldsymbol{H}^{r+1}(\Omega) \cap \boldsymbol{H}_0^1(\Omega), \tag{2.7}$$

$$\|\boldsymbol{v} - \boldsymbol{R}_h(\boldsymbol{v})\|_{0,E_j} \leq Ch_{E_j}^{r+1} |\boldsymbol{v}|_{r+1,\Delta_{E_j}} \quad \forall \boldsymbol{v} \in \boldsymbol{H}^{r+1}(\Omega),\ 1 \leq j \leq N_h, \tag{2.8}$$

where Δ_{E_j} is a suitable macro element containing E_j. Note that for $r = 1, 2$, and 3, the existence of this interpolant follows from [8, 7, 9]. The bounds (2.7) and (2.8) are proved in [12] and in [13], respectively.

Also, for each integer $r \geq 1$, there is an operator $r_h \in \mathcal{L}(L_0^2(\Omega); Q_h)$ such that for any E_j in $\mathcal{K}_h$

$$\int_{E_j} z_h(r_h(q) - q) = 0 \quad \forall z_h \in \mathbb{P}_{r-1}(E_j), \forall q \in L_0^2(\Omega), \tag{2.9}$$

$$\|q - r_h(q)\|_{m,E_j} \leq Ch_{E_j}^{r-m} |q|_{r,E_j} \quad \forall q \in H^r(\Omega) \cap L_0^2(\Omega), m = 0, 1. \tag{2.10}$$

Finally, we recall some standard trace and inverse inequalities, which hold true on each element E in $\mathcal{K}_h$, with diameter h_E (see [11]):

$$\|\boldsymbol{v}\|_{0,e} \le C(h_E^{-1/2}\|\boldsymbol{v}\|_{0,E} + h_E^{1/2}\|\nabla \boldsymbol{v}\|_{0,E}) \quad \forall e \in \partial E, \quad \forall \boldsymbol{v} \in \boldsymbol{X}, \tag{2.11}$$

$$\|\nabla\boldsymbol{v}\|_{0,e} \le C(h_E^{-1/2}\|\nabla\boldsymbol{v}\|_{0,E} + h_E^{1/2}\|\nabla^2 \boldsymbol{v}\|_{0,E}) \quad \forall e \in \partial E, \quad \forall \boldsymbol{v} \in \boldsymbol{X}, \tag{2.12}$$

$$\|\boldsymbol{v}\|_{L^4(e)} \le Ch_E^{-3/4}(\|\boldsymbol{v}\|_{0,E} + h_E\|\nabla \boldsymbol{v}\|_{0,E}) \quad \forall e \in \partial E, \quad \forall \boldsymbol{v} \in \boldsymbol{X}, \tag{2.13}$$

$$\|\boldsymbol{v}^h\|_{0,e} \le Ch_E^{-1/2}\|\boldsymbol{v}^h\|_{0,E} \quad \forall e \in \partial E, \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h, \tag{2.14}$$

$$\|\nabla\boldsymbol{v}^h\|_{0,e} \le Ch_E^{-1/2}\|\nabla\boldsymbol{v}^h\|_{0,E} \quad \forall e \in \partial E, \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h, \tag{2.15}$$

$$\|\nabla\boldsymbol{v}^h\|_{0,E} \le Ch_E^{-1}\|\boldsymbol{v}^h\|_{0,E} \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h, \tag{2.16}$$

$$\|\boldsymbol{v}^h\|_{L^4(E)} \le Ch_E^{-1/2}\|\boldsymbol{v}^h\|_{0,E} \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h. \tag{2.17}$$

3. Variational formulation and scheme.

Let us first define the bilinear forms $a : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$ and $b : \boldsymbol{X} \times Q \to \mathbb{R}$:

$$a(\boldsymbol{v}, \boldsymbol{w}) = \sum_{j=1}^{N_h} \int_{E_j} \nabla \boldsymbol{v} : \nabla \boldsymbol{w} - \sum_{k=1}^{M_h} \int_{e_k} (\{\nabla \boldsymbol{v}\}\mathbf{n}_k \cdot [\boldsymbol{w}] - \epsilon_0 \{\nabla \boldsymbol{w}\}\mathbf{n}_k \cdot [\boldsymbol{v}]), \tag{3.1}$$

$$b(\boldsymbol{v}, q) = -\sum_{j=1}^{N_h} \int_{E_j} q \nabla \cdot \boldsymbol{v} + \sum_{k=1}^{M_h} \int_{e_k} \{p\}[\boldsymbol{v}] \cdot \mathbf{n}_k, \tag{3.2}$$

where ϵ_0 takes the constant value 1 or -1. Throughout the paper, we will assume the following hypothesis: if $\epsilon_0 = 1$, the jump parameter σ is chosen to be equal to 1; if $\epsilon_0 = -1$, the jump parameter σ is bounded below by $\sigma_0 > 0$ and σ_0 is sufficiently large. Based on this assumption, we can easily prove the following lemma.

LEMMA 3.1. *There is a constant $\kappa > 0$ such that*

$$a(\boldsymbol{v}^h, \boldsymbol{v}^h) + J(\boldsymbol{v}^h, \boldsymbol{v}^h) \ge \kappa \|\boldsymbol{v}^h\|_X^2 \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h. \tag{3.3}$$

In addition to these bilinear forms, we consider the following upwind discretization of the term $\boldsymbol{u} \cdot \nabla \boldsymbol{z}$:

$$\begin{aligned} c(\boldsymbol{u}, \boldsymbol{z}, \boldsymbol{\theta}) = &\sum_{j=1}^{N_h} \left(\int_{E_j} (\boldsymbol{u} \cdot \nabla \boldsymbol{z}) \cdot \boldsymbol{\theta} + \int_{\partial E_j^-} |\{\boldsymbol{u}\} \cdot \boldsymbol{n}_{E_j}| (\boldsymbol{z}^{\text{int}} - \boldsymbol{z}^{\text{ext}}) \cdot \boldsymbol{\theta}^{int} \right) \\ &+ \frac{1}{2} \sum_{j=1}^{N_h} \int_{E_j} (\nabla \cdot \boldsymbol{u}) \boldsymbol{z} \cdot \boldsymbol{\theta} - \frac{1}{2} \sum_{k=1}^{M_h} \int_{e_k} [\boldsymbol{u}] \cdot \boldsymbol{n}_k \{\boldsymbol{z} \cdot \boldsymbol{\theta}\} \end{aligned} \tag{3.4}$$

for all $\boldsymbol{u}, \boldsymbol{z}, \boldsymbol{\theta}$ in $\mathbf{X}$ and where on each element the inflow boundary is

$$\partial E_j^- = \{\boldsymbol{x} \in \partial E_j : \{\boldsymbol{u}\} \cdot \mathbf{n}_{E_j} < 0\},$$

and the superscript int (resp., ext) refers to the trace of the function on a side of E_j coming from the interior of E_j (resp., coming from the exterior of E_j on that side). Note that the form c is not linear with respect to its first argument but is linear with respect to its second and third arguments. To avoid any confusion, if necessary, in the analysis, we will explicitly write $c(\boldsymbol{u}, \boldsymbol{z}, \boldsymbol{\theta}) = c_{\boldsymbol{w}}(\boldsymbol{u}, \boldsymbol{z}, \boldsymbol{\theta})$ when the inflow boundaries ∂E_j^- are defined with respect to the velocity $\{\boldsymbol{w}\}$. We finally recall the positivity of c proved in [12]:

$$c(\boldsymbol{u}, \boldsymbol{z}, \boldsymbol{z}) \ge 0 \quad \forall \boldsymbol{u}, \boldsymbol{z} \in \boldsymbol{X}. \tag{3.5}$$

With these forms, we consider a variational problem of (2.1)–(2.4): for all $t > 0$ find $\boldsymbol{u}(t) \in \boldsymbol{X}$ and $p(t) \in Q$ satisfying

$$(\boldsymbol{u}_t(t), \boldsymbol{v}) + \nu(a(\boldsymbol{u}(t), \boldsymbol{v}) + J(\boldsymbol{u}(t), \boldsymbol{v})) \\ + c(\boldsymbol{u}(t), \boldsymbol{u}(t), \boldsymbol{v}) + b(\boldsymbol{v}, p(t)) = (\boldsymbol{f}(t), \boldsymbol{v}) \quad \forall \boldsymbol{v} \in \boldsymbol{X}, \tag{3.6}$$

$$b(\boldsymbol{u}(t), q) = 0 \quad \forall q \in Q, \tag{3.7}$$

$$(\boldsymbol{u}(0), \boldsymbol{v}) = (\boldsymbol{u}_0, \boldsymbol{v}) \quad \forall \boldsymbol{v} \in \boldsymbol{X}. \tag{3.8}$$

We shall now show the equivalence of the strong and weak solutions.

Lemma 3.2. *Every strong solution of* (2.1)–(2.4) *is also a solution of* (3.6)–(3.8) *and conversely.*

Proof. Fix $t > 0$. Let $(\boldsymbol{u}, p)$ be the solution of (2.1)–(2.4). Since $\boldsymbol{u}(t) \in \boldsymbol{H}_0^1(\Omega)$, by the trace theorem $[\boldsymbol{u}(t)] \cdot \boldsymbol{n}_k = 0$ on each edge. Also, $\nabla \cdot \boldsymbol{u}(t) = 0$; thus $\boldsymbol{u}$ satisfies (3.7). Multiplying the Navier–Stokes equation (2.1) by $\boldsymbol{v} \in \boldsymbol{X}$, integrating over each element, and summing over all elements yield

$$\sum_{j=1}^{N_h} \int_{E_j} (\boldsymbol{u}_t \cdot \boldsymbol{v} + \nu \nabla \boldsymbol{u} : \nabla \boldsymbol{v}) - \nu \sum_{k=1}^{M_h} \int_{e_k} [\nabla \boldsymbol{u} \mathbf{n}_k \cdot \boldsymbol{v}] + \sum_{j=1}^{N_h} \int_{E_j} \boldsymbol{u} \cdot \nabla \boldsymbol{u} \cdot \boldsymbol{v} \\ - \sum_{j=1}^{N_h} \int_{E_j} p \nabla \cdot \boldsymbol{v} + \sum_{k=1}^{M_h} \int_{e_k} [p \boldsymbol{v} \cdot \mathbf{n}_k] = \int_\Omega \boldsymbol{f} \cdot \boldsymbol{v}.$$

The boundary terms are rewritten as

$$\sum_{k=1}^{M_h} \int_{e_k} [\nabla \boldsymbol{u} \boldsymbol{n}_k . \boldsymbol{v}] = \sum_{k=1}^{M_h} \int_{e_k} \{\nabla \boldsymbol{u}\} \boldsymbol{n}_k \cdot [\boldsymbol{v}] + \sum_{k=1}^{M_h} \int_{e_k} [\nabla \boldsymbol{u}] \boldsymbol{n}_k \cdot \{\boldsymbol{v}\}.$$

The first part of the lemma is then obtained because the jumps of $\boldsymbol{u}, \nabla \boldsymbol{u} \boldsymbol{n}_k$, and p are zero almost everywhere.

Conversely, let $(\boldsymbol{u}, p)$ be a solution to (3.6)–(3.8). First, let E belong to $\mathcal{K}_h$ and choose $\boldsymbol{v} \in \mathcal{D}(E)^2$, extended by zero outside E. Then, $(\boldsymbol{u}, p)$ satisfy in the sense of distributions

$$\boldsymbol{u}_t - \nu \Delta \boldsymbol{u} + \boldsymbol{u} \cdot \nabla \boldsymbol{u} + \nabla p = \boldsymbol{f}, \quad \nabla \cdot \boldsymbol{u} = 0 \quad \text{in} \quad E. \tag{3.9}$$

Next consider $\boldsymbol{v} \in \mathcal{C}^1(\bar{E})$ such that $\boldsymbol{v} = \boldsymbol{0}$ on ∂E, extended by zero outside E, and $\nabla \boldsymbol{v} \cdot \boldsymbol{n} = 0$ on ∂E except on one side e_k. We multiply (3.9) by $\boldsymbol{v}$ and integrate by parts. We then obtain

$$\int_{e_k} \{\nabla \boldsymbol{v}\} \boldsymbol{n}_k \cdot [\boldsymbol{u}] = 0,$$

which implies that $[\boldsymbol{u}] = \boldsymbol{0}$ almost everywhere on e_k. If e_k belongs to the boundary $\partial\Omega$, this implies that $\boldsymbol{u}|_{e_k} = \boldsymbol{0}$. Thus, $\boldsymbol{u} \in \boldsymbol{H}_0^1(\Omega)$. Finally, choose $\boldsymbol{v} \in \mathcal{C}^1(\bar{E})$, with $\boldsymbol{v} = \boldsymbol{0}$ on ∂E except on one side e_k, extended by zero outside of E. Multiplying (3.9) by v and integrating by parts, we have

$$\int_{e_k} (-\nu \nabla \boldsymbol{u} \boldsymbol{n}_E + p \boldsymbol{n}_E) \cdot \boldsymbol{v} = \int_{e_k} \{-\nu \nabla \boldsymbol{u} \boldsymbol{n}_E + p \boldsymbol{n}_E\} \cdot \boldsymbol{v}.$$

Since $\boldsymbol{v}$ is arbitrary, this means that the quantity $-\nu\nabla\boldsymbol{u}\boldsymbol{n}_k + p\boldsymbol{n}_k$ is continuous across e_k. Therefore, (3.9) is satisfied over the entire domain Ω. The initial condition (2.3) is straightforward. □

We recall a discrete inf-sup condition and a property satisfied by R_h (see [12]).

LEMMA 3.3. *There exists a positive constant β_0, independent of h such that*

$$\inf_{q^h\in Q^h} \sup_{\boldsymbol{v}^h\in\mathbf{X}^h} \frac{b(\boldsymbol{v}^h, q^h)}{\|\boldsymbol{v}^h\|_X \, \|q^h\|_0} \geq \beta_0. \tag{3.10}$$

Furthermore, the operator R_h satisfies

$$b(R_h(\boldsymbol{v}) - \boldsymbol{v}, q^h) = 0 \quad \forall q^h \in Q^h, \quad \forall \boldsymbol{v} \in \boldsymbol{H}_0^1(\Omega). \tag{3.11}$$

In order to subtract the artificial diffusion introduced by the eddy viscosity on the coarse grid, we consider a coarsening of the mesh $\mathcal{K}_h$, namely $\mathcal{K}_H$, such that the fine mesh $\mathcal{K}_h$ is a refinement of $\mathcal{K}_H$ (so typically $h \ll H$). Denote by $\boldsymbol{L}$ the space of tensors $L^2(\Omega)^{2\times 2}$ and consider the finite-dimensional subspace of $\boldsymbol{L}$:

$$\boldsymbol{L}_H = \{\boldsymbol{S} \in \boldsymbol{L} : S_{ij}|_\Sigma \in \mathbb{P}_{r-1}(\Sigma)\, \forall \Sigma \in \mathcal{K}_H\}.$$

Let $P_H : \boldsymbol{L} \to \boldsymbol{L}_H$ denote the L^2 orthogonal projection on $\boldsymbol{L}_H$ and let I denote the identity mapping. Since P_H is a projection, we have the following properties:

$$\|I - P_H\| \leq 1, \tag{3.12}$$

$$\|(I - P_H)\nabla\boldsymbol{v}\|_{0,\Omega} \leq CH^r|\boldsymbol{v}|_{r+1,\Omega} \quad \forall \boldsymbol{v} \in \boldsymbol{H}^{r+1}(\Omega). \tag{3.13}$$

Throughout the paper, the variable C will denote a generic positive constant that will take different values at different places but will be independent of h, H, ν, and ν_T. Define the following bilinear $g : \boldsymbol{X} \times \boldsymbol{X} \to \mathbb{R}$:

$$g(\boldsymbol{v}, \boldsymbol{w}) = \sum_{j=1}^{N_h} \int_{E_j} (I - P_H)\nabla\boldsymbol{v} : (I - P_H)\nabla\boldsymbol{w} \quad \forall \boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{X}.$$

For all $t > 0$, we seek a discontinuous approximation $(\boldsymbol{u}^h(t), p^h(t)) \in \mathbf{X}^h \times Q^h$ such that

$$\begin{aligned} &(\boldsymbol{u}_t^h(t), \boldsymbol{v}^h) + \nu(a(\boldsymbol{u}^h(t), \boldsymbol{v}^h) + J(\boldsymbol{u}^h(t), \boldsymbol{v}^h)) + \nu_T g(\boldsymbol{u}^h(t), \boldsymbol{v}^h) \\ &+ c(\boldsymbol{u}^h(t), \boldsymbol{u}^h(t), \boldsymbol{v}^h) + b(\boldsymbol{v}^h, p^h(t)) = (\boldsymbol{f}(t), \boldsymbol{v}^h) \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h, \end{aligned} \tag{3.14}$$

$$b(\boldsymbol{u}^h(t), q^h) = 0 \quad \forall q^h \in Q^h, \tag{3.15}$$

$$(\boldsymbol{u}^h(0), \boldsymbol{v}^h) = (\boldsymbol{u}_0, \boldsymbol{v}^h) \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h. \tag{3.16}$$

LEMMA 3.4. *There exists a unique solution to* (3.14)–(3.16).

Proof. Equations (3.14) and (3.15) reduce to the ordinary differential system

$$\frac{d\boldsymbol{u}^h}{dt} + \nu A\boldsymbol{u}^h + B\boldsymbol{u}^h + \nu_T G\boldsymbol{u}^h = \mathbf{F}.$$

By continuity, a solution exists. To prove uniqueness, we choose $\boldsymbol{v}^h = \boldsymbol{u}^h$ in (3.14) and $q^h = p^h$ in (3.15); we apply the coercivity equation (3.3) and the generalized Cauchy–Schwarz

$$\frac{1}{2}\frac{d}{dt}\|\boldsymbol{u}^h\|_{0,\Omega}^2 + \nu\kappa\|\boldsymbol{u}^h\|_X^2 \leq \|\boldsymbol{f}\|_{L^{4/3}(\Omega)}\|\boldsymbol{u}^h\|_{L^4(\Omega)} \leq \frac{\nu\kappa}{2}\|\boldsymbol{u}^h\|_X^2 + \frac{C}{\nu\kappa}\|\boldsymbol{f}\|_{L^{4/3}(\Omega)}^2.$$

Integrating over $[0,t]$ yields

$$\|\boldsymbol{u}^h(t)\|^2_{L^\infty(0,T;L^2(\Omega))} + \nu\kappa\|\boldsymbol{u}^h\|^2_{L^2(0,T;X)} \leq \|\boldsymbol{u}^h(0)\|^2_0 + \frac{C}{\nu\kappa}\|\boldsymbol{f}\|^2_{L^2(0,T;L^{4/3}(\Omega))}.$$

Since $\boldsymbol{u}^h$ is bounded in $L^\infty(0,T;L^2(\Omega)^2)$, it is unique [4]. The existence and uniqueness of p^h are obtained from the inf-sup condition stated above. □

Remark 1. From a continuum mechanics point of view, it might be advantageous to consider the symmetrized velocity tensor. In this case, the bilinear form a is replaced by

$$a(\boldsymbol{v},\boldsymbol{w}) = \sum_{j=1}^{N_h}\int_{E_j}\nabla^s\boldsymbol{v} : \nabla^s\boldsymbol{w} - \sum_{k=1}^{M_h}\int_{e_k}(\{\nabla^s\boldsymbol{v}\}\mathbf{n}_k\cdot[\boldsymbol{w}] - \epsilon_0\{\nabla^s\boldsymbol{w}\}\mathbf{n}_k\cdot[\boldsymbol{v}]),$$

where $\nabla^s\boldsymbol{v} = 0.5(\nabla\boldsymbol{v} + \nabla\boldsymbol{v}^T)$ and the term relating the coarse and refined meshes is replaced by $\sum_{j=1}^{N_h}\int_{E_j}(I-P_H)\nabla^s\boldsymbol{u} : (I-P_H)\nabla^s\boldsymbol{v}^h$. It is easy to check that all the results proved in this paper also hold true for the symmetrized tensor formulation.

4. Semidiscrete a priori error estimate. In this section, a priori error estimates for the continuous in time problem are derived. The estimates are optimal in the fine mesh size h. The effects of the coarse scale appear as higher order terms.

THEOREM 4.1. *Let $(\boldsymbol{u},p)$ be the solution of* (2.1)–(2.4) *satisfying $\boldsymbol{u}\in L^\infty(0,T;\boldsymbol{H}^2(\Omega)), p\in L^2(0,T;H^1(\Omega))$. In addition, we assume that $\boldsymbol{u}_t\in L^2(0,T;\boldsymbol{H}^{r+1}(\Omega))$, $\boldsymbol{u}\in L^\infty(0,T;\boldsymbol{H}^{r+1}(\Omega))$, and $p\in L^2(0,T;\boldsymbol{H}^r(\Omega))$. Then, the continuous in time solution $\boldsymbol{u}_h$ satisfies*

$$\begin{aligned}&\|\boldsymbol{u}-\boldsymbol{u}^h\|_{L^\infty(0,T;L^2(\Omega))} + \kappa^{1/2}\nu^{1/2}\|\boldsymbol{u}-\boldsymbol{u}^h\|_{L^2(0,T;X)}\\ &\qquad + \nu_T^{1/2}\|(I-P_H)\nabla(\boldsymbol{u}-\boldsymbol{u}^h)\|_{L^2(0,T;L^2(\Omega))}\\ &\leq Ce^{CT(\nu^{-1}+1)}[h^r((\nu+\nu^{-1}+\nu_T)^{1/2}|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))} + \nu^{-1/2}|p|_{L^2(0,T;H^r(\Omega))}\\ &\qquad + |\boldsymbol{u}_t|_{L^2(0,T;H^{r+1}(\Omega))}) + \nu_T^{1/2}H^r|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}] + Ch^r|\boldsymbol{u}_0|_{r+1,\Omega},\end{aligned}$$

where C is a positive constant independent of h, H, ν and ν_T.

Proof. We fix $t>0$ and for simplicity, we drop the argument in t. Defining $\boldsymbol{e}^h = \boldsymbol{u}-\boldsymbol{u}^h$ and subtracting (3.14), (3.15), (3.16) from (3.6), (3.7), (3.8), respectively, yields

$$\begin{aligned}&(\boldsymbol{e}^h_t,\boldsymbol{v}^h) + \nu a(\boldsymbol{e}^h,\boldsymbol{v}^h) + \nu J(\boldsymbol{e}^h,\boldsymbol{v}^h) + \nu_T g(\boldsymbol{e}^h,\boldsymbol{v}^h) + c(\boldsymbol{u},\boldsymbol{u},\boldsymbol{v}^h)\\ &\quad - c(\boldsymbol{u}^h,\boldsymbol{u}^h,\boldsymbol{v}^h) = -b(\boldsymbol{v}^h,p-p^h) + \nu_T g(\boldsymbol{u},\boldsymbol{v}^h) \quad \forall\boldsymbol{v}^h\in\boldsymbol{X}_h, \quad \forall t>0,\end{aligned} \tag{4.1}$$

$$b(\boldsymbol{e}^h,q^h) = 0 \quad \forall q^h\in Q^h, \quad \forall t>0, \tag{4.2}$$

$$(\boldsymbol{e}^h(0),\boldsymbol{v}^h) = 0, \quad \forall\boldsymbol{v}^h\in\boldsymbol{X}^h. \tag{4.3}$$

Decompose the error $\boldsymbol{e}^h = \boldsymbol{\eta} - \boldsymbol{\phi}^h$, where $\boldsymbol{\phi}^h = \boldsymbol{u}^h - R_h(\boldsymbol{u})$ and $\boldsymbol{\eta}$ is the interpolation error $\boldsymbol{\eta} = \boldsymbol{u} - R_h(\boldsymbol{u})$. Set $\boldsymbol{v}^h = \boldsymbol{\phi}^h$ in (4.1) and $q^h = r_h(p) - p_h$ in (4.2):

$$\begin{aligned}&(\boldsymbol{\phi}^h_t,\boldsymbol{\phi}^h) + \nu a(\boldsymbol{\phi}^h,\boldsymbol{\phi}^h) + \nu J(\boldsymbol{\phi}^h,\boldsymbol{\phi}^h) + \nu_T g(\boldsymbol{\phi}^h,\boldsymbol{\phi}^h)\\ &\quad + c_{\boldsymbol{u}^h}(\boldsymbol{u}^h,\boldsymbol{u}^h,\boldsymbol{\phi}^h) - c_{\boldsymbol{u}}(\boldsymbol{u},\boldsymbol{u},\boldsymbol{\phi}^h) = (\boldsymbol{\eta}_t,\boldsymbol{\phi}^h) + \nu a(\boldsymbol{\eta},\boldsymbol{\phi}^h) + \nu J(\boldsymbol{\eta},\boldsymbol{\phi}^h)\\ &\quad + \nu_T g(\boldsymbol{\eta},\boldsymbol{\phi}^h) + b(\boldsymbol{\phi}^h,p-r_h(p)) - \nu_T g(\boldsymbol{u},\boldsymbol{\phi}^h) \quad \forall t>0.\end{aligned} \tag{4.4}$$

We now bound the terms on the right hand-side of (4.4). The first three terms are rewritten as

$$\begin{aligned}(\boldsymbol{\eta}_t, \boldsymbol{\phi}^h) + \nu a(\boldsymbol{\eta}, \boldsymbol{\phi}^h) + \nu J(\boldsymbol{\eta}, \boldsymbol{\phi}^h) &= (\boldsymbol{\eta}_t, \boldsymbol{\phi}^h) + \nu \sum_{j=1}^{N_h} \int_{E_j} \nabla \boldsymbol{\eta} : \nabla \boldsymbol{\phi}^h \\ - \nu \sum_{k=1}^{M_h} \int_{e_k} \{\nabla \boldsymbol{\eta}\} \boldsymbol{n}_k \cdot [\boldsymbol{\phi}^h] &+ \nu \epsilon_0 \sum_{k=1}^{M_h} \int_{e_k} \{\nabla \boldsymbol{\phi}^h\} \boldsymbol{n}_k \cdot [\boldsymbol{\eta}] + \nu J(\boldsymbol{\eta}, \boldsymbol{\phi}^h) \\ &= S_1 + \cdots + S_5.\end{aligned}$$

Using the Cauchy–Schwarz and Young's inequalities and the approximation result (2.7), the first two terms are bounded as follows:

$$\begin{aligned}S_1 &\le \|\boldsymbol{\eta}_t\|_{0,\Omega} \|\boldsymbol{\phi}^h\|_{0,\Omega} \le \frac{1}{2} \|\boldsymbol{\phi}^h\|_{0,\Omega}^2 + C h^{2r+2} |\boldsymbol{u}_t|_{r+1,\Omega}^2, \\ S_2 &\le \nu \sum_{j=1}^{N_h} \|\nabla \boldsymbol{\eta}\|_{0,E_j} \|\nabla \boldsymbol{\phi}^h\|_{0,E_j} \le \frac{\kappa \nu}{8} \|\nabla \boldsymbol{\phi}^h\|_0^2 + C \nu h^{2r} |\boldsymbol{u}|_{r+1,\Omega}^2.\end{aligned}$$

To bound the third term, we insert the standard Lagrange interpolant of degree r, denoted by $L_h(\boldsymbol{u})$:

$$\begin{aligned}- \nu \sum_{k=1}^{M_h} \int_{e_k} \{\nabla \boldsymbol{\eta}\} \boldsymbol{n}_k \cdot [\boldsymbol{\phi}^h] &= - \nu \sum_{k=1}^{M_h} \int_{e_k} \{\nabla (\boldsymbol{u} - L_h(\boldsymbol{u}))\} \boldsymbol{n}_k \cdot [\boldsymbol{\phi}^h] \\ &\quad - \nu \sum_{k=1}^{M_h} \int_{e_k} \{\nabla (L_h(\boldsymbol{u}) - R_h(\boldsymbol{u}))\} \boldsymbol{n}_k \cdot [\boldsymbol{\phi}^h].\end{aligned}$$

By using inequalities (2.12) and (2.15), the definition of the jump (2.5), and the approximation results (2.7), the third term can be bounded by

$$S_3 \le \frac{\kappa \nu}{12} J(\boldsymbol{\phi}^h, \boldsymbol{\phi}^h) + C \nu h^{2r} |\boldsymbol{u}|_{r+1,\Omega}^2.$$

Then, from the trace inequalities (2.11) and (2.15) and the approximation result (2.7), we have

$$\begin{aligned}S_4 &\le C \nu \left(\sum_{k=1}^{M_h} \frac{\sigma}{|e|} \|[\boldsymbol{\eta}]\|_{0,e_k}^2 \right)^{1/2} \left(\sum_{k=1}^{M_h} \frac{|e|}{\sigma} \|\{\nabla \boldsymbol{\phi}^h\}\|_{0,e_k}^2 \right)^{1/2} \\ &\le \frac{\kappa \nu}{8} \|\nabla \boldsymbol{\phi}^h\|_0^2 + C \nu h^{2r} |\boldsymbol{u}|_{r+1,\Omega}^2.\end{aligned}$$

The jump term is bounded by the approximation result (2.7) as follows:

$$S_5 \le \frac{\kappa \nu}{12} J(\boldsymbol{\phi}^h, \boldsymbol{\phi}^h) + C \nu J(\boldsymbol{\eta}, \boldsymbol{\eta}) \le \frac{\kappa \nu}{12} J(\boldsymbol{\phi}^h, \boldsymbol{\phi}^h) + C \nu h^{2r} |\boldsymbol{u}|_{r+1,\Omega}^2.$$

The eddy viscosity term in the right-hand side of (4.4) is bounded by (3.12) and (2.7):

$$\nu_T g(\boldsymbol{\eta}, \boldsymbol{\phi}^h) \le \frac{\nu_T}{4} \|(I - P_H) \nabla \boldsymbol{\phi}^h\|_0^2 + C \nu_T h^{2r} |\boldsymbol{u}|_{r+1,\Omega}^2.$$

Because of (2.9), the pressure term is reduced to

$$b(\boldsymbol{\phi}^h, p - r_h(p)) = \sum_{k=1}^{M_h} \int_{e_k} \{p - r_h(p)\}[\boldsymbol{\phi}^h] \cdot \boldsymbol{n}_k,$$

which is bounded by using the Cauchy–Schwarz inequality, trace inequality (2.11), and the approximation result (2.10):

$$\begin{aligned} b(\boldsymbol{\phi}^h, p - r_h(p)) &\leq C \left(\|p - r_h(p)\|_0^2 + \sum_{j=1}^{N_h} h_{E_j}^2 |p - r_h(p)|_{1,E_j}^2 \right)^{1/2} J(\boldsymbol{\phi}^h, \boldsymbol{\phi}^h)^{1/2} \\ &\leq \frac{\kappa\nu}{12} J(\boldsymbol{\phi}^h, \boldsymbol{\phi}^h) + C\frac{h^{2r}}{\nu} |p|_{r,\Omega}^2. \end{aligned}$$

The last term on the right-hand side of (4.4), corresponding to the consistency error, is bounded using the Cauchy–Schwarz inequality and the bound (3.13):

$$\nu_T g(\boldsymbol{u}, \boldsymbol{\phi}^h) \leq \frac{\nu_T}{4} \|\!|(I - P_H)\nabla \boldsymbol{\phi}^h\|\!|_0^2 + C\nu_T H^{2r} |\boldsymbol{u}|_{r+1,\Omega}^2.$$

Thus far, the terms in the right-hand side of (4.4) are bounded by

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{\phi}^h\|_0^2 + Ch^{2r}|\boldsymbol{u}_t|_{r+1,\Omega}^2 + C(\nu + \nu_T)h^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2 + C\frac{h^{2r}}{\nu}|p|_{r,\Omega}^2 \\ &\quad + C\nu_T H^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2 + \frac{\kappa\nu}{4}\|\boldsymbol{\phi}^h\|_X^2 + \frac{\nu_T}{2}\|\!|(I - P_H)\nabla\boldsymbol{\phi}^h\|\!|_0^2. \end{aligned}$$

Consider now the nonlinear terms in (4.4). We first note that since $\boldsymbol{u}$ is continuous, the second term in (3.4) vanishes and can be replaced by a similar quantity with a different domain of integration:

$$c_{\boldsymbol{u}}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{\phi}^h) = c_{\boldsymbol{u}^h}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{\phi}^h).$$

Therefore, adding and subtracting the interpolant $R_h(\boldsymbol{u})$ yields

$$\begin{aligned} c_{\boldsymbol{u}^h}(\boldsymbol{u}^h, \boldsymbol{u}^h, \boldsymbol{\phi}^h) - c_{\boldsymbol{u}^h}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{\phi}^h) &= c_{\boldsymbol{u}^h}(\boldsymbol{u}^h, \boldsymbol{\phi}^h, \boldsymbol{\phi}^h) + c_{\boldsymbol{u}^h}(\boldsymbol{\phi}^h, \boldsymbol{u}, \boldsymbol{\phi}^h) \\ &- c_{\boldsymbol{u}^h}(\boldsymbol{\phi}^h, \boldsymbol{\eta}, \boldsymbol{\phi}^h) - c_{\boldsymbol{u}^h}(\boldsymbol{\eta}, R_h(\boldsymbol{u}), \boldsymbol{\phi}^h) - c_{\boldsymbol{u}^h}(\boldsymbol{u}, \boldsymbol{\eta}, \boldsymbol{\phi}^h). \end{aligned}$$

To simplify the writing, we drop the subscript $\boldsymbol{u}_h$ and write $c(\cdot,\cdot,\cdot)$ for $c_{\boldsymbol{u}_h}(\cdot,\cdot,\cdot)$. From inequality (3.5), the first term is positive. We then bound the other terms. We first note that we can rewrite the form c as

$$c(\boldsymbol{\phi}^h, \boldsymbol{u}, \boldsymbol{\phi}^h) = \sum_{j=1}^{N_h} \int_{E_j} (\boldsymbol{\phi}^h \cdot \nabla \boldsymbol{u}) \cdot \boldsymbol{\phi}^h - \frac{1}{2} b(\boldsymbol{\phi}^h, \boldsymbol{u} \cdot \boldsymbol{\phi}^h). \tag{4.5}$$

The first term, using the L^p bound (2.6), is bounded by

$$\begin{aligned} \sum_{j=1}^{N_h} \int_{E_j} (\boldsymbol{\phi}^h \cdot \nabla \boldsymbol{u}) \cdot \boldsymbol{\phi}^h &\leq \|\boldsymbol{\phi}^h\|_{L^4(\Omega)} \|\nabla \boldsymbol{u}\|_{L^4(\Omega)} \|\boldsymbol{\phi}^h\|_{L^2(\Omega)} \\ &\leq \frac{\kappa\nu}{64}\|\boldsymbol{\phi}^h\|_X^2 + \frac{C}{\nu}\|\boldsymbol{u}\|_{L^\infty(0,T;W^{2,4/3}(\Omega))}^2 \|\boldsymbol{\phi}^h\|_{0,\Omega}^2. \end{aligned}$$

Let $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ be the piecewise constant vectors such that

$$\boldsymbol{c}_1|_{E_j} = \frac{1}{|E_j|}\int_{E_j} \boldsymbol{u}, \quad \boldsymbol{c}_2|_{E_j} = \frac{1}{|E_j|}\int_{E_j} \boldsymbol{\phi}^h, \quad 1 \le j \le N_h.$$

We rewrite using (4.2) and (3.11):

$$b(\boldsymbol{\phi}^h, \boldsymbol{u}\cdot\boldsymbol{\phi}^h) = b(\boldsymbol{\phi}^h, \boldsymbol{u}\cdot\boldsymbol{\phi}^h - \boldsymbol{c}_1\cdot\boldsymbol{c}_2) = b(\boldsymbol{\phi}^h, (\boldsymbol{u}-\boldsymbol{c}_1)\cdot\boldsymbol{\phi}^h) + b(\boldsymbol{\phi}^h, \boldsymbol{c}_1\cdot(\boldsymbol{\phi}^h - \boldsymbol{c}_2)).$$

Then, expanding the first term,

$$\begin{aligned} b(\boldsymbol{\phi}^h, (\boldsymbol{u}-\boldsymbol{c}_1)\cdot\boldsymbol{\phi}^h) &= -\sum_{j=1}^{N_h}\int_E (\boldsymbol{u}-\boldsymbol{c}_1)\cdot\boldsymbol{\phi}^h \nabla\cdot\boldsymbol{\phi}^h \\ &+ \sum_{k=1}^{M_h}\int_{e_k} \{(\boldsymbol{u}-\boldsymbol{c}_1)\cdot\boldsymbol{\phi}^h\}[\boldsymbol{\phi}^h]\cdot\boldsymbol{n}_k = S_6 + S_7. \end{aligned}$$

The first term is bounded, for $s > 2$, using the inverse inequality (2.16) and (2.6):

$$\begin{aligned} S_6 &\le C\sum_{j=1}^{N_h} \|\boldsymbol{u}-\boldsymbol{c}_1\|_{L^s(E_j)}\|\boldsymbol{\phi}^h\|_{L^{\frac{2s}{s-2}}(E_j)}\|\nabla\boldsymbol{\phi}^h\|_{L^2(E_j)} \\ &\le C\|\boldsymbol{\phi}^h\|_{0,\Omega}|\boldsymbol{u}|_{W^{1,s}(\Omega)}\|\boldsymbol{\phi}^h\|_{L^{\frac{2s}{s-2}}(\Omega)} \\ &\le C\|\boldsymbol{\phi}^h\|_{0,\Omega}|\boldsymbol{u}|_{W^{1,s}(\Omega)}\|\boldsymbol{\phi}^h\|_X \le \frac{\kappa\nu}{64}\|\boldsymbol{\phi}^h\|_X^2 + \frac{C}{\nu}\|\boldsymbol{u}\|^2_{L^\infty(0,T;W^{2,4/3}(\Omega))}\|\boldsymbol{\phi}^h\|_0^2. \end{aligned}$$

The bound for the second term is more technical. First, passing to the reference element $\hat{E}$ and using the trace inequality (2.14), we obtain

$$\begin{aligned} S_7 &\le C\sum_{k=1}^{M_h} |e_k||E|^{-1/2}\|\boldsymbol{\phi}^h\|_{0,E}\|(\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1)\cdot\hat{\boldsymbol{\phi}}^h\|_{\hat{e}} \\ &\le C\sum_{k=1}^{M_h} |e_k||E|^{-1/2}\|\boldsymbol{\phi}^h\|_{0,E}(\|(\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1)\cdot\hat{\boldsymbol{\phi}}^h\|_{0,\hat{E}} + \|\hat{\nabla}((\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1)\cdot\hat{\boldsymbol{\phi}}^h)\|_{0,\hat{E}}). \end{aligned}$$

The L^2 term is bounded, for $s > 2$, as

$$\begin{aligned} \|(\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1)\cdot\hat{\boldsymbol{\phi}}^h\|_{0,\hat{E}} &\le \|\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1\|_{L^s(\hat{E})}\|\hat{\boldsymbol{\phi}}^h\|_{L^{\frac{2s}{s-2}}(\hat{E})} \\ &\le h|E|^{-1/s-(s-2)/(2s)}|\boldsymbol{u}|_{W^{1,s}(E)}\|\boldsymbol{\phi}^h\|_{L^{\frac{2s}{s-2}}(E)} \le C|\boldsymbol{u}|_{W^{1,s}(E)}\|\boldsymbol{\phi}^h\|_{L^{\frac{2s}{s-2}}(E)}. \end{aligned}$$

Note that for the gradient term we write

$$\|\hat{\nabla}((\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1)\cdot\hat{\boldsymbol{\phi}}^h)\|_{0,\hat{E}} = \|(\hat{\nabla}\hat{\boldsymbol{u}}\cdot\hat{\boldsymbol{\phi}}^h + (\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1)\cdot\nabla\hat{\boldsymbol{\phi}}^h)\|.$$

Let us first bound

$$\begin{aligned} \|\hat{\nabla}\hat{\boldsymbol{u}}\cdot\hat{\boldsymbol{\phi}}^h\|_{0,\hat{E}} &\le \|\hat{\nabla}\hat{\boldsymbol{u}}\|_{L^s(\hat{E})}\|\hat{\boldsymbol{\phi}}^h\|_{L^{\frac{2s}{s-2}}(\hat{E})} \\ &\le Ch|E|^{-1/s}\|\nabla\boldsymbol{u}\|_{L^s(E)}|E|^{-(s-2)/2s}\|\boldsymbol{\phi}^h\|_{L^{\frac{2s}{s-2}}(E)} \le C\|\nabla\boldsymbol{u}\|_{L^s(E)}\|\boldsymbol{\phi}^h\|_{L^{\frac{2s}{s-2}}(E)}. \end{aligned}$$

Now the other term is

$$\|(\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1)\cdot\hat{\nabla}\hat{\boldsymbol{\phi}}^h\|_{0,\hat{E}} \le \|\hat{\boldsymbol{u}}-\hat{\boldsymbol{c}}_1\|_{L^\infty(\hat{E})}\|\hat{\nabla}\hat{\boldsymbol{\phi}}^h\|_{0,\hat{E}} \le Ch\|\boldsymbol{u}\|_{L^\infty(E)}\|\nabla\boldsymbol{\phi}^h\|_{0,E}.$$

Combining all the bounds above and using (2.6), we have

$$S_7 \le C\sum_{j=1}^{N_h}\|\boldsymbol{\phi}^h\|_{0,E_j}\Big[|\boldsymbol{u}|_{W^{1,s}(E_j)}\|\boldsymbol{\phi}^h\|_{L^{\frac{2s}{s-2}}(E_j)}$$
$$+\|\nabla\boldsymbol{u}\|_{L^s(E_j)}\|\boldsymbol{\phi}^h\|_{L^{\frac{2s}{s-2}}(E_j)} + h|\boldsymbol{u}|_{L^\infty(E_j)}\|\nabla\boldsymbol{\phi}^h\|_{L^2(E_j)}\Big] \le \frac{\kappa\nu}{32}\|\boldsymbol{\phi}^h\|_X^2 + \frac{C}{\nu}\|\boldsymbol{\phi}^h\|_0^2.$$

Now,

$$b(\boldsymbol{\phi}^h, \boldsymbol{c}_1\cdot(\boldsymbol{\phi}^h-\boldsymbol{c}_2)) = -\sum_{j=1}^{N_h}\int_E \boldsymbol{c}_1\cdot(\boldsymbol{\phi}^h-\boldsymbol{c}_2)\nabla\cdot\boldsymbol{\phi}^h$$
$$+\sum_{k=1}^{M_h}\int_{e_k}\{\boldsymbol{c}_1\cdot(\boldsymbol{\phi}^h-\boldsymbol{c}_2)\}[\boldsymbol{\phi}^h]\cdot\boldsymbol{n}_k = S_8+S_9.$$

The first term is bounded by (2.16):

$$S_8 \le C\sum_{j=1}^{N_h}\|\boldsymbol{c}_1\|\|\boldsymbol{\phi}^h-\boldsymbol{c}_2\|_{0,E_j}h^{-1}\|\boldsymbol{\phi}^h\|_{0,E_j}$$
$$\le C\sum_{j=1}^{N_h}\|\boldsymbol{c}_1\|\|\nabla\boldsymbol{\phi}^h\|_{0,E_j}\|\boldsymbol{\phi}^h\|_{0,E_j} \le \frac{\kappa\nu}{64}\|\boldsymbol{\phi}^h\|_X^2 + \frac{C}{\nu}\|\boldsymbol{u}\|^2_{L^\infty([0,T]\times\Omega)}\|\boldsymbol{\phi}^h\|^2_{0,\Omega}.$$

Similarly, the second term is bounded as

$$S_9 \le C\sum_{j=1}^{N_h}\|\boldsymbol{c}_1\|\|\nabla\boldsymbol{\phi}^h\|_{0,E_j}\|\boldsymbol{\phi}_h\|_{0,E_j} \le \frac{\kappa\nu}{64}\|\boldsymbol{\phi}^h\|_X^2 + \frac{C}{\nu}\|\boldsymbol{u}\|^2_{L^\infty([0,T]\times\Omega)}\|\boldsymbol{\phi}^h\|^2_{0,\Omega}.$$

Thus,

$$c(\boldsymbol{\phi}^h,\boldsymbol{u},\boldsymbol{\phi}^h) \le \frac{5\kappa\nu}{64}\|\boldsymbol{\phi}^h\|_X^2 + \frac{C}{\nu}\|\boldsymbol{\phi}^h\|^2_{0,\Omega}.$$

Let us now bound $c(\boldsymbol{\phi}^h,\boldsymbol{\eta},\boldsymbol{\phi}^h)$:

$$c(\boldsymbol{\phi}^h,\boldsymbol{\eta},\boldsymbol{\phi}^h) = \sum_{j=1}^{N_h}\left(\int_{E_j}(\boldsymbol{\phi}^h\cdot\nabla\boldsymbol{\eta})\cdot\boldsymbol{\phi}^h + \int_{\partial E_j^-}|\{\boldsymbol{\phi}^h\}\cdot\boldsymbol{n}_{E_j}|(\boldsymbol{\eta}^{\text{int}}-\boldsymbol{\eta}^{\text{ext}})\cdot\boldsymbol{\phi}^{h,int}\right)$$
$$-\frac{1}{2}b(\boldsymbol{\phi}^h,\boldsymbol{\eta}\cdot\boldsymbol{\phi}^h).$$

The first term is easily bounded:

$$\sum_{j=1}^{N_h}\int_{E_j}(\boldsymbol{\phi}^h\cdot\nabla\boldsymbol{\eta})\cdot\boldsymbol{\phi}^h \le \sum_{j=1}^{N_h}\|\boldsymbol{\phi}^h\|_{0,E_j}\|\boldsymbol{\phi}^h\|_{L^4(E_j)}\|\nabla\boldsymbol{\eta}\|_{L^4(E_j)}$$
$$\le \frac{\kappa\nu}{32}\|\boldsymbol{\phi}^h\|_X^2 + \frac{C}{\nu}\|\boldsymbol{u}\|^2_{L^\infty(0,T;W^{2,4/3}(\Omega))}\|\boldsymbol{\phi}^h\|^2_{0,\Omega}.$$

The second term is bounded using inequalities (2.13), (2.16), (2.6), and (2.8):

$$\sum_{j=1}^{N_h}\int_{\partial E_j^-}|\{\boldsymbol{\phi}^h\}\cdot\boldsymbol{n}_{E_j}|(\boldsymbol{\eta}^{\rm int}-\boldsymbol{\eta}^{\rm ext})\cdot\boldsymbol{\phi}^{h,int}\le C\sum_{j=1}^{N_h}\|\boldsymbol{\phi}^h\|_{L^4(\partial E_j)}\|\boldsymbol{\eta}\|_{L^4(\partial E_j)}\|\boldsymbol{\phi}^h\|_{L^2(\partial E_j)}$$
$$\le C\sum_{j=1}^{N_h}h^{-3/2}h^{r+1}|\boldsymbol{u}|_{r+1,\Omega}\|\boldsymbol{\phi}^h\|_{0,\Omega}^2\le\frac{\kappa\nu}{64}\|\boldsymbol{\phi}^h\|_X^2+C\|\boldsymbol{u}\|_{L^\infty(0,T;H^{r+1}(\Omega))}^2\|\boldsymbol{\phi}^h\|_{0,\Omega}^2.$$

The last term in $c(\boldsymbol{\phi}^h,\boldsymbol{\eta},\boldsymbol{\phi}^h)$ is bounded like the terms S_6, S_7, S_8, and S_9 of $c(\boldsymbol{\phi}^h,\boldsymbol{u},\boldsymbol{\phi}^h)$. The remaining nonlinear terms are bounded in a similar fashion:

$$c_{\boldsymbol{u}^h}(\boldsymbol{\eta},R_h(\boldsymbol{u}),\boldsymbol{\phi}^h)=\sum_{j=1}^{N_h}\int_{E_j}(\boldsymbol{\eta}\cdot\nabla R_h(\boldsymbol{u}))\cdot\boldsymbol{\phi}^h$$
$$+\sum_{j=1}^{N_h}\int_{\partial E_j^-}|\{\boldsymbol{\eta}\}\cdot\boldsymbol{n}_{E_j}|(R_h(\boldsymbol{u})^{\rm int}-R_h(\boldsymbol{u})^{\rm ext})\cdot\boldsymbol{\phi}^{h,{\rm int}}+\frac{1}{2}\sum_{j=1}^{N_h}\int_{E_j}(\nabla\cdot\boldsymbol{\eta})R_h(\boldsymbol{u})\cdot\boldsymbol{\phi}^h$$
$$-\frac{1}{2}\sum_{k=1}^{M_h}\int_{e_k}[\boldsymbol{\eta}]\cdot\boldsymbol{n}_k\{R_h(\boldsymbol{u})\cdot\boldsymbol{\phi}^h\}=S_{10}+\cdots+S_{13}.$$

Using the bound (2.6) and the approximation result (2.7), we have

$$S_{10}\le\|\boldsymbol{\eta}\|_{L^2(\Omega)}\|\nabla R_h(\boldsymbol{u})\|_{L^4(\Omega)}\|\boldsymbol{\phi}^h\|_{L^4(\Omega)}\le\frac{\kappa\nu}{64}\|\boldsymbol{\phi}^h\|_X^2+C\|\boldsymbol{u}\|_{L^\infty([0,T]\times\Omega)}^2h^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2.$$

The inequalities (2.11), (2.14), and (2.6) and the approximation result (2.7) yield

$$S_{11}\le C\sum_{j=1}^{N_h}h_{E_j}^{-1/2}(\|\boldsymbol{\eta}\|_{0,E_j}+h_{E_j}\|\nabla\boldsymbol{\eta}\|_{0,E_j})h_{E_j}^{-1/2}\|\boldsymbol{\phi}^h\|_{0,E_j}$$
$$\le C\|\boldsymbol{\phi}^h\|_{0,\Omega}^2+C\|\boldsymbol{u}\|_{L^\infty([0,T]\times\Omega)}^2h^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2.$$

Similarly, we have

$$S_{12}\le\sum_{j=1}^{N_h}\|\boldsymbol{u}\|_{L^\infty([0,T]\times\Omega)}\|\boldsymbol{\phi}^h\|_{0,E_j}\|\nabla\cdot\boldsymbol{\eta}\|_{0,E_j}$$
$$\le C\|\boldsymbol{\phi}^h\|_{0,\Omega}^2+C\|\boldsymbol{u}\|_{L^\infty([0,T]\times\Omega)}^2h^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2.$$

Note that S_{13} is bounded exactly like S_{11}. The other nonlinear term is bounded using (2.7) and (2.14):

$$c_{\boldsymbol{u}^h}(\boldsymbol{u},\boldsymbol{\eta},\boldsymbol{\phi}^h)=\sum_{j=1}^{N_h}\int_{E_j}(\boldsymbol{u}\cdot\nabla\boldsymbol{\eta})\cdot\boldsymbol{\phi}^h+\sum_{j=1}^{N_h}\int_{\partial E_j^-}|\{\boldsymbol{u}\}\cdot\boldsymbol{n}_{E_j}|(\boldsymbol{\eta}^{\rm int}-\boldsymbol{\eta}^{\rm ext})\cdot\boldsymbol{\phi}^{h,{\rm int}}$$
$$\le C\sum_{j=1}^{N_h}\|\boldsymbol{u}\|_{L^\infty([0,T]\times\Omega)}\|\nabla\boldsymbol{\eta}\|_{0,E_j}\|\boldsymbol{\phi}^h\|_{0,E_j}+C\sum_{j=1}^{N_h}\|\boldsymbol{u}\|_{L^\infty([0,T]\times\Omega)}\|\boldsymbol{\eta}\|_{0,\partial E_j}\|\boldsymbol{\phi}^h\|_{0,\partial E_j}$$
$$\le C\|\boldsymbol{\phi}^h\|_{0,\Omega}^2+C\|\boldsymbol{u}\|_{L^\infty([0,T]\times\Omega)}^2h^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2.$$

Combining all bounds above and using (3.3), we obtain

$$\frac{1}{2}\frac{d}{dt}\|\boldsymbol{\phi}^h\|_0^2 + \frac{\kappa\nu}{2}\|\boldsymbol{\phi}^h\|_X^2 + \frac{\nu_T}{2}\|(I-P_H)\nabla\boldsymbol{\phi}^h\|_0^2 \le C\left(\frac{1}{\nu}+1\right)\|\boldsymbol{\phi}^h\|_0^2$$
$$+ Ch^{2r}\left(\nu + \frac{1}{\nu} + \nu_T\right)|\boldsymbol{u}|_{r+1,\Omega}^2 + C\frac{h^{2r}}{\nu}|p|_{r,\Omega}^2 + Ch^{2r}|\boldsymbol{u}_t|_{r+1,\Omega}^2 + C\nu_T H^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2.$$

Integrating from 0 to t, noting that $\|\boldsymbol{\phi}^h(0)\|_0$ is of the order h^r, and using Gronwall's lemma, yield

$$\|\boldsymbol{\phi}^h(t)\|_0^2 + \kappa\nu\|\boldsymbol{\phi}^h\|_{L^2(0,t;X)}^2 + \nu_T\|(I-P_H)\nabla\boldsymbol{\phi}^h\|_{L^2(0,t;L^2(\Omega))}^2$$
$$\le Ce^{C(1+\nu^{-1})}h^{2r}[(\nu+\nu^{-1}+\nu_T)|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}^2 + \nu^{-1}|p|_{L^2(0,T;H^r(\Omega))}^2$$
$$+|\boldsymbol{u}_t|_{L^2(0,T;H^{r+1}(\Omega))}^2 + \nu_T H^{2r}|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}^2] + Ch^r|\boldsymbol{u}_0|_{r+1,\Omega}^2,$$

where the constant C is independent of ν, ν_T, h, H but depends on $\|\boldsymbol{u}\|_{L^\infty(0,T;W^{2,4/3}(\Omega))}$. The theorem is obtained using the approximation results (2.7) and (2.8) and the following inequality:

$$\|\boldsymbol{u}(t)-\boldsymbol{u}^h(t)\|_0^2 + \kappa\nu\|\boldsymbol{u}(t)-\boldsymbol{u}^h(t)\|_{L^2(0,T;X)}^2 + \nu_T\|(I-P_H)\nabla(\boldsymbol{u}(t)-\boldsymbol{u}^h(t))\|_{L^2(0,T;L^2(\Omega))}^2$$
$$\le \|\boldsymbol{\phi}^h(t)\|_0^2 + \kappa\nu\|\boldsymbol{\phi}^h\|_{L^2(0,T;X)}^2 + \nu_T\|(I-P_H)\nabla\boldsymbol{\phi}^h\|_{L^2(0,T;L^2(\Omega))}^2$$
$$+\|\boldsymbol{\eta}(t)\|_0^2 + \kappa\nu\|\boldsymbol{\eta}\|_{L^2(0,T;X)}^2 + \nu_T\|(I-P_H)\nabla\boldsymbol{\eta}\|_{L^2(0,T;L^2(\Omega))}^2. \qquad \square$$

Remark 2. One of the most important properties of Theorem 4.1 is that the new method improves its robustness with respect to the Reynolds number. In most cases, error estimations of Navier–Stokes equations give a Gronwall constant that depends on the Reynolds number as $1/\nu^3$. In contrast, this approach leads to a better error estimate with a Gronwall constant depending on $1/\nu$. Optimal convergence rates are obtained for Theorem 4.1 if ν_T and H are appropriately chosen.

COROLLARY 4.2. *Assume that $\nu_T = h^\beta$ and $H = h^{1/\alpha}$. If the relation $\beta \ge 2r(\alpha-1)/\alpha$ is satisfied, then the estimate becomes*

$$\|\boldsymbol{u}-\boldsymbol{u}^h\|_{L^\infty(0,T;L^2(\Omega))} + \|\boldsymbol{u}-\boldsymbol{u}^h\|_{L^2(0,T;X)} = \mathcal{O}(h^r).$$

For example, one may choose for a linear approximation the pair $(\nu_T, H) = (h, h^{1/2})$, for quadratic approximation $(\nu_T, H) = (h, h^{3/4})$ or $(\nu_T, H) = (h^2, h^{1/2})$, and for cubic approximation $(\nu_T, H) = (h, h^{5/6})$ or $(\nu_T, H) = (h^2, h^{2/3})$.

THEOREM 4.3. *Under the assumptions of Theorem* 4.1 *and if $a(\cdot,\cdot)$ is symmetric ($\epsilon_0 = -1$), the following estimate holds true:*

$$\|\boldsymbol{u}_t - \boldsymbol{u}_t^h\|_{L^2(0,T;L^2(\Omega))} + \nu^{1/2}\|\boldsymbol{u}-\boldsymbol{u}^h\|_{L^\infty(0,T;X)} \le Ce^{CT\nu^{-1}}[h^r|\boldsymbol{u}_0|_{r+1,\Omega}$$
$$+ h^r|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))} + h^r|\boldsymbol{u}_t|_{L^2(0,T;H^{r+1}(\Omega))} + C\nu_T H^r h^{-1}|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}],$$

where C is a positive constant independent of h, H, ν and ν_T. If $a(\cdot,\cdot)$ is nonsymmetric ($\epsilon_0 = 1$), the estimate is suboptimal, of order h^{r-1}.

Proof. We just give the outline of the proof. We introduce the modified Stokes problem: for any $t > 0$, find $(\boldsymbol{u}^S(t), p^S(t)) \in \boldsymbol{X}^h \times Q^h$ such that

$$\nu(a(\boldsymbol{u}^S(t),\boldsymbol{v}^h) + J(\boldsymbol{u}^S(t),\boldsymbol{v}^h)) + \nu_T g(\boldsymbol{u}^S(t),\boldsymbol{v}^h) + b(\boldsymbol{v}^h, p^S(t))$$
$$= \nu(a(\boldsymbol{u}(t),\boldsymbol{v}^h) + J(\boldsymbol{u}(t),\boldsymbol{v}^h)) + \nu_T g(\boldsymbol{u}(t),\boldsymbol{v}^h) + b(\boldsymbol{v}^h, p(t)) \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h, \tag{4.6}$$
$$b(\boldsymbol{u}^S(t), q^h) = 0 \quad \forall q^h \in Q^h. \tag{4.7}$$

For any $t > 0$, there exists a unique solution to (4.6), (4.7). Furthermore, it is easy to show that the solution satisfies the error estimate

$$\begin{aligned} &\kappa^{1/2}\nu^{1/2}\|\boldsymbol{u}(t) - \boldsymbol{u}^S(t)\|_X + \nu_T^{1/2}\|(I - P_H)\nabla(\boldsymbol{u} - \boldsymbol{u}^S)\|_{0,\Omega} \\ &\le h^r((\nu + \nu^{-1} + \nu_T)^{1/2}|\boldsymbol{u}|_{r+1,\Omega} + \nu^{-1/2}|p|_{r,\Omega} + |\boldsymbol{u}_t|_{r+1,\Omega}) + \nu_T^{1/2}H^r|\boldsymbol{u}|_{r+1,\Omega} \quad \forall t > 0. \end{aligned}$$

Define $\boldsymbol{\eta} = \boldsymbol{u} - \boldsymbol{u}^S$ and $\boldsymbol{\xi} = \boldsymbol{u}^h - \boldsymbol{u}^S$, and choose the test function $\boldsymbol{v}^h = \boldsymbol{\xi}_t$. The resulting error equation is

$$\begin{aligned} &\|\boldsymbol{\xi}_t\|^2_{0,\Omega} + \nu a(\boldsymbol{\xi}, \boldsymbol{\xi}_t) + \frac{\nu}{2}\frac{d}{dt}J(\boldsymbol{\xi}, \boldsymbol{\xi}) + \frac{\nu_T}{2}\frac{d}{dt}g(\boldsymbol{\xi}, \boldsymbol{\xi}) \\ &= (\boldsymbol{\eta}_t, \boldsymbol{\xi}_t) - \nu_T g(\boldsymbol{u}, \boldsymbol{\xi}_t) + c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{\xi}_t) - c(\boldsymbol{u}^h, \boldsymbol{u}^h, \boldsymbol{\xi}_t). \end{aligned} \tag{4.8}$$

The first two terms in the right-hand side of (4.8) are bounded as in Theorem 4.1. A detailed argument is given in [23]. Let us rewrite the nonlinear terms

$$\begin{aligned} c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{\xi}_t) - c(\boldsymbol{u}^h, \boldsymbol{u}^h, \boldsymbol{\xi}_t) &= c(\boldsymbol{\xi}, \boldsymbol{\xi}, \boldsymbol{\xi}_t) - c(\boldsymbol{\xi}, \boldsymbol{\eta}, \boldsymbol{\xi}_t) + c(\boldsymbol{\xi}, \boldsymbol{u}, \boldsymbol{\xi}_t) \\ &\quad - c(\boldsymbol{\eta}, \boldsymbol{u}^h, \boldsymbol{\xi}_t) + c(\boldsymbol{u}, \boldsymbol{\xi}, \boldsymbol{\xi}_t) - c(\boldsymbol{u}, \boldsymbol{\eta}, \boldsymbol{\xi}_t). \end{aligned}$$

We assume that $\boldsymbol{\xi}$ belongs to $L^\infty((0,T) \times \Omega)$. L^p bounds, inverse inequality, and approximation results give the bounds for each nonlinear term as in Theorem 4.1. Collecting all the bounds with (4.8) gives

$$\begin{aligned} &\|\boldsymbol{\xi}_t\|^2_{0,\Omega} + \nu a(\boldsymbol{\xi}, \boldsymbol{\xi}_t) + \frac{\nu}{2}\frac{d}{dt}J(\boldsymbol{\xi}, \boldsymbol{\xi}) + \frac{\nu_T}{2}\frac{d}{dt}g(\boldsymbol{\xi}, \boldsymbol{\xi}) \\ &\le \frac{1}{2}\|\boldsymbol{\xi}_t\|^2_{0,\Omega} + C\|\boldsymbol{\xi}\|^2_X + Ch^{2r}|\boldsymbol{u}|^2_{r+1,\Omega} + Ch^{2r}|\boldsymbol{u}_t|^2_{r+1,\Omega} + C\nu_T^2 H^{2r}h^{-2}|\boldsymbol{u}|^2_{r+1,\Omega}. \end{aligned} \tag{4.9}$$

In the case where the bilinear form a is symmetric ($\epsilon_0 = -1$), the inequality becomes

$$\begin{aligned} &\frac{1}{2}\|\boldsymbol{\xi}_t\|^2_{0,\Omega} + \frac{\nu}{2}\frac{d}{dt}\|\boldsymbol{\xi}\|^2_X + \frac{\nu_T}{2}\frac{d}{dt}g(\boldsymbol{\xi}, \boldsymbol{\xi}) \\ &\le C\|\boldsymbol{\xi}\|^2_X + Ch^{2r}|\boldsymbol{u}|^2_{r+1,\Omega} + Ch^{2r}|\boldsymbol{u}_t|^2_{r+1,\Omega} + C\nu_T^2 H^{2r}h^{-2}|\boldsymbol{u}|^2_{r+1,\Omega}. \end{aligned} \tag{4.10}$$

Integrating from 0 to t and using Gronwall's lemma yield

$$\begin{aligned} &\|\boldsymbol{\xi}_t\|^2_{L^2(0,T;L^2(\Omega))} + \nu\|\boldsymbol{\xi}\|^2_{L^\infty(0,T;X)} + \nu_T \max_{0\le t\le T} g(\boldsymbol{\xi}, \boldsymbol{\xi}) \le Ce^{CT\nu^{-1}}[h^{2r}|\boldsymbol{u}_0|^2_{r+1,\Omega} \\ &+ Ch^{2r}|\boldsymbol{u}|^2_{L^2(0,T;H^{r+1}(\Omega))} + Ch^{2r}|\boldsymbol{u}_t|^2_{L^2(0,T;H^{r+1}(\Omega))} + C\nu_T^2 H^{2r}h^{-2}|\boldsymbol{u}|^2_{L^2(0,T;H^{r+1}(\Omega))}]. \end{aligned}$$

In the case where the bilinear form a is nonsymmetric, we rewrite (4.9) as

$$a(\boldsymbol{\xi}, \boldsymbol{\xi}_t) = \frac{1}{2}\frac{d}{dt}|\!|\!|\nabla\boldsymbol{\xi}|\!|\!|^2_0 - \sum_{k=1}^{M_h}\int_{e_k}\{\nabla\boldsymbol{\xi}\}\boldsymbol{n}_k\cdot[\boldsymbol{\xi}_t] + \sum_{k=1}^{M_h}\int_{e_k}\{\nabla\boldsymbol{\xi}_t\}\boldsymbol{n}_k\cdot[\boldsymbol{\xi}].$$

The bound is then suboptimal: $\mathcal{O}(h^{r-1})$. □

We now derive an error estimate for the pressure.

THEOREM 4.4. *We keep the assumptions of Theorem* 4.1 *and we consider the case where* $a(\cdot,\cdot)$ *is symmetric* ($\epsilon_0 = -1$) *and* $\nu \le 1$. *Then the solution* p^h *satisfies* *the following error estimate:*

$$\begin{aligned}\|p^h - r_h(p)\|_{L^2(0,T;L^2(\Omega))} \le Ce^{CT\nu^{-1}}[\nu h^r|\boldsymbol{u}_0|_{r+1,\Omega}\\ + \nu h^r|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))} + \nu h^r|\boldsymbol{u}_t|_{L^2(0,T;H^{r+1}(\Omega))} + C\nu\nu_T H^r h^{-1}|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}]\\ + C\nu^{1/2}h^r|\boldsymbol{u}_0|_{r+1,\Omega} + C\nu h^r|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))} + C\nu h^r|p|_{L^2(0,T;H^r(\Omega))}\\ + C\nu_T H^r|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}\\ + Ce^{CT(\nu^{-1}+1)}[h^r((\nu+\nu^{-1}+\nu_T)^{1/2}|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))} + \nu^{-1/2}|p|_{L^2(0,T;H^r(\Omega))}\\ + |\boldsymbol{u}_t|_{L^2(0,T;H^{r+1}(\Omega))}) + \nu_T^{1/2}H^r|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}] + Ch^r|\boldsymbol{u}_0|_{r+1,\Omega},\end{aligned}$$

where C is independent of $h, H, \nu,$ and ν_T. Again, if $a(\cdot,\cdot)$ is nonsymmetric ($\epsilon_0 = 1$), the estimate is suboptimal.

Proof. The error equation can be written for all $\boldsymbol{v}^h$ in $\boldsymbol{X}^h$:

$$\begin{aligned}-b(\boldsymbol{v}^h, p^h - r_h(p)) = (\boldsymbol{u}_t^h - \boldsymbol{u}_t, \boldsymbol{v}^h) + \nu a(\boldsymbol{u}^h - \boldsymbol{u}, \boldsymbol{v}^h) + \nu J(\boldsymbol{u}^h - \boldsymbol{u}, \boldsymbol{v}^h)\\ + \nu_T g(\boldsymbol{u}^h - \boldsymbol{u}, \boldsymbol{v}^h) + c(\boldsymbol{u}^h, \boldsymbol{u}^h, \boldsymbol{v}^h) - c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}^h) + \nu_T g(\boldsymbol{u}, \boldsymbol{v}^h) - b(\boldsymbol{v}^h, p - r_h(p)).\end{aligned}$$

From the inf-sup condition (3.10), there is $\boldsymbol{v}^h \in \boldsymbol{X}^h$ such that

$$b(\boldsymbol{v}^h, p^h - r_h(p)) = -\|p^h - r_h(p)\|_0^2, \quad \|\boldsymbol{v}^h\|_X \le \frac{1}{\beta_0}\|p^h - r_h(p)\|_{0,\Omega}.$$

Thus, we have

$$\begin{aligned}\|p^h - r_h(p)\|_{0,\Omega}^2 = (\boldsymbol{u}_t^h - \boldsymbol{u}_t, \boldsymbol{v}^h) + \nu\sum_{j=1}^{N_h}\int_{E_j}\nabla(\boldsymbol{u}^h - \boldsymbol{u}) : \nabla\boldsymbol{v}^h\\ - \nu\sum_{k=1}^{M_h}\int_{e_k}\{\nabla(\boldsymbol{u}^h - \boldsymbol{u})\}\boldsymbol{n}_k\cdot[\boldsymbol{v}^h] + \nu\epsilon_0\sum_{k=1}^{M_h}\int_{e_k}\{\nabla\boldsymbol{v}^h\}\boldsymbol{n}_k\cdot[\boldsymbol{u}^h - \boldsymbol{u}] + \nu J(\boldsymbol{u}^h - \boldsymbol{u}, \boldsymbol{v}^h)\\ + \nu_T g(\boldsymbol{u}^h - \boldsymbol{u}, \boldsymbol{v}^h) + c(\boldsymbol{u}^h, \boldsymbol{u}^h, \boldsymbol{v}^h) - c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}^h) + \nu_T g(\boldsymbol{u}, \boldsymbol{v}^h) - b(\boldsymbol{v}^h, p - r_h(p)).\end{aligned}$$

All the terms above can be handled as in Theorem 4.1. The resulting inequality is

$$\begin{aligned}\|p^h - r_h(p)\|_{0,\Omega}^2 \le C\nu^2\|\boldsymbol{u}_t^h - \boldsymbol{u}_t\|_{0,\Omega}^2 + C\nu^2\|\boldsymbol{u}^h - \boldsymbol{u}\|_X^2 + C\nu^2 h^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2\\ + C\nu^2 h^{2r}|p|_{r,\Omega}^2 + C\nu_T^2 H^{2r}|\boldsymbol{u}|_{r+1,\Omega}^2 + C\nu_T^2 g(\boldsymbol{u}^h - \boldsymbol{u}, \boldsymbol{u}^h - \boldsymbol{u}) + C\|\boldsymbol{u}^h - \boldsymbol{u}\|_{0,\Omega}^2.\end{aligned}$$

We now integrate from 0 to T and use Theorem 4.1 and Theorem 4.3 to conclude. □

5. Fully discrete scheme. In this section, we formulate two fully discrete finite element schemes for the discontinuous eddy viscosity method. Let Δt denote the time step, let $M = T/\Delta t$, and let $0 = t_0 < t_1 < \cdots < t_M = T$ be a subdivision of the interval $(0,T)$. We denote the function ϕ evaluated at the time t_m by ϕ_m and the average of ϕ at two successive time levels by $\phi_{m+\frac{1}{2}} = \frac{1}{2}(\phi_m + \phi_{m+1})$.

Scheme 1: Given $\boldsymbol{u}_0^h$, find $(\boldsymbol{u}_m^h)_{m\ge1}$ in $\boldsymbol{X}^h$ and $(p_m^h)_{m\ge1}$ in Q^h such that

$$\begin{aligned}\frac{1}{\Delta t}(\boldsymbol{u}_{m+1}^h - \boldsymbol{u}_m^h, \boldsymbol{v}^h) + \nu(a(\boldsymbol{u}_{m+1}^h, \boldsymbol{v}^h) + J(\boldsymbol{u}_{m+1}^h, \boldsymbol{v}^h)) + c(\boldsymbol{u}_m^h, \boldsymbol{u}_{m+1}^h, \boldsymbol{v}^h)\\ + \nu_T g(\boldsymbol{u}_{m+1}^h, \boldsymbol{v}^h) + b(\boldsymbol{v}^h, p_{m+1}^h) = (\boldsymbol{f}_{m+1}, \boldsymbol{v}^h) \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h,\end{aligned} \tag{5.1}$$

$$b(\boldsymbol{u}_{m+1}^h, q^h) = 0 \quad \forall q^h \in Q^h. \tag{5.2}$$

Scheme 2: Given $\tilde{\boldsymbol{u}}_0^h, \tilde{\boldsymbol{u}}_1^h, \tilde{p}_1^h$, find $(\tilde{\boldsymbol{u}}_m^h)_{m\geq 2}$ in $\boldsymbol{X}^h$ and $(\tilde{p}_m^h)_{m\geq 2}$ in Q^h such that

$$\begin{aligned}&\frac{1}{\Delta t}(\tilde{\boldsymbol{u}}_{m+1}^h - \tilde{\boldsymbol{u}}_m^h, \boldsymbol{v}^h) + \nu(a(\tilde{\boldsymbol{u}}_{m+\frac{1}{2}}^h, \boldsymbol{v}^h) + J(\tilde{\boldsymbol{u}}_{m+\frac{1}{2}}^h, \boldsymbol{v}^h)) + c(\tilde{\boldsymbol{u}}_{m+\frac{1}{2}}^h, \tilde{\boldsymbol{u}}_{m+\frac{1}{2}}^h, \boldsymbol{v}^h)\\ &\quad + \nu_T g(\tilde{\boldsymbol{u}}_{m+\frac{1}{2}}^h, \boldsymbol{v}^h) + b(\boldsymbol{v}^h, \tilde{p}_{m+\frac{1}{2}}^h) = (\boldsymbol{f}_{m+\frac{1}{2}}, \boldsymbol{v}^h) \quad \forall \boldsymbol{v}^h \in \boldsymbol{X}^h,\end{aligned} \tag{5.3}$$

$$b(\tilde{\boldsymbol{u}}_{m+1}^h, q^h) = 0 \quad \forall q^h \in Q^h. \tag{5.4}$$

For both schemes, the initial velocity is defined to be the L^2 projection of $\boldsymbol{u}_0$. Scheme 1 is based on a backward Euler discretization. Scheme 2 is based on a Crank–Nicolson discretization, and requires the velocity and pressure at the first step. The approximations $\tilde{\boldsymbol{u}}_1^h$ and $\tilde{p}_1^h$ can be obtained by a first order scheme (see [2]). We will show that Scheme 1 is first order in time and Scheme 2 is second order in time. First, we prove the stability of the schemes.

Lemma 5.1. *The solution* $(\boldsymbol{u}_m^h)_m$ *of* (5.1), (5.2) *remains bounded in the following sense:*

$$\|\boldsymbol{u}_m^h\|_{0,\Omega}^2 \leq K, \quad m = 0, \dots, M,$$
$$\Delta t \sum_{m=0}^{M-1} \|\boldsymbol{u}_{m+1}^h\|_X^2 \leq \frac{K}{2\nu}, \quad \Delta t \sum_{m=0}^{M-1} |||(I - P_H)\nabla \boldsymbol{u}_{m+1}^h|||_0^2 \leq \frac{K}{2\nu_T},$$

where $K = \|\boldsymbol{u}_0\|_{0,\Omega}^2 + \|\boldsymbol{f}\|_{L^2([0,T]\times\Omega)}^2$.

The solution $(\tilde{\boldsymbol{u}}_m^h)_m$ *of* (5.3), (5.4) *remains bounded in the following sense:*

$$\|\tilde{\boldsymbol{u}}_m^h\|_{0,\Omega}^2 \leq \tilde{K}, \quad m = 0, \dots, M,$$
$$\Delta t \sum_{m=0}^{M-1} \|\tilde{\boldsymbol{u}}_{m+1}^h\|_X^2 \leq \frac{\tilde{K}}{2\nu}, \quad \Delta t \sum_{m=0}^{M-1} \|(I - P_H)\nabla \tilde{\boldsymbol{u}}_{m+1}^h\|_{0,\Omega}^2 \leq \frac{\tilde{K}}{2\nu_T},$$

where $\tilde{K} = \|\boldsymbol{u}_0\|_{0,\Omega}^2 + 2\|\boldsymbol{f}\|_{L^2([0,T]\times\Omega)}^2$.

Proof. Choose $\boldsymbol{v}^h = \boldsymbol{u}_{m+1}^h$ in (5.1) and $q^h = p_{m+1}^h$ in (5.2). We multiply by $2\Delta t$ and sum over m. Then, from the positivity of c and (3.3), we have

$$\begin{aligned}&\|\boldsymbol{u}_m^h\|_{0,\Omega}^2 - \|\boldsymbol{u}_0^h\|_{0,\Omega}^2 + 2\kappa\nu\Delta t \sum_{j=0}^{m-1} \|\boldsymbol{u}_{j+1}^h\|_X^2 + 2\nu_T \Delta t \sum_{j=0}^{m-1} |||(I - P_H)\nabla \boldsymbol{u}_{j+1}^h|||_0^2\\ &\qquad \leq \Delta t \sum_{j=0}^{m-1} \|\boldsymbol{f}_{j+1}\|_{0,\Omega}^2 + \Delta t \sum_{j=0}^{m-1} \|\boldsymbol{u}_{j+1}^h\|_{0,\Omega}^2.\end{aligned}$$

The result is obtained by using a discrete version of Gronwall's lemma [15] and the fact that $\|\boldsymbol{u}_0^h\|_{0,\Omega} \leq \|\boldsymbol{u}_0\|_{0,\Omega}$.

For Scheme 2, the proof is similar. Choose $\boldsymbol{v}^h = \tilde{\boldsymbol{u}}_{m+\frac{1}{2}}$ in (5.3) and $q^h = \tilde{p}_{m+\frac{1}{2}}^h$ in (5.4). The rest of the proof follows as above. See [23] for more details. ∎

THEOREM 5.2. *Under the assumptions of Theorem* 4.1 *and if* $\boldsymbol{u}_t$ *and* $\boldsymbol{u}_{tt}$ *belong to* $L^\infty(0,T;L^2(\Omega))$*, there is a constant* C *independent of* h, H, ν*, and* ν_T *such that*

$$\max_{m=0,\dots,M}\|\boldsymbol{u}_m-\boldsymbol{u}_m^h\|_{0,\Omega}+\left(\nu\kappa\Delta t\sum_{m=0}^{M-1}\|\boldsymbol{u}_{m+1}-\boldsymbol{u}_{m+1}^h\|_X^2\right)^{1/2}$$
$$+\left(\nu_T\Delta t\sum_{m=0}^{M}|\!|\!|(I-P_H)(\nabla\boldsymbol{u}_{m+1}-\boldsymbol{u}_{m+1}^h)|\!|\!|_0^2\right)^{1/2}\leq Ch^r|\boldsymbol{u}_0|_{r+1,\Omega}$$
$$+Ce^{CT\nu^{-1}}[h^r(\nu+\nu^{-1}+\nu_T)^{1/2}|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}+\nu_T^{1/2}H^r|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}$$
$$+\nu^{-1/2}\Delta t(\|\boldsymbol{u}_t\|_{L^\infty(0,T;L^2(\Omega))}+\|\boldsymbol{u}_{tt}\|_{L^\infty(0,T;L^2(\Omega))})+h^r\nu^{-1/2}|p|_{L^2(0,T;H^r(\Omega))}].$$

Proof. As in the continuous case, we set $\boldsymbol{e}_m=\boldsymbol{u}_m-\boldsymbol{u}_m^h$. We subtract from (5.1) and (5.2) equations (3.14) and (3.15) evaluated at time $t=t_{m+1}$.

$$(\boldsymbol{u}_t(t_{m+1}),\boldsymbol{v}^h)-\frac{1}{\Delta t}(\boldsymbol{u}_{m+1}^h-\boldsymbol{u}_m^h,\boldsymbol{v}^h)+\nu[a(\boldsymbol{e}_{m+1},\boldsymbol{v}^h)+J(\boldsymbol{e}_{m+1},\boldsymbol{v}^h)]$$
$$+\nu_Tg(\boldsymbol{e}_{m+1},\boldsymbol{v}^h)+c(\boldsymbol{u}_{m+1},\boldsymbol{u}_{m+1},\boldsymbol{v}^h)-c(\boldsymbol{u}_m^h,\boldsymbol{u}_{m+1}^h,\boldsymbol{v}^h)$$
$$+b(\boldsymbol{v}^h,p_{m+1}-p_{m+1}^h)=\nu_Tg(\boldsymbol{u}_{m+1},\boldsymbol{v}^h)\quad\forall\boldsymbol{v}^h\in\boldsymbol{X}^h,\tag{5.5}$$
$$b(\boldsymbol{e}_{m+1},q^h)=0\quad\forall q^h\in Q^h.\tag{5.6}$$

Define $\boldsymbol{\phi}_m=\boldsymbol{u}_m^h-(R_h(\boldsymbol{u}))_m$, $\boldsymbol{\eta}_m=\boldsymbol{u}_m-(R_h(\boldsymbol{u}))_m$. Choose $\boldsymbol{v}^h=\boldsymbol{\phi}_{m+1}$ in (5.5) and $q^h=p_{m+1}^h$ in (5.6). Adding and subtracting the interpolant and using (3.3) yield the following error equation:

$$\frac{1}{2\Delta t}(\|\boldsymbol{\phi}_{m+1}\|_{0,\Omega}^2-\|\boldsymbol{\phi}_m\|_{0,\Omega}^2)+\nu\kappa\|\boldsymbol{\phi}_{m+1}\|_X^2+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+1}|\!|\!|_0^2$$
$$+c(\boldsymbol{u}_m^h,\boldsymbol{u}_{m+1}^h,\boldsymbol{\phi}_{m+1})-c(\boldsymbol{u}_{m+1},\boldsymbol{u}_{m+1},\boldsymbol{\phi}_{m+1})+b(\boldsymbol{\phi}_{m+1},p_{m+1}^h-p_{m+1})$$
$$\leq\left\|\frac{\partial\boldsymbol{u}}{\partial t}(t_{m+1})-\frac{1}{\Delta t}(\boldsymbol{u}_{m+1}-\boldsymbol{u}_m)\right\|_{0,\Omega}\|\boldsymbol{\phi}_{m+1}\|_{0,\Omega}+\frac{1}{\Delta t}\left\|\boldsymbol{\eta}_{m+1}-\boldsymbol{\eta}_m\right\|_{0,\Omega}\|\boldsymbol{\phi}_{m+1}\|_{0,\Omega}$$
$$+\nu|a(\boldsymbol{\eta}_{m+1},\boldsymbol{\phi}_{m+1})+J(\boldsymbol{\eta}_{m+1},\boldsymbol{\phi}_{m+1})|+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{\eta}_{m+1}|\!|\!|_0|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+1}|\!|\!|_0$$
$$+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{u}_{m+1}|\!|\!|_0|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+1}|\!|\!|_0.$$

We rewrite the nonlinear terms

$$c_{\boldsymbol{u}_m^h}(\boldsymbol{u}_m^h,\boldsymbol{u}_{m+1}^h,\boldsymbol{\phi}_{m+1})-c_{\boldsymbol{u}_{m+1}}(\boldsymbol{u}_{m+1},\boldsymbol{u}_{m+1},\boldsymbol{\phi}_{m+1})$$
$$=c_{\boldsymbol{u}_m^h}(\boldsymbol{u}_m^h,\boldsymbol{u}_{m+1}^h,\boldsymbol{\phi}_{m+1})-c_{\boldsymbol{u}_m^h}(\boldsymbol{u}_{m+1},\boldsymbol{u}_{m+1},\boldsymbol{\phi}_{m+1}).$$

We now drop the subscript $\boldsymbol{u}_m^h$:

$$c_{\boldsymbol{u}_m^h}(\boldsymbol{u}_m^h,\boldsymbol{u}_{m+1}^h,\boldsymbol{\phi}_{m+1})-c_{\boldsymbol{u}_m^h}(\boldsymbol{u}_{m+1},\boldsymbol{u}_{m+1},\boldsymbol{\phi}_{m+1})$$
$$=c(\boldsymbol{u}_m^h,\boldsymbol{\phi}_{m+1},\boldsymbol{\phi}_{m+1})-c(\boldsymbol{\phi}_m,\boldsymbol{\eta}_{m+1},\boldsymbol{\phi}_{m+1})+c(\boldsymbol{\phi}_m,\boldsymbol{u}_{m+1},\boldsymbol{\phi}_{m+1})$$
$$-c(\boldsymbol{\eta}_m,\boldsymbol{u}_{m+1}^I,\boldsymbol{\phi}_{m+1})-c(\boldsymbol{u}_m,\boldsymbol{\eta}_{m+1},\boldsymbol{\phi}_{m+1})-c(\boldsymbol{u}_{m+1}-\boldsymbol{u}_m,\boldsymbol{u}_{m+1},\boldsymbol{\phi}_{m+1}).$$

Thus, we rewrite the error equation as

$$\begin{aligned}&\frac{1}{2\Delta t}(\|\boldsymbol{\phi}_{m+1}\|^2_{0,\Omega}-\|\boldsymbol{\phi}_m\|^2_{0,\Omega})+\nu\kappa\|\boldsymbol{\phi}_{m+1}\|^2_X+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+1}|\!|\!|^2_0\\&+c(\boldsymbol{u}^h_m,\boldsymbol{\phi}_{m+1},\boldsymbol{\phi}_{m+1})\leq|c(\boldsymbol{\phi}_m,\boldsymbol{\eta}_{m+1},\boldsymbol{\phi}_{m+1})|+|c(\boldsymbol{\phi}_m,\boldsymbol{u}_{m+1},\boldsymbol{\phi}_{m+1})|\\&+|c(\boldsymbol{\eta}_m,\boldsymbol{u}^I_{m+1},\boldsymbol{\phi}_{m+1})|+|c(\boldsymbol{u}_m,\boldsymbol{\eta}_{m+1},\boldsymbol{\phi}_{m+1})|+|c(\boldsymbol{u}_{m+1}-\boldsymbol{u}_m,\boldsymbol{u}_{m+1},\boldsymbol{\phi}_{m+1})|\\&+|b(\boldsymbol{\phi}_{m+1},p^h_{m+1}-p_{m+1})|+\left\|\frac{\partial\boldsymbol{u}}{\partial t}(t_{m+1})-\frac{1}{\Delta t}(\boldsymbol{u}_{m+1}-\boldsymbol{u}_m)\right\|_{0,\Omega}\|\boldsymbol{\phi}_{m+1}\|_{0,\Omega}\\&+\frac{1}{\Delta t}\|\boldsymbol{\eta}_{m+1}-\boldsymbol{\eta}_m\|_{0,\Omega}\|\boldsymbol{\phi}_{m+1}\|_{0,\Omega}+\nu|a(\boldsymbol{\eta}_{m+1},\boldsymbol{\phi}_{m+1})+J(\boldsymbol{\eta}_{m+1},\boldsymbol{\phi}_{m+1})|\\&+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{\eta}_{m+1}|\!|\!|_0|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+1}|\!|\!|_0\\&+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{u}_{m+1}|\!|\!|_0|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+1}|\!|\!|_0\leq|T_0|+\cdots+|T_{10}|.\end{aligned}$$

We want to bound the terms $T_0,T_2,\ldots,T_{10}$. T_0 can be handled as in Theorem 4.1. Then, T_0 is bounded as

$$T_0\leq\frac{\kappa\nu}{6}\|\boldsymbol{\phi}_{m+1}\|^2_X+C\nu^{-1}(\|\boldsymbol{u}\|^2_{L^\infty(0,T;H^{r+1}(\Omega))}+\|\boldsymbol{u}\|^2_{L^\infty(0,T;W^{2,4/3}(\Omega))})\|\boldsymbol{\phi}_m\|^2_{0,\Omega}.$$

Also, the term T_1 is bounded exactly like the term (4.5) in the proof of Theorem 4.1. Here, the constant vectors are

$$\boldsymbol{c}_1=\frac{1}{|E_j|}\int_{E_j}\boldsymbol{u}_{m+1},\quad\boldsymbol{c}_2=\frac{1}{|E_j|}\int_{E_j}\boldsymbol{\phi}_{m+1}.$$

Then, T_1 can be rewritten as

$$\begin{aligned}T_1=&\sum_{j=1}^{N_h}\int_{E_j}(\boldsymbol{\phi}_m\cdot\nabla\boldsymbol{u}_{m+1})\cdot\boldsymbol{\phi}_{m+1}-\frac{1}{2}b(\boldsymbol{\phi}_m,(\boldsymbol{u}_{m+1}-\boldsymbol{c}_1)\cdot\boldsymbol{\phi}_{m+1})\\&-\frac{1}{2}b(\boldsymbol{\phi}_m,\boldsymbol{c}_1\cdot(\boldsymbol{\phi}_{m+1}-\boldsymbol{c}_2))\leq\frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|^2_X+C\nu^{-1}\|\boldsymbol{\phi}_m\|^2_{0,\Omega}.\end{aligned}$$

Expanding T_2, we obtain

$$\begin{aligned}T_2=&\sum_{j=1}^{N_h}\int_{E_j}(\boldsymbol{\eta}_m\cdot\nabla\boldsymbol{u}^I_{m+1})\cdot\boldsymbol{\phi}_{m+1}+\sum_{j=1}^{N_h}\int_{\partial E_j^-}|\{\boldsymbol{\eta}_m\}\cdot\boldsymbol{n}_{E_j}|(\boldsymbol{u}^{I,\text{int}}_{m+1}-\boldsymbol{u}^{I,\text{ext}}_{m+1})\cdot\boldsymbol{\phi}^{int}_{m+1}\\&+\frac{1}{2}\sum_{j=1}^{N_h}\int_{E_j}(\nabla\cdot\boldsymbol{\eta}_m)\boldsymbol{u}^I_{m+1}\cdot\boldsymbol{\phi}_{m+1}-\frac{1}{2}\sum_{k=1}^{P_h}\int_{e_k}[\boldsymbol{\eta}_m]\cdot\boldsymbol{n}_k\{\boldsymbol{u}^I_{m+1}\cdot\boldsymbol{\phi}_{m+1}\}\\&=T_{21}+\cdots+T_{24}.\end{aligned}$$

The bound for T_{21} is obtained using (2.6) and (2.8):

$$\begin{aligned}T_{21}&\leq\|\boldsymbol{\eta}_m\|_{0,\Omega}\|\nabla\boldsymbol{u}^I_{m+1}\|_{L^4(\Omega)}\|\boldsymbol{\phi}_{m+1}\|_{L^4(\Omega)}\\&\leq\frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|^2_X+C\nu^{-1}h^{2r}\|\boldsymbol{u}\|^2_{L^\infty(0,T;W^{2,4/3}(\Omega))}|\boldsymbol{u}_m|^2_{r+1,\Omega}.\end{aligned}$$

Similarly for the term T_{22}, the inequalities (2.7) and (2.14) give

$$\begin{aligned}T_{22}&\leq C\sum_{j=1}^{N_h}\|\boldsymbol{\eta}_m\|_{L^2(\partial E_j)}\|\boldsymbol{u}^I_{m+1}\|_{L^\infty(\Omega)}\|\boldsymbol{\phi}_{m+1}\|_{L^2(\partial E_j)}\\&\leq\frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|^2_X+C\nu^{-1}h^{2r}\|\boldsymbol{u}\|^2_{L^\infty([0,T]\times\Omega)}|\boldsymbol{u}_m|^2_{r+1,\Omega}.\end{aligned}$$

The estimate of T_{23} is obtained by using a bound on interpolant, the Cauchy–Schwarz inequality, the approximation result (2.7), Young's inequality, and L^p bound (2.6):

$$T_{23} \leq \frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|_X^2 + C\nu^{-1}h^{2r}\|\boldsymbol{u}\|^2_{L^\infty([0,T]\times\Omega)}|\boldsymbol{u}_m|^2_{r+1,\Omega}.$$

The term T_{24} is bounded exactly as for T_{22}. Because of the regularity of $\boldsymbol{u}$ and the approximation result (2.7), we can bound T_3:

$$\begin{aligned} T_3 &\leq C\|\boldsymbol{u}_m\|_{L^\infty(\Omega)}h^r|\boldsymbol{u}_{m+1}|_{r+1,\Omega}\|\boldsymbol{\phi}_{m+1}\|_{0,\Omega} \\ &\leq \frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|_X^2 + C\nu^{-1}h^{2r}\|\boldsymbol{u}\|^2_{L^\infty([0,T]\times\Omega)}|\boldsymbol{u}_m|^2_{r+1,\Omega}. \end{aligned}$$

The term T_4 is bounded using the estimate (2.6):

$$\begin{aligned} T_4 &\leq \Delta t\|\boldsymbol{u}_t\|_{L^\infty(t_m,t_{m+1};L^2(\Omega))}\|\nabla\boldsymbol{u}_{m+1}\|_{L^4(\Omega)}\|\boldsymbol{\phi}_{m+1}\|_{L^4(\Omega)} \\ &\leq \frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|_X^2 + C\nu^{-1}\Delta t^2\|\boldsymbol{u}_t\|^2_{L^\infty(t_m,t_{m+1};L^2(\Omega))}\|\boldsymbol{u}\|^2_{L^\infty(0,T;W^{2,4/3}(\Omega))}. \end{aligned}$$

By property of the interpolant (3.11) and properties of $r_h(p)$, (2.9), and (2.10), we now bound T_5:

$$\begin{aligned} T_5 &= b(\boldsymbol{\phi}_{m+1}, p^h_{m+1} - (r_h(p))_{m+1}) - b(\boldsymbol{\phi}_{m+1}, p_{m+1} - (r_h(p))_{m+1}) \\ &= -b(\boldsymbol{\phi}_{m+1}, p_{m+1} - (r_h(p))_{m+1}) = \sum_{k=1}^{M_h}\int_{e_k}\{p_{m+1} - (r_h(p))_{m+1}\}[\boldsymbol{\phi}_{m+1}]\cdot\boldsymbol{n}_k \\ &\leq \sum_{k=1}^{M_h}\|[\boldsymbol{\phi}_{m+1}]\|_{0,e_k}|e_k|^{1/2-1/2}\|p_{m+1}\|_{0,e_k} \leq \frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|_X^2 + C\nu^{-1}h^{2r}|p_{m+1}|^2_{r,\Omega}. \end{aligned}$$

From a Taylor expansion, we have

$$T_6 \leq C\Delta t\|\boldsymbol{\phi}_{m+1}\|_X\|\boldsymbol{u}_{tt}(t^*)\|_{0,\Omega} \leq \frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|_X^2 + C\nu^{-1}\Delta t^2\|\boldsymbol{u}_{Tm}\|^2_{L^\infty(0,T;L^2(\Omega))}.$$

To bound T_7, we assume that $h \leq \Delta t$ and we use (2.8) and (2.6):

$$\begin{aligned} T_7 &\leq \frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|_X^2 + C\nu^{-1}\frac{h^{2r+2}}{\Delta t^2}(|\boldsymbol{u}_{m+1}|^2_{r+1,\Omega} + |\boldsymbol{u}_m|^2_{r+1,\Omega}) \\ &\leq \frac{\kappa\nu}{24}\|\boldsymbol{\phi}_{m+1}\|_X^2 + C\nu^{-1}h^{2r}(|\boldsymbol{u}_{m+1}|^2_{r+1,\Omega} + |\boldsymbol{u}_m|^2_{r+1,\Omega}). \end{aligned}$$

The terms T_8, T_9, and T_{10} are exactly bounded as in Theorem 4.1. (See [23] for details.) Combining all the bounds of the terms $T_0, \ldots, T_{10}$, multiplying by $2\Delta t$, and summing over m, we obtain

$$\begin{aligned} &\|\boldsymbol{\phi}_{m+1}\|^2_{0,\Omega} - \|\boldsymbol{\phi}_0\|^2_{0,\Omega} + \nu\kappa\Delta t\sum_{i=0}^m\|\boldsymbol{\phi}_{i+1}\|_X^2 + \nu_T\Delta t\sum_{i=0}^m\|(I - P_H)\nabla\boldsymbol{\phi}_{i+1}\|_0^2 \\ &\leq Ce^{CT\nu^{-1}}[h^{2r}(\nu + \nu^{-1} + \nu_T)|\boldsymbol{u}|^2_{L^2(0,T;H^{r+1}(\Omega))} + \nu_T H^{2r}|\boldsymbol{u}|^2_{L^2(0,T;H^{r+1}(\Omega))} \\ &+ \nu^{-1}\Delta t^2(\|\boldsymbol{u}_t\|^2_{L^\infty(0,T;L^2(\Omega))} + \|\boldsymbol{u}_{tt}\|^2_{L^\infty(0,T;L^2(\Omega))}) + h^{2r}\nu^{-1}|p|^2_{L^2(0,T;H^r(\Omega))}]. \end{aligned}$$

The final result is obtained by noting that $\|\boldsymbol{\phi}_0\|_{0,\Omega}$ is of order h^r and by using approximation results and a triangle inequality. ∎

THEOREM 5.3. *Assume that* $\boldsymbol{u}_{tt}\in L^\infty(0,T;(H^1(\Omega))^2)$, $p_{tt}\in L^\infty(0,T;H^1(\Omega))$, $\boldsymbol{u}_{ttt}\in L^\infty(0,T;(H^2(\Omega))^2)$, *and* $\boldsymbol{f}_{tt}\in L^\infty(0,T;(L^2(\Omega))^2)$. *Under the assumptions of Theorem* 4.1, *there is a constant* C *independent of* h,H,ν, *and* ν_T *such that*

$$\begin{aligned}&\max_{m=0,\ldots,M}\|\boldsymbol{u}_m-\tilde{\boldsymbol{u}}_m\|_{0,\Omega}+\left(\nu\kappa\Delta t\sum_{m=0}^{M-1}\|\boldsymbol{u}_{m+1}-\tilde{\boldsymbol{u}}_{m+1}\|_X^2\right)^{1/2}\\&+\left(\nu_T\Delta t\sum_{m=0}^{M-1}|\!|\!|(I-P_H)\nabla\boldsymbol{u}_{m+1}-\tilde{\boldsymbol{u}}_{m+1}|\!|\!|_0^2\right)^{1/2}\le Ce^{CT\nu^{-1}}[h^r\nu^{-1/2}\|p\|_{L^2(0,T;H^r(\Omega))}\\&\quad+h^r(\nu+\nu^{-1}+\nu_T)^{1/2}\|\boldsymbol{u}\|_{L^2(0,T;H^{r+1}(\Omega))}+\Delta t^2\nu^{1/2}\|\boldsymbol{u}_{ttt}\|_{L^\infty(0,T;H^2(\Omega))}\\&\quad+\Delta t^2\nu^{-1/2}(\|\boldsymbol{u}_{tt}\|_{L^\infty(0,T;H^1(\Omega))}+\|p_{tt}\|_{L^\infty(0,T;H^1(\Omega))}+\|\boldsymbol{u}_{ttt}\|_{L^\infty(0,T;L^2(\Omega))}\\&\quad+\|\boldsymbol{f}_{tt}\|_{L^\infty(0,T;L^2(\Omega))})+\nu_T^{1/2}H^r|\boldsymbol{u}|_{L^2(0,T;H^{r+1}(\Omega))}]+Ch^r|\boldsymbol{u}_0|_{r+1,\Omega}.\end{aligned}$$

Proof. The proof is derived in a similar fashion as for the backward Euler scheme. Using the same notation, the error equation is obtained by subtracting (3.6) evaluated at the time $t=t_{m+1/2}$ from (5.3) and adding and subtracting the interpolant $(R_h(\boldsymbol{u}))_{m+1/2}$. After some manipulation, we obtain

$$\begin{aligned}&\frac{1}{2\Delta t}(\|\boldsymbol{\phi}_{m+1}\|_{0,\Omega}^2-\|\boldsymbol{\phi}_m\|_{0,\Omega}^2)+\nu\kappa\|\boldsymbol{\phi}_{m+\frac12}\|_X^2+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+\frac12}|\!|\!|_0^2\\&+c(\tilde{\boldsymbol{u}}^h_{m+\frac12},\boldsymbol{\phi}^h_{m+\frac12},\boldsymbol{\phi}_{m+\frac12})\le|c(\boldsymbol{\phi}_{m+\frac12},\boldsymbol{\eta}_{m+\frac12},\boldsymbol{\phi}_{m+\frac12})|+|c(\boldsymbol{\phi}_{m+\frac12},\boldsymbol{u}_{m+\frac12},\boldsymbol{\phi}_{m+\frac12})|\\&\qquad+|c(\boldsymbol{\eta}_{m+\frac12},\boldsymbol{u}^I_{m+\frac12},\boldsymbol{\phi}_{m+\frac12})|+|c(\boldsymbol{u}_{m+\frac12},\boldsymbol{\eta}_{m+\frac12},\boldsymbol{\phi}_{m+\frac12})|\\&+|c(\boldsymbol{u}_{m+\frac12}-\boldsymbol{u}(t_{m+\frac12}),\boldsymbol{u}_{m+\frac12},\boldsymbol{\phi}_{m+\frac12})|+|c(\boldsymbol{u}(t_{m+\frac12}),\boldsymbol{u}_{m+\frac12}-\boldsymbol{u}(t_{m+\frac12}),\boldsymbol{\phi}_{m+\frac12})|\\&+|b(\boldsymbol{\phi}_{m+\frac12},\tilde{p}^h_{m+\frac12}-p(t_{m+\frac12}))|+\left\|\boldsymbol{u}_t(t_{m+\frac12})-\frac{1}{\Delta t}(\boldsymbol{u}_{m+1}-\boldsymbol{u}_m)\right\|_{0,\Omega}\|\boldsymbol{\phi}_{m+\frac12}\|_{0,\Omega}\\&+\frac{1}{\Delta t}\|\boldsymbol{\eta}_{m+1}-\boldsymbol{\eta}_m\|_{0,\Omega}\|\boldsymbol{\phi}_{m+\frac12}\|_{0,\Omega}+\|\boldsymbol{f}_{m+\frac12}-\boldsymbol{f}(t_{m+\frac12})\|_{0,\Omega}\|\boldsymbol{\phi}_{m+\frac12}\|_{0,\Omega}\\&\quad+\nu|a(\boldsymbol{u}(t_{m+\frac12})-\boldsymbol{u}^I_{m+\frac12},\boldsymbol{\phi}_{m+1})+J(\boldsymbol{u}(t_{m+\frac12})-\boldsymbol{u}^I_{m+\frac12},\boldsymbol{\phi}_{m+1})|\\&\qquad+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{\eta}_{m+\frac12}|\!|\!|_0|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+\frac12}|\!|\!|_0\\&\quad+\nu_T|\!|\!|(I-P_H)\nabla\boldsymbol{u}_{m+\frac12}|\!|\!|_0|\!|\!|(I-P_H)\nabla\boldsymbol{\phi}_{m+\frac12}|\!|\!|_0\le A_0+\cdots+A_{13}.\end{aligned}$$

The terms $A_0,A_1,A_2,A_3,A_8,A_{11}$, and A_{12} are bounded exactly like the terms T_0,T_1,T_2, T_3,T_7,T_9, and T_{10}, respectively. From a Taylor expansion, we bound the terms A_4 and A_5:

$$\begin{aligned}A_4+A_5&=\sum_{j=1}^{N_h}\int_{E_j}((\boldsymbol{u}_{m+\frac12}-\boldsymbol{u}(t_{m+\frac12}))\cdot\nabla\boldsymbol{u}_{m+\frac12})\cdot\boldsymbol{\phi}_{m+\frac12}\\&\quad+\sum_{j=1}^{N_h}\int_{E_j}\boldsymbol{u}(t_{m+\frac12})\cdot\nabla(\boldsymbol{u}_{m+\frac12}-\boldsymbol{u}(t_{m+\frac12}))\cdot\boldsymbol{\phi}_{m+\frac12}\\&=\frac{\Delta t^2}{8}\sum_{j=1}^{N_h}\int_{E_j}(\boldsymbol{u}_{tt}(t^*)\cdot\nabla\boldsymbol{u}_{m+\frac12})\cdot\boldsymbol{\phi}_{m+\frac12}+\frac{\Delta t^2}{8}\sum_{j=1}^{N_h}\int_{E_j}\boldsymbol{u}(t_{m+\frac12})\cdot\nabla(\boldsymbol{u}_{tt}(t^*))\cdot\boldsymbol{\phi}_{m+\frac12}\\&\le\frac{\kappa\nu}{64}\|\boldsymbol{\phi}_{m+\frac12}\|_X^2+C\nu^{-1}\Delta t^4\|\boldsymbol{u}_{tt}\|^2_{L^\infty(0,T;H^1(\Omega))}\|\boldsymbol{u}\|^2_{L^\infty(0,T;W^{2,4/3}(\Omega))}.\end{aligned}$$

With (3.7), (3.11), and (5.4), the pressure term can be rewritten as

$$\begin{aligned}
A_6 &= b(\boldsymbol{\phi}_{m+\frac{1}{2}}, \tilde{p}^h_{m+\frac{1}{2}} - p_{m+\frac{1}{2}}) + b(\boldsymbol{\phi}_{m+\frac{1}{2}}, p_{m+\frac{1}{2}} - p(t_{m+\frac{1}{2}}))\\
&= -\, b(\boldsymbol{\phi}_{m+\frac{1}{2}}, p_{m+\frac{1}{2}} - (r_h(p))_{m+\frac{1}{2}}) + b(\boldsymbol{\phi}_{m+\frac{1}{2}}, p_{m+\frac{1}{2}} - p(t_{m+\frac{1}{2}}))\\
&= \sum_{k=1}^{M_h} \int_{e_k} \{p_{m+\frac{1}{2}} - (r_h(p))_{m+\frac{1}{2}}\}[\boldsymbol{\phi}_{m+\frac{1}{2}}] \cdot \boldsymbol{n}_k - \sum_{j=1}^{N_h} \int_{E_j} (p_{m+\frac{1}{2}} - p(t_{m+\frac{1}{2}}))\nabla \cdot \boldsymbol{\phi}_{m+\frac{1}{2}}\\
&\quad + \sum_{k=1}^{M_h} \int_{e_k} \{p_{m+\frac{1}{2}} - p(t_{m+\frac{1}{2}})\}[\boldsymbol{\phi}_{m+\frac{1}{2}}] \cdot \mathbf{n}_k\\
&\le \frac{\kappa\nu}{64}\|\boldsymbol{\phi}_{m+\frac{1}{2}}\|_X^2 + C\nu^{-1}h^{2r}(|p_{m+1}|^2_{r,\Omega} + |p_m|^2_{r,\Omega}) + C\nu^{-1}\Delta t^4 \|p_{tt}\|^2_{L^\infty(0,T;H^1(\Omega))}.
\end{aligned}$$

We now bound A_7, using a Taylor expansion:

$$A_7 \le C\Delta t^2 \|\boldsymbol{u}_{ttt}(t^*)\|_{0,\Omega}\|\boldsymbol{\phi}_{m+\frac{1}{2}}\|_{0,\Omega} \le \frac{\kappa\nu}{64}\|\boldsymbol{\phi}_{m+\frac{1}{2}}\|_X^2 + C\nu^{-1}\Delta t^4\|\boldsymbol{u}_{ttt}\|^2_{L^\infty(0,T;L^2(\Omega))}.$$

Also using a Taylor expansion, we bound A_9:

$$A_9 \le C\nu^{-1}\Delta t^4 \|\boldsymbol{f}_{tt}\|^2_{L^\infty(0,T;L^2(\Omega))} + \frac{\kappa\nu}{64}\|\boldsymbol{\phi}_{m+\frac{1}{2}}\|_X^2.$$

Finally the last term A_{10} is handled as follows:

$$\begin{aligned}
A_{10} &= \nu[a(\boldsymbol{\eta}_{m+\frac{1}{2}}, \boldsymbol{\phi}_{m+\frac{1}{2}}) + J(\boldsymbol{\eta}_{m+\frac{1}{2}}, \boldsymbol{\phi}_{m+\frac{1}{2}})]\\
&\quad + \nu[a(\boldsymbol{u}(t_{m+\frac{1}{2}}) - \boldsymbol{u}_{m+\frac{1}{2}}, \boldsymbol{\phi}_{m+\frac{1}{2}}) + J(\boldsymbol{u}(t_{m+\frac{1}{2}}) - \boldsymbol{u}_{m+\frac{1}{2}}, \boldsymbol{\phi}_{m+\frac{1}{2}})] = A_{101} + A_{102}.
\end{aligned}$$

The term A_{101} is bounded like T_8. The term A_{102} reduces to

$$\begin{aligned}
A_{102} &= \nu \sum_{j=1}^{N_h} \int_{E_j} \nabla(\boldsymbol{u}(t_{m+\frac{1}{2}}) - \boldsymbol{u}_{m+\frac{1}{2}}) : \nabla\boldsymbol{\phi}_{m+\frac{1}{2}}\\
&\quad - \nu \sum_{k=1}^{M_h} \int_{e_k} \{\nabla(\boldsymbol{u}(t_{m+\frac{1}{2}}) - \boldsymbol{u}_{m+\frac{1}{2}})\boldsymbol{n}_k\}[\boldsymbol{\phi}_{m+\frac{1}{2}}] \le \frac{\kappa\nu}{64}\|\boldsymbol{\phi}_{m+\frac{1}{2}}\|_X^2\\
&\quad + C\nu\Delta t^4 \|\boldsymbol{u}_{tt}\|^2_{L^\infty(0,T;H^2(\Omega))}.
\end{aligned}$$

Combining all the bounds above yields

$$\begin{aligned}
&\frac{1}{2\Delta t}(\|\boldsymbol{\phi}_{m+1}\|^2_{0,\Omega} - \|\boldsymbol{\phi}_m\|^2_{0,\Omega}) + \frac{\nu\kappa}{2}\|\boldsymbol{\phi}_{m+\frac{1}{2}}\|_X^2 + \frac{\nu_T}{2}\vert\vert\vert(I - P_H)\nabla\boldsymbol{\phi}_{m+\frac{1}{2}}\vert\vert\vert_0^2\\
&\le C\nu^{-1}(\|\boldsymbol{\phi}_m\|^2_{0,\Omega} + \|\boldsymbol{\phi}_{m+1}\|^2_{0,\Omega}) + Ch^{2r}(\nu + \nu^{-1} + \nu_T)(|\boldsymbol{u}_{m+1}|^2_{r+1,\Omega} + |\boldsymbol{u}_m|^2_{r+1,\Omega})\\
&\quad + Ch^{2r}\nu^{-1}(|p_{m+1}|^2_{r,\Omega} + |p_m|^2_{r,\Omega}) + C\Delta t^4\nu\|\boldsymbol{u}_{ttt}\|^2_{L^\infty(0,T;H^2(\Omega))}\\
&\quad + C\Delta t^4\nu^{-1}(\|\boldsymbol{u}_{tt}\|^2_{L^\infty(0,T;H^1(\Omega))} + \|p_{tt}\|^2_{L^\infty(0,T;H^1(\Omega))} + \|\boldsymbol{u}_{ttt}\|^2_{L^\infty(0,T;L^2(\Omega))}\\
&\quad + \|\boldsymbol{f}_{tt}\|^2_{L^\infty(0,T;L^2(\Omega))}) + C\nu_T H^{2r}(|\boldsymbol{u}_{m+1}|^2_{r+1,\Omega} + |\boldsymbol{u}_m|^2_{r+1,\Omega}).
\end{aligned}$$

The end of the proof is similar to that of Theorem 5.2. □

Corollary 5.4. *Assume that $\nu_T = h^\beta$ and $H = h^{1/\alpha}$, where $\beta \geq 2r(\alpha-1)/\alpha$ (see Corollary 4.2); then the estimates in Theorems 5.2 and 5.3 are optimal:*

$$\max_{m=0,\ldots,M} \|\boldsymbol{u}_m - \boldsymbol{u}_m^h\|_{0,\Omega} + \left(\Delta t \sum_{m=0}^{M-1} \|\boldsymbol{u}_{m+1} - \boldsymbol{u}_{m+1}^h\|_X^2\right)^{1/2} = \mathcal{O}(h^r + \Delta t),$$

$$\max_{m=0,\ldots,M} \|\boldsymbol{u}_m - \tilde{\boldsymbol{u}}_m\|_{0,\Omega} + \left(\Delta t \sum_{m=0}^{M-1} \|\boldsymbol{u}_{m+1} - \tilde{\boldsymbol{u}}_{m+1}\|_X^2\right)^{1/2} = \mathcal{O}(h^r + \Delta t^2).$$

Remark 3. The analysis presented in this paper is applicable to the three-dimensional Navier–Stokes equations assuming that the L^p bound (2.6) and the inf-sup condition (3.10) hold true.

6. Conclusion. In this paper, we have analyzed the stability and convergence of totally discontinuous schemes for solving the time-dependent Navier–Stokes equations. Both semidiscrete approximation and fully discrete approximation are constructed for velocity. In addition, semidiscrete approximation of pressure is obtained. We showed that these estimations are optimal. Numerical experiments are currently under investigation.

REFERENCES

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

[2] G. Baker, *Galerkin approximations for the Navier-Stokes equations*, manuscript, Harvard University, Cambridge, MA, 1976.

[3] C. E. Baumann, *An h-p Adaptive Discontinuous Finite Element Method for Computational Fluid Dynamics*, Ph.D. thesis, The University of Texas, Austin, TX, 1997.

[4] G. Birkhoff and G. Rota, *Ordinary Differential Equations*, Ginn, Boston, 1962.

[5] J. Boussinesq, *Théorie de l'écoulement tourbillant*, Mem. Pres. Acad. Sci. Paris, 23 (1877), pp. 46–50.

[6] B. Cockburn, G. Karniadakis, and C. W. Shu, *Discontinuous Galerkin Methods: Theory, Computation, and Applications*, Springer-Verlag, Berlin, 2000.

[7] M. Crouzeix and R. Falk, *Nonconforming finite elements for the Stokes problem*, Math. Comp., 52 (1989), pp. 437–456.

[8] M. Crouzeix and P. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO Ser. Rouge, (1973), pp. 33–75.

[9] M. Fortin and M. Soulie, *A non-conforming piecewise quadratic finite element on triangles*, Internat. J. Numer. Methods, 19 (1983), pp. 505–520.

[10] U. Frisch and S. A. Orszag, *Turbulence: Challenges for theory and experiment*, Phys. Today, (1990), pp. 24–32.

[11] V. Girault and P.-A. Raviart, *Finite element approximation of the Navier-Stokes equations*, Lecture Notes in Math. 749, Springer-Verlag, Berlin, 1979.

[12] V. Girault, B. Rivière, and M. F. Wheeler, *A discontinuous Galerkin method with nonoverlapping domain decomposition for the Stokes and Navier-Stokes problems*, Math. Comp., 74 (2005), pp. 53–84.

[13] V. Girault and R. Scott, *A quasi-local interpolation operator preserving the discrete divergence*, Calcolo, 40 (2003), pp. 1–19.

[14] J.-L. Guermond, *Stabilization of Galerkin approximations of transport equations by subgrid modeling*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1293–1316.

[15] J. G. Heywood and R. Rannacher, *Finite-element approximation of the nonstationary Navier–Stokes problem part* IV: *Error analysis for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.

[16] T. J. R. Hughes, *The multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid-scale models, bubbles and the origin of stabilized methods*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 387–401.

[17] T. J. R. Hughes, L. Mazzei, and K. E. Jansen, *Large eddy simulation and the variational multiscale method*, Comput. Visual. Sci., 3 (2000), pp. 47–59.

[18] T. J. R. Hughes, A. A. Oberai, and L. Mazzei, *Large eddy simulation of turbulent channel flows by the variational multiscale method*, Phys. Fluids, 13 (2001), pp. 1784–1799.

[19] T. Iliescu and W. J. Layton, *Approximating the larger eddies in fluid motion.* III. *The Boussinesq model for turbulent fluctuations*, An. Stiint. Univ. Al. I. Cuza Ias., Mat. (N.S.), 44 (1998), pp. 245–261.

[20] V. John and S. Kaya, *A finite element variational multiscale method for the Navier–Stokes equations*, SIAM J. Sci. Comput., 26 (2005), pp. 1485–1503.

[21] O. A. Karakashian and W. N. Jureidini, *Nonconforming finite element method for the stationary Navier–Stokes equations*, SIAM J. Numer. Anal., 35 (1998), pp. 93–120.

[22] S. Kaya and W. Layton, *Subgrid-scale eddy viscosity methods are variational multiscale methods*, Tech. report TR-MATH 03-05, University of Pittsburgh, Pittsburgh, PA, 2003.

[23] S. Kaya and B. Rivière, *Analysis of a discontinuous Galerkin and eddy viscosity method for Navier-Stokes*, Tech. report TR-MATH 03-14, University of Pittsburgh, Pittsburgh, PA, 2003.

[24] W. J. Layton, *A connection between subgrid scale eddy viscosity and mixed methods*, Appl. Math. Comput., 133 (2002), pp. 147–157.

[25] P. Lesaint and P. A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Element Methods in Partial Differential Equations, C. A. deBoor, ed., Academic Press, New York, 1974, pp. 89–123.

[26] R. Lewandowski, *Analyse Mathématique et Oceanographie*, Masson, Paris, 1997.

[27] Y. Maday and E. Tadmor, *Analysis of spectral vanishing viscosity method for periodic conservation laws*, SIAM J. Numer. Anal., 26 (1989), pp. 854–870.

[28] W. H. Reed and T. R. Hill, *Triangular mesh methods for the neutron transport equation*, Tech. report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[29] J. Smagorinsky, *General circulation experiments with the primitive equation,* I: *The basic experiment*, Month. Weath. Rev., 91 (1963), pp. 99–164.

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

[31] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1596–1626

A FINITE DIFFERENCE SCHEME FOR OPTION PRICING IN JUMP DIFFUSION AND EXPONENTIAL LÉVY MODELS*

RAMA CONT† AND EKATERINA VOLTCHKOVA†

Abstract. We present a finite difference method for solving parabolic partial integro-differential equations with possibly singular kernels which arise in option pricing theory when the random evolution of the underlying asset is driven by a Lévy process or, more generally, a time-inhomogeneous jump-diffusion process. We discuss localization to a finite domain and provide an estimate for the localization error under an integrability condition on the Lévy measure. We propose an explicit-implicit finite difference scheme which can be used to price European and barrier options in such models. We study stability and convergence of the scheme proposed and, under additional conditions, provide estimates on the rate of convergence. Numerical tests are performed with smooth and nonsmooth initial conditions.

Key words. parabolic integro-differential equations, finite difference methods, Lévy process, jump-diffusion models, option pricing, viscosity solutions

AMS subject classifications. 47G20, 65M06, 65M12, 49L25, 60H30, 60G51

DOI. 10.1137/S0036142903436186

1. Introduction.

The shortcomings of diffusion models in representing the risk related to large market movements have led to the development of various option pricing models with jumps, where large returns are represented as discontinuities in prices as a function of time. Models with jumps allow for a more realistic representation of price dynamics and greater flexibility in modelling and have been the focus of much recent work [11].

Exponential Lévy models, where the market price of an asset is represented as the exponential $S_t = \exp(rt + X_t)$ of a Lévy process X_t, offer analytically tractable examples of positive jump processes which are simple enough to allow a detailed study both in terms of statistical properties and as models for risk-neutral dynamics, i.e., option pricing models. Option pricing with exponential Lévy models is discussed in [11, 17, 24]. The flexibility of choice of the Lévy process X allows us to calibrate the model to market prices of options and reproduce a wide variety of implied volatility skews/smiles [12]. The Markov property of the price allows us to express prices of European and barrier options in terms of solutions of partial integro-differential equations (PIDEs) that involve, in addition to a (possibly degenerate) second-order differential operator, a nonlocal integral term that requires specific treatment at both the theoretical and numerical levels [13].

In this paper, we propose a finite difference scheme for solving such PIDEs. Our numerical solution is based on splitting the operator into a local and a nonlocal part: we treat the local term using an implicit step and the nonlocal term using an explicit step. This idea, previously used for nonlinear PDEs [3], allows for an efficient numerical implementation. Some difficulties arise due to the nonlocal character of the

*Received by the editors October 13, 2003; accepted for publication (in revised form) October 1, 2004; published electronically October 31, 2005. This work was presented at the Bachelier Seminar (Paris, 2003), Workshop on Lévy Processes and Partial Integro-Differential Equations (Palaiseau, France, 2003), IFIP 2003 (Nice, France), Workshop on Computational Finance (Zürich, 2003), and ECCOMAS 2004 (Jyväskylä, Finland).

http://www.siam.org/journals/sinum/43-4/43618.html

†CMAP-Ecole Polytechnique, F 91128 Palaiseau, France (Rama.Cont@polytechnique.org, voltchko@cmapx.polytechnique.fr).

integral operator, nonsmoothness of initial conditions, the singularity at zero of the integral kernel, and the possible degeneracy of the diffusion coefficient. We resolve these difficulties in the framework of viscosity solutions and provide error estimates in each case, under assumptions which are easily verified on the Lévy density. We study the consistency and stability of this scheme, show its convergence to the solution of the PIDE, and study its numerical performance in two examples, the Merton model with Gaussian jumps and the infinite activity variance Gamma model. Our scheme can be used for European and barrier options and can also be extended to the case of nonconstant coefficients.

1.1. Relation to previous literature. Various numerical methods for solving such parabolic integro-differential equations have been proposed in the recent literature [2, 25, 16, 31]. In the case where the characteristic function of the log-price is known analytically, the fast Fourier transform of Carr and Madan [9] can be used for pricing European options. Though our finite difference method requires more operations than the fast Fourier transform [9], our method does not require a closed form expression for the characteristic function of the log-price and can also handle barrier options (i.e., boundary value problems).

Finite difference schemes for PIDEs have been proposed in [2, 16, 30], but a rigorous analysis of consistency, stability, and convergence is absent from these studies. By appealing to the formalism of viscosity solutions, our analysis allows fairly general hypotheses on the model and applies to models based on pure-jump Lévy processes such as the variance Gamma model [23], hyperbolic models [17], and the normal inverse Gaussian (NIG) model [7].

For jump-diffusion models with finite jump intensity, Andersen and Andreasen [2] proposed an operator splitting method where the differential part is treated using a Crank–Nicholson step and the jump integral is computed using an explicit time step. Our method applies more generally to models with infinite activity, i.e., singular integral kernels; in addition, we propose an analysis of the convergence of our algorithm, which is absent in [2].

Using a variational formulation of the integro-differential equation, Zhang [31] studied a finite difference scheme in the case of jump-diffusion models having finite intensity and possessing all exponential moments (see also [16]). These conditions rule out all models in the literature except the Merton model: our analysis does not require such restrictive conditions. The variational formulation has been recently extended by Matache, von Petersdorff, and Schwab [25] to the infinite activity case using a wavelet Galerkin method. While the approach of [25] is more general than the above approaches, it does not allow us to treat singular cases such as the variance Gamma model [23].

1.2. Outline. Section 2 starts by recalling facts about Lévy processes and exponential Lévy models. In section 3 we briefly discuss, following [13], the characterization of prices of European and barrier options in exponential Lévy models in terms of viscosity solutions of PIDEs.

Solving such PIDEs by finite difference methods involves several approximations: localization of the equation to a bounded domain, treatment of the singularity due to small jumps, discretization of the equation in space, and iteration in time. We discuss localization errors in section 4 and provide an estimate for the localization errors under an integrability condition on the Lévy measure. In section 5 we propose an explicit-implicit finite difference scheme and study consistency, stability, and convergence of the scheme proposed. Convergence properties of the scheme are studied in section

6: we show that the scheme is monotone, unconditionally stable, and consistent and exhibit conditions for its convergence to the solution of the PIDE. Under further conditions on the scheme, we are able to give an estimate of the rate of convergence in section 6.4. Finally, in section 7, numerical tests are performed for smooth and nonsmooth initial conditions to assess the effect of various numerical parameters on the accuracy of the scheme.

2. Exponential Lévy models.

We consider here the class of exponential Lévy models: the risk-neutral dynamics of the underlying asset is given by $S_t = \exp(rt + X_t)$, where X_t is a time-homogeneous jump-diffusion (Lévy) process.

2.1. Lévy processes: Definitions.

A Lévy process is a stochastic process X_t with stationary independent increments which is continuous in probability. Without loss of generality we assume that $X_0 = 0$. The characteristic function of X_t has the following form, called the Lévy–Khinchin representation [27]:

$$E[e^{izX_t}] = e^{-t\psi(z)} = \exp\left\{t\left(-\frac{\sigma^2 z^2}{2} + i\gamma z + \int_{-\infty}^{\infty}(e^{izx} - 1 - izx1_{|x|\le 1})\nu(dx)\right)\right\}, \tag{2.1}$$

where $\sigma > 0$ and γ are real constants and ν is a positive measure verifying

$$\int_{-1}^{+1} x^2\nu(dx) < \infty, \qquad \int_{|x|>1}\nu(dx) < \infty. \tag{2.2}$$

The random process X can be interpreted as the superposition of a Brownian motion with drift and an infinite superposition of independent (compensated) Poisson processes with various jump sizes x, $\nu(dx)$ being the intensity of jumps of size x. In general ν is not a finite measure: $\int \nu(dx)$ need not be finite. In the case where $\lambda = \int \nu(dx) < +\infty$, the measure ν can be normalized to define a *probability measure* μ, which can now be interpreted as the distribution of jump sizes:

$$\mu(dx) = \frac{\nu(dx)}{\lambda}.$$

The jumps of X are then described by a *compound Poisson* process with λ as jump intensity (average number of jumps per unit time) and $\mu(.)$ as jump size distribution. In this case the truncation of small jumps is not needed, and the Lévy–Khinchin representation reduces to

$$E[e^{izX_t}] = \exp\left\{t\left(-\frac{\sigma^2 z^2}{2} + i\gamma_0(\nu)z + \int_{-\infty}^{\infty}(e^{izx} - 1)\nu(dx)\right)\right\}.$$

A Lévy process is a Markov process; its infinitesimal generator $L^X : f \to L^X f$ is an integro-differential operator defined by the expression

$$\begin{aligned} L^X f(x) &= \lim_{t\to 0}\frac{E[f(x+X_t)] - f(x)}{t} \\ &= \frac{\sigma^2}{2}\frac{\partial^2 f}{\partial x^2} + \gamma\frac{\partial f}{\partial x} + \int \nu(dy)\left[f(x+y) - f(x) - y1_{\{|y|\le 1\}}\frac{\partial f}{\partial x}(x)\right], \end{aligned} \tag{2.3}$$

which is well defined for $f \in C^2(\mathbb{R})$ with compact support.

2.2. Exponential Lévy models. Let $(S_t)_{t\in[0,T]}$ be the price of a financial asset modelled as a stochastic process on a filtered probability space $(\Omega,\mathcal{F},\mathcal{F}_t,\mathbb{Q})$. Under the hypothesis of absence of arbitrage there exists a measure equivalent to $\mathbb{Q}$ under which (S_t) is a martingale. We will assume in what follows that $\mathbb{Q}$ is a martingale measure.

Exponential Lévy models assume that the (risk-neutral) dynamics of S_t under $\mathbb{Q}$ is represented as the exponential of a Lévy process:

$$S_t = S_0 e^{rt+X_t}. \tag{2.4}$$

Here X_t is a Lévy process with characteristic triplet (σ,γ,ν), and the interest rate r is included for ease of notation. Different exponential Lévy models proposed in the financial modelling literature simply correspond to different choices for the Lévy measure ν; see [11, Chap. 3] for a review. The absence of arbitrage then imposes that $\hat{S}_t = S_t e^{-rt} = \exp X_t$ is a martingale, which is equivalent to the following conditions on the triplet (σ,γ,ν):

$$\int_{|y|>1} \nu(dy)e^y < \infty, \tag{2.5}$$

$$\gamma = \gamma(\sigma,\nu) = -\frac{\sigma^2}{2} - \int (e^y - 1 - y1_{|y|\leq 1})\nu(dy). \tag{2.6}$$

We will assume this relation holds in what follows. The infinitesimal generator L^X then becomes

$$L^X f(x) = \frac{\sigma^2}{2}\left[\frac{\partial^2 f}{\partial x^2} - \frac{\partial f}{\partial x}\right] + \int_{-\infty}^{\infty} \nu(dy)\ \left[f(x+y) - f(x) - (e^y-1)\frac{\partial f}{\partial x}(x)\right]. \tag{2.7}$$

We will also use the notation $Y_t = rt + X_t$. The infinitesimal generator of Y_t is

$$Lf = L^X f + r\frac{\partial f}{\partial x}. \tag{2.8}$$

3. Partial integro-differential equation for option prices. The value of an option is defined as a discounted conditional expectation of its terminal payoff H_T under the risk-adjusted martingale measure (sometimes called risk-neutral probability) $\mathbb{Q}$:

$$C_t = E[e^{-r(T-t)}H_T|\mathcal{F}_t].$$

For a European call or put, $H_T = H(S_T)$. From the Markov property, $C_t = C(t,S)$, where

$$C(t,S) = E[e^{-r(T-t)}H(S_T)|S_t = S]. \tag{3.1}$$

Introducing the change of variable $\tau = T-t$, $x = \ln(S/S_0)$, and defining $h(x) = H(S_0e^x)$ and $u(\tau,x) = e^{r\tau}C(T-\tau,S_0e^x)$, then

$$u(\tau,x) = E[h(x+Y_\tau)]. \tag{3.2}$$

If u is sufficiently smooth—for example, $u \in C^{1,2}$ with bounded derivatives—then by applying Ito's formula to $u(t,X_t)$ between 0 and T one can show [8] that it is a classical solution of the Cauchy problem:

$$\frac{\partial u}{\partial \tau} = Lu \quad \text{on } (0,T]\times\mathbb{R}, \qquad u(0,x) = h(x), \quad x\in\mathbb{R}. \tag{3.3}$$

Barrier options lead to initial-boundary value problems. Consider, for instance, an up-and-out call option with maturity T, strike K, and (upper) barrier $U > S_0$. The terminal payoff is given by

$$H_T = (S_T - K)^+ 1_{T<\theta},$$

where $\theta = \inf\{t \geq 0 \mid S_t \geq U\}$, the first moment when the barrier is crossed. Due to the strong Markov property of Lévy processes, it is possible to express the value of the option $C_t = e^{-r(T-t)} E[H_T|\mathcal{F}_t]$ as a deterministic function of time t and current stock value S_t before the barrier is crossed. Namely, for any $(t, S) \in [0, T] \times (0, \infty)$ we can define

$$C_b(t, S) = e^{-r(T-t)}\ E[H(Se^{Y_{T-t}})1_{T<\theta_t}], \tag{3.4}$$

where $H(S) = (S-K)^+$, $\{Y_{s-t},\ s \geq t\}$ is a Lévy process, and $\theta_t = \inf\{s \geq t \mid Se^{Y_{s-t}} \geq U\}$, the first exit time after t. Then,

$$C_t = C_b(t,\ S_t)1_{t\leq\theta} \tag{3.5}$$

for all $t \leq T$. Note that outside of the set $\{t \leq \theta\}$ the objects C_t and $C_b(t, S_t)$ are different: if the barrier has already been crossed, C_t will always be zero, but $C_b(t, S_t)$ may become positive if the stock returns to the region below the barrier. By going to the log variables we define

$$u_b(\tau, x) = e^{r\tau} C_b(T - \tau, S_0e^x). \tag{3.6}$$

Again, if u_b is smooth the Itô formula can be used to show [8] that u_b is a solution of the following initial-boundary value problem:

$$\begin{aligned} &\frac{\partial u}{\partial \tau} = Lu \qquad \text{on } (0, T] \times (-\infty, \log(U/S_0)), \\ u(0, x) = h(x), \quad x < \log(U/S_0); \qquad &u(\tau, x) = 0, \quad x \geq \log(U/S_0). \end{aligned}$$

Prices of down-and-out or double barrier options are defined similarly. In the case of pure jump models where $\sigma = 0$, these smoothness conditions can fail to hold; counterexamples are given in [13]. In this case the option price should be seen as a viscosity solution of the PIDE, as discussed below.

3.1. Viscosity solutions for integro-differential equations.

Existence and uniqueness of (classical) solutions for the PIDEs considered above in Sobolev–Hölder spaces have been studied in [8, 18] in the case where the diffusion component is nondegenerate: for a Lévy process this simply means $\sigma > 0$, but more generally these results apply to jump diffusion where the diffusion coefficient is bounded away from zero. However, many of the models in the financial modelling literature are pure jump models with $\sigma = 0$, for which such results are not available. In fact, in pure jump models with finite variation (3.3) is formally a *first* order in the price variable so the effect of the jump term is more like a convection term rather than a diffusion term. A notion of solution that yields existence and uniqueness for such equations without requiring nondegeneracy of coefficients or a priori knowledge of smoothness of solutions is the notion of viscosity solution, introduced by Crandall and Lions for PDEs (see [14]) and extended to integro-differential equations of the type considered here in [1, 4, 26, 28, 29].

Denote by USC (respectively, LSC) the class of upper semicontinuous (respectively, lower semicontinuous) functions $u : (0,T] \times \mathbb{R} \to \mathbb{R}$ and by $C_p^+([0,T] \times \mathbb{R})$ the set of measurable functions on $[0,T] \times \mathbb{R}$ with polynomial growth of degree p at $+\infty$ and bounded on $[0,T] \times \mathbb{R}^-$:

$$\varphi \in C_p^+([0,T] \times \mathbb{R}) \iff \exists C > 0,\ |\varphi(t,x)| \le C(1 + |x|^p\, 1_{x>0}). \tag{3.7}$$

Let $O = (l,u) \subseteq \mathbb{R}$ be an open interval, $\partial O = \{l,u\}$ its boundary, and $g \in C_p^+([0,T] \times \mathbb{R} \setminus O)$ a continuous function. Consider the following initial-boundary value problem on $[0,T] \times \mathbb{R}$:

$$\frac{\partial u}{\partial \tau} = Lu \qquad \text{on } (0,T] \times O, \tag{3.8}$$

$$u(0,x) = h(x), \quad x \in O; \qquad u(\tau,x) = g(\tau,x), \quad x \notin O. \tag{3.9}$$

Definition 3.1 (viscosity solution). *A function $u \in USC$ is a viscosity subsolution of* (3.8)–(3.9) *if for any test function $\varphi \in C^2([0,T] \times \mathbb{R}) \cap C_p^+([0,T] \times \mathbb{R})$ and any global maximum point $(\tau,x) \in [0,T] \times \mathbb{R}$ of $u - \varphi$, the following properties are verified:*

$$\text{if } (\tau,x) \in (0,T] \times O, \qquad \left(\frac{\partial \varphi}{\partial \tau} - L\varphi\right)(\tau,x) \le 0, \tag{3.10}$$

$$\text{if } \tau = 0,\ x \in \overline{O}, \qquad \min\left\{\left(\frac{\partial \varphi}{\partial \tau} - L\varphi\right)(\tau,x),\ u(\tau,x) - h(x)\right\} \le 0,$$

$$\text{if } \tau \in (0,T],\ x \in \partial O, \qquad \min\left\{\left(\frac{\partial \varphi}{\partial \tau} - L\varphi\right)(\tau,x),\ u(\tau,x) - g(\tau,x)\right\} \le 0,$$

$$\text{if } x \notin \overline{O}, \qquad u(\tau,x) \le g(\tau,x). \tag{3.11}$$

A function $u \in LSC$ is a viscosity supersolution of (3.8)–(3.9) *if, for any test function $\varphi \in C^2([0,T] \times \mathbb{R}) \cap C_p^+([0,T] \times \mathbb{R})$ and any global minimum point $(\tau,x) \in [0,T] \times \mathbb{R}$ of $u - \varphi$, we have*

$$\text{if } (\tau,x) \in (0,T] \times O, \qquad \left(\frac{\partial \varphi}{\partial \tau} - L\varphi\right)(\tau,x) \ge 0,$$

$$\text{if } \tau = 0,\ x \in \overline{O}, \qquad \max\left\{\left(\frac{\partial \varphi}{\partial \tau} - L\varphi\right)(\tau,x),\ u(\tau,x) - h(x)\right\} \ge 0,$$

$$\text{if } \tau \in (0,T],\ x \in \partial O, \qquad \max\left\{\left(\frac{\partial \varphi}{\partial \tau} - L\varphi\right)(\tau,x),\ u(\tau,x) - g(\tau,x)\right\} \ge 0,$$

$$\text{if } x \notin \overline{O}, \qquad u(\tau,x) \ge g(\tau,x).$$

A function $u \in C_p^+([0,T] \times \mathbb{R})$ is called a viscosity solution of (3.8)–(3.9) *if it is both a subsolution and a supersolution. This function is then continuous on* $(0,T] \times \mathbb{R}$.

Note that the initial and boundary conditions are verified in a viscosity sense. The definition also includes the case of initial value problems: $O = \mathbb{R}$. Existence and uniqueness of viscosity solutions for such parabolic integro-differential equations in the case $O = \mathbb{R}$ are discussed in [1] in the case where ν is a finite measure and in [4] and [26] for general Lévy measures. Growth conditions other than $u \in C_p^+$ can be considered (see, e.g., [1, 4]) with additional conditions on the Lévy measure ν. The main tool for showing uniqueness is the comparison principle: if u, v are viscosity solutions and $u(0,x) \geq v(0,x)$, then for all $\tau \in [0,T]$, $u(\tau,x) \geq v(\tau,x)$. This property can be extended to subsolutions and supersolutions in the following sense [1, 20].

PROPOSITION 3.2 (comparison principle for semicontinuous solutions [1, 20]). *If $u \in USC$ is a subsolution and $v \in LSC$ is a supersolution of* (3.8)–(3.9) *with $O = \mathbb{R}$ and h is a continuous function, then $u \leq v$ on $[0,T] \times \mathbb{R}$.*

Proofs and extensions can be found in [1] for the case where ν is a bounded measure; the case of a general Lévy measure has recently been treated in [20].

3.2. Option prices as viscosity solutions of PIDEs.

The following result, whose proof in a more general setting is given in [13], shows that values of European and barrier options with Lipschitz payoff function can be expressed in terms of (viscosity) solutions of (3.8)–(3.9).

PROPOSITION 3.3 (option prices as viscosity solutions). *Let the payoff function H verify the Lipschitz condition on its domain of definition,*

$$|H(S_1) - H(S_2)| \leq C|S_1 - S_2| \qquad \forall S_1, S_2 \in (S_0e^l, S_0e^u), \tag{3.12}$$

and let $h(x) = H(S_0e^x)$ have polynomial growth at infinity. Then the following hold:

1. *The forward value of a European option $u(\tau,x)$ defined by* (3.2) *is the unique viscosity solution of the Cauchy problem* (3.3).
2. *If the forward value $u_b(\tau,x)$ of a knockout (single or double) barrier option defined by* (3.6) *is continuous, then it is a viscosity solution of* (3.8)–(3.9) *(with $g \equiv 0$).*

The assumptions on the payoff function apply to put options, single-barrier knockout puts, double barrier knockout options and also to the log-contract. One can then retrieve call options by put-call parity. For barrier options, continuity holds in particular if $\nu(\mathbb{R}) < \infty$ and $\sigma > 0$ but also in pure jump models with infinite activity [13]. For barrier options with rebate, the zero boundary condition has to be replaced by the value of the rebate, as in the case of diffusion models.

4. Localization estimates.

In order to solve numerically the PIDE, we first localize the variables and the integral term to bounded domains. This section discusses estimates for the localization error using a probabilistic approach.

4.1. Localization to a bounded domain.

To solve numerically the initial-boundary value problem (3.8)–(3.9) in the case of an unbounded domain O, we first truncate the domain to an interval $x \in (-A, A)$. Usually, this leads us to define some boundary conditions at $x = -A$ and $x = A$. As noted above, the operator L is nonlocal: computing the integral term at a point $x \in (-A, A)$ requires knowledge of $u(\tau, \cdot)$ on $\{x + y \mid y \in \operatorname{supp} \nu\}$, which in most examples is equal to the whole real line $\mathbb{R}$. In the case of knock-out barrier options, a natural boundary condition is given by the zero extension (or the rebate). In other cases, this extension is done by imposing a numerical boundary condition. Choosing $u(\tau,x) = g(\tau,x)$ for some given continuous function g with polynomial growth will lead to a probabilistic interpretation of the solution of the localized problem.

Although many choices are possible for the boundary condition g, we will consider here two cases. The simplest choice is $g = 0$, i.e., extend the solution by zero outside the domain. Another extension is given by the payoff function (the initial condition) itself, $g(\tau,x) = h(x)$, which is asymptotically close to the solution at infinity. We will see that both choices lead to a localization error that exponentially decreases with the domain size. Let us define $u_A(\tau,x)$ as the solution of the localized problem

$$\frac{\partial u_A}{\partial \tau} = Lu_A, \qquad (0,T]\times(-A,A), \tag{4.1}$$
$$u_A(0,x)=h(x), \qquad x\in(-A,A); \qquad u_A(\tau,x)=g(\tau,x), \qquad x\notin(-A,A),$$

where $g=0$ or $g(\tau,x)=h(x)$.

PROPOSITION 4.1. *Assume that h is bounded ($||h||_\infty<\infty$) and*

$$\exists\alpha>0, \int_{|x|>1} e^{\alpha|x|}\nu(dx)<\infty. \tag{4.2}$$

Let $u(\tau,x)$ be the solution of (3.3) *and let $u_A(\tau,x)$ be the (viscosity) solution of* (4.1) *with boundary condition $g=0$ or $g(\tau,x)=h(x)$. Then*

$$|u(\tau,x)-u_A(\tau,x)|\le 2C_{\tau,\alpha}||h||_\infty e^{-\alpha(A-|x|)} \qquad \forall x\in(-A,A), \tag{4.3}$$

where the constant $C_{\tau,\alpha}$ does not depend on A.

Proof. The proof is based on the probabilistic representation [8, 13] of the solutions of (5.1) and (4.1). Let us define $M_\tau^x=\sup_{t\in[0,\tau]}|Y_t+x|$. Then

$$\begin{aligned} u(\tau,x) &= \mathbb{E}[h(Y_\tau+x)],\\ u_A(\tau,x) &= \mathbb{E}[h(Y_\tau+x)1_{\{M_\tau^x<A\}}] \qquad \text{if } g=0,\\ \text{or } u_A(\tau,x) &= \mathbb{E}[h(Y_\tau+x)1_{\{M_\tau^x<A\}}+h(Y_{\theta(x)}+x)1_{\{M_\tau^x\ge A\}}], \end{aligned}$$

where $\theta(x)=\inf\{t\ge0,|Y_t+x|\ge A\}$ is the first exit time of Y_t+x from $[-A,A]$. Subtracting u_A from u gives

$$\begin{aligned} |u(\tau,x)-u_A(\tau,x)| &= |\mathbb{E}h(Y_\tau+x)1_{\{M_\tau^x\ge A\}}|\\ &\le ||h||_\infty\mathbb{Q}(M_\tau^x\ge A) \quad \text{for } g=0, \end{aligned}$$

and in the case $g(\tau,x)=h(x)$ we obtain

$$\begin{aligned} |u(\tau,x)-u_A(\tau,x)| &\le \mathbb{E}|h(Y_\tau+x)1_{\{M_\tau^x\ge A\}}|+\mathbb{E}|h(Y_{\theta(x)}+x)1_{\{M_\tau^x\ge A\}}|\\ &\le 2||h||_\infty\mathbb{Q}(M_\tau^x\ge A). \end{aligned}$$

So, in both cases

$$|u(\tau,x)-u_A(\tau,x)|\le 2||h||_\infty\mathbb{Q}(M_\tau^x\ge A). \tag{4.4}$$

Theorem 25.18 of [27] together with (4.2) implies

$$C_{\tau,\alpha}=\mathbb{E}e^{\alpha M_\tau^0}<\infty. \tag{4.5}$$

Therefore, Chebyshev's inequality applies, and we obtain

$$\mathbb{Q}(M_\tau^0\ge A)\le C_{\tau,\alpha}e^{-\alpha A}. \tag{4.6}$$

Now, to pass from M_τ^0 to M_τ^x, we use the following implications:

$$\begin{array}{rrcl} & \sup|Y_t+x| & \le & \sup|Y_t|+|x|\\ \Rightarrow & (\sup|Y_t+x|\ge A & \Rightarrow & \sup|Y_t|+|x|\ge A)\\ \Rightarrow & \mathbb{Q}(M_\tau^x\ge A) & \le & \mathbb{Q}(M_\tau^0\ge A-|x|)\\ & & \le & C_{\tau,\alpha}e^{-\alpha(A-|x|)} \quad \text{by (4.6)}. \end{array}$$

Combining the last inequality with (4.4) gives the desired result. □

Remark 1. In the case of a put option, $\|h\|_\infty < \infty$ and Proposition 4.1 applies. In other examples, h may be unbounded; in this case it is still possible to obtain an exponentially decreasing localization error under additional restrictions on ν. Note that for the call option, although the payoff grows exponentially, one can transform the problem into pricing a put using put-call parity in order to obtain a smaller localization error.

Remark 2. An exponential bound on localization error in the L_2-norm is given in [25] using analytical methods. The advantage of the probabilistic approach is to provide a local (pointwise) estimate. For instance, our estimate (4.3) reflects the intuitive fact that the localization error is more pronounced near the boundary.

The above result implies that the localization error decreases uniformly on each closed subinterval of $(-A, A)$:

$$|u(\tau,x) - u_A(\tau,x)| \le k e^{-\alpha\delta A} \qquad \text{for } |x| \le (1-\delta)A,$$

where $0 < \delta < 1$.

Assumption (4.2) means that the tails of ν have to decrease exponentially, which is true in all examples considered in the option pricing literature (except Carr and Wu's log-stable model [10]). Note that in an exponential Lévy model we already have $\int_1^{+\infty} e^{\alpha x}\nu(\mathrm{d}x) < \infty$ for all $\alpha \le 1$, because of the martingale condition, so (4.2) is a condition on the negative jumps.

4.2. Truncation of the integral. To compute numerically the integral term, we need to reduce the region of integration to a bounded interval. In terms of the jump process, this amounts to the truncation of large jumps. We will now give an estimate for the error resulting from this approximation. Recall that the solution of the Cauchy problem (3.3) (in the European case where $O = \mathbb{R}$) for a Lipschitz payoff function H is

$$u(\tau,x) = \mathbb{E}[H(S_0 e^{x+r\tau+X_\tau})], \tag{4.7}$$

where X_τ is a Lévy process with the triplet (γ,σ,ν). Let us define a new process $\tilde{X}_\tau$ characterized by the Lévy triplet $(\tilde{\gamma},\sigma,\nu\mathbf{1}_{x\in[B_l,B_r]})$, where $\tilde{\gamma}$ is such that $\exp(rt+\tilde{X}_t)$ remains a martingale:

$$\tilde{\gamma} = -\frac{\sigma^2}{2} - \int_{B_l}^{B_r} (e^y - 1 - y1_{|y|\le 1})\nu(dy).$$

We now define

$$\tilde{u}(\tau,x) = \mathbb{E}[H(S_0 e^{x+r\tau+\tilde{X}_\tau})] \tag{4.8}$$

and we estimate the difference between $\tilde{u}$ and the true solution u.

PROPOSITION 4.2. *Let H be Lipschitz: $|H(S_1)-H(S_2)| \le c|S_1-S_2|$. Assume that there exists $\alpha_r, \alpha_l > 0$, such that $\int_1^\infty e^{(1+\alpha_r)y}\nu(dy) < \infty$ and $\int_{-\infty}^{-1} |y|e^{\alpha_l|y|}\nu(dy) < \infty$.*

If u and $\tilde{u}$ are defined by (4.7) *and* (4.8), *respectively, then*

$$|u(\tau,x) - \tilde{u}(\tau,x)| \le 2c\, S_0 e^{x+r\tau}\tau(C_1 e^{-\alpha_l|B_l|} + C_2 e^{-\alpha_r|B_r|}). \tag{4.9}$$

Proof. Let us denote $R_\tau = X_\tau - \tilde{X}_\tau$, $R_\tau \perp\!\!\!\perp \tilde{X}_\tau$. We have

$$\begin{aligned}|u(\tau,x) - \tilde{u}(\tau,x)| &= |\mathbb{E}[H(S_0 e^{x+r\tau+\tilde{X}_\tau+R_\tau})] - \mathbb{E}[H(S_0 e^{x+r\tau+\tilde{X}_\tau})]| \\ &\le c\, S_0 e^{x+r\tau}\mathbb{E}[e^{\tilde{X}_\tau}|e^{R_\tau}-1|] = c\, S_0 e^{x+r\tau}\mathbb{E}|e^{R_\tau}-1|.\end{aligned}$$

By construction, $\mathbb{E}[e^{R_\tau}-1]=0$. Since $|e^{R_\tau}-1|=(e^{R_\tau}-1)+2(1-e^{R_\tau})^+$ and $(1-e^{R_\tau})^+\le|R_\tau|$, we obtain

$$|u(\tau,x)-\tilde{u}(\tau,x)|\le 2c\,S_0e^{x+r\tau}\mathbb{E}|e^{R_\tau}|. \tag{4.10}$$

The Lévy triplet of R_τ is $(\gamma-\tilde{\gamma},0,\nu\mathbf{1}_{x\notin[B_l,B_r]})$ with

$$\gamma-\tilde{\gamma}=-\int_{y\notin[B_l,B_r]}(e^y-1)\nu(dy).$$

One can write $R_\tau=P_\tau+N_\tau$, where P_τ and N_τ are characterized by $(\int_{-\infty}^{B_l}(1-e^y)\nu(\mathrm{d}y),0,\nu\mathbf{1}_{x>B_r})$ and $(-\int_{B_r}^{\infty}(e^y-1)\nu(\mathrm{d}y),0,\nu\mathbf{1}_{x<B_l})$, respectively. We assume without loss of generality that $B_l<-1$, $B_r>1$.[1] Since P_τ has nonnegative drift, no Brownian component, and only positive jumps bounded from below by $B_r>0$, we have $P_\tau\ge 0$ (recall that $P_0=0$). Conversely, N_τ has only negative jumps (bounded from above by $B_l<0$) and nonpositive drift. In consequence, $N_\tau\le 0$. Therefore,

$$\begin{aligned}\mathbb{E}|R_\tau|\le\mathbb{E}|P_\tau|+\mathbb{E}|N_\tau|\;&=\;\mathbb{E}P_\tau-\mathbb{E}N_\tau\\&=\tau\left[\int_{-\infty}^{B_l}(1-e^y-y)\nu(\mathrm{d}y)+\int_{B_r}^{\infty}(e^y-1+y)\nu(\mathrm{d}y)\right]\\&\le\tau\left[2\int_{-\infty}^{B_l}|y|\nu(\mathrm{d}y)+2\int_{B_r}^{\infty}e^y\nu(\mathrm{d}y)\right].\end{aligned}\tag{4.11}$$

Using the hypotheses on ν, we obtain

$$\begin{aligned}\mathbb{E}|R_\tau|&\le\tau\left(2e^{-\alpha_l|B_l|}\int_{-\infty}^{B_l}|y|e^{\alpha_l|y|}\nu(\mathrm{d}y)+2e^{-\alpha_r|B_r|}\int_{B_r}^{\infty}e^{(1+\alpha_r)y}\nu(\mathrm{d}y)\right)\\&\le\tau(C_1e^{-\alpha_l|B_l|}+C_2e^{-\alpha_r|B_r|}),\end{aligned}$$

which we substitute into (4.10). □

Remark 3. The hypotheses on ν in Proposition 4.2 are a little stronger than (4.2). We require them to obtain an exponential decay of the truncation error. However, we can use estimate (4.11) directly. In other words, existence of the integrals in (4.11) suffices to obtain a convergence of $\tilde{u}$ to u as $|B_l|$ and $|B_r|$ grow to infinity, but this convergence does not necessarily occur at an exponential rate.

Remark 4. The requirements are different for the left and right tails of ν. For example, in the variance Gamma model with $\nu(x)=a\exp(-\eta_\pm|x|)/|x|$ one needs η_+ to be greater than 1, and η_- only positive. Proposition 4.9 then applies with $\alpha_l<\eta_-$ and $\alpha_r<\eta_+-1$.

Using Propositions 4.1 and 4.2 we can fix in advance $[-A,A]$ and $[B_l,B_r]$ to have a given bound on the respective errors. In what follows we will assume this has been done and concentrate on the numerical solution of the localized problem.

[1] Clearly, if (4.9) is true for such values, it is true for all B_l, B_r, up to change of the constants. On the other hand, this estimate is not useful if ν has a bounded support. For example, if there are only negative jumps, we will take $B_r=0$, but in this case $\nu\mathbf{1}_{x\le B_r}=\nu$, and there is no truncation error due to B_r.

5. An explicit-implicit finite difference scheme. We now present a numerical procedure for solving the PIDE:

$$\frac{\partial u}{\partial \tau} = Lu, \qquad (0,T]\times O, \tag{5.1}$$

$$u(\tau,x) = g(\tau,x), \quad x \in O^c, \tag{5.2}$$

$$u(0,x) = h(x), \qquad x \in O, \tag{5.3}$$

where L is defined by (2.8), $O^c = \mathbb{R}\setminus O$, and $g \in C_p^+([0,T]\times\mathbb{R}\setminus O)$ is a continuous function.

Our method is based on splitting the operator L into two parts:

$$\frac{\partial u}{\partial \tau} = Du + Ju,$$

where D and J stand for the differential and integral parts of L, respectively. We replace Du with a finite difference approximation $D_\Delta u$ and Ju with the trapezoidal quadrature approximation $J_\Delta u$ and use the following explicit-implicit time-stepping scheme:

$$\frac{u^{n+1}-u^n}{\Delta t} = D_\Delta u^{n+1} + J_\Delta u^n.$$

We treat the integral part in an explicit time stepping in order to avoid the inversion of the nonsparse matrix J_Δ. We show that this does not affect the stability of the scheme: it is unconditionally stable like the fully implicit scheme but does not require us to invert the dense matrix J_Δ. We first describe the space discretization and the time-stepping scheme in the case of a jump-diffusion model where the jump intensity is finite. Next, we deal with the singular case $\nu(\mathbb{R}) = +\infty$ using an approach similar to the "vanishing viscosity" method [14].

5.1. Explicit-implicit scheme: Finite intensity case. We suppose here that $\nu(\mathbb{R}) = \lambda < +\infty$. Then the integro-differential operator can be written as $Lu \equiv Du + Ju$, where

$$Du = \frac{\sigma^2}{2}\frac{\partial^2 u}{\partial x^2} - \left(\frac{\sigma^2}{2} - r + \alpha\right)\frac{\partial u}{\partial x} - \lambda u, \qquad Ju = \int_{B_l}^{B_r} \nu(dy) u(\tau, x+y), \tag{5.4}$$

and $\alpha = \int_{B_l}^{B_r}(e^y - 1)\nu(\mathrm{d}y)$.

We introduce a uniform grid on $[0,T]\times[-A,A]$: $\tau_n = n\Delta t$, $n = 0,\dots,M$, $x_i = -A + i\Delta x$, $i \in \{0,\dots,N\}$, with $\Delta t = T/M$, $\Delta x = 2A/N$. Let $\{u_i^n\}$ be the solution of the numerical scheme, to be defined below.

To approximate the integral terms we use the trapezoidal quadrature rule with the same step Δx. Let K_l, K_r be such that $[B_l, B_r] \subset [(K_l - 1/2)\Delta x, (K_r + 1/2)\Delta x]$. Then

$$\int_{B_l}^{B_r} \nu(dy) u(\tau, x_i + y) \approx \sum_{j=K_l}^{K_r} \nu_j u_{i+j}, \lambda \approx \hat{\lambda} = \sum_{j=K_l}^{K_r} \nu_j,$$

$$\alpha \approx \hat{\alpha} = \sum_{j=K_l}^{K_r} (e^{y_j} - 1)\nu_j, \qquad \text{where } \nu_j = \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} \nu(\mathrm{d}y). \tag{5.5}$$

The space derivatives are discretized using finite differences:

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_i \approx \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta x)^2}, \tag{5.6}$$

$$\left(\frac{\partial u}{\partial x}\right)_i \approx \begin{cases} \frac{u_{i+1}-u_i}{\Delta x} & \text{if} \quad \sigma^2/2 - r + \hat{\alpha} < 0 \\ \frac{u_i - u_{i-1}}{\Delta x} & \text{if} \quad \sigma^2/2 - r + \hat{\alpha} \geq 0. \end{cases} \tag{5.7}$$

The choice of approximation for the first-order derivative is determined by stability requirement and will be discussed later (section 6). Since the two cases are treated similarly, let us suppose without loss of generality that $\sigma^2/2 - r + \hat{\alpha} < 0$.

Using (5.5)–(5.7) we obtain $Du \approx D_\Delta u$, $Ju \approx J_\Delta u$, where

$$(D_\Delta u)_i = \frac{\sigma^2}{2} \frac{u_{i+1} - 2u_i + u_{i-1}}{(\Delta x)^2} - \left(\frac{\sigma^2}{2} - r + \hat{\alpha}\right) \frac{u_{i+1} - u_i}{\Delta x} - \hat{\lambda} u_i, \tag{5.8}$$

$$(J_\Delta u)_i = \sum_{j=K_l}^{K_r} \nu_j u_{i+j}. \tag{5.9}$$

Finally, we replace problem (4.1) with the following time-stepping scheme:

Initialization:

$$u_i^0 = h(x_i), \qquad i \in \{0, \ldots, N\}, \tag{5.10}$$

$$u_i^0 = g(0, x_i) \qquad \text{otherwise.} \tag{5.11}$$

(**S**) For $n = 0, \ldots, M-1$,

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = (D_\Delta u^{n+1})_i + (J_\Delta u^n)_i \quad \text{if } i \in \{0, \ldots, N\}, \tag{5.12}$$

$$u_i^{n+1} = g(\ (n+1)\Delta t, x_i) \qquad \text{if } i \notin \{0, \ldots, N\}. \tag{5.13}$$

5.2. Explicit-implicit scheme: Infinite intensity case. If $\nu(\mathbb{R}) = +\infty$, the above method cannot be applied directly. The idea is to come down to a nonsingular case by approximating the process X_τ by an appropriate finite activity process with a modified diffusion coefficient.

The procedure is similar to the one described in section 4.2, but this time we deal with small jumps. Given $\varepsilon > 0$ let us define a process X_τ^ε characterized by the Lévy triplet $(\gamma(\varepsilon), \sqrt{\sigma^2 + \sigma^2(\varepsilon)}, \nu \mathbf{1}_{|x| \geq \varepsilon})$, where

$$\sigma^2(\varepsilon) = \int_{-\varepsilon}^{\varepsilon} y^2 \nu(\mathrm{d}y),$$

and $\gamma(\varepsilon)$ is determined by the martingale condition

$$\gamma(\varepsilon) = -\frac{\sigma^2 + \sigma^2(\varepsilon)}{2} - \int_{|y| \geq \varepsilon} (e^y - 1 - y 1_{|y| \leq 1}) \nu(dy).$$

This means that we replace the jumps of size smaller than ε by a Brownian motion $\sigma(\varepsilon) W_\tau$. Therefore, X_τ^ε has jumps of finite intensity. The function u^ε defined as

$$u^\varepsilon(\tau, x) = \mathbb{E}[h(x + r\tau + X_\tau^\varepsilon)] \equiv \mathbb{E}[h(x + Y_\tau^\varepsilon)] \tag{5.14}$$

satisfies the following Cauchy problem:

$$\begin{aligned} \frac{\partial u^\varepsilon}{\partial \tau} &= L^\varepsilon u^\varepsilon, \qquad (0,T]\times\mathbb{R}, \\ u^\varepsilon(0,x) &= h(x), \qquad x\in\mathbb{R}, \end{aligned} \tag{5.15}$$

where

$$\begin{aligned} L^\varepsilon f = \frac{\sigma^2+\sigma^2(\varepsilon)}{2}\frac{\partial^2 f}{\partial x^2} - \left(\frac{\sigma^2+\sigma^2(\varepsilon)}{2} - r + \alpha(\varepsilon)\right)\frac{\partial f}{\partial x} - \lambda(\varepsilon)f(x) \\ + \int_{|y|\geq\varepsilon}\nu(dy)f(x+y), \end{aligned} \tag{5.16}$$

and $\alpha(\varepsilon) = \int_{|y|\geq\varepsilon}(e^y-1)\nu(\mathrm{d}y)$, $\lambda(\varepsilon) = \int_{|y|\geq\varepsilon}\nu(\mathrm{d}y)$.

Note that even if $\sigma = 0$, L^ε contains a nonzero diffusion term, as in the vanishing viscosity method, with the difference that we also have additional terms in the first- and zeroth-order terms in order to conserve the martingale property. The next theorem gives an estimate of the rate of convergence for this "compensated vanishing viscosity" approximation.

THEOREM 5.1. *Let h be Lipschitz: $|h(x)-h(y)| \leq c\,|x-y|$. Let u and u^ε be defined by* (4.7) *and* (5.14)*, respectively. Then*

$$|u(\tau,x) - u^\varepsilon(\tau,x)| \leq C\frac{\int_{-\varepsilon}^{\varepsilon}|y|^3\nu(\mathrm{d}y)}{\sigma^2(\varepsilon)}. \tag{5.17}$$

Proof. We essentially use [11, Proposition 6.2] with the only difference being that we also adjust the drift parameter γ to preserve the martingale property. Let us define $Z_\tau = Y_\tau - (\gamma-\gamma(\varepsilon))\tau$. Then,

$$\begin{aligned} |u(\tau,x) - u^\varepsilon(\tau,x)| &= |\mathbb{E}[h(x+Y_\tau)] - \mathbb{E}[h(x+Y_\tau^\varepsilon)]| \\ &\leq |\mathbb{E}[h(x+Z_\tau)] - \mathbb{E}[h(x+Y_\tau^\varepsilon)]| + \\ &\quad + |\mathbb{E}[h(x+Z_\tau+(\gamma-\gamma(\varepsilon))\tau)] - \mathbb{E}[h(x+Z_\tau)]|. \end{aligned} \tag{5.18}$$

Since h is Lipschitz, it is almost everywhere differentiable with $|h'| \leq c$. By [11, Proposition 6.2] we have

$$|\mathbb{E}[h(x+Z_\tau)] - \mathbb{E}[h(x+Y_\tau^\varepsilon)]| \leq K\,c\,\frac{\int_{-\varepsilon}^{\varepsilon}|y|^3\nu(\mathrm{d}y)}{\sigma^2(\varepsilon)} \tag{5.19}$$

with $K < 16.5$. The second term may be estimated as follows:

$$|\mathbb{E}[h(x+Z_\tau+(\gamma-\gamma(\varepsilon))\tau)] - \mathbb{E}[h(x+Z_\tau)]| \leq c\,|\gamma-\gamma(\varepsilon)|\tau,$$

where

$$\begin{aligned} |\gamma-\gamma(\varepsilon)| &= \left|\frac{\sigma^2(\varepsilon)}{2} - \int_{|y|<\varepsilon}(e^y-1-y)\nu(\mathrm{d}y)\right| \\ &= \left|\frac{1}{2}\int_{|y|<\varepsilon}\nu(\mathrm{d}y)\int_0^y e^s(y-s)^2ds\right| \leq \frac{e^\varepsilon}{6}\int_{-\varepsilon}^{\varepsilon}|y|^3\nu(\mathrm{d}y). \end{aligned} \tag{5.20}$$

Since $\sigma^2(\varepsilon) \to 0$ as $\varepsilon \to 0$, (5.20) converges faster than (5.19) and therefore may be neglected. □

Remark 5. If $\lim_{x\to 0} \nu(x)|x|^{1+\beta} = a > 0$, with $0 \le \beta < 2$, then (5.17) gives

$$|u(\tau, x) - u^{\varepsilon}(\tau, x)| \le C(\beta)\varepsilon,$$

so the approximation error is proportional to ε. This case includes all practical examples used in option pricing such as variance Gamma, NIG, and tempered stable processes.

6. Convergence.

We study in this section the convergence of the finite difference scheme presented above. In the usual approach to the convergence of finite difference schemes for PDEs, consistency and stability ensure convergence under regularity assumptions on the solution such as uniform boundedness of the derivatives. This approach is not feasible here because, even for a European put option, the second derivative (Gamma of the option) is *never* uniformly bounded in t.

6.1. Monotonicity.

Monotonicity is an important property for pricing applications: it guarantees that a (discrete) comparison principle holds for the numerical solution so arbitrage inequalities will be verified exactly and not only as $\Delta t, \Delta x \to 0$, leading to arbitrage-free approximations. The following result shows that the scheme above is monotone.

Proposition 6.1 (monotonicity). *Scheme* (S) *is monotone: if* u^0, v^0 *are two bounded initial conditions, then*

$$u^0 \ge v^0 \quad \Rightarrow \quad \forall n \ge 1,\ u^n \ge v^n.$$

Proof. We start by rewriting (5.12) in the following form:

$$-c\Delta t u_{i-1}^{n+1} + (1 + a\Delta t)u_i^{n+1} - b\Delta t u_{i+1}^{n+1} = u_i^n + \Delta t \sum_j \nu_j u_{i+j}^n, \tag{6.1}$$

where[2]

$$\begin{aligned} a &= \frac{\sigma^2}{(\Delta x)^2} - \left(\frac{\sigma^2}{2} - r + \hat{\alpha}\right)\frac{1}{\Delta x} + \hat{\lambda} \ge 0, \\ b &= \frac{\sigma^2}{2(\Delta x)^2} - \left(\frac{\sigma^2}{2} - r + \hat{\alpha}\right)\frac{1}{\Delta x} \ge 0, \\ c &= \frac{\sigma^2}{2(\Delta x)^2} \ge 0. \end{aligned} \tag{6.2}$$

Notice that $a = b + c + \hat{\lambda}$.

Let u^n and v^n be two solutions of (S) corresponding to the initial conditions $h(x)$ and $f(x)$, respectively, and let $h(x) \ge f(x)$ for all $x \in \mathbb{R}$. Let us show by induction that $w^n = u^n - v^n \ge 0$ for all $n \ge 0$. We have $w_i^0 = h(x_i) - f(x_i) \ge 0$ for all $i \in \mathbb{Z}$. Let $w^n \ge 0$ and suppose that $\inf_{i\in\mathbb{Z}} w_i^{n+1} < 0$. Since for all $i \in \mathbb{Z}\backslash\{0, \dots, N\}$,

[2] We recall that the case $\sigma^2/2 - r + \hat{\alpha} < 0$ is considered. If $\sigma^2/2 - r + \hat{\alpha} \ge 0$, we change the approximation of the first-order derivative (see (5.7)) to have $a, b, c \ge 0$, which is needed for stability and monotonicity.

$w_i^{n+1} = h(x_i) - f(x_i) \geq 0$, this implies that there exists $i_0 \in \{0, \dots, N\}$, such that $w_{i_0}^{n+1} = \inf_{i\in\mathbb{Z}} w_i^{n+1}$. Using (6.1) we obtain that

$$\begin{aligned}\inf_{i\in\mathbb{Z}} w_i^{n+1} = w_{i_0}^{n+1} &= -c\Delta t w_{i_0}^{n+1} + (1 + a\Delta t)w_{i_0}^{n+1} - b\Delta t w_{i_0}^{n+1} - \hat{\lambda}\Delta t w_{i_0}^{n+1}\\ &\geq -c\Delta t w_{i_0-1}^{n+1} + (1 + a\Delta t)w_{i_0}^{n+1} - b\Delta t w_{i_0+1}^{n+1}\\ &= w_{i_0}^n + \Delta t \sum_j \nu_j w_{i_0+j}^n \geq 0, \qquad (6.3)\end{aligned}$$

which contradicts the assumption. Therefore, $\inf_{i\in\mathbb{Z}} w_i^{n+1} \geq 0$, and, consequently $w^{n+1} \geq 0$. □

It is convenient to rewrite the scheme as follows:

$$\begin{aligned}u(\tau_n, x_i) &= F_\Delta[u(\tau_n - \Delta t, \cdot)](x_i), \quad n = 1, \dots, M,\ i \in \{0, \dots, N\},\\ u(\tau_n, x_i) &= g(\tau_n, x_i), \quad n = 0, \dots, M,\ i \notin \{0, \dots, N\},\\ u(0, x_i) &= h(x_i), \quad i \in \{0, \dots, N\}.\end{aligned} \qquad (6.4)$$

Let u^Δ be the solution of the scheme (6.4) defined on the grid $Q_\Delta = \{(\tau_n, x_i) \mid n = 0, \dots, M,\ i \in \mathbb{Z}\}$. One can define super- and subsolutions for the scheme by analogy with the definitions in section 3.

DEFINITION 6.2. *A function w^Δ defined on Q_Δ is a supersolution of the scheme* (6.4) *if*

$$\begin{aligned}w^\Delta(\tau_n, x_i) &\geq F_\Delta[w^\Delta(\tau_n - \Delta t, \cdot)](x_i), \quad n = 1, \dots, M,\ i \in \{0, \dots, N\},\\ w^\Delta(\tau_n, x_i) &\geq g(\tau_n, x_i), \quad n = 0, \dots, M,\ i \notin \{0, \dots, N\},\\ w^\Delta(0, x_i) &\geq h(x_i), \quad i \in \{0, \dots, N\}.\end{aligned}$$

A function z^Δ on Q_Δ is a subsolution of (6.4) *if*

$$\begin{aligned}z^\Delta(\tau_n, x_i) &\leq F_\Delta[z^\Delta(\tau_n - \Delta t, \cdot)](x_i), \quad n = 1, \dots, M,\ i \in \{0, \dots, N\},\\ z^\Delta(\tau_n, x_i) &\leq g(\tau_n, x_i), \quad n = 0, \dots, M,\ i \notin \{0, \dots, N\},\\ z^\Delta(0, x_i) &\leq h(x_i), \quad i \in \{0, \dots, N\}.\end{aligned}$$

In particular, the scheme is unconditionally stable in the sup norm. The following result extends the discrete comparison principle to sub- and supersolutions.

LEMMA 1. *For any supersolution w and any subsolution z of* (6.4) *we have*

$$z^\Delta \leq u^\Delta \leq w^\Delta \quad \text{on } Q_\Delta.$$

Proof. For $n = 0$ or $i \notin \{0, \dots, N\}$ the above inequalities are satisfied by definition. For $n > 0$, $i \in \{0, \dots, N\}$, they follow directly from the monotonicity of the scheme. Indeed, if $z^\Delta(\tau_n - \Delta t, \cdot) \leq u^\Delta(\tau_n - \Delta t, \cdot) \leq w^\Delta(\tau_n - \Delta t, \cdot)$, then, for $i \in \{0, \dots, N\}$, we obtain

$$\begin{aligned}z^\Delta(\tau_n, x_i) &\leq F_\Delta[z^\Delta(\tau_n - \Delta t, \cdot)](x_i) \leq F_\Delta[u^\Delta(\tau_n - \Delta t, \cdot)](x_i) = u^\Delta(\tau_n, x_i)\\ &= F_\Delta[u^\Delta(\tau_n - \Delta t, \cdot)](x_i) \leq F_\Delta[w^\Delta(\tau_n - \Delta t, \cdot)](x_i) \leq w^\Delta(\tau_n, x_i). \qquad □\end{aligned}$$

6.2. Consistency. We now show the consistency of the scheme in the uniform norm for the following class of test functions $v : [0,T] \times \mathbb{R} \mapsto \mathbb{R}$:

$$H = \Big\{ v \in C^{1,2}((0,T] \times O), v \text{ uniformly continuous on}[0,T] \times \mathbb{R},$$

$$\frac{\partial v}{\partial \tau}, \frac{\partial v}{\partial x}, \frac{\partial^2 v}{\partial x^2} \text{ uniformly continuous on}(0,T] \times O \Big\}. \tag{6.5}$$

Proposition 6.3 (consistency in the uniform norm). *For any $v \in H$ and any $\varepsilon > 0$,*

$$\exists \Delta > 0, \quad \left| \frac{v(\tau_n, x_i) - v(\tau_{n-1}, x_i)}{\Delta t} - L_\Delta v(\tau_n, x_i) - \left(\frac{\partial v}{\partial \tau} - Lv \right)(\tau, x) \right| < \varepsilon \tag{6.6}$$

for all Δt, $\Delta x > 0$, $(\tau, x) \in (0,T] \times O$, $n \geq 1$, $i \in \{0, \ldots, N\}$, *such that* $\sup\{\Delta t, \Delta x, |\tau_n - \tau|, |x_i - x|\} < \Delta$.

Proof. Let $\sup\{\Delta t, \Delta x, |\tau_n - \tau|, |x_i - x|\} < \Delta$. We have to prove that the expression in (6.6) is bounded by $\alpha(\Delta)$ independently of $(\tau, x), (\tau_n, x_i) \in (0,T] \times O$, such that $\alpha(\Delta) \to 0$ as $\Delta \to 0$. We have

$$\left| \frac{v(\tau_n, x_i) - v(\tau_{n-1}, x_i)}{\Delta t} - \frac{\partial v}{\partial \tau}(\tau, x) \right| = \left| \frac{1}{\Delta t} \int_{\tau_{n-1}}^{\tau_n} \left(\frac{\partial v}{\partial \tau}(t, x_i) - \frac{\partial v}{\partial \tau}(\tau, x) \right) dt \right| \tag{6.7}$$

$$\leq \sup_{t \in (\tau_{n-1}, \tau_n)} \left| \frac{\partial v}{\partial \tau}(t, x_i) - \frac{\partial v}{\partial \tau}(\tau, x) \right| \leq \sup_{\substack{t,\, \tau \,\in\, (0,T],\, y,\, x \,\in\, O \\ |t - \tau| \,\leq\, 2\Delta \\ |y - x| \,\leq\, \Delta}} \left| \frac{\partial v}{\partial \tau}(t, y) - \frac{\partial v}{\partial \tau}(\tau, x) \right| \xrightarrow{\Delta \downarrow 0} 0,$$

since $\frac{\partial v}{\partial \tau}$ is uniformly continuous by assumption.

Consider now the terms in Dv (the differential part of Lv). Using Taylor expansion up to the second order we obtain

$$\begin{aligned} &\left| \frac{v(\tau_n, x_{i-1}) - 2v(\tau_n, x_i) + v(\tau_n, x_{i+1})}{\Delta x^2} - \frac{\partial^2 v}{\partial x^2}(\tau, x) \right| \\ &\qquad = \left| \frac{1}{\Delta x^2} \int_{x_{i-1}}^{x_{i+1}} \left(\frac{\partial^2 v}{\partial x^2}(\tau_n, y) - \frac{\partial^2 v}{\partial x^2}(\tau, x) \right) (\Delta x - |x_i - y|) dy \right| \\ &\qquad \leq 2 \sup_{y \in (x_{i-1}, x_{i+1})} \left| \frac{\partial^2 v}{\partial x^2}(\tau_n, y) - \frac{\partial^2 v}{\partial x^2}(\tau, x) \right| \\ &\qquad \leq 2 \sup_{\substack{t,\, \tau \,\in\, (0,T],\, y,\, x \,\in\, O \\ |t - \tau| \,\leq\, \Delta \\ |y - x| \,\leq\, 2\Delta}} \left| \frac{\partial^2 v}{\partial x^2}(t, y) - \frac{\partial^2 v}{\partial x^2}(\tau, x) \right| \xrightarrow{\Delta \downarrow 0} 0, \end{aligned}$$

as $\frac{\partial^2 v}{\partial x^2}$ is uniformly continuous on $(0,T] \times O$. In the same way, we show that

$$\begin{aligned} \left| \frac{v(\tau_n, x_{i+1}) - v(\tau_n, x_i)}{\Delta x} - \frac{\partial v}{\partial x}(\tau, x) \right| &= \left| \frac{1}{\Delta x} \int_{x_i}^{x_{i+1}} \left(\frac{\partial v}{\partial x}(\tau_n, y) - \frac{\partial v}{\partial x}(\tau, x) \right) dy \right| \\ &\leq \sup_{\substack{t,\, \tau \,\in\, (0,T],\, y,\, x \,\in\, O \\ |t - \tau| \,\leq\, \Delta \\ |y - x| \,\leq\, 2\Delta}} \left| \frac{\partial v}{\partial x}(t, y) - \frac{\partial v}{\partial x}(\tau, x) \right| \xrightarrow{\Delta \downarrow 0} 0. \end{aligned}$$

Since $|1 - e^{y_j - y}| \leq \Delta x$ if $\Delta x \leq 1$ and $|y_j - y| \leq \Delta x/2$, we also have

$$
\begin{aligned}
(6.8)\quad |\alpha - \hat{\alpha}| &= \left| \int_{B_l}^{B_r} (e^y - 1)\nu(\mathrm{d}y) - \sum_{j=K_l}^{K_r} (e^{y_j} - 1)\nu_j \right| \\
&= \left| \sum_{j=K_l}^{K_r} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} (1 - e^{y_j - y}) e^y \nu(\mathrm{d}y) \right| \leq \Delta x \left| \sum_{j=K_l}^{K_r} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} e^y \nu(\mathrm{d}y) \right| \\
&= \Delta x \int_{B_l}^{B_r} e^y \nu(\mathrm{d}y) = (\alpha + \lambda)\Delta x.
\end{aligned}
$$

Therefore, using previous estimates and uniform boundedness of $\frac{\partial v}{\partial x}$ on $(0, T] \times O$, we obtain

$$
\begin{aligned}
|(D_\Delta v)(\tau_n, x_i) - (Dv)(\tau, x)| &= \left| \frac{\sigma^2}{2} \left[\frac{v(\tau_n, x_{i-1}) - 2v(\tau_n, x_i) + v(\tau_n, x_{i+1})}{\Delta x^2} - \frac{\partial^2 v}{\partial x^2}(\tau, x) \right] \right. \\
&\left. + \left(\frac{\sigma^2}{2} - r + \hat{\alpha} \right) \left[\frac{v(\tau_n, x_{i+1}) - v(\tau_n, x_i)}{\Delta x} - \frac{\partial v}{\partial x}(\tau, x) \right] + (\alpha - \hat{\alpha}) \frac{\partial v}{\partial x}(\tau, x) \right| \overset{\Delta \downarrow 0}{\longrightarrow} 0.
\end{aligned}
$$

The integral part can be estimated as follows:

$$
\begin{aligned}
&(6.9)\\
&|(J_\Delta v)(\tau_{n-1}, x_i) - (Jv)(\tau, x)| = \left| \sum_{j=K_l}^{K_r} v(\tau_{n-1}, x_i + y_j)\nu_j - \int_{B_l}^{B_r} v(\tau, x + y)\nu(\mathrm{d}y) \right| \\
&\qquad = \left| \sum_{j=K_l}^{K_r} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} (v(\tau_{n-1}, x_i + y_j) - v(\tau, x + y))\nu(\mathrm{d}y) \right| \\
&\qquad \leq \lambda \Delta \sup_{\substack{\theta_1, \theta_2 \in [0,T],\ \xi_1, \xi_2 \in \mathbb{R} \\ |\theta_1 - \theta_2| \leq 2\Delta \\ |\xi_1 - \xi_2| \leq 3\Delta/2}} |v(\theta_1, \xi_1) - v(\theta_2, \xi_2)| \overset{\Delta \downarrow 0}{\longrightarrow} 0,
\end{aligned}
$$

since v is uniformly continuous on $[0, T] \times \mathbb{R}$. Combining all these estimates gives (6.6). □

6.3. Convergence. A general approach for obtaining convergence of finite difference schemes in the case of nonsmooth solutions was developed by Barles and Souganidis [6] who showed that, when subsolutions/supersolutions verify a strong comparison principle, the solution of any monotone, locally consistent scheme converges uniformly on compact sets to the solution of the limit equation. The idea of [6] is to show that the pointwise limits

$$
(6.10)\qquad \underline{u}(\tau, x) = \liminf_{(\Delta t, \Delta x) \to 0\ (\tau_n, x_i) \to (\tau, x)} u^{(\Delta t, \Delta x)}(\tau_n, x_i),
$$

$$
(6.11)\qquad \overline{u}(\tau, x) = \limsup_{(\Delta t, \Delta x) \to 0\ (\tau_n, x_i) \to (\tau, x)} u^{(\Delta t, \Delta x)}(\tau_n, x_i)
$$

define sub-/supersolutions and then apply the comparison principle to conclude $\overline{u} = \underline{u} = u$. As noted above, our scheme (5.12) is monotone and locally consistent, but the comparison principles [1, 20] available for the PIDEs considered here require uniform continuity properties which may not hold for $\overline{u}$ and $\underline{u}$ defined as in (6.10)–(6.11). Therefore we cannot directly apply the Barles–Souganidis result to obtain convergence.[3]

We avoid this problem by considering smooth sub-/supersolutions and deriving inequalities linking them with $\overline{u}, \underline{u}$.

Definition 6.4. *A function $w \in H$ is a smooth supersolution of the problem* (5.1)–(5.3) *if it verifies the following inequalities:*

$$\frac{\partial w}{\partial \tau}(\tau, x) - Lu(\tau, x) \geq 0, \quad (\tau, x) \in (0, T] \times O, \tag{6.12}$$

$$w(\tau, x) \geq g(\tau, x), \quad x \in O^c, \tag{6.13}$$

$$w(0, x) \geq h(x), \quad x \in O. \tag{6.14}$$

A function $z \in H$ is a smooth subsolution of the problem (5.1)–(5.3) *if it satisfies* (6.12)–(6.14) *with the reverse inequalities.*

Lemma 2. *Let $w(\tau, x)$ be a smooth supersolution and let $z(\tau, x)$ be a smooth subsolution of the problem* (5.1)–(5.3). *Then for all $\varepsilon > 0$, there exists $\Delta > 0$ such that*

$$\forall \Delta t, \Delta x \leq \Delta, \forall n \geq 0, \forall i \in \mathbb{Z}, \; z(\tau_n, x_i) - \varepsilon < u^{\Delta}(\tau_n, x_i) < w(\tau_n, x_i) + \varepsilon. \tag{6.15}$$

Proof. Choose b such that $0 < b(T+1) < \varepsilon$ and let $\bar{w}(\tau, x) = w(\tau, x) + b(\tau + 1)$. If $i \notin \{0, \ldots, N\}$, we have

$$\bar{w}(\tau_n, x_i) = w(\tau_n, x_i) + b(\tau_n + 1) \geq g(\tau_n, x_i) + b(\tau_n, x_i) \geq g(\tau_n, x_i). \tag{6.16}$$

If $i \in \{0, \ldots, N\}$,

$$\bar{w}(0, x_i) = w(0, x_i) + b \geq h(x_i) + b \geq h(x_i). \tag{6.17}$$

If $n \geq 1$, $i \in \{0, \ldots, N\}$, we obtain by Proposition 6.3

$$\begin{aligned} &\frac{\bar{w}(\tau_n, x_i) - \bar{w}(\tau_n - \Delta t, x_i)}{\Delta t} - L_\Delta \bar{w}(\tau_n, x_i) \\ &\quad = \frac{w(\tau_n, x_i) - w(\tau_n - \Delta t, x_i)}{\Delta t} - L_\Delta w(\tau_n, x_i) + b(1 + \hat{\lambda}\Delta t) \\ &\qquad\qquad \to \left(\frac{\partial w}{\partial \tau}(\tau, x) - Lu(\tau, x) \right) + b > 0, \end{aligned} \tag{6.18}$$

as $\Delta t, \Delta x \to 0$, $(\tau_n, x_i) \to (\tau, x)$, uniformly on $(0, T] \times O$. Therefore, for any sufficiently small $\Delta > 0$, for all $\Delta t, \Delta x \leq \Delta$, we have

$$\frac{\bar{w}(\tau_n, x_i) - \bar{w}(\tau_n - \Delta t, x_i)}{\Delta t} - L_\Delta \bar{w}(\tau_n, x_i) \geq 0$$

or, equivalently,

$$\bar{w}(\tau_n, x_i) \geq F_\Delta[\bar{w}(\tau_n - \Delta t, \cdot)](x_i) \qquad \forall n \geq 1, \; \forall i \in \{0, \ldots, N\}. \tag{6.19}$$

[3] Note also that the Barles–Souganidis method does not yield a rate of convergence either: this issue is treated in section 6.4 below.

By (6.16), (6.17), and (6.19), the function $\bar{w}$ is a supersolution of (6.4). So, Lemma 1 implies that

$$u^{\Delta}(\tau_n, x_i) \le \bar{w}(\tau_n, x_i) \le w(\tau_n, x_i) + b(T+1) < w(\tau_n, x_i) + \varepsilon \qquad \forall n \ge 0,\ \forall i \in \mathbb{Z},$$

which is the desired property. The lower bound $z(\tau_n, x_i) - \varepsilon$ can be proved in the same manner. □

Corollary 6.5. *Let $\underline{u}$ and $\overline{u}$ be the functions defined by* (6.10)–(6.11). *For any smooth subsolution $z(\tau, x)$ and any supersolution $w(\tau, x)$ of the problem* (5.1)–(5.3), *we have, for all $(\tau, x) \in [0, T] \times \mathbb{R}$,*

$$z(\tau, x) \le \underline{u}(\tau, x) \le \overline{u}(\tau, x) \le w(\tau, x). \tag{6.20}$$

Proof. By the definition of upper and lower limits, (6.15) implies (6.20). □

We are now ready to give our main result on the convergence of the scheme. We assume $\sigma > 0$ since, as explained in section 5.2, the scheme introduces a viscosity term $\sigma(\varepsilon) > 0$ even in the pure jump case where $\sigma = 0$.

Theorem 6.6. *In the European case ($O = \mathbb{R}$) with $\sigma > 0$, if h is a bounded piecewise continuous function on $\mathbb{R}$, then for all $\tau > 0$ and all $x \in \mathbb{R}$, the discrete solution converges to the solution of the continuous problem:*

$$\lim_{\substack{(\Delta t, \Delta x) \to 0 \\ (\tau_n, x_i) \to (\tau, x)}} u^{(\Delta t, \Delta x)}(\tau_n, x_i) = u(\tau, x).$$

Proof. If $\underline{h}, \overline{h} \in C^\infty(\mathbb{R})$ are such that $\underline{h} \le h \le \overline{h}$, then $z(\tau, x) = \mathbb{E}[\underline{h}(x + Y_\tau)]$ and $w(\tau, x) = \mathbb{E}[\overline{h}(x + Y_\tau)]$ are solutions in H of (5.1) and therefore, respectively, a subsolution and a supersolution of the Cauchy problem (5.1), (5.3). By Corollary 6.5, we obtain (6.20). If $w(\tau, x) - u(\tau, x)$ and $u(\tau, x) - z(\tau, x)$ could be made arbitrarily small, this would imply that there exists a limit of $u^{(\Delta t, \Delta x)}(\tau_n, x_i)$ equal to $u(\tau, x)$. So, it remains to construct appropriate smooth approximations $\underline{h}$ and $\overline{h}$ of h.

Let $\xi_1, \dots, \xi_I$ be the discontinuity points of h. We will suppose that the jumps of h are bounded by K. Given $\varepsilon > 0$, we choose $\underline{h}, \overline{h} \in C^\infty(\mathbb{R})$ such that

$$\begin{aligned} \underline{h}(x) \le h(x) \le \overline{h}(x) \quad & \forall x \in \mathbb{R}, \\ |\overline{h}(x) - \underline{h}(x)| \le \varepsilon \quad & \forall x \notin \bigcup_{i=1}^{I} (\xi_i - \varepsilon, \xi_i + \varepsilon), \\ |\overline{h}(x) - \underline{h}(x)| \le K \quad & \forall x \in \bigcup_{i=1}^{I} (\xi_i - \varepsilon, \xi_i + \varepsilon). \end{aligned}$$

Then we have

$$\begin{aligned} w(\tau, x) - z(\tau, x) &= \mathbb{E}[\overline{h}(x + Y_\tau) - \underline{h}(x + Y_\tau)] \\ &\le \varepsilon \mathbb{Q}\left(x + Y_\tau \notin \bigcup_{i=1}^{I} (\xi_i - \varepsilon, \xi_i + \varepsilon) \right) + K \mathbb{Q}\left(x + Y_\tau \in \bigcup_{i=1}^{I} (\xi_i - \varepsilon, \xi_i + \varepsilon) \right) \\ &\le \varepsilon + K \mathbb{Q}\left(x + Y_\tau \in \bigcup_{i=1}^{I} (\xi_i - \varepsilon, \xi_i + \varepsilon) \right). \end{aligned} \tag{6.21}$$

Denoting $\Omega_\varepsilon = \{x + Y_\tau \in \bigcup_{i=1}^{I}(\xi_i - \varepsilon, \xi_i + \varepsilon)\}$ we obtain $\bigcap_{\varepsilon>0}\Omega_\varepsilon = \{x + Y_\tau \in \{\xi_1, \dots, \xi_I\}\}$. Since σ is strictly positive, Y_τ has an absolutely continuous distribution for $\tau > 0$, so we have $\mathbb{Q}(x + Y_\tau \in \{\xi_1, \dots, \xi_I\}) = 0$. Consequently,

$$\mathbb{Q}\left(x + Y_\tau \in \bigcup_{i=1}^{I}(\xi_i - \varepsilon, \xi_i + \varepsilon)\right) \xrightarrow{\varepsilon\downarrow 0} 0.$$

We have shown that the expression in (6.21) goes to zero as $\varepsilon \to 0$. The inequalities (6.20) together with $z(\tau, x) \le u(\tau, x) \le w(\tau, x)$ then imply that $\underline{u}(\tau, x) = \overline{u}(\tau, x) = u(\tau, x)$. The proof is complete. □

Remark 6. If $\tau = 0$, we obtain

$$\mathbb{Q}\left(x + Y_\tau \in \bigcup_{i=1}^{I}(\xi_i - \varepsilon, \xi_i + \varepsilon)\right) \xrightarrow{\varepsilon\downarrow 0} \mathbb{Q}(x \in \{\xi_1, \dots, \xi_I\}) = 1_{\{x\in\{\xi_1,\dots,\xi_I\}\}}.$$

So, the scheme does not converge to the initial condition at the discontinuity points of h. However, this has no practical importance since we do not need to calculate the solution numerically at $\tau = 0$.

Together with Proposition 5.1, Theorem 6.6 allows us to compute option prices in all the examples of exponential Lévy models given in section 2. Theorem 6.6 does not yield the rate of convergence; we now use a different approach to obtain an estimate on the rate of convergence under a further condition on the scheme.

6.4. Rate of convergence. In the viscosity solution framework, results on convergence rates for numerical schemes have been obtained by Crandall and Lions [15] for first-order equations and by Krylov [21, 22] for second-order parabolic PDEs; see also [5]. Our approach is inspired by [22] but is somewhat simpler and yields better bounds, given that we are dealing with a linear equation.

First, let us write the scheme in the following form: $u(\tau_n, x_i) = u(\tau_{n-1}, x_i) + L_\Delta u(\tau_n, x_i)\Delta t$, where (in the case $\sigma^2/2 - r + \hat{\alpha} < 0$)

$$\begin{aligned} L_\Delta u(\tau, x) = {}& \frac{\sigma^2}{2}\frac{u(\tau, x + \Delta x) - 2u(\tau, x) + u(\tau, x - \Delta x)}{\Delta x^2} \\ & - \left(\frac{\sigma^2}{2} - r + \hat{\alpha}\right)\frac{u(\tau, x + \Delta x) - u(\tau, x)}{\Delta x} - \hat{\lambda} u(\tau, x) + \sum_{j=K_l}^{K_r} \nu_j u(\tau - \Delta t, x + j\Delta x). \end{aligned}$$

As previously, we assume $\sigma > 0$.[4]

Consider the grid $\{(\tau_n, x_i),\ n = 0, \dots, M,\ i \in \mathbb{N}\}$, where $\tau_n = n\Delta t$, $x_i = x_0 + i\Delta x$. Given $J \subset \{0, \dots, M\}$, we denote by $\mathcal{B}_{J\times\mathbb{N}}$ the space of bounded functions on $\{(\tau_n, x_i),\ n \in J,\ i \in \mathbb{N}\}$ with the norm

$$\|v\|_J = \sup_{n\in J,\ i\in\mathbb{N}} |v(\tau_n, x_i)|. \tag{6.22}$$

For $(v(x_i),\ i \in \mathbb{N})$ we write

$$\|v\| = \sup_{i\in\mathbb{N}} |v(x_i)|. \tag{6.23}$$

The following (see appendix) is a slightly improved version of a result in [22].

[4] The infinite activity pure-jump case being reduced to this one; see section 5.2.

LEMMA 3. *Consider, for $\xi \in \mathcal{B}_{\{1,\dots,M\}\times\mathbb{N}}$, the problem*

$$v(0,x_i) = h(x_i), \qquad i \in \mathbb{N}, \tag{6.24}$$

$$v(\tau_n,x_i) = v(\tau_{n-1},x_i) + L_\Delta v(\tau_n,x_i)\Delta t + \xi(\tau_n,x_i), \quad n=1,\dots,M,\ i\in\mathbb{N}. \tag{6.25}$$

(i) *For any $\xi \in \mathcal{B}_{\{1,\dots,M\}\times\mathbb{N}}$ the problem* (6.24)–(6.25) *has a unique solution $v(\xi) \in \mathcal{B}_{\{0,\dots,M\}\times\mathbb{N}}$.*

(ii) *If $v_i = v(\xi_i)$, where $\xi_i \in \mathcal{B}_{\{1,\dots,M\}\times\mathbb{N}}$, $i=1,2$, then*

$$\|v_1(\tau_n,\cdot) - v_2(\tau_n,\cdot)\| \le \sum_{k=1}^{n} \|\xi_1(\tau_k,\cdot) - \xi_2(\tau_k,\cdot)\|, \quad n=1,\dots,M. \tag{6.26}$$

We will consider payoff functions h which verify the following conditions.

ASSUMPTION 6.1. *$h : \mathbb{R} \mapsto \mathbb{R}$ is continuous on $\mathbb{R}$ and there exists $\xi_1,\dots,\xi_I \in \mathbb{R}$ such that h is C^∞ on $]\xi_i,\xi_{i+1}[$ for $i=0,\dots,I-1$ and for all $n\ge 0$, for all $x \notin \{\xi_1,\dots,\xi_I\}$, $|h^{(n)}(x)| \le K$.*

For example, the payoff of a put option $h(x) = (1-e^x)^+$ verifies these conditions.

LEMMA 4. *Let h verify Assumption* 6.1 *and let $u(\tau,x) = \mathbb{E}[h(x+X_\tau)]$ be the viscosity solution of the problem*

$$\frac{\partial u}{\partial \tau}(\tau,x) = Lu(\tau,x), \qquad (\tau,x) \in (0,T]\times\mathbb{R}, \tag{6.27}$$

$$u(0,x) = h(x), \qquad x \in \mathbb{R}. \tag{6.28}$$

Then, for all $\tau > 0$ and $m,n \in \mathbb{N}$, $m+n>0$, we have

$$\left\| \frac{\partial^{m+n} u}{\partial \tau^m \partial x^n}(\tau,\cdot) \right\| \le \frac{C}{(\sqrt{\tau})^{2m+n-1}}. \tag{6.29}$$

The constant C depends only on m, n, K, T, and coefficients of the operator L (σ, r, α, and λ).

Using this lemma we prove the following consistency result.

LEMMA 5. *If u solves* (6.27)–(6.28)*, then, for all $k \ge 1$,*

$$\|Lu(\tau_k,\cdot) - L_\Delta u(\tau_k,\cdot)\| \le C\left[\frac{\Delta x^2}{\tau_k^{3/2}} + \frac{\Delta x}{\tau_k^{1/2}} + \Delta x + 2(\sqrt{\tau_k} - \sqrt{\tau_{k-1}})\right], \tag{6.30}$$

with $C = C(K,T,r,\sigma,\lambda,\alpha)$.

Under a CFL-type condition, we obtain the following rate of convergence.

THEOREM 6.7. *Assume that the initial condition h verifies Assumption* 6.1. *Let u be the unique solution of* (6.27)–(6.28) *and let u_Δ be the solution of* (6.24)–(6.25) *with $\xi = 0$. If $c_1 \le \Delta t/\Delta x^2 \le c_2$, then*

$$\|u - u_\Delta\|_{\{0,\dots,M\}} \le C\Delta x, \tag{6.31}$$

where C depends only on T, K, and coefficients of the operator L (σ, r, α, and λ).

Proof. The function $u(\tau_n,x_i)$ solves (6.24)–(6.25) with $\xi(\tau_k,x_i) = \int_{\tau_{k-1}}^{\tau_k} Lu(s,x_i)ds - L_\Delta u(\tau_k,x_i)\Delta t$. From Lemma 3,

$$\begin{aligned}\|u(\tau_n,\cdot) - u_\Delta(\tau_n,\cdot)\| &\le \sum_{k=1}^{n} \left\| \int_{\tau_{k-1}}^{\tau_k} Lu(s,\cdot)ds - L_\Delta u(\tau_k,\cdot)\Delta t \right\| \\ &\le \sum_{k=1}^{n} \left\| \int_{\tau_{k-1}}^{\tau_k} (Lu(s,\cdot) - Lu(\tau_k,\cdot))ds \right\| + \sum_{k=1}^{n} \|Lu(\tau_k,\cdot) - L_\Delta u(\tau_k,\cdot)\|\,\Delta t.\end{aligned} \tag{6.32}$$

Let us look at the first term:

$$\left|\int_{\tau_{k-1}}^{\tau_k}(Lu(s,x)-Lu(\tau_k,x))ds\right| = \left|\int_{\tau_{k-1}}^{\tau_k}ds\int_s^{\tau_k}\frac{\partial^2 u}{\partial\tau^2}(\theta,x)d\theta\right|$$
$$\leq C\int_{\tau_{k-1}}^{\tau_k}ds\int_s^{\tau_k}\frac{d\theta}{\theta^{3/2}} \leq C\Delta t\int_{\tau_{k-1}}^{\tau_k}\frac{ds}{s^{3/2}}$$

by Lemma 4. Therefore,

$$\sum_{k=1}^{n}\left\|\int_{\tau_{k-1}}^{\tau_k}(Lu(s,\cdot)-Lu(\tau_k,\cdot))ds\right\| \leq C\left[\int_0^{\Delta t}ds\int_s^{\Delta t}\frac{d\theta}{\theta^{3/2}}+\Delta t\int_{\Delta t}^{T}\frac{ds}{s^{3/2}}\right]$$
$$= C\left[2\Delta t+\Delta t\frac{2(\sqrt{T}-\sqrt{\Delta t})}{\sqrt{T}\sqrt{\Delta t}}\right] \leq 4C\sqrt{\Delta t}. \tag{6.33}$$

To estimate the second term in (6.32), note that

$$\sum_{k=1}^{n}\frac{\Delta t}{\sqrt{\tau_k}} = \sum_{k=1}^{n}\int_{\tau_{k-1}}^{\tau_k}\frac{ds}{\sqrt{\tau_k}} \leq \sum_{k=1}^{n}\int_{\tau_{k-1}}^{\tau_k}\frac{ds}{\sqrt{s}} = 2\sqrt{\tau_n}$$

and

$$\sum_{k=1}^{n}\frac{\Delta t}{\tau_k^{3/2}} = \frac{1}{\sqrt{\Delta t}}\sum_{k=1}^{n}\frac{1}{k^{3/2}} \leq \frac{C}{\sqrt{\Delta t}},$$

since the series $\sum_{k=1}^{\infty}1/k^{3/2}$ converges. Using Lemma 5, we obtain

$$\sum_{k=1}^{n}\|Lu(\tau_k,\cdot)-L_\Delta u(\tau_k,\cdot)\|\Delta t \leq C\left[\Delta x^2/\sqrt{\Delta t}+2\sqrt{\tau_n}\Delta x+\tau_n\Delta x+2\sqrt{\tau_n}\Delta t\right]$$
$$\leq C\left[\Delta x^2/\sqrt{\Delta t}+\Delta x+\Delta t\right]. \tag{6.34}$$

If $c_1 \leq \Delta t/\Delta x^2 \leq c_2$, the estimates (6.32), (6.33), and (6.34) imply (6.31). □

7. Numerical results. We now illustrate the performance of the scheme proposed above with two examples. The computations were done in variance Gamma models with Lévy density

$$\nu(x) = a\frac{\exp(-\eta_\pm|x|)}{|x|}$$

and two sets of parameters: $a = 6.25$, $\eta_- = 14.4$, $\eta_+ = 60.2$ (VG1) and $a = 0.5$, $\eta_- = 2.7$, $\eta_+ = 5.9$ (VG2), and a Merton model with Gaussian jumps in log-price with volatility $\sigma = 15\%$ and Lévy density

$$\nu(x) = 0.1\frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

for a put of maturity of $T = 1$ year. Both of these models satisfy the hypotheses of sections 4.1 and 4.2: the Lévy densities have exponentially decreasing tails. The Black–Scholes implied volatilities computed from prices of put options are shown in Figure 7.1: as expected, they display a "smile" (convex) feature as a function of the

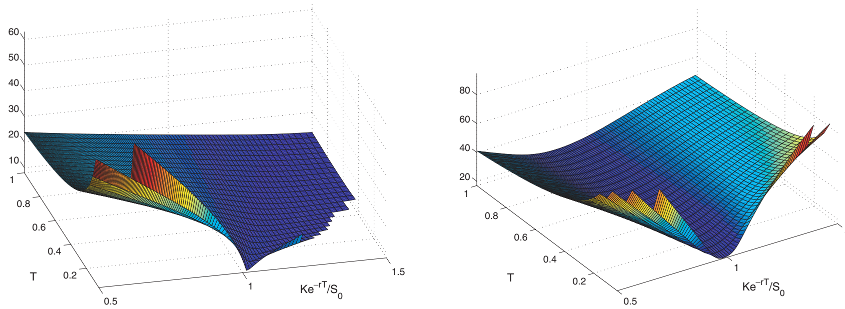


FIG. 7.1. *Black–Scholes implied volatilities for put options, as a function of strike and maturity. Left: variance Gamma model. Right: Merton jump-diffusion model.*

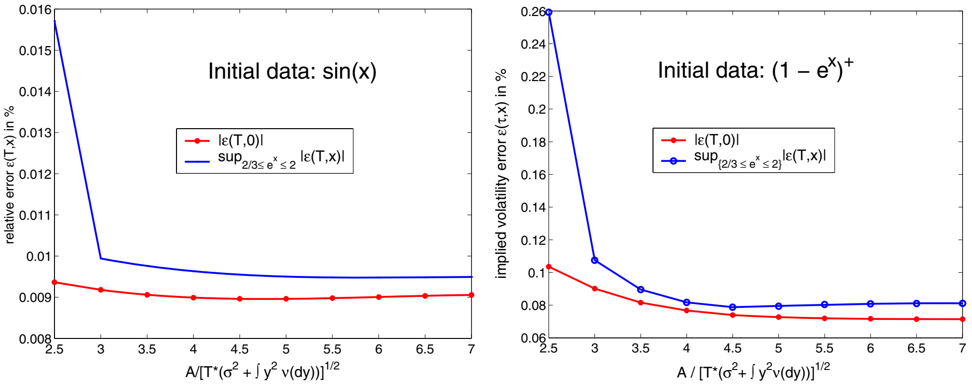


FIG. 7.2. *Influence of domain size on localization error for the explicit-implicit finite difference scheme. Left: smooth initial condition* $h(x) = \sin(x)$ *in Merton jump-diffusion model. Right: put option in Merton jump-diffusion model.*

strike of the option, which flattens out with increasing maturity. These two models have been chosen since in each case there is an alternative method for computing the solution of the PIDE: in the Merton model the solution can be expressed as a series expansion, and in the variance Gamma model a closed form expression for the characteristic function is available, so one can use Carr and Madan's FFT method [9]. Comparing our finite difference solution to these alternative solutions (computed with high precision) allows us to study the behavior of various error terms in relation with the parameters of the scheme.

The error metric which is relevant from the point of view of financial applications is the error in terms of Black–Scholes implied volatility:

$$\varepsilon(\tau, x) = |\Sigma^{\mathrm{PIDE}}(\tau, x) - \Sigma^{\mathrm{FFT}}(\tau, x)| \qquad \text{in} \quad \%,$$

where Σ denotes the Black–Scholes implied volatility computed by inverting the Black–Scholes formula with respect to the volatility parameter and applying it to the computed option price. We have computed both pointwise errors at $x = 0$ (i.e., forward at-the-money options) and uniform errors on the computational range $x \in [\log(2/3), \log(2)]$. This range contains all options prices quoted on the market.

The localization error is shown in Figure 7.2 for the Merton model: domain size A is represented in terms of its ratio to the standard deviation of X_T. An

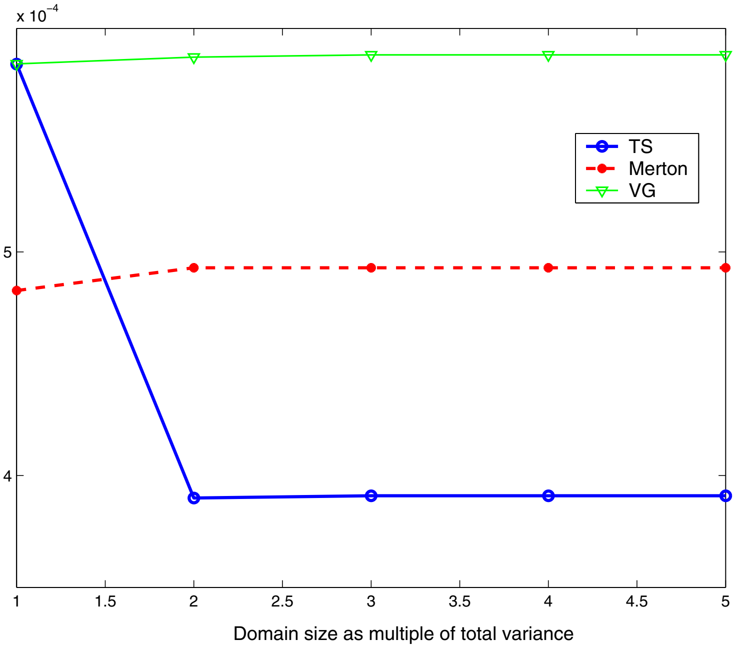


FIG. 7.3. *Influence of domain size on localization error for the explicit-implicit finite difference scheme: put option in tempered stable, Merton, and variance Gamma models.*

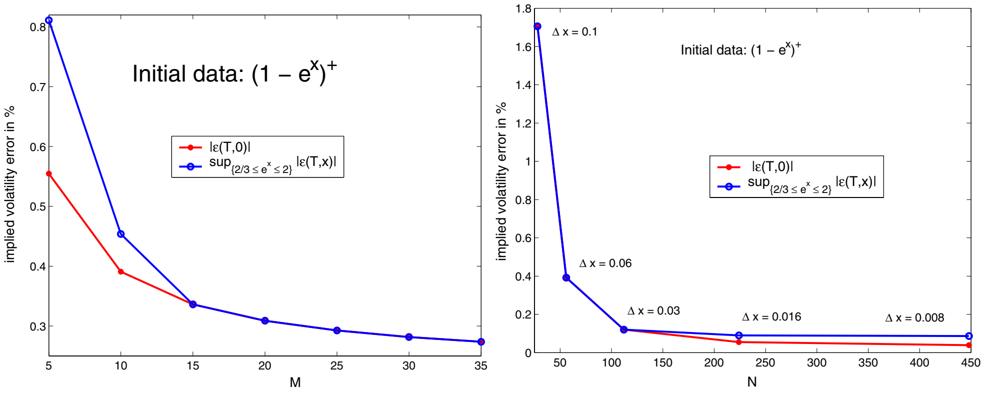


FIG. 7.4. *Numerical accuracy for a put option in the Merton model. Left: influence of number of time steps M, $\Delta x = 0.05$, $\Delta t = T/M$. Right: influence of number of space steps N, $\Delta x = 2A/N$, $\Delta t = 0.02$.*

acceptable level is obtained for values of order $\simeq 5$. Notice that as soon as this ratio is greater than or equal to 3, the uniform and pointwise errors are quite close to each other, indicating that we are out of the zone of influence of the numerical boundary conditions. Figure 7.3 shows the same analysis for the variance Gamma model.

Figure 7.4 illustrates the decay of numerical error when $\Delta t, \Delta x \to 0$, i.e., when the number of time/space steps is increased. The behavior is quite similar to the case of the Black–Scholes model.

Figure 7.5 illustrates the behavior of the error (for a fixed grid size) as a function of maturity for two initial conditions: a smooth one (forward contract) and a

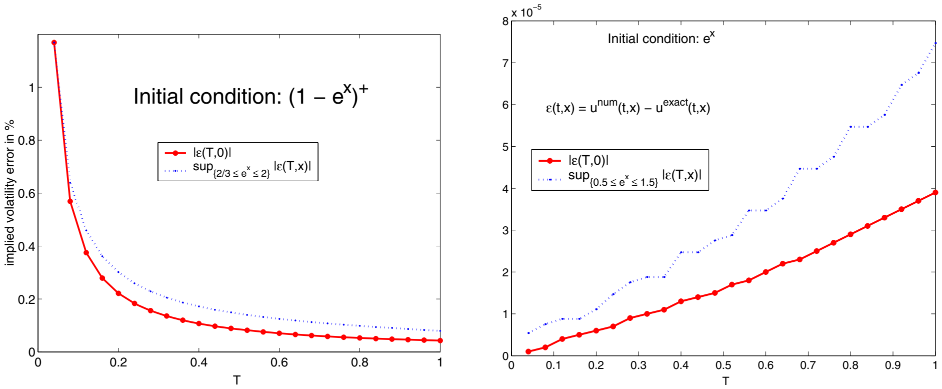


FIG. 7.5. *Decrease of error with maturity in the Merton model. Left: nonsmooth initial condition (put option) $h(x) = (1 - e^x)^+$. Right: smooth initial condition $h(x) = e^x$.*

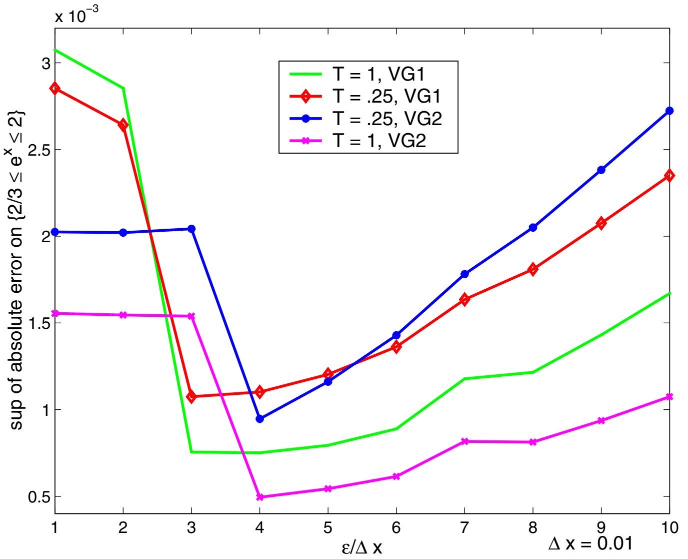


FIG. 7.6. *Influence of truncation of small jumps on numerical error in various variance Gamma models. Put option.*

nonsmooth one (put option). We observe that a nonsmooth initial condition leads to a lack of accuracy for small T. This phenomenon, which is not specific to models with jumps, can be overcome using an irregular time-stepping scheme which exploits the smoothness in time of the solution. Matache, von Petersdorff, and Schwab [25] have suggested using irregularly (logarithmically) spaced time stepping, more refined near maturity, in order to improve this convergence. Note that scheme introduces a "numerical viscosity" $\sigma(\varepsilon)$, so even when the underlying model is a pure-jump Lévy process this numerical diffusion term has a regularizing effect.

In the case of infinite activity models, an additional parameter which influences the solution is the truncation parameter ε for the small jumps. Whereas the error in Proposition 5.1 vanishes when $\varepsilon \to 0$, for a fixed Δx the discretization error *increases* as $\varepsilon \to 0$: the constant in (6.31) is of order $\lambda(\varepsilon)$, which increases with ε. This suggests that there is an optimal choice $\varepsilon(\Delta x) > 0$ for a given Δx. The optimal choice of ε is not universal and depends on the growth of the Lévy density near zero. Figure 7.6

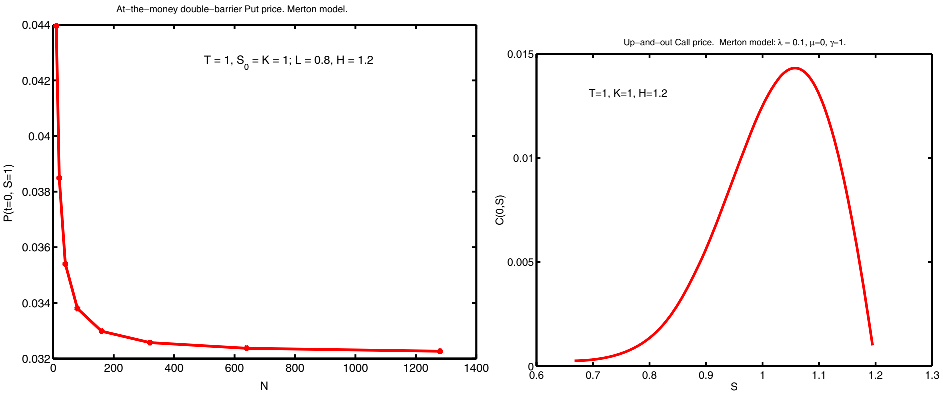


Fig. 7.7. *Left: at-the-money double-barrier put price as a function of the number of space steps. Barrier levels: $L = 0.8$, $H = 1.2$. Right: up-and-out call price in the Merton model. Barrier level: $H = 1.2$.*

Table 7.1

Examples of numerical values for at-the-money option prices. $S = K = 100$, $T = 1$, $A = 5\sqrt{\sigma^2 + \int y^2\nu(dy)}$, $dt = 0.02$, $dx = 0.01$. *The truncation parameter ε is chosen according to Figure* 7.6.

Model	Put	t	Up-and-out call	t	Double-barrier put	t
		sec.	$H = 120$	sec.	$L = 80$, $H = 120$	sec.
VG1	6.72	0.5	2.73	0.2	2.42	0.1
VG2	8.38	0.9	3.34	0.5	1.68	0.1
Merton	11.04	1.2	1.17	0.5	3.35	4

illustrates this phenomenon for the variance Gamma model.

The last two figures give examples of boundary value problems for barrier options in the Merton model. Figure 7.7 (right) shows the price of an up-and-out call. Figure 7.7 (left) illustrates the numerical convergence of a double-barrier put price as the number N of space steps increases.

In Table 7.1, we give some examples of option values obtained with our numerical scheme as well as the corresponding computation time in seconds.

Appendix A. Proofs of lemmas.

A.1. Proof of Lemma 3. To prove (i), we rewrite (6.24)–(6.25) as $v(\tau, x) = \Psi_\Delta v(\tau, x)$, where Ψ_Δ is a contraction, and apply the fixed point theorem. Define

$$p_\Delta = \frac{1}{\Delta t} + \frac{\sigma^2}{\Delta x^2} + \left|\frac{\sigma^2}{2} - r + \hat{\alpha}\right| + \hat{\lambda} + 1 \geq 1$$

such that $\Phi_\Delta v(\tau_n, x_i) = -\frac{v(\tau_n, x_i) - v(\tau_{n-1}, x_i)}{\Delta t} + L_\Delta v(\tau_n, x_i) + p_\Delta v(\tau_n, x_i)$ is monotone: if $v \geq 0$, then $\Phi_\Delta v \geq 0$. Φ_Δ verifies

$$\text{(A.1)} \qquad \Phi_\Delta e^{-2\tau} = e^{-2\tau}\left[p_\Delta - (1 - e^{-2\Delta t})\left(\frac{1}{\Delta t} + \hat{\lambda}\right)\right] \leq e^{-2\tau}(p_\Delta - 1)$$

for small Δt (for example, this property holds as soon as $\Delta t \le 0.7$). We now define $\tilde{v}(\tau,x) = e^{2\tau} v(\tau,x)$, $\tilde{\xi}(\tau,x) = (\Delta t)^{-1} p_\Delta^{-1} e^{2\tau}\xi(\tau,x)$ and rewrite (6.24)–(6.25) as

$$\tilde{v}(0,x_i) = h(x_i), \quad i \in \mathbb{N},$$

$$\text{(A.2)}\quad \tilde{v}(\tau_n,x_i) = p_\Delta^{-1} e^{2\tau_n}\Phi_\Delta[e^{-2\tau_n}\tilde{v}(\tau_n,x_i)] + \tilde{\xi}(\tau_n,x_i), \quad n=1,\dots,M,\ i \in \mathbb{N},$$

or, equivalently, $\tilde{v}(\tau_n,x_i) = \Psi_\Delta \tilde{v}(\tau_n,x_i)$, $\quad n = 0,\dots,M,\ \ i \in \mathbb{N}$, with

$$\Psi_\Delta \tilde{v}(\tau_n,x_i) = 1_{n\in\{1,\dots,M\}}(p_\Delta^{-1} e^{2\tau_n}\Phi_\Delta[e^{-2\tau_n}\tilde{v}(\tau_n,x_i)] + \tilde{\xi}(\tau_n,x_i)) + 1_{n=0}h(x_i).$$

Ψ_Δ defines an operator on $\mathcal{B}_{\{0,\dots,M\}\times\mathbb{N}}$. Moreover, Ψ_Δ is a contraction,

$$\begin{aligned}|\Psi_\Delta\tilde{v}_1(\tau_n,x_i) - \Psi_\Delta\tilde{v}_2(\tau_n,x_i)| &= 1_{n\in\{1,\dots,M\}} p_\Delta^{-1} e^{2\tau_n}|\Phi_\Delta[e^{-2\tau_n}(\tilde{v}_1-\tilde{v}_2)](\tau_n,x_i)|\\ &\le 1_{n\in\{1,\dots,M\}} p_\Delta^{-1} e^{2\tau_n}\|\tilde{v}_1-\tilde{v}_2\|_{\{0,\dots,M\}}\Phi_\Delta[e^{-2\tau_n}],\end{aligned}$$

since $v_1 - v_2 \le \|\tilde{v}_1-\tilde{v}_2\|_{\{0,\dots,M\}}$ and Φ_Δ is monotone. Using (A.1), we obtain

$$\|\Psi_\Delta\tilde{v}_1 - \Psi_\Delta\tilde{v}_2\|_{\{0,\dots,M\}} \le (1-p_\Delta^{-1})\|\tilde{v}_1-\tilde{v}_2\|_{\{0,\dots,M\}},$$

so Ψ_Δ is a contraction and the fixed point theorem entails (i).

To prove (ii), first assume that for all k, i, $\xi_1(\tau_k,x_i) \ge \xi_2(\tau_k,x_i)$ and show that in this case, $v_1 \ge v_2$. Let $v = v_1 - v_2$. We have $v(0,x_i) = 0$. Suppose that $v(\tau_{n-1},x_i) \ge 0$. Given $\varepsilon > 0$, let us take $i(\varepsilon)$ such that $v(\tau_n,x_{i(\varepsilon)}) \le \inf_i v(\tau_n,x_i) + \varepsilon \le v(\tau_n,x_i)+\varepsilon$, for all i. Then, for all $i \in \mathbb{N}$, we have

$$\begin{aligned}&\inf_i v(\tau_n,x_i) \ge v(\tau_n,x_{i(\varepsilon)}) - \varepsilon\\ &= -c\Delta t v(\tau_n,x_{i(\varepsilon)}) + (1+a\Delta t)v(\tau_n,x_{i(\varepsilon)}) - b\Delta t v(\tau_n,x_{i(\varepsilon)}) - \hat{\lambda}\Delta t v(\tau_n,x_{i(\varepsilon)}) - \varepsilon\\ &\ge -c\Delta t(v(\tau_n,x_{i(\varepsilon)-1})+\varepsilon) + (1+a\Delta t)v(\tau_n,x_{i(\varepsilon)}) - b\Delta t(v(\tau_n,x_{i(\varepsilon)+1})+\varepsilon)\\ &\qquad -\hat{\lambda}\Delta t(\inf_i v(\tau_n,x_i)+\varepsilon) - \varepsilon\\ &= v(\tau_{n-1},x_{i(\varepsilon)}) + \Delta t\sum_j \nu_j v(\tau_{n-1},x_{i(\varepsilon)+j}) + (\xi_1-\xi_2)(\tau_n,x_{i(\varepsilon)}) - \hat{\lambda}\Delta t\inf_i v(\tau_n,x_i)\\ &\qquad -(1+a\Delta t)\varepsilon \ge -\hat{\lambda}\Delta t\inf_i v(\tau_n,x_i) - (1+a\Delta t)\varepsilon.\end{aligned}$$

Therefore $\inf_i v(\tau_n,x_i) \ge -\frac{1+a\Delta t}{1+\hat{\lambda}\Delta t}\varepsilon$.

Taking the limit $\varepsilon \to 0$, we obtain $\inf_i v(\tau_n,x_i) \ge 0$, so $v = v_1 - v_2 \ge 0$.

Consider now the general case. Let $v(\tau_n,x_i) = v_1(\tau_n,x_i) + \sum_{k=1}^n \|\xi_1(\tau_k,\cdot) - \xi_2(\tau_k,\cdot)\|$. Then v solves (6.24)–(6.25) with $\xi(\tau_n,x_i) = v(\tau_n,x_i) - v(\tau_{n-1},x_i) - L_\Delta v(\tau_n,x_i)\Delta t$. We have

$$\begin{aligned}\xi(\tau_n,x_i) &= \xi_1(\tau_n,x_i) + \|\xi_1(\tau_n,\cdot)-\xi_2(\tau_n,\cdot)\| - L_\Delta\left[\sum_{k=1}^n \|\xi_1(\tau_k,\cdot)-\xi_2(\tau_k,\cdot)\|\right]\Delta t\\ &= \xi_1(\tau_n,x_i) + (1+\hat{\lambda}\Delta t)\|\xi_1(\tau_n,\cdot)-\xi_2(\tau_n,\cdot)\| \ge \xi_2(\tau_n,x_i).\end{aligned}$$

But as shown above this implies $v(\tau_n,x_i) \ge v_2(\tau_n,x_i)$, so

$$v_2(\tau_n,x_i) - v_1(\tau_n,x_i) \le \sum_{k=1}^n \|\xi_1(\tau_k,\cdot)-\xi_2(\tau_k,\cdot)\|.$$

Interchanging v_1 with v_2, we obtain (6.26).

A.2. Proof of Lemma 4. By definition,

$$u(\tau,x) = h(x) * \tilde{p}_\tau(x) = h(x) * \tilde{p}_\tau^W(x) * \tilde{p}_\tau^{(X-W)}(x), \tag{A.3}$$

where $p_\tau(x) = \tilde{p}_\tau(-x)$ is the density of X_τ, and $\tilde{p}_\tau^W(-x)$, $\tilde{p}_\tau^{(X-W)}(-x)$ are densities of the processes σW_τ and $(X_\tau - \sigma W_\tau)$, respectively. Therefore,

$$\left\| \frac{\partial^{m+n} u}{\partial \tau^m \partial x^n}(\tau,\cdot) \right\| \le \left\| \frac{\partial^{m+n}(h * \tilde{p}_\tau^W)}{\partial \tau^m \partial x^n} \right\|. \tag{A.4}$$

The derivatives of h may have jumps at $\xi_1,\dots,\xi_I$; denote these jumps by $a_i^{(n)} = h^{(n)}(\xi_i+) - h^{(n)}(\xi_i-)$. Then, for all $n \ge 1$, $i = 1,\dots,I$, $|a_i^{(n)}| \le 2K$. Using standard properties of convolution products we obtain

$$\begin{aligned}\frac{\partial^n (h * \tilde{p}_\tau^W)}{\partial x^n}(x) = (h^{(n)} * \tilde{p}_\tau^W)(x) + \sum_{i=1}^{I} \Bigg[a_i^{(n-1)} \tilde{p}_\tau^W(x-\xi_i) + a_i^{(n-2)} \frac{\partial \tilde{p}_\tau^W}{\partial x}(x-\xi_i) + {} \\ \cdots + a_i^{(1)} \frac{\partial^{n-2} \tilde{p}_\tau^W}{\partial x^{n-2}}(x-\xi_i) \Bigg],\end{aligned} \tag{A.5}$$

where $h^{(n)}$ is the nth pointwise derivative of h. For all $n \ge 0$, we have

$$\frac{\partial^n \tilde{p}_\tau^W}{\partial x^n}(x) = \frac{1}{(\sqrt{\tau})^{n+1}} \frac{\partial^n \tilde{p}_1^W}{\partial x^n}\left(\frac{x}{\sqrt{\tau}} \right). \tag{A.6}$$

Consequently, for all $n \ge 1$,

$$\begin{aligned}\left\| \frac{\partial^n (h * \tilde{p}_\tau^W)}{\partial x^n} \right\| &\le \|h^{(n)}\| + 2KI \left(\|\tilde{p}_\tau^W\| + \left\| \frac{\partial \tilde{p}_\tau^W}{\partial x} \right\| + \cdots + \left\| \frac{\partial^{n-2} \tilde{p}_\tau^W}{\partial x^{n-2}} \right\| \right) \\ &\le K + 2KI \left(\frac{1}{\sqrt{\tau}} \|\tilde{p}_1^W\| + \frac{1}{(\sqrt{\tau})^2} \left\| \frac{\partial \tilde{p}_1^W}{\partial x} \right\| + \cdots + \frac{1}{(\sqrt{\tau})^{n-1}} \left\| \frac{\partial^{n-2} \tilde{p}_1^W}{\partial x^{n-2}} \right\| \right) \\ &\le \frac{K}{(\sqrt{\tau})^{n-1}} \left[T^{\frac{n-1}{2}} + 2I \left(T^{\frac{n-2}{2}} \|\tilde{p}_1^W\| + T^{\frac{n-1}{2}} \left\| \frac{\partial \tilde{p}_1^W}{\partial x} \right\| + \cdots + \left\| \frac{\partial^{n-2} \tilde{p}_1^W}{\partial x^{n-2}} \right\| \right) \right] \\ &= \frac{C}{(\sqrt{\tau})^{n-1}},\end{aligned}$$

so (6.29) is verified for $m = 0$ and $n \ge 1$. Proceed by induction on m: assume (6.29) for $m-1$ and $n \ge 1$. For any $f \in C^\infty(\mathbb{R})$, we have

$$\begin{aligned}|Lf(x)| &= \left| \frac{\sigma^2}{2} \frac{\partial^2 f}{\partial x^2}(x) - \left(\frac{\sigma^2}{2} - r + \alpha \right) \frac{\partial f}{\partial x}(x) - \lambda f(x) + \int_{B_l}^{B_r} \nu(dy) f(x+y) \right| \\ &\le \frac{\sigma^2}{2} \left| \frac{\partial^2 f}{\partial x^2}(x) \right| + |\sigma^2/2 - r + \alpha| \left| \frac{\partial f}{\partial x}(x) \right| + 2\lambda \|f\|.\end{aligned}$$

Applying this result to $f = \partial^{m+n-1} u^{(\mu)} / \partial \tau^{m-1} \partial x^n$ and using $\partial_\tau u = Lu$, we obtain

$$\begin{aligned}&\left| \frac{\partial^{m+n} u}{\partial \tau^m \partial x^n}(\tau,x) \right| = \left| \frac{\partial}{\partial \tau} \left(\frac{\partial^{m-1+n} u}{\partial \tau^{m-1} \partial x^n} \right)(\tau,x) \right| = \left| L \left(\frac{\partial^{m-1+n} u}{\partial \tau^{m-1} \partial x^n} \right)(\tau,x) \right| \\ &\le \frac{\sigma^2}{2} \left\| \frac{\partial^{m-1+n+2} u}{\partial \tau^{m-1} \partial x^{n+2}}(\tau,\cdot) \right\| + \left| \frac{\sigma^2}{2} - r + \alpha \right| \left\| \frac{\partial^{m-1+n+1} u}{\partial \tau^{m-1} \partial x^{n+1}}(\tau,\cdot) \right\| + 2\lambda \left\| \frac{\partial^{m-1+n} u}{\partial \tau^{m-1} \partial x^n}(\tau,\cdot) \right\| \\ &\le \frac{C(K,T,m,n,r,\sigma,\lambda,\alpha)}{(\sqrt{\tau})^{2(m-1)+(n+2)-1}} = \frac{C}{(\sqrt{\tau})^{2m+n-1}}.\end{aligned}$$

We have thus shown (6.29) for $m \geq 0$, $n \geq 1$. For $m \geq 1$, $n = 0$, we proceed similarly, by induction on m starting from

$$\left\|\frac{\partial u}{\partial \tau}(\tau,\cdot)\right\| = \|Lu(\tau,\cdot)\| \leq \frac{\sigma^2}{2}\left\|\frac{\partial^2 u}{\partial x^2}(\tau,\cdot)\right\| + \left|\frac{\sigma^2}{2} - r + \alpha\right| \left\|\frac{\partial u}{\partial x}(\tau,\cdot)\right\| + 2\lambda \|u(\tau,\cdot)\| \leq \frac{C}{\sqrt{\tau}},$$

since $\|u(\tau,\cdot)\| = \|h * \tilde{p}_\tau^W\| \leq \|h\| \leq K$.

A.3. Proof of Lemma 5. We have

$$|Lu(\tau_k,x_i) - L_\Delta u(\tau_k,x_i)| \leq |Du(\tau_k,x_i) - D_\Delta u(\tau_k,x_i)| + |Ju(\tau_k,x_i) - J_\Delta u(\tau_{k-1},x_i)|,$$

where D and J are the differential and integral parts of L defined by (5.4), and D_Δ, J_Δ are their approximations given by (5.8)–(5.9). Recall that $|\alpha - \hat{\alpha}| \leq (\alpha+\lambda)\Delta x$ by (6.8). From Taylor's formula, there exist $\xi_1, \eta_1, \xi_2 \in [x_{i-1}, x_{i+1}]$ such that

$$\begin{aligned}
|Du(\tau_k,x_i) - D_\Delta u(\tau_k,x_i)| &= \left| \frac{\sigma^2}{2}\frac{\Delta x^2}{24}\left[\frac{\partial^4 u}{\partial x^4}(\tau_k,\xi_1) + \frac{\partial^4 u}{\partial x^4}(\tau_k,\eta_1)\right]\right. \\
&\quad \left. + \left(\frac{\sigma^2}{2} - r + \hat{\alpha}\right)\frac{\Delta x}{2}\frac{\partial^2 u}{\partial x^2}(\tau_k,\xi_2) + (\alpha - \hat{\alpha})\frac{\partial u}{\partial x}(\tau_k,x_i)\right| \\
\leq \frac{\Delta x^2}{12}\frac{\sigma^2}{2}\left\|\frac{\partial^4 u}{\partial x^4}(\tau_k,\cdot)\right\| &+ \frac{\Delta x}{2}|\sigma^2/2 - r + \alpha| \left\|\frac{\partial^2 u}{\partial x^2}(\tau_k,\cdot)\right\| + \Delta x(\alpha+\lambda)\left\|\frac{\partial u}{\partial x}(\tau_k,\cdot)\right\| \\
&\leq C[\Delta x^2/\tau_k^{3/2} + \Delta x/\tau_k^{1/2} + \Delta x]
\end{aligned}$$

by Lemma 4. The integral part can be estimated as follows:

$$\begin{aligned}
|Ju(\tau_k,x_i) - J_\Delta u(\tau_{k-1},x_i)| &= \left|\sum_{j=K_l}^{K_r} u(\tau_{k-1},x_{i+j})\nu_j - \int_{B_l}^{B_r} u(\tau_k,x_i+y)\nu(dy)\right| \\
&= \left|\sum_{j=K_l}^{K_r}\int_{y_{j-1/2}}^{y_{j+1/2}} [u(\tau_{k-1},x_i+y_j) - u(\tau_k,x_i+y)]\nu(dy)\right| \\
&\leq \left|\sum_{j=K_l}^{K_r}[u(\tau_{k-1},x_i+y_j) - u(\tau_k,x_i+y_j)]\nu_j\right| + \left|\sum_{j=K_l}^{K_r}\int_{y_{j-1/2}}^{y_{j+1/2}}[u(\tau_k,x_i+y_j) \right. \\
&\qquad \left. - u(\tau_k,x_i+y)]\nu(dy)\right| \\
&\leq \left|\sum_{j=K_l}^{K_r}\nu_j\int_{\tau_{k-1}}^{\tau_k}\frac{\partial u}{\partial \tau}(s,x_i+y_j)ds\right| + \left|\sum_{j=K_l}^{K_r}\int_{y_{j-1/2}}^{y_{j+1/2}}\nu(dy)\int_{x_i+y_j}^{x_i+y}\frac{\partial u}{\partial x}(\tau_k,\xi)d\xi\right| \\
&\leq \lambda\int_{\tau_{k-1}}^{\tau_k}\left\|\frac{\partial u}{\partial \tau}(s,\cdot)\right\|ds + \frac{\lambda\Delta x}{2}\left\|\frac{\partial u}{\partial x}(\tau_k,\cdot)\right\| \leq C\left[\int_{\tau_{k-1}}^{\tau_k}\frac{ds}{\sqrt{s}} + \Delta x\right] \\
&= C\left[2(\sqrt{\tau_k} - \sqrt{\tau_{k-1}}) + \Delta x\right].
\end{aligned}$$

Assembling the various terms we obtain (6.30).

Acknowledgments. We thank Yves Achdou, Peter Forsyth, Daniel Gabay, Huyen Pham, Olivier Pironneau, Christoph Schwab, Peter Tankov, and an anonymous referee for helpful suggestions.

REFERENCES

[1] O. Alvarez and A. Tourin, *Viscosity solutions of nonlinear integro-differential equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, 13 (1996), pp. 293–317.

[2] L. Andersen and J. Andreasen, *Jump-diffusion models: Volatility smile fitting and numerical methods for pricing*, Rev. Derivatives Research, 4 (2000), pp. 231–262.

[3] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton, *Implicit-explicit methods for time-dependent partial differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 797–823.

[4] G. Barles, R. Buckdahn, and E. Pardoux, *BSDEs and integral-partial differential equations*, Stochastics Stochastics Rep., 60 (1997), pp. 57–83.

[5] G. Barles and E. R. Jakobsen, *On the convergence rate of approximation schemes for Hamilton–Jacobi–Bellman equations*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 33–54.

[6] G. Barles and P. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptotic Anal., 4 (1991), pp. 271–283.

[7] O. Barndorff-Nielsen, *Processes of normal inverse Gaussian type*, Finance Stoch., 2 (1998), pp. 41–68.

[8] A. Bensoussan and J.-L. Lions, *Contrôle Impulsionnel et Inéquations Quasi-Variationnelles*, Dunod, Paris, 1982.

[9] P. Carr and D. Madan, *Option valuation using the fast Fourier transform*, J. Comput. Finance, 2 (1998), pp. 61–73.

[10] P. Carr and L. Wu, *The finite moment logstable process and option pricing*, J. Finance, 58 (2003), pp. 753–778.

[11] R. Cont and P. Tankov, *Financial Modelling with Jump Processes*, Chapman & Hall/CRC, Boca Raton, FL, 2004.

[12] R. Cont and P. Tankov, *Nonparametric calibration of jump-diffusion option pricing models*, J. Comput. Finance, 7 (2004), pp. 1–49.

[13] R. Cont and E. Voltchkova, *Integrodifferential equations for option prices in exponential Lévy models*, Finance Stoch., 9 (2005), pp. 299–325.

[14] M. G. Crandall, H. Ishii, and P.-L. Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.), 27 (1992), pp. 1–67.

[15] M. Crandall and P.-L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[16] Y. d'Halluin, P. A. Forsyth, and G. Labahn, *A penalty method for American options with jump diffusion processes*, Numer. Math., 97 (2004), pp. 321–352.

[17] E. Eberlein, *Application of generalized hyperbolic Lévy motions to finance*, in Lévy Processes—Theory and Applications, O. Barndorff-Nielsen, T. Mikosch, and S. Resnick, eds., Birkhäuser Boston, Boston, 2001, pp. 319–336.

[18] M. G. Garroni and J. L. Menaldi, *Second Order Elliptic Integro-Differential Problems*, Chapman & Hall/CRC, Boca Raton, FL, 2002.

[19] N. Jacob, *Pseudo Differential Operators and Markov Processes. Volume* I. *Fourier Analysis and Semigroups*, Imperial College Press, London, 2001.

[20] E. Jakobsen and K. Karlsen, *A Maximum Principle for Semicontinuous Functions Applicable to Integro-Partial Differential Equations*, working paper, University of Oslo, Oslo, Norway, 2004.

[21] N. Krylov, *On the rate of convergence of finite difference approximations for Bellman's equations*, St. Petersburg Math. J., 9 (1997), pp. 245–256.

[22] N. Krylov, *On the rate of convergence of finite difference approximations for Bellman's equations with variable coefficients*, Probab. Theory Related Fields, 117 (1997), pp. 1–16.

[23] D. Madan and F. Milne, *Option pricing with variance gamma martingale components*, Math. Finance, 1 (1991), pp. 39–55.

[24] D. Madan, *Purely discontinuous asset price processes*, in Option Pricing, Interest Rates and Risk Management, J. Cvitanic, E. Jouini, and M. Musiela, eds., Cambridge University Press, Cambridge, UK, 2001, pp. 105–153.

[25] A.-M. Matache, T. von Petersdorff, and C. Schwab, *Fast deterministic pricing of options on Lévy driven assets*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 37–71.

[26] H. Pham, *Optimal stopping of controlled jump-diffusion processes: A viscosity solution approach*, J. Math. Systems Estim. Control, 8 (1998), pp. 1–27.

[27] K. Sato, *Lévy Processes and Infinitely Divisible Distributions*, Cambridge University Press, Cambridge, UK, 1999.

[28] A. Sayah, *Equations d'Hamilton Jacobi du premier ordre avec termes integro-differentiels*, Comm. Partial Differential Equations, 16 (1991), pp. 1057–1093.

[29] H. Soner, *Optimal control of jump-Markov processes and viscosity solutions*, in Stochastic Differential Systems, Stochastic Control Theory and Applications, IMA Vol. Math. Appl. 10, Springer-Verlag, New York, 1988, pp. 501–511.

[30] D. Tavella and C. Randall, *Pricing Financial Instruments*, Wiley, New York, 2000.

[31] X. Zhang, *Valuation of American options in a jump-diffusion model*, in Numerical Methods in Finance, Cambridge University Press, Cambridge, UK, 1997, pp. 93–114.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1627–1650

INCOMPRESSIBLE FINITE ELEMENTS VIA HYBRIDIZATION. PART I: THE STOKES SYSTEM IN TWO SPACE DIMENSIONS*

BERNARDO COCKBURN† AND JAYADEEP GOPALAKRISHNAN‡

Abstract. In this paper, we introduce a new and efficient way to compute exactly divergence-free velocity approximations for the Stokes equations in two space dimensions. We begin by considering a mixed method that provides an exactly divergence-free approximation of the velocity and a continuous approximation of the vorticity. We then rewrite this method solely in terms of the tangential fluid velocity and the pressure on mesh edges by means of a new hybridization technique. This novel formulation bypasses the difficult task of constructing an exactly divergence-free basis for velocity approximations. Moreover, the discrete system resulting from our method has fewer degrees of freedom than the original mixed method since the pressure and the tangential velocity variables are defined just on the mesh edges. Once these variables are computed, the velocity approximation satisfying the incompressibility condition exactly, as well as the continuous numerical approximation of the vorticity, can at once be obtained locally. Moreover, a discontinuous numerical approximation of the pressure within elements can also be obtained locally. We show how to compute the matrix system for our tangential velocity-pressure formulation on general meshes and present in full detail such computations for the lowest-order case of our method.

Key words. divergence-free finite element, mixed method, velocity, vorticity, pressure, hybridized method, fluid flow, Stokes flow, Lagrange multipliers

AMS subject classifications. 65N30, 76D07

DOI. 10.1137/04061060X

1. Introduction.

In this paper, we introduce a new and efficient way to compute exactly divergence-free velocity approximations for the Stokes equations in two space dimensions. We proceed as follows. First, we consider the mixed method for the Stokes equations studied in [12, 13, 20]. This method provides a continuous approximation for the vorticity and an exactly divergence-free approximation of the velocity. Then we introduce a new hybridization technique that allows us to reduce the original method to a *mixed method for the Lagrange multipliers* arising from the hybridization, namely, the tangential fluid velocity and the pressure along mesh edges. This novel implementation of the method requires neither the introduction of stream function variables (as in [12, 13, 20]) nor the construction of a globally divergence-free finite element basis. We thus avoid the difficulties in construction of a globally divergence-free basis as well as the increase in degrees of freedom that accompanies the introduction of the stream function. Our new *tangential velocity-pressure* formulation has fewer degrees of freedom as both the unknowns are defined only on mesh edges. Moreover, after solving for these unknowns, the original exactly divergence-free numerical approximation of the fluid velocity and the original continuous numerical approximation of the vorticity can be easily computed in an element-by-element fashion. An

*Received by the editors June 25, 2004; accepted for publication (in revised form) January 28, 2005; published electronically November 14, 2005.
http://www.siam.org/journals/sinum/43-4/61060.html

†School of Mathematics, University of Minnesota, Vincent Hall, Minneapolis, MN 55455 (cockburn@math.umn.edu). The research of this author was supported in part by National Science Foundation grant DMS-0411254 and by the University of Minnesota Supercomputing Institute.

‡Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (jayg@math.ufl.edu). The research of this author was supported in part by National Science Foundation grant DMS-0410030.

approximation to the pressure inside the elements can also be computed in this way, a feature made possible by the hybridization procedure.

Let us describe the hybridization technique we propose. Recall that the Stokes equations couple the fluid velocity $\boldsymbol{u}$ and the pressure p by the equations

$$-\boldsymbol{\Delta u} + \mathbf{grad}\, p = \boldsymbol{f} \qquad \text{on } \Omega, \tag{1.1}$$

$$\operatorname{div} \boldsymbol{u} = 0 \qquad \text{on } \Omega, \tag{1.2}$$

$$\boldsymbol{u} = \boldsymbol{g} \qquad \text{on } \partial\Omega. \tag{1.3}$$

Here, $\boldsymbol{f} \in L^2(\Omega)^2$ and $\boldsymbol{g} \in H^{1/2}(\partial\Omega)^2$ are given data. For simplicity, we assume that $\Omega \subseteq \mathbb{R}^2$ is a (bounded connected) polygon. To define the mixed method, we introduce the vorticity

$$\omega = \operatorname{curl} \boldsymbol{u} := \frac{\partial}{\partial x} u_y - \frac{\partial}{\partial y} u_x,$$

where $\boldsymbol{u} = (u_x, u_y)$, and rewrite the Stokes system as

$$\omega - \operatorname{curl} \boldsymbol{u} = 0 \qquad \text{on } \Omega, \tag{1.4}$$

$$\mathbf{curl}\, \omega + \mathbf{grad}\, p = \boldsymbol{f} \qquad \text{on } \Omega, \tag{1.5}$$

$$\operatorname{div} \boldsymbol{u} = 0 \qquad \text{on } \Omega, \tag{1.6}$$

$$\boldsymbol{u} \cdot \boldsymbol{t} = g_t \qquad \text{on } \partial\Omega, \tag{1.7}$$

$$\boldsymbol{u} \cdot \boldsymbol{n} = g_n \qquad \text{on } \partial\Omega. \tag{1.8}$$

Here, $g_t = \boldsymbol{g} \cdot \boldsymbol{t}$ and $g_n = \boldsymbol{g} \cdot \boldsymbol{n}$, where $\boldsymbol{n}$ denotes the outward unit normal on $\partial\Omega$ and $\boldsymbol{t}$ the unit tangent vector on $\partial\Omega$ oriented such that Ω is on the left as we move in the direction of $\boldsymbol{t}$ along $\partial\Omega$. Note that to obtain (1.5), we made use of the identity

$$-\boldsymbol{\Delta u} = \mathbf{curl}\, \operatorname{curl} \boldsymbol{u} - \mathbf{grad}\, \operatorname{div} \boldsymbol{u},$$

where

$$\mathbf{curl}\, \omega = \left(\frac{\partial \omega}{\partial y}, -\frac{\partial \omega}{\partial x} \right).$$

To give a weak formulation of the above problem, define the spaces

$$\begin{aligned} \mathcal{W} \quad &= H^1(\Omega), \\ \mathcal{V} \quad &= \{\boldsymbol{v} \in H(\operatorname{div}, \Omega) : \operatorname{div} \boldsymbol{v} = 0\}, \\ \mathcal{V}(b) &= \{\boldsymbol{v} \in \mathcal{V} : \boldsymbol{v} \cdot \boldsymbol{n}|_{\partial\Omega} = b\} \end{aligned}$$

for any $b \in H^{-1/2}(\partial\Omega)$. The weak formulation seeks the pair of functions satisfying

$$(\omega, \tau)_\Omega - (\boldsymbol{u}, \mathbf{curl}\, \tau)_\Omega = (g_t, \tau)_{\partial\Omega} \qquad \text{for all } \tau \in \mathcal{W}, \tag{1.9}$$

$$(\boldsymbol{v}, \mathbf{curl}\, \omega)_\Omega = (\boldsymbol{f}, \boldsymbol{v})_\Omega \qquad \text{for all } \boldsymbol{v} \in \mathcal{V}(0). \tag{1.10}$$

Here, $(\cdot, \cdot)_\Omega$ denotes the $L^2(\Omega)$ (or $L^2(\Omega)^2$) inner product. Note that since the velocity test functions are taken in the space $\mathcal{V}(0)$, the pressure is no longer present in this variational formulation. By classical existence results for the Stokes system, it is easy to show that there is a unique solution for the above system of equations, provided that the compatibility condition

$$(g_n, 1)_{\partial\Omega} = 0 \tag{1.11}$$

is satisfied. We assume throughout that (1.11) holds.

Now the approximate solution is sought in the finite element subspaces of the above defined spaces:

$$\begin{aligned} \mathcal{W}_h &= \{w \in \mathcal{W} : w|_K \in P_{k+1}(K) \text{ for all } K \in \mathcal{T}\}, \\ \mathcal{V}_h &= \{\boldsymbol{v} \in \mathcal{V} : \boldsymbol{v}|_K \in P_k(K)^2 \text{ for all } K \in \mathcal{T}\}. \end{aligned}$$

Here $\mathcal{T}$ denotes a finite element triangulation of Ω. Let $\mathcal{V}_h(b) = \mathcal{V}(b) \cap \mathcal{V}_h$ and $g_{n,h}$ be the $L^2(\partial\Omega)$-orthogonal projection of the boundary data g_n onto the space

$$\{\boldsymbol{v}_h \cdot \boldsymbol{n}|_{\partial\Omega} : \boldsymbol{v}_h \in \mathcal{V}_h\}.$$

Then the discrete mixed formulation seeks $(\omega_h, \boldsymbol{u}_h)$ in $\mathcal{W}_h \times \mathcal{V}_h(g_{n,h})$ satisfying

$$(\omega_h, \tau)_\Omega - (\boldsymbol{u}_h, \mathbf{curl}\,\tau)_\Omega = (g_t, \tau)_{\partial\Omega} \qquad \text{for all } \tau \in \mathcal{W}_h, \tag{1.12}$$

$$(\boldsymbol{v}, \mathbf{curl}\,\omega_h)_\Omega = (\boldsymbol{f}, \boldsymbol{v})_\Omega \qquad \text{for all } \boldsymbol{v} \in \mathcal{V}_h(0). \tag{1.13}$$

We assume that the latter space is nonempty. A three-dimensional version of the above mixed discretization was studied in [12, 13, 20, 22], where the existence of a unique solution was established. Note that this is a conforming method since

$$\mathcal{W}_h \times \mathcal{V}_h(0) \subset \mathcal{W} \times \mathcal{V}(0) \subset H^1(\Omega) \times \mathcal{V}.$$

This implies, in particular, that in order to implement the method in the above form, we must face the difficult task of constructing bases for the finite-dimensional space of globally divergence-free velocities $\mathcal{V}_h(0)$.

The construction of an exactly divergence-free finite element basis has been a long-standing research question [14]. Piecewise divergence-free approximations have been investigated in [2, 17, 18], but their normal components are not continuous in general. Basis functions for finite-dimensional spaces of *weakly* divergence-free functions were constructed in [15, 16, 24]. However, this construction proved to be extremely difficult to extend to spaces of polynomials of higher degree. Exactly divergence-free finite element spaces have been studied, but known results require the use of polynomials of degree four or higher for the two-dimensional case [19, 23], and no similar result exists for the three-dimensional case. The difficulty of constructing exactly incompressible finite element spaces was overcome in [12] by setting the divergence-free spaces as the curl of an appropriate space of stream functions. Unfortunately, the introduction of the stream function increases degrees of freedom. In contrast, our approach to overcoming this difficulty via hybridization actually results in a reduction in degrees of freedom.

Recently, globally divergence-free approximations were devised by using discontinuous Galerkin methods with polynomials of degree one or higher in the framework of the Navier–Stokes equations [10]. To achieve this, the fact that the divergence-free condition is enforced element-by-element is exploited to construct an element-by-element postprocessing of the discontinuous approximation that automatically results in an exactly divergence-free velocity. A similar technique in the framework of discontinuous Galerkin methods for Darcy flow was developed in [3]. Unfortunately, such approaches cannot be used for conforming mixed methods since they rely on the fact that the discontinuous Galerkin methods enforce the equations element-by-element.

The main idea of our procedure is to look for approximations in discrete spaces that have *no* continuity constraints across mesh interfaces and introduce new sets of equations that guarantee that the new approximation *coincides* with the original approximation $(\omega_h, \boldsymbol{u}_h)$ given by (1.12)–(1.13). This approach is inspired by hybridization techniques used in the context of mixed methods for second-order elliptic problems [1, 5, 8, 11]. We proceed in two steps. The objective of the first step is to circumvent construction of divergence-free finite element bases. Hence, in this step, we relax the continuity of the normal components of the approximate velocity across interelement boundaries and use a velocity space of functions with no interelement continuity. As a direct consequence, the pressure reappears in the equations, but only on the edges if the approximate velocities are divergence-free inside each element. Then, new equations are introduced to enforce the continuity of the normal component of the velocity across interelement boundaries. A similar hybridization technique, but in the framework of discontinuous Galerkin methods for the Stokes problem, is explored in [7].

The objective of the second step is the eventual elimination of both the original unknowns (velocity and vorticity) from the equations. To do this, we must develop a new hybridization technique for the vorticity. Such a hybridization is far more involved than the previous one since the vorticity is continuous across interelement boundaries. Indeed, all the previously known hybridization procedures relaxed continuity of spaces with edge (or face in three dimensions) degrees of freedom. Examples include hybridization techniques for the Raviart–Thomas and Brezzi–Douglas–Marini (BDM) methods for scalar second-order elliptic problems which involve finite element subspaces of $H(\mathrm{div}, \Omega)$. Hybridization of the Morley element method for the biharmonic problem [1] also involved such spaces with edge degrees of freedom. However, hybridization techniques to relax continuity constraints of finite element subspaces of $H^1(\Omega)$ with vertex degrees of freedom have remained unknown until now. While this may have led to a widespread belief that methods using spaces of this type are not amenable to hybridization, in this paper we show otherwise. We show how one can approximate vorticity in a space of functions which have no continuity conditions across element interfaces while imposing the natural continuity properties of the vorticity as an equation of the method.

After the above mentioned hybridizations, we proceed to adapt the methodology introduced in [8] to eliminate the vorticity and velocity from the hybridized method. This elimination is far from obvious but is greatly facilitated by the fact that both the vorticity and the velocity are in spaces of functions with no interelement continuity and by the fact that both the pressure and the tangential velocity are defined only on the mesh edges. This allows us to express the vorticity and velocity in terms of the pressure and tangential velocity. Then, we show how to characterize these Lagrange multipliers as the only solution of a new mixed method. We view this method as a “tangential velocity-pressure discretization” for the Stokes equation wherein the unknowns are all on the mesh edges.

Notice that since the unknowns are defined only on the edges, this system is smaller than the original one. Moreover, once the Lagrange multipliers are obtained, vorticity and velocity approximations can be obtained by *local* element-by-element computations. An interesting feature of our mixed method for the Lagrange multipliers is that it is possible to further eliminate the pressure Lagrange multiplier and form one Schur complement equation for the tangential velocity Lagrange multiplier. This equation can be easily solved using well-established iterative techniques for symmetric positive definite systems.

We should note that ours is not the first paper to give hybridized methods for the Stokes problem. A hybrid formulation involving deviatoric stresses, hydrostatic pressure, and velocity was given in [6, 25]. Note also that some domain decomposition methods result from hybridization performed at the subdomain level. For example, in [4], the method gives rise to an indefinite system for the velocity nodes on the subdomain boundaries and the mean values of the pressure on the subdomains. However, none of the above mentioned methods provide incompressible velocities.

The paper is organized as follows. In section 2, we give a detailed description of the hybridization of the original conforming mixed method. The resulting method is written as a method for the two original variables *and* two additional Lagrange multipliers. Then, in section 3, we show how to eliminate the former two variables from the equations and characterize the Lagrange multipliers alone as the unique solution of a mixed method. This characterization (Theorem 3.1) is an extension to the Stokes system of what was done for hybridized mixed methods for second-order elliptic problems in [8] and is one of our main results. In section 4, we construct the bases for the Lagrange multipliers, and in section 5, we discuss some key implementation aspects of the method. These include the construction of the Schur complement matrix for the tangential velocity and the detailed computation of the matrices of the method for the lowest-order case. Section 6 concludes the paper.

2. The hybridized mixed method.

In this section, we present the hybridization of the mixed method in full detail as described in the introduction. Let us emphasize once again that this is carried out in two steps. The objective of the first is to avoid having to construct finite-dimensional spaces of divergence-free velocities. The objective of the second is the eventual elimination of the original variables from the equations. Note that the actual elimination is not carried out until section 3.

2.1. First hybridization: Introduction of pressure on the mesh edges.

We begin by relaxing the continuity of the normal component of the approximate velocity $\boldsymbol{u}_h$ across interelement boundaries. Thus, instead of seeking velocity approximations in the space $\mathcal{V}_h$, we seek approximations in the space

$$V_h = \{\boldsymbol{v} : \boldsymbol{v}|_K \in P_k(K)^2 \text{ and } \operatorname{div}(\boldsymbol{v}|_K) = 0 \text{ for all } K \in \mathcal{T}\}.$$

This forces us to weakly impose (1.5) in a different way. Indeed, if we multiply (1.5) by a test function $\boldsymbol{v}_h \in V_h$ and integrate over the element K, we obtain

$$(\mathbf{curl}\,\omega, \boldsymbol{v}_h)_K + (\mathbf{grad}\,p, \boldsymbol{v}_h)_K = (\boldsymbol{f}, \boldsymbol{v}_h)_K,$$

and hence,

$$(\mathbf{curl}\,\omega, \boldsymbol{v}_h)_K + (p, \boldsymbol{v}_h \cdot \boldsymbol{n})_{\partial K} = (\boldsymbol{f}, \boldsymbol{v}_h)_K.$$

Replacing ω and p by their respective approximations, ω_h and p_h, and adding over the elements of the triangulation, we obtain one equation of the method:

$$(\boldsymbol{v}_h, \mathbf{curl}\,\omega_h)_\Omega + \sum_{e\in\mathcal{E}} (p_h, [\![\boldsymbol{v}_h \cdot \boldsymbol{n}]\!])_e = (\boldsymbol{f}, \boldsymbol{v}_h)_\Omega \quad \text{for all } \boldsymbol{v}_h \in V_h.$$

Here, we are using the following notation: For $\boldsymbol{v} \in V_h$ the jump of the normal component of $\boldsymbol{v}$ across interelement boundaries, denoted by $[\![\boldsymbol{v} \cdot \boldsymbol{n}]\!]$, is defined on the set

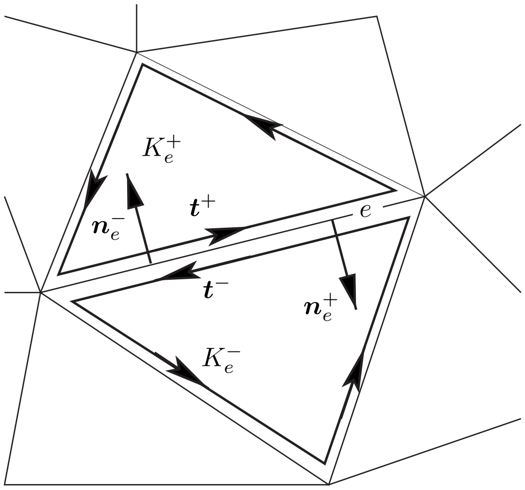


FIG. 1. *Notation for elements, normals, and tangents near an edge e.*

$\mathcal{E}$ of all edges of the triangulation $\mathcal{T}$ as follows. On every interior edge e in $\mathcal{E}$ shared by two mesh triangles K_e^+ and K_e^- we define

$$[\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]_e = \boldsymbol{v}_e^+\cdot\boldsymbol{n}_e^+ + \boldsymbol{v}_e^-\cdot\boldsymbol{n}_e^-,$$

where $\boldsymbol{n}_e^+$ and $\boldsymbol{n}_e^-$ denote the outward unit normals on the boundaries of K_e^+ and K_e^-, respectively (see Figure 1) and $\boldsymbol{v}_e^\pm(\boldsymbol{x}) = \lim_{\epsilon\downarrow 0}\boldsymbol{v}(\boldsymbol{x}-\epsilon\,\boldsymbol{n}_e^\pm)$. On edges $e\subset\partial\Omega$, we set $[\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]_e = \boldsymbol{v}|_{\partial\Omega}\cdot\boldsymbol{n}$. By $[\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]$ (without any subscript) we mean the function that is defined on the union of all edges in $\mathcal{E}$ and equals $[\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]_e$ on each edge $e\in\mathcal{E}$.

Now, in accordance with the hybridization paradigm, we impose the continuity of the normal component of the velocity $\boldsymbol{u}_h$ across interelement boundaries through the equation

$$\sum_{e\in\mathcal{E}}(q_h,\,[\![\boldsymbol{u}_h\cdot\boldsymbol{n}]\!])_e = (g_n,q_h)_{\partial\Omega}\quad\text{for all } q_h\in P_h,$$

where P_h is defined naturally by

$$P_h = \{p : p = [\![\boldsymbol{v}\cdot\boldsymbol{n}]\!] \text{ for some } \boldsymbol{v}\in V_h\}. \tag{2.1}$$

Notice that in the above equation, we are also incorporating the boundary condition on the normal component of the velocity.

Thus, after the first hybridization of the mixed method, we are seeking an approximation $(\omega_h,\boldsymbol{u}_h,p_h)\in\mathcal{W}_h\times V_h\times P_h$ satisfying

$$(\omega_h,\tau_h)_\Omega - (\boldsymbol{u}_h,\mathbf{curl}\,\tau_h)_\Omega = (g_t,\tau_h)_{\partial\Omega}\quad\text{for all } \tau_h\in\mathcal{W}_h, \tag{2.2}$$

$$(\boldsymbol{v}_h,\mathbf{curl}\,\omega_h)_\Omega + \sum_{e\in\mathcal{E}}(p_h,\,[\![\boldsymbol{v}_h\cdot\boldsymbol{n}]\!])_e = (\boldsymbol{f},\boldsymbol{v}_h)_\Omega\quad\text{for all } \boldsymbol{v}_h\in V_h, \tag{2.3}$$

$$\sum_{e\in\mathcal{E}}(q_h,\,[\![\boldsymbol{u}_h\cdot\boldsymbol{n}]\!])_e = (g_n,q_h)_{\partial\Omega}\quad\text{for all } q_h\in P_h. \tag{2.4}$$

Note that although the original mixed method (1.12)–(1.13) did not involve the pressure variable, the pressure reappears upon hybridization, but only along the mesh edges. We shall call p_h the "pressure Lagrange multiplier." The above discrete formulation has a unique solution, as we show next.

PROPOSITION 2.1. *There is a unique solution* $(\omega_h, \boldsymbol{u}_h, p_h) \in \mathcal{W}_h \times V_h \times P_h$ *for the hybridized mixed method* (2.2)–(2.4)*, and the solution components* ω_h *and* $\boldsymbol{u}_h$ *are the same as the solution of* (1.12)–(1.13).

Proof. We show that if g_n, g_t, and $\boldsymbol{f}$ are equal to zero, then $\omega_h, \boldsymbol{u}_h$, and p_h are also zero. First, (2.4) implies $\boldsymbol{u}_h \in \mathcal{V}_h(0)$. Moreover, (2.2)–(2.3) implies that ω_h and $\boldsymbol{u}_h$ satisfy (1.12)–(1.13) with zero data. By uniqueness of solutions of (1.12)–(1.13), we find that $\omega_h = 0$ and $\boldsymbol{u}_h = 0$. This together with (2.3) implies that p_h is zero. Hence there is a unique solution for (2.2)–(2.4). It is easy to see that if ω_h and $\boldsymbol{u}_h$ satisfy (2.2)–(2.4), then they also satisfy (1.12)–(1.13); hence the equivalence of both problems. □

Before proceeding to describe the second hybridization, let us point out that the first allows us to recover an approximation for the pressure inside the elements in an element-by-element fashion. To define such an approximation, we follow a technique of [7]. We define the pressure π_h on the triangle K as the element of $P_k(K)$ such that

$$-(\pi_h, \operatorname{div} \boldsymbol{v})_K = (\boldsymbol{f}, \boldsymbol{v})_K - (\mathbf{curl}\, \omega_h, \boldsymbol{v})_K - (\boldsymbol{v} \cdot \boldsymbol{n}, p_h)_{\partial K} \tag{2.5}$$

for all $\boldsymbol{v}$ in $P_k(K)^2 + \boldsymbol{x}\, P_k(K)$, where $\boldsymbol{n}$ denotes the outward unit normal to K. That (2.5) uniquely defines π_h follows from two facts:

(i) div : $P_k(K)^2 + \boldsymbol{x}\, P_k(K) \mapsto P_k(K)$ is a surjection;

(ii) If $\operatorname{div} \boldsymbol{v} = 0$ for a $\boldsymbol{v}$ in $P_k(K)^2 + \boldsymbol{x}\, P_k(K)$, then $\boldsymbol{v} \in P_k(K)^2$ and the right-hand side of the above equation is zero by the definition of the hybridized method.

The idea of recovering pressure approximations a posteriori as in (2.5) from approximations of other variables is old (see, e.g., [15]), but because hybridization provides p_h, we are able to compute π_h locally in our case. Thus our method can simultaneously provide approximations to the *velocity*, *vorticity*, and *pressure*.

2.2. Second hybridization: Introduction of the tangential velocity variable. Now we relax the continuity of the approximate vorticity ω_h across mesh edges in the interior of the domain. Thus, instead of considering continuous approximations in the space $\mathcal{W}_h$, we formulate a method using the space

$$W_h = \{w : w|_K \in P_{k+1}(K) \text{ for all } K \in \mathcal{T}\}.$$

This forces us to weakly impose (1.4) in a different way. Indeed, if we multiply that equation by a test function $\tau_h \in W_h$ and integrate over the element K, we obtain

$$(\omega, \tau_h)_K - (\boldsymbol{u}, \mathbf{curl}\, \tau_h)_K - (\boldsymbol{u} \cdot \boldsymbol{t}, \tau_h)_{\partial K} = 0,$$

where $\boldsymbol{t}$ denotes the unit tangent vector along ∂K oriented as in Figure 1. Here and elsewhere to simplify notation, we do not explicitly indicate the dependence of $\boldsymbol{t}$ on the underlying boundary (such as ∂K above). Denoting the tangential component of the velocity $\boldsymbol{u}$ on the interelement boundaries by

$$\boldsymbol{\lambda} = (\boldsymbol{u} \cdot \boldsymbol{t})\, \boldsymbol{t},$$

we can rewrite the above equation as

$$(\omega, \tau_h)_K - (\boldsymbol{u}, \mathbf{curl}\, \tau_h)_K - (\boldsymbol{\lambda}, \tau_h \boldsymbol{t})_{\partial K} = 0.$$

Next, replacing ω, $\boldsymbol{u}$, and $\boldsymbol{\lambda}$ by their respective approximations ω_h, $\boldsymbol{u}_h$, and $\boldsymbol{\lambda}_h$, we obtain, after adding over the elements K of the triangulation,

$$(\omega_h, \tau_h)_\Omega - (\boldsymbol{u}_h, \mathbf{curl}\, \tau_h)_\Omega - \sum_{e \in \mathcal{E} \setminus \partial\Omega} (\boldsymbol{\lambda}_h,\, [\![\tau_h \boldsymbol{t}]\!]_e)_e = (g_t, \tau_h)_{\partial\Omega}. \tag{2.6}$$

Here, the "tangential jump" of τ across interelement boundaries, $[\![\tau \boldsymbol{t}]\!]$, is defined as follows. For every interior edge $e \in \mathcal{E}$ shared by triangles K_e^+ and K_e^-, let

$$[\![\tau \boldsymbol{t}]\!]_e = \tau_e^+ \boldsymbol{t}^+ + \tau_e^- \boldsymbol{t}^-,$$

where, as before, $\tau_e^\pm(\boldsymbol{x}) = \lim_{\epsilon \downarrow 0} \tau(\boldsymbol{x} - \epsilon\, \boldsymbol{n}_e^\pm)$, and $\boldsymbol{t}^+$ and $\boldsymbol{t}^-$ are unit tangent vectors along the boundaries of K_e^+ and K_e^-, respectively, oriented in accordance with our previous notation: Unit tangent vectors along the boundary of a domain are given the orientation that leaves the domain on its left (see Figure 1). Hence $\boldsymbol{t}^+ = -\boldsymbol{t}^-$ on e. It is convenient to adopt the convention that the jump $[\![\tau \boldsymbol{t}]\!]$ on the boundary of Ω vanishes:

$$[\![\tau \boldsymbol{t}]\!]_e = 0 \qquad \text{for edges } e \subset \partial\Omega.$$

By $[\![\tau \boldsymbol{t}]\!]$ (without any subscript), we mean the function defined on the union of all edges in $\mathcal{E}$ that equals $[\![\tau \boldsymbol{t}]\!]_e$ on each edge $e \in \mathcal{E}$. With these conventions, we can now write (2.6) as

$$(\omega_h, \tau_h)_\Omega - (\boldsymbol{u}_h, \mathbf{curl}\, \tau_h)_\Omega - \sum_{e \in \mathcal{E}} (\boldsymbol{\lambda}_h,\, [\![\tau_h \boldsymbol{t}]\!])_e = (g_t, \tau_h)_{\partial\Omega}.$$

Now, proceeding as in the first hybridization, we impose the continuity of the vorticity by using the equation

$$\sum_{e \in \mathcal{E}} (\boldsymbol{\mu}_h,\, [\![\omega_h \boldsymbol{t}]\!])_e = 0 \qquad \text{for all } \boldsymbol{\mu}_h \in M_h, \tag{2.7}$$

where the space M_h is given by

$$M_h = \{\boldsymbol{\mu} : \boldsymbol{\mu} = [\![\tau \boldsymbol{t}]\!] \text{ for some } \tau \in W_h\}.$$

The above choice is dictated by the fact that a function $w \in W_h$ is continuous if and only if $[\![w \boldsymbol{t}]\!] = 0$. Clearly, if ω_h satisfies (2.7), then it belongs to the space $\mathcal{W}_h \subset H^1(\Omega)$.

Summarizing our considerations so far, the hybridized mixed method gives an approximation $(\omega_h, \boldsymbol{u}_h, \boldsymbol{\lambda}_h, p_h) \in W_h \times V_h \times M_h \times P_h$ defined by

$$(\omega_h, \tau_h)_\Omega - (\boldsymbol{u}_h, \mathbf{curl}\, \tau_h)_\Omega - \sum_{e \in \mathcal{E}} (\boldsymbol{\lambda}_h,\, [\![\tau_h \boldsymbol{t}]\!])_e = (g_t, \tau_h)_{\partial\Omega}, \tag{2.8}$$

$$(\boldsymbol{v}_h, \mathbf{curl}\, \omega_h)_\Omega + \sum_{e \in \mathcal{E}} (p_h,\, [\![\boldsymbol{v}_h \cdot \boldsymbol{n}]\!])_e = (\boldsymbol{f}, \boldsymbol{v}_h)_\Omega, \tag{2.9}$$

$$\sum_{e \in \mathcal{E}} (q_h,\, [\![\boldsymbol{u}_h \cdot \boldsymbol{n}]\!])_e = (g_n, q_h)_{\partial\Omega}, \tag{2.10}$$

$$\sum_{e \in \mathcal{E}} (\boldsymbol{\mu}_h,\, [\![\omega_h \boldsymbol{t}]\!])_e = 0 \tag{2.11}$$

for all $\tau_h \in W_h, \boldsymbol{v}_h \in V_h, q_h \in P_h$, and $\boldsymbol{\mu}_h \in M_h$. By arguments similar to those used in the proof of Proposition 2.1, it is easy to prove the following result.

Proposition 2.2. *There is a unique solution* $(\omega_h, \boldsymbol{u}_h, \boldsymbol{\lambda}_h, p_h) \in W_h \times V_h \times M_h \times P_h$ *for the hybridized mixed method* (2.8)–(2.11)*, and the solution components* ω_h *and* $\boldsymbol{u}_h$ *satisfy* (1.12)–(1.13).

At this point, the number of unknowns of our method seems to have proliferated, and it is far from evident that the hybridization we just described has any advantage at all. However, in the next section we show that the structure of this hybridized method allows us to easily eliminate the velocity $\boldsymbol{u}_h$ and the vorticity ω_h from the above equations.

3. A characterization of the Lagrange multipliers.

In this section, we eliminate the velocity and vorticity unknowns from the equations of the previously given hybridized mixed method using the methodology developed in [8]. As a result, we obtain a characterization of the tangential velocity and pressure Lagrange multipliers.

3.1. The main result.

We begin by defining *local* maps that lift functions defined on the boundary of the elements of the triangulation into functions on the domain Ω: Define $(w(\boldsymbol{\lambda}), \boldsymbol{u}(\boldsymbol{\lambda})) \in W_h \times V_h$ and $(\mathcal{w}(p), \mathbf{u}(p)) \in W_h \times V_h$ by

$$(w(\boldsymbol{\lambda}), \tau)_K - (\boldsymbol{u}(\boldsymbol{\lambda}), \mathbf{curl}\,\tau)_K = (\boldsymbol{\lambda}, \tau \boldsymbol{t})_{\partial K} \qquad \text{for all } \tau \in W_h, \tag{3.1}$$

$$(\boldsymbol{v}, \mathbf{curl}\; w(\boldsymbol{\lambda}))_K = 0 \qquad \text{for all } \boldsymbol{v} \in V_h, \tag{3.2}$$

$$(\mathcal{w}(p), \tau)_K - (\mathbf{u}(p), \mathbf{curl}\,\tau)_K = 0 \qquad \text{for all } \tau \in W_h, \tag{3.3}$$

$$(\boldsymbol{v}, \mathbf{curl}\; \mathcal{w}(p))_K = -(p, \boldsymbol{v} \cdot \boldsymbol{n})_{\partial K} \qquad \text{for all } \boldsymbol{v} \in V_h. \tag{3.4}$$

In addition it is convenient to define the local mappings $(w(g_t), \boldsymbol{u}(g_t))$ and $(\mathsf{w}(\boldsymbol{f}), \mathbf{u}(\boldsymbol{f}))$ in $W_h \times V_h$ as follows:

$$(w(g_t), \tau)_K - (\boldsymbol{u}(g_t), \mathbf{curl}\,\tau)_K = (g_t, \tau)_{\partial K \cap \partial\Omega} \qquad \text{for all } \tau \in W_h, \tag{3.5}$$

$$(\boldsymbol{v}, \mathbf{curl}\; w(g_t))_K = 0 \qquad \text{for all } \boldsymbol{v} \in V_h, \tag{3.6}$$

$$(\mathsf{w}(\boldsymbol{f}), \tau)_K - (\mathbf{u}(\boldsymbol{f}), \mathbf{curl}\,\tau)_K = 0 \qquad \text{for all } \tau \in W_h, \tag{3.7}$$

$$(\boldsymbol{v}, \mathbf{curl}\; \mathsf{w}(\boldsymbol{f}))_K = (\boldsymbol{f}, \boldsymbol{v})_K \qquad \text{for all } \boldsymbol{v} \in V_h. \tag{3.8}$$

Note that all four pairs of local maps above are given as solutions of a single mixed problem, but with different right-hand sides. That all four maps are well defined follows from the unique solvability of the mixed problem (which is the original mixed problem restricted to one element). Although all four maps use the same mixed problem, we have chosen to explicitly distinguish each of them so as to delineate the dependence of the final solution on the data components and the Lagrange multipliers.

The main result of this section characterizes the Lagrange multipliers as the unique solution of a variational equation involving the bilinear forms

$$a(\boldsymbol{\lambda}, \boldsymbol{\mu}) = (w(\boldsymbol{\lambda}), w(\boldsymbol{\mu}))_\Omega, \tag{3.9}$$

$$b(\boldsymbol{\mu}, p) = -\sum_{K \in \mathcal{T}} (\boldsymbol{u}(\boldsymbol{\mu}), \mathbf{curl}\; \mathcal{w}(p))_K, \tag{3.10}$$

$$c(p, q) = (\mathcal{w}(p), \mathcal{w}(q))_\Omega \tag{3.11}$$

and the linear functionals

$$\ell_1(\boldsymbol{\mu}) = (\boldsymbol{f}, \boldsymbol{u}(\boldsymbol{\mu}))_\Omega - (g_t, w(\boldsymbol{\mu}))_{\partial\Omega}, \tag{3.12}$$

$$\ell_2(q) = (\boldsymbol{f}, \mathbf{u}(q))_\Omega + (g_n, q)_{\partial\Omega} - (g_t, \mathcal{w}(q))_{\partial\Omega}. \tag{3.13}$$

Theorem 3.1. *The Lagrange multiplier* $(\boldsymbol{\lambda}_h, p_h) \in M_h \times P_h$ *of the hybridized mixed method* (2.8)–(2.11) *is the unique solution of*

$$a(\boldsymbol{\lambda}_h, \boldsymbol{\mu}) + b(\boldsymbol{\mu}, p_h) = \ell_1(\boldsymbol{\mu}) \qquad \text{for all } \boldsymbol{\mu} \in M_h \text{ and} \tag{3.14}$$

$$b(\boldsymbol{\lambda}_h, q) - c(p_h, q) = \ell_2(q) \qquad \text{for all } q \in P_h. \tag{3.15}$$

Moreover, the solution components ω_h *and* $\boldsymbol{u}_h$ *of the hybridized mixed method* (2.8)–(2.11) *can be determined locally as follows:*

$$\omega_h = w(\boldsymbol{\lambda}_h) + \mathcal{W}(p_h) + w(g_t) + \mathsf{w}(\boldsymbol{f}), \tag{3.16}$$

$$\boldsymbol{u}_h = \boldsymbol{u}(\boldsymbol{\lambda}_h) + \boldsymbol{\mathfrak{u}}(p_h) + \boldsymbol{u}(g_t) + \boldsymbol{\mathsf{u}}(\boldsymbol{f}). \tag{3.17}$$

3.2. Proof. To prove the above result, we follow the approach introduced in [8]. Accordingly, the first step will be to use the local maps to rewrite the first two equations of the hybridized method, namely (2.8) and (2.9). This will yield (3.16) and (3.17). Next, the two remaining equations of the hybridized method, namely (2.10) and (2.11), will be used to characterize the pressure and tangential velocity Lagrange multipliers of the method. In order to carry out these steps, we need to obtain a few identities involving the local mappings. This is done in the first lemma below. Then, in a second lemma, we show how to rewrite (2.10) and (2.11) solely in terms of the multipliers. In this way, we eliminate the vorticity and velocity and at the same time obtain a variational characterization of the Lagrange multipliers. Let us now state and prove the lemmas.

Lemma 3.2 (elementary identities). *On any mesh element* $K \in \mathcal{T}$, *for any* $\boldsymbol{\lambda} \in M_h, \boldsymbol{\mu} \in M_h, p \in P_h$, *and* $q \in P_h$, *we have the following orthogonality properties for the local vorticity maps:*

$$(w(\boldsymbol{\lambda}), \mathcal{W}(p))_K = 0, \tag{3.18}$$

$$(w(\boldsymbol{\lambda}), \mathsf{w}(\boldsymbol{f}))_K = 0, \tag{3.19}$$

$$(w(g_t), \mathcal{W}(p))_K = 0, \tag{3.20}$$

$$(w(g_t), \mathsf{w}(\boldsymbol{f}))_K = 0. \tag{3.21}$$

Moreover, we have the following identities for the bilinear forms a, b, and c:

$$a_K(\boldsymbol{\lambda}, \boldsymbol{\mu}) := (w(\boldsymbol{\lambda}), w(\boldsymbol{\mu}))_K = (\boldsymbol{\lambda}, w(\boldsymbol{\mu})\boldsymbol{t})_{\partial K}, \tag{3.22}$$

$$b_K(\boldsymbol{\lambda}, p) := -(\boldsymbol{u}(\boldsymbol{\lambda}), \mathbf{curl}\ \mathcal{W}(p))_K = (\boldsymbol{u}(\boldsymbol{\lambda}) \cdot \boldsymbol{n}, p)_{\partial K} = (\boldsymbol{\lambda}, \mathcal{W}(p)\boldsymbol{t})_{\partial K}, \tag{3.23}$$

$$c_K(p, q) := (\mathcal{W}(p), \mathcal{W}(q))_K = -(q, \boldsymbol{\mathfrak{u}}(p) \cdot \boldsymbol{n})_{\partial K}. \tag{3.24}$$

Finally, we have the following identities related to the linear forms ℓ_1 *and* ℓ_2*:*

$$(\boldsymbol{f}, \boldsymbol{u}(\boldsymbol{\mu}))_K = -(\mathsf{w}(\boldsymbol{f})\boldsymbol{t}, \boldsymbol{\mu})_{\partial K}, \tag{3.25}$$

$$(w(\boldsymbol{\mu}), g_t)_{\partial K \cap \partial \Omega} = (\boldsymbol{\mu}, w(g_t)\, \boldsymbol{t})_{\partial K}, \tag{3.26}$$

$$(\boldsymbol{f}, \boldsymbol{\mathfrak{u}}(q))_K = -(\boldsymbol{\mathsf{u}}(\boldsymbol{f}) \cdot \boldsymbol{n}, q)_{\partial K}, \tag{3.27}$$

$$(\mathcal{W}(q), g_t)_{\partial K \cap \partial \Omega} = (q, \boldsymbol{u}(g_t) \cdot \boldsymbol{n})_{\partial K}. \tag{3.28}$$

Proof. Let us begin by proving the orthogonality identities. Equation (3.18) is obtained by setting $\tau = w(\boldsymbol{\lambda})$ in (3.3) and using (3.2). The proof of (3.19) is analogous. Equations (3.20) and (3.21) follow from similar arguments, as the equations defining the liftings of g_t and $\boldsymbol{\lambda}$ have the same structure.

Next, let us prove the identities associated with the bilinear forms $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, and $c(\cdot,\cdot)$. Equation (3.22) is obtained as follows. Setting $\tau = w(\boldsymbol{\mu})$ in the definition of the liftings (3.1), we get

$$\begin{aligned}(w(\boldsymbol{\lambda}), \boldsymbol{w}(\boldsymbol{\mu}))_K &= (\boldsymbol{u}(\boldsymbol{\lambda}), \mathbf{curl}\; w(\boldsymbol{\mu}))_K + (\boldsymbol{\lambda}, w(\boldsymbol{\mu})\,\boldsymbol{t})_{\partial K}\\ &= (\boldsymbol{\lambda}, w(\boldsymbol{\mu})\,\boldsymbol{t})_{\partial K},\end{aligned}$$

by (3.2) with $\boldsymbol{\lambda} = \boldsymbol{\mu}$ and $\boldsymbol{v} = \boldsymbol{u}(\boldsymbol{\lambda})$. Let us prove (3.23). Taking $\boldsymbol{v} = \boldsymbol{u}(\boldsymbol{\lambda})$ in (3.4), we get

$$\begin{aligned}(\boldsymbol{u}(\boldsymbol{\lambda})\cdot\boldsymbol{n}, p)_{\partial K} &= -(\boldsymbol{u}(\boldsymbol{\lambda}), \mathbf{curl}\; \mathrm{w}(p))_K\\ &= (\boldsymbol{\lambda}, \mathrm{w}(p)\,\boldsymbol{t})_{\partial K} - (w(\boldsymbol{\lambda}), \mathrm{w}(p))_K \quad \text{by (3.1) with } \tau = \mathrm{w}(p),\\ &= (\boldsymbol{\lambda}, \mathrm{w}(p)\,\boldsymbol{t})_{\partial K},\end{aligned}$$

by the orthogonality property (3.18). Now let us prove (3.24). We have, by (3.3) with $\tau = \mathrm{w}(q)$,

$$\begin{aligned}(\mathrm{w}(p), \mathrm{w}(q))_K &= (\mathbf{u}(p), \mathbf{curl}\; \mathrm{w}(q))_K\\ &= -(q, \mathbf{u}(p)\cdot\boldsymbol{n})_{\partial K},\end{aligned}$$

by (3.4) with $p = q$ and $\boldsymbol{v} = \mathbf{u}(p)$.

Finally, let us consider the last set of identities. We first prove (3.25). Setting $\boldsymbol{v} = \boldsymbol{u}(\boldsymbol{\mu})$ in (3.8), we get

$$\begin{aligned}(\boldsymbol{f}, \boldsymbol{u}(\boldsymbol{\mu}))_K &= (\boldsymbol{u}(\boldsymbol{\mu}), \mathbf{curl}\; \mathsf{w}(\boldsymbol{f}))_K\\ &= (w(\boldsymbol{\mu}), \mathsf{w}(\boldsymbol{f}))_K - (\boldsymbol{\mu}, \mathsf{w}(\boldsymbol{f})\,\boldsymbol{t})_{\partial K},\end{aligned}$$

by (3.1) with $\boldsymbol{\lambda} = \boldsymbol{\mu}$ and $\tau = \mathsf{w}(\boldsymbol{f})$. The desired equation follows by using the already established orthogonality property (3.19). Next, let us prove (3.26). Setting $\tau = w(\boldsymbol{\mu})$ in (3.5) and then using (3.2), we get

$$\begin{aligned}(g_t, w(\boldsymbol{\mu}))_{\partial K\cap\partial\Omega} &= (w(\boldsymbol{\mu}), w(g_t))_K\\ &= (\boldsymbol{\mu}, w(g_t)\boldsymbol{t})_{\partial K} + (\boldsymbol{u}(\boldsymbol{\mu}), \mathbf{curl}\,(w(g_t)))_K,\end{aligned}$$

by (3.1) with $\boldsymbol{\lambda} = \boldsymbol{\mu}$ and $\tau = w(g_t)$. Equation (3.26) follows from (3.6). Equation (3.27) is obtained as follows:

$$\begin{aligned}(\boldsymbol{f}, \mathbf{u}(q))_K &= (\mathbf{u}(q), \mathbf{curl}\; \mathsf{w}(\boldsymbol{f}))_K && \text{by (3.8) with } \boldsymbol{v} = \mathbf{u}(q),\\ &= (\mathrm{w}(q), \mathsf{w}(\boldsymbol{f}))_K && \text{by (3.3) with } p = q \text{ and } \tau = \mathsf{w}(\boldsymbol{f}),\\ &= (\mathbf{u}(\boldsymbol{f}), \mathbf{curl}\; \mathrm{w}(q))_K && \text{by (3.7) with } \tau = \mathrm{w}(q),\\ &= -(q, \mathbf{u}(\boldsymbol{f}))_K,\end{aligned}$$

by (3.4) with $p = q$ and $\boldsymbol{v} = \mathbf{u}(\boldsymbol{f})$. Finally, let us prove (3.28). By (3.5) with $\tau = \mathrm{w}(q)$, we have

$$\begin{aligned}(\mathrm{w}(q), g_t)_{\partial K\cup\partial\Omega} &= (w(g_t), \mathrm{w}(q))_K - (\boldsymbol{u}(g_t), \mathbf{curl}\; \mathrm{w}(q))_K,\\ &= -(\boldsymbol{u}(g_t), \mathbf{curl}\; \mathrm{w}(q))_K \qquad \text{by (3.20)},\\ &= (q, \boldsymbol{u}(g_t)\cdot\boldsymbol{n})_{\partial K},\end{aligned}$$

by (3.4) with $p = q$ and $\boldsymbol{v} = \boldsymbol{u}(g_t)$. This completes the proof. $\square$

LEMMA 3.3 (the jump conditions). *For arbitrary $\boldsymbol{\lambda} \in M_h$ and $p \in P_h$ set*

$$\widetilde{\omega}_h^{\boldsymbol{\lambda},p} = w(\boldsymbol{\lambda}) + \mathcal{W}(p) + w(g_t) + \mathsf{w}(\boldsymbol{f}),$$
$$\widetilde{\boldsymbol{u}}_h^{\boldsymbol{\lambda},p} = \boldsymbol{u}(\boldsymbol{\lambda}) + \boldsymbol{\mathfrak{u}}(p) + \boldsymbol{u}(g_t) + \boldsymbol{\mathsf{u}}(\boldsymbol{f}).$$

Let $(\omega_h, \boldsymbol{u}_h, \boldsymbol{\lambda}_h, p_h)$ be the unique solution of (2.8)–(2.11). *Then the following statements are equivalent:*

A. *For all $\boldsymbol{\mu} \in M_h$ and $q \in P_h$,*

$$\sum_{e\in\mathcal{E}} (\boldsymbol{\mu},\, \llbracket \widetilde{\omega}_h^{\boldsymbol{\lambda},p} \boldsymbol{t} \rrbracket)_e = 0 \quad \textit{and} \quad \sum_{e\in\mathcal{E}} (q,\, \llbracket \widetilde{\boldsymbol{u}}_h^{\boldsymbol{\lambda},p} \cdot \boldsymbol{n} \rrbracket)_e = (g_n, q)_{\partial\Omega}.$$

B. $\widetilde{\omega}_h^{\boldsymbol{\lambda},p} = \omega_h$ *and* $\widetilde{\boldsymbol{u}}_h^{\boldsymbol{\lambda},p} = \boldsymbol{u}_h$.

C. $\boldsymbol{\lambda} = \boldsymbol{\lambda}_h$ *and* $p = p_h$.

D. $a(\boldsymbol{\lambda}, \boldsymbol{\mu}) + b(\boldsymbol{\mu}, p) = \ell_1(\boldsymbol{\mu})$ *for all* $\boldsymbol{\mu} \in M_h$ *and* $b(\boldsymbol{\lambda}, q) - c(p, q) = \ell_2(q)$ *for all* $q \in P_h$.

Proof. A $\Longrightarrow$ B: By adding the equations defining $(w(\boldsymbol{\lambda}), \boldsymbol{u}(\boldsymbol{\lambda}))$, $(\mathcal{W}(p), \boldsymbol{\mathfrak{u}}(p))$, $(\mathsf{w}(\boldsymbol{f}), \boldsymbol{\mathsf{u}}(\boldsymbol{f}))$, and $(w(g_t), \boldsymbol{u}(g_t))$, we find that $\widetilde{\omega}_h^{\boldsymbol{\lambda},p}$ and $\widetilde{\boldsymbol{u}}_h^{\boldsymbol{\lambda},p}$ satisfy the first two equations of our hybridized mixed method, i.e.,

$$(\widetilde{\omega}_h^{\boldsymbol{\lambda},p}, \tau_h)_\Omega - (\widetilde{\boldsymbol{u}}_h^{\boldsymbol{\lambda},p}, \mathbf{curl}\,\tau_h)_\Omega - \sum_{e\in\mathcal{E}} (\boldsymbol{\lambda},\, \llbracket \tau_h \boldsymbol{t} \rrbracket)_e = (g_t, \tau_h)_{\partial\Omega},$$
$$(\boldsymbol{v}_h, \mathbf{curl}\,\widetilde{\omega}_h^{\boldsymbol{\lambda},p})_\Omega + \sum_{e\in\mathcal{E}} (p,\, \llbracket \boldsymbol{v}_h \cdot \boldsymbol{n} \rrbracket)_e = (\boldsymbol{f}, \boldsymbol{v}_h)_\Omega,$$

for all $\tau_h \in W_h$ and $\boldsymbol{v}_h \in V_h$. Since statement A holds, they also satisfy the remaining equations of the method. By uniqueness of solutions of the hybridized mixed method (as given by Proposition 2.2) we get statement B.

B $\Longrightarrow$ C: By linear superposition,

$$\omega_h = w(\boldsymbol{\lambda}_h) + \mathcal{W}(p_h) + w(g_t) + \mathsf{w}(\boldsymbol{f}), \tag{3.29}$$
$$\boldsymbol{u}_h = \boldsymbol{u}(\boldsymbol{\lambda}_h) + \boldsymbol{\mathfrak{u}}(p_h) + \boldsymbol{u}(g_t) + \boldsymbol{\mathsf{u}}(\boldsymbol{f}). \tag{3.30}$$

Comparing these equations with the definitions of $\widetilde{\omega}_h^{\boldsymbol{\lambda},p}$ and $\widetilde{\boldsymbol{u}}_h^{\boldsymbol{\lambda},p}$, we find that statement B implies

$$w(\boldsymbol{\lambda}_h) + \mathcal{W}(p_h) = w(\boldsymbol{\lambda}) + \mathcal{W}(p), \tag{3.31}$$
$$\boldsymbol{u}(\boldsymbol{\lambda}_h) + \boldsymbol{\mathfrak{u}}(p_h) = \boldsymbol{u}(\boldsymbol{\lambda}) + \boldsymbol{\mathfrak{u}}(p). \tag{3.32}$$

In particular,

$$w(\boldsymbol{\lambda}_h - \boldsymbol{\lambda}) + \mathcal{W}(p_h - p) = 0.$$

Since the two terms on the left-hand side above are $L^2(\Omega)$-orthogonal by (3.18), they both must vanish. Moreover, by the definition of $\boldsymbol{\mathfrak{u}}(\cdot)$ (see (3.3)), $w(\boldsymbol{\lambda}_h - \boldsymbol{\lambda}) = 0$ implies

$$(\boldsymbol{\mathfrak{u}}(p_h - p), \mathbf{curl}\,\tau)_K = 0 \quad \text{for all } K \in \mathfrak{T}, \tau \in W_h.$$

Hence $\boldsymbol{\mathfrak{u}}(p_h - p) = 0$. By (3.32) we also get $\boldsymbol{u}(\boldsymbol{\lambda}_h - \boldsymbol{\lambda}) = 0$. Thus,

$$w(\boldsymbol{\lambda}_h - \boldsymbol{\lambda}) = \mathcal{W}(p_h - p) = 0, \qquad \boldsymbol{u}(\boldsymbol{\lambda}_h - \boldsymbol{\lambda}) = \boldsymbol{\mathfrak{u}}(p_h - p) = \mathbf{0},$$

so $\boldsymbol{\lambda}_h - \boldsymbol{\lambda} = 0$ and $p_h - p = 0$.

C $\implies$ D: We know by (3.29)–(3.30) and the last two equations of the hybridized mixed method that

$$\Theta := \sum_{e\in\mathcal{E}} \big(\boldsymbol{\mu},\, [\![\big(w(\boldsymbol{\lambda}_h) + \mathscr{w}(p_h) + \mathsf{w}(\boldsymbol{f}) + w(g_t)\big)\boldsymbol{t}]\!]\big)_e = 0,$$

$$\Psi := \sum_{e\in\mathcal{E}} \big(q,\, [\![\big(\boldsymbol{u}(\boldsymbol{\lambda}_h) + \mathbf{u}(p_h) + \mathbf{u}(\boldsymbol{f}) + \boldsymbol{u}(g_t)\big)\cdot\boldsymbol{n}]\!]\big)_e - (g_n, q)_{\partial\Omega} = 0.$$

Hence, it suffices to show that

$$\Theta = a(\boldsymbol{\lambda},\boldsymbol{\mu}) + b(\boldsymbol{\mu}, p) - \ell_1(\boldsymbol{\mu}), \tag{3.33}$$

$$\Psi = b(\boldsymbol{\lambda}, q) - c(p,q) - \ell_2(q). \tag{3.34}$$

To do this, let us split $\Theta =: \theta_1 + \theta_2 + \theta_3 + \theta_4$, where

$$\begin{aligned}
\theta_1 &:= \sum_{e\in\mathcal{E}} \big(\boldsymbol{\mu},\, [\![w(\boldsymbol{\lambda})]\!]\boldsymbol{t}\big)_e = (w(\boldsymbol{\lambda}), w(\boldsymbol{\mu}))_\Omega && \text{by (3.22)},\\
\theta_2 &:= \sum_{e\in\mathcal{E}} \big(\boldsymbol{\mu},\, [\![\mathscr{w}(p)\boldsymbol{t}]\!]\big)_e = -\sum_{K\in\mathcal{T}} (\boldsymbol{u}(\boldsymbol{\mu}), \mathbf{curl}\ \mathscr{w}(p))_K && \text{by (3.23)},\\
\theta_3 &:= \sum_{e\in\mathcal{E}} \big(\boldsymbol{\mu},\, [\![\mathsf{w}(\boldsymbol{f})\boldsymbol{t}]\!]\big)_e = -(\boldsymbol{f}, \boldsymbol{u}(\boldsymbol{\mu}))_\Omega && \text{by (3.25)},\\
\theta_4 &:= \sum_{e\in\mathcal{E}} \big(\boldsymbol{\mu},\, [\![w(g_t)\boldsymbol{t}]\!]\big)_e = (g_t, w(\boldsymbol{\mu}))_{\partial\Omega} && \text{by (3.26)}.
\end{aligned}$$

Hence

$$\begin{aligned}
\theta_1 &= a(\boldsymbol{\lambda},\boldsymbol{\mu}) && \text{by (3.9)},\\
\theta_2 &= b(\boldsymbol{\mu}, p) && \text{by (3.10)},\\
\theta_3 + \theta_4 &= -\ell_1(\boldsymbol{\mu}) && \text{by (3.12)}.
\end{aligned}$$

This proves (3.33).

To prove (3.34), we split $\Psi =: \psi_1 + \psi_2 + \psi_3 + \psi_4 + \psi_5$, where

$$\begin{aligned}
\psi_1 &:= \sum_{e\in\mathcal{E}} \big(q,\, [\![\boldsymbol{u}(\boldsymbol{\lambda})\cdot\boldsymbol{n}]\!]\big)_e = -\sum_{K\in\mathcal{T}} (\boldsymbol{u}(\boldsymbol{\lambda}), \mathbf{curl}\ \mathscr{w}(q))_K && \text{by (3.23)},\\
\psi_2 &:= \sum_{e\in\mathcal{E}} \big(q,\, [\![\mathbf{u}(p)\cdot\boldsymbol{n}]\!]\big)_e = -(\mathscr{w}(p), \mathscr{w}(q))_\Omega && \text{by (3.24)},\\
\psi_3 &:= \sum_{e\in\mathcal{E}} \big(q,\, [\![\mathbf{u}(\boldsymbol{f})\cdot\boldsymbol{n}]\!]\big)_e = -(\boldsymbol{f}, \mathbf{u}(q))_\Omega && \text{by (3.27)},\\
\psi_4 &:= \sum_{e\in\mathcal{E}} \big(q,\, [\![\boldsymbol{u}(g_t)\cdot\boldsymbol{n}]\!]\big)_e = (g_t, \mathscr{w}(q))_{\partial\Omega} && \text{by (3.28)},\\
\psi_5 &:= -(g_n, q)_{\partial\Omega}.
\end{aligned}$$

Hence

$$\begin{aligned}
\psi_1 &= b(\boldsymbol{\lambda}, q) && \text{by (3.10)},\\
\psi_2 &= -c(p,q) && \text{by (3.11)},\\
\psi_3 + \psi_4 + \psi_5 &= -\ell_2(\boldsymbol{\mu}) && \text{by (3.13)}.
\end{aligned}$$

Adding the above equations, we obtain (3.34).

D $\Longrightarrow$ A: If statement D holds, then, by the previous step, we have

$$\sum_{e\in\mathcal{E}} \big(\boldsymbol{\mu},\, [\![\big(w(\boldsymbol{\lambda}) + \mathcal{w}(p) + \mathsf{w}(\boldsymbol{f}) + w(g_t)\big)\boldsymbol{t}]\!]\big)_e = 0,$$
$$\sum_{e\in\mathcal{E}} \big(q,\, [\![\big(\boldsymbol{u}(\boldsymbol{\lambda}) + \mathbf{u}(p) + \mathbf{u}(\boldsymbol{f}) + \boldsymbol{u}(g_t)\big)\cdot\boldsymbol{n}]\!]\big)_e = (g_n, q)_{\partial\Omega},$$

which is statement A. □

Proof of Theorem 3.1. The proof of the theorem is immediate from the previous lemmas: The first assertion of the theorem follows from the equivalence of statements C and D of Lemma 3.3. The second follows from the first by linear superposition. Thus Theorem 3.1 is proved. □

4. Local bases for Lagrange multipliers.

For the hybridized method to be of practical use, it is imperative that we develop computable bases of locally supported functions for the multiplier spaces P_h and M_h.

4.1. The pressure space.

We begin with a characterization of the space of pressure Lagrange multipliers arising from the first hybridization.

PROPOSITION 4.1. *The space P_h defined in* (2.1) *is characterized by*

$$P_h = \left\{ p : p|_e \in P_k(e) \text{ for all } e\in\mathcal{E} \text{ and } \sum_{e\in\mathcal{E}} (p,1)_e = 0 \right\}.$$

Proof. Let Q_h denote the set in the right-hand side above. To show that $P_h \subseteq Q_h$, consider any $\boldsymbol{v}_h \in V_h$ and let $p_h = [\![\boldsymbol{v}_h\cdot\boldsymbol{n}]\!]$. Then $[\![\boldsymbol{v}_h\cdot\boldsymbol{n}]\!]_e \in P_k(e)$ and

$$\sum_{e\in\mathcal{E}} (p_h,1)_e \,\mathrm{d}s = \sum_{K\in\mathcal{T}} (\boldsymbol{v}_h\cdot\boldsymbol{n},1)_{\partial K} = \sum_{K\in\mathcal{T}} (\operatorname{div}\boldsymbol{v}_h,1)_K = 0.$$

Hence $P_h \subseteq Q_h$.

To show the reverse inclusion, consider any $p_h \in Q_h$. Then there is a function $\widetilde{\boldsymbol{v}}_h \in \widetilde{V}_h := \{\boldsymbol{r} : \boldsymbol{r}|_K = \boldsymbol{x}p_k(\boldsymbol{x}) + \boldsymbol{q}_k$ for some $p_k \in P_k(K)$ and $\boldsymbol{q}_k \in P_k(K)^2\}$ such that

$$[\![\widetilde{\boldsymbol{v}}_h\cdot\boldsymbol{n}]\!]_e = p_h|_e \quad \text{for all } e\in\mathcal{E}.$$

Note that $\operatorname{div}(\widetilde{\boldsymbol{v}}_h|_K)$ is not zero in general. Let S_h be the space of functions whose average on Ω is zero and whose restriction to each mesh element $K\in\mathcal{T}$ is in $P_k(K)$. The function $s_h(\boldsymbol{x})$, defined by

$$s_h|_K = \operatorname{div}(\widetilde{\boldsymbol{v}}_h|_K) \quad \text{for all } K\in\mathcal{T},$$

is in S_h because p_h is in Q_h:

$$(s_h,1)_\Omega = \sum_{K\in\mathcal{T}} (\operatorname{div}\widetilde{\boldsymbol{v}}_h,1)_K = \sum_{K\in\mathcal{T}} (\widetilde{\boldsymbol{v}}_h\cdot\boldsymbol{n},1) = \sum_{e\in\mathcal{E}} (p_h,1)_e = 0.$$

Now, the space $\widetilde{V}_h \cap H_0(\operatorname{div},\Omega)$ is a standard Raviart–Thomas space, and by its well-known properties, $\operatorname{div} : \widetilde{V}_h \cap H_0(\operatorname{div},\Omega) \mapsto S_h$ is a surjection. Hence, there is a $\boldsymbol{z}_h \in \widetilde{V}_h \cap H_0(\operatorname{div},\Omega)$ such that

$$\operatorname{div}\boldsymbol{z}_h = s_h.$$

Then $\boldsymbol{v}_h = \widetilde{\boldsymbol{v}}_h - \boldsymbol{z}_h$ is in V_h and $[\![\boldsymbol{v}_h \cdot \boldsymbol{n}]\!]_e = [\![\widetilde{\boldsymbol{v}}_h \cdot \boldsymbol{n}]\!]_e = p_h|_e$. Hence $Q_h \subseteq P_h$. □

In view of Proposition 4.1, the function q_h belongs to the space P_h if and only if it belongs to

$$\widetilde{P}_h = \{p : p|_e \in P_k(e) \text{ for all } e \in \mathcal{E}\}$$

and satisfies

$$\sum_{e\in\mathcal{E}} (q_h, 1)_e = 0. \tag{4.1}$$

Thus we need only construct a local basis for $\widetilde{P}_h$ and then enforce the last equation. Obviously, we can construct a basis for $\widetilde{P}_h$ by taking the union of local bases for $P_k(e)$, say Legendre polynomials, on every edge $e \in \mathcal{E}$. In practice, the constraint (4.1) can be handled a posteriori in a very simple way, as shown in section 5.

4.2. The lowest-order tangential velocity space. In the remainder of this section, we construct a local basis for the space M_h of tangential velocity Lagrange multipliers. In this subsection, we study the lowest-order case. In the next, we show how our considerations here generalize to the higher-order case.

In order to explicitly give a local basis for M_h, we introduce some more notation. Let K be a mesh triangle and let x be one of its vertices. We denote by $\mathit{\Lambda}_{x,K}$ the union of the two edges of K that are connected to the vertex x. Let

$$\hat{\mathit{\Lambda}}_h = \{\mathit{\Lambda}_{x,K} : x \text{ is a vertex of } \mathcal{T} \text{ and } K \in \mathcal{T}\}.$$

For all $\mathit{\Lambda} \in \hat{\mathit{\Lambda}}_h$, we denote by $K_\mathit{\Lambda}$ the (unique) triangle $K \in \mathcal{T}$ such that $\mathit{\Lambda} \subseteq \partial K$, and by $x_\mathit{\Lambda}$ we denote the common vertex of $\mathit{\Lambda}$ and $K_\mathit{\Lambda}$. Let $\phi_\mathit{\Lambda}$ denote the function (that is discontinuous in general) which vanishes on all $K \in \mathcal{T}$ except on $K_\mathit{\Lambda}$, where it equals the linear function that is one on $x_\mathit{\Lambda}$ and zero on the remaining two vertices of $K_\mathit{\Lambda}$. We define a basis for M_h using the functions

$$\boldsymbol{\psi}_\mathit{\Lambda} = [\![\phi_\mathit{\Lambda}\, \boldsymbol{t}]\!].$$

Obviously $\boldsymbol{\psi}_\mathit{\Lambda} \in M_h$, but not all of $\boldsymbol{\psi}_\mathit{\Lambda}, \mathit{\Lambda} \in \hat{\mathit{\Lambda}}_h$, are linearly independent; e.g., the functions $\boldsymbol{\psi}_\mathit{\Lambda}$ for all $\mathit{\Lambda}$ connected to one vertex are linked by one equation. Therefore, for every mesh vertex x (including $x \in \partial\Omega$), we arbitrarily pick one element $\mathit{\Lambda} \in \hat{\mathit{\Lambda}}_h$ with vertex $x_\mathit{\Lambda} = x$, denote it by $\triangledown_x$ (see Figure 2), and "omit" it: Define

$$\mathit{\Lambda}_h = \hat{\mathit{\Lambda}}_h \setminus \{\triangledown_x : \text{ for all mesh vertices } x\}.$$

Proposition 4.2. *The set $\mathcal{B} = \{\boldsymbol{\psi}_\mathit{\Lambda} : \mathit{\Lambda} \in \mathit{\Lambda}_h\}$ is a basis for M_h when $k = 0$.*

Proof. Obviously the span of $\mathcal{B}$ is contained in M_h. Hence it suffices to prove that

$$\operatorname{card} \mathcal{B} = \dim M_h \tag{4.2}$$

and

$$\mathcal{B} \text{ is a linearly independent set.} \tag{4.3}$$

To prove (4.2), let us first count the dimension of M_h. Defining $T_h : W_h \mapsto M_h$ by

$$T_h \tau = [\![\tau \boldsymbol{t}]\!],$$

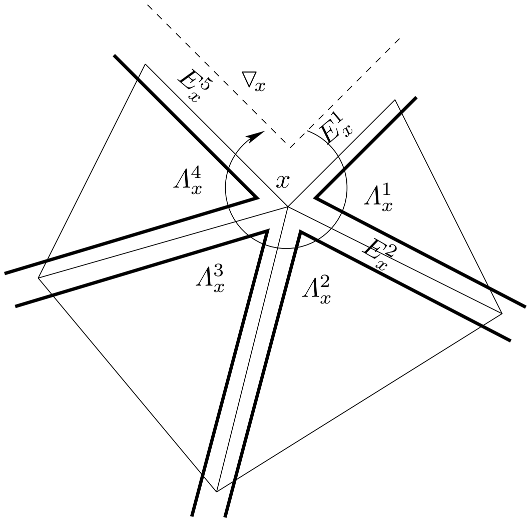


FIG. 2. *Construction of basis functions supported near a mesh vertex x.*

we note that M_h is the range of T_h. Since the null space of T_h is $\mathcal{W}_h$, by the rank-nullity theorem, we find that

$$\dim(M_h) = \operatorname{rank}(T_h) = \dim(W_h) - \dim(\mathcal{W}_h). \tag{4.4}$$

In the lowest-order case, this easily gives

$$\dim(M_h) = 3n_K - n_V,$$

where n_K and n_V are the number of triangles and vertices of the mesh, respectively. Now, since

$$\operatorname{card}\mathcal{B} = \operatorname{card}\Lambda_h = \operatorname{card}\hat{\Lambda}_h - n_V = 3n_K - n_V,$$

we immediately see that (4.2) holds.

To prove (4.3), let $\boldsymbol{\mu}$ be any linear combination of the basis elements:

$$\boldsymbol{\mu} = \sum_{\Lambda\in\Lambda_h} c_\Lambda \boldsymbol{\psi}_\Lambda. \tag{4.5}$$

Then, consider $\boldsymbol{\mu}|_{\nabla_x}$ for any mesh vertex x (including $x \in \partial\Omega$). Enumerate all $\Lambda \in \Lambda_h$ with vertex x as $\Lambda_x^1, \Lambda_x^2, \ldots, \Lambda_x^{N_x}$ and all edges in $\mathcal{E}$ connected to x as $E_x^1, E_x^2, \ldots E_x^{N_x+1}$, as in Figure 2. The enumerations are such that the two edges of Λ_x^j are E_x^j and E_x^{j+1}. Let μ_x^i be the function defined on E_x^i that equals the magnitude of $\boldsymbol{\mu}|_{E_x^i}$. Observe that the limit of $\mu_x^1(y)$ as y approaches x along the edge E_x^1 is $|c_{\Lambda_x^1}|$. Similarly, the limit of $\mu_x^{N_x}(y)$ as y approaches x along the edge $E_x^{N_x+1}$ is $|c_{\Lambda_x^{N_x}}|$. Also note that the limit of $\mu_x^j(y)$ as y approaches x along the edge E_x^j is $|c_{\Lambda_x^j} - c_{\Lambda_x^{j-1}}|$ for all $j = 2, 3, \ldots, N_x - 1$.

Now suppose $\boldsymbol{\mu} \equiv \mathbf{0}$. We have to show that all the coefficients c_Λ in (4.5) are zero. Since $\boldsymbol{\mu}$ vanishes everywhere, in particular, for a mesh vertex x, the function $\mu_x^j(y)$ vanishes on the edge E_x^j. Hence its limit as y approaches x along the edge E_x^j

equals zero. Thus,

$$\begin{aligned} |c_{\Lambda_x^1}| = |c_{\Lambda_x^{N_x}}| &= 0 && \text{and} \\ |c_{\Lambda_x^j} - c_{\Lambda_x^{j-1}}| &= 0 && \text{for all } j = 2, \ldots, N_x - 1. \end{aligned}$$

This implies that $c_{\Lambda_x^j} = 0$ for all j. The above argument applies to every mesh vertex, so all the coefficients c_Λ in (4.5) are zero. Hence (4.3) follows. □

4.3. Basis for the space of tangential velocities. By augmenting the basis $\mathcal{B}$ for the lowest-order case constructed above with some locally supported functions, it is possible to construct a basis for M_h of any order. Define $\mathcal{B}_e^{(k+1)}$ to be any basis for the set of polynomials on edge e of degree at most $k+1$ that vanishes at both endpoints of e. Let $\mathcal{E}_0$ denote the set of all *interior* edges of the mesh $\mathcal{T}$. Then we have the following result.

THEOREM 4.3. *The set*

$$\mathcal{B}^{(k+1)} = \left(\bigcup_{e \in \mathcal{E}_0} \mathcal{B}_e^{(k+1)} \right) \cup \mathcal{B}$$

is a basis for M_h.

Proof. It is easy to see that each element of $\mathcal{B}_e^{(k+1)}$ can be written as $[\![\phi \boldsymbol{t}]\!]$ for some $\phi \in W_h$. Hence the span of $\mathcal{B}^{(k+1)}$ is contained in M_h. As in the proof of Proposition 4.2, it now suffices to prove that

$$\operatorname{card} \mathcal{B}^{(k+1)} = \dim(M_h) \tag{4.6}$$

and that $\mathcal{B}^{(k+1)}$ is a linearly independent set. Since functions in $\mathcal{B}_e^{(k+1)}$ vanish at endpoints of their edge of support, by a minor modification of the arguments in the proof of Proposition 4.2, the linear independence of $\mathcal{B}^{(k+1)}$ follows.

To prove (4.6), observe that $\operatorname{card} \mathcal{B}_e^{(k+1)} = \dim(P_{k+1}(e)) - 2 = k$. Since

$$\operatorname{card} \mathcal{E}_0 = 3n_K - n_E,$$

where n_E denotes the number of all edges of $\mathcal{T}$ (including boundary edges), we have

$$\begin{aligned} \operatorname{card} \mathcal{B}^{(k+1)} &= \operatorname{card} \mathcal{B} + \sum_{e \in \mathcal{E}_0} \operatorname{card} \mathcal{B}_e^{(k+1)} \\ &= (3n_K - n_V) + (3n_K - n_E)k \\ &= 3n_K(k+1) - n_V - kn_E. \end{aligned} \tag{4.7}$$

Now, let us show that this equals $\dim(M_h)$. Since, by (4.4), $\dim(M_h) = \dim(W_h) - \dim(\mathcal{W}_h)$, we need to compute the dimension of the spaces W_h and $\mathcal{W}_h$. The number of degrees of freedom of $\mathcal{W}_h$ can be computed by splitting them into vertex degrees of freedom (one per vertex), edge degrees of freedom (k per edge), and interior degrees of freedom ($\dim(P^{k-2})$ per triangle):

$$\dim(\mathcal{W}_h) = n_K \left(\frac{1}{2}(k-1)k \right) + kn_E + n_V.$$

Now, since

$$\dim(W_h) = n_K \frac{1}{2}(k+2)(k+3) = n_K \left(3(k+1) + \frac{1}{2}(k-1)k \right),$$

the dimension of M_h can immediately be seen to be equal to $\operatorname{card} \mathcal{B}^{(k+1)}$, as calculated in (4.7). Hence (4.6) follows. □

5. Some implementation aspects. In this section, we first point out some general issues in implementing the Lagrange multiplier system. Afterward we specialize to a detailed discussion of the lowest-order case. We exhibit explicit expressions for all the local mappings in the lowest-order case. We also show how traditional finite element ideas such as matrix assembly through local element stiffness matrices apply for the Lagrange multiplier system, provided that the local matrices are properly defined.

5.1. The matrix equations. In order to solve for $\boldsymbol{\lambda}_h$ and p_h satisfying (3.14)–(3.15), we use the previously introduced local basis. Let $\boldsymbol{\psi}^{(i)}$, $i = 1, 2, \ldots, N_M$, be an enumeration of the basis for M_h introduced in section 4. Let $p^{(l)}$, $l = 1, 2, \ldots, N_P$, denote any basis for $\widetilde{P}_h$ with the property that a basis function is supported on just one edge. With respect to these bases, let A, B, and C denote the matrices associated to the bilinear forms $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, and $c(\cdot,\cdot)$, respectively:

$$\begin{aligned} \mathsf{A}_{ij} &= (w(\boldsymbol{\psi}^{(j)}), w(\boldsymbol{\psi}^{(i)}))_\Omega, \\ \mathsf{B}_{lj} &= -(\mathbf{curl}\, \mathcal{w}(p^{(l)}), \boldsymbol{u}(\boldsymbol{\psi}^{(j)}))_\Omega, \\ \mathsf{C}_{lm} &= (\mathcal{w}(p^{(m)}), \mathcal{w}(p^{(l)}))_\Omega. \end{aligned}$$

Then the Lagrange multiplier system (3.14)–(3.15) takes the following matrix form:

$$\begin{bmatrix} \mathsf{A} & \mathsf{B}^t \\ \mathsf{B} & -\mathsf{C} \end{bmatrix} \begin{bmatrix} \Lambda \\ \mathsf{P} \end{bmatrix} = \begin{bmatrix} \mathsf{L}_1 \\ \mathsf{L}_2 \end{bmatrix}. \tag{5.1}$$

Here the Λ and P are vectors of coefficients of $\boldsymbol{\lambda}_h$ and p_h, respectively, i.e.,

$$\boldsymbol{\lambda}_h = \sum_{i=1}^{N_M} \Lambda_i\, \boldsymbol{\psi}^{(i)} \quad \text{and} \quad p_h = \sum_{l=1}^{N_P} \mathsf{P}_l\, p^{(l)}.$$

Notice that we have used a basis for the space $\widetilde{P}_h$ and not for the space P_h. In view of Proposition 4.1, we therefore anticipate the pressure to be given only up to a constant. This, of course, reflects the fact that the pressure in the Stokes system is also defined up to a constant.

To clarify how one can deal with this in practical implementations, let us examine the null space of

$$\mathsf{M} := \begin{bmatrix} \mathsf{A} & \mathsf{B}^t \\ \mathsf{B} & -\mathsf{C} \end{bmatrix}.$$

If $\mathsf{M} \left[\begin{smallmatrix} \Lambda \\ \mathsf{P} \end{smallmatrix} \right] = 0$, then

$$a(\boldsymbol{\lambda}_h, \boldsymbol{\mu}) + b(\boldsymbol{\mu}, p_h) = 0 \qquad \text{for all } \boldsymbol{\mu} \in M_h \text{ and} \tag{5.2}$$

$$b(\boldsymbol{\lambda}_h, q) - c(p_h, q) = 0 \qquad \text{for all } q \in \widetilde{P}_h. \tag{5.3}$$

Now, an immediate consequence of the definition of the liftings is that for any constant function $\kappa \in \widetilde{P}_h$

$$\mathcal{w}(\kappa) = 0 \quad \text{and} \quad \mathbf{u}(\kappa) = \mathbf{0}. \tag{5.4}$$

Any $q \in \widetilde{P}_h$ can be decomposed as $q = \mathring{q} + \bar{q}$, where $\mathring{q} \in P_h$ and $\bar{q}$ is a constant function ($\bar{q}$ equals the global mean of q). Decomposing both p_h and q this way in (5.2)–(5.3), we find that

$$\begin{aligned} a(\boldsymbol{\lambda}_h, \boldsymbol{\mu}) + b(\boldsymbol{\mu}, \mathring{p}_h) &= 0 \qquad && \text{for all } \boldsymbol{\mu} \in M_h \text{ and} \\ b(\boldsymbol{\lambda}_h, \mathring{q}) - c(\mathring{p}_h, \mathring{q}) &= 0 \qquad && \text{for all } \mathring{q} \in P_h. \end{aligned}$$

By the unique solvability of (3.14)–(3.15) asserted by Theorem 3.1, we conclude that both $\boldsymbol{\lambda}_h$ and $\mathring{p}_h$ vanish. Thus, $\mathsf{M} \left[\begin{smallmatrix} \mathsf{\Lambda} \\ \mathsf{P} \end{smallmatrix}\right] = 0$ if and only if $\boldsymbol{\lambda}_h = \mathbf{0}$ and p_h equals a constant function. The null space of M is therefore equal to the span of $\left[\begin{smallmatrix} 0 \\ 1_{\mathsf{P}} \end{smallmatrix}\right]$, where 1_{P} denotes the vector of coefficients of $\kappa \equiv 1 \in \widetilde{P}_h$. Note that if b denotes the vector $\left[\begin{smallmatrix} \mathsf{L}_1 \\ \mathsf{L}_2 \end{smallmatrix}\right]$ on the right-hand side of (5.1), then by (5.4),

$$\mathsf{b} \cdot \left[\begin{smallmatrix} 0 \\ 1_{\mathsf{P}} \end{smallmatrix}\right] = \ell_1(0) + \ell_2(\kappa) = 0.$$

Thus (5.1) has a solution, and if $\left[\begin{smallmatrix} \mathsf{\Lambda} \\ \mathsf{P} \end{smallmatrix}\right]$ is a solution, then all solutions are of the form $\left[\begin{smallmatrix} \mathsf{\Lambda} \\ \mathsf{P} \end{smallmatrix}\right] + \alpha \left[\begin{smallmatrix} 0 \\ 1_{\mathsf{P}} \end{smallmatrix}\right]$ for some $\alpha \in \mathbb{R}$.

To compute one solution to (5.1), one can now apply variations of standard techniques. For example, if one uses a Krylov space iteration such as MINRES for solving (5.1), then the nth iterate x_n is in $\mathsf{x}_0 + \operatorname{span}\{\mathsf{r}_0, \mathsf{M}\mathsf{r}_0, \mathsf{M}^2\mathsf{r}_0, \ldots, \mathsf{M}^{n-1}\mathsf{r}_0\}$, where $\mathsf{r}_0 = \mathsf{b} - \mathsf{M}\mathsf{x}_0$ and x_0 is the initial iterate. Since $\left[\begin{smallmatrix} 0 \\ 1_{\mathsf{P}} \end{smallmatrix}\right] \cdot (\mathsf{M}^j \mathsf{r}_0) = 0$ for all $j \geq 0$, if the initial iterate x_0 satisfies $\mathsf{x}_0 \cdot \left[\begin{smallmatrix} 0 \\ 1_{\mathsf{P}} \end{smallmatrix}\right] = 0$, then all further iterates x_n satisfy $\mathsf{x}_n \cdot \left[\begin{smallmatrix} 0 \\ 1_{\mathsf{P}} \end{smallmatrix}\right] = 0$. Hence by adjusting the final pressure iterate by a scalar multiple of 1_{P}, we can obtain the pressure Lagrange multiplier of zero mean. If one uses a direct solver instead, one can convert (5.1) to an invertible system by simply deleting the row and column of M corresponding to one fixed pressure degree of freedom.

5.2. The Schur complement matrix for the tangential velocity.

Many standard stable choices of mixed finite elements for Stokes equations result in a velocity-pressure discretization of the form (5.1). There is often a preference for solving the discrete system using a positive definite Schur complement system obtained by eliminating the velocity variable (the Schur complement matrix being $\mathsf{C} + \mathsf{B}\mathsf{A}^{-1}\mathsf{B}^t$), because iterative solvers for positive definite systems are well developed. However, this is not feasible for our method, because in contrast to the standard methods, our matrix A in (5.1) is not invertible in general.

But we can obtain an alternate Schur complement system for our discretization by utilizing a feature of our method that is usually not found in standard methods for Stokes equations, namely, the invertibility of the other diagonal block (C) on a subspace. More precisely, we have the following result.

Proposition 5.1. *For any $q_h \in \widetilde{P}_h$, $c(q_h, q_h) = 0$ if and only if q_h is constant.*

Proof. It is obvious from (5.4) that if q_h is constant, then $c(q_h, q_h) = 0$. To prove the converse, we observe that $c(q_h, q_h) = 0$ implies $\mathcal{W}(q_h) = 0$, so from (3.4) it follows that

$$\sum_{e \in \mathcal{E}} (q_h,\, \llbracket \boldsymbol{v} \cdot \boldsymbol{n} \rrbracket)_e = \sum_{e \in \mathcal{E}} (q_h - \bar{q}_h,\, \llbracket \boldsymbol{v} \cdot \boldsymbol{n} \rrbracket)_e = 0 \quad \text{for all } \boldsymbol{v} \in V_h, \tag{5.5}$$

where

$$\bar{q}_h = \frac{\sum_{e \in \mathcal{E}} \int_e q_h \, \mathrm{d}s}{\sum_{e \in \mathcal{E}} \int_e \mathrm{d}s}$$

is the global mean of q_h. By Proposition 4.1, there is a $\boldsymbol{v} \in V_h$ such that $q_h - \bar{q}_h = \llbracket \boldsymbol{v} \cdot \boldsymbol{n} \rrbracket$. Hence (5.5) implies that $q_h - \bar{q}_h \equiv 0$. □

The above proposition readily implies that the matrix C restricted to the orthogonal complement $1_{\mathsf{P}}^{\perp} := \{\mathsf{Q} : \mathsf{Q} \cdot 1_{\mathsf{P}} = 0\}$ is invertible. Therefore, rewriting (5.1) as

$$\begin{bmatrix} \mathsf{C} & -\mathsf{B} \\ -\mathsf{B}^t & -\mathsf{A} \end{bmatrix} \begin{bmatrix} \mathsf{P} \\ \Lambda \end{bmatrix} = - \begin{bmatrix} \mathsf{L}_2 \\ \mathsf{L}_1 \end{bmatrix} \tag{5.6}$$

and eliminating P, we get an alternate Schur complement system:

$$(\mathsf{B}^t\mathsf{C}^{-1}\mathsf{B} + \mathsf{A})\Lambda = \mathsf{L}_1 + \mathsf{B}^t\mathsf{C}^{-1}\mathsf{L}_2. \tag{5.7}$$

Note that the two applications of C^{-1} above make sense because Range$(\mathsf{B}) \subseteq 1_{\mathsf{P}}^{\perp}$ (since $1_{\mathsf{P}} \cdot \mathsf{B}\Lambda = b(\boldsymbol{\lambda}_h, \kappa) = 0$) and $\mathsf{L}_2 \in 1_{\mathsf{P}}^{\perp}$ (since $1_{\mathsf{P}} \cdot \mathsf{L}_2 = \ell_2(\kappa) = 0$). The Schur complement matrix in (5.7) is invertible because (5.6) uniquely determines Λ. Thus (5.7) is a symmetric and positive definite system, well suited to solution by minimization algorithms such as conjugate gradients.

5.3. The local mappings for lowest-order case. We now give explicit expressions for the local maps which define the Lagrange multiplier bilinear forms in the lowest-order case, i.e., $k = 0$. A simple computation gives that, on any triangle T, we have

$$\begin{aligned} w(\boldsymbol{\lambda}) &= \frac{1}{|T|}\int_{\partial T\setminus\partial\Omega} \boldsymbol{\lambda}\cdot\boldsymbol{t}\, \mathrm{d}s, & \boldsymbol{u}(\boldsymbol{\lambda}) &= \frac{1}{|T|}\int_{\partial T\setminus\partial\Omega} \boldsymbol{\lambda}\cdot\boldsymbol{t}\,(\boldsymbol{x}-\boldsymbol{x}_T)^{\perp}\,\mathrm{d}s, \\ w(g_t) &= \frac{1}{|T|}\int_{\partial T\cap\partial\Omega} g_t\,\mathrm{d}s, & \boldsymbol{u}(g_t) &= \frac{1}{|T|}\int_{\partial T\cap\partial\Omega} g_t\,(\boldsymbol{x}-\boldsymbol{x}_T)^{\perp}\,\mathrm{d}s, \\ \mathcal{W}(p) &= \boldsymbol{\mathcal{W}}_p^T \times (\boldsymbol{x}-\boldsymbol{x}_T), & \mathbf{u}(p) &= -\frac{1}{|T|}\int_T (\boldsymbol{x}-\boldsymbol{x}_T)^{\perp}\,\mathcal{W}(p)\,\mathrm{d}x, \\ \mathsf{w}(\boldsymbol{f}) &= \mathbf{w}_{\boldsymbol{f}}^T \times (\boldsymbol{x}-\boldsymbol{x}_T), & \mathbf{u}(\boldsymbol{f}) &= -\frac{1}{|T|}\int_T (\boldsymbol{x}-\boldsymbol{x}_T)^{\perp}\,\mathsf{w}(\boldsymbol{f})\,\mathrm{d}x, \end{aligned}$$

where the point $\boldsymbol{x}_T$ denotes the barycenter of the triangle T,

$$\boldsymbol{\mathcal{W}}_p^T = -\frac{1}{|T|}\int_{\partial T} p\,\boldsymbol{n}\,\mathrm{d}s, \qquad \mathbf{w}_{\boldsymbol{f}}^T = \frac{1}{|T|}\int_T \boldsymbol{f}\,\mathrm{d}x.$$

We have used standard notation for vector operations above, e.g., for vectors $\boldsymbol{a} = (a_1, a_2)$ and $\boldsymbol{b} = (b_1, b_2)$,

$$\boldsymbol{a}\times\boldsymbol{b} = a_1\,b_2 - a_2\,b_1, \qquad \boldsymbol{a}^{\perp} = (-a_2, a_1).$$

It is easy to simplify the above expressions to obtain the local mappings of our Lagrange multiplier basis functions. We first give the liftings of $\boldsymbol{\psi}_{\mathit{\Lambda}}$ for a basis function $\boldsymbol{\psi}_{\mathit{\Lambda}}$ associated with a $\mathit{\Lambda} \in \mathit{\Lambda}_h$. Let K be the triangle formed by vertices $\boldsymbol{x}_1, \boldsymbol{x}_2$, and $\boldsymbol{x}_3$. Let e_i denote the edge of K opposite to vertex $\boldsymbol{x}_i$, and let $\boldsymbol{n}_i$ denote the outward unit normal of K on edge e_i. These notations, when superscripted by L, R, $-$, or $+$, denote the corresponding geometrical parameters of adjacent triangles K^L, K^R, K_e^-, or $K \equiv K_e^+$, respectively, as illustrated in Figure 3.

Consider the basis function associated to $\mathit{\Lambda} \in \mathit{\Lambda}_h$ with vertex $\boldsymbol{x}_3$ and $\mathit{\Lambda} \subseteq \partial K$ as marked in Figure 3. The liftings $w_{\mathit{\Lambda}} := w(\boldsymbol{\psi}_{\mathit{\Lambda}})$ and $\boldsymbol{u}_{\mathit{\Lambda}} := \boldsymbol{u}(\boldsymbol{\psi}_{\mathit{\Lambda}})$ are supported on

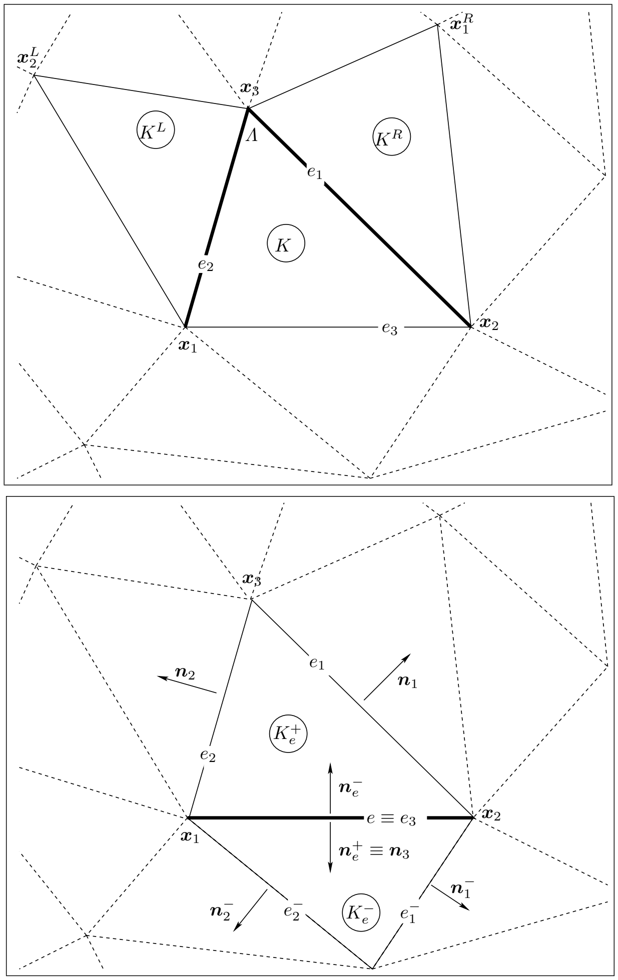


FIG. 3. *Illustration of triangles where liftings associated to a wedge Λ and an edge e are nonzero.*

$K \cup K^R \cup K^L$ and are given by

$$
\begin{aligned}
w_\Lambda &= \frac{|e_1| + |e_2|}{2|K|}, &\quad \boldsymbol{u}_\Lambda &= \frac{|e_1|}{6|K|}(\boldsymbol{x}_3 - \boldsymbol{x}_1)^\perp + \frac{|e_2|}{6|K|}(\boldsymbol{x}_3 - \boldsymbol{x}_2)^\perp &\quad &\text{on } K,\\
w_\Lambda &= -\frac{|e_2|}{2|K^L|}, &\quad \boldsymbol{u}_\Lambda &= -\frac{|e_2|}{6|K^L|}(\boldsymbol{x}_3 - \boldsymbol{x}_2^L)^\perp &\quad &\text{on } K^L,\\
w_\Lambda &= -\frac{|e_1|}{2|K^R|}, &\quad \boldsymbol{u}_\Lambda &= -\frac{|e_1|}{6|K^R|}(\boldsymbol{x}_3 - \boldsymbol{x}_1^R)^\perp &\quad &\text{on } K^R.
\end{aligned}
$$

Here, $|e|$ denotes the length of the edge e and $|K|$ denotes the area of the triangle K. The points $\boldsymbol{x}_1^R$ and $\boldsymbol{x}_2^L$ are shown in Figure 3.

Next, let us display the liftings associated with the pressure. To treat this case, consider an edge e shared by $K \equiv K_e^+$ and another triangle K_e^-. Let p_e denote the indicator function of edge e. The liftings $\mathrm{w}_e := \mathrm{w}(p_e)$ and $\mathbf{u}_e := \mathbf{u}(p_e)$ are supported on $K_e^+ \cup K_e^-$. Using the notation of Figure 3 wherein $e \equiv e_3$, we can express the

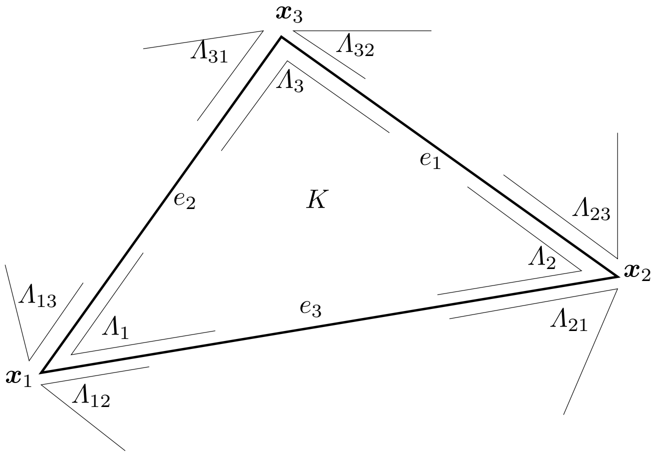


FIG. 4. *Geometry in local element matrix calculations.*

liftings on $K_e^{\pm}$ by

$$\mathsf{w}_e(\boldsymbol{x}) = \boldsymbol{\mathsf{w}}_e^{\pm} \times (\boldsymbol{x} - \boldsymbol{x}_{K_e^{\pm}}), \qquad \mathbf{u}_e(\boldsymbol{x}) = \frac{1}{36}\sum_{\ell=1}^{3} (\boldsymbol{\mathsf{w}}_e^{\pm} \cdot \boldsymbol{E}_\ell^{\pm})\boldsymbol{E}_\ell^{\pm},$$

where, in accordance with our previous notation, $\boldsymbol{x}_{K_e^{\pm}}$ denotes the barycenter of $K_e^{\pm}$,

$$\boldsymbol{\mathsf{w}}_e^{\pm} = -\frac{1}{|K_e^{\pm}|}\boldsymbol{n}_e^{\pm}|e|,$$

and $\boldsymbol{E}_\ell^{\pm} = \boldsymbol{n}_\ell^{\pm}|e_\ell^{\pm}|$. Here $\boldsymbol{n}_e^{\pm}$ is as illustrated in Figures 1 and 3.

Finally, we give formulae for the local mappings associated with the body force on the triangle K. If $\boldsymbol{f}$ is supported only on K, then $\mathsf{w}(\boldsymbol{f})$ and $\mathbf{u}(\boldsymbol{f})$ are supported only on K. Their values on K are given by

$$\mathsf{w}(\boldsymbol{f}) = \boldsymbol{\mathsf{w}} \times (\boldsymbol{x} - \boldsymbol{x}_K), \qquad \mathbf{u}(\boldsymbol{f}) = \frac{1}{36}\sum_{\ell=1}^{3} (\boldsymbol{\mathsf{w}} \cdot \boldsymbol{E}_\ell)\boldsymbol{E}_\ell,$$

where, as before, $\boldsymbol{E}_\ell = |e_\ell|\boldsymbol{n}_\ell$ and

$$\boldsymbol{\mathsf{w}} = \frac{1}{|K|}\int_K \boldsymbol{f}\,\mathrm{d}x.$$

5.4. The local element matrices for the lowest-order case. It is possible to “assemble” the global stiffness matrix of the Lagrange multiplier equations (3.14)–(3.15) just as one does for traditional finite element methods, provided that appropriate local element stiffness matrices are defined for our method. We illustrate this in the lowest-order case.

First, we enumerate the degrees of freedom local to an element as in Figure 4. In this enumeration, we include the omitted elements of $\hat{\Lambda}_h$. The omissions can be taken care of during assembly simply by not assembling the rows and columns corresponding to the omitted elements of $\hat{\Lambda}_h$ (just as one would handle zero Dirichlet boundary conditions when solving the Dirichlet problem with standard finite elements). Figure 4 shows nine elements of $\hat{\Lambda}_h$ connected to K, which we have enumerated as Λ_1, Λ_2, Λ_3, Λ_{12}, Λ_{21}, Λ_{13}, Λ_{31}, Λ_{23}, and Λ_{32}, or in short, Λ_I for all I in the index set $\mathcal{I} := \{1, 2, 3, 12, 21, 13, 31, 23, 32\}$. The local matrices are made using nine functions in M_h whose local mappings are nonzero on K, namely $\boldsymbol{\psi}_{\Lambda_I}$ for all $I \in \mathcal{I}$. The local stiffness matrix of an element K has the form

$$\begin{bmatrix} \mathsf{A}^{(K)} & (\mathsf{B}^{(K)})^t \\ \mathsf{B}^{(K)} & -\mathsf{C}^{(K)} \end{bmatrix},$$

where

$$\begin{aligned}
\mathsf{A}^{(K)}_{IJ} &= \int_K w(\boldsymbol{\psi}_{\Lambda_I})\, w(\boldsymbol{\psi}_{\Lambda_J})\, \mathrm{d}x, && I, J \in \mathcal{I},\\
\mathsf{B}^{(K)}_{LJ} &= -\int_K \mathbf{curl}\, \mathrm{w}(p_L) \cdot \boldsymbol{u}(\boldsymbol{\psi}_{\Lambda_J})\, \mathrm{d}x, && J \in \mathcal{I}, \quad L \in \{1, 2, 3\},\\
\mathsf{C}^{(K)}_{LM} &= \int_K \mathrm{w}(p_L)\, \mathrm{w}(p_M)\, \mathrm{d}x, && L, M \in \{1, 2, 3\}.
\end{aligned}$$

Here p_L denotes the characteristic function of the edge e_L in Figure 4. We can calculate the integrals above after substituting the previously given expressions for the liftings of the basis functions in the integrands.

In order to give explicit expressions for $\mathsf{A}^{(K)}$, $\mathsf{B}^{(K)}$, and $\mathsf{C}^{(K)}$, suppose that $\{i, j, k\}$ is any permutation of $\{1, 2, 3\}$. Let σ_j equal zero if the edge e_j is contained in the boundary $\partial\Omega$, and let σ_j equal one otherwise. Define

$$\begin{aligned}
W_I &= \begin{cases} \sigma_i |e_i| + \sigma_j |e_j| & \text{if } I = k,\\ -|e_k| & \text{if } I = ij, \end{cases}\\
\boldsymbol{U}_I &= \begin{cases} \sigma_i |e_i| (\boldsymbol{x}_k - \boldsymbol{x}_i)^\perp + \sigma_j |e_j| (\boldsymbol{x}_k - \boldsymbol{x}_j)^\perp & \text{if } I = k,\\ -|e_k| (\boldsymbol{x}_i - \boldsymbol{x}_k)^\perp & \text{if } I = ij. \end{cases}
\end{aligned}$$

Then,

$$\begin{aligned}
\mathsf{A}^{(K)}_{IJ} &= \frac{1}{4|K|} W_I W_J,\\
\mathsf{B}^{(K)}_{LJ} &= \frac{1}{6|K|} \boldsymbol{E}_L \cdot \boldsymbol{U}_J,\\
\mathsf{C}^{(K)}_{LM} &= \frac{1}{36|K|} \sum_{\ell=1}^{3} (\boldsymbol{E}_L \cdot \boldsymbol{E}_\ell)(\boldsymbol{E}_M \cdot \boldsymbol{E}_\ell),
\end{aligned}$$

where, as before, $\boldsymbol{E}_L = \boldsymbol{n}_L |e_L|$ for all $L \in \{1, 2, 3\}$. With these local matrices, one can assemble all the global matrices of our method as simply as those of any other finite element method.

6. Conclusion.

We have developed new hybridization techniques which, when applied to a well-known conforming mixed method for the Stokes problem, result in a new "tangential velocity-pressure" discretization. The advantages of the new method include fewer globally coupled degrees of freedom and numerical velocity approximations that satisfy the incompressibility condition exactly. Our results are achieved by using the methodology introduced in [8] to study hybridized mixed methods for second-order elliptic problems.

In a forthcoming sequel [9], we will discuss the extension of the ideas here to the Stokes problem in three space dimensions, variable degree incompressible finite elements, and other boundary conditions.

REFERENCES

[1] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[2] G. A. Baker, W. N. Jureidini, and O. A. Karakashian, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1466–1485.

[3] P. Bastian and B. Rivière, *Superconvergence and $H(\text{div})$ projection for discontinuous Galerkin methods*, Internat. J. Numer. Methods Fluids, 42 (2003), pp. 1043–1057.

[4] J. H. Bramble and J. E. Pasciak, *A domain decomposition technique for Stokes problems*, Appl. Numer. Math., 6 (1990), pp. 251–261.

[5] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[6] C. Brătianu and S. N. Atluri, *A hybrid finite element method for Stokes flow.* I. *Formulation and numerical studies*, Comput. Methods Appl. Mech. Engrg., 36 (1983), pp. 23–37.

[7] J. Carrero, B. Cockburn, and D. Schötzau, *Hybridized, globally divergence-free LDG methods. Part* I: *The Stokes problem*, Math. Comp., to appear.

[8] B. Cockburn and J. Gopalakrishnan, *A characterization of hybridized mixed methods for second order elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 283–301.

[9] B. Cockburn and J. Gopalakrishnan, *Incompressible finite elements via hybridization. Part* II. *The Stokes system in three space dimensions*, SIAM J. Numer. Anal., 43 (2005), pp. 1651–1672.

[10] B. Cockburn, G. Kanschat, and D. Schötzau, *A locally conservative LDG method for the incompressible Navier-Stokes equations*, Math. Comp., 74 (2005), pp. 1067–1095.

[11] B. M. Fraejis de Veubeke, *Displacement and equilibrium models in the finite element method*, in Stress Analysis, O. Zienkiewicz and G. Holister, eds., Wiley, New York, 1977, pp. 145–197.

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

[13] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer Ser. Comput. Math. 5, Springer-Verlag, New York, 1986.

[14] M. D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows. A Guide to Theory, Practice, and Algorithms*, Computer Science and Scientific Computing, Academic Press, Boston, MA, 1989.

[15] D. Griffiths, *Finite elements for incompressible flow*, Math. Methods Appl. Sci., 1 (1979), pp. 16–31.

[16] F. Hecht, *Construction d'une base P_1 non conforme à divergence nulle*, RAIRO Modél. Math. Anal. Numér., 15 (1981), pp. 119–150.

[17] O. A. Karakashian and W. N. Jureidini, *A nonconforming finite element method for the stationary Navier–Stokes equations*, SIAM J. Numer. Anal., 35 (1998), pp. 93–120.

[18] O. A. Karakashian and T. Katsaounis, *A discontinuous Galerkin method for the incompressible Navier-Stokes equations*, in Discontinuous Galerkin Methods. Theory, Computation and Applications, Lect. Notes Comput. Sci. Eng. 11, B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., Springer-Verlag, Berlin, 2000, pp. 157–166.

[19] J. Morgan and R. Scott, *A nodal basis for C^1 piecewise polynomials of degree $n \geq 5$*, Math. Comput., 29 (1975), pp. 736–740.

[20] J.-C. Nédélec, *Éléments finis mixtes incompressibles pour l'équation de Stokes dans $\mathbf{R}^3$*, Numer. Math., 39 (1982), pp. 97–112.

[21] J.-C. Nédélec, *A new family of mixed finite elements in $\mathbf{R}^3$*, Numer. Math., 50 (1986), pp. 57–81.

[22] R. Scholz, *A mixed method for* 4*th order problems using linear finite elements*, RAIRO Anal. Numér., 12 (1978), pp. 85–90, iii.

[23] L. R. Scott and M. Vogelius, *Conforming finite element methods for incompressible and nearly incompressible continua*, in Large-Scale Computations in Fluid Mechanics, Part 2 (La Jolla, CA, 1983), Lectures in Appl. Math. 22-2, AMS, Providence, RI, 1985, pp. 221–244.

[24] F. Thomasset, *Implementation of Finite Element Methods for Navier-Stokes Equations*, Springer Ser. Comput. Phys., Springer-Verlag, New York, 1981.

[25] L. A. Ying and S. N. Atluri, *A hybrid finite element method for Stokes flow.* II. *Stability and convergence studies*, Comput. Methods Appl. Mech. Engrg., 36 (1983), pp. 39–60.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1651–1672

INCOMPRESSIBLE FINITE ELEMENTS VIA HYBRIDIZATION. PART II: THE STOKES SYSTEM IN THREE SPACE DIMENSIONS[*]

BERNARDO COCKBURN[†] AND JAYADEEP GOPALAKRISHNAN[‡]

Abstract. We introduce a method that gives exactly incompressible velocity approximations to Stokes flow in three space dimensions. The method is designed by extending the ideas in Part I [B. Cockburn and J. Gopalakrishnan, *SIAM J. Numer. Anal.*, 43 (2005), pp. 1627–1650] of this series, where the Stokes system in two space dimensions was considered. Thus we hybridize a vorticity-velocity formulation to obtain a new mixed method coupling approximations of tangential velocity and pressure on mesh faces. Once this relatively small tangential velocity-pressure system is solved, it is possible to recover a globally divergence-free numerical approximation of the fluid velocity, an approximation of the vorticity whose tangential component is continuous across interelement boundaries, and a discontinuous numerical approximation of the pressure. The main difference between our method here and that of the two-dimensional case treated in Part I is in the use of Nédélec elements, which necessitates development of new hybridization techniques. We also generalize the method to allow for varying polynomial degrees on different mesh elements and to incorporate certain nonstandard but physically relevant boundary conditions.

Key words. divergence-free finite element, mixed method, hybridized method, Nédélec element, fluid flow, Stokes flow, velocity, vorticity, pressure, Lagrange multipliers

AMS subject classifications. 65N30, 76D07

DOI. 10.1137/040610659

1. Introduction.

This is a sequel to our paper [7] in which we introduced a new hybridized method for the Stokes equations in two space dimensions. Here we generalize the ideas presented in [7] to the Stokes system in three space dimensions. We also extend the method to allow variable degrees of approximation on different mesh elements. As in [7], the three-dimensional version of our method simultaneously yields an exactly divergence-free numerical approximation of the fluid velocity and a continuous numerical approximation of the vorticity. A discontinuous numerical approximation of the pressure can also be recovered separately. These three approximations are obtained in an element-by-element fashion after one global system for certain Lagrange multipliers arising from the hybridization is solved. This global system represents a new "tangential velocity-pressure" discretization of the Stokes system on the mesh faces because the Lagrange multipliers are approximations to the pressure and tangential fluid velocity on element interfaces.

We are hybridizing a mixed formulation that has previously appeared in the literature [9, 14] (cf. [1, 3]). However, the previous works resort to introduction of a stream function variable to obtain exactly divergence-free numerical velocities. This approach is beset with significant difficulties in three dimensions: (i) While the stream function is a scalar function in two dimensions, in three dimensions it is a vector function, so its introduction into the method, as in [14], leads to a significant increase in

[*] Received by the editors June 27, 2004; accepted for publication (in revised form) January 28, 2005; published electronically November 14, 2005.
http://www.siam.org/journals/sinum/43-4/61065.html

[†] School of Mathematics, University of Minnesota, Vincent Hall, Minneapolis, MN 55455 (cockburn @math.umn.edu). The research of this author was supported in part by National Science Foundation grant DMS-0411254 and by the University of Minnesota Supercomputing Institute.

[‡] Department of Mathematics, University of Florida, Gainesville, FL 32611-8105 (jayg@math. ufl.edu). The research of this author was supported in part by National Science Foundation grant DMS-0410030.

number of degrees of freedom. (ii) The stream function is not uniquely defined. While in two dimensions it is defined up to a constant, in three dimensions one has to impose a nontrivial "gauge condition." (iii) The definition of the stream function must take into account the topology of the three-dimensional domain. For domains that are not simply connected, one must find "cuts" and base the definition of finite element spaces for the stream function on them (see [1]). Finding such cuts in automatic computation is not easy. (iv) Formulations involving the stream function alone lead to fourth-order problems (see, e.g., [1, 9]) and hence to badly conditioned matrices. Notwithstanding these difficulties, the use of the stream function has hitherto been the only successful approach in obtaining exactly incompressible approximations of all orders in three dimensions. The search for exactly incompressible numerical approximations to Stokes flow has a rich history. References to some previous attempts can be found in [4, 7, 11].

All the above-mentioned difficulties disappear in our approach via hybridization. Because we do not introduce the stream function, our method requires nothing special to be done when the computational domain has nontrivial topology. For the same reason we never encounter a fourth-order operator—our matrices represent discretizations of operators of second order only. Moreover, while the introduction of the stream function results in an increase in degrees of freedom in some of the previous works, our approach using hybridization actually results in a decrease in degrees of freedom, as we shall see in section 3.

As we move from two to three space dimensions, the main difference we encounter is in the treatment of vorticity. When considering finite element approximations to vorticity, we now have to use the $H(\mathbf{curl}\,,\Omega)$-conforming Nédélec elements [13], while in two dimensions we used the simpler $H^1(\Omega)$-conforming finite elements. However, the velocity approximation is treated in exactly the same way as in the two-dimensional case—it continues to be in an $H(\mathrm{div},\Omega)$-conforming subspace of exactly divergence-free functions. Another important similarity between the two- and three-dimensional cases is in the structure of the method and equations, so we are able to easily adapt the elimination procedure which we developed in [7] to three dimensions. The result is a Lagrange multiplier system that is completely analogous to the two-dimensional case.

The introduction of Nédélec spaces necessitates development of new hybridization techniques in three dimensions. Indeed, the Nédélec space has edge degrees of freedom, and none of the existing hybridization techniques handle them. To elaborate, consider the following sequence of spaces:

$$H^1(\Omega)/\mathbb{R} \xrightarrow{\mathbf{grad}} H(\mathbf{curl}\,,\Omega) \xrightarrow{\mathbf{curl}} H(\mathrm{div},\Omega) \xrightarrow{\mathrm{div}} L^2(\Omega).$$

As we traverse the sequence from right to left, the continuity conditions on the spaces become more complex. Finite element subspaces of $H(\mathrm{div},\Omega)$ consist of functions whose normal component is continuous across element interfaces. Hybridization techniques to relax such continuity are well known, and they are the basis for the hybridized Raviart–Thomas and Brezzi–Douglas–Marini (BDM)-type methods [2, 5]. Such hybridizations relaxed continuity of finite element subspaces across interior mesh faces using traces from (just) two elements sharing an interior mesh face. However, once we move on to finite element subspaces of $H(\mathbf{curl}\,,\Omega)$, the continuity constraints are more complicated, as reflected by the fact that these spaces have edge degrees of freedom which are connected to multiple elements. Moving further left to $H^1(\Omega)$, we find finite element subspaces having vertex degrees of freedom, adding another layer of complexity. Since all previously known hybridization techniques relaxed continuity across mesh faces, we find a widespread belief that methods using edge and vertex degrees of freedom are not amenable to hybridization. In this paper, we dispel this belief by hybridizing a method that uses Nédélec spaces having edge degrees of freedom. It is also possible to hybridize methods that use $H^1(\Omega)$-subspaces, as we demonstrated in [7].

We make two other extensions in this paper. The first extends to the Stokes system what was done for second-order elliptic equations in [6]. Thus, we exploit the ease of construction of variable degree methods via hybridization to give a variable degree version of the original mixed method. Our hybridized variable degree method does not require one to implement transition elements. This is quite convenient considering that transitional Nédélec elements are not trivial to implement. Second, we show how one can incorporate boundary conditions involving the pressure and tangential vorticity into our method. Although such boundary conditions are physically relevant, few methods are known that can incorporate them naturally.

We have kept the organization of this paper very similar to that of Part I [7] to render the analogies and differences with the two-dimensional case transparent. We introduce the variable degree method in section 2. In section 3, we briefly present the elimination strategy to obtain a reduced Lagrange multiplier system. A computable basis for the space of Lagrange multipliers of variable degree is given in section 4, and full details of the lowest-order case are given in section 5. Finally, in section 6, we show how to incorporate other boundary conditions.

2. The variable degree hybridized mixed method.

The three-dimensional Stokes problem is to find a fluid velocity field $\boldsymbol{u}$ and pressure p satisfying

$$-\boldsymbol{\Delta u} + \mathbf{grad}\, p = \boldsymbol{f} \qquad \text{on } \Omega, \tag{2.1}$$

$$\operatorname{div} \boldsymbol{u} = 0 \qquad \text{on } \Omega, \tag{2.2}$$

$$\boldsymbol{u} = \boldsymbol{g} \qquad \text{on } \partial\Omega. \tag{2.3}$$

Here we assume that Ω is a bounded connected domain with polyhedral boundary $\partial\Omega$ such that Ω lies on only one side of $\partial\Omega$ locally, the data $\boldsymbol{f}$ is in $L^2(\Omega)^3$, and $\boldsymbol{g} \in H^{1/2}(\partial\Omega)^3$. We do not assume that Ω is simply connected. We also do not assume that $\partial\Omega$ is connected. We require the data $\boldsymbol{g}$ to satisfy the compatibility condition

$$(g_n, 1)_{\partial\Omega} = 0,$$

where $g_n = \boldsymbol{g} \cdot \boldsymbol{n}$ and $\boldsymbol{n}$ is the outward unit normal on $\partial\Omega$. Under this assumption, it is well known that the Stokes problem has a unique solution.

Let us reformulate the Stokes problem by introducing vorticity $\boldsymbol{\omega} = \mathbf{curl}\, \boldsymbol{u}$. Using the identity

$$-\boldsymbol{\Delta u} = \mathbf{curl}\, \mathbf{curl}\, \boldsymbol{u} - \mathbf{grad} \operatorname{div} \boldsymbol{u},$$

the Stokes system (2.1)–(2.3) can be rewritten as

$$\boldsymbol{\omega} - \mathbf{curl}\, \boldsymbol{u} = 0 \qquad \text{on } \Omega, \tag{2.4}$$

$$\mathbf{curl}\, \boldsymbol{\omega} + \mathbf{grad}\, p = \boldsymbol{f} \qquad \text{on } \Omega, \tag{2.5}$$

$$\operatorname{div} \boldsymbol{u} = 0 \qquad \text{on } \Omega, \tag{2.6}$$

$$\boldsymbol{u}_{\mathsf{T}} = \boldsymbol{g}_{\mathsf{T}} \qquad \text{on } \partial\Omega, \tag{2.7}$$

$$\boldsymbol{u} \cdot \boldsymbol{n} = \boldsymbol{g} \cdot \boldsymbol{n} \qquad \text{on } \partial\Omega, \tag{2.8}$$

where we have split (2.3) into two equations, one in the direction of the outward unit normal $\boldsymbol{n}$ on $\partial\Omega$, and the other in the tangent plane; i.e., $\boldsymbol{g}_{\mathsf{T}} := \boldsymbol{g} - (\boldsymbol{g}\cdot\boldsymbol{n})\boldsymbol{n}$ denotes the tangential component of $\boldsymbol{g}$.

There is a well-known weak problem based on this reformulation. Define $\mathcal{W} = H(\mathbf{curl}\,,\Omega)$ and

$$\mathcal{V}(b) = \{\boldsymbol{v} \in H(\mathrm{div},\Omega) : \mathrm{div}\,\boldsymbol{v} = 0 \text{ and } \boldsymbol{v}\cdot\boldsymbol{n}|_{\partial\Omega} = b\}$$

for any $b \in H^{-1/2}(\partial\Omega)$. Then $(\boldsymbol{\omega},\boldsymbol{u})$ is the only element of $\mathcal{W}\times\mathcal{V}(g_n)$ satisfying

$$(2.9) \qquad (\boldsymbol{\omega},\boldsymbol{\tau})_\Omega - (\boldsymbol{u},\mathbf{curl}\,\boldsymbol{\tau})_\Omega = (\boldsymbol{g}_{\mathsf{T}},\boldsymbol{\tau})_{\partial\Omega} \qquad \text{for all } \boldsymbol{\tau}\in\mathcal{W},$$

$$(2.10) \qquad (\boldsymbol{v},\mathbf{curl}\,\boldsymbol{\omega})_\Omega = (\boldsymbol{f},\boldsymbol{v})_\Omega \qquad \text{for all } \boldsymbol{v}\in\mathcal{V}(0).$$

Here $(\cdot,\cdot)_\Omega$ denotes the $L^2(\Omega)$ (or $L^2(\Omega)^3$) innerproduct. Note that the pressure has disappeared in this mixed formulation.

One way to develop a hybridized mixed method that discretizes (2.9)–(2.10) is to first approximate the weak formulation by a conforming mixed method and then relax the continuity constraints of the discrete spaces. Here, we motivate the construction of our variable degree hybridized mixed method (2.9)–(2.10) by another equivalent approach using the differential problem (2.4)–(2.8). Suppose the domain Ω is meshed by a tetrahedral mesh $\mathcal{T}$ (satisfying the usual finite element assumptions). To each tetrahedron K we associate a degree $k(K)$ and the following pair of spaces:

$$\begin{aligned} W(K) &= P_{k(K)}(K)^3 \oplus S_{k(K)+1}(K),\\ V(K) &= \{\boldsymbol{v}\in P_{k(K)}(K)^3 :\ \mathrm{div}\,\boldsymbol{v} = 0\}, \end{aligned}$$

where $P_\ell(K)^3$ denotes the set of vector functions whose (three) components are polynomials of degree at most ℓ and $S_\ell(K)$ is the set of all vector functions $\boldsymbol{p}_\ell(\boldsymbol{x})$ whose components are homogeneous polynomials of degree ℓ satisfying $\boldsymbol{p}_\ell(\boldsymbol{x})\cdot\boldsymbol{x} = 0$. Define the variable degree Nédélec space with no continuity conditions by

$$W_h = \{\boldsymbol{w} : \boldsymbol{w}|_K \in W(K) \text{ for all } K\in\mathcal{T}\}.$$

While the vorticity is approximated in W_h, the velocity is approximated in

$$V_h = \{\boldsymbol{v} : \boldsymbol{v}|_K \in V(K) \text{ for all } K\in\mathcal{T}\}.$$

The numerical method is motivated by requiring that (2.4) and (2.5) be satisfied weakly on each element K: Multiplying (2.4) and (2.5) by test functions $\boldsymbol{\tau}\in W(K)$ and $\boldsymbol{v}\in V(K)$ and integrating by parts,

$$\begin{aligned} (\boldsymbol{\omega},\boldsymbol{\tau})_K - (\boldsymbol{u},\mathbf{curl}\,\boldsymbol{\tau})_K - (\boldsymbol{u}_{\mathsf{T}},\boldsymbol{n}\times\boldsymbol{\tau})_{\partial K} &= 0,\\ (\boldsymbol{v},\mathbf{curl}\,\boldsymbol{\omega})_K + (\boldsymbol{v}\cdot\boldsymbol{n},p)_{\partial K} &= (\boldsymbol{f},\boldsymbol{v})_K, \end{aligned}$$

where $\boldsymbol{u}_{\mathsf{T}}$ denotes the tangential component of $\boldsymbol{u}$ on ∂K. Therefore we require that the discrete approximations to vorticity and velocity, namely, $\boldsymbol{\omega}_h$ and $\boldsymbol{u}_h$, respectively, satisfy

$$\begin{aligned} (\boldsymbol{\omega}_h,\boldsymbol{\tau})_K - (\boldsymbol{u}_h,\mathbf{curl}\,\boldsymbol{\tau})_K - (\boldsymbol{\lambda}_h,\boldsymbol{n}\times\boldsymbol{\tau})_{\partial K} &= 0,\\ (\boldsymbol{v},\mathbf{curl}\,\boldsymbol{\omega}_h)_K + (\boldsymbol{v}\cdot\boldsymbol{n},p_h)_{\partial K} &= (\boldsymbol{f},\boldsymbol{v})_K, \end{aligned}$$

where we have introduced two additional approximations $\boldsymbol{\lambda}_h \approx \boldsymbol{u}_{\mathsf{T}}$ and $p_h \approx p$, which we shall call Lagrange multiplier approximations of the tangential velocity and pressure, respectively.

The description of the method is completed by adding appropriate continuity conditions for $\boldsymbol{\omega}_h$ and $\boldsymbol{u}_h$ at the element interfaces. Since $\boldsymbol{\omega}_h$ and $\boldsymbol{u}_h$ are to approximate $\boldsymbol{\omega}$ and $\boldsymbol{u}$ in (2.9)–(2.10), the functional setting of (2.9)–(2.10) clarifies the continuity constraints to be put on $\boldsymbol{\omega}_h$ and $\boldsymbol{u}_h$. To make this precise, let us introduce some more notation: Let $\mathcal{F}$ denote the set of all faces of the triangulation $\mathcal{T}$. On every interior face in $F \in \mathcal{F}$ shared by two tetrahedra K_F^+ and K_F^- we define

$$\begin{aligned} [\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]_F &= \boldsymbol{v}_F^+\cdot\boldsymbol{n}_F^+ + \boldsymbol{v}_F^-\cdot\boldsymbol{n}_F^-,\\ [\![\boldsymbol{n}\times\boldsymbol{v}]\!]_F &= \boldsymbol{n}_F^+\times\boldsymbol{v}_F^+ + \boldsymbol{n}_F^-\times\boldsymbol{v}_F^-, \end{aligned}$$

where $\boldsymbol{n}_F^+$ and $\boldsymbol{n}_F^-$ denote the outward unit normals on the boundaries of K_F^+ and K_F^-, respectively, and $\boldsymbol{v}_F^\pm(\boldsymbol{x}) = \lim_{\epsilon\downarrow 0}\boldsymbol{v}(\boldsymbol{x}-\epsilon\,\boldsymbol{n}_F^\pm)$. On faces $e \subset \partial\Omega$ we set

$$[\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]_F = \boldsymbol{v}|_{\partial\Omega}\cdot\boldsymbol{n} \quad\text{and}\quad [\![\boldsymbol{n}\times\boldsymbol{v}]\!]_F = 0.$$

By $[\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]$ (without subscripts) we mean the function that is defined on the union of all the faces and equals $[\![\boldsymbol{v}\cdot\boldsymbol{n}]\!]_F$ on each face $e \in \mathcal{F}$. The function $[\![\boldsymbol{n}\times\boldsymbol{v}]\!]$ is similarly defined. Then here are our spaces of Lagrange multipliers:

$$P_h = \{p: \quad p = [\![\boldsymbol{v}\cdot\boldsymbol{n}]\!] \text{ for some } \boldsymbol{v}\in V_h\}, \tag{2.11}$$

$$M_h = \{\boldsymbol{\mu}: \quad \boldsymbol{\mu} = [\![\boldsymbol{n}\times\boldsymbol{v}]\!] \text{ for some } \boldsymbol{v}\in W_h\}. \tag{2.12}$$

They are ideal for imposing the natural continuity conditions of the Sobolev spaces $\mathcal{W}$ and $\mathcal{V}$ on the discrete approximations $\boldsymbol{\omega}_h$ and $\boldsymbol{v}_h$; e.g.,

$$\sum_{F\in\mathcal{F}} (\boldsymbol{\mu},\, [\![\boldsymbol{n}\times\boldsymbol{\omega}_h]\!])_F = 0 \quad \text{for all } \boldsymbol{\mu}\in M_h$$

implies that $\boldsymbol{\omega}_h \in H(\mathbf{curl}\,)$.

Thus we have motivated the following definition of our variable degree hybridized mixed method: Find $(\boldsymbol{\omega}_h, \boldsymbol{u}_h, \boldsymbol{\lambda}_h, p_h) \in W_h \times V_h \times M_h \times P_h$ satisfying

$$(\boldsymbol{\omega}_h, \boldsymbol{\tau}_h)_\Omega - (\boldsymbol{u}_h, \mathbf{curl}\,\boldsymbol{\tau}_h)_\Omega - \sum_{F\in\mathcal{F}} (\boldsymbol{\lambda}_h,\, [\![\boldsymbol{n}\times\boldsymbol{\tau}_h]\!])_F = (\boldsymbol{g}_{\mathsf{T}}, \boldsymbol{n}\times\boldsymbol{\tau}_h)_{\partial\Omega}, \tag{2.13}$$

$$(\boldsymbol{v}_h, \mathbf{curl}\,\boldsymbol{\omega}_h)_\Omega + \sum_{F\in\mathcal{F}} (p_h,\, [\![\boldsymbol{v}_h\cdot\boldsymbol{n}]\!])_F = (\boldsymbol{f}, \boldsymbol{v}_h)_\Omega, \tag{2.14}$$

$$\sum_{F\in\mathcal{F}} (q_h,\, [\![\boldsymbol{u}_h\cdot\boldsymbol{n}]\!])_F = (g_n, q_h)_{\partial\Omega}, \tag{2.15}$$

$$\sum_{F\in\mathcal{F}} (\boldsymbol{\mu}_h,\, [\![\boldsymbol{n}\times\boldsymbol{\omega}_h]\!])_F = 0 \tag{2.16}$$

for all $\boldsymbol{\tau}_h \in W_h$, $\boldsymbol{v}_h \in V_h$, $q_h \in P_h$, and $\boldsymbol{\mu}_h \in M_h$.

Proposition 2.1. *There is a unique solution for the system* (2.13)–(2.16).

Proof. Since the system is square, we need only verify that when $\boldsymbol{f}$ and $\boldsymbol{g}$ are zero, all solution components vanish. Zero data implies that $\boldsymbol{\omega}_h$ and $\boldsymbol{u}_h$ lie in the following two spaces, respectively:

$$\mathcal{W}_h = W_h \cap \mathcal{W}, \qquad \mathcal{V}_h(0) = V_h \cap \mathcal{V}(0).$$

Therefore we find from (2.13) and (2.14) that

$$(\boldsymbol{\omega}_h, \boldsymbol{\tau}_h)_\Omega - (\boldsymbol{u}_h, \mathbf{curl}\, \boldsymbol{\tau}_h)_\Omega = 0 \quad \text{for all } \boldsymbol{\tau}_h \in \mathcal{W}_h, \tag{2.17}$$

$$(\boldsymbol{v}_h, \mathbf{curl}\, \boldsymbol{\omega}_h)_\Omega = 0 \quad \text{for all } \boldsymbol{v}_h \in \mathcal{V}_h(0). \tag{2.18}$$

Note that for the mixed method (2.17)–(2.18) to make sense, the spaces therein must be nonempty, as we have tacitly assumed. Setting $\boldsymbol{v}_h = \boldsymbol{u}_h$ in (2.18) and adding these equations, one immediately finds that $(\boldsymbol{\omega}_h, \boldsymbol{\tau}_h)_\Omega = 0$ for all $\boldsymbol{\tau}_h \in \mathcal{W}_h$, and thus $\boldsymbol{\omega}_h = 0$. Now that $\boldsymbol{\omega}_h$ vanishes from (2.17), we have

$$(\boldsymbol{u}_h, \mathbf{curl}\, \boldsymbol{\tau}_h)_\Omega = 0 \quad \text{for all } \boldsymbol{\tau}_h \in \mathcal{W}_h. \tag{2.19}$$

By a well-known property of the Nédélec space, we have that on each element **curl** $W(K) = V(K)$. Moreover, $[\![\boldsymbol{n} \cdot \mathbf{curl}\, \boldsymbol{w}]\!]_F = 0$ whenever $[\![\boldsymbol{n} \times \boldsymbol{w}]\!]_F = 0$ for every interior mesh face F. Hence, it is easy to see that for the variable degree spaces $\mathcal{W}_h$ and $\mathcal{V}_h(0)$ we have (cf. [9, Lemma III.5.1])

$$\mathcal{V}_h(0) \subset \mathbf{curl}\ \mathcal{W}_h.$$

Therefore, in (2.19) we can choose $\boldsymbol{\tau}_h$ such that $\mathbf{curl}\, \boldsymbol{\tau}_h = \boldsymbol{u}_h$, and thus $\boldsymbol{u}_h$ vanishes. Finally, since $\boldsymbol{\omega}_h$ and $\boldsymbol{u}_h$ vanish from (2.13) and (2.14), we find that the Lagrange multipliers $\boldsymbol{\lambda}_h$ and p_h must vanish as well. □

In the uniform degree case, our hybridized mixed method is equivalent to the mixed method considered in [14] in the following sense: Our $\boldsymbol{\omega}_h$ and $\boldsymbol{u}_h$ coincide with vorticity and velocity approximations discussed there. Therefore, the error estimates proven there apply to our solution components $\boldsymbol{\omega}_h$ and $\boldsymbol{u}_h$. It may appear at this point that our method has too many unknowns. But as we shall see in the next section, it is possible to eliminate all but the Lagrange multiplier variables from (2.13)–(2.16), thus making our formulation more attractive.

Before proceeding to the above-mentioned elimination, let us note one advantage that results from hybridization: Since hybridization provides an approximation to the pressure on the mesh faces through the Lagrange multiplier p_h, we can compute an approximation to the pressure inside mesh elements in a completely local (element-by-element) fashion. Borrowing an idea from [4], we define the pressure π_h on the triangle K as the element of $P_{k(K)}(K)$ such that

$$-(\pi_h, \operatorname{div} \boldsymbol{v})_K = (\boldsymbol{f}, \boldsymbol{v})_K - (\mathbf{curl}\, \boldsymbol{\omega}_h, \boldsymbol{v})_K - (\boldsymbol{v} \cdot \boldsymbol{n}, p_h)_{\partial K} \tag{2.20}$$

for all $\boldsymbol{v}$ in $P_{k(K)}(K)^3 + \boldsymbol{x}\, P_{k(K)}(K)$, where $\boldsymbol{n}$ denotes the outward unit normal to K. That (2.20) uniquely defines π_h follows from two facts: (i) div : $P_{k(K)}(K)^3 + \boldsymbol{x}\, P_{k(K)}(K) \mapsto P_{k(K)}(K)$ is a surjection, and (ii) if $\operatorname{div} \boldsymbol{v} = 0$ for a $\boldsymbol{v}$ in $P_{k(K)}(K)^3 + \boldsymbol{x}\, P_{k(K)}(K)$, then $\boldsymbol{v} \in P_{k(K)}(K)^3$ and the right-hand side of (2.20) is zero by the definition of the hybridized method. Thus our method can simultaneously provide approximations to the velocity, vorticity, and pressure.

3. A characterization of the Lagrange multipliers.

3.1. The Lagrange multiplier equation. We now show how one can eliminate the vorticity as well as the velocity variables from our hybridized mixed method (2.13)–(2.16) and arrive at a system of equations involving the Lagrange multipliers alone. Our arguments here are a straightforward generalization of the arguments in [7].

We define *lifting* maps that map functions defined on element interfaces into functions on Ω: Define $(\boldsymbol{w}(\boldsymbol{\lambda}), \boldsymbol{u}(\boldsymbol{\lambda})) \in W_h \times V_h$ and $(\mathsf{w}(p), \mathbf{u}(p)) \in W_h \times V_h$ element-by-element as follows:

$$(\boldsymbol{w}(\boldsymbol{\lambda}), \boldsymbol{\tau})_K - (\boldsymbol{u}(\boldsymbol{\lambda}), \mathbf{curl}\,\boldsymbol{\tau})_K = (\boldsymbol{\lambda}, \boldsymbol{n} \times \boldsymbol{\tau})_{\partial K} \qquad \text{for all } \boldsymbol{\tau} \in W(K), \tag{3.1}$$

$$(\boldsymbol{v}, \mathbf{curl}\ \boldsymbol{w}(\boldsymbol{\lambda}))_K = 0 \qquad \text{for all } \boldsymbol{v} \in V(K), \tag{3.2}$$

$$(\mathsf{w}(p), \boldsymbol{\tau})_K - (\mathbf{u}(p), \mathbf{curl}\,\boldsymbol{\tau})_K = 0 \qquad \text{for all } \boldsymbol{\tau} \in W(K), \tag{3.3}$$

$$(\boldsymbol{v}, \mathbf{curl}\ \mathsf{w}(p))_K = -(p, \boldsymbol{v} \cdot \boldsymbol{n})_{\partial K} \qquad \text{for all } \boldsymbol{v} \in V(K). \tag{3.4}$$

In addition, define $(\mathbf{w}(\boldsymbol{f}), \mathbf{u}(\boldsymbol{f}))$ and $(\boldsymbol{w}(\boldsymbol{g}_{\mathsf{T}}), \boldsymbol{u}(\boldsymbol{g}_{\mathsf{T}}))$ in $W_h \times V_h$ by

$$(\mathbf{w}(\boldsymbol{f}), \boldsymbol{\tau})_K - (\mathbf{u}(\boldsymbol{f}), \mathbf{curl}\,\boldsymbol{\tau})_K = 0 \qquad \text{for all } \boldsymbol{\tau} \in W(K), \tag{3.5}$$

$$(\boldsymbol{v}, \mathbf{curl}\ \mathbf{w}(\boldsymbol{f}))_K = (\boldsymbol{f}, \boldsymbol{v})_K \qquad \text{for all } \boldsymbol{v} \in V(K), \tag{3.6}$$

$$(\boldsymbol{w}(\boldsymbol{g}_{\mathsf{T}}), \boldsymbol{\tau})_K - (\boldsymbol{u}(\boldsymbol{g}_{\mathsf{T}}), \mathbf{curl}\,\boldsymbol{\tau})_K = (\boldsymbol{g}_{\mathsf{T}}, \boldsymbol{n} \times \boldsymbol{\tau})_{\partial K \cap \partial\Omega} \qquad \text{for all } \boldsymbol{\tau} \in W(K), \tag{3.7}$$

$$(\boldsymbol{v}, \mathbf{curl}\ \boldsymbol{w}(\boldsymbol{g}_{\mathsf{T}}))_K = 0 \qquad \text{for all } \boldsymbol{v} \in V(K). \tag{3.8}$$

Note that all of the above local problems are uniquely solvable. Hence, these local maps are well defined.

The main result of this section characterizes the Lagrange multipliers as the unique solution of a variational equation involving the bilinear forms

$$\begin{aligned} a(\boldsymbol{\lambda}, \boldsymbol{\mu}) &= (\boldsymbol{w}(\boldsymbol{\lambda}), \boldsymbol{w}(\boldsymbol{\mu}))_{\Omega}, \\ c(p, q) &= (\mathsf{w}(p), \mathsf{w}(q))_{\Omega}, \\ b(\boldsymbol{\mu}, p) &= -\sum_{K \in \mathcal{T}} (\boldsymbol{u}(\boldsymbol{\mu}), \mathbf{curl}\ \mathsf{w}(p))_K \end{aligned}$$

and the functionals

$$\ell_1(\boldsymbol{\mu}) = (\boldsymbol{f}, \boldsymbol{u}(\boldsymbol{\mu}))_{\Omega} - (\boldsymbol{g}_{\mathsf{T}}, \boldsymbol{w}(\boldsymbol{\mu}))_{\partial\Omega}, \tag{3.9}$$

$$\ell_2(q) = (\boldsymbol{f}, \mathbf{u}(q))_{\Omega} + (g_n, q)_{\partial\Omega} - (\boldsymbol{g}_{\mathsf{T}}, \mathsf{w}(q))_{\partial\Omega}. \tag{3.10}$$

THEOREM 3.1. *The Lagrange multiplier* $(\boldsymbol{\lambda}_h, p_h) \in M_h \times P_h$ *of the hybridized mixed method* (2.13)–(2.16) *is the unique solution of*

$$a(\boldsymbol{\lambda}_h, \boldsymbol{\mu}) + b(\boldsymbol{\mu}, p_h) = \ell_1(\boldsymbol{\mu}) \qquad \textit{for all } \boldsymbol{\mu} \in M_h \textit{ and} \tag{3.11}$$

$$b(\boldsymbol{\lambda}_h, q) - c(p_h, q) = \ell_2(q) \qquad \textit{for all } q \in P_h. \tag{3.12}$$

Moreover, the solution components ω_h *and* $\boldsymbol{u}_h$ *of the hybridized mixed method* (2.13)–(2.16) *can be determined locally as follows:*

$$\omega_h = \boldsymbol{w}(\boldsymbol{\lambda}_h) + \mathsf{w}(p_h) + \boldsymbol{w}(g_t) + \mathbf{w}(\boldsymbol{f}), \tag{3.13}$$

$$\boldsymbol{u}_h = \boldsymbol{u}(\boldsymbol{\lambda}_h) + \mathbf{u}(p_h) + \boldsymbol{u}(g_t) + \mathbf{u}(\boldsymbol{f}). \tag{3.14}$$

The proof of this theorem proceeds exactly along the lines of the proof of the analogous theorem in [7].

4. Local bases for Lagrange multipliers.

It is clear from section 3 that one should, in practice, implement our hybridized mixed method not in its direct form (2.13)–(2.16), but rather in the reduced form (3.11)–(3.12). This requires a computable basis for the Lagrange multiplier spaces M_h and P_h. Local bases for W_h and V_h are obvious as they do not have continuity constraints across mesh faces. But bases for the Lagrange multiplier spaces are not immediate from their definition, so we develop local bases for P_h and M_h in this section. Note that the construction of the basis for the space of tangential velocities in three space dimensions differs significantly from that of the two-dimensional case.

4.1. The space of interface pressures.

We begin with a characterization of the space of pressure Lagrange multipliers arising from the first hybridization. To state it, define

$$k(F) = \max\{k(K) : K \in \mathfrak{T} \text{ and } K \text{ has } F \text{ as a face}\} \tag{4.1}$$

for every $F \in \mathfrak{F}$, and set $P(F)$ equal to the space of polynomials of degree at most $k(F)$ on the face F.

PROPOSITION 4.1. *The space P_h defined in* (2.11) *is characterized by*

$$P_h = \left\{ p : p|_F \in P(F) \textit{ for all } F \in \mathfrak{F} \textit{ and } \sum_{F\in\mathfrak{F}} (p,1)_F = 0 \right\}.$$

Note that the use of variable degree spaces requires the pressure Lagrange multiplier to have the *maximum* of the degrees from adjacent elements.

The proof of this proposition is quite similar to that of the two-dimensional case considered in [7]. The two main steps of the proof are as follows. In the first, one constructs a local extension $\widetilde{\boldsymbol{v}}_h$ of any given $p \in P_h$ into the Raviart–Thomas space

$$R_h = \{\boldsymbol{r} : \boldsymbol{r}|_K = \boldsymbol{x}p(\boldsymbol{x}) + \boldsymbol{q} \text{ for some } p \in P_{k(K)}(K) \text{ and } \boldsymbol{q} \in P_{k(K)}(K)^3\}$$

such that $[\![\widetilde{\boldsymbol{v}}_h \cdot \boldsymbol{n}]\!] = p$. In the second, one uses a global correction $\boldsymbol{z}_h \in R_h \cap H_0(\text{div}, \Omega)$ such that $\boldsymbol{v}_h = \widetilde{\boldsymbol{v}}_h - \boldsymbol{z}_h$ is in V_h and satisfies $[\![\boldsymbol{v}_h \cdot \boldsymbol{n}]\!] = p$. This is possible by the surjectivity of the divergence map

$$\text{div} : R_h \cap H_0(\text{div}, \Omega) \mapsto S_h,$$

where $S_h = \{v : v|_K \in P_{k(K)}(K) \text{ and average of } v \text{ on } \Omega \text{ is zero}\}$. While this surjectivity is a well-known property for uniform degree spaces, for the variable degree Raviart–Thomas space, it follows from our results in [6]. The remaining details of the proof of Proposition 4.1 are identical to its two-dimensional analogue in [7], so we omit them.

By Proposition 4.1, the Lagrange multiplier space P_h can be identified with $\widetilde{P}_h/\mathbb{R}$, where

$$\widetilde{P}_h = \{p : p|_F \in P_k(F) \text{ for all } F \in \mathfrak{F}\}.$$

Obviously, we can construct a basis for $\widetilde{P}_h$ by taking the union of local bases for $P_k(F)$, say Legendre polynomials, on every edge $F \in \mathfrak{F}$. It is enough to construct such a basis in computations.

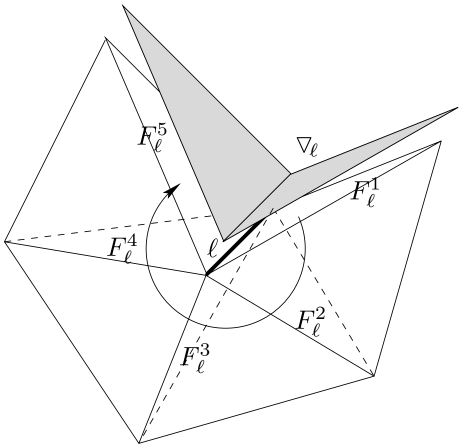


FIG. 1. *Construction of basis functions supported near a mesh edge ℓ.*

4.2. The lowest-order tangential velocity space. Now, we begin the construction of a local basis for the space M_h of tangential velocity Lagrange multipliers. In this subsection, we study the lowest-order case, which is easier to describe. In the next subsection, we consider the general case.

In order to explicitly give a local basis for M_h, we introduce some more notation. Let K be a tetrahedron in $\mathcal{T}$ and let ℓ be one of its edges. We denote by $\varLambda_{\ell,K}$ the union of the two faces of K that share the edge ℓ. Define the collection of such *wedges* by

$$\hat{\varLambda}_h = \{\varLambda_{\ell,K} : \ell \text{ is an edge of } \mathcal{T} \text{ and } K \in \mathcal{T}\}.$$

For all $\varLambda \in \hat{\varLambda}_h$, we denote by $K_\varLambda$ the (unique) tetrahedron $K \in \mathcal{T}$ such that $\varLambda \subseteq \partial K$. The edge of a wedge $\varLambda$ is the common edge of the two faces that form $\varLambda$. This edge is denoted by $\ell_\varLambda$. Let β_i and β_j be the barycentric coordinate functions (with respect to the tetrahedron $K_\varLambda$) associated with the two endpoints of $\ell_\varLambda$. Set

$$\boldsymbol{\phi}_\varLambda = \begin{cases} \beta_i \,\boldsymbol{\nabla}\, \beta_j - \beta_j \,\boldsymbol{\nabla}\, \beta_i & \text{on } K_\varLambda, \\ 0 & \text{on all other } K \in \mathcal{T}. \end{cases}$$

We define a basis for M_h using the functions

$$\boldsymbol{\psi}_\varLambda = [\![\boldsymbol{n} \times \boldsymbol{\phi}_\varLambda]\!].$$

Since $\boldsymbol{\phi}_\varLambda \in W_h$, the functions $\boldsymbol{\psi}_\varLambda$ are in M_h by definition. But not all of $\boldsymbol{\psi}_\varLambda, \varLambda \in \hat{\varLambda}_h$ are linearly independent; e.g., the functions $\boldsymbol{\psi}_\varLambda$ for all $\varLambda$ connected to one edge are linked by one equation. Therefore, for every mesh edge ℓ (including edges $\ell \subset \partial\Omega$), we arbitrarily pick one wedge $\varLambda \in \hat{\varLambda}_h$ with edge $\ell_\varLambda = \ell$, denote it by $\triangledown_\ell$ (see Figure 1), and "omit" it: Define

$$\varLambda_h = \hat{\varLambda}_h \setminus \{\triangledown_\ell : \text{ for all mesh edges } \ell\}.$$

PROPOSITION 4.2. *The set $\mathcal{B}_0 = \{\boldsymbol{\psi}_\varLambda : \varLambda \in \varLambda_h\}$ is a basis for M_h whenever $k(K) = 0$ for all $K \in \mathcal{T}$.*

Proof. Since the span of $\mathcal{B}_0$ is contained in M_h, it suffices to prove that

$$\operatorname{card}\mathcal{B}_0 = \dim M_h \tag{4.2}$$

and

$$\mathcal{B}_0 \text{ is a linearly independent set.} \tag{4.3}$$

To prove (4.2), let us first count the dimension of M_h. Defining $T_h : W_h \mapsto M_h$ by

$$T_h \boldsymbol{\tau} = [\![\boldsymbol{n} \times \boldsymbol{\tau}]\!],$$

we note that M_h is the range of T_h. Since the null space of T_h is $\mathcal{W}_h$, by the rank-nullity theorem, we find that

$$\dim(M_h) = \operatorname{rank}(T_h) = \dim(W_h) - \dim(\mathcal{W}_h). \tag{4.4}$$

Now, $W(K)$ in the lowest-order case is a space of dimension six. Since the number of degrees of freedom of the conforming lowest-order Nédélec space $\mathcal{W}_h$ equals the number of edges n_E in the mesh, we find that

$$\dim(M_h) = 6n_K - n_E,$$

where n_K is the number of tetrahedra in the mesh $\mathcal{T}$. Thus

$$\operatorname{card}\mathcal{B}_0 = \operatorname{card}\Lambda_h = \operatorname{card}\hat{\Lambda}_h - n_E = 6n_K - n_E,$$

which coincides with $\dim M_h$.

Now, let us prove (4.3). We want to show that if

$$\boldsymbol{\mu} = \sum_{\Lambda \in \Lambda_h} c_\Lambda \boldsymbol{\psi}_\Lambda \tag{4.5}$$

vanishes, then all the coefficients c_Λ are zero. Notice that the function $\boldsymbol{\mu}$, in general, is not well defined at the edge ℓ, as the limits of $\boldsymbol{\mu}$ from various faces sharing the edge ℓ can differ. In order to examine these limits, we introduce the following notation. Enumerate all $\Lambda \in \Lambda_h$ with edge ℓ as $\Lambda_\ell^1, \Lambda_\ell^2, \ldots, \Lambda_\ell^{N_\ell}$ and all faces in $\mathcal{F}$ sharing the edge ℓ as $F_\ell^1, F_\ell^2, \ldots, F_\ell^{N_\ell+1}$ (see Figure 1) in such a way that the two faces of Λ_ℓ^j are F_ℓ^j and F_ℓ^{j+1}, and the two faces of $\triangledown_\ell$ are F_ℓ^1 and $F_\ell^{N_\ell+1}$. Let $\boldsymbol{t}_F$ be the unit tangent vector along ∂F fixed by arbitrarily choosing one of the two possible orientations. Let $\boldsymbol{n}_F$ be a unit vector normal to F chosen by the right-hand rule and

$$\boldsymbol{\nu}_F = \boldsymbol{t}_F \times \boldsymbol{n}_F. \tag{4.6}$$

Note that both the choices of orientation for $\boldsymbol{t}_F$ yield the same $\boldsymbol{\nu}_F$, which represents the outward unit normal of F relative to the plane containing F.

Our proof proceeds by examining the following functions on the edge ℓ:

$$\mu_\ell^i := (\boldsymbol{\mu}|_{F_\ell^i}) \cdot \boldsymbol{\nu}_{F_\ell^i}\big|_\ell = \sum_{\Lambda \in \Lambda_h} c_\Lambda \Big(\boldsymbol{\psi}_\Lambda|_{F_\ell^i} \cdot \boldsymbol{\nu}_{F_\ell^i} \Big)\Big|_\ell .$$

Now, there are at most five $\varLambda \in \varLambda_h$ such that $\boldsymbol{\psi}_\varLambda$ is nonzero on the face F_ℓ^1. Moreover, only one of them has nonzero normal trace $\boldsymbol{\psi}_\varLambda \cdot \boldsymbol{\nu}_{F_\ell^i}$ on ℓ, namely, $\boldsymbol{\psi}_{\varLambda_\ell^1}$. Hence

$$\mu_\ell^1 = c_{\varLambda_\ell^1} \left(\boldsymbol{\psi}_{\varLambda_\ell^1}|_{F_\ell^i} \cdot \boldsymbol{\nu}_{F_\ell^1} \right)\Big|_\ell .$$

It then follows that

$$|\mu_\ell^1| = \left|\boldsymbol{\nu}_{F_\ell^1} \cdot (c_{\varLambda_\ell^1} \boldsymbol{\psi}_{\varLambda_\ell^1})\big|_\ell\right| = \left|c_{\varLambda_\ell^1}\, (\boldsymbol{\nu}_{F_\ell^1} \times \boldsymbol{n}_{F_\ell^1}) \cdot \boldsymbol{\phi}_{\varLambda_\ell^1}\big|_\ell\right| = \left|c_{\varLambda_\ell^1}\, (\boldsymbol{t}_{F_\ell^1} \cdot \boldsymbol{\phi}_{\varLambda_\ell^1})\big|_\ell\right| = \frac{1}{h_\ell} |c_{\varLambda_\ell^1}|,$$

where h_ℓ denotes the length of the edge ℓ. Similarly, we also find that $|\mu_\ell^{N_\ell+1}| = |c_\varLambda^{N_\ell+1}|$ and

$$|\mu_\ell^j| = \frac{1}{h_\ell} |c_{\varLambda_x^j} - c_{\varLambda_x^{j-1}}| \qquad \text{for all } j = 2, \dots, N_\ell.$$

If $\boldsymbol{\mu}$ vanishes everywhere, then for any mesh edge ℓ the function μ_ℓ^j defined above must vanish on the edge ℓ. Hence

$$\begin{aligned} |c_{\varLambda_\ell^1}| = |c_{{\varLambda_\ell}^{N_\ell+1}}| &= 0, && \text{and} \\ |c_{\varLambda_\ell^j} - c_{\varLambda_\ell^{j+1}}| &= 0 && \text{for all } j = 2, \dots, N_\ell. \end{aligned}$$

Hence $c_{\varLambda_\ell^j} = 0$ for all j. This argument applies to every mesh edge, so all the coefficients $c_\varLambda$ in (4.5) are zero. Hence (4.3) follows. □

4.3. The higher-order space of tangential velocities. In this subsection we show how to construct a local basis for the Lagrange multiplier space M_h in the general case of the higher-order spaces and the variable degree method. Here there is one important difference compared to the two-dimensional case. In the two-dimensional case [7], we were able to obtain a basis for the higher-order space by augmenting the lowest-order basis with some edge basis functions. In the three-dimensional case, however, we cannot expect to get a basis for the higher-order space by just augmenting $\mathcal{B}_0$ with some face basis functions. This is because while in two dimensions a vertex represents at most one degree of freedom, in three dimensions an edge can have more than one degree of freedom associated to it. Thus we must add to $\mathcal{B}_0$ functions that represent face degrees of freedom as well as functions that represent the additional edge degrees of freedom.

In order to give a basis explicitly, as well as to understand the nature of our space of tangential velocities, it is convenient to recall a basis for the Nédélec space given in [10]. For any integer $k \geq 0$ and any N-simplex D ($N = 2$ or 3 for our purposes), the Nédélec space is

$$W_k(D) = P_k(D)^N \oplus S_{k+1}(D).$$

Let $\beta_1, \dots, \beta_{N+1}$ denote the $N+1$ barycentric coordinate functions of the N-simplex D. Let $I_{lm}(N, k)$ denote the set of all multi-indices $\boldsymbol{\alpha} := (\alpha_1, \dots, \alpha_N)$ (where α_i are nonnegative integers) such that $\alpha_i = 0$ for all i not equal to l or m and $\alpha_l + \alpha_m = k$. Similarly, $I_{lmn}(N, k)$ is the set of multi-indices $\boldsymbol{\alpha}$ with $\alpha_i = 0$ for all i not equal to l, m, or n, and $\alpha_l + \alpha_m + \alpha_n = k$. Using powers of barycentric coordinates (for

$\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{N+1})$, we define $\boldsymbol{\beta}^{\boldsymbol{\alpha}} := \beta_1^{\alpha_1} \dots \beta_{N+1}^{\alpha_{N+1}})$, we introduce the following sets of functions:

$$\mathcal{B}_{lm}^{(D)} = \bigcup_{\boldsymbol{\alpha} \in I_{lm}(N+1,k)} \big\{\boldsymbol{\beta}^{\boldsymbol{\alpha}}(\beta_l \boldsymbol{\nabla} \beta_m - \beta_m \boldsymbol{\nabla} \beta_l)\big\}, \tag{4.7}$$

$$\begin{aligned} \mathcal{B}_{lmn}^{(D)} = \bigcup_{\boldsymbol{\alpha} \in I_{lmn}(N+1,k-1)} & \big\{\boldsymbol{\beta}^{\boldsymbol{\alpha}}(\beta_l\beta_m \boldsymbol{\nabla} \beta_n - \beta_m\beta_n \boldsymbol{\nabla} \beta_l), \\ & \boldsymbol{\beta}^{\boldsymbol{\alpha}}(\beta_m\beta_n \boldsymbol{\nabla} \beta_l - \beta_n\beta_l \boldsymbol{\nabla} \beta_m)\big\}. \end{aligned} \tag{4.8}$$

From the results of [10], it now follows that if D is a triangle, then the union of the sets $\mathcal{B}_{12}^{(D)}$, $\mathcal{B}_{23}^{(D)}$, $\mathcal{B}_{31}^{(D)}$, and $\mathcal{B}_{123}^{(D)}$ forms a basis for the Nédélec space $W_k(D)$. If D is a tetrahedron instead, then a basis for $W_k(D)$ is

$$\mathcal{B}_{12}^{(D)} \cup \mathcal{B}_{13}^{(D)} \cup \mathcal{B}_{14}^{(D)} \cup \mathcal{B}_{23}^{(D)} \cup \mathcal{B}_{24}^{(D)} \cup \mathcal{B}_{34}^{(D)} \cup \mathcal{B}_{123}^{(D)} \cup \mathcal{B}_{124}^{(D)} \cup \mathcal{B}_{134}^{(D)} \cup \mathcal{B}_{234}^{(D)} \cup \mathcal{B}_{1234}^{(D)},$$

where

$$\begin{aligned} \mathcal{B}_{1234}^{(D)} := \bigcup \{ & \boldsymbol{\beta}^{\boldsymbol{\alpha}}(\beta_1\beta_2\beta_3 \boldsymbol{\nabla} \beta_4 - \beta_2\beta_3\beta_4 \boldsymbol{\nabla} \beta_1), \boldsymbol{\beta}^{\boldsymbol{\alpha}}(\beta_2\beta_3\beta_4 \boldsymbol{\nabla} \beta_1 - \beta_3\beta_4\beta_1 \boldsymbol{\nabla} \beta_2), \\ & \boldsymbol{\beta}^{\boldsymbol{\alpha}}(\beta_3\beta_4\beta_1 \boldsymbol{\nabla} \beta_2 - \beta_4\beta_3\beta_2 \boldsymbol{\nabla} \beta_1) : \boldsymbol{\alpha} \in I_{1234}(N+1, k-2)\}. \end{aligned}$$

Note that the basis functions in (4.7) are "edge" basis functions, those in (4.8) are "face" basis functions, and those in $\mathcal{B}_{1234}^{(D)}$ are "interior" basis functions, in the sense explained in [10].

Since the Lagrange multiplier space M_h is obtained using the tangential traces of functions in W_h, it is instructive to study the space of tangential traces of the Nédélec space on one tetrahedron K. Let $\boldsymbol{a}_1$, $\boldsymbol{a}_2$, $\boldsymbol{a}_3$, and $\boldsymbol{a}_4$ be the vertices of K; let F_{lmn} be the face formed by $\boldsymbol{a}_l$, $\boldsymbol{a}_m$, and $\boldsymbol{a}_n$; and let e_{ij} be the edge formed by $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$. We denote by $\boldsymbol{n} \times W_k(K)$ the space of functions on ∂K of the form $\boldsymbol{n} \times \boldsymbol{w}$ for some $\boldsymbol{w} \in W_k(K)$. Recall that for any N-dimensional domain D, the Raviart–Thomas space of polynomials is $R_k(D) = \boldsymbol{x}P_k(D) + P_k(D)^N$, where $\boldsymbol{x}$ is the coordinate vector on D. Define the Raviart–Thomas space on the manifold ∂K by

$$\begin{aligned} R_k(\partial K) = \{\boldsymbol{r} : \ & \boldsymbol{r}|_{F_{ijl}} \in R_k(F_{ijl}) \text{ and} \\ & (\boldsymbol{r}|_{F_{ijl}}) \cdot \boldsymbol{\nu}_{F_{ijl}} + (\boldsymbol{r}|_{F_{ijm}}) \cdot \boldsymbol{\nu}_{F_{ijm}} = 0 \text{ on } e_{ij} \text{ for all } i,j,l,m\}, \end{aligned}$$

where we have used the notation in (4.6). Then we have the following result.

PROPOSITION 4.3. *The space of tangential traces of the Nédélec space* $\boldsymbol{n} \times W_k(K)$ *is the Raviart–Thomas space* $R_k(\partial K)$.

Proof. We begin by proving that $\boldsymbol{n} \times W_k(K) \subseteq R_k(\partial K)$. Let the tangential component of $\boldsymbol{w} \in W_k(K)$ on ∂K be denoted by $\boldsymbol{w}_\mathsf{T}$. We first prove that $\boldsymbol{w}_\mathsf{T}$ on face F_{lmn} is in $R_k(F_{lmn})$. It is easy to see from the structure of the basis functions in (4.7) that if $\boldsymbol{w} \in \mathcal{B}_{ij}^{(K)}$, then $\boldsymbol{w}_\mathsf{T}|_{F_{lmn}}$ is zero if i or j does not belong to $\{l,m,n\}$. If both i and j are in $\{l,m,n\}$, then $\boldsymbol{w}_\mathsf{T}|_{F_{lmn}} \in \mathcal{B}_{ij}^{(F_{lmn})}$. Therefore, we find that $\boldsymbol{w}_\mathsf{T}|_{F_{lmn}}$ is in the Nédélec space $W_k(F_{lmn})$.

In two dimensions, the Nédélec space is the "rotated" Raviart–Thomas space. Indeed, if D is a triangle in the x-y plane, then

$$W_k(D) = P_k(K)^2 \oplus S_{k+1}(K) = P_k(K)^2 \oplus \begin{pmatrix} -y \\ x \end{pmatrix} P_k(K).$$

Since $\boldsymbol{n}\times\boldsymbol{w}|_{F_{lmn}}$ is $\boldsymbol{w}_{\mathsf{T}}|_{F_{lmn}}$ rotated (by an angle of $\pi/2$), it follows that the tangential trace $\boldsymbol{n}\times\boldsymbol{w}$ on F_{lmn} is in the Raviart–Thomas space $R_k(F_{lmn})$.

To show the continuity of the normal components of $\boldsymbol{n}\times\boldsymbol{w}$ across edges of ∂K, consider an edge e_{ij} shared by two faces F_{ijl} and F_{ijm}. Then by (4.6) and the continuity of the tangential components of the Nédélec space, we have the following equalities on e_{ij}:

$$\begin{aligned}(\boldsymbol{n}\times\boldsymbol{w})|_{F_{ijl}}\cdot\boldsymbol{\nu}_{F_{ijl}} &= (\boldsymbol{\nu}_{F_{ijl}}\times\boldsymbol{n})\cdot\boldsymbol{w}|_{Fijl} = -\boldsymbol{t}_{F_{ijl}}\cdot\boldsymbol{w}|_{Fijl}\\ &= \boldsymbol{t}_{F_{ijm}}\cdot\boldsymbol{w}|_{Fijm} = -(\boldsymbol{n}\times\boldsymbol{w})|_{Fijm}\cdot\boldsymbol{\nu}_{F_{ijm}}.\end{aligned}$$

Thus $\boldsymbol{n}\times\boldsymbol{w}$ is in $R_k(\partial K)$ whenever $\boldsymbol{w}\in W_k(K)$. It is easy to see that all functions in $R_k(\partial K)$ can be obtained as tangential traces of $W_k(K)$. $\square$

Now we are ready to describe the building blocks of a basis for the general higher-order M_h arising from the variable degree Nédélec spaces. The basis is divided into two parts: one corresponding to the interior faces of the mesh and another corresponding to the wedges in Λ_h. The former is easy to describe: Let $\mathcal{F}_0$ denote the set of all interior faces of the mesh $\mathcal{T}$. For any face $F\in\mathcal{F}_0$, define

$$\mathring{V}(F) = \{\boldsymbol{w}\in R_{k(F)}(F) : \boldsymbol{w}|_{\partial F}\cdot\boldsymbol{\nu}_F = 0 \text{ on } \partial F\},$$

where $k(F)$ is the *maximum* of the degrees from either side of F, as defined in (4.1). Let $\mathring{\mathcal{B}}_F$ be a basis for $\mathring{V}(F)$.

To describe the wedge basis functions, recall the notations introduced in the previous subsection. Now we additionally require that for every mesh edge ℓ, the "omitted wedge" $\triangledown_\ell$ is associated to a tetrahedron (having ℓ as an edge and) having the *minimal* degree: More precisely, we choose $\triangledown_\ell$ such that

$$k(K_{\triangledown_\ell}) = \min_{i=1,\ldots,N_\ell} k(K_{\Lambda^i_\ell}). \tag{4.9}$$

For all the remaining $\Lambda\in\Lambda_h$, we define the following Raviart–Thomas-type space:

$$R(\Lambda) = \{\boldsymbol{r}\in R_{k(K_\Lambda)}(\partial K_\Lambda) :\ \boldsymbol{r} \text{ is supported on } \Lambda\}.$$

Just as we decompose the standard Raviart–Thomas space, we can decompose $R(\Lambda)$ into subspaces corresponding to interior and boundary degrees of freedom: If F_Λ^+ and F_Λ^- denote the two faces of Λ and $\mathring{R}(F_\Lambda^\pm) = \{\boldsymbol{r}\in R_{k(K_\Lambda)}(\partial K_\Lambda) :\ \boldsymbol{r}$ is supported on $F_\Lambda^\pm\}$, we can decompose $R(\Lambda) = \mathring{R}(F_\Lambda^+)\oplus\mathring{R}(F_\Lambda^-)\oplus V(\Lambda)$, where $V(\Lambda)$ is a subspace that is linearly independent, to $\mathring{R}(F_\Lambda^+)\oplus\mathring{R}(F_\Lambda^-)$; e.g., we can choose $V(\Lambda)$ to be the $L^2(\Lambda)$-orthogonal complement of $\mathring{R}(F_\Lambda^+)\oplus\mathring{R}(F_\Lambda^-)$ in $R(\Lambda)$. (Another alternative is suggested in the next paragraph.) Let $\mathcal{B}_\Lambda$ be a basis for $V(\Lambda)$. Our next theorem shows that such wedge basis functions together with the face basis functions form a basis for the global space M_h.

Particular examples of $\mathcal{B}_\Lambda$ and $\mathring{\mathcal{B}}_F$ are easy to exhibit. We give one conveniently implementable choice that follows from the previous results of [10]. Let $\Lambda\in\Lambda_h$ and let β_i, $i=1,2,3,4$, denote the barycentric coordinates of K_Λ such that β_i and β_j are associated to the two endpoints of the edge ℓ_Λ. Define

$$\boldsymbol{\phi}_\Lambda^{(\boldsymbol{\alpha})} = \begin{cases}\boldsymbol{\beta}^{\boldsymbol{\alpha}}(\beta_i\,\boldsymbol{\nabla}\,\beta_j - \beta_j\,\boldsymbol{\nabla}\,\beta_i) & \text{on } K_\Lambda,\\ 0 & \text{on all other } K\in\mathcal{T}\end{cases}$$

for all $\boldsymbol{\alpha} \in I_{ij}(4, k(K_\Lambda))$ and

$$\boldsymbol{\psi}_\Lambda^{(\boldsymbol{\alpha})} = [\![\boldsymbol{n} \times \boldsymbol{\phi}_\Lambda^{(\boldsymbol{\alpha})}]\!].$$

We can choose

$$\mathcal{B}_\Lambda = \{\boldsymbol{\psi}_\Lambda^{(\boldsymbol{\alpha})} : \boldsymbol{\alpha} \in I_{ij}(4, k(K_\Lambda))\}.$$

For an example of a face basis, let $F \in \mathcal{F}_0$. If β_i, β_j, and β_k are the three barycentric coordinate functions of the face F, then we may choose

$$\mathring{\mathcal{B}}_F = \bigcup_{\boldsymbol{\alpha} \in I_{ijk}(3, k(F)-1)} \Big\{ \boldsymbol{\beta}^{\boldsymbol{\alpha}} (\beta_i \beta_j \,\boldsymbol{\nabla}\, \beta_k - \beta_j \beta_k \,\boldsymbol{\nabla}\, \beta_i) \times \boldsymbol{n}_F, \quad \boldsymbol{\beta}^{\boldsymbol{\alpha}} (\beta_j \beta_k \,\boldsymbol{\nabla}\, \beta_i - \beta_k \beta_i \,\boldsymbol{\nabla}\, \beta_j) \times \boldsymbol{n}_F \Big\}. \tag{4.10}$$

The following theorem gives a basis for M_h.

THEOREM 4.4. *The set*

$$\mathcal{B} = \Big(\bigcup_{\Lambda \in \Lambda_h} \mathcal{B}_\Lambda \Big) \cup \Big(\bigcup_{F \in \mathcal{F}_0} \mathring{\mathcal{B}}_F \Big)$$

is a basis for M_h.

Proof. It follows from Proposition 4.3 that elements of $\mathring{\mathcal{B}}_F$ and $\mathcal{B}_\Lambda$ can be written as $[\![\boldsymbol{n} \times \boldsymbol{\phi}]\!]$ for some $\boldsymbol{\phi} \in W_h$. Hence the span of $\mathcal{B}$ is contained in M_h. It now suffices to prove that

$$\operatorname{card} \mathcal{B} = \dim(M_h) \tag{4.11}$$

and that $\mathcal{B}$ is a linearly independent set. For any $\boldsymbol{\mu} \in \mathring{\mathcal{B}}_F$, the normal trace from F on ∂F vanishes:

$$(\boldsymbol{\mu}|_{\partial F}) \cdot \boldsymbol{\nu}_F = 0.$$

The normal traces of functions in $\mathcal{B}_\Lambda$ from Λ on ℓ_Λ are linearly independent. Hence by a minor modification of the arguments in the proof of Proposition 4.2, the linear independence of $\mathcal{B}$ follows from the linear independence of functions within $\mathcal{B}_\Lambda$ and $\mathring{\mathcal{B}}_F$.

To prove (4.11), let us first count the number of elements in $\mathcal{B}$. The dimension of $\mathring{V}(F)$ can be calculated easily (either directly or using (4.10)). It equals

$$\operatorname{card} \mathring{\mathcal{B}}_F = 2 \operatorname{card} I_{123}(3, k(F)-1) = k(F)\big(k(F)+1\big).$$

Moreover,

$$\operatorname{card} \mathcal{B}_\Lambda = \operatorname{card} I_{12}(4, k(K_\Lambda)) = k(K_\Lambda) + 1.$$

Thus,

$$\operatorname{card} \mathcal{B} = \sum_{\Lambda \in \Lambda_h} (k(K_\Lambda) + 1) + \sum_{F \in \mathcal{F}_0} k(F)\big(k(F)+1\big). \tag{4.12}$$

Now let us compute the dimension of M_h by using the identity (see (4.4))

$$\dim(M_h) = \dim(W_h) - \dim(\mathcal{W}_h).$$

By the tangential continuity conditions on the variable degree space $\mathcal{W}_h$, we find that the space of traces $\boldsymbol{n}_F \times \boldsymbol{w}$ on a face $F \in \mathcal{F}$ for $\boldsymbol{w} \in \mathcal{W}_h$ is $R_{\underset{\sim}{k}(F)}(F)$, where

$$\underset{\sim}{k}(F) = \min\{k(K) : K \in \mathcal{T} \text{ and } K \text{ has } F \text{ as a face}\}.$$

Furthermore, the tangential component $\boldsymbol{w} \cdot \boldsymbol{t}$ on an edge E is in $P_{\underset{\sim}{k}(E)}(E)$, where

$$\underset{\sim}{k}(E) = \min\{k(K) : K \in \mathcal{T} \text{ and } K \text{ has } E \text{ as an edge}\}.$$

Splitting the global degrees of freedom of $\mathcal{W}_h$ as edge degrees of freedom, face degrees of freedom, and interior degrees of freedom, we find that

$$\dim(\mathcal{W}_h) = \sum_{E\in\mathcal{E}} \big(\underset{\sim}{k}(E)+1\big) + \sum_{F\in\mathcal{F}} \underset{\sim}{k}(F)\big(\underset{\sim}{k}(F)+1\big) + \sum_{K\in\mathcal{T}} \frac{1}{2}\big(k(K)-1\big)k(K)\big(k(K)+1\big).$$

Consequently,

$$\begin{aligned} \dim(W_h) - \dim(\mathcal{W}_h) = {}& \left(\sum_{K\in\mathcal{T}} 6\big(k(K)+1\big) - \sum_{E\in\mathcal{E}} \big(\underset{\sim}{k}(E)+1\big)\right) \\ & + \left(\sum_{K\in\mathcal{T}} 4k(K)\big(k(K)+1\big) - \sum_{F\in\mathcal{F}} \underset{\sim}{k}(F)\big(\underset{\sim}{k}(F)+1\big)\right). \end{aligned} \tag{4.13}$$

Because of (4.4), it suffices to show that the above equals card $\mathcal{B}$.

In order to do this, we simplify the right-hand side of (4.13). Observe that by rearrangement,

$$\sum_{K\in\mathcal{T}} 4k(K)\big(k(K)+1\big) = \sum_{F\in\mathcal{F}} \Big(k(K_F^+)\big(k(K_F^+)+1\big) + k(K_F^-)\big(k(K_F^-)+1\big)\Big),$$

where $K_F^\pm$ is as defined earlier and one of $k(K_F^\pm)$ is understood to vanish if $F \subseteq \partial\Omega$. Hence

$$\sum_{K\in\mathcal{T}} 4k(K)\big(k(K)+1\big) - \sum_{F\in\mathcal{F}} \underset{\sim}{k}(F)\big(\underset{\sim}{k}(F)+1\big) = \sum_{F\in\mathcal{F}_0} k(F)\big(k(F)+1\big).$$

Similarly, denoting by K_ℓ^i, $i = 1, 2, \ldots, N_\ell$, the tetrahedra in $\mathcal{T}$ which have ℓ as an edge, the rearrangement

$$\sum_{K\in\mathcal{T}} 6\big(k(K)+1\big) = \sum_{\ell\in\mathcal{E}} \sum_{i=1}^{N_\ell} \big(k(K_\ell^i)+1\big)$$

implies, in view of (4.9), that

$$\begin{aligned} \sum_{K\in\mathcal{T}} 6\big(k(K)+1\big) - \sum_{E\in\mathcal{E}} \big(\underset{\sim}{k}(E)+1\big) &= \sum_{\ell\in\mathcal{E}} \left(\sum_{i=1}^{N_\ell} k(K_\ell^i)+1\right) - \sum_{\ell\in\mathcal{E}} \big(k(K_{\triangledown_\ell})+1\big) \\ &= \sum_{\Lambda\in\Lambda_h} \big(k(K_\Lambda)+1\big). \end{aligned}$$

Using these identities in (4.13), we obtain

$$\dim(W_h) - \dim(\mathcal{W}_h) = \sum_{\mathit{\Lambda} \in \mathit{\Lambda}_h} \big(k(K_{\mathit{\Lambda}}) + 1\big) + \sum_{F \in \mathcal{F}_0} k(F)\big(k(F) + 1\big),$$

which coincides with card $\mathcal{B}$ as computed in (4.12). Hence (4.11) follows. $\square$

5. Formulae for the lowest-order case. In [7], we discussed a few implementation techniques to implement and solve the two-dimensional analogue of the Lagrange multiplier system (3.11)–(3.12). The considerations there apply to the three-dimensional case as well. In particular, one can form the stiffness matrix of (3.11)–(3.12) and then perform one further elimination (of the pressure multiplier) to obtain a Schur complement system involving the tangential velocity variable $\boldsymbol{\lambda}_h$ alone. We do not repeat this and other details discussed in [7]. However, since the formulae for the liftings change in three dimensions, we give here new formulae for the liftings as well as local stiffness matrices for the lowest-order case.

First, consider the local maps which define the linear and bilinear forms in (3.11)–(3.12) for the lowest-order case (i.e., $k(K) = 0$ for all $K \in \mathcal{T}$). Let K be any tetrahedron in $\mathcal{T}$. Simple computations show that

$$\begin{aligned}
\boldsymbol{w}(\boldsymbol{\lambda}) &= \frac{1}{|K|}\int_{\partial K\setminus\partial\Omega} \boldsymbol{\lambda} \times \boldsymbol{n}\,\mathrm{d}s, &\quad \boldsymbol{u}(\boldsymbol{\lambda}) &= \frac{1}{2|K|}\int_{\partial K\setminus\partial\Omega} (\boldsymbol{x} - \boldsymbol{x}_K) \times (\boldsymbol{n} \times \boldsymbol{\lambda})\,\mathrm{d}s,\\
\boldsymbol{w}(\boldsymbol{g}_{\mathsf{T}}) &= \frac{1}{|K|}\int_{\partial K\cap\partial\Omega} \boldsymbol{g}_{\mathsf{T}} \times \boldsymbol{n}\,\mathrm{d}s, &\quad \boldsymbol{u}(\boldsymbol{g}_{\mathsf{T}}) &= \frac{1}{2|K|}\int_{\partial K\cap\partial\Omega} (\boldsymbol{x} - \boldsymbol{x}_K) \times (\boldsymbol{n} \times \boldsymbol{g}_{\mathsf{T}})\,\mathrm{d}s,\\
\mathcal{w}(p) &= \mathcal{w}_p^K \times (\boldsymbol{x} - \boldsymbol{x}_K), &\quad \mathbf{u}(p) &= \frac{1}{2|K|}\int_K (\boldsymbol{x} - \boldsymbol{x}_K) \times \mathcal{w}(p)\,\mathrm{d}x,\\
\mathsf{w}(\boldsymbol{f}) &= \mathsf{w}_{\boldsymbol{f}}^K \times (\boldsymbol{x} - \boldsymbol{x}_K), &\quad \mathsf{u}(\boldsymbol{f}) &= \frac{1}{2|K|}\int_K (\boldsymbol{x} - \boldsymbol{x}_K) \times \mathsf{w}(\boldsymbol{f})\,\mathrm{d}x,
\end{aligned}$$

where the point $\boldsymbol{x}_K$ denotes the barycenter of the tetrahedron K,

$$\mathcal{w}_p^K = -\frac{1}{2|K|}\int_{\partial K} p\,\boldsymbol{n}\,\mathrm{d}s, \quad \text{and} \quad \mathsf{w}_{\boldsymbol{f}}^K = \frac{1}{2|K|}\int_K \boldsymbol{f}\,\mathrm{d}x.$$

Here and elsewhere we use $|X|$ to denote the measure of X.

In order to implement (3.11)–(3.12), one uses the basis for M_h and P_h described previously, applies the above local lifting maps to the basis functions, and forms local stiffness matrices of the bilinear forms of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$. In line with these steps, we next simplify the above expressions in the case of a lowest-order basis function of M_h and P_h. Let K be the tetrahedron formed by vertices $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$. Let F_i denote the face of K opposite to vertex x_i, and let $\boldsymbol{n}_i$ denote the outward unit normal of K on the face F_i. We first give the lifting of $\boldsymbol{\psi}_{\mathit{\Lambda}}$, a basis function associated with $\mathit{\Lambda} \in \mathit{\Lambda}_h$ with $K_{\mathit{\Lambda}} = K$ (see Figure 2). Since we are considering the lowest-order case, by definition,

$$\boldsymbol{\psi}_{\mathit{\Lambda}} = \begin{cases} \boldsymbol{n}_3 \times (\beta_1\,\boldsymbol{\nabla}\,\beta_2 - \beta_2\,\boldsymbol{\nabla}\,\beta_1) & \text{on face } F_3,\\ \boldsymbol{n}_4 \times (\beta_1\,\boldsymbol{\nabla}\,\beta_2 - \beta_2\,\boldsymbol{\nabla}\,\beta_1) & \text{on face } F_4,\\ 0 & \text{on all other mesh faces.} \end{cases}$$

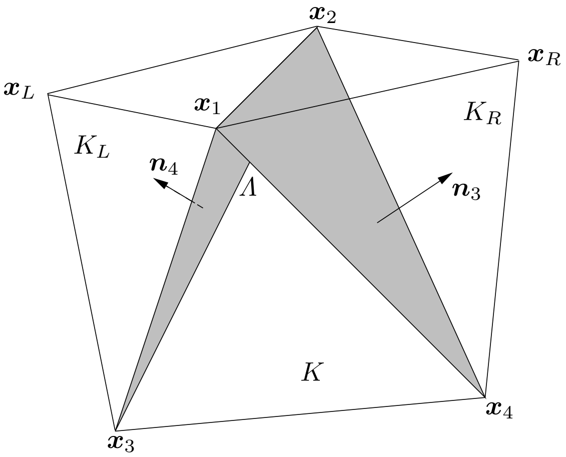


FIG. 2. *The lifting of the basis function from Λ is supported on three mesh tetrahedra K, K_L, and K_R.*

It is easily seen that the above expression is equal to the following:

$$\boldsymbol{\psi}_\Lambda = \begin{cases} -\dfrac{1}{2|F_3|}(\boldsymbol{x}-\boldsymbol{x}_4) & \text{on face } F_3, \\ -\dfrac{1}{2|F_4|}(\boldsymbol{x}-\boldsymbol{x}_3) & \text{on face } F_4, \\ 0 & \text{on all other mesh faces.} \end{cases}$$

The computations are simplified by working with the latter expression for $\boldsymbol{\psi}_\Lambda$, which also illustrates the connection of the tangential traces with the Raviart–Thomas space. The liftings $\boldsymbol{w}_\Lambda := \boldsymbol{w}(\boldsymbol{\psi}_\Lambda)$ and $\boldsymbol{u}_\Lambda := \boldsymbol{u}(\boldsymbol{\psi}_\Lambda)$ are supported on three tetrahedra, unless Λ intersects $\partial\Omega$. Since the formulae one obtains when Λ intersects $\partial\Omega$ are similar to the remaining cases, we consider only the case shown in Figure 2, where the lifting is supported on the three tetrahedra shown, namely, K, K_L, and K_R. Letting $\boldsymbol{x}_{ij} = \boldsymbol{x}_i - \boldsymbol{x}_j$ for any subscripts i and j, we have

$$\begin{aligned}
\boldsymbol{w}_\Lambda &= -\frac{(\boldsymbol{x}_{31}+\boldsymbol{x}_{32})\times\boldsymbol{n}_4}{6|K_L|} \quad \text{and} \\
\boldsymbol{u}_\Lambda &= \frac{-1}{48|K_L|}\Big[(\boldsymbol{x}_{13}\times\boldsymbol{n}_4)\times\boldsymbol{x}_{L1} + (\boldsymbol{x}_{23}\times\boldsymbol{n}_4)\times\boldsymbol{x}_{L2}\Big] \quad \text{on } K_L, \\
\boldsymbol{w}_\Lambda &= -\frac{(\boldsymbol{x}_{41}+\boldsymbol{x}_{42})\times\boldsymbol{n}_3}{6|K_R|} \quad \text{and} \\
\boldsymbol{u}_\Lambda &= \frac{-1}{48|K_R|}\Big[(\boldsymbol{x}_{14}\times\boldsymbol{n}_3)\times\boldsymbol{x}_{R1} + (\boldsymbol{x}_{24}\times\boldsymbol{n}_3)\times\boldsymbol{x}_{R2}\Big] \quad \text{on } K_R, \\
\boldsymbol{w}_\Lambda &= \left[\frac{(\boldsymbol{x}_{31}+\boldsymbol{x}_{32})\times\boldsymbol{n}_4}{6|K|} + \frac{(\boldsymbol{x}_{41}+\boldsymbol{x}_{42})\times\boldsymbol{n}_3}{6|K|}\right] \quad \text{and} \\
\boldsymbol{u}_\Lambda &= \frac{1}{48|K|}\Big[(\boldsymbol{x}_{13}\times\boldsymbol{n}_4)\times\boldsymbol{x}_{41} + (\boldsymbol{x}_{23}\times\boldsymbol{n}_4)\times\boldsymbol{x}_{42} + (\boldsymbol{x}_{14}\times\boldsymbol{n}_3)\times\boldsymbol{x}_{31} \\
&\qquad + (\boldsymbol{x}_{24}\times\boldsymbol{n}_3)\times\boldsymbol{x}_{32}\Big] \quad \text{on } K.
\end{aligned}$$

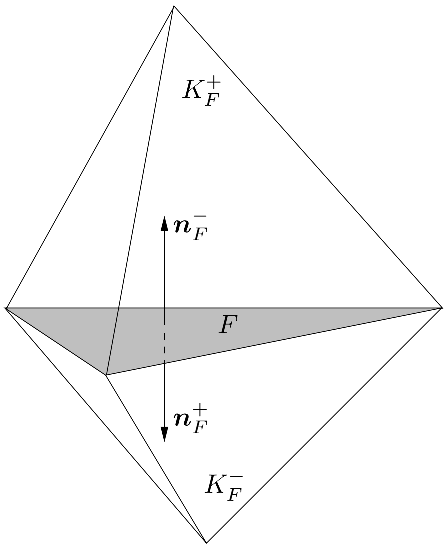


FIG. 3. *The lifting of the pressure basis function from a face F is supported on the tetrahedra adjacent to the face F.*

Next, let us derive the liftings associated with the pressure. To treat this case, consider a face F (shared by the tetrahedra K_F^+ and K_F^-; see Figure 3). Let p_F denote the indicator function of edge F. The liftings $\mathcal{w}_F := \mathcal{w}(p_F)$ and $\mathbf{u}_F := \mathbf{u}(p_F)$ are supported on $K_F^+ \cup K_F^-$. Let $\boldsymbol{x}_i^\pm$, $i = 1, \dots, 4$, denote any enumeration of the four vertices of $K_F^\pm$. In accordance with our previous notation, set $\boldsymbol{x}_{K_F^\pm}$ equal to the barycenter of $K_F^\pm$ and $\boldsymbol{x}_{iK} = \boldsymbol{x}_i^\pm - \boldsymbol{x}_{K_F^\pm}$. We can express the liftings on $K_F^\pm$ by

$$\mathcal{w}_F(\boldsymbol{x})|_{K_F^\pm} = \mathcal{w}^\pm \times (\boldsymbol{x} - \boldsymbol{x}_{K_F^\pm}), \qquad \mathbf{u}_F(\boldsymbol{x})|_{K_F^\pm} = \frac{1}{40}\sum_{i=1}^{4} \boldsymbol{x}_{iK} \times (\mathcal{w}^\pm \times \boldsymbol{x}_{iK}),$$

where

$$\mathcal{w}^\pm = -\frac{|F|}{2|K_F^\pm|}\boldsymbol{n}_F^\pm$$

and $\boldsymbol{n}_F^\pm$ denotes the outward unit normal of $K_F^\pm$ on F (see Figure 3).

The formulae for the maps associated with the body force are similar. If $\boldsymbol{f}$ is supported only on K, then $\mathsf{w}(\boldsymbol{f})$ and $\mathsf{u}(\boldsymbol{f})$ are supported only on K. Their values on K are given by

$$\mathsf{w}(\boldsymbol{f}) = \mathsf{w} \times (\boldsymbol{x} - \boldsymbol{x}_K), \qquad \mathsf{u}(\boldsymbol{f}) = \frac{1}{40}\sum_{i=1}^{4} \boldsymbol{x}_{iK} \times (\mathsf{w} \times \boldsymbol{x}_{iK}),$$

where

$$\mathsf{w} = \frac{1}{2|K|}\int_K \boldsymbol{f}\, \mathrm{d}x.$$

Now that we have expressions for the liftings of the basis functions, we can easily compute the local stiffness matrices of the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ with respect to the basis. Once the local matrices are made, one assembles them to get the global matrices in much the same way as one does for standard finite element methods. To compute the local stiffness matrix, we first list the degrees of freedom local to an element. In this list, we include the omitted elements of $\hat{\Lambda}_h$. The omissions can be taken care of after assembly by simply deleting the rows and columns corresponding to the omitted elements of $\hat{\Lambda}_h$. To geometrically identify the degrees of freedom on an element K, let $\boldsymbol{x}_i$ denote the vertices of K and let E_{ij} denote the edge of K with endpoints $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. There are six wedge degrees of freedom interior to K, which we denote by Λ_{ij} for $ij \in \mathcal{I}_0 := \{12, 13, 14, 23, 24, 34\}$. The wedge Λ_{ij} is geometrically identified as the wedge contained in ∂K with edge E_{ij}. In addition, there are twelve degrees of freedom from wedges "exterior" to K that contribute to the local stiffness matrix of K. We denote these as Λ_{ijk} for $ijk \in \mathcal{I}_1 := \{ijk : ij \in \mathcal{I}_0$ and k does not equal i or $j\}$ (cf. [7, Figure 4]). The wedge Λ_{ijk} is the (unique) wedge with edge E_{ij}, whose one face coincides with the face of K formed by vertices $\boldsymbol{x}_i$, $\boldsymbol{x}_j$, and $\boldsymbol{x}_k$, and whose other face is not contained in ∂K. Thus all wedge degrees of freedom within an element can be identified using the index set $\mathcal{I} = \mathcal{I}_0 \cup \mathcal{I}_1$. The pressure degrees of freedom are easier to enumerate: There is one for each face of K, so they can be identified using the index set $\mathcal{L} := \{1, 2, 3, 4\}$. The local stiffness matrices associated to an element K can now be given by

$$\mathsf{A}^{(K)}_{IJ} = \int_K \boldsymbol{w}(\boldsymbol{\psi}_{\Lambda_I}) \cdot \boldsymbol{w}(\boldsymbol{\psi}_{\Lambda_J})\, \mathrm{d}x, \qquad I, J \in \mathcal{I},$$

$$\mathsf{B}^{(K)}_{LJ} = -\int_K \mathbf{curl}\, \mathsf{w}(p_L) \cdot \boldsymbol{u}(\boldsymbol{\psi}_{\Lambda_J})\, \mathrm{d}x, \qquad J \in \mathcal{I}, \quad L \in \mathcal{L},$$

$$\mathsf{C}^{(K)}_{LM} = \int_K \mathsf{w}(p_L) \cdot \mathsf{w}(p_M)\, \mathrm{d}x, \qquad L, M \in \mathcal{L}.$$

Here, as before, p_L denotes the characteristic function of the face F_L for all $L \in \mathcal{L}$.

We can calculate the integrals above after substituting the previously given expressions for the liftings of the basis functions into the integrands. To take into account modifications required near the boundary $\partial\Omega$, let σ_j equal zero if the face F_j is contained in the boundary $\partial\Omega$ and let σ_j equal one otherwise. The simplified expressions for $\mathsf{A}^{(K)}$, $\mathsf{B}^{(K)}$, and $\mathsf{C}^{(K)}$ for any element K are given below. Suppose that $\{i, j, k, l\}$ is any permutation of $\{1, 2, 3, 4\}$. Then define

$$\boldsymbol{W}_I = \begin{cases} \sigma_l(\boldsymbol{x}_{ki} + \boldsymbol{x}_{kj}) \times \boldsymbol{n}_l + \sigma_k(\boldsymbol{x}_{li} + \boldsymbol{x}_{lj}) \times \boldsymbol{n}_k & \text{if } I = ij, \\ -\sigma_l(\boldsymbol{x}_{ki} + \boldsymbol{x}_{kj}) \times \boldsymbol{n}_l & \text{if } I = ijk, \end{cases}$$

$$\boldsymbol{U}_I = \begin{cases} \sigma_l(\boldsymbol{x}_{ik} \times \boldsymbol{n}_l) \times \boldsymbol{x}_{li} + (\boldsymbol{x}_{jk} \times \boldsymbol{n}_l) \times \boldsymbol{x}_{kj} & \\ \qquad + \sigma_k(\boldsymbol{x}_{il} \times \boldsymbol{n}_k) \times \boldsymbol{x}_{ki} + (\boldsymbol{x}_{jl} \times \boldsymbol{n}_k) \times \boldsymbol{x}_{lj} & \text{if } I = ij, \\ -\sigma_l(\boldsymbol{x}_{ik} \times \boldsymbol{n}_l) \times \boldsymbol{x}_{li} + (\boldsymbol{x}_{jk} \times \boldsymbol{n}_l) \times \boldsymbol{x}_{kj} & \text{if } I = ijk. \end{cases}$$

After a few simplifications, one finds that

$$\mathsf{A}^{(K)}_{IJ} = \frac{1}{36|K|} \boldsymbol{W}_I \cdot \boldsymbol{W}_J,$$

$$\mathsf{B}^{(K)}_{LJ} = \frac{1}{48|K|} \boldsymbol{E}_L \cdot \boldsymbol{U}_J,$$

$$\mathsf{C}^{(K)}_{LM} = \frac{1}{80|K|} \sum_{\ell=1}^{4} (\boldsymbol{E}_L \times \boldsymbol{x}_{\ell K}) \cdot (\boldsymbol{E}_M \times \boldsymbol{x}_{\ell K}),$$

where $\boldsymbol{E}_L = \boldsymbol{n}_L|F_L|$ for all $L \in \mathcal{L}$. Using these local matrices, it is quite easy to implement the lowest-order case of our method, even for general tetrahedral meshes. For the variable degree and higher-order cases, one would need to select a good basis for the polynomial spaces involved on one element and then perform the above steps within a computer implementation. Our calculations above, besides showing the essential simplicity of our discretization in the lowest-order case, also clarify the data structures one would need in implementing the method.

6. Extension to other boundary conditions.

Although the previously considered Dirichlet boundary condition on velocity is the most commonly occurring boundary condition in the Stokes problem, other types of boundary conditions are also encountered in practice. One can have boundary conditions on pressure of the form

$$p = s$$

and boundary conditions on tangential vorticity of the form

$$\boldsymbol{n} \times \boldsymbol{\omega} = \boldsymbol{r}.$$

Here s and $\boldsymbol{r}$ are functions prescribed on parts of the boundary $\partial\Omega$. We now show how one may incorporate such boundary conditions into our hybridized discretization. Note that the above types of boundary conditions are difficult to impose in a natural fashion in many existing methods—see remarks in [11, section 4.3] and [12]. They are often practically important. For example, pressure is often used as an outflow condition. The tangential vorticity boundary condition is useful when matching an exterior potential flow since vorticity is known to decay faster than velocity. The tangential vorticity boundary condition has been considered previously in [8] in formulations with the stream function.

Assume that the polyhedral boundary $\partial\Omega$ is partitioned into three disjoint subsets Γ_1, Γ_2, and Γ_3 such that each mesh face $F \in \mathcal{F}$ on the boundary $\partial\Omega$ is contained in one and only one of these three subsets. We consider the Stokes equations (2.1)–(2.2) with the following boundary conditions:

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{g} && \text{on } \Gamma_1,\\ \left.\begin{aligned}\boldsymbol{n}\times\boldsymbol{\omega} &= \boldsymbol{r}\\ \boldsymbol{u}\cdot\boldsymbol{n} &= g_n\end{aligned}\right\} && && \text{on } \Gamma_2,\\ \left.\begin{aligned}p &= s\\ \boldsymbol{u}_{\mathsf{T}} &= \boldsymbol{g}_{\mathsf{T}}\end{aligned}\right\} && && \text{on } \Gamma_3.\end{aligned}$$

A straightforward generalization of our method can be obtained in this case.

To describe this generalization, we first redefine the jump-functions as follows: The functions $[\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]$ and $[\![\boldsymbol{n}\times\boldsymbol{\tau}]\!]$ are defined just as before on the interior faces, but for mesh faces F on the boundary we set

$$[\![\boldsymbol{n}\cdot\boldsymbol{v}]\!]_F = \begin{cases} 0 \text{ for all faces } F \subseteq \Gamma_3,\\ \boldsymbol{n}\cdot\boldsymbol{v} \text{ for the remaining faces } F \subseteq \partial\Omega\setminus\Gamma_3\end{cases}$$

and

$$[\![\boldsymbol{n}\times\boldsymbol{\tau}]\!]_F = \begin{cases} 0 \text{ for all faces } F \subseteq \Gamma_3\cup\Gamma_1,\\ \boldsymbol{n}\times\boldsymbol{\tau} \text{ for the remaining faces } F \subseteq \partial\Omega\setminus(\Gamma_3\cup\Gamma_1).\end{cases}$$

Then along the lines of the derivation of (2.13)–(2.16), we can derive the following hybridized mixed formulation: Find $(\boldsymbol{\omega}_h, \boldsymbol{u}_h, \boldsymbol{\lambda}_h, p_h) \in W_h \times V_h \times M_h \times P_h$ satisfying

$$\begin{aligned}(\boldsymbol{\omega}_h, \boldsymbol{\tau}_h)_\Omega - (\boldsymbol{u}_h, \mathbf{curl}\,\boldsymbol{\tau}_h)_\Omega - \sum_{F\in\mathcal{F}} (\boldsymbol{\lambda}_h, [\![\boldsymbol{n}\times\boldsymbol{\tau}_h]\!])_F &= (\boldsymbol{g}_{\mathsf{T}}, \boldsymbol{n}\times\boldsymbol{\tau}_h)_{\Gamma_1\cup\Gamma_3},\\ (\boldsymbol{v}_h, \mathbf{curl}\,\boldsymbol{\omega}_h)_\Omega + \sum_{F\in\mathcal{F}} (p_h, [\![\boldsymbol{v}_h\cdot\boldsymbol{n}]\!])_F &= (\boldsymbol{f}, \boldsymbol{v}_h)_\Omega - (s, \boldsymbol{v}_h\cdot\boldsymbol{n})_{\Gamma_3},\\ \sum_{F\in\mathcal{F}} (q_h, [\![\boldsymbol{u}_h\cdot\boldsymbol{n}]\!])_F &= (g_n, q_h)_{\Gamma_1\cup\Gamma_2},\\ \sum_{F\in\mathcal{F}} (\boldsymbol{\mu}_h, [\![\boldsymbol{n}\times\boldsymbol{\omega}_h]\!])_F &= (\boldsymbol{\mu}_h, \boldsymbol{r})_{\Gamma_2}\end{aligned}$$

for all $\boldsymbol{\tau}_h \in W_h, \boldsymbol{v}_h \in V_h, q_h \in P_h, \boldsymbol{\mu}_h \in M_h$. Here W_h and V_h are the same spaces as before. The spaces of Lagrange multipliers P_h and M_h continue to be defined by (2.11) and (2.12), but now with the revised definition of jump-functions.

For this formulation, we can prove, by a minor modification of the argument used in Proposition 2.1, that there is one and only one solution. Moreover, the entire analysis of section 3 goes through with minor changes. We obtain a reduced Lagrange multiplier system and can formulate a theorem entirely analogous to Theorem 3.1. The discussion of the liftings and the basis functions in the previous sections continues to apply for these boundary conditions.

The method we presented in this paper gives a powerful alternative for problems in computational fluid mechanics which require exactly divergence-free solutions for their successful treatment. Applications to such problems, the error analysis of the method, and the design of good preconditioners for solving the resulting matrix equations are subjects of ongoing work.

REFERENCES

[1] C. Amrouche, C. Bernardi, M. Dauge, and V. Girault, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.

[2] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[3] A. Bendali, J. M. Domínguez, and S. Gallic, *A variational approach for the vector potential formulation of the Stokes and Navier-Stokes problems in three-dimensional domains*, J. Math. Anal. Appl., 107 (1985), pp. 537–560.

[4] J. Carrero, B. Cockburn, and D. Schötzau, *Hybridized, globally divergence-free LDG methods. Part* I: *The Stokes problem*, Math. Comp., to appear.

[5] B. Cockburn and J. Gopalakrishnan, *A characterization of hybridized mixed methods for the second order elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 283–301.

[6] B. Cockburn and J. Gopalakrishnan, *Error analysis of variable degree mixed methods for elliptic problems via hybridization*, Math. Comp., 74 (2005), pp. 1653–1677.

[7] B. Cockburn and J. Gopalakrishnan, *Incompressible finite elements for the Stokes system via hybridization. Part* I: *The Stokes system in two space dimensions*, SIAM J. Numer. Anal., 43 (2005), pp. 1627–1650.

[8] V. Girault, *Incompressible finite element methods for Navier-Stokes equations with nonstandard boundary conditions in* $\mathbf{R}^3$, Math. Comp., 51 (1988), pp. 55–74.

[9] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer Ser. Comput. Math. 5, Springer-Verlag, New York, 1986.

[10] J. Gopalakrishnan, L. E. García-Castillo, and L. F. Demkowicz, *Nédélec spaces in affine coordinates*, Comput. Math. Appl., 49 (2005), pp. 1285–1294.

[11] M. D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows, A Guide to Theory, Practice, and Algorithms*, Computer Science and Scientific Computing, Academic Press, Boston, MA, 1989.

[12] M. D. Gunzburger, R. A. Nicolaides, and C. H. Liu, *Algorithmic and theoretical results on computation of incompressible viscous flows by finite element methods*, Comput. & Fluids, 13 (1985), pp. 361–373.

[13] J.-C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[14] J.-C. Nédélec, *Éléments finis mixtes incompressibles pour l'équation de Stokes dans* $\mathbf{R}^3$, Numer. Math., 39 (1982), pp. 97–112.

SIAM J. Numer. Anal.
Vol. 43, No. 4, pp. 1673–1692

FINITE VOLUME METHODS ON SPHERES AND SPHERICAL CENTROIDAL VORONOI MESHES*

QIANG DU† AND LILI JU‡

Abstract. We study in this paper a finite volume approximation of linear convection-diffusion equations defined on a sphere using the spherical Voronoi meshes, in particular the spherical centroidal Voronoi meshes. The high quality of spherical centroidal Voronoi meshes is illustrated through both theoretical analysis and computational experiments. In particular, we show that the L^2 error of the approximate solution is of quadratic order when the underlying mesh is given by a spherical centroidal Voronoi mesh. We also demonstrate numerically the high accuracy and the superconvergence of the approximate solutions.

Key words. finite volume method, spherical Voronoi tessellations, spherical centroidal Voronoi tessellations, error estimates, convection-diffusion equations

AMS subject classifications. 65N15, 65N50, 65D17

DOI. 10.1137/S0036142903425410

1. Introduction.

The numerical solution of partial differential equations defined on spheres is an active research subject in the scientific community. The subject is related to a number of important applications such as weather forecasting and climate modeling. For example, the numerical solution of linear convection-diffusion equations and nonlinear shallow water equations in spherical geometry can be used to test numerical algorithms for more complex atmospheric circulation models. Though these models were often solved with spectral methods or traditional finite difference methods in spherical coordinates, methods that use quasi-uniform tessellations of the sphere are gradually gaining popularity as the grid-based methods offer great potential when combined with massive parallelism and local adaptivity.

To get efficient and accurate numerical solutions of PDEs, it is well known that grid quality plays an important role and high quality grid generation is often a significant part of the overall solution process. In this regard, there were many recent studies on the approximations of PDEs defined on spheres using various spherical grids, such as grids based on Bucky-balls [19], icosahedral grids [2, 32, 33], skipped grids [22], grids from a gnomonic (cubed sphere) mapping [24], etc. In standard Euclidean geometry, the so-called Voronoi–Delaunay grids have always been very popular grids used in both finite element and finite volume methods [28]. Other spherical grids have also been studied; see, for example, [16].

In [8, 9], we proposed a high quality spherical grid based on the spherical centroidal Voronoi tessellation (SCVT), which can be used for both data assimilation purposes and for the numerical solution of PDEs on spheres. A very recent study made in [31] on both the global and the local uniformity of spherical grids indicated

*Received by the editors April 3, 2003; accepted for publication (in revised form) March 18, 2005; published electronically November 14, 2005. This work was supported in part by grants NSF-DMS 0409297 and NSF-ITR 0205232.

http://www.siam.org/journals/sinum/43-4/42541.html

†Department of Mathematics, Pennsylvania State University, University Park, PA 16802 (qdu@math.psu.edu).

‡Department of Mathematics, University of South Carolina, Columbia, SC 29208 (ju@math.sc.edu). Part of the work was completed while the author was an Industrial post-doc at the IMA, University of Minnesota.

that the SCVT grid with a uniform density measures better than many other variations, and, when used to discretize a model Poisson equation on the sphere, the SCVT-based grid tends to produces the smallest local truncation errors among all the grids under consideration. In [9], a finite volume approximation to a second order linear elliptic equation using the spherical Voronoi meshes was studied, and a first order error estimate for the discrete H^1 norm was obtained under some grid regularity assumptions. Preliminary numerical experiments demonstrated the good performance of the finite volume scheme when implemented with the spherical centroidal Voronoi meshes (SCVMs) that include both the SCVT and its dual (Delaunay) triangular grid. The SCVM enjoys some optimization properties [8] and they can also be defined with a nonuniform density function. They offer excellent local grid regularity and global mesh conformity as well as flexible mesh adaptivity. Thus, the SCVM naturally becomes an optimal grid in some sense, or at least a practically *safe* choice for discretizing PDEs on the sphere.

In this paper, we make further attempts to substantiate the optimality of SCVMs both theoretically and computationally. Our main results include a carefully designed finite volume scheme for a general second order convection-diffusion equation defined on a sphere. When implemented with the SCVM, we present a rigorous quadratic order L^2 error estimate for such a discrete scheme whose proof relies critically on the geometric properties of the SCVT. We further demonstrate through experiments the superconvergent properties of the numerical solutions and their gradients solved using our modified finite volume scheme and the SCVT-based grid. All these findings provide compelling reasons for regarding the SCVTs with the uniform density as arguably the best alternative for near uniform partitions of the sphere and the SCVT-based grids the optimal triangular grids to use for the numerical solution of many PDEs defined on spheres.

We point out that the conclusions given in this paper can be readily adapted to problems defined on the two-dimensional (2d) Euclidean plane. The analysis for the spherical case is somewhat more involved than the planar case since we must deal with the differences between spherical triangles and planar triangles.

The paper is organized as follows: we first introduce the model equation, along with some notation used in the paper. Then in section 2, we briefly recall the basic theory of the spherical centroidal Voronoi meshes. Some discrete function spaces and a finite volume scheme for linear convection-diffusion equations on the sphere given in [9] are discussed in section 3. With a suitable modification to the finite volume scheme, a rigorous L^2 error estimate is given in section 4 for SCVMs. In section 5, a superconvergent gradient recovery scheme is provided, and in section 6 we present some numerical experiments. Some concluding remarks are given in section 7.

We now introduce the model equation to be considered. First, let $\mathbb{S}^2$ denote the sphere (surface of the ball) having radius $r > 0$, i.e., $\mathbb{S}^2 = \big\{\mathbf{x} \in \mathbb{R}^3 \mid |\mathbf{x}| = r\big\}$. Let ∇_s denote the tangential gradient operator [13, 17] on $\mathbb{S}^2$ defined by

$$\nabla_s u(\mathbf{x}) = (\nabla_{s,1}, \nabla_{s,2}, \nabla_{s,3})u(\mathbf{x}) = \nabla u(\mathbf{x}) - (\nabla u(\mathbf{x}) \cdot \vec{\mathbf{n}}_{\mathbb{S}^2,\mathbf{x}})\vec{\mathbf{n}}_{\mathbb{S}^2,\mathbf{x}}\,,$$

where $\nabla = (D_1, D_2, D_3)$ denotes the general gradient operator in $\mathbb{R}^3$ and $\vec{\mathbf{n}}_{\mathbb{S}^2,\mathbf{x}}$ is the unit outer normal vector to $\mathbb{S}^2$ at $\mathbf{x} = (x_1, x_2, x_3)$. We consider the second order elliptic equation on the sphere given by

$$\nabla_s \cdot \big(-a(\mathbf{x})\nabla_s u(\mathbf{x}) + \vec{\mathbf{v}}(\mathbf{x})u(\mathbf{x})\big) + b(\mathbf{x})u(\mathbf{x}) = f(\mathbf{x}) \qquad \text{for } \mathbf{x} \in \mathbb{S}^2\,. \tag{1.1}$$

Note that since $\mathbb{S}^2$ has no boundary, there is no boundary condition imposed.

We use the standard notation $L^p(\mathbb{S}^2)$, $W^{m,p}(\mathbb{S}^2)$ for Sobolev spaces on $\mathbb{S}^2$ (viewed as a compact, 2d Riemannian manifold) [17], equipped with norms $\|\cdot\|_{L^p(\mathbb{S}^2)}$ and $\|\cdot\|_{W^{m,p}(\mathbb{S}^2)}$. We set $H^m(\mathbb{S}^2) = W^{m,2}(\mathbb{S}^2)$ and use the standard inner product $(u, v) = \int_{\mathbb{S}^2} u(\mathbf{x})v(\mathbf{x})\, ds(\mathbf{x})$ for $u, v \in L^2(\mathbb{S}^2)$.

Let the data in (1.1) satisfy the following assumptions.

Assumption 1. $f \in L^2(\mathbb{S}^2)$, $a \in C^1(\mathbb{S}^2)$, $b \in L^\infty(\mathbb{S}^2)$, and $\vec{\mathbf{v}} \in C^1(\mathbb{S}^2, \mathbb{R}^3)$ such that $a(\mathbf{x}) \geq \alpha_1 > 0$, $b(\mathbf{x}) \geq 0$, and $\nabla_s \cdot \vec{\mathbf{v}}(\mathbf{x}) + b(\mathbf{x}) \geq \alpha_2 > 0$, a.e.

For any $u, v \in H^1(\mathbb{S}^2)$, define the bilinear functional $\mathcal{A}$ such that

$$\begin{aligned}\mathcal{A}(u,v) = &\int_{\mathbb{S}^2} a(\mathbf{x})\big(\nabla_s u(\mathbf{x}) \cdot \nabla_s v(\mathbf{x})\big) + u(\mathbf{x})\big(\vec{\mathbf{v}}(\mathbf{x}) \cdot \nabla_s v(\mathbf{x})\big)\, ds(\mathbf{x}) \\ &+ \int_{\mathbb{S}^2} b(\mathbf{x})u(\mathbf{x})v(\mathbf{x})\, ds(\mathbf{x}).\end{aligned} \tag{1.2}$$

We easily see, for some constant $C > 0$, that

$$\mathcal{A}(u, v) \leq C\|u\|_{H^1(S^2)}\|v\|_{H^1(S^2)}.$$

The problem (1.1) has a unique weak solution $u \in H^2(\mathbb{S}^2)$ such that

$$\mathcal{A}(u, v) = (f, v), \quad \forall\, v \in H^1(\mathbb{S}^2) \tag{1.3}$$

and u satisfies the H^2 regularity estimate $\|u\|_{H^2(\mathbb{S}^2)} \leq C\|f\|_{L^2(\mathbb{S}^2)}$ for some constant $C > 0$. Though the same conclusion holds under weaker conditions on $\vec{\mathbf{v}}$ and b (and $\nabla_s \cdot \vec{\mathbf{v}}(\mathbf{x}) + b(\mathbf{x})$), for simplicity Assumption 1 is made throughout the paper.

2. Spherical centroidal Voronoi meshes. Let $d(\mathbf{x}, \mathbf{y})$ denote the geodesic distance between $\mathbf{x}$ and $\mathbf{y}$ on $\mathbb{S}^2$, i.e., $d(\mathbf{x}, \mathbf{y}) = r \arccos[(\mathbf{x} \cdot \mathbf{y})/r^2]$, where arccos denotes the inverse cosine. We also use $m(\cdot)$ to denote the standard measure (surface area or curve length) of the argument. Given a set of distinct points $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{S}^2$, the corresponding spherical Voronoi regions $\{V_i\}_{i=1}^n$ are defined by

$$V_i = \left\{\mathbf{x} \in \mathbb{S}^2 \mid d(\mathbf{x}_i, \mathbf{x}) < d(\mathbf{x}_j, \mathbf{x}) \text{ for } j = 1, \dots, n \text{ and } j \neq i\right\}, \quad 1 \leq i \leq n\,.$$

$\{V_i\}_{i=1}^n$ forms a *Voronoi tessellation* or *Voronoi diagram* of $\mathbb{S}^2$ associated with the generators $\{\mathbf{x}_i\}_1^n$. Each Voronoi cell V_i is an open convex spherical polygon on $\mathbb{S}^2$ with geodesic arcs making up its boundary. It is also well known that the dual tessellation (in a graph-theoretical sense) to a Voronoi tessellation of $\mathbb{S}^2$ consists of spherical triangles which form the *Delaunay triangulation.*

Given a density function ρ defined on $\mathbb{S}^2$, for any spherical region $V \subset \mathbb{S}^2$, the *constrained mass centroid* $\mathbf{x}^c$ of V on $\mathbb{S}^2$ is given by the solution of

$$\min_{\mathbf{x} \in V} F(\mathbf{x})\,, \qquad \text{where} \qquad F(\mathbf{x}) = \int_V \rho(\mathbf{y})|\mathbf{y} - \mathbf{x}|^2\, ds(\mathbf{y})\,. \tag{2.1}$$

As in [7, 8, 9], a Voronoi tessellation of $\mathbb{S}^2$ is called a *constrained centroidal Voronoi tessellation* (CCVT) of $\mathbb{S}^2$ or, specifically, SCVT if and only if the points $\{\mathbf{x}_i\}_{i=1}^m$ which serve as the generators of the associated spherical Voronoi tessellation $\{V_i\}_{i=1}^k$ are also the constrained mass centroids of those Voronoi regions. For any set of points $\{\widetilde{\mathbf{x}}_i\}_{i=1}^n$ on $\mathbb{S}^2$ and any spherical tessellation $\{\widetilde{V}_i\}_{i=1}^n$ of $\mathbb{S}^2$, the corresponding *energy*

$$\mathcal{K}\big(\{\widetilde{\mathbf{x}}_i, \widetilde{V}_i\}_{i=1}^n\big) = \sum_{i=1}^n \int_{\widetilde{V}_i} \rho(\mathbf{x})\|\mathbf{x} - \widetilde{\mathbf{x}}_i\|^2\, ds(\mathbf{x})$$

is minimized only if $\{\widetilde{\mathbf{x}}_i, \widetilde{V}_i\}_{i=1}^n$ are a SCVT [8]. Consequently, SCVMs have many good geometric properties [8, 9]. A constant density function ρ leads to *uniformly* distributed SCVTs, and a nonconstant density function provides systematically a nonuniform distribution of points while the accumulation of SCVT generators still remains locally regular.

Constructing a constrained mass centroid from (2.1) may be cumbersome. In [8], it has been shown that one can compute first the standard centroid $\mathbf{x}_i^*$ of V_i in $\mathbb{R}^3$, then compute $\mathbf{x}_i^c$ using the fact that it is the projection of $\mathbf{x}_i^*$ onto $\mathbb{S}^2$ along the normal direction at $\mathbf{x}_i^c$. We refer to [8, 9, 21] for both deterministic and probabilistic algorithms for the construction of SCVTs. Figure 2.1 shows some examples of SCVTs associated with a constant density. More examples, including SCVTs with nonuniform densities, can be found in [9].

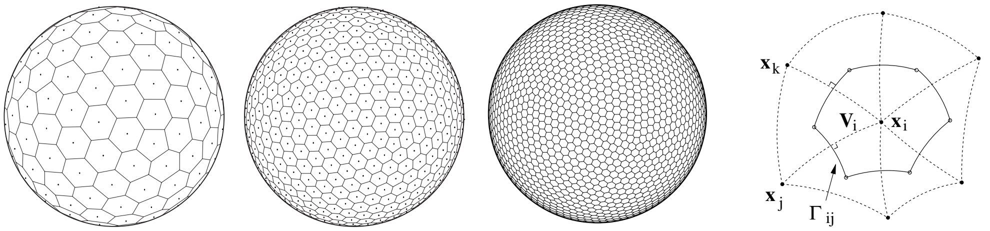


FIG. 2.1. *SCVTs for a constant density function with* 162, 642, *and* 2562 *generators, and an illustration of a spherical Voronoi region and its dual triangles.*

Given a spherical Voronoi mesh $\mathcal{W} = \{\mathbf{x}_i, V_i\}_{i=1}^n$, following [9], we refer to a pair of generators $\mathbf{x}_i$ and $\mathbf{x}_j$ as *neighbors* if and only if $\Gamma_{i,j} = \overline{V}_i \cap \overline{V}_j \neq \emptyset$. For Voronoi meshes, $\Gamma_{i,j}$ can only be a point or a geodesic arc on the sphere. For each $\mathbf{x}_i$, let χ_i be the set of the indices of its neighbors $\mathbf{x}_j$'s such that $m(\Gamma_{i,j}) > 0$. Let $\overline{\mathbf{x}_i\mathbf{x}_j}$ be the vector from $\mathbf{x}_i$ to $\mathbf{x}_j$, and let $\widetilde{\mathbf{x}_i\mathbf{x}_j}$ be the geodesic arc joining $\mathbf{x}_i$ and $\mathbf{x}_j$. From the construction of spherical Voronoi tessellations, it is known that $\widetilde{\mathbf{x}_i\mathbf{x}_j}$ is perpendicular to Γ_{ij} and the plane determined by $\Gamma_{i,j}$ and the origin bisects $\widetilde{\mathbf{x}_i\mathbf{x}_j}$ at its midpoint $\mathbf{x}_{ij}$ [9]; see Figure 2.1. Thus, $|\mathbf{x}_i - \mathbf{x}| = |\mathbf{x}_j - \mathbf{x}|$ for $\mathbf{x} \in \Gamma_{ij}$ and for $k = i, j$, $\vec{\mathbf{n}}_{\mathbf{x},V_k}$ is parallel to $\overline{\mathbf{x}_i\mathbf{x}_j}$ where $\vec{\mathbf{n}}_{\mathbf{x},V_k}$ is the outer unit normal vector to the boundary of V_k, taken to lie in the tangent plane of $\mathbb{S}^2$ at $\mathbf{x}$.

Let $h_i = \max_{\mathbf{y}\in V_i} d(\mathbf{x}_i, \mathbf{y})$, and we define the *mesh quality norm* by $h = \max_i h_i$. h gives the maximum geodesic distance between any particular generator $\mathbf{x}_i$ and the points in its associated cell V_i, and it has been used in [8] for the polynomial interpolation on the sphere.

Given a Voronoi mesh $\mathcal{W} = \{\mathbf{x}_i, V_i\}_{i=1}^m$, we define the *mesh regularity norm* σ by

$$\sigma = \min_{1\le i\le n} \sigma_i, \quad \text{where} \quad \sigma_i = \min_{j\in\chi_i} \sigma_{ij} \quad \text{and} \quad \sigma_{ij} = d(\mathbf{x}_i, \mathbf{x}_j)/(2h_i)\,. \tag{2.2}$$

If $\mathbf{x}_i, \mathbf{x}_j$, and $\mathbf{x}_k$ are neighbors for each other in $\mathcal{W}$, we denote by $\tilde{T}_{ijk}$ the spherical triangle determined by $\mathbf{x}_i, \mathbf{x}_j$, and $\mathbf{x}_k$, and by T_{ijk} the corresponding planar triangle (see Figure 2.1). In addition, let $\tilde{\mathcal{T}} = \{\tilde{T}_{ijk} \,|\, ijk \in \Sigma\}$, $\mathcal{T} = \{T_{ijk} \,|\, ijk \in \Sigma\}$ where $\Sigma = \{\, ijk \,|\, i, j, k \text{ are neighbors in } \mathcal{W}\}$. $\tilde{\mathcal{T}}$ gives the spherical Delaunay triangulation of $\mathbb{S}^2$ associated with the generators $\{\mathbf{x}_i\}_{i=1}^n$.

Meshes of the type $\mathcal{W} = \{\mathbf{x}_i, V_i\}_{i=1}^n$ are used as control (or finite) volumes for the discretization method discussed below. Our use of these meshes is particularly motivated by the covolume mesh approaches in the numerical solution of PDEs [28, 29] and for applications to nonlinear problems [6, 12, 30].

3. A finite volume method on spherical Voronoi tessellations. Without loss of generality, we consider the case of the unit sphere S^2 in what follows.

3.1. Some definitions and geometric properties. For any $\mathbf{x}$, let $\mathcal{P}(\mathbf{x}) = \mathbf{x}/|\mathbf{x}|$ be the projection onto the unit sphere S^2. $\mathcal{P}$ is also a one-to-one smooth function that maps $\mathbf{S}^* = \cup_{T_{ijk}\in\mathcal{T}} T_{ijk}$ to $\mathbb{S}^2 = \cup_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \tilde{T}_{ijk}$.

Clearly, $\mathbf{S}^* \cap \mathbb{S}^2 = \{\mathbf{x}_i\}_{i=1}^n$. Moreover, for any $\mathbf{x} \in \mathbb{S}^2$ and $\mathbf{x}', \mathbf{x}'' \in \tilde{T}_{ijk}$, we have

$$\begin{cases} |\mathbf{x} - \mathcal{P}^{-1}(\mathbf{x})| \le ch^2, \\ (1-ch^2)d(\mathbf{x}',\mathbf{x}'') \le |\mathcal{P}^{-1}(\mathbf{x}') - \mathcal{P}^{-1}(\mathbf{x}'')| \le (1+ch^2)d(\mathbf{x}',\mathbf{x}''), \\ m(T_{ijk}) \le m(\tilde{T}_{ijk}) \le (1+ch^2)m(T_{ijk}), \end{cases} \tag{3.1}$$

where c is a generic constant for h small.

Let $\Omega = \{\mathbf{x} \,|\, 1-ch^2 < |\mathbf{x}| < 1+ch^2\}$; then $\cup T_{ijk} \subset \Omega$ and $\cup\tilde{T}_{ijk} \subset \Omega$. For any $u \in H^2(\mathbb{S}^2)$, define the function Eu, the extension of u in Ω, by $Eu(\mathbf{y}) = u(\mathbf{y}/|\mathbf{y}|)$ for any $\mathbf{y} \in \Omega$. The following results have been shown in [9].

Proposition 1. *For any $\mathbf{y} \in \Omega$ and $\mathbf{x} = \mathbf{y}/|\mathbf{y}| \in \mathbb{S}^2$, and $i, j = 1, 2, 3$,*

$$\begin{cases} \nabla_s u(\mathbf{x}) = \nabla Eu(\mathbf{x}), \quad \nabla(D_i Eu)(\mathbf{x}) = \nabla_s(\nabla_{s,i} u)(\mathbf{x}) - (\nabla_{s,i} u(\mathbf{x}))\vec{\mathbf{n}}_{\mathbb{S}^2,\mathbf{x}}, \\ |\mathbf{y}|\ \nabla Eu(\mathbf{y}) = \nabla Eu(\mathbf{x}), \quad |\mathbf{y}|^2 D_i D_j Eu(\mathbf{y}) = D_i D_j Eu(\mathbf{x}). \end{cases} \tag{3.2}$$

By (3.1) and Proposition 1, the following result can be obtained using a proof similar to that used in Lemma 1 in [13].

Proposition 2. *There exists a generic constant $c > 0$ such that for any $ijk \in \Sigma$,*

$$\begin{cases} C_1\|u\|_{L^2(\tilde{T}_{ijk})} \le \|Eu|_{\mathbf{S}^*}\|_{L^2(T_{ijk})} \le C_2\|u\|_{L^2(\tilde{T}_{ijk})}, \\ C_3\|u\|_{H^1(\tilde{T}_{ijk})} \le \|Eu|_{\mathbf{S}^*}\|_{H^1(T_{ijk})} \le C_4\|u\|_{H^1(\tilde{T}_{ijk})}, \\ \|Eu|_{\mathbf{S}^*}\|_{H^2(T_{ijk})} \le C_5\|u\|_{H^2(\tilde{T}_{ijk})}. \end{cases} \tag{3.3}$$

We call u^L a piecewise linear function on $\mathbf{S}^*$ if and only if

$$u^L(\mathbf{x}^*) = \lambda_i u^L(\mathbf{x}_i) + \lambda_j u^L(\mathbf{x}_j) + \lambda_k u^L(\mathbf{x}_k), \quad \forall\, \mathbf{x}^* \in T_{ijk},$$

where $\lambda_i, \lambda_j, \lambda_k$ are the barycentric coordinates of $\mathbf{x}^*$ in the planar triangle T_{ijk}.

Let $\mathcal{V}_{\mathcal{W}}$ be the space of piecewise constant functions associated with a spherical Voronoi mesh $\mathcal{W} = \{\mathbf{x}_i, V_i\}_{i=1}^n$,

$$\mathcal{V}_{\mathcal{W}} = \{u \mid u(\mathbf{x}) \ \text{ is constant on each cell } \ V_i\}, \tag{3.4}$$

and denote by $\mathcal{U}_{\mathcal{W}}$ the space of all functions u_h on $\mathbb{S}^2$ such that $u_h(\mathbf{x}) = u^L(\mathcal{P}^{-1}(\mathbf{x}))$ for $\mathbf{x} \in \mathbb{S}^2$, where u^L is a piecewise linear function on $\mathbf{S}^*$ with $\{u^L(\mathbf{x}_i) = u_h(\mathbf{x}_i)\}_{i=1}^n$, i.e., $Eu_h(\mathbf{x}^*) = u^L(\mathbf{x}^*)$ for any $\mathbf{x}^* \in \mathbf{S}^*$.

If we interpret the Sobolev space on $\mathbf{S}^*$ in the piecewise sense, then it is easy to get $u_h \in H^1(\mathbb{S}^2)$ for any $u_h \in \mathcal{U}_{\mathcal{W}}$ using (3.2) and the fact that $Eu_h = u^L \in H^1(\mathbf{S}^*)$.

We now state some standard estimates on $\mathcal{P}_{\mathcal{U}}(u)$ and $\mathcal{P}_{\mathcal{V}}(u)$ which are the interpolants on $\mathcal{U}_{\mathcal{W}}$ and $\mathcal{V}_{\mathcal{W}}$, respectively, of a function u defined on $\mathbb{S}^2$.

Proposition 3. *For any $u \in H^2(\mathbb{S}^2)$, there exists a generic constant $C > 0$ such that*

$$\begin{cases} \|u - \mathcal{P}_{\mathcal{U}}(u)\|_{L^2(\mathbb{S}^2)} + h\|u - \mathcal{P}_{\mathcal{U}}(u)\|_{H^1(\mathbb{S}^2)} \le Ch^2\|u\|_{H^2(\mathbb{S}^2)}, \\ \|u - \mathcal{P}_{\mathcal{V}}(u)\|_{L^2(\mathbb{S}^2)} \le Ch\|u\|_{H^2(\mathbb{S}^2)}. \end{cases} \tag{3.5}$$

Proof. Note that $\mathcal{P}_{\mathcal{U}}(u)(\mathbf{x}) = u^L(\mathcal{P}^{-1}(\mathbf{x}))$ with

$$u^L(\mathbf{x}^*) = \lambda_i u(\mathbf{x}_i) + \lambda_j u(\mathbf{x}_j) + \lambda_k u(\mathbf{x}_k) \quad \forall\, \mathbf{x}^* \in T_{ijk}.$$

Using the estimate for the linear interpolation on planar triangles and the relation

$$u(\mathbf{x}) - \mathcal{P}_{\mathcal{U}}(u)(\mathbf{x}) = \tilde{u}(\mathcal{P}^{-1}(\mathbf{x})) - u^L(\mathcal{P}^{-1}(\mathbf{x})),$$

where $\tilde{u} = Eu|_{\mathbf{S}^*}$, we obtain by (3.1) and Proposition 2 that

$$\begin{aligned}
\|u - \mathcal{P}_{\mathcal{U}}(u)\|_{L^2(\mathbb{S}^2)} &= \left(\sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} |u(\mathbf{x}) - \mathcal{P}_{\mathcal{U}}(u)(\mathbf{x})|^2 \, ds(\mathbf{x}) \right)^{1/2} \\
&= \left(\sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} |\tilde{u}(\mathcal{P}^{-1}(\mathbf{x})) - u^L(\mathcal{P}^{-1}(\mathbf{x}))|^2 \, ds(\mathbf{x}) \right)^{1/2} \\
&\le C \left(\sum_{T_{ijk} \in \mathcal{T}} \int_{T_{ijk}} |\tilde{u}(\mathbf{x}^*) - u^L(\mathbf{x}^*)|^2 \, ds(\mathbf{x}^*) \right)^{1/2} \qquad (3.6) \\
&\le Ch^2 \left(\sum_{T_{ijk} \in \mathcal{T}} \|\tilde{u}\|^2_{H^2(T_{ijk})} \right)^{1/2} \\
&\le Ch^2 \left(\sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \|u\|^2_{H^2(\tilde{T}_{ijk})} \right)^{1/2} = Ch^2 \|u\|_{H^2(\mathbb{S}^2)}\,.
\end{aligned}$$

The other estimates can be proved in similar manners. We omit the details. □

For given functions $u, v \in \mathcal{V}_{\mathcal{W}}$, or $\mathcal{U}_{\mathcal{W}}$, we define, similar to [27], the discrete inner products and norms associated with a spherical Voronoi mesh $\mathcal{W} = \{\mathbf{x}_i, V_i\}_{i=1}^n$ by the following:

$$\begin{cases}
(u,v)_{\mathcal{W}} = \displaystyle\sum_{i=1}^n m(V_i) u(\mathbf{x}_i) v(\mathbf{x}_i)\,, \quad \|u\|^2_{0,\mathcal{W}} = (u,u)_{\mathcal{W}}, \\
|u|^2_{1,\mathcal{W}} = \displaystyle\frac{1}{2} \sum_{i=1}^n \sum_{j \in \chi_i} m(\Gamma_{ij}) d(\mathbf{x}_i, \mathbf{x}_j) \left(\frac{u(\mathbf{x}_i) - u(\mathbf{x}_j)}{|\mathbf{x}_i - \mathbf{x}_j|} \right)^2, \\
\|u\|^2_{1,\mathcal{W}} = \|u\|^2_{0,\mathcal{W}} + |u|^2_{1,\mathcal{W}}.
\end{cases}$$

Norms for general function spaces can also be defined.

We conclude with some norm equivalence results under mesh regularity assumptions. For convenience, we assume that all three angles of T_{ijk} are less than 90°. This is generally valid for the triangles in the SCVMs with sufficiently large (no smaller than 42, for example, for the constant density) number of vertices (generators) or, equivalently, sufficient small h. Using (3.1), Propositions 1 and 2, and similar arguments as those in Proposition 1 of section 2.1 in [26], we have the following.

PROPOSITION 4. *For any $u_h \in \mathcal{U}_{\mathcal{W}}$, there exist some constants $\{C_i > 0\}_{i=1}^4$,*

$$\begin{cases}
C_1 \|u_h\|_{0,\mathcal{W}} \le \|u_h\|_{L^2(\mathbb{S}^2)} \le C_2 \|u_h\|_{0,\mathcal{W}}, \\
C_3 \|u_h\|_{1,\mathcal{W}} \le \|u_h\|_{H^1(\mathbb{S}^2)} \le C_4 \|u_h\|_{1,\mathcal{W}}.
\end{cases} \qquad (3.7)$$

The results of Proposition 4 are in fact valid for more general Voronoi–Delaunay meshes that satisfy the local mesh regular properties. Let $l(\tilde{T}_{ijk})$ be the maximum number of spherical Voronoi regions V_m having nonempty intersection $V_m \cap \tilde{T}_{ijk}$ for any spherical triangle $\tilde{T}_{ijk}$, and let $l(V_m)$ be the maximum number of spherical triangles $\tilde{T}_{ijk}$ needed to cover any spherical Voronoi region V_m; we need all the $\{l(\tilde{T}_{ijk})\}$ and $\{l(V_m)\}$ to be bounded above by a constant integer independent of h. Under those conditions and the mesh regularity conditions, the above equivalence of norms still holds.

3.2. A finite volume discretization scheme. Based on Green's formula, a finite volume method for (1.1) was proposed in [9]. Set $\{u_i^h = u^h(\mathbf{x}_i)\}_{i=1}^n$ and let the approximate flux $\mathcal{F}_{ij}$ be defined by

$$\mathcal{F}_{ij} = -m(\Gamma_{ij}) a_{ij} \frac{u_j^h - u_i^h}{|\mathbf{x}_j - \mathbf{x}_i|} \approx \int_{\Gamma_{ij}} (-a(\mathbf{x}) \nabla_s u(\mathbf{x})) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i} \, d\gamma(\mathbf{x}) \,, \tag{3.8}$$

where $a_{ij} m(\Gamma_{ij}) = \int_{\Gamma_{ij}} a(\mathbf{x}) \, d\gamma(\mathbf{x})$. An up-wind approximate convection flux $\mathcal{V}_{i,j}$ was defined in [18] by

$$\mathcal{V}_{ij} = \beta_{ij}^+ u_i^h + \beta_{ij}^- u_j^h \approx \int_{\Gamma_{ij}} (\vec{\mathbf{v}}(\mathbf{x}) u(\mathbf{x})) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i} \, d\gamma(\mathbf{x}) \,, \tag{3.9}$$

where $\beta_{ij}^+ = (\beta_{ij} + |\beta_{ij}|)/2$, $\beta_{ij}^- = (\beta_{ij} - |\beta_{ij}|)/2$, and $\beta_{ij} = \int_{\Gamma_{ij}} \vec{\mathbf{v}}(\mathbf{x}) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i} \, d\gamma(\mathbf{x})$.

For all V_i, let f_i and b_i denote, respectively, the mean value of f and b on V_i; i.e.,

$$f_i = \frac{1}{m(V_i)} \int_{V_i} f(\mathbf{x}) \, ds(\mathbf{x}) \qquad \text{and} \qquad b_i = \frac{1}{m(V_i)} \int_{V_i} b(\mathbf{x}) \, ds(\mathbf{x}) \,. \tag{3.10}$$

The finite volume scheme given in [9] is defined as follows: find $u^h \in \mathcal{V}_{\mathcal{W}}$ such that

$$(\mathcal{L}^h u^h)_i = \frac{1}{m(V_i)} \sum_{j \in \chi_i} \left(\mathcal{F}_{ij} + \mathcal{V}_{ij} \right) + b_i u_i^h = f_i \qquad \text{for} \quad i = 1, \dots, n \,. \tag{3.11}$$

Since $\mathcal{F}_{ij} = -\mathcal{F}_{ji}$ and $\mathcal{V}_{ij} = -\mathcal{V}_{ji}$ for neighboring $\mathbf{x}_i$ and $\mathbf{x}_j$ with $m(\Gamma_{ij}) > 0$, the above scheme satisfies the discrete conservation law

$$\sum_{i=1}^n \sum_{j \in \chi_i} (\mathcal{F}_{ij} + \mathcal{V}_{ij}) = 0 \,.$$

Note that an approximate convection flux of the form $\mathcal{V}_{i,j} = (u_i^h + u_j^h)\beta_{ij}/2$ leads to a central difference scheme. A stability condition such as

$$P_i = \max_{j \in \chi_i} \frac{|\beta_{ij}| \cdot |\mathbf{x}_i - \mathbf{x}_j|}{2m(\Gamma_{ij}) a_{ij}} \le 1 \quad \text{for} \quad i = 1, \dots, n$$

is needed in such a case. P_i is called the local Peclet number [18, 27].

3.3. Previous results and a modified scheme. Assuming that $\mathcal{W}$ is *regular* in the sense that σ is not *too small*, i.e., it remains bounded from below as $h \to 0$, then the following result has been proved in [9].

THEOREM 1. *Let Assumption* 1 *be satisfied and the mesh be regular, and let* $\mathcal{F}_{ij}$*,* $\mathcal{V}_{ij}$*,* f_i*, and* b_i *be defined by* (3.8)–(3.10)*. Then the discrete system* (3.11) *has a unique* *solution $u^h \in \mathcal{V}_\mathcal{W}$. Furthermore, assume that the unique solution u of* (1.1) *belongs to $H^2(\mathbb{S}^2)$; then there exists a constant $C > 0$ only depending on a, $\vec{\mathbf{v}}$, b, and σ such that*

$$\|e^h\|_{1,\mathcal{W}} \le Ch\|u\|_{H^2(\mathbb{S}^2)}\ , \tag{3.12}$$

where $e^h = \{e_i^h = u(\mathbf{x}_i) - u_i^h\}$.

Note that Theorem 1 holds for general regular spherical Voronoi meshes. For more existing studies on the finite volume methods, especially when applied to solve second order elliptic on the 2d plane, we refer to [1, 3, 4, 5, 12, 14, 15, 20, 26, 25, 28, 29, 34, 35].

To get second order accuracy for the L^2 estimates, the order of approximation for the convection term used in the original scheme needs to be improved with better integration rules. For this purpose, let us define the bilinear functionals $\mathcal{A}^*$ and $\mathcal{A}_\mathcal{W}$ such that

$$\mathcal{A}^*(u, v^h) = \sum_{i=1}^n v^h(\mathbf{x}_i)\mathcal{A}^*(u, \psi_i)\ , \quad \mathcal{A}_\mathcal{W}(u, v^h) = \sum_{i=1}^n v^h(\mathbf{x}_i)\mathcal{A}_\mathcal{W}(u, \psi_i) \tag{3.13}$$

for any $u \in H^2(\mathbb{S}^2) \cup \mathcal{U}_\mathcal{W}$ and $v^h \in \mathcal{V}_\mathcal{W}$, where

$$\mathcal{A}^*(u, \psi_i) = \int_{\partial V_i} (-\nabla_s u(\mathbf{x}) + \vec{\mathbf{v}}(\mathbf{x})u(\mathbf{x})) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i}\ d\gamma(\mathbf{x}) + \int_{V_i} b(\mathbf{x})P_\mathcal{V}(u)(\mathbf{x})\ ds(\mathbf{x})\ ,$$

$$\mathcal{A}_\mathcal{W}(u, \psi_i) = \sum_{j\in\chi_i} \mathcal{F}_{ij}(u) + \int_{\partial V_i} \mathcal{P}_\mathcal{U}(u)(\mathbf{x})(\vec{\mathbf{v}}(\mathbf{x}) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i})\ d\gamma(\mathbf{x}) + m(V_i)b_i u(\mathbf{x}_i)\ .$$

Comparing $\mathcal{A}_\mathcal{W}$ with the finite volume scheme (3.11), we have in fact replaced only the convection term $\mathcal{V}_{ij}$ by $\int_{\partial V_i} \mathcal{P}_\mathcal{U}(u)(\mathbf{x})(\vec{\mathbf{v}}(\mathbf{x})\cdot\vec{\mathbf{n}}_{\mathbf{x},V_i})\ d\gamma(\mathbf{x})$. Note that no change is made for a pure diffusion problem containing only the second order terms $\nabla_s \cdot (a(\mathbf{x})\nabla_s u(\mathbf{x}))$. Our discrete problem here is then as follows: find $u_h \in \mathcal{U}_\mathcal{W}$ such that

$$\mathcal{A}_\mathcal{W}(u_h, v^h) = (f, v^h) \quad \forall\ v^h \in \mathcal{V}_\mathcal{W}\ , \tag{3.14}$$

i.e.,

$$\mathcal{A}_\mathcal{W}(u_h, \psi_i) = f_i \qquad \text{for} \quad i = 1, 2, \dots, n.$$

Formulations like the above for finite volume methods have been used, for instance, in [26]. Combining Proposition 3 and Proposition 4, it can be shown that the error estimate of Theorem 1 still holds for the above u_h using analysis similar to that used in [9].

THEOREM 2. *Suppose that Assumption* 1 *is satisfied. Let $\mathcal{F}_{ij}$ be defined by* (3.8). *Then the discrete system* (3.14) *has a unique solution $u_h \in \mathcal{U}_\mathcal{W}$. Furthermore, assume that the unique solution u of* (1.1) *belongs to $H^2(\mathbb{S}^2)$; then there exists a constant $C > 0$ only depending on a, $\vec{\mathbf{v}}$, b, and σ such that for $e_h = u - u_h$, we have*

$$\|e_h\|_{H^1(\mathbb{S}^2)} \le Ch\|u\|_{H^2(\mathbb{S}^2)}\ . \tag{3.15}$$

Proof. Notice that $\|\mathcal{P}_\mathcal{U}(u) - u_h\|_{1,\mathcal{W}} = \|e_h\|_{1,\mathcal{W}}$; then we have

$$\begin{aligned}\|e_h\|_{H^1(\mathbb{S}^2)} &= \|u - u_h\|_{H^1(\mathbb{S}^2)}\\ &\le \|u - \mathcal{P}_\mathcal{U}(u)\|_{H^1(\mathbb{S}^2)} + \|\mathcal{P}_\mathcal{U}(u) - u_h\|_{H^1(\mathbb{S}^2)}\\ &\le C_1 h\|u\|_{H^2(\mathbb{S}^2)} + C_2\|\mathcal{P}_\mathcal{U}(u) - u_h\|_{1,\mathcal{W}}\\ &= C_1 h\|u\|_{H^2(\mathbb{S}^2)} + C_2\|e_h\|_{1,\mathcal{W}} \le Ch\|u\|_{H^2(\mathbb{S}^2)},\end{aligned}$$

where the conclusion of Theorem 1 has been used. ☐

4. L^2 error estimate on SCVMs.

An improved error estimate in the L^2 norm is generally expected for our finite volume approximations of second order elliptic equations. However, it is shown here that the quadratic order error estimate can only be proved when the grid satisfies certain geometric constraints. In fact, a part of the estimate depends critically on the property that if $\mathcal{W} = \{\mathbf{x}_i, V_i\}_{i=1}^n$ is an SCVT of $\mathbb{S}^2$ corresponding to a density function ρ, then

$$\int_{V_i} \rho(\mathbf{x})(\mathbf{x}_i^* - \mathbf{x})\, ds(\mathbf{x}) = 0, \quad \forall\, i = 1, 2, \dots, n,$$

where $\mathbf{x}_i^*$ is the standard mass centroid of V_i, whose projection $\mathbf{x}_i^c$ (through the standard map $\mathcal{P}$) onto the sphere coincides with $\mathbf{x}_i$. We note that so far we have not been able to extend the elegant analysis of the covolume schemes for planar Poisson equations in [28, 29] to our context, nor we have found any improvement of the results there for the SCVT-based meshes. Thus, we resort to a more traditional approach of obtaining estimates through appropriate weak forms.

For the rest of the section, only those schemes based on SCVMs are analyzed.

4.1. A technical lemma.

For the interpolation operator $\mathcal{P}_{\mathcal{V}}$, we present a better approximation result that requires the properties of the SCVMs.

Lemma 1. *Suppose that $\mathcal{W} = \{\mathbf{x}_i, V_i\}_{i=1}^n$ is an SCVT of $\mathbb{S}^2$ with the density function ρ satisfying $\rho \in C^1(\mathbb{S}^2)$ and $\rho(\mathbf{x}) > 0$ for any $\mathbf{x} \in \mathbb{S}^2$. Then, for any $w \in H^2(\mathbb{S}^2)$, there exits a constant $C > 0$ such that*

$$\left| \int_{V_i} (w - \mathcal{P}_{\mathcal{V}}(w))\ ds(\mathbf{x}) \right| \le C h^2 m(V_i)^{1/2} \|w\|_{H^2(V_i)}, \quad i = 1, \dots, n. \tag{4.1}$$

Proof. Let us assume that $w \in C^2(\mathbb{S}^2)$; then it is easy to see that $Ew \in C^2(\Omega)$. Consider the spherical Voronoi region V_i associated with $\mathbf{x}_i$, for any $\mathbf{x} \in V_i$. We have

$$\begin{aligned} w(\mathbf{x}_i) - w(\mathbf{x}) &= Ew(\mathbf{x}_i) - Ew(\mathbf{x}) \\ &= \nabla Ew(\mathbf{x}) \cdot (\mathbf{x}_i - \mathbf{x}) + \int_0^1 H(Ew)(t\mathbf{x} + (1-t)\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}) \cdot (\mathbf{x}_i - \mathbf{x}) t dt, \end{aligned}$$

where $H(Ew)(\mathbf{x})$ denotes the Hessian matrix of Ew at $\mathbf{x}$. Thus

$$\left| \int_{V_i} w - \mathcal{P}_{\mathcal{V}}(w)\ ds(\mathbf{x}) \right| \le E_1 + E_2 + E_3,$$

where, with $\mathbf{x}_i^*$ being the mass centroid of V_i in R^3 with the density ρ, we have

$$\begin{aligned} E_1 &= \left| \int_{V_i} \nabla Ew(\mathbf{x}) \cdot (\mathbf{x}_i^* - \mathbf{x})\ ds(\mathbf{x}) \right| = \left| \int_{V_i} \nabla_s w(\mathbf{x}) \cdot (\mathbf{x}_i^* - \mathbf{x})\ ds(\mathbf{x}) \right|, \\ E_2 &= \left| \int_{V_i} \nabla Ew(\mathbf{x}) \cdot (\mathbf{x}_i - \mathbf{x}_i^*)\ ds(\mathbf{x}) \right| = \left| \int_{V_i} \nabla_s w(\mathbf{x}) \cdot (\mathbf{x}_i - \mathbf{x}_i^*)\ ds(\mathbf{x}) \right|, \\ E_3 &= \int_{V_i} \int_0^1 |H(Ew)(t\mathbf{x} + (1-t)\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}) \cdot (\mathbf{x}_i - \mathbf{x})| t dt\ ds(\mathbf{x}). \end{aligned}$$

Using the property of the SCVT that $\int_{V_i} \rho(\mathbf{x})(\mathbf{x}_i^* - \mathbf{x})\, ds(\mathbf{x}) = 0$, we have

$$\begin{aligned} E_1 &= \left| \int_{V_i} \nabla_s w(\mathbf{x}) \cdot (\mathbf{x}_i^* - \mathbf{x}) - \frac{\rho(\mathbf{x})}{\rho(\mathbf{x}_i)} \Pi_{\mathcal{V}}(\nabla_s w) \cdot (\mathbf{x}_i^* - \mathbf{x})\, ds(\mathbf{x}) \right| \\ &\leq \left| \int_{V_i} \frac{\rho(\mathbf{x}) - \rho(\mathbf{x}_i)}{\rho(\mathbf{x}_i)} \nabla_s w(\mathbf{x}) \cdot (\mathbf{x}_i^* - \mathbf{x})\, ds(\mathbf{x}) \right| \\ &\quad + \left| \int_{V_i} \frac{\rho(\mathbf{x})}{\rho(\mathbf{x}_i)} (\nabla_s w(\mathbf{x}) - \Pi_{\mathcal{V}}(\nabla_s w)) \cdot (\mathbf{x}_i^* - \mathbf{x})\, ds(\mathbf{x}) \right|, \end{aligned}$$

where $\Pi_{\mathcal{V}}$ denotes the L^2 projection on $\mathcal{V}_{\mathcal{W}}$. Denoting the two terms on the right-hand side of the last equation by E_4 and E_5, respectively, we have

$$\begin{aligned} E_4 &\leq h \int_{V_i} \frac{\max_{\mathbf{x}\in\Omega} |\nabla E\rho(\mathbf{x})|}{|\rho(\mathbf{x}_i)|} |\nabla_s w(\mathbf{x})|\, \|\mathbf{x}_i^* - \mathbf{x}\|\, ds(\mathbf{x}) \\ &\leq 2h^2 \int_{V_i} \frac{\max_{\mathbf{x}\in\Omega} |\nabla_s \rho(\mathbf{x})|}{|\rho(\mathbf{x}_i)|} |\nabla_s w(\mathbf{x})|\, ds(\mathbf{x}) \qquad (4.2) \\ &\leq Ch^2 \int_{V_i} |\nabla_s w(\mathbf{x})|\, ds(\mathbf{x}) \leq Ch^2 m(V_i)^{1/2} \|w\|_{H^2(V_i)} \end{aligned}$$

and

$$\begin{aligned} E_5 &\leq \int_{V_i} \frac{\max_{\mathbf{x}\in\Omega} |\rho(\mathbf{x})|}{|\rho(\mathbf{x}_i)|} |\nabla_s w(\mathbf{x}) - \Pi_{\mathcal{V}}(\nabla_s w)|\, \|\mathbf{x}_i^* - \mathbf{x}\|\, ds(\mathbf{x})| \\ &\leq C \|\nabla_s w - \Pi_{\mathcal{V}}(\nabla_s w)\|_{L^2(V_i)} \left(\int_{V_i} \|\mathbf{x}_i^* - \mathbf{x}\|^2\, ds(\mathbf{x}) \right)^{1/2} \qquad (4.3) \\ &\leq Ch^2 m(V_i)^{1/2} \|w\|_{H^2(V_i)}. \end{aligned}$$

Combining (4.2) and (4.3), we get

$$E_1 \leq Ch^2 m(V_i)^{1/2} \|w\|_{H^2(V_i)}\ . \qquad (4.4)$$

Consider E_2. Since $\mathbf{x}_i = \mathbf{x}_i^*/|\mathbf{x}_i^*|$, by (3.1) we know that $|\mathbf{x}_i - \mathbf{x}_i^*| < ch^2$. Thus

$$\begin{aligned} E_2 &\leq Ch^2 \left(\int_{V_i} |\nabla_s w(\mathbf{x})|^2\, ds(\mathbf{x}) \right)^{1/2} \left(\int_{V_i} ds(\mathbf{x}) \right)^{1/2} \qquad (4.5) \\ &\leq Ch^2 m(V_i)^{1/2} \|w\|_{H^1(V_i)}. \end{aligned}$$

On the other hand, for E_3, let $V_i^t = \{\mathbf{x}^* = t\mathbf{x} + (1-t)\mathbf{x}_i \mid \mathbf{x} \in V_i\}$. By changing variable $\mathbf{x}^* = t\mathbf{x} + (1-t)\mathbf{x}_i$ and using $ds(\mathbf{x}) \leq 2ds(\mathbf{x}^*)/t^2$, we get

$$E_3 \leq 2h^2 \int_0^1 \int_{V_i^t} (|H(Ew)|/t)\, ds(\mathbf{x}^*) dt\ .$$

Obviously, $m(V_i^t) \leq t^2 m(V_i)$. By a proof similar to that given in [9], we get

$$\begin{aligned} E_3 &\leq 2h^2 \int_0^1 \left(\int_{V_i^t} |H(Ew)|^2\, ds(\mathbf{x}^*) \right)^{1/2} m(V_i)^{1/2} dt \\ &\leq Ch^2 m(V_i)^{1/2} \|w\|_{H^2(V_i)}\ . \qquad (4.6) \end{aligned}$$

Finally, we obtain (4.1) for $u \in H^2(\mathbb{S}^2)$ by combining (4.4) and (4.5) with (4.6) and invoking a density argument. □

Note that in the planar case we have $E_2 = 0$ instead of (4.5).

4.2. Estimates for the weak forms. For simplicity, we assume $a(\mathbf{x}) = 1$. We first compare the bilinear forms (1.2) and (3.13). Let $\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}}$ be the unit outer normal at $\mathbf{x} \in \partial\tilde{T}_{ijk}$ of the boundary of $\tilde{T}_{ijk}$ that is tangent to $\mathbb{S}^2$. By Green's formula, for $\mathcal{W} \in H^2(\mathbb{S}^2)$ we have

$$\begin{aligned}
\mathcal{A}(u-u_h, \mathcal{P}_{\mathcal{U}}(w)) &= \int_{\mathbb{S}^2} \nabla_s(u-u_h)\cdot\nabla_s\mathcal{P}_{\mathcal{U}}(w)\; ds(\mathbf{x}) \\
&= \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \left(\int_{\tilde{T}_{ijk}} \nabla_s(u-u_h)\cdot\nabla_s\mathcal{P}_{\mathcal{U}}(w) + (u-u_h)(\vec{\mathbf{v}}\cdot\nabla_s\mathcal{P}_{\mathcal{U}}(w))\; ds(\mathbf{x}) \right) \\
&\qquad + \int_{\mathbb{S}^2} b(u-u_h)\mathcal{P}_{\mathcal{U}}(w)\; ds(\mathbf{x}) \\
&= \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \left(\int_{\tilde{T}_{ijk}} -\triangle_s(u-u_h)\mathcal{P}_{\mathcal{U}}(w) + (\nabla_s\cdot(u-u_h)\vec{\mathbf{v}})\mathcal{P}_{\mathcal{U}}(w)\; ds(\mathbf{x}) \right. \\
&\qquad \left. + \int_{\partial\tilde{T}_{ijk}} (\nabla_s(u-u_h)\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}})\mathcal{P}_{\mathcal{U}}(w) - (u-u_h)(\vec{\mathbf{v}}\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}})\mathcal{P}_{\mathcal{U}}(w)\; d\gamma(\mathbf{x}) \right) \\
&\qquad + \int_{\mathbb{S}^2} b(u-u_h)\mathcal{P}_{\mathcal{U}}(w)\; ds(\mathbf{x})\ ,
\end{aligned} \tag{4.7}$$

$$\begin{aligned}
\mathcal{A}^*(u-u_h, \mathcal{P}_{\mathcal{V}}(w)) &= \sum_{i=1}^{n} \mathcal{P}_{\mathcal{V}}(w)(\mathbf{x}_i)\mathcal{A}^*(u-u_h, \psi_i) \\
&= \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \left(\int_{\tilde{T}_{ijk}} -\triangle_s(u-u_h)\mathcal{P}_{\mathcal{V}}(w) + (\nabla_s\cdot(u-u_h)\vec{\mathbf{v}})\mathcal{P}_{\mathcal{V}}(w)\; ds(\mathbf{x}) \right. \\
&\qquad \left. + \int_{\partial\tilde{T}_{ijk}} (\nabla_s(u-u_h)\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}})\mathcal{P}_{\mathcal{V}}(w) - (u-u_h)(\vec{\mathbf{v}}\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}})\mathcal{P}_{\mathcal{V}}(w)\; d\gamma(\mathbf{x}) \right) \\
&\qquad + \int_{\mathbb{S}^2} bP_{\mathcal{V}}(u-u_h)\mathcal{P}_{\mathcal{V}}(w)\; ds(\mathbf{x}).
\end{aligned} \tag{4.8}$$

We now compare the first term of each functional.

LEMMA 2. *There is a constant $C > 0$ such that for $u \in H^3(\mathbb{S}^2)$ and $w \in H^2(\mathbb{S}^2)$,*

$$\left| \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} \triangle_s u(\mathcal{P}_{\mathcal{U}}(w) - \mathcal{P}_{\mathcal{V}}(w))\; ds(\mathbf{x}) \right| \le Ch^2 \|u\|_{H^3(\mathbb{S}^2)} \|w\|_{H^2(\mathbb{S}^2)}\ . \tag{4.9}$$

Proof. Letting E denote the left-hand side of (4.9), we have

$$\begin{aligned}
E &= \left| \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} \triangle_s u(\mathcal{P}_{\mathcal{U}}(w) - \mathcal{P}_{\mathcal{V}}(w))\; ds(\mathbf{x}) \right| \\
&\le \left| \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} \triangle_s u(\mathcal{P}_{\mathcal{U}}(w) - w)\; ds(\mathbf{x}) \right| + \left| \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} \triangle_s u(\mathcal{P}_{\mathcal{V}}(w) - w)\; ds(\mathbf{x}) \right| \\
&\le \left| \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} \triangle_s u(\mathcal{P}_{\mathcal{U}}(w) - w)\; ds(\mathbf{x}) \right| + \left| \sum_{i=1}^{n} \int_{V_i} \Pi_{\mathcal{V}}(\triangle_s u)(\mathcal{P}_{\mathcal{V}}(w) - w)\; ds(\mathbf{x}) \right| \\
&\qquad + \left| \sum_{i=1}^{n} \int_{V_i} (\triangle_s u(\mathbf{x}) - \Pi_{\mathcal{V}}(\triangle_s u))(\mathcal{P}_{\mathcal{V}}(w) - w)\; ds(\mathbf{x}) \right|,
\end{aligned} \tag{4.10}$$

where $\Pi_{\mathcal{V}}$ denotes the L^2 projection on $\mathcal{V}_{\mathcal{W}}$. Using Proposition 3 and the Cauchy–Schwarz inequality, we get

$$\left| \sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} \triangle_s u(\mathcal{P}_{\mathcal{U}}(w) - w) \, ds(\mathbf{x}) \right| \le Ch^2 \|u\|_{H^2(\mathbb{S}^2)} \|w\|_{H^2(\mathbb{S}^2)}.$$

Using Lemma 1, we have

$$\begin{aligned} \left| \sum_{i=1}^n \int_{V_i} \Pi_{\mathcal{V}}(\triangle_s u)(\mathcal{P}_{\mathcal{V}}(w) - w) \, ds(\mathbf{x}) \right| &\le Ch^2 \sum_{i=1}^n |\Pi_{\mathcal{V}}(\triangle_s u)|_{V_i}| \, m(V_i)^{1/2} \|w\|_{H^2(V_i)} \\ &= Ch^2 \|\Pi_{\mathcal{V}}(\triangle_s u)\|_{L^2(\mathbb{S}^2)} \|w\|^2_{H^2(\mathbb{S}^2)} \\ &\le Ch^2 \|u\|_{H^2(\mathbb{S}^2)} \|w\|_{H^2(\mathbb{S}^2)} \end{aligned}$$

$$\left| \sum_{i=1}^n \int_{V_i} (\triangle_s u(\mathbf{x}) - \Pi_{\mathcal{V}}(\triangle_s u))(\mathcal{P}_{\mathcal{V}}(w) - w) \, ds(\mathbf{x}) \right| \le Ch^2 \|u\|_{H^3(\mathbb{S}^2)} \|w\|_{H^2(\mathbb{S}^2)} \; .$$

Thus we obtain the estimate (4.9) in the lemma. □

Now we are ready to show the following.

LEMMA 3. *Let $u_h \in \mathcal{U}_{\mathcal{W}}$ be the unique solution of the discrete system* (3.14) *and assume that the unique variational solution u of* (1.1) *belongs to $H^3(\mathbb{S}^2)$. Then, for any $w \in H^2(\mathbb{S}^2)$, there exists a constant $C > 0$ such that*

$$|\mathcal{A}(u - u_h, \mathcal{P}_{\mathcal{U}}(w)) - \mathcal{A}^*(u - u_h, \mathcal{P}_{\mathcal{V}}(w))| \le Ch^2 \|u\|_{H^3(\mathbb{S}^2)} \|w\|_{H^2(\mathbb{S}^2)}. \tag{4.11}$$

Proof. By equation (4.8) we obtain

$$\mathcal{A}(u - u_h, \mathcal{P}_{\mathcal{U}}(w)) - \mathcal{A}^*(u - u_h, \mathcal{P}_{\mathcal{V}}(w)) = E_1 + E_2 + E_3 + E_4 + E_5 + E_6, \tag{4.12}$$

where

$$\begin{aligned} E_1 &= - \sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} \triangle_s u(\mathcal{P}_{\mathcal{U}}(w) - \mathcal{P}_{\mathcal{V}}(w)) \, ds(\mathbf{x}) \; , \\ E_2 &= \sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} \triangle_s u_h(\mathcal{P}_{\mathcal{U}}(w) - \mathcal{P}_{\mathcal{V}}(w)) \, ds(\mathbf{x}) \; , \\ E_3 &= \sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \int_{\partial \tilde{T}_{ijk}} (\nabla_s (u - u_h) \cdot \vec{\mathbf{n}}_{\mathbf{x}, \tilde{T}_{ijk}})(\mathcal{P}_{\mathcal{V}}(w) - \mathcal{P}_{\mathcal{U}}(w)) \, d\gamma(\mathbf{x}) \; , \\ E_4 &= \sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} (\nabla_s \cdot (u - u_h)\vec{\mathbf{v}})(\mathcal{P}_{\mathcal{U}}(w) - \mathcal{P}_{\mathcal{V}}(w)) \, ds(\mathbf{x}) \; , \\ E_5 &= \sum_{\tilde{T}_{ijk} \in \tilde{\mathcal{T}}} \int_{\partial \tilde{T}_{ijk}} (u - u_h)(\vec{\mathbf{v}} \cdot \vec{\mathbf{n}}_{\mathbf{x}, \tilde{T}_{ijk}})(\mathcal{P}_{\mathcal{V}}(w) - \mathcal{P}_{\mathcal{U}}(w)) \, d\gamma(\mathbf{x}) \; , \\ E_6 &= \int_{\mathbb{S}^2} b((u - u_h)\mathcal{P}_{\mathcal{U}}(w) - P_{\mathcal{V}}(u - u_h)\mathcal{P}_{\mathcal{V}}(w)) \, ds(\mathbf{x}) \; . \end{aligned}$$

Consider E_1. By Lemma 2 we have

$$|E_1| \le Ch^2 \|u\|_{H^3(\mathbb{S}^2)} \|w\|_{H^2(\mathbb{S}^2)}. \tag{4.13}$$

As for E_2, using Proposition 3 and Theorem 2 we have

$$\begin{aligned} |E_2| &\le \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} |\triangle_s u_h(\mathcal{P}_{\mathcal{U}}(w)-\mathcal{P}_{\mathcal{V}}(w))|\, ds(\mathbf{x}) \\ &\le Ch \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} |\nabla_s u_h|\, |(\mathcal{P}_{\mathcal{U}}(w)-\mathcal{P}_{\mathcal{V}}(w))|\, ds(\mathbf{x}) \qquad (4.14) \\ &\le Ch^2\|u_h\|_{H^1(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)} \le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)}. \end{aligned}$$

According to the continuity of $\nabla_s u$ on each $\partial\tilde{T}_{ijk}$, we have

$$\sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\partial\tilde{T}_{ijk}} (\nabla_s u\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}})(\mathcal{P}_{\mathcal{V}}(w)-\mathcal{P}_{\mathcal{U}}(w))\, d\gamma(\mathbf{x}) = 0,$$

and thus we get

$$E_3 = \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\partial\tilde{T}_{ijk}} (\nabla_s u_h\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}})(\mathcal{P}_{\mathcal{V}}(w)-\mathcal{P}_{\mathcal{U}}(w))\, d\gamma(\mathbf{x})\ . \qquad (4.15)$$

Additionally, on each edge $\tilde{L}$ of $\tilde{T}_{ijk}$, by symmetry with respect to the midpoint of $\tilde{L}$, we have that $\nabla_s u_h(z_1)\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}}$ is an even function for $\mathbf{x}\in\tilde{L}$ while $\mathcal{P}_{\mathcal{V}}(w)-\mathcal{P}_{\mathcal{U}}(w)$ is odd. Thus,

$$\int_{\tilde{L}} (\nabla_s u_h\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}})(\mathcal{P}_{\mathcal{V}}(w)-\mathcal{P}_{\mathcal{U}}(w))\, d\gamma(\mathbf{x}) = 0.$$

Thus we have

$$E_3 = 0. \qquad (4.16)$$

About E_4, we have by Theorem 2 that

$$\begin{aligned} |E_4| &\le \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} |(\nabla_s\cdot(u-u_h)\vec{\mathbf{v}})(\mathcal{P}_{\mathcal{U}}(w)-\mathcal{P}_{\mathcal{V}}(w))|\, ds(\mathbf{x}) \\ &\le \sup_{\mathbf{x}\in\mathbb{S}^2}(|\vec{\mathbf{v}}|+|\nabla_s\vec{\mathbf{v}}|) \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\tilde{T}_{ijk}} |\nabla_s(u-u_h)|\, |\mathcal{P}_{\mathcal{U}}(w)-\mathcal{P}_{\mathcal{V}}(w)|\, ds(\mathbf{x}) \qquad (4.17) \\ &\le Ch\|u-u_h\|_{H^1(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)} \le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)}. \end{aligned}$$

About E_5, using Trace theorem [17] and Theorem 2 we have

$$\begin{aligned} |E_5| &\le \sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\partial\tilde{T}_{ijk}} |(u-u_h)(\vec{\mathbf{v}}\cdot\vec{\mathbf{n}}_{\mathbf{x},\tilde{T}_{ijk}})(\mathcal{P}_{\mathcal{V}}(w)-\mathcal{P}_{\mathcal{U}}(w))|\, d\gamma(\mathbf{x}) \\ &\le \sup_{\mathbf{x}\in\mathbb{S}^2}(|\vec{\mathbf{v}}|)\left(\sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\partial\tilde{T}_{ijk}} |u-u_h|^2\, d\gamma(\mathbf{x})\right)^{1/2} \\ &\quad \cdot\left(\sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \int_{\partial\tilde{T}_{ijk}} |\mathcal{P}_{\mathcal{V}}(w)-\mathcal{P}_{\mathcal{U}}(w)|^2\, d\gamma(\mathbf{x})\right)^{1/2} \qquad (4.18) \\ &\le C\left(\sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} \|u-u_h\|^2_{H^1(\tilde{T}_{ijk})}\right)^{1/2}\left(\sum_{\tilde{T}_{ijk}\in\tilde{\mathcal{T}}} h^2\|w\|^2_{H^2(\tilde{T}_{ijk})}\right)^{1/2}. \\ &\le Ch\|u-u_h\|_{H^1(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)} \le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)}. \end{aligned}$$

About E_6, by Proposition 3 and Theorem 2 we have

$$
\begin{aligned}
|E_6| \le & \left| \int_{\mathbb{S}^2} b(u-u_h)(\mathcal{P}_{\mathcal{U}}(w) - \mathcal{P}_{\mathcal{V}}(w))\, ds(\mathbf{x}) \right| \\
& + \left| \int_{\mathbb{S}^2} b((u-u_h) - \mathcal{P}_{\mathcal{V}}(u-u_h))(\mathcal{P}_{\mathcal{V}}(w))\, ds(\mathbf{x}) \right| \\
\le & C\|u-u_h\|_{L^2(\mathbb{S}^2)}\, \|\mathcal{P}_{\mathcal{U}}(w) - \mathcal{P}_{\mathcal{V}}(w)\|_{L^2(\mathbb{S}^2)} \\
& + Ch\|u-u_h\|_{H^1(\mathbb{S}^2)}\, \|\mathcal{P}_{\mathcal{V}}(w)\|_{L^2(\mathbb{S}^2)} \\
\le & Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)}.
\end{aligned} \tag{4.19}
$$

Combining (4.12)–(4.14) and (4.16)–(4.19), we get (4.11). □

We see from the above proof that the extra regularity of $u \in H^3(\mathbb{S}^2)$ is only required for estimating the term E_1 (which is given in Lemma 2); the other terms merely require $u \in H^2(\mathbb{S}^2)$.

LEMMA 4. *Let $u_h \in \mathcal{U}_{\mathcal{W}}$ be the unique solution of the discrete system* (3.14) *and assume that the unique variational solution u of* (1.1) *belongs to $H^2(\mathbb{S}^2)$. Then, for any $w \in H^2(\mathbb{S}^2)$, there exists a constant $C > 0$ such that*

$$
|\mathcal{A}_{\mathcal{W}}(u_h, \mathcal{P}_{\mathcal{V}}(w)) - \mathcal{A}^*(u_h, \mathcal{P}_{\mathcal{V}}(w))| \le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)}\ . \tag{4.20}
$$

Proof. Since $\mathcal{P}_{\mathcal{V}}(u_h)|_{V_i} = u_h(\mathbf{x}_i)$ and $\mathcal{P}_{\mathcal{V}}(w)|_{V_i} = w(\mathbf{x}_i)$, we have

$$
\sum_{i=1}^n \int_{\partial V_i} (u_h - \mathcal{P}_{\mathcal{U}}(u_h))(\vec{\mathbf{v}} \cdot n_{\mathbf{x},V_i})\mathcal{P}_{\mathcal{V}}(w)\, d\gamma(\mathbf{x}) = 0,
$$

$$
\sum_{i=1}^n \left(\int_{V_i} b\mathcal{P}_{\mathcal{V}}(u_h)\mathcal{P}_{\mathcal{V}}(w)\, ds(\mathbf{x}) - m(V_i) b_i u_h(\mathbf{x}_i) w(\mathbf{x}_i) \right) = 0.
$$

Thus

$$
\begin{aligned}
& \mathcal{A}^*(u_h, \mathcal{P}_{\mathcal{V}}(w)) - \mathcal{A}_{\mathcal{W}}(u_h, \mathcal{P}_{\mathcal{V}}(w)) \\
& \quad = \sum_{i=1}^n \left(\int_{\partial V_i} (-\nabla_s u_h(\mathbf{x}) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i})\mathcal{P}_{\mathcal{V}}(w)\, d\gamma(\mathbf{x}) - \sum_{j\in\chi_i} \mathcal{F}_{ij}(u_h)\mathcal{P}_{\mathcal{V}}(w) \right) \\
& \quad = \sum_{i=1}^n \sum_{j\in\chi_i} m(\Gamma_{ij})\xi_{ij} w(x_i)\, ,
\end{aligned} \tag{4.21}
$$

with

$$
\begin{aligned}
\xi_{ij} &= -\frac{1}{m(\Gamma_{ij})} \int_{\partial\Gamma_{ij}} \nabla_s u_h(\mathbf{x}) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i}\, d\gamma(\mathbf{x}) + \frac{u_h(\mathbf{x}_i) - u_h(\mathbf{x}_j)}{|\mathbf{x}_i - \mathbf{x}_j|} \\
&= -\frac{1}{m(\Gamma_{ij})} \int_{\partial\Gamma_{ij}} \nabla_s u_h(\mathbf{x}) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i}\, d\gamma(\mathbf{x}) + \nabla E u_h(\mathbf{x}^*) \cdot \vec{\mathbf{n}}_{\mathbf{x},V_i}\, ,
\end{aligned}
$$

for any $\mathbf{x} \in \Gamma_{ij}$, $\mathbf{x}^* = \mathcal{P}^{-1}(\mathbf{x})$. Since $|\mathbf{x} - \mathbf{x}^*| \le Ch^2$, by Proposition 1

$$
|\nabla E u_h(\mathbf{x}^*) - \nabla_s u_h(\mathbf{x})| \le Ch^2 |\nabla E u_h(\mathbf{x}^*)|.
$$

Then we get

$$
\xi_{ij} \le \frac{Ch^2 |u_h(\mathbf{x}_i) - u_h(\mathbf{x}_j)|}{|\mathbf{x}_i - \mathbf{x}_j|}\ . \tag{4.22}
$$

It is also easy to find that

$$E=\sum_{i=1}^{n}\sum_{j\in\chi_i} m(\Gamma_{ij})\xi_{ij}w(\mathbf{x}_i)=-\frac{1}{2}\sum_{i=1}^{n}\sum_{j\in\chi_i} m(\Gamma_{ij})\xi_{ij}|\mathbf{x}_i-\mathbf{x}_j|\frac{w(\mathbf{x}_i)-w(\mathbf{x}_j)}{|\mathbf{x}_i-\mathbf{x}_j|}\ .$$

By Proposition 4 and Theorem 2 we have that

$$\begin{aligned}|E| &\le \sum_{i=1}^{n}\sum_{j\in\chi_i} m(\Gamma_{ij})\xi_{ij}d(\mathbf{x}_i,\mathbf{x}_j)\frac{|w(\mathbf{x}_i)-w(\mathbf{x}_j)|}{|\mathbf{x}_i-\mathbf{x}_j|}\\ &\le 2\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j\in\chi_i} m(\Gamma_{ij})d(\mathbf{x}_i,\mathbf{x}_j)\xi_{ij}^2\right)^{1/2}\\ &\quad\cdot\left(\frac{1}{2}\sum_{i=1}^{n}\sum_{j\in\chi_i} m(\Gamma_{ij})d(\mathbf{x}_i,\mathbf{x}_j)\left(\frac{w(\mathbf{x}_i)-w(\mathbf{x}_j)}{|\mathbf{x}_i-\mathbf{x}_j|}\right)^2\right)^{1/2}\\ &\le Ch^2|u_h|_{1,\mathcal{W}}|w|_{1,\mathcal{W}}\le Ch^2\|u_h\|_{H^1(\mathbb{S}^2)}\|\mathcal{P}_{\mathcal{U}}(w)\|_{H^1(\mathbb{S}^2)}\\ &\le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)}.\end{aligned}\tag{4.23}$$

Combining (4.21) and (4.23), we thus get (4.20). □

We note that the results of the above lemmas hold for more general $a=a(\mathbf{x})$ as well, but some slight modifications of the proofs are needed.

4.3. Main result.

We now present our main result on the L^2 error estimate.

THEOREM 3. *Let Assumption* 1 *be satisfied and additionally we assume that* $b\in H^1(\mathbb{S}^2)$. *Suppose that* $\mathcal{W}=\{\mathbf{x}_i,V_i\}_{i=1}^n$ *is an SCVM of* $\mathbb{S}^2$ *with the density function* ρ *satisfying* $\rho\in C^1(\mathbb{S}^2)$ *and* $\rho(\mathbf{x})>0$ *for any* $\mathbf{x}\in\mathbb{S}^2$. *Let* $\mathcal{F}_{ij}$ *be defined by* (3.8). *Then the discrete system* (3.14) *has a unique solution* $u_h\in\mathcal{U}_{\mathcal{W}}$. *Furthermore, assume that the unique solution* u *of* (1.1) *belongs to* $H^3(\mathbb{S}^2)$. *Then there exists a constant* $C>0$ *only depending on* ρ, a, $\vec{\mathbf{v}}$, b, *and* σ *such that*

$$\|e_h\|_{L^2(\mathbb{S}^2)}\le Ch^2\|u\|_{H^3(\mathbb{S}^2)},\tag{4.24}$$

where $e_h(\mathbf{x})=u(\mathbf{x})-u_h(\mathbf{x})$.

Proof. Since $u-u_h\in H^1(\mathbb{S}^2)$, according to (1.3) we know that there exists a weak solution $w\in H^2(\mathbb{S}^2)$ satisfying

$$\mathcal{A}(w,v)=(u-u_h,v)\quad\forall\, v\in H^1(\mathbb{S}^2).$$

Putting $v=u-u_h$ in the above equality, we get

$$||u-u_h||^2_{L^2(\mathbb{S}^2)}=(u-u_h,u-u_h)=\mathcal{A}(w,u-u_h).\tag{4.25}$$

Furthermore, from the H^2 regularity estimate, we have

$$||w||_{H^2(\mathbb{S}^2)}\le C||u-u_h||_{L^2(\mathbb{S}^2)}\tag{4.26}$$

for some constant $C>0$.

For the interpolants $\mathcal{P}_{\mathcal{U}}(w)$ and $\mathcal{P}_{\mathcal{V}}(w)$, we have

$$\mathcal{A}^*(u,\mathcal{P}_{\mathcal{V}}(w))+\int_{\mathbb{S}^2} b(u-P_{\mathcal{V}}(u))\mathcal{P}_{\mathcal{V}}(w)\,ds(\mathbf{x})=(f,\mathcal{P}_{\mathcal{V}}(w))$$

and

$$\mathcal{A}_{\mathcal{W}}(u_h, \mathcal{P}_{\mathcal{V}}(w)) = (f, \mathcal{P}_{\mathcal{V}}(w))\ .$$

Consequently, we get

$$\begin{aligned} \|u-u_h\|^2_{L^2(\mathbb{S}^2)} \le\ & |\mathcal{A}(u-u_h, w-\mathcal{P}_{\mathcal{U}}(w))| \\ & + |\mathcal{A}(u-u_h, \mathcal{P}_{\mathcal{U}}(w)) - \mathcal{A}^*(u-u_h, \mathcal{P}_{\mathcal{V}}(w))| \\ & + |\mathcal{A}^*(u_h, \mathcal{P}_{\mathcal{V}}(w)) - \mathcal{A}_{\mathcal{W}}(u_h, \mathcal{P}_{\mathcal{V}}(w))| \\ & + \left| \int_{\mathbb{S}^2} b(u - P_{\mathcal{V}}(u))\mathcal{P}_{\mathcal{V}}(w)\ ds(\mathbf{x}) \right|. \end{aligned} \tag{4.27}$$

According to Theorem 2, Proposition 3, and (4.26), we get

$$\begin{aligned} |\mathcal{A}(u-u_h, w-\mathcal{P}_{\mathcal{U}}(w))| &\le C\|u-u_h\|_{H^1(\mathbb{S}^2)}\ \|w-\mathcal{P}_{\mathcal{U}}(w)\|_{H^1(\mathbb{S}^2)} \\ &\le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\ \|w||_{H^2(\mathbb{S}^2)} \\ &\le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\ \|u-u_h\|_{L^2(\mathbb{S}^2)}. \end{aligned} \tag{4.28}$$

By Lemma 3 and (4.26), we get

$$\begin{aligned} |\mathcal{A}(u-u_h, \mathcal{P}_{\mathcal{U}}(w)) - \mathcal{A}^*(u-u_h, \mathcal{P}_{\mathcal{V}}(w))| &\le Ch^2\|u\|_{H^3(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)} \\ &\le Ch^2\|u\|_{H^3(\mathbb{S}^2)}\|u-u_h\|_{L^2(\mathbb{S}^2)}\ . \end{aligned} \tag{4.29}$$

Again, by Lemma 4 and (4.26), we have

$$\begin{aligned} |\mathcal{A}_{\mathcal{W}}(u_h, \mathcal{P}_{\mathcal{V}}(w)) - \mathcal{A}^*(u_h, \mathcal{P}_{\mathcal{V}}(w))| &\le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)} \\ &\le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|u-u_h\|_{L^2(\mathbb{S}^2)}\ . \end{aligned} \tag{4.30}$$

It is easy to see that

$$\begin{aligned} \left| \int_{\mathbb{S}^2} b(u-P_{\mathcal{V}}(u))\mathcal{P}_{\mathcal{V}}(w)\ ds(\mathbf{x}) \right| &= \left| \sum_{i=1}^n \int_{V_i} b(u-P_{\mathcal{V}}(u))\mathcal{P}_{\mathcal{V}}(w)\ ds(\mathbf{x}) \right| \\ &\le \left| \sum_{i=1}^n \int_{V_i} \Pi_{\mathcal{V}}(b)(u-P_{\mathcal{V}}(u))\mathcal{P}_{\mathcal{V}}(w)\ ds(\mathbf{x}) \right| \\ &\quad + \left| \sum_{i=1}^n \int_{V_i} (b-\Pi_{\mathcal{V}}(b))(u-P_{\mathcal{V}}(u))\mathcal{P}_{\mathcal{V}}(w)\ ds(\mathbf{x}) \right|. \end{aligned}$$

Using a proof similar to Lemmas 1 and 2, we can get

$$\begin{aligned} \left| \int_{\mathbb{S}^2} b(u-P_{\mathcal{V}}(u))\mathcal{P}_{\mathcal{V}}(w)\ ds(\mathbf{x}) \right| &\le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|w\|_{H^2(\mathbb{S}^2)} \\ &\le Ch^2\|u\|_{H^2(\mathbb{S}^2)}\|u-u^h\|_{L^2(\mathbb{S}^2)}. \end{aligned} \tag{4.31}$$

Combining (4.28), (4.29), and (4.30), we get

$$\|u-u_h\|^2_{L^2(\mathbb{S}^2)} \le Ch^2\|u\|_{H^3(\mathbb{S}^2)}\|u-u_h\|_{L^2(\mathbb{S}^2)},$$

which means

$$\|e_h\|_{L^2(\mathbb{S}^2)} = \|u-u_h\|_{L^2(\mathbb{S}^2)} \le Ch^2\|u\|_{H^3(\mathbb{S}^2)}.$$

We end the proof by noting that the case of general $a(\mathbf{x})$ can be verified similarly. □

Remark 1. Extra regularity on the exact solution is required to get the quadratic order error estimates, though in the standard finite element literature such a requirement is not needed in general. This is seen as a consequence of the quadrature approximations to the standard weak forms of the equations. From the proof of Lemma 3, it is easy to see that the regularity in $H^3(\mathbb{S}^2)$ can in fact be further weakened to, for instance, $W^{3,p}(\mathbb{S}^2)$ for $p > 1$.

Remark 2. The quadratic order error estimates depend critically on the properties of the SCVMs. The proof is not valid for a general spherical Voronoi mesh. Such an L^2 error estimate has not been given in the literature even for the planar finite volume methods based on the general Voronoi–Delaunay meshes. With other choices of the covolumes which are not of the Voronoi–Delaunay type, a quadratic order estimate has been proved in [26] for 2d diffusion equations. A first order L^2 error estimate has been given in [20].

5. Superconvergent gradient recovery. In this section, we discuss how to postprocess the finite volume solutions to obtain their tangential gradients in the longitude and latitude directions. Let us rewrite (1.1) in the spherical coordinate system (ϕ, θ) defined by $\mathbf{x} = (r \sin\phi \cos\theta, r \sin\phi \sin\theta, r \cos\phi)$ for $\phi \in [0, \pi]$ and $\theta \in [0, 2\pi)$. Ignoring the radial component, we may denote $\vec{\mathbf{v}}(\phi, \theta) = (v_1(\phi, \theta), v_2(\phi, \theta))$, where v_1 and v_2 are the (orthogonal) components of $\vec{\mathbf{v}}$ in the ϕ and θ directions, respectively, on the tangential surface of $\mathbb{S}^2$ at $\mathbf{x}$. We also have

$$\nabla_s u(\phi, \theta) = \left(\frac{1}{r} \frac{\partial u}{\partial \phi}, \ \frac{1}{r \sin\phi} \frac{\partial u}{\partial \theta} \right), \quad \nabla_s \cdot \vec{\mathbf{v}}(\phi, \theta) = \frac{1}{r \sin\phi} \left(\frac{\partial}{\partial \phi}(v_1 \sin\phi) + \frac{\partial v_2}{\partial \theta} \right).$$

For any generator $\mathbf{x}_i$ in $\mathcal{W}$, let $V_{\mathbf{x}_i} = \cup_{\mathbf{x}_i \in T_{ijk}} T_{ijk}$ and $\vec{\mathbf{n}}_{N,\mathbf{x}_i}$ be the unit vector on $S_{\mathbf{x}_i}$ along the ϕ direction. A map $H_{\mathbf{x}_i} : V_{\mathbf{x}_i} \to \mathbb{R}^2$ is defined by first projecting $V_{\mathbf{x}_i}$ onto the tangential plane $S_{\mathbf{x}_i}$ of $\mathbb{S}^2$ at $\mathbf{x}_i$ ($S_{\mathbf{x}_i} \perp \vec{\mathbf{n}}_{\mathbb{S}^2,\mathbf{x}_i}$ at $\mathbf{x}_i$) and then moving $S_{\mathbf{x}_i}$ to the (x, y)-plane via an affine map such that $\vec{\mathbf{n}}_{\mathbb{S}^2,\mathbf{x}_i}$ is mapped to the z-axis and $\vec{\mathbf{n}}_{N,\mathbf{x}_i}$ to the x-axis; see Figure 5.1. Let $V_i' = H_{\mathbf{x}_i}(V_{\mathbf{x}_i})$.

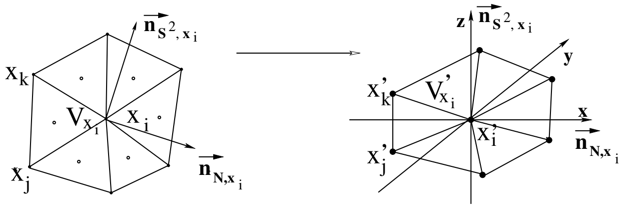

$\vec{\mathbf{n}}_{S^2,x_i}$
x_k
V_{x_i}
x_i
x_j
$\vec{\mathbf{n}}_{N,x_i}$
z
$\vec{\mathbf{n}}_{S^2,x_i}$
y
x_k'
V_{x_i}'
x
x_i'
$\vec{\mathbf{n}}_{N,x_i}$
x_j'

FIG. 5.1. *The mapping $H_{\mathbf{x}_i}$.*

On each planar triangle $\triangle \mathbf{x}_i' \mathbf{x}_j' \mathbf{x}_k' = H_{\mathbf{x}_i}(T_{ijk})$, we uniquely determine a linear function $\overline{u}_{T_{ijk}}$ by setting $\overline{u}_{T_{ijk}}(\mathbf{x}_i') = u_h(\mathbf{x}_i)$, $\overline{u}_{T_{ijk}}(\mathbf{x}_j') = u_h(\mathbf{x}_j)$, and $\overline{u}_{T_{ijk}}(\mathbf{x}_k') = u_h(\mathbf{x}_k)$. Now we define

$$\overline{\nabla}_s u_h(\mathbf{x}_i) = \frac{1}{q} \sum_{T_{ijk} \subset V_{\mathbf{x}_i}} \left(\frac{\partial \overline{u}_{T_{ijk}}}{\partial x}, \ \frac{\partial \overline{u}_{T_{ijk}}}{\partial y} \right), \tag{5.1}$$

where $q = \text{Card}(\{T_{ijk} \mid T_{ijk} \subset V_{\mathbf{x}_i}\})$. We also let

$$D(u, u_h) = \left(\sum_{i \in I} |\nabla_s u(\mathbf{x}_i) - \overline{\nabla}_s u_h(\mathbf{x}_i)|^2 m(V_i) \right)^{1/2}. \tag{5.2}$$

The index set I may be taken to be the set of all Voronoi generators or a large portion of the generator set. In light of the recent studies on the finite element gradient recovery [36] at mesh symmetric points, the close relationship between finite element and finite volume schemes [3, 34], and the nice properties of SCVMs, we expect that for the finite volume solution with SCVMs, there exists the estimate $D(u, u_h) = O(h^2)$. Such results are to be numerically investigated in the next section.

6. Numerical experiments.

Let $\mathbb{S}^2$ be the unit sphere. We now present numerical results that are summarized in the following two examples, with each example containing two separate experiments (corresponding to two different exact solutions) but with one identical exact solution. In our experiments, the finite volume meshes are taken to be the SCVMs corresponding to a constant density function with various different numbers of generators.

For our first example, we choose the exact solution to be

$$u_1(\phi, \theta) = \sin^2 \phi \cos^2 \theta \tag{6.1}$$

and study two different model problems whose data are given in Table 6.1.

Table 6.1

Data for model problems		$a(\phi,\theta)$	$v_1(\phi,\theta)$	$v_2(\phi,\theta)$	$b(\phi,\theta)$
I	no convection	1	0	0	1
II	convection dominated	0.05	$1+\sin\phi$	$1+\sin\theta$	$3.0+\sin^2\phi$

Approximate solutions were obtained using the finite volume scheme (3.11) with the central difference scheme and the uniformly distributed SCVM based on the constant density function $\rho = 1$ (as in Figure 2.1). In Table 6.2, errors in the approximate solution are listed against the number of generators.

Table 6.2

n		$\|u_1 - u_{1,h}\|_{L^2(\mathbb{S}^2)}$	$D(u_1, u_{1,h})$	$\|u_2 - u_{2,h}\|_{L^2(\mathbb{S}^2)}$	$D(u_2, u_{2,h})$
162	I	4.038E-02	1.331E-01	4.032E-01	4.313E-00
	II	6.406E-02	1.655E-01	5.962E-01	4.314E-00
642	I	1.021E-02	3.444E-02	1.370E-01	1.178E-00
	II	1.612E-02	4.214E-02	1.241E-01	1.115E-00
2562	I	2.556E-03	8.788E-03	2.687E-02	3.115E-01
	II	3.994E-03	1.161E-02	3.033E-02	2.908E-01
10242	I	6.362E-04	2.221E-03	7.445E-03	7.902E-02
	II	1.004E-03	3.072E-03	7.577E-03	7.346E-02
40962	I	1.631E-04	5.269E-04	2.080E-04	1.745E-02
	II	2.375E-04	8.132E-04	2.072E-04	1.797E-02

For the second example, the exact solution of (1.1) is chosen to be

$$u_2(\phi, \theta)) = \sin^2(2\phi) \cos(4\theta). \tag{6.2}$$

Errors in the approximate solution are again given in Table 6.2. As the exact solution (6.2) is more complex than (6.1), the largest 2% of the pointwise gradient errors $|\nabla_s u_2(\mathbf{x}_i) - \overline{\nabla}_s u_{2,h}(\mathbf{x}_i)|$ was removed from the estimate when computing the $D(u_2, u_{2,h})$. These relatively larger errors concentrate near the 12 defect points of the SCVM (i.e., those Voronoi cells with only 5 neighbors) where the mesh lacks perfect symmetry.

From the numerical values given in the tables we see that, for both the L^2 errors and the gradient recovery errors $D(u, u_h)$, the trend of quadratic order convergence is very evident as we refine the mesh.

7. Conclusion. High quality spherical grids have many applications. Many strategies have already been studied in atmospheric and geophysical simulations for producing good spherical grids [31]. Though many of these choices produce good quality grids, in general the recently proposed concept of SCVT [8, 9] yields grids superior to most of existing ones. Our study here on a finite volume approximation of linear convection diffusion equations based on the SCVM demonstrated further their optimality from both theoretical and computational standpoints.

Further studies can be carried out to explore the local energy equipartition property and hierarchical SCVMs for multiresolution analysis, to validate superconvergent gradient recovery through analytical means. The application of the SCVM to Ginzburg–Landau models has been studied recently [10, 11] and we expect to find many more applications to other complex physical problems in the future.

REFERENCES

[1] L. Baughman and N. Walkington, *Co-volume methods for degenerate parabolic problems*, Numer. Math., 64 (1993), pp. 45–67.

[2] J. R. Baumgardner and P. Frederickson, *Icosahedral discretization of the two-sphere*, SIAM J. Numer. Anal., 22 (1985), pp. 1107–1115.

[3] Z. Cai, *On the finite volume element method*, Numer. Math., 58 (1991), pp. 713–735.

[4] S. Chou and Q. Li, *Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach*, Math. Comp., 69 (2000), pp. 103–120.

[5] Y. Coudière, T. Gallouët, and R. Herbin, *Discrete Sobolev inequalities and L^p error estimates for finite volume solutions of convection diffusion equations*, Math. Model. Numer. Anal., 35 (2001), pp. 767–778.

[6] Q. Du, *Convergence analysis of a numerical method for a mean field model of superconducting vortices*, SIAM J. Numer. Anal., 37 (2000), pp. 911–926.

[7] Q. Du, V. Faber, and M. Gunzburger, *Centroidal Voronoi tessellations: Applications and algorithms*, SIAM Rev., 41 (1999), pp. 637–676.

[8] Q. Du, M. D. Gunzburger, and L. Ju, *Constrained centroidal Voronoi tessellations for surfaces*, SIAM J. Sci. Comput., 24 (2003), pp. 1488–1506.

[9] Q. Du, M. Gunzburger, and L. Ju, *Voronoi-based finite volume methods, optimal Voronoi meshes and PDEs on the sphere*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 3933–3957.

[10] Q. Du and L. Ju, *Numerical simulation of the quantized vortices on a thin superconducting hollow sphere*, J. Comput. Phys., 201 (2004), pp. 511–530.

[11] Q. Du and L. Ju, *Approximations of a Ginzburg-Landau model for superconducting hollow spheres based on spherical centroidal Voronoi tessellations*, Math. Comp., 74 (2005), pp. 1257–1281.

[12] Q. Du, R. A. Nicolaides, and X. Wu, *Analysis and convergence of a covolume approximation of the Ginzburg–Landau model of superconductivity*, SIAM J. Numer. Anal., 35 (1998), pp. 1049–1072.

[13] G. Dziuk, *Finite elements for the Beltrami operator on arbitrary surfaces*, in Partial Differential Equations and Calculus of Variations, Lecture Notes in Math. 1357, S. Hildebrandt and R. Leis, eds., Springer-Verlag, Berlin, 1988, pp. 142–155.

[14] R. Ewing, R. Lazarov, and P. Vassilevski, *Local refinement techniques for elliptic problems on cell-centered grids I. Error analysis*, Math. Comp., 56 (1991), pp. 437–461.

[15] T. Gallouët, R. Herbin, and M. H. Vignal, *Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions*, SIAM J. Numer. Anal., 37 (2000), pp. 1935–1972.

[16] D. Hardin and E. Saff, *Discretizing manifolds via minimum energy points*, Notices of Amer. Math. Soc., 51 (2004), pp. 1186–1194.

[17] E. Hebey, *Sobolev Spaces on Riemannian Manifolds*, Springer-Verlag, Berlin, 1991.

[18] J. Heinrich, P. Huyakorn, O. Zienkiewicz, and A. Mitchell, *An upwind finite element scheme for 2d convective transport equations*, Int. Num. Meth. Engrg., 11 (1977), pp. 131–143.

[19] T. Heinze and A. Hense, *The shallow water equations on the sphere and their Lagrange-Galerkin-Solution*, Meteorol. Atmos. Phys., 81 (2002), pp. 129–137.

[20] R. Herbin, *An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 165–173.

[21] L. Ju, Q. Du, and M. Gunzburger, *Probabilistic methods for centroidal Voronoi tessellations and their parallel implementations*, Parallel Comput., 28 (2002), pp. 1477–1500.

[22] A. Layton, *Cubic spline collocation method for the shallow water equations on the sphere*, J. Comput. Phys., 179 (2002), pp. 578–592.

[23] R. D. Lazarov, I. D. Mishev, and P. S. Vassilevski, *Finite volume methods for convection-diffusion problems*, SIAM J. Numer. Anal., 33 (1996), pp. 31–55.

[24] R. LeVeque and J. Rossmanith, *A wave propagation algorithm for the solution of PDEs on the surface of a sphere*, Internat. Ser. Numer. Math. Hyperbolic Problems, 141 (2001), pp. 643–652.

[25] R. Li, *Generalized difference methods for a nonlinear Dirichlet problem*, SIAM J. Numer. Anal., 24 (1987), pp. 77–88.

[26] R. Li, Z. Chen, and W. Wu, *Generalized difference methods for differential equations: Numerical analysis of finite volume methods*, Marcel Dekker, New York, 2000.

[27] I. Mishev, *Finite volume methods on Voronoi meshes*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 193–212.

[28] R. A. Nicolaides, *Direct discretization of planar div-curl problems*, SIAM J. Numer. Anal., 29 (1992), pp. 32–56.

[29] R. A. Nicolaides, *Analysis and convergence of the MAC scheme.* I: *The linear problem*, SIAM J. Numer. Anal., 29 (1992), pp. 1579–1591.

[30] R. Nicolaides and X. Wu, *Analysis and convergence of the MAC scheme.* II. *Navier–Stokes equations*, Math. Comp., 65 (1996), pp. 29–44.

[31] T. Ringler, *Comparing truncation error to PDE solution error on spherical Voronoi Tessellations*, Tech report, Dept. Atmospheric Sci., Colorado State Univ., 2003.

[32] G. Stuhne and W. Peltier, *New icosahedral grid-point discretizations of the shallow water equations on the sphere*, J. Comput. Phys., 148 (1999), pp. 23–58.

[33] H. Tomita, M. Tsugawa, M.Satoh, and K.Goto, *Shallow water model on a modified icosahedral geodesic grid by using spring dynamics*, J. Comput. Phys., 174 (2001), pp. 579–613.

[34] R. Vanselow, *Relations between FEM and FVM*, in Finite Volumes for Complex Applications: Problems and Perspectives, F. Benkhaldoun and R. Vilsmerier, eds., Hermes, Paris, 1996.

[35] P. S. Vassilevski, S. I. Petrova, and R. D. Lazarov, *Finite difference schemes on triangular cell-centered grids with local refinement*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1287–1313.

[36] Z. Zhang and R. Lin, *Ultraconvergence of ZZ patch recovery at mesh symmetry points*, Numerische Mathematik, 95 (2003), pp. 781–801.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1693–1714

ERROR ESTIMATION FOR REDUCED-ORDER MODELS OF DYNAMICAL SYSTEMS[*]

CHRIS HOMESCU[†], LINDA R. PETZOLD[†], AND RADU SERBAN[‡]

Abstract. The use of reduced-order models to describe a dynamical system is pervasive in science and engineering. Often these models are used without an estimate of their error or range of validity. In this paper we consider dynamical systems and reduced models built using proper orthogonal decomposition. We show how to compute estimates and bounds for these errors by a combination of small sample statistical condition estimation and error estimation using the adjoint method. Most importantly, the proposed approach allows the assessment of *regions of validity* for reduced models, i.e., ranges of perturbations in the original system over which the reduced model is still appropriate. Numerical examples validate our approach: the error norm estimates approximate well the forward error, while the derived bounds are within an order of magnitude.

Key words. model reduction, proper orthogonal decomposition, small sample statistical condition estimation, adjoint method

AMS subject classifications. 65L10, 65L99

DOI. 10.1137/040603541

1. Introduction.

Model reduction of dynamical systems described by differential equations is ubiquitous in science and engineering [2]. Reduced models are used for efficient simulation [17, 31] and control [18, 28]. Moreover, the process of creating low-order models forces the researcher to isolate and quantify the dominant physical mechanisms, revealing effective design decisions that would not have been identified through numerical simulation, experiments, or "black box" optimization methods [30].

The proper orthogonal decomposition (POD) method has been used extensively in a variety of fields including fluid dynamics [23], identification of coherent structures [12, 21], and control [27] and inverse problems [19]. The method has been employed for industrial applications such as supersonic jet modeling [5], turbine flows [6], thermal processing of foods [3], and study of the dynamic wind pressures acting on buildings [16], to name only a few.

Depending on the field of research, POD is also known as principal component analysis (statistics [14]), Karhunen–Loève decomposition (signal analysis and pattern recognition [9]), and the method of empirical orthogonal functions (EOFs) in geophysical fluid dynamics [7, 24] and meteorology [1, 8, 29]. Principal components related techniques (PCAs) are the main dimension-reduction methods in analysis of multivariate data, addressing the need to compress or decompose data for eliminating the redundancy of high throughput measurements such as spatial, spectra, or image data. PCA involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called principal

[*] Received by the editors January 26, 2004; accepted for publication (in revised form) February 28, 2005; published electronically November 22, 2005. This work was supported by grants DOE DE-FG03-00ER25430, NSF/NCSA ACI-9619019, and NSF/ITR ACI-0086061.

http://www.siam.org/journals/sinum/43-4/060354.html

[†] Department of Computer Science, University of California, Santa Barbara, CA 93106 (homescu@cs.ucsb.edu, petzold@engineering.ucsb.edu).

[‡] Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551 (radu@llnl.gov). The work of this author was performed under the auspices of the U.S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under contract W-7405-Eng-48.

components. The PCA components account for as much of the variability in the data as possible. The EOFs form a set of basis functions which specify a transformation on a set of empirical signals. The result is a set of signals that, phenomenologically speaking, are statistically independent, i.e., have maximum variance. Thus, information is evenly distributed among the signals, as well as the equally measurable values of each signal, resulting in maximum information entropy and robustness to noise.

All the above reduction methods attempt to maximize the expectation of the energy in a basis set. It was shown that such an optimal basis is given by the eigenfunctions of the integral equation whose kernel is the averaged autocorrelation function. In practice, the covariance matrix is constructed based on measurements of the state, and the existing model projected onto those eigenvectors which correspond to the largest eigenvalues. Assessing the optimality of these reduction methods (POD, PCA, and EOF) is a *norm dependent* statement. For example, it was shown in [12] that for a given number of modes, POD is the most efficient choice among all linear decompositions in the sense that it retains, on average, the greatest possible kinetic energy.

As soon as one contemplates the use of a reduced model, questions concerning the quality of the approximation become paramount. To judge the quality of the reduced model, it is important to estimate its error. An algorithm for estimating the error of a class of reduction methods based on projection techniques was presented in [32]. In this approach, the original problem is linearized around the initial time. The resulting first-order error estimates are valid for only a small number of time steps (during which the Jacobian matrix can be considered constant). First-order estimates of POD errors were used in [20] to extend the concept of domain decomposition as a dynamic a posteriori verification and, if necessary, correction of the approximate solution. Error estimates for reduced models, more precisely the error for certain functionals of the solution, were obtained in [25]. The authors employed the dual-weighted-residual method, which makes use of the solution of an adjoint system.

In the context of fluid dynamics, bounds for the errors resulting from POD model reduction of 2-dimensional (2-D) Navier–Stokes equations were computed in [19]. In that work, the approximation error was decomposed into a contribution that arises due to the POD spatial approximation (measured in terms of the spectral properties specifying the POD basis) and the approximation error due to the backward Euler scheme for time integration. The resulting estimates made use of certain inequalities that, although valid for the nonlinear evolution problem considered, may not be satisfied for other examples. For models that contain discontinuities, for example, if the solution involves shocks, it was found in [22] that the POD reduced model was able to represent a shock in a given location only if one of the snapshots used to build the model has a discontinuity in the same location. This may require an unacceptably large number of snapshots to achieve sufficient accuracy of the approximate solution. To overcome this limitation a domain decomposition technique was introduced, using a reduced-order model over the majority of the computational domain while solving the full equations in a small region. Given an approximate solution (with unknown accuracy) generated with a set of POD basis functions, the error is estimated by augmenting the POD basis with top hat basis functions and computing the first-order change in the solution due to the additional basis functions. By comparing against the results from a solution of known accuracy, such as one of the snapshots used to generate the POD basis, the need for domain decomposition and its spatial extent can be determined.

Bounds of POD errors, but not estimates, were considered in [26], as well as effects (on the reduced-order model) of small perturbations in the ensemble of data from which the POD-reduced order model was constructed.

In the present work we take the analysis of reduced models one step further by analyzing the influence of perturbations to the original system on the quality of the approximation given by the reduced model. This question is of particular interest in applications (such as control and inverse problems) in which reduced models are used not just to approximate the solution of the original system that provided the data used in constructing the reduced model, but rather to approximate the solution of systems perturbed from the original one. To the best of our knowledge, there are no published results to address the estimation of the model reduction error of such perturbed systems.

We base our approach on a combination of the small sample statistical condition estimation (SCE) method [15] and error estimation using the adjoint method. Using this framework, we define *regions of validity* of the reduced models, that is, ranges of perturbations in the original system over which the reduced model is still appropriate. We consider perturbations in both the initial conditions and in parameters describing the dynamical system itself. The proposed approach is particularly attractive because the resulting error bounds do not rely on the solution of the perturbed system. In this sense, we provide an a priori assessment of the validity of the model-reduction approximation. We note that our approach is based on linearization. For large enough perturbations, knowledge of the solution of the perturbed system would be required.

Unlike the method presented in [32], our estimates and bounds are valid over the entire time interval considered, not in a neighborhood of the initial time. Moreover, we obtain estimates for the continuous error, as opposed to its discrete approximation. Although we study only a particular projection-based model reduction technique (POD) among those considered in [32], the methodology developed here for POD can be easily extended to other types of projection. Compared to the approach taken in [19], our method is applicable to a larger class of problems, our main requirement being that the norm of the POD-based error is small enough for the linearized error equation to be a good enough approximation. Furthermore, our estimates are independent of the time integration method. We note also that our use of adjoint models for error estimation is similar to that employed in [25]. However, as will be seen below, the use of the SCE method enables the derivation of error "condition numbers" and allows effective treatment of the region of validity problem.

In the context of integration of ordinary differential equations (ODE), the SCE method combined with the adjoint approach has been used in [4] for estimation and control of the global integration error.

The remainder of this paper is organized as follows. In sections 2 and 3 we briefly describe the use of POD for model reduction and, respectively, the SCE method for norm estimation. In section 3.1 we motivate our proposed approach of using SCE, combined with error estimation using the adjoint method, to estimate the errors due to the use of a reduced-order model. In section 4 we analyze errors arising purely from the model reduction itself: the total approximation error and the subspace integration error. In section 5 we analyze regions of validity of POD reduced models. In section 6 we present numerical results for two example problems. The first one is obtained from the semidiscretization of time-dependent partial differential equation (PDE), namely advection-diffusion, while the second example models a pollution chemical reaction mechanism. Finally, section 7 summarizes our results and describes our plans for future research.

2. POD-based reduced models. POD provides a method for finding the best approximating affine subspace to a given set of data. When using POD for model reduction of dynamical systems, the data are time snapshots of the solution obtained via numerical simulations or from experiments. Consider the ODE system

$$\frac{dy}{dt} = f(y,t)\,, \quad y(t_0) = y_0\,, \tag{2.1}$$

for $t \in [t_0, t_f]$, with $y, y_0 \in R^n$ and $f : R^n \times R \to R^n$. Consider next the solutions of (2.1) at m time points, collected in the $n \times m$ matrix $\mathcal{Y} = [y(t_1) - \bar{y}, y(t_2) - \bar{y}, \dots y(t_m) - \bar{y}]$, where $\bar{y}$ is the mean of these observations. POD seeks a subspace $S \in R^n$ and the corresponding projection matrix P_S so that the total square distance

$$\|\mathcal{Y} - P\mathcal{Y}\|^2 = \sum_{i=1}^{m} \|\,(y(t_i) - \bar{y}) - P\,(y(t_i) - \bar{y})\,\|^2$$

is minimized. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ be the ordered eigenvalues of the *correlation matrix* $R = \mathcal{Y}\mathcal{Y}^T$. Then the minimum value of $\|\mathcal{Y} - P\mathcal{Y}\|^2$ over all k-dimensional subspaces S, with $k \leq n$, is given by $\sum_{j=k+1}^{n} \lambda_j$. Moreover, the minimizing S is the invariant subspace corresponding to the eigenvalues $\lambda_1, \dots, \lambda_k$. Using the singular value decomposition (SVD) [10] of the observation matrix, $U^T \mathcal{Y} V = \Sigma$, the projection matrix corresponding to the optimal POD subspace S is obtained as

$$P = \rho\rho^T \in R^{n \times n}\ , \tag{2.2}$$

where ρ is the matrix of projection onto S, the subspace spanned by the reduced basis obtained from the SVD. The matrix $\rho \in R^{n \times k}$ consists of the columns V_i $(i = 1, \dots, k)$, the singular vectors corresponding to the k largest singular values.

Without loss of generality, for the sake of simplicity in presentation we assume in what follows that $\bar{y} = 0$.

In a coordinate system embedded in S, the projection of a point $y \in R^n$ onto S is represented by $z = \rho^T y \in R^k$, while in the full space the same projection is expressed as $\rho z = Py \in R^n$.

A POD-based reduced model that approximates the original problem (2.1) can then be constructed [26] by projecting onto S the vector field $f(y,t)$ at each point $y \in S$. Therefore

$$\frac{dz}{dt} = \rho^T f(\rho z, t)\,, \quad z(t_0) = \rho^T y_0\,. \tag{2.3}$$

In full space, the approximate solution $\widetilde{y}$ is the solution of the ODE initial-value problem (IVP)

$$\frac{d\widetilde{y}}{dt} = Pf(\widetilde{y}, t)\,, \quad \widetilde{y}(t_0) = Py_0\,. \tag{2.4}$$

3. Small sample statistical method for condition estimation. The SCE method, originally proposed in [15], offers an efficient means for condition estimation for general matrix functions, at the cost of allowing moderate relative errors in the estimate. The basic idea is described below (for complete details, see [11, 15]).

For any vector $x \in R^n$, if u is selected uniformly and randomly from the unit sphere S_{n-1}, the expected value of $u^T x$ is proportional to the norm of x:

$$E(|u^T x|) = W_n \|x\| \,.$$

The *Wallis factor* W_n is defined as

$$W_1 = 1\,, \quad W_n = \begin{cases} \dfrac{1 \cdot 3 \cdots (n-2)}{2 \cdot 4 \cdots (n-1)} & n \text{ odd} \\ \dfrac{2}{\pi} \dfrac{2 \cdot 4 \cdots (n-2)}{1 \cdot 3 \cdots (n-1)} & n \text{ even} \end{cases}$$

and can be approximated with $W_n \approx \sqrt{2/(\pi(n-1/2))}$. Therefore $\xi = |u^T x|/W_n$ is an estimate for the norm $\|x\|$. This estimate is first order in the sense that the probability of a relative error in the estimate is inversely proportional to the size of the error. That is, for $\gamma > 1$,

$$\Pr\left(\frac{\|x\|}{\gamma} \leq \xi \leq \gamma \|x\|\right) \geq 1 - \frac{2}{\pi\gamma} + O\left(\gamma^{-2}\right) \,.$$

Additional function evaluations can improve the estimation procedure. Suppose that we obtain estimates $\xi_1, \xi_2, \ldots, \xi_q$ corresponding to orthogonal vectors $u_1, u_2, \ldots, u_q$ selected uniformly and randomly from the unit sphere S_{n-1}. The expected value of the norm of the projection of x onto the span $\mathcal{U}$ generated by $u_1, u_2, \ldots, u_q$ is

$$E\left(\sqrt{|u_1^T x|^2 + |u_2^T x|^2 + \cdots + |u_q^T x|^2}\right) = \frac{W_n}{W_q} \|x\| \,.$$

The analysis in [15] shows that the estimate $\nu(q) = (W_q/W_n)\sqrt{|u_1^T x|^2 + \cdots + |u_q^T x|^2}$ is qth order accurate; i.e., a relative error of size γ in the estimate occurs with probability proportional to γ^{-q}. For example,

$$\begin{aligned} \Pr\left(\frac{\|x\|}{\gamma} \leq \nu(2) \leq \gamma \|x\|\right) &\approx 1 - \frac{\pi}{4\gamma^2}\,, \\ \Pr\left(\frac{\|x\|}{\gamma} \leq \nu(3) \leq \gamma \|x\|\right) &\approx 1 - \frac{32}{3\pi^2\gamma^3}\,, \\ \Pr\left(\frac{\|x\|}{\gamma} \leq \nu(4) \leq \gamma \|x\|\right) &\approx 1 - \frac{81\pi^2}{512\gamma^4}\,. \end{aligned}$$

3.1. SCE for estimation of approximation errors in model reduction.

All error estimates derived in this paper begin with the linearizations of one of the ODEs, (2.1), (2.3), or (2.4), or perturbations of these. Thus the error estimates are based on solutions of linear error equations. To estimate the norm $\|e(t_f)\|$ of an error vector $e(t) \in R^n$ at $t = t_f$, we need to evaluate quantities $u_j^T e(t_f)$ for some random vector u_j selected uniformly from the unit sphere S_{n-1}. The norm estimate is then

$$||e(t_f)|| \approx \frac{W_q}{W_n} \sqrt{\sum_{j=1}^{q} |u_j^T e(t_f)|^2}\,. \tag{3.1}$$

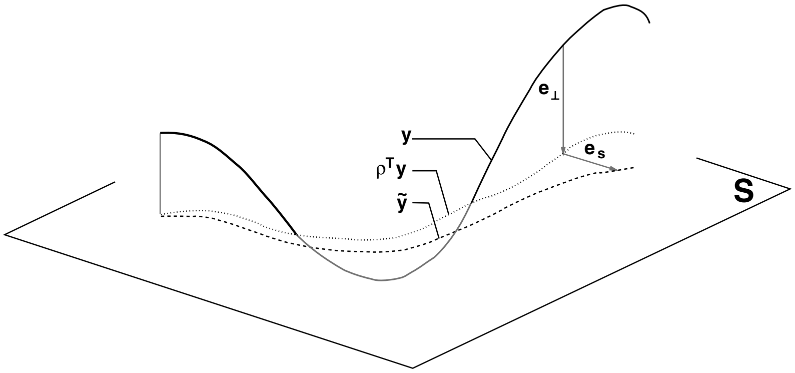


FIG. 4.1. *Solution and error components for POD-reduced models. y is the solution of the original ODE, $z = \rho^T y$ is its projection on the subspace S, and $\widetilde{y}$ is the solution of the reduced model. The error component $e_\perp \in S^\perp$, while the subspace integration error component $e_S \in S$.*

The scalar products $u_j^T e(t_f)$ can be computed efficiently using an adjoint model (to the corresponding linear error equation) with final conditions at t_f based on the vector u_j. However, this approach naturally raises the question: *"What is the advantage of using (typically more than one) solution(s) of the adjoint system to estimate the norm of a quantity that can be otherwise obtained with only one forward ODE solution (of the error equation)?"* Our method is motivated by the fact that we are interested not only in estimating the error for one given ODE system, but rather in estimating (as efficiently as possible) the behavior of such errors for families of related ODE systems, based on different values of problem parameters. In section 5 we study the concept of *regions of validity of reduced models*, i.e., the range of perturbations in the original ODE (2.1) over which the reduced model (2.3) is still appropriate. An approach based on forward error equations involves solving repeatedly such error equations (for each value of interest of the perturbation). On the other hand, an approach combining SCE estimates and adjoint models (as described in our paper) can be used to define what we term "condition numbers" for these error equations. While these condition numbers can provide only approximate upper bounds for the norms of the errors under investigation, they have the undeniable advantage of allowing a priori estimates of the errors induced by perturbations, i.e., before having to solve such a perturbed system (or even a reduced perturbed system).

4. Estimation of the approximation error. We begin by estimating the difference between the solution of the POD-reduced model (2.4) and the solution of the original equation (2.1). The total approximation error $e = \widetilde{y} - y$ can be split [26] into the subspace approximation error $e_\perp = \rho^T y - y$ and the error introduced by the integration in the subspace S, $e_S = \widetilde{y} - \rho^T y$:

$$e = \widetilde{y} - y = \left(\widetilde{y} - \rho^T y\right) + \left(\rho^T y - y\right) = e_S + e_\perp\,. \tag{4.1}$$

The error component $e_\perp$ is orthogonal to S, while the component e_S is parallel to S (see Figure 4.1). Algebraically, this is expressed as $Pe_\perp(t) = 0$ and $Pe_S(t) = e_S(t)$.

4.1. Total approximation error. Subtracting (2.1) from (2.4) yields an equation for the total error e,

$$\frac{de}{dt} = Pf(\widetilde{y},t) - f(y,t) = Pf(\widetilde{y},t) - f(\widetilde{y},t) + f(\widetilde{y},t) - f(y,t)$$

$$= (P - I)f(\widetilde{y}, t) - \mathbf{J}(\widetilde{y}, t)(y - \widetilde{y}) + O(\|e\|)\,,$$

where $\mathbf{J}$ is the Jacobian of the function f, i.e., $\mathbf{J} = \partial f/\partial y$, and we define $Q = I - P$. Thus, to a first-order approximation, the error function satisfies

$$\frac{de}{dt} = \mathbf{J}(\widetilde{y}, t)e(t) - Qf(\widetilde{y}, t)\,, \quad e(t_0) = -Qy_0\,. \tag{4.2}$$

Let the matrix function $\Phi(t) \in R^{n\times n}$ satisfy

$$\frac{d\Phi}{dt} = \mathbf{J}(\widetilde{y}, t)\Phi\,, \quad \Phi(t_0) = I_n\,.$$

Then

$$e(t_f) = -\int_{t_0}^{t_f} \Phi(t_f)\Phi^{-1}(\tau)Qf(\widetilde{y}\,(\tau), \tau)\,d\tau - \Phi(t_f)Qy_0\,.$$

For a random vector u uniformly selected from the unit sphere S_{n-1}, we have

$$u^Te(t_f) = -\int_{t_0}^{t_f} u^T\Phi(t_f)\Phi^{-1}(\tau)Qf(\widetilde{y}\,(\tau), \tau)\,d\tau - u^T\Phi(t_f)Qy_0\,.$$

It is straightforward to verify that the solution $\lambda \in R^n$ of the adjoint system,

$$\frac{d\lambda}{dt} = -\mathbf{J}^T(\widetilde{y}, t)\lambda\,, \quad \lambda(t_f) = u, \tag{4.3}$$

satisfies $\lambda^T(s) = u^T\Phi(t_f)\Phi^{-1}(s)$ and $\lambda^T(t_0) = z^T\Phi(t_f)$. Therefore the quantity $u^Te(t_f)$ is simply

$$u^Te(t_f) = -\int_{t_0}^{t_f} \lambda^T(\tau)Qf(\widetilde{y}\,(\tau), \tau)\,d\tau - \lambda^T(t_0)Qy_0\,. \tag{4.4}$$

The SCE estimate for the norm of $e(t_f)$ is obtained by combining (3.1) and (4.4):

$$\|e(t_f)\| \approx \frac{W_q}{W_n}\sqrt{\sum_{j=1}^{q}\left|\int_{t_0}^{t_f} \lambda^T(\tau)Qf(\widetilde{y}\,(\tau), \tau)\,d\tau + \lambda^T(t_0)^TQy_0\right|^2}\,. \tag{4.5}$$

The value of the integral is $\xi(t_0)$, where ξ satisfies the quadrature equation

$$\frac{d\xi}{dt} = -\lambda^T(t)Qf(\widetilde{y}\,(t), t)\,, \quad \xi(t_f) = 0\,. \tag{4.6}$$

Algorithm 1 summarizes our approach.

It may seem more efficient to compute the SCE norm estimate using a POD-reduced adjoint system to evaluate λ in (4.5). Although the same projection can be used to model-reduce the adjoint system, this approach still requires knowledge of the mean of the adjoint solution, which is unavailable without a solution of the adjoint system (4.3). In other words, the approximation subspace is parallel to S but not identical to it. This issue can be circumvented if we are not considering error components outside the subspace S. This estimate is presented next. Its main advantage is given by the fact that the differential equations are solved in a space of dimension $k \ll n$, where n is the dimension of the solution for the original problem.

Algorithm 1 Estimate for the total approximation error

```
Provide the matrix of measurement data $\mathcal{Y}$
Set the POD dimension $k$
Construct POD projection matrices $\rho$ and $P$
Select uniformly and randomly $q$ orthogonal vectors $u_i$ from the unit sphere $S_{n-1}$
Solve (2.3) for $z$ and compute $\widetilde{y}(t) = \rho z(t)$
Initialize $s = 0$
for $i = 1$ to $q$ do
  Set $\lambda(t_f) = u_i$ and $\xi(t_f) = 0$
  Solve (4.3)+(4.6) for $\lambda$ and $\xi$
  Update $s \leftarrow s + \left[\psi(t_0) + \lambda^T(t_0)^T Q y_0\right]^2$
end for
Compute Wallis factors $W_q$ and $W_n$
Compute the SCE norm estimate $\|e\| = (W_q/W_n) \cdot \sqrt{s}$
```

4.2. Subspace integration error. Starting with its definition, $e_S = \widetilde{y} - \rho^T y$, the subspace integration error is readily found to obey, in a first-order approximation, the following ODE:

$$\frac{de_S}{dt} = \frac{d\,\widetilde{y}}{dt} - P\frac{dy}{dt} = P\left(f(\widetilde{y},t) - f(y,t)\right)$$
$$\approx P\mathbf{J}(\widetilde{y},t)e(t) = P\mathbf{J}(\widetilde{y},t)\left(e_S + e_\perp\right)\,.$$

The starting point $\widetilde{y}(t_0)$ is the projection $\rho^T y(t_0)$ of $y(t_0)$ onto S, yielding the initial condition $e_S(t_0) = 0$. Thus, the subspace integration error is governed by an ODE with the subspace approximation error $e_\perp(t)$ as forcing term,

$$\frac{de_S}{dt} = P\mathbf{J}(\widetilde{y},t)e_S + P\mathbf{J}(\widetilde{y},t)e_\perp\,, \quad e_S(t-0) = 0\,. \tag{4.7}$$

We note that the linearization in (4.7) is directly related (through the projection matrix P) to the linearization of the full model, $f(\widetilde{y},t) - f(y,t) \approx \mathbf{J}(\widetilde{y},t)(\widetilde{y} - y)$. Since we assume that we operate in a region where the full model linearization is valid, this implies that the linearization in (4.7) is valid for the region considered.

If h are the S-coordinates of e_S, i.e., $h = \rho^T e_S \in R^k$, we have $e_S = \rho h$ and therefore

$$\frac{dh}{dt} = \rho^T\mathbf{J}(\widetilde{y},t)\rho h + \rho^T\mathbf{J}(\widetilde{y},t)e_\perp\,, \quad h(t_0) = 0\,, \tag{4.8}$$

where we have used that $\rho^T\rho = I_k$. Now let $\psi \in R^{k\times k}$ be the fundamental matrix of (4.8); i.e,

$$\frac{d\psi}{dt} = \rho^T\mathbf{J}(\widetilde{y},t)\rho\psi\,, \quad \psi(t_0) = I_k\,.$$

Then, for a random vector v uniformly selected from the unit sphere S_{k-1}, we have

$$v^T h(t_f) = \int_{t_0}^{t_f} v^T\psi(t_f)\psi^{-1}(\tau)\rho^T\mathbf{J}(\widetilde{y}\,(\tau),\tau)e_\perp(\tau)\,d\tau\,.$$

The solution μ of the adjoint system

$$\frac{d\mu}{dt} = -\rho^T\mathbf{J}^T(\widetilde{y},t)\rho\mu\,, \quad \mu(t_f) = v, \tag{4.9}$$

satisfies $\mu^T(\tau) = v^T\psi(t_f)\psi^{-1}(\tau)$, for all $\tau \in [t_0, t_f]$, and therefore

$$v^T h(t_f) = \int_{t_0}^{t_f} \mu^T(\tau)\rho^T\mathbf{J}(\widetilde{y}\,(\tau),\tau)e_\perp(\tau)\,d\tau\,,$$

yielding the following SCE estimate for the norm of the subspace integration error:

$$\|e_S(t_f)\| = \|h(t_f)\| \approx \frac{W_q}{W_n}\sqrt{\sum_{j=1}^{q}\left|\int_{t_0}^{t_f}\mu_j^T(\tau)\rho^T\mathbf{J}(\widetilde{y}\,(\tau),\tau)e_\perp(\tau)\,d\tau\right|^2}\,, \tag{4.10}$$

where μ_j is the solution of (4.9) with final condition $\mu(t_f) = v_j$.

Bounds for the subspace integration error can be obtained as follows. We have

$$\begin{aligned}\left|\int_{t_0}^{t_f}\mu^T(\tau)\rho^T\mathbf{J}(\widetilde{y}\,(\tau),\tau)e_\perp(\tau)\,d\tau\right| &\le \int_{t_0}^{t_f}\left|\mu^T(\tau)\rho^T\mathbf{J}(\widetilde{y}\,(\tau),\tau)e_\perp(\tau)\right|\,d\tau\\ &\le \|\mathbf{J}^T\rho\mu\|_{L_1}\cdot\|e_\perp\|_{L_\infty}\,,\end{aligned}$$

where the last inequality is Hölder's inequality, $\|f^Tg\|_{L_1} \le \|f\|_{L_p}\cdot\|g\|_{L_q}$, $1/p+1/q = 1$, for $p = 1$ and $q = \infty$, applied to vector-valued functions $f, g : [t_0, t_f] \to R^n$ for which the L_p norm is defined as

$$||f||_{L_p} = \left(\int_{t_0}^{t_f}||f(\tau)||_p^p\,d\tau\right)^{1/p}\,,\ \text{where } ||f(\tau)||_p = \left(\sum_{i=1}^{n}|f_i(\tau)|^p\right)^{1/p}\,.$$

Therefore

$$\|e_S(t_f)\| \le \kappa(e_S)\cdot\|e_\perp\|_{L_\infty}\,, \tag{4.11}$$

where

$$\kappa(e_S) = \frac{W_q}{W_n}\sqrt{\sum_{j=1}^{q}\|J^T\rho\mu_j\|_{L_1}^2} = \sqrt{\sum_{j=1}^{q}\left(\int_{t_0}^{t_f}\left|J^T(\widetilde{y}\,(\tau),\tau)\rho\mu_j(\tau)\right|\,d\tau\right)^2}\,.$$

The quantity $\kappa(e_S)$ can be seen as a "condition number" for the subspace integration error.

The expressions derived above require knowledge of the projection error $e_\perp$ at all times in $[t_0, t_f]$. While the projection error may not be readily available, its norm can be easily related to the error associated with the choice of the POD subspace. For this, a more convenient formulation of the POD approximation is to find a subspace $S \subset R^n$ which minimizes the total square distance defined as

$$d^2 = \|y - Py\|_{L_2}^2 = \int_{t_0}^{t_f}\|y(\tau) - Py(\tau)\|_2^2\,d\tau\,. \tag{4.12}$$

The solution to this problem requires the construction of the correlation matrix $R = \int_{t_0}^{t_f} y(\tau)y(\tau)^T\,d\tau$. If $\lambda_1 \ge \cdots \ge \lambda_m \ge 0$ are the ordered eigenvalues of the symmetric positive semidefinite matrix R, then the minimum value of d^2 over all k-dimensional affine subspaces S passing through $\bar{y}$ is given by $\sum_{j=k+1}^{n}\lambda_j$. The minimizing S is the invariant subspace corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_k$, while the projection matrix ρ consists of the unit eigenvectors corresponding to these k largest eigenvalues. We also have that

$$||e_\perp||_{L_\infty} \leq ||e_\perp||_{L_2} \equiv \sqrt{\sum_{j=k+1}^{n} \lambda_j}\,.$$

Employing observations as data points for a trapezoidal approximation for the integral (4.12) leads to the same subspace S as the one obtained with the POD definition in section 2, while the corresponding optimal total square distances will be proportional.

5. Regions of validity for POD-reduced models.

Once a reduced model is constructed, we wish to apply it to simulate systems that are close in some sense to the system that was used for generating the reduced model. This raises the issue of defining the range of initial conditions and parameters over which the reduced model can be used with acceptable accuracy.

In the following section we denote by a small letter (e.g., y) any solution of the unperturbed system and by a capital letter (e.g., Y) any solution of a perturbed system.

If $Y \in R^n$ is the solution of an ODE obtained by applying a perturbation to (2.1), either in the initial conditions or in the right-hand side, the issue of the errors introduced by this perturbation, in addition to the model reduction error $e(t)$, can be addressed from two different perspectives:

- When the reduced model, with a POD projection matrix based on the solution of the unperturbed ODE, is used to approximate the perturbed solution Y, it is of interest to estimate the error $E_1 = \widetilde{Y} - Y$, where $\widetilde{Y}$ is the solution of an ODE of the form (2.4), with P based on y.
- Alternatively, we may want estimates for the cumulative error (due to the POD model reduction and the perturbation in the original ODE), $E_2 = \widetilde{Y} - y$. Note that calculating $E_2 = \widetilde{Y} - y$ is completely equivalent to computing $\widetilde{y} - Y$ (by considering y to be a perturbation to Y).

It is important to realize that *useful* estimates should not rely on the solution Y (or $\widetilde{Y}$) of the perturbed system (or its POD reduction). Indeed, such error estimates are desired with the sole objective of deciding whether or not to solve these systems.

In this section we begin by analyzing the errors E_1 and E_2 induced by a perturbation δy_0 in the initial conditions of (2.1) and then by treating the case of perturbations δp in model parameters affecting the right-hand side. For each of these two cases, Figure 5.1 illustrates the solutions of the unperturbed and perturbed full- and reduced-order models, as well as the corresponding errors e, E_1, and E_2.

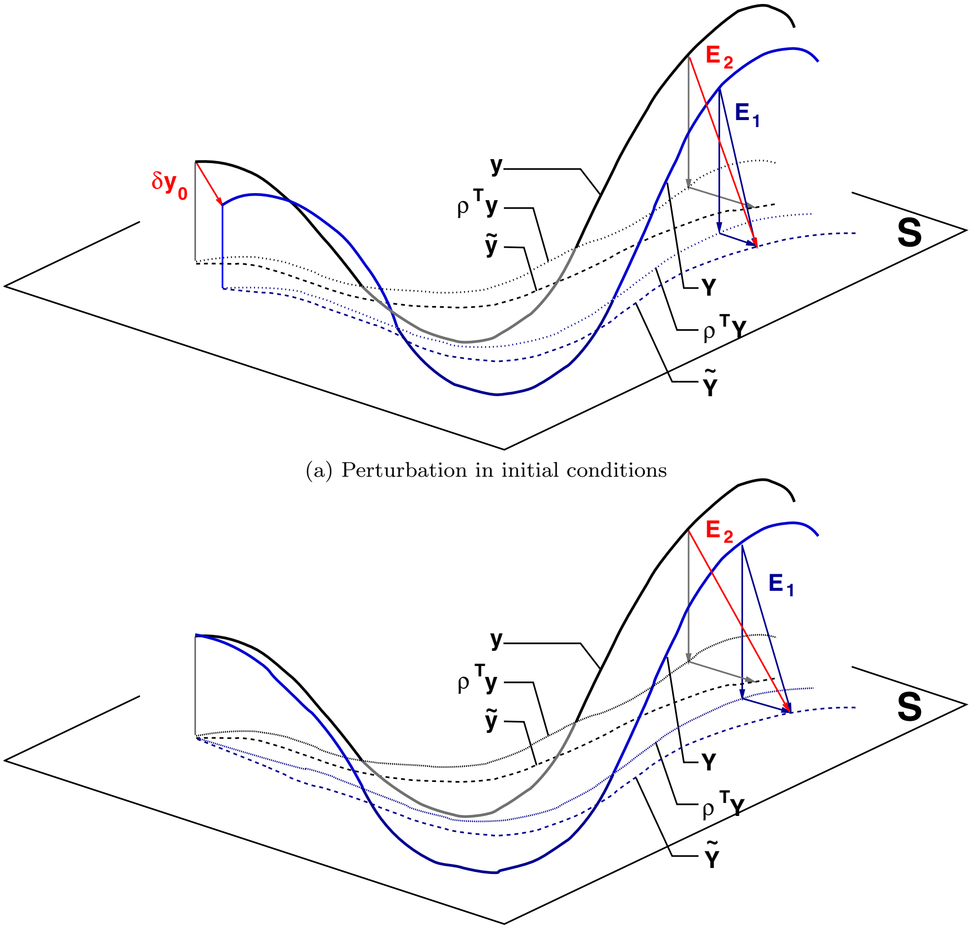
5.1. Perturbations in initial conditions.

Here, Y and $\widetilde{Y}$ are solutions of the ODEs

$$\frac{dY}{dt} = f(Y,t)\,, \quad Y(t_0) = Y_0 = y_0 + \delta y_0\,, \tag{5.1}$$

$$\frac{d\widetilde{Y}}{dt} = Pf(\widetilde{Y},t)\,, \quad \widetilde{Y}(t_0) = PY_0 = P(y_0 + \delta y_0), \tag{5.2}$$

which were obtained by perturbing the initial conditions of (2.1).



(a) Perturbation in initial conditions

(b) Perturbation in right-hand side

Fig. 5.1. *Error components in model reduction of perturbed systems. The solution of the perturbed system and the solution of the reduced perturbed system are denoted by Y and $\widetilde{Y}$, respectively. The error E_1 represents the error committed in reducing the perturbed model, while E_2 is the cumulative error (perturbation + model reduction).*

5.1.1. Estimation of $E_1 = \widetilde{Y} - Y$. An SCE estimate like (4.5) is not useful in the sense described above, as it would be based on the error equation

$$\frac{dE_1}{dt} = \mathbf{J}(\widetilde{Y}, t)E_1 - Qf(\widetilde{Y}, t)\,, \quad E_1(t_0) = -Q(y_0 + \delta y_0 - \bar{y})\,, \tag{5.3}$$

which is a linearization around the (unknown) trajectory $\widetilde{Y}(t)$.

Instead, let us focus on estimating the norm of $\Delta(t_f) = E_1(t_f) - e(t_f)$, with which the norm $\|E_1(t_f)\|$ could be bounded by

$$\big|\|e(t_f)\| - \|\Delta(t_f)\|\big| \le \|E_1(t_f)\| \le \|e(t_f)\| + \|\Delta(t_f)\|\,. \tag{5.4}$$

Any estimates of $\|\Delta(t_f)\|$ would require solving the POD-reduced perturbed system (5.2). However, as in section 4.2, this problem can be circumvented by splitting the error Δ into two components: $\Delta_\perp$ orthogonal to S and Δ_S parallel to S. Using the fact that $Q\,\widetilde{Y} = Q\,\widetilde{y} = 0$, we have

$$\Delta_\perp = Q\Delta = Q(\widetilde{Y} - \widetilde{y}) - (Y - y) = -Q(Y - y)$$

and

$$\Delta_S = \Delta - \Delta_\perp = (\widetilde{Y} - \widetilde{y}) - P(Y - y)\,.$$

We evaluate the influence of δy_0 on each component separately. Retaining only the first-order term of a Taylor series for $\Delta_\perp$ around $\delta y_0 = 0$ and using the fact that $\Delta_\perp = 0$ for $\delta y_0 = 0$, we get

$$\Delta_\perp = -Q\,\frac{dY}{d\delta y_0}\bigg|_{\delta y_0 = 0} \delta y_0\ .$$

The sensitivity matrix dY/dy_0 is nothing but the fundamental matrix corresponding to the linearization of (2.1). It is then easy to see that if λ is now the solution of

$$\frac{d\lambda}{dt} = -\mathbf{J}^T(y,t)\lambda\,, \quad \lambda(t_f) = Qu, \quad \text{for some } u \in R^n\,, \tag{5.5}$$

then $u^T\Delta_\perp(t_f) = -\lambda^T(t_0)\cdot\delta y_0$.

Therefore, an SCE estimate of $\|\Delta_\perp(t_f)\|$ can be based on the solutions of systems (5.5) with vectors u_j uniformly and randomly selected from the unit sphere S_{n-1}. However, taking into account that $\Delta_\perp$ is orthogonal to S, a more accurate estimate can be obtained by using vectors from the sphere S_{n-k-1} embedded in $S^\perp$, instead of selecting vectors $u \in S_{n-1}$ and projecting them onto $S^\perp$, the orthogonal complement of S. If u' is the representation in R^n of such a vector, then $Qu' = u'$. Thus we have the same adjoint system (5.5), but the probability that the estimate lies within a given factor γ of the true norm $\|\Delta_\perp(t_f)\|$ is now higher (see section 3).

In practice we use the approximation $y \approx \widetilde{y}$ in evaluating the Jacobian in (5.5), with $\widetilde{y}$ computed from the solution z of the k-dimensional ODE (2.3) and obtain the following SCE estimate:

$$\|\Delta_\perp(t_f)\| \approx \frac{W_q}{W_n}\sqrt{\sum_{j=1}^{q}|u_j'^T\Delta_\perp(t_f)|^2} = \frac{W_q}{W_n}\sqrt{\sum_{j=1}^{q}|\lambda^T(t_0)\delta y_0|^2}\,,$$

where λ is the solution of $d\lambda/dt = -\mathbf{J}^T(\widetilde{y},t)\lambda$, $\lambda(t_f) = Qu_j'$. Hölder's inequality (for $p = q = 2$) gives $|\lambda^T(t_0)\delta y_0| \le \|\lambda(t_0)\|_2\cdot\|\delta y_0\|_2$, which implies

$$\|\Delta_\perp(t_f)\| \le \kappa_1\cdot\|\delta y_0\|\,, \tag{5.6}$$

where the "condition number" for the orthogonal component of Δ is defined as

$$\kappa_1 = \frac{W_q}{W_n}\sqrt{\sum_{j=1}^{q}\|\lambda(t_0)\|_2^2}\,.$$

With the assumption $\mathbf{J}(y,t) \approx \mathbf{J}(\widetilde{y},t)$, the Δ component parallel to S, $\Delta_S = \Delta - \Delta_\perp = (\widetilde{Y} - \widetilde{y}) - P(Y - y)$, satisfies, up to first order,

$$\frac{d\Delta_S}{dt} = \left(Pf(\widetilde{Y},t) - Pf(\widetilde{y},t)\right) - P\left(f(Y,t) - f(y,t)\right) \approx P\mathbf{J}(\widetilde{y},t)\Delta_S\ .$$

Since at the initial time $\Delta_S(t_0) = 0$, to a first-order approximation $\Delta_S(t) = 0$ for all $t \geq t_0$. In other words, a perturbation to the initial conditions of the original ODE does not introduce additional subspace integration errors. As a consequence, $\Delta(t_f) \approx \Delta_\perp(t_f)$ and, combining (5.4) and (5.6), we have

$$\|E_1(t_f)\| \leq \|e(t_f)\| + \kappa_1 \cdot \|\delta y_0\| \,. \tag{5.7}$$

Note that when using SCE estimates for the norms involved in the above bounds, the true value of $\|E_1(t_f)\|$ may not be bracketed by these bounds.

5.1.2. Estimation of $\boldsymbol{E_2 = \widetilde{y} - Y}$. Subtracting the ODEs satisfied by $\widetilde{y}$ and Y, the error E_2 satisfies, to a first-order approximation,

$$\frac{dE_2}{dt} = \mathbf{J}(\widetilde{y}, t)E_2 - Qf(\widetilde{y}, t)\,, \quad E_2(t_0) = -Qy_0 - \delta y_0\,. \tag{5.8}$$

For a uniformly selected random vector $u \in S_{n-1}$ and with λ the solution of (4.3), we have

$$\begin{aligned} u^T E_2(t_f) &= -\int_{t_0}^{t_f} \lambda^T(\tau) Q f(\widetilde{y}\,(\tau), \tau)\, d\tau - \lambda^T(t_0)\,(Qy_0 + \delta y_0) \\ &= u^T e(t_f) - \lambda^T(t_0)\delta y_0\,, \end{aligned} \tag{5.9}$$

where $e(t_f)$ is the approximation error for the original system, defined by (4.1). Straightforward calculations yield

$$\|E_2(t_f)\| \leq \|e(t_f)\| + \kappa_2 \cdot \|\delta y_0\|\,, \tag{5.10}$$

where

$$\kappa_2 = \frac{W_q}{W_n} \sqrt{\sum_{j=1}^{q} \|\lambda(t_0)\|_2^2}\,.$$

We first note that the new condition number κ_2 has the exact same form as κ_1 obtained in section 5.1.1, the only difference being in the final conditions used for the adjoint variables λ_j. Secondly, the SCE bound estimate (5.10) is more accurate than the SCE bound estimate for the norm of $E_1(t_f)$ (which is based on the additional approximation $E_1 \approx e + \Delta_\perp$, ignoring Δ_S and using $y \approx \widetilde{y}$ in the adjoint system). Furthermore, as seen from (5.9), an SCE estimate for $\|E_2(t_f)\|$ can be computed without need for Y or $\widetilde{Y}$, unlike for $\|E_1(t_f)\|$.

5.2. Perturbations in model parameters. Now let Y be the solution of the ODE system

$$\frac{dY}{dt} = f(Y, t, p + \delta p)\,, \quad Y(t_0) = y_0\,, \tag{5.11}$$

representing a perturbation in some model parameters affecting the right-hand side of (2.1). As in section 5.1, let $\widetilde{Y}$ be the solution of a POD-based reduced-order model obtained from (5.11) using the same POD projection matrix as for the model reduction of the unperturbed system. Then $\widetilde{Y}$ satisfies

$$\frac{d\widetilde{Y}}{dt} = Pf(\widetilde{Y}, t, p + \delta p)\,, \quad \widetilde{Y}(t_0) = Py_0\,.$$

5.2.1. Estimation of $\boldsymbol{E_1 = \widetilde{Y} - Y}$. Similar to section 5.1.1, we decompose the error $\Delta = E_1 - e$ into its components $\Delta_\perp \in S^\perp$ and $\Delta_S \in S$. We retain only the first-order term from the Taylor expansion of $\Delta_\perp$ around $\delta p = 0$,

$$\Delta_\perp = -Q \left.\frac{dY}{d\delta p}\right|_{\delta p=0} \delta p\,.$$

The sensitivity matrix $\Psi = dY/d\delta p$ satisfies

$$\frac{d\Psi}{dt} = \mathbf{J}(y,t,p)\Psi + \mathbf{K}(y,t,p)\,, \quad \Psi(t_0) = 0\,,$$

where $\mathbf{K} = \partial f/\partial p$ is the Jacobian of f with respect to p. In terms of the fundamental matrix Φ of the linearization of (2.1), we have

$$\Psi(t_f) = \int_{t_0}^{t_f} \Phi(t_f)\Phi^{-1}(\tau)\mathbf{K}(y(\tau),\tau,p)\,d\tau,$$

and thus

$$u^T\Delta_\perp(t_f) = -\left(\int_{t_0}^{t_f} \lambda^T(\tau)\mathbf{K}(y(\tau),\tau,p)\,d\tau\right)\cdot \delta p\,,$$

where λ is the solution of (5.5) and $u \in R^n$.

The observations in section 5.1.1 remain valid: (a) using vectors u' from the S_{n-k-1} sphere embedded in $S^\perp$ gives a more accurate SCE error norm estimate; (b) a more efficient adjoint solution can be obtained assuming $\mathbf{J}(y,t,p) \approx \mathbf{J}(\widetilde{y},t,p)$ and $\mathbf{K}(y,t,p) \approx \mathbf{K}(\widetilde{y},t,p)$.

The SCE estimate of the norm of $\Delta_\perp$ is then

$$\|\Delta_\perp(t_f)\| \approx \frac{W_q}{W_n}\sqrt{\sum_{j=1}^{q}\left|\int_{t_0}^{t_f} \lambda^T(\tau)\mathbf{K}(\widetilde{y}\,(\tau),\tau,p)\,\delta p\,d\tau\right|^2}\,,$$

bounded by $\|\Delta_\perp(t_f)\| \le \kappa_1 \cdot \|\delta p\|$, where κ_1 is now defined as

$$\kappa_1 = \frac{W_q}{W_n}\sum_{j=1}^{q} \|\lambda^T\mathbf{K}\|_{L_1}^2.$$

In complete analogy with section 5.1.1, if $\mathbf{J}(y,t,p) \approx \mathbf{J}(\widetilde{y},t,p)$ and $\mathbf{K}(y,t,p) \approx \mathbf{K}(\widetilde{y},t,p)$, the component Δ_S parallel to S satisfies to a first-order approximation

$$\frac{d\Delta_S}{dt} = P\mathbf{J}(\widetilde{y},t,p)\Delta_S\,, \quad \Delta_S(t_0) = 0,$$

and therefore $\Delta_S(t) = 0$, for all $t \ge t_0$. As a consequence, $\Delta(t_f) \approx \Delta_\perp(t_f)$ and

$$\|E_1(t_f)\| \le \|e(t_f)\| + \kappa_1 \cdot \|\delta p\|_\infty\,.$$

5.2.2. Estimation of $\boldsymbol{E_2 = \widetilde{y} - Y}$. Following a similar approach to section 5.2.1, for a uniformly selected random vector $u \in S_{n-1}$ and with $\lambda_{\widetilde{y}}$ the solution of

(4.3),

$$\begin{aligned} u^T E_2(t_f) &= -\int_{t_0}^{t_f} \lambda^T(\tau)\, \Big[Q f(\widetilde{y}(\tau), \tau, p) + \mathbf{K}(\widetilde{y}(\tau), \tau, p)\, \delta p \Big]\, d\tau \\ &\qquad\qquad -\lambda^T(t_0) Q y_0 \\ &= u^T e(t_f) - \int_{t_0}^{t_f} \lambda^T(\tau) \mathbf{K}(\widetilde{y}(\tau), \tau, p)\, \delta p\, d\tau\,, \end{aligned} \tag{5.12}$$

where $e(t_f)$ is the approximation error for the original system, defined by (4.1). As in section 5.1.2, it follows that

$$\|E_2(t_f)\| \le \|e(t_f)\| + \kappa_2 \cdot \|\delta p\|_\infty\,, \tag{5.13}$$

where the condition number κ_2 is now

$$\kappa_2 = \frac{W_q}{W_n} \sqrt{\sum_{j=1}^{q} \|\lambda^T \mathbf{K}\|_{L_1}^2}\,.$$

The above SCE bound estimate for the norm of $E_2(t_f)$ is again more accurate than the one derived in section 5.2.1 for the bound on the norm of $E_1(t_f)$. Furthermore, starting from (5.12), an SCE estimate for $\|E_2(t_f)\|$ can be computed without need for Y or $\widetilde{Y}$, unlike for $\|E_1(t_f)\|$ in section 5.2.1.

6. Examples. We consider reduced-order ODE examples that are representative of problems derived from spatial discretization of PDEs (linear advection-diffusion) or directly obtained from physical phenomena (a pollution model). Additional examples are described in [13].

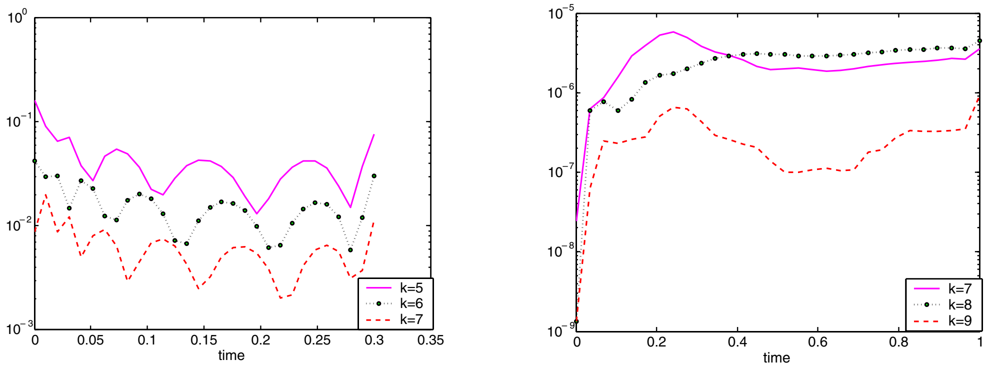
For each example, two figures with numerical results are provided (Figures 6.2 and 6.3 for the first example and Figures 6.4 and 6.5 for the second one). The estimates (and bounds) were obtained using $q = 1$ (blue), $q = 2$ (green), and $q = 3$ (red), where q is the number of orthogonal vectors used by the SCE.

Figure 6.2 contains POD approximation errors as functions of the dimension of the subspace S. The norm of the total approximation error at the final time, $\|e(t_f)\| = \|\widetilde{y}(t_f) - y(t_f)\|$, is given in plot (a), while the norm of the subspace integration error at the final time, computed in the subspace S, i.e., $\|e_S(t_f)\|$, is presented in plot (b). The solid (black) lines represent the corresponding norms computed by the forward integration of the error equations (4.2) and (4.8), respectively. The dotted (colored) lines describe SCE estimates (4.5) and (4.10), respectively, for different values of q. The dashed (colored) lines appear only in plot (b) and represent the bounds of (4.11) for different values of q.

The first four plots in Figure 6.3 contain estimates of errors induced by a perturbation δy_0 in the initial conditions. Plot (a) presents the norm of the total approximation error of the perturbed system at the final time, $\|E_1(t_f)\| = \|\,\widetilde{Y}(t_f) - Y(t_f)\|$, as a function of the subspace dimension k. Plot (b) contains the norm of the cumulative error of the perturbed system at the final time, $\|E_2(t_f)\| = \|\widetilde{y}(t_f) - Y(t_f)\|$, as a function of the subspace dimension k. Plots (c) and (d) present the error bounds for $\|E_1(t_f)\|$ and $\|E_2(t_f)\|$, respectively, as predicted by the condition numbers κ_1 and κ_2 over a range of perturbations δy_0, for a given value of k. The solid (black) line represents the norm computed by the forward integration of the error equations (5.3) and

Table 6.1
The sum of ignored eigenvalues Λ_k and their relative size Λ.

	Example 1: Advection-diffusion		Example 2: Pollution model	
k	Λ_k	Λ	Λ_k	Λ
5	1.803561e-01	5.890413e-06	6.341930e-13	2.652438e-12
6	2.831234e-02	9.246781e-07	6.971282e-14	2.915657e-13
7	4.193422e-03	1.369567e-07	1.139176e-15	4.764470e-15
8	5.662276e-04	1.849294e-08	1.175776e-16	4.917547e-16
9	6.944298e-05	2.268001e-09	4.938977e-17	2.065669e-16
10	7.716002e-06	2.520038e-10	9.158667e-18	3.830506e-17



(a) 1-D advection-diffusion example

(b) Pollution example

Fig. 6.1. *The norm of the total model-reduction error* $\|e\| = \|y - \widetilde{y}\|$ *vs. time, with* y *the solution of the full model and* $\widetilde{y}$ *the solution of the reduced model.*

(5.8), respectively. For different values of q, the dashed (colored) lines represent SCE estimates of the upper bound of (5.7) in plots (a) and (b), and of (5.10) in plots (c) and (d). For different values of q, the dotted (colored) lines represent SCE estimates for $\|E_1(t_f)\|$ in plot (a) and for $\|E_2(t_f)\|$ in plot (b).

The last four plots in Figure 6.3 contain estimates of errors induced by a perturbation δp in the model parameters. The corresponding plots (e), (f), (g), and (h) are in a format which is analogous to the one above.

The (blue) line made of circles represents the norm of the true (nonlinear) error, $e(t) = \widetilde{y}(t) - y(t)$, where $\widetilde{y}$ is the solution of (2.4) and y is the solution of (2.1).

Dimension of the POD subspace. Let $\Lambda_k = \sum_{i=k+1}^{n} \lambda_i$ be the sum of the eigenvalues ignored in the construction of the POD-reduced model and $\Lambda = \Lambda_k / \sum_{i=1}^{n} \lambda_i$ be its relative size compared to the sum of all eigenvalues. The POD subspace dimension k is selected such that the relative error is very close to one, yet k is sufficiently small. A relative error near zero means that a high percentage of the energy for the full model was captured by the reduced-order model. The values of Λ_k and Λ, for the numerical examples considered in this paper, are presented in Table 6.1.

To assess how well the full model is approximated by the POD-based reduced model, we present in Figure 6.1 the behavior of the norm of the total error over the given time interval for both examples considered in the paper. The dimension of the POD subspace is denoted by k and has values $(5, 6, 7)$ for the 1-D advection-diffusion example and $(7, 8, 9)$ for the pollution example.

Number of orthogonal vectors for the SCE estimate. We considered one, two, and three SCE vectors for our numerical examples. As expected, having just one SCE vector yielded the worst estimate in most of the cases. Nevertheless, even that estimate was, in many cases, good enough to warrant its inclusion in our results.

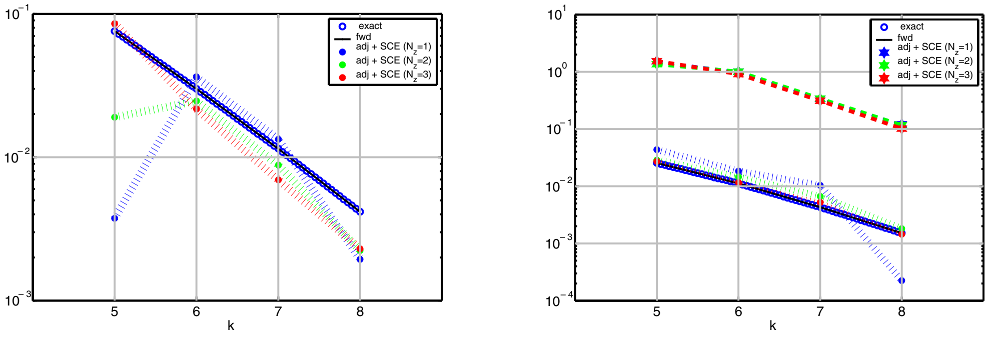
6.1. Linear advection-diffusion model. We consider the 1-D problem

$$\begin{aligned} & u_t = p_1 u_{xx} + p_2 u_x \\ \text{with BC} \quad & u(0,t) = u(2,t) = 0 \\ \text{and IC} \quad & u(x,0) = u_0(x) = x(2-x)e^{2x} \,. \end{aligned}$$

The PDE is discretized on a uniform grid of size $n+2$ with central differencing. With $y_i(t) = u(x_i, t)$ and eliminating boundary values, we obtain the following size n ODE system:

$$\frac{dy_i}{dt} = p_1 \frac{y_{i+1} - 2y_i + y_{i-1}}{\Delta x^2} + p_2 \frac{y_{i+1} - y_{i-1}}{2\Delta x} \,, \quad y_i(0) = u_0(x_i) \,.$$

The problem parameters were $p_1 = 0.5$, $p_2 = 1.0$, and $N = 100$. Results for this problem are shown in Figures 6.2 and 6.3. The POD projection matrices were based on $m = 100$ data points equally spaced in the interval $[t_0, t_f] = [0.0, 0.3]$. The estimate of the total error is consistently close to the exact value, with the estimates corresponding to $q = 2, 3$ almost identical to the subspace integration error. The bounds are within an order of magnitude for both IC and RHS perturbations. The RHS perturbation increases the distance between the bounds and the forward error. That was expected, since the RHS perturbation changes the advection coefficient p_2, which is dominant for the time window considered.



(a) Total error

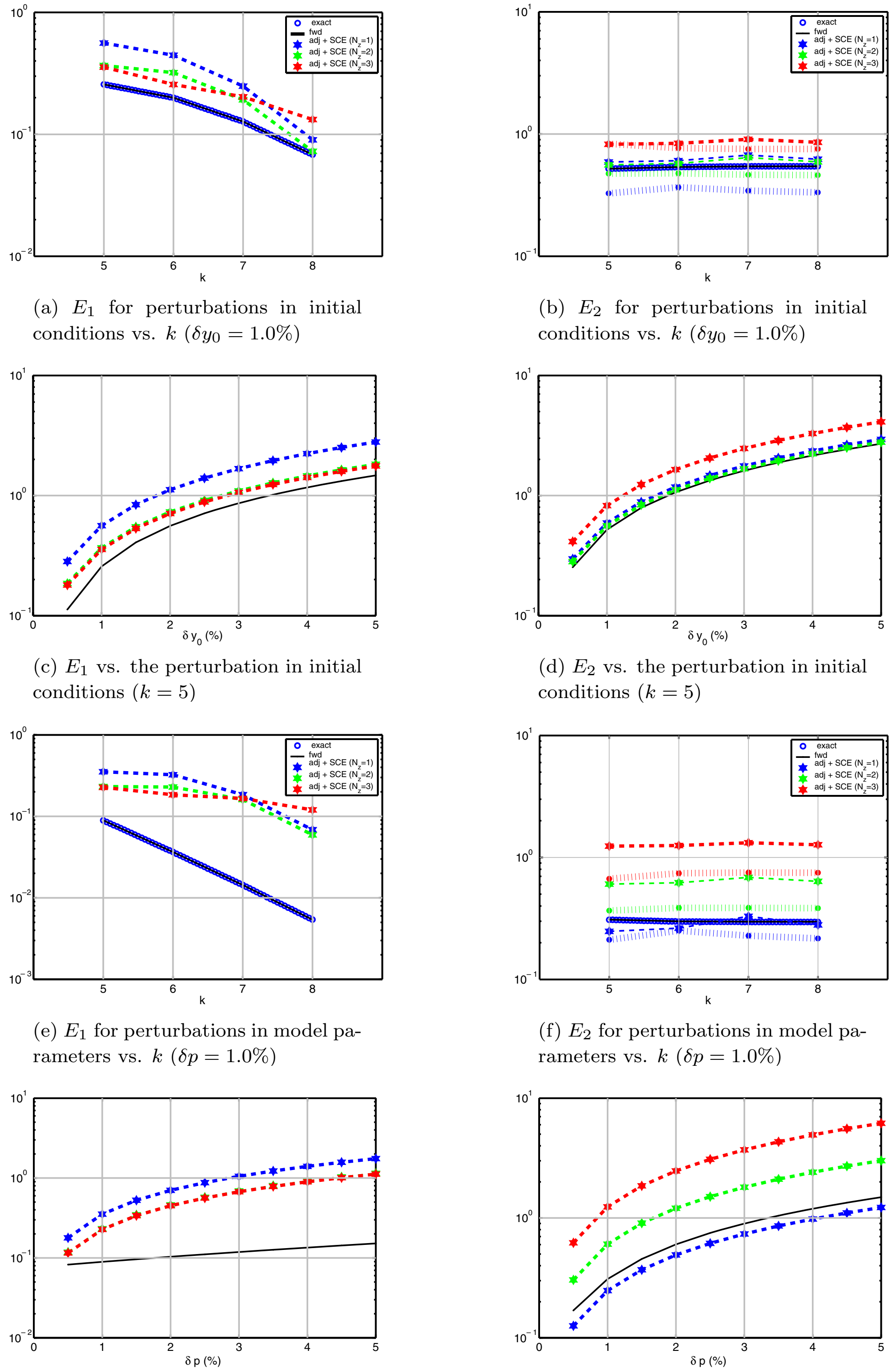
(b) Subspace integration error

FIG. 6.2. *1-D advection-diffusion example. Model-reduction error.*

6.2. Pollution model. Next we consider the chemical reactions from an air pollution model described in [33]. This is a highly nonlinear stiff ODE system consisting of 25 reactions and 20 species. The problem is of the form

$$\frac{dy}{dt} = f(y) \,, \quad y(0) = y_0, \quad y \in R^{20} \,,$$

where the function $f(y)$ is defined by



(a) E_1 for perturbations in initial conditions vs. k ($\delta y_0 = 1.0\%$)

(b) E_2 for perturbations in initial conditions vs. k ($\delta y_0 = 1.0\%$)

(c) E_1 vs. the perturbation in initial conditions ($k = 5$)

(d) E_2 vs. the perturbation in initial conditions ($k = 5$)

(e) E_1 for perturbations in model parameters vs. k ($\delta p = 1.0\%$)

(f) E_2 for perturbations in model parameters vs. k ($\delta p = 1.0\%$)

(g) E_1 vs. the perturbation in model parameters ($k = 5$)

(h) E_2 vs. the perturbation in model parameters ($k = 5$)

FIG. 6.3. *1-D advection-diffusion example. Regions of validity.*

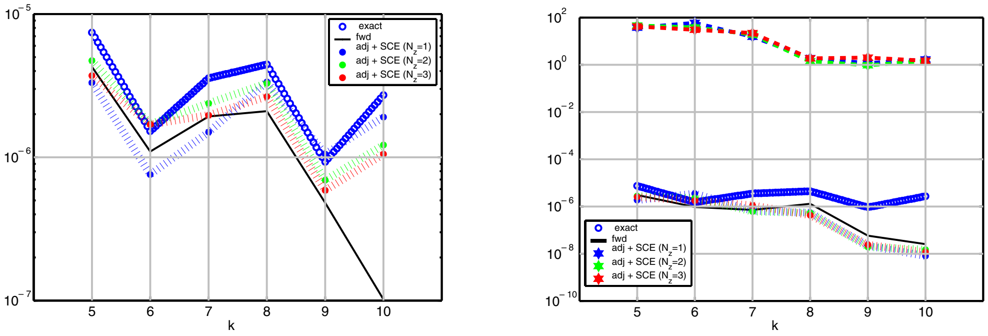
$$\begin{array}{ll}
f_1 = -\sum_{j\in\{1,10,14,23,24\}} r_j + \sum_{j\in\{2,3,9,11,12,22,25\}} r_j & f_{12} = r_9 \\
f_2 = -r_2 - r_3 - r_9 - r_{12} + r_1 + r_{21} & f_{14} = -r_{13} + r_{12} \\
f_3 = -r_{15} + r_1 + r_{17} + r_{19} + r_{22} & f_{18} = r_{20} \\
f_4 = -r_2 - r_{16} - r_{17} - r_{23} + r_{15} & f_{13} = -r_{11} + r_{10} \\
f_5 = -r_3 + r_4 + r_4 + r_6 + r_7 + r_{13} + r_{20} & f_{17} = -r_{20} \\
f_6 = -r_6 - r_8 - r_{14} - r_{20} + r_3 + 2r_{18} & f_{15} = r_{14} \\
f_7 = -r_4 - r_5 - r_6 + r_{13} & f_{16} = -r_{18} - r_{19} + r_{16} \\
f_8 = r_4 + r_5 + r_6 + r_7 & f_{10} = -r_{12} + r_7 + r_9 \\
f_{11} = -r_9 - r_{10} + r_8 + r_{11} & f_9 = -r_7 - r_8 \\
f_{19} = -r_{21} - r_{22} + r_{22} - r_{24} + r_{25} & f_{20} = -r_{25} + r_{24}
\end{array}$$

and $y_0 = [0, 0.2, 0, 0.04, 0, 0, 0.1, 0.3, 0.01, 0.0, 0, 0, 0, 0, 0, 0.007, 0, 0, 0]^T$. The auxiliary variables r_j and the model parameters k_j are given in Table 6.2.

TABLE 6.2
Auxiliary variables (r_j) and model parameters (k_j) for the pollution model.

$r_1 = k_1y_1$ $r_2 = k_2y_2y_4$ $r_3 = k_3y_5y_2$ $r_4 = k_4y_7$ $r_5 = k_5y_7$ $r_6 = k_6y_7y_6$	$r_7 = k_7y_9$ $r_8 = k_8y_9y_6$ $r_9 = k_9y_{11}y_2$ $r_{10} = k_{10}y_{11}y_1$ $r_{11} = k_{11}y_{13}$ $r_{12} = k_{12}y_{10}y_2$	$r_{13} = k_{13}y_{14}$ $r_{14} = k_{14}y_1y_6$ $r_{15} = k_{15}y_3$ $r_{16} = k_{16}y_4$ $r_{17} = k_{17}y_4$ $r_{18} = k_{18}y_{16}$	$r_{19} = k_{19}y_{16}$ $r_{20} = k_{20}y_{17}y_6$ $r_{21} = k_{21}y_{19}$ $r_{22} = k_{22}y_{19}$ $r_{23} = k_{23}y_1y_4$ $r_{24} = k_{24}y_{19}y_1$ $r_{25} = k_{25}y_{20}$
$k_1 = 0.350 \cdot 10^0$ $k_2 = 0.266 \cdot 10^2$ $k_3 = .123 \cdot 10^5$ $k_4 = .860 \cdot 10^{-3}$ $k_5 = .820 \cdot 10^{-3}$ $k_6 = .150 \cdot 10^5$	$k_7 = .130 \cdot 10^{-3}$ $k_8 = .240 \cdot 10^5$ $k_9 = .165 \cdot 10^5$ $k_{10} = .900 \cdot 10^4$ $k_{11} = .220 \cdot 10^{-1}$ $k_{12} = .120 \cdot 10^5$	$k_{13} = .188 \cdot 10^1$ $k_{14} = .163 \cdot 10^5$ $k_{15} = .480 \cdot 10^7$ $k_{16} = .350 \cdot 10^{-3}$ $k_{17} = .175 \cdot 10^{-1}$ $k_{18} = .100 \cdot 10^9$	$k_{19} = .444 \cdot 10^{12}$ $k_{20} = .124 \cdot 10^4$ $k_{21} = .210 \cdot 10^1$ $k_{22} = .578 \cdot 10^1$ $k_{23} = .474 \cdot 10^{-1}$ $k_{24} = .178 \cdot 10^4$ $k_{25} = .312 \cdot 10^1$

Numerical results depicting the approximation errors and the regions of validity at $t_f = 1.0$ are presented in Figures 6.4 and 6.5, respectively. The POD projection matrix was based on $m = 1000$ data points equally spaced in the interval $[t_0, t_f] = [0.0, 1.0]$.

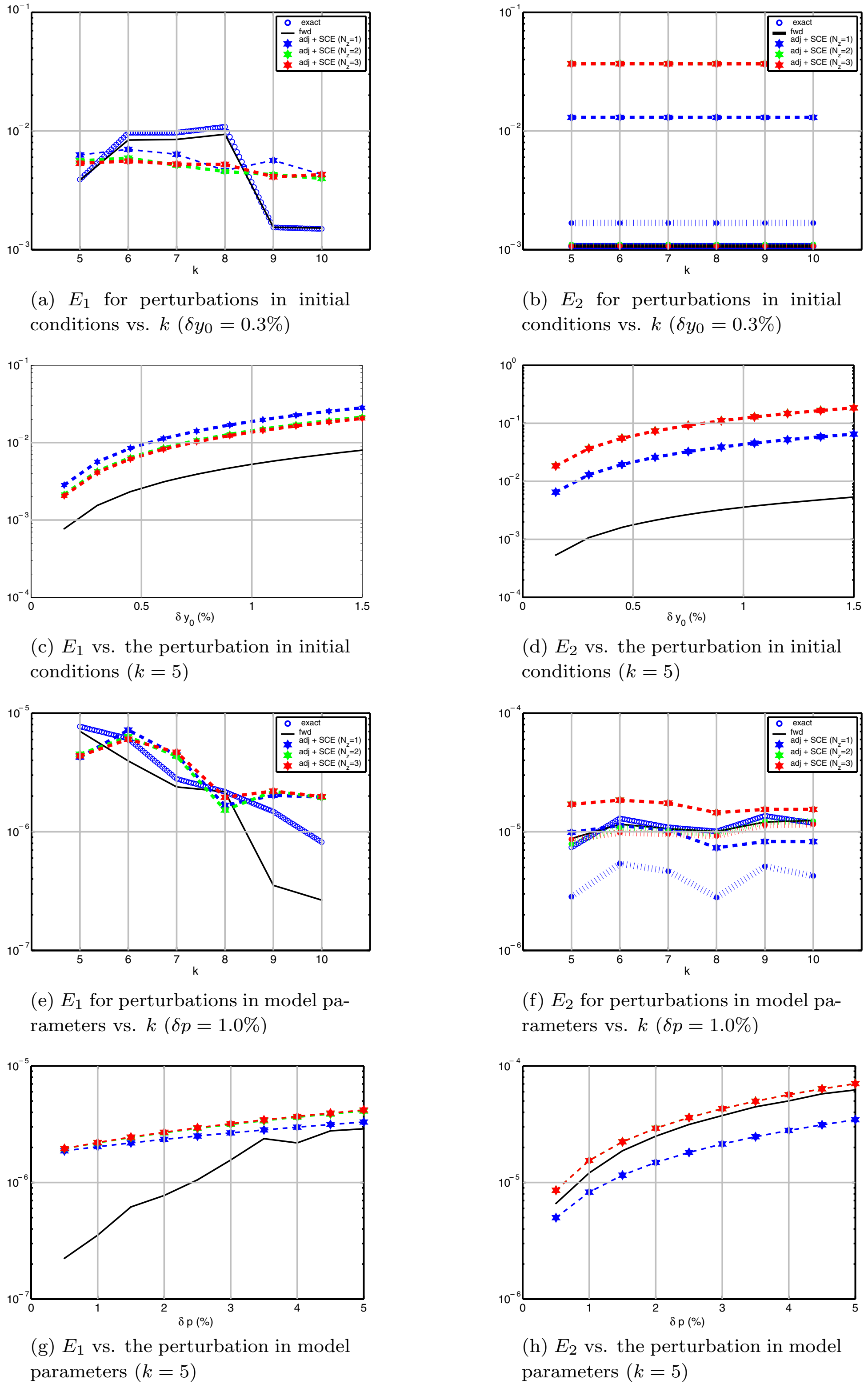


(a) Total error

(b) Subspace integration error

FIG. 6.4. *Pollution example. Model-reduction error.*

For $k = 5, 6, 7$ the total error and the subspace integration error are very well approximated by estimates corresponding to $q = 2$ or 3. For $k = 8, 9, 10$ the estimates are not as good, although they remain within an order of magnitude. We believe that



(a) E_1 for perturbations in initial conditions vs. k ($\delta y_0 = 0.3\%$)

(b) E_2 for perturbations in initial conditions vs. k ($\delta y_0 = 0.3\%$)

(c) E_1 vs. the perturbation in initial conditions ($k = 5$)

(d) E_2 vs. the perturbation in initial conditions ($k = 5$)

(e) E_1 for perturbations in model parameters vs. k ($\delta p = 1.0\%$)

(f) E_2 for perturbations in model parameters vs. k ($\delta p = 1.0\%$)

(g) E_1 vs. the perturbation in model parameters ($k = 5$)

(h) E_2 vs. the perturbation in model parameters ($k = 5$)

FIG. 6.5. *Pollution example. Regions of validity.*

this behavior is related to the fact that the POD error (either absolute or relative) is very small. We note that the problem was solved using relative tolerances of 10^{-4} and absolute tolerance of 10^{-7}. Thus one can expect a less uniform behavior if the results are in the neighborhood of 10^{-7}.

Finally, we note that due to the fact that the problem parameters k_j have orders of magnitude ranging from 10^{-3} to 10^{12}, we have limited the RHS perturbation only to perturbations in k_4, k_5, and k_7.

6.3. More examples. In [13] we include more test problems derived from spatial discretization of PDEs (Burgers's PDE or the Brusselator PDE) or directly obtained from physical phenomena (HIRES High Irradiance Response). The results obtained for those examples confirm our approach, in the sense that the SCE estimates offer a good approximation for the errors of the POD-reduced models. For more details, the reader may consult [13], which is available online.

7. Conclusions and future work.

We have presented effective methods for estimating approximation errors due to the use of POD-based reduced-order models and for evaluating regions of validity of such reduced models. The bounds defining these regions of validity are a priori, in the sense that they do not rely on the solution of the perturbed system. The proposed approach, based on SCE norm estimates combined with the adjoint method, allows the definition and construction of so-called error condition numbers which can be used to assess the size of errors induced by perturbations (in initial conditions or in the model itself) without having to solve the perturbed system. The effectiveness of the proposed methods was demonstrated on several test problems.

We are currently investigating the applicability of this technique to the estimation of errors from other types of reduced-order models, as well as considering more complex models than those presented both in this paper and in [13]. Thus we will consider models which exhibit more interesting (e.g., oscillatory or chaotic) behavior in their POD-reduced model. For example, we think that our method to efficiently compute the error corresponding to different perturbations may be useful in conjunction with reduced models in oceanography or atmospheric sciences (recent advances [21] present POD-based reduced models that can approximate well even the bifurcation behavior of the flow).

REFERENCES

[1] U. Achatz and G. Branstator, *A two-layer model with empirical linear corrections and reduced order for studies of internal climate variability*, J. Atmospheric Sci., 56 (1999), pp. 3140–3160.

[2] A.C. Antoulas and D.C. Sorensen, *Approximation of Large-Scale Dynamical Systems: An Overview*, Tech. report TR0101, Rice University, Houston, TX, 2001.

[3] E. Balsa-Canto, A.A. Alonso, and J.R. Banga, *A novel, efficient and reliable method for thermal process design and optimization*, J. Food Engrg., 52 (2002), pp. 227–247.

[4] Y. Cao and L.R. Petzold, *A posteriori error estimation and global error control for ordinary differential equations by the adjoint method*, SIAM J. Sci. Comput., 26 (2004), pp. 359–374.

[5] E. Caraballo, M. Saminy, J. Scott, S. Narayan, and J. DeBonis, *Application of proper orthogonal decomposition to a supersonic axisymmetric jet*, AIAA J., 41 (2003), pp. 866–877.

[6] P.G.A. Cizmas and A. Palacios, *Proper orthogonal decomposition of turbine rotor-stator interaction*, J. Propul. Power, 19 (2003), pp. 268–281.

[7] D.T. Crommelin and A.J. Majda, *Strategies for model reduction: Comparing different optimal bases*, J. Atmospheric Sci., 61 (2004), pp. 2206–2217.

[8] F. D'Andrea and R. Vautard, *Extratropical low-frequency variability as a low-dimensional problem* I: *A simplified model*, Q.J.R. Meterol. Soc., 127 (2001), pp. 1357–1375.

[9] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, San Diego, CA, 1990.

[10] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1996.

[11] T. Gudmundsson, C.S. Kenney, and A.J. Laub, *Small-sample statistical estimates for matrix norms*, SIAM J. Matrix Anal. Appl., 16 (1995), pp. 776–792.

[12] P. Holmes, J.L. Lumley, and G. Berkooz, *Turbulence, Coherent Structures, Dynamical Systems and Symmetry*, Cambridge University Press, Cambridge, UK, 1998.

[13] C. Homescu, L.R. Petzold, and R. Serban, *Error Estimation for Reduced Order Models of Dynamical Systems*, Tech. report UCRL-TR-201494, Lawrence Livermore National Laboratory, Livermore, CA, 2003.

[14] I.T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 2002.

[15] C.S. Kenney and A.J. Laub, *Small-sample statistical condition estimates for general matrix functions*, SIAM J. Sci. Comput., 15 (1994), pp. 36–61.

[16] H. Kikuchi, Y. Tamura, H. Ueda, and K. Hibi, *Dynamic wind pressures acting on a tall building model - proper orthogonal decomposition*, J. Wind. Eng. Ind. Aerod., 71 (1997), pp. 631–646.

[17] M.E. Kowalski and H.M. Jin, *Model-order reduction of nonlinear models of electromagnetic phased-array hyperthermia*, IEEE T. Bio-Med. Eng., 50 (2003), pp. 1243–1254.

[18] K. Kunisch and S. Volkwein, *Control of the Burgers equation by a reduced-order approach using proper orthogonal decomposition*, J. Optim. Theory Appl., 102 (1999), pp. 345–371.

[19] K. Kunisch and S. Volkwein, *Galerkin proper orthogonal decomposition methods for a general equation in fluid dynamics*, SIAM J. Numer. Anal., 40 (2002), pp. 492–515.

[20] P.A. LeGresley and J.J. Alonso, *Dynamic Domain Decomposition and Error Correction for Reduced Order Models*, AIAA Paper 2003-0250, 41st AIAA Aerospace Sciences Meeting & Exhibit, Reno, NV, 2003.

[21] C. Lopez and E. Garcia-Hernández, *Low-dimensional dynamical system model for observed coherent structures in ocean satellite data*, Phys. A, 328 (2003), pp. 233–250.

[22] D.J. Lucia, P.I. King, M.E. Oxley, and P.S. Beran, *Reduced Order Modeling for a One-Dimensional Nozzle Flow with Moving Shocks*, AIAA Paper 01-2602, AIAA 15th Computational Fluid Dynamics Conference, Anaheim, CA, 2001.

[23] X. Ma and G.E. Karniadakis, *A low-dimensional model for simulating three-dimensional cylinder flow*, J. Fluid Mech., 458 (2002), pp. 181–190.

[24] A.J. Majda, I. Timofeyev, and E. Vanden-Eijnden, *Systematic strategies for stochastic mode reduction in climate*, J. Atmospheric Sci., 60 (2003), pp. 1705–1722.

[25] M. Meyer and H.G. Matthies, *Efficient model reduction in non-linear dynamics using the Karhunen–Loeve expansion and dual-weighted-residual methods*, Comput. Mech., 31 (2003), pp. 179–191.

[26] M. Rathinam and L.R. Petzold, *A new look at proper orthogonal decomposition*, SIAM J. Numer. Anal., 41 (2003), pp. 1893–1925.

[27] S.S. Ravindran, *A reduced-order approach for optimal control of fluids using proper orthogonal decomposition*, Internat. J. Numer. Methods Fluids, 34 (2000), pp. 425–448.

[28] J.A. Rule, R.E. Richard, and R.L. Clark, *Design of an aeroelastic delta wing model for active flutter control*, J. Guid. Control Dynam., 24 (2001), pp. 918–924.

[29] F.M. Selten, *Baroclinic empirical orthogonal functions as basis functions in an atmospheric model*, J. Atmospheric Sci., 54 (1997), pp. 2100–2114.

[30] B. Shapiro, *Creating compact models of complex electronic systems: An overview and suggested use of existing model reduction and experimental system identification tools*, IEEE T. Comp. Pack. T., 26 (2003), pp. 165–172.

[31] Y. Shin and T. Sakurai, *Power distribution analysis of VLSI interconnects using model order reduction*, IEEE T. Comput. Aid. D., 21 (2002), pp. 739–745.

[32] S. Utku, J.L.M. Clemente, and M. Salama, *Errors in reduction methods*, Comput. & Structures, 21 (1985), pp. 1153–1157.

[33] J.G. Verwer, *Gauss–Seidel iteration for stiff ODEs from chemical kinetics*, SIAM J. Sci. Comput., 15 (1994), pp. 1243–1250.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1715–1727

B-SERIES AND ORDER CONDITIONS FOR EXPONENTIAL INTEGRATORS*

HÅVARD BERLAND[†], BRYNJULF OWREN[†], AND BÅRD SKAFLESTAD[†]

Abstract. We introduce a general format of numerical ODE-solvers which include many of the recently proposed exponential integrators. We derive a general order theory for these schemes in terms of *B*-series and bicolored rooted trees. To ease the construction of specific schemes we generalize an idea of Zennaro [*Math. Comp.,* 46 (1986), pp. 119–133] and define natural continuous extensions in the context of exponential integrators. This leads to a relatively easy derivation of some of the most popular recently proposed schemes. The general format of schemes considered here makes use of coefficient functions which will usually be selected from some finite dimensional function spaces. We will derive lower bounds for the dimension of these spaces in terms of the order of the resulting schemes. Finally, we illustrate the presented ideas by giving examples of new exponential integrators of orders 4 and 5.

Key words. exponential integrators, splitting methods, natural continuous extensions, Runge–Kutta schemes

AMS subject classifications. 65L05, 65L20, 65M99

DOI. 10.1137/040612683

1. Introduction.

Numerical integration schemes which use the matrix exponential go back all the way to Certaine [4], but there are also early papers by Lawson [15], Nørsett [20], Ehle and Lawson [6], and Friedli [7] to mention just a few. Recently there has been a revived interest in these schemes, in particular for the solution of nonlinear partial differential equations; see for instance [11, 17, 5, 3, 14, 13]. For a thorough review of the history of exponential integrators; see [16] and the references therein. The integrators found in these papers are derived in rather different ways, and they are formulated for different types of systems of differential equations. On this note, we consider the autonomous nonlinear system of ordinary differential equations

$$\dot{u} = Lu + N(u), \qquad u(0) = u_0. \tag{1.1}$$

Here L is a matrix and $N(u)$ a nonlinear mapping. The order theory we consider is valid for a large class of exponential integrators, including the Runge–Kutta–Munthe-Kaas (RKMK) schemes [17], the commutator-free Lie group integrators [3], and those schemes of Cox and Matthews [5], as well as Krogstad [14] which reduce to classical Runge–Kutta schemes when $L = 0$.

We present the general format for integrators of (1.1) as

$$N_r = N\big(\exp(c_r hL)\, u_0 + h \sum_{j=1}^{s} a_r^j(hL)\, N_j\big), \quad r = 1, \dots, s \tag{1.2}$$

$$u_1 = \exp(hL)\, u_0 + h \sum_{r=1}^{s} b^r(hL)\, N_r. \tag{1.3}$$

*Received by the editors August 2, 2004; accepted for publication (in revised form) March 23, 2005; published electronically November 22, 2005.

http://www.siam.org/journals/sinum/43-4/61268.html

[†]Department of Mathematical Sciences, NTNU, N-7491 Trondheim, Norway (havard.berland@math.ntnu.no, brynjulf.owren@math.ntnu.no, bard.skaflestad@math.ntnu.no).

Here we assume that the functions $a_r^j(z)$ and $b^r(z)$ are at least p times continuously differentiable at $z=0$ for integration schemes of order p.

Table 1
Examples of schemes in general format for exponential integrators.

(a) RKMK, order 4

0				
$\frac{1}{2}$	$\frac{1}{2}\phi_0(z/2)$			
$\frac{1}{2}$	$\frac{z}{8}\phi_0(z/2)$	$\frac{1}{2}(1-\frac{z}{4})\phi_0(z/2)$		
1			$\phi_0(z)$	
	$\frac{1}{6}\phi_0(z)(1+\frac{z}{2})$	$\frac{1}{3}\phi_0(z)$	$\frac{1}{3}\phi_0(z)$	$\frac{1}{6}\phi_0(z)(1-\frac{z}{2})$

(b) Commutator-free, order 4

0				
$\frac{1}{2}$	$\frac{1}{2}\phi_0(z/2)$			
$\frac{1}{2}$		$\frac{1}{2}\phi_0(z/2)$		
1	$\frac{z}{4}\phi_0(z/2)^2$		$\phi_0(z/2)$	
	$\frac{1}{2}\phi_0(z)-\frac{1}{3}\phi_0(z/2)$	$\frac{1}{3}\phi_0(z)$	$\frac{1}{3}\phi_0(z)$	$-\frac{1}{6}\phi_0(z)+\frac{1}{3}\phi_0(z/2)$

Table 1 gives the coefficient functions $a_r^j(z)$ and $b^r(z)$ for the fourth order RKMK scheme introduced in [18] in this general format when applied to the problem (1.1) with an affine Lie group action, and the commutator-free scheme of order 4 from [3]; in both tables $\phi_0(z) = (\mathrm{e}^z - 1)/z$.

For deriving order conditions, we expand the coefficient functions in powers of z,

$$a_r^j(z) = \sum_{k\geq 0} \alpha_r^{j,k} z^k \quad \text{and} \quad b^r(z) = \sum_{k\geq 0} \beta^{r,k} z^k,$$

where the sum may terminate with a remainder term. For the schemes we consider here, these functions are in fact all entire. If $N(u) = 0$ in (1.1), then any scheme in the above class will reproduce the exact solution in every step. Whereas if $L = 0$, the scheme (1.2)–(1.3) reduces to a classical Runge–Kutta method with coefficients $a_r^j = \alpha_r^{j,0}$ and $b^r = \beta^{r,0}$. This scheme is henceforth called *the underlying Runge–Kutta scheme*. We will always assume that $c_r = \sum_j \alpha_r^{j,0},\ 1 \leq r \leq s$.

The schemes proposed by Friedli [7] closely resemble the format (1.2)–(1.3), the difference being that the coefficient functions a_{rj} (resp. b_r) are evaluated in $c_r hL$ rather than in hL, thus a nontrivial discrepancy may occur whenever $c_r = 0$. And even though Friedli explicitly requires that the functions $a_{rj}(z)$ and $b_r(z)$ be of the form

$$\int_0^1 \mathrm{e}^{(1-\theta)z} p(\theta)\,\mathrm{d}\theta, \qquad p(\theta) \text{ polynomial},$$

his analysis holds also for the case of more general coefficient functions, so that the order conditions he obtains for $p \leq 4$ are almost identical to those derived in section 2 here. However, the order theory presented here is general.

We will discuss conditions on the coefficients $\alpha_r^{j,k}$ and $\beta^{r,k}$ under which the scheme (1.2)–(1.3) has order of consistency p for problems of the type (1.1). We will use the well known approach involving rooted trees; see, for instance, [9, 2]. The conditions we find will depend only on the first $\alpha_r^{j,k}$ for $k \le p-2$ and on $\beta^{j,k}$ for $k \le p-1$. On this note we will not address issues related to the behavior of the coefficient functions $a_r^j(z)$ and $b^r(z)$ for large values of z.

In the recent paper [12], an order theory for explicit exponential integrators is presented and its application to semilinear parabolic problems is discussed. While classical or nonstiff order conditions are usually derived by assuming that a Lipschitz constant exists, one needs to account for the unboundedness of the operator L whenever PDEs are considered. It is found that a set of additional order conditions must be satisfied to guarantee convergence order p under suitable assumptions; one requires the linear operator L to be the infinitesimal generator of an analytic semigroup, and that the nonlinear function satisfies a Lipschitz condition. The authors are also able to give an example where order reduction is seen numerically for schemes not satisfying the additional conditions. But the conditions are rather restrictive, and in [13] exponential integrators of (nonstiff) order four are tried out numerically for a number of well-known semilinear PDEs, and no order reduction is seen, despite the fact that these integrators do not satisfy all the required conditions for order four as given in [12]. This shows that the issue of determining the order behavior of exponential integrators for PDEs is indeed a subtle one, and remains today in an unsatisfactory state of nonresolution.

2. *B*-series and order conditions. Repeated differentiation of (1.1) with respect to time yields

$$\begin{aligned}
\frac{\mathrm{d}^2u}{\mathrm{d}t^2} &= L\dot{u} + N'(\dot{u}) \\
&= L^2u + LN + N'(Lu) + N'(N) \\
\frac{\mathrm{d}^3u}{\mathrm{d}t^3} &= L^3u + L^2N + LN'(Lu) + LN'(N) \\
&\quad + N''(Lu, Lu) + 2N''(Lu, N) + N'(L^2u) \\
&\quad + N'(LN) + N''(N, N) + N'N'(Lu) + N'N'(N),
\end{aligned}$$

etc. The exact solution of (1.1) has a formal expansion

$$u(h) = \sum_{q=0}^{\infty} \frac{h^q}{q!} \left.\frac{\mathrm{d}^q}{\mathrm{d}h^q}\right|_{h=0} u(h),$$

where each term in the qth derivative corresponds in an obvious way to a rooted bicolored tree. Let for instance $\bullet \sim F(\bullet) = N(u)$ and $\circ \sim F(\circ) = Lu$ be the two trees with one node. Next, define B_+ as the operation which takes a finite set of trees $\{\tau_1, \dots, \tau_\mu\}$ and connects their roots to a new common black root. Similarly, $\tau = W_+(\tau')$ connects the root of τ' to a new white root resulting in the tree τ associated to $F(\tau) = L \cdot F(\tau')$. It suffices here to allow W_+ to act on a single tree and not on a set of trees. To each tree τ with q nodes formed this way, there exists precisely one term, $F(\tau)$ called an elementary differential, in the qth derivative of the solution of (1.1). For $q > 1$ it is defined recursively as

$$F(B_+(\tau_1, \dots, \tau_\mu))(u) = N^{(\mu)}(F(\tau_1), \dots, F(\tau_\mu))(u) \tag{2.1}$$

$$F(W_+(\tau'))(u) = LF(\tau')(u). \tag{2.2}$$

We may denote by T the set of all bicolored trees such that each white node has at most one child, and set $T = T_b \cup T_w$ the union of trees with black and white roots, respectively. Introducing the empty set $\emptyset$, and using the convention $B_+(\emptyset) = \bullet$, $W_+(\emptyset) = \circ$, we may write

$$T \cup \emptyset = \bigcup_{m \geq 0} W_+^m(T_b \cup \emptyset), \quad T_w = \bigcup_{m \geq 1} W_+^m(T_b \cup \emptyset). \tag{2.3}$$

The same bicolored trees used here also appear in the linearly implicit W-methods; see Steihaug and Wolfbrandt [21] as well as the text [10] by Hairer and Wanner. Following, for instance, the text by Hairer, Lubich, and Wanner [8], we may work with formal B-series. For an arbitrary map $\boldsymbol{c} : T \cup \emptyset \to \mathbf{R}$, we let the formal series

$$B(\boldsymbol{c}, u) = \boldsymbol{c}(\emptyset)u + \sum_{\tau \in T} \frac{h^{|\tau|}}{\sigma(\tau)} \boldsymbol{c}(\tau) F(\tau)(u) \tag{2.4}$$

be a B-series, where $\sigma(\tau)$ is the symmetry coefficient defined as $\sigma(\bullet) = \sigma(\circ) = 1$, and for $\tau = B_+(\tau_1, \ldots, \tau_\mu)$,

$$\sigma(\tau) = \sigma(\tau_1) \cdots \sigma(\tau_\mu)\, m_1! \cdot m_2! \cdots ,$$

where the m_is count the number of equal trees among $\tau_1, \ldots, \tau_\mu$.

The further derivation of order conditions is based on the assumption that both the exact and numerical solution possess B-series of the form (2.4), say $B(\boldsymbol{e}, u_0)$ and $B(\boldsymbol{u}_1, u_0)$, respectively. We refer to [1] for details, and present only the final result.

Theorem 2.1. *Let $T' \subset T$ be the set of bicolored rooted trees such that every white node has precisely one child. An exponential integrator defined by* (1.2)–(1.3) *has order of consistency p if*

$$\boldsymbol{u}_1(\tau) = \frac{1}{\gamma(\tau)} \quad \textit{for all } \tau \in T' \textit{ such that } |\tau| \leq p,$$

where

$$\begin{aligned} \boldsymbol{u}_1(\emptyset) = \boldsymbol{U}_r(\emptyset) &= 1, \quad 1 \leq r \leq s, \\ \boldsymbol{u}_1(W_+^m B_+(\tau_1, \ldots, \tau_\mu)) &= \sum_{r=1}^{s} \beta^{r,m} \boldsymbol{U}_r(\tau_1) \cdots \boldsymbol{U}_r(\tau_\mu) \\ \boldsymbol{U}_r(W_+^m B_+(\tau_1, \ldots, \tau_\mu)) &= \sum_{j=1}^{s} \alpha_r^{j,m} \boldsymbol{U}_j(\tau_1) \cdots \boldsymbol{U}_j(\tau_\mu). \end{aligned}$$

The trees in T' with at most four nodes are listed in Table 2. Note that even though all trees in the set T feature in the B-series for the exact and numerical solutions, it suffices to consider a subset T' consisting of all trees in T except those with a terminal white node. There is an interesting connection between the set of trees T' and the trees used to develop the order theory for composition methods in [19]. White nodes appear as connected strings of nodes which, except for the root, have exactly one parent and one child, and always terminate in a black node. Therefore one can remove all white nodes and assign to the terminating black node the number of removed nodes plus one. Black nodes not connected to a white node are assigned the number one. These multilabelled trees are precisely those appearing in [19], they can also be identified as the set of rooted trees of nonempty sets. The generating function for these trees is well known,

$$M(x) = \frac{x}{1-x}\ \exp\left(M(x) + \frac{M(x^2)}{2} + \frac{M(x^3)}{3} + \cdots\right).$$

The number of order conditions for each order 1 to 9 is 1, 2, 5, 13, 37, 108, 332, 1042, and 3360.

3. Construction of exponential integrators.

The schemes of Lawson [15] are exponential integrators derived simply by introducing a change of variable, $w(t) = \mathrm{e}^{-tL}\, u(t)$ in (1.1), and by applying a standard Runge–Kutta scheme to the resulting ODE. This approach results in a formula for w_1 in terms of w_0. By setting $u_n = \mathrm{e}^{tL}\, w_n$ one gets a scheme of the form (1.2)–(1.3) in which

$$a_r^j(z) = \alpha_r^{j,0}\,\mathrm{e}^{(c_r - c_j)z} \quad \text{and} \quad b^r(z) = \beta^{r,0}\,\mathrm{e}^{(1-c_r)z},$$

as noted by Lawson in [15].

This scheme has order p if the underlying scheme determined by $\alpha_r^{j,0}$ and $\beta^{r,0}$ is of order p. This gives us a very useful tool for constructing exponential integrators with given underlying Runge–Kutta schemes. We express this in the following proposition.

PROPOSITION 3.1. *Suppose that the coefficients $\alpha_r^{j,0}$ and $\beta^{r,0}$, $1 \le r, j \le s$ define a Runge–Kutta scheme of order p. Then, any exponential integrator of the form* (1.2)–(1.3) *satisfying*

$$\alpha_r^{j,m} = \frac{1}{m!}(a_r^j)^{(m)}(0) = \frac{1}{m!}\alpha_r^{j,0}(c_r - c_j)^m, \quad 0 \le m \le p-2, \tag{3.1}$$

$$\beta^{r,m} = \frac{1}{m!}(b_r^j)^{(m)}(0) = \frac{1}{m!}\beta^{r,0}(1 - c_r)^m, \quad 0 \le m \le p-1, \tag{3.2}$$

is of order p. In the above expression we use $0^0 := 1$.

Proof. Order conditions for exponential integrators of order p involve $\alpha_r^{j,m}$ for $0 \le m \le p-2$ and $\beta^{r,m}$ for $0 \le m \le p-1$. On the other hand, the Lawson schemes must satisfy the order conditions for exponential integrators, and their values for these coefficients are precisely those specified in the proposition. □

It is convenient to introduce finite dimensional function spaces V_a and V_b to which the respective coefficient functions $a_r^j(z)$ and $b^r(z)$ will belong. For the purpose of calculations, it is also useful to work with basis functions ψ_k for these spaces,

$$a_r^j(z) = \sum_{k=0}^{K_a - 1} A_r^{j,k}\psi_k(z) \quad \text{and} \quad b^r(z) = \sum_{k=0}^{K_b - 1} B^{r,k}\psi_k(z), \tag{3.3}$$

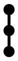
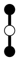
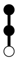
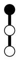
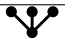
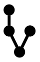
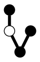
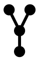
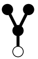
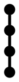
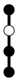
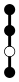
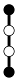
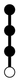
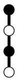
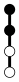
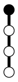
where $K_a = \dim(V_a)$ and $K_b = \dim(V_b)$. There is a technical assumption that we will adopt to the end of this note.

ASSUMPTION 3.2. *Any finite dimensional function space V of dimension K used for coefficient functions $a_r^j(z)$ or $b^r(z)$ has the property that the map from V to $\mathbf{R}^K$ defined by*

$$f \mapsto\ (f(0), f'(0), \ldots, f^{(K-1)}(0))^T$$

is injective. Equivalently, any function in V is uniquely determined by its first K Taylor coefficients.

TABLE 2
Trees, elementary differentials and coefficients for $\tau \in T'$ with $|\tau| \leq 4$.

	$\|\tau\|$	Tree	$F(\tau)$	$\gamma(\tau)$	$\boldsymbol{u}_1(\tau)$	$\sigma(\tau)$
1	1		N	1	$\sum_r \beta^{r,0}$	1
2	2		$N'N$	2	$\sum_r \beta^{r,0} c_r$	1
3	2		LN	2	$\sum_r \beta^{r,1}$	1
4	3		$N''(N,N)$	3	$\sum_r \beta^{r,0} c_r^2$	2
5	3		$N'N'N$	6	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,0} c_j$	1
6	3		$N'(LN)$	6	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,1}$	1
7	3		$LN'N$	6	$\sum_r \beta^{r,1} c_r$	1
8	3		L^2N	6	$\sum_r \beta^{r,2}$	1
9	4		$N'''(N,N,N)$	4	$\sum_r \beta^{r,0} c_r^3$	6
10	4		$N''(N'N,N)$	8	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,0} c_j c_r$	1
11	4		$N''(LN,N)$	8	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,1} c_r$	1
12	4		$N'N''(N,N)$	12	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,0} c_j^2$	2
13	4		$LN''(N,N)$	12	$\sum_r \beta^{r,1} c_r^2$	2
14	4		$N'N'N'N$	24	$\sum_{r,j,k} \beta^{r,0} \alpha_r^{j,0} \alpha_j^{k,0} c_k$	1
15	4		$N'N'(LN)$	24	$\sum_{r,j,k} \beta^{r,0} \alpha_r^{j,0} \alpha_j^{k,1}$	1
16	4		$N'(LN'N)$	24	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,1} c_j$	1
17	4		$N'(L^2N)$	24	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,2}$	1
18	4		$LN'N'N$	24	$\sum_{r,j} \beta^{r,1} \alpha_r^{j,0} c_j$	1
19	4		$LN'(LN)$	24	$\sum_{r,j} \beta^{r,1} \alpha_r^{j,1}$	1
20	4		$L^2N'N$	24	$\sum_r \beta^{r,2} c_r$	1
21	4		L^3N	24	$\sum_r \beta^{r,3}$	1

3.1. Deriving schemes with natural continuous extensions. The approach of Krogstad in [14] is to approximate the nonlinear function $N(u(t_0 + \theta h))$, $0 < \theta < 1$

with a polynomial in θ. Assuming that the functions $a_r^j(z)$ for the internal stages are given, one lets $N(u(t_n+\theta h))$ be approximated by

$$\bar{N}(t_0+\theta h)=\sum_{r=1}^{s} w_r'(\theta)N_r, \tag{3.4}$$

where $N_r = N(U_r)$ are the stage derivatives and $w_r(\theta)$ are polynomials of degree d, with $w(0) = 0$, such that $\bar{N}(t_0+\theta h)$ approximates $N(u(t_0+\theta h))$ uniformly for $0<\theta<1$ to a given order. Replacing the exact problem with the approximate one, $\dot{v}=Lv+\bar{N}(t),\ v(t_0)=u_0$ one finds

$$u_1 := v(t_0+h)=\mathrm{e}^{hL}u_0+\sum_{r=1}^{s} b^r(hL)N_r, \quad \text{where } b^r(z)=\int_0^1 \mathrm{e}^{(1-\theta)z}w_r'(\theta)\,\mathrm{d}\theta;$$

we then define the functions

$$\phi_k(z)=\int_0^1 \mathrm{e}^{(1-\theta)z}\theta^k\,\mathrm{d}\theta,\ k=0,1,\dots. \tag{3.5}$$

Thus, here the function space $V_b=\operatorname{span}\{\phi_0,\dots,\phi_{d-1}\}$, so $\psi_k=\phi_k$ and $K_b=d$ in (3.3). Cox and Matthews [5] presented a fourth order scheme using these basis functions with $K_b=3$. Krogstad [14] also derived a variant of their method by using a continuous extension as just explained. In [22] Zennaro developed a theory which generalizes the collocation polynomial idea to arbitrary Runge–Kutta schemes. The approach was called natural continuous extensions (NCE). By making a slight modification to the approach of Zennaro, one can find a useful way of deriving exponential integrators as well as providing them with a continuous extension.

Suppose $w_1(\theta),\dots,w_s(\theta)$ are given polynomials of degree d, and that the stage derivatives $N_1,\dots,N_s$ of an exponential integrator are given from (1.2). We define the $d-1$ degree polynomial $\bar{N}(t)$ by (3.4).

Definition 3.3. *We call $\bar{N}(t)$ of* (3.4) *a natural continuous n-extension (NCNE) of degree d of the exponential integrator* (1.2)–(1.3) *if*

1. $$w_r(0)=0,\quad w_r(1)=b^r(0),\qquad r=1,\dots,s,$$

2. $$\max_{t_0\le t\le t_1}|N(u(t))-\bar{N}(t)|=\mathcal{O}(h^{d-1}), \tag{3.6}$$

   where $u(t)$ is the exact solution of (1.1) *satisfying $u(t_0)=u_0$;*

3. $$\int_{t_0}^{t_1} G(t)(N(u(t))-\bar{N}(t))\,\mathrm{d}t=\mathcal{O}(h^{p+1}) \tag{3.7}$$

   for every smooth matrix-valued function $G(t)$.

It is important to note that the polynomial $\bar{N}(t)$ only depends on the stages N_r and the weights $b^r(0)=\beta^{r,0}$ corresponding to the underlying Runge–Kutta scheme. We also observe that since the $w_r(\theta)$ does not depend on L, an NCNE as defined above is also an NCE in the sense of Zennaro for the system $\dot{u}=N(u)$. Before discussing the existence of NCNEs, we motivate their usefulness in designing exponential integrators. Suppose an underlying Runge–Kutta method has been chosen, and that an NCNE has been found. Then we can determine the functions $b^r(z)$ in order to obtain an exponential Runge–Kutta method of the same order as the underlying scheme.

THEOREM 3.4. *If $\bar{N}(t)$ defined from* (3.4) *is an NCNE of degree d for a pth order scheme, then the functions*

$$b^r(z) = \int_0^1 \mathrm{e}^{(1-\theta)z} w'(\theta)\,\mathrm{d}\theta = \beta^{r,0} + z\int_0^1 \mathrm{e}^{(1-\theta)z} w(\theta)\,\mathrm{d}\theta,$$

define the weights of an exponential integrator of order p.

Proof. The exponential integrator we consider is obtained by replacing (1.1) by

$$\dot{v} = Lv + \bar{N}(t), \qquad v(t_0) = u_0 \tag{3.8}$$

over the interval $[t_0, t_1]$ and by solving (3.8) exactly. We subtract (3.8) from (1.1) to obtain

$$\dot{u} - \dot{v} = L(\dot{u} - \dot{v}) + \big(N(u) - \bar{N}(t)\big).$$

We may solve this equation to obtain

$$u(t_1) - v(t_1) = \int_{t_0}^{t_1} \mathrm{e}^{(t_1-t)L}\big(N(u(t)) - \bar{N}(t)\big)\,\mathrm{d}t = \mathcal{O}(h^{p+1}),$$

the last equality is thanks to (3.7). □

A reinterpretation of a result by Zennaro [22] combined with Proposition 3.1 leads to the following theorem.

THEOREM 3.5. *Suppose that an underlying Runge–Kutta scheme with coefficients $\alpha_r^{j,0}$ and $\beta^{r,0}$ of order p is given. Then it is possible to find a set of coefficient functions $a_r^j(z)$ with $a_r^j(0) = \alpha_r^{j,0}$ such that an NCNE of degree $d = \lfloor\frac{p+1}{2}\rfloor$ exists. Moreover, if $\bar{N}(t)$ is a NCNE of degree d, then*

$$\left\lfloor \frac{p+1}{2} \right\rfloor \le d \le \min(\nu^*, p),$$

where ν^* is the number of distinct elements among $c_1, \dots, c_s$.

COROLLARY 3.6. *For every underlying Runge–Kutta scheme, there exists an exponential integrator whose coefficient functions $b^r(z)$ are in the linear span of the functions $\{\phi_0(z), \dots, \phi_{d-1}(z)\}$, where $d = \lfloor\frac{p+1}{2}\rfloor$.*

Note, in particular, that one can derive fourth order exponential integrators using linear combinations of just $\phi_0(z)$ and $\phi_1(z)$ for $b^r(z)$, which is one less than what Cox and Matthews used; we present a specific example in section 4.

3.2. Lower bounds for K_a and K_b. We start establishing lower bounds for the number of necessary basis functions ψ_k by proving an ancillary result.

LEMMA 3.7. *Let $q \ge 0$ be an integer. The matrix $T_q \in \mathbf{R}^{d\times d}$ with elements*

$$(T_q)_{m+1,k+1} = \frac{1}{(q+m+k+1)!}, \quad 0 \le m, k \le d-1$$

is invertible.

Proof. Let $w = (w_1, \ldots, w_d)^T \in \mathbf{R}^d$ be arbitrary, and consider the polynomial

$$p(x) = \sum_{k=0}^{d-1} w_{k+1} \frac{x^{q+d+k}}{(q+d+k)!}.$$

We compute

$$p^{(d-m-1)}(1) = \sum_{k=0}^{d-1} w_{k+1} \frac{1}{(q+m+k+1)!} = (T_q w)_{m+1}, \quad 0 \le m \le d-1.$$

So $T_q w = 0$ is equivalent to $p^{(j)}(1) = 0$ for $0 \le j \le d-1$. Since $p(x)$ is of the form $x^{q+d} r(x)$ where $r(x)$ is a polynomial of degree at most $d-1$, it follows that $p(x) \equiv 0$ so that $w = 0$. □

As $\phi_k^{(m)}(0) = m!k!/(m+k+1)!$ for ϕ_k defined by (3.5), we get as an immediate consequence of this lemma that the function spaces $V = \mathrm{span}(\phi_q, \ldots, \phi_{q+K-1})$, $q \ge 0$ satisfy Assumption 3.2.

Theorem 3.8. *For an exponential integrator of order p, the dimension of the function spaces V_a and V_b are bounded from below as follows:*

$$K_a = \dim V_a \ge \left\lfloor \frac{p}{2} \right\rfloor, \quad K_b = \dim V_b \ge \left\lfloor \frac{p+1}{2} \right\rfloor. \tag{3.9}$$

Proof. We will show that using smaller values of K_a or K_b than dictated by (3.9) is incompatible with the order conditions for a scheme of order p. Let V_a and V_b be arbitrary function spaces, satisfying Assumption 3.2, let V denote either of them, and let $d = \dim V$. If $f \in V$, then there are numbers $w_0, \ldots, w_{d-1}$ such that

$$f^{(d)}(0) = \sum_{m=0}^{d-1} w_m \, f^{(m)}(0). \tag{3.10}$$

Suppose now that $d_a := \dim V_a = \lfloor p/2 \rfloor - 1$ and $d_b := \dim V_b = \lfloor (p+1)/2 \rfloor - 1$.

Consider the bicolored trees $\tau_q^{m,k}$ defined by

$$\tau_q^{m,k} = B_+^q \big(W_+^m B_+ (\overbrace{\bullet, \ldots, \bullet}^{k}) \big)$$

which consist of a string of $q \ge 0$ black nodes followed by a string of $m > 0$ white nodes with a bushy tree of $k+1$ black nodes grafted onto the topmost leaf of the white nodes. We shall use these trees with $q = 0$ for proving the bound on K_b and with $q = 1$ for K_a. The density of $\tau_q^{m,k}$ is given by

$$\gamma(\tau_q^{m,k}) = \frac{(q+m+k+1)!}{\mathrm{k}!}.$$

The trees corresponding to order conditions for a scheme of order p have at most p nodes, $|\tau_q^{m,k}| = q+m+k+1 \le p \Rightarrow 0 \le k \le p-m-1-q$. The definition of d_a and d_b implies that $p-2 \ge 2d_a$ and $p-1 \ge 2d_b$. If we set $q = 1$, $m = d_a$ we thus obtain conditions for $0 \le k \le d_a$, whereas $q = 0$, $m = d_b$ results in $0 \le k \le d_b$.

The conditions corresponding to $\tau_1^{d_a,k}$ can be expressed as

$$\frac{1}{d_a!} \sum_{r,j=1}^{s} \beta^{r,0} (a_r^j)^{(d_a)}(0) c_j^k = \frac{k!}{(d_a+k+2)!}, \quad 0 \le k \le d_a$$

which, upon insertion of $(a_r^j)^{(d_a)}(0) = \sum w_m (a_r^j)^{(m)}(0)$ as in (3.10), yields

$$\frac{k!\,d_a!}{(k+d_a+2)!} = \sum_{m=0}^{d_a-1} w_m \left(\sum_{r=1}^{s} \beta^{r,0} (a_r^j)^{(m)}(0) c_j^k \right) = \sum_{m=0}^{d_a-1} w_m \, \frac{m!\,k!}{(m+k+2)!}.$$

The conditions for $\tau_0^{d_b,k}$ similarly yield

$$\frac{k!\,d_b!}{(k+d_b+1)!} = \sum_{m=0}^{d_b-1} w_m \left(\sum_{r=1}^{s} (b^r)^{(m)}(0)\, c_r^k \right) = \sum_{m=0}^{d_b-1} w_m \frac{m!\,k!}{(m+k+1)!}.$$

In both cases ($d = d_a$ or d_b), we end up with a $(d+1) \times d$ linear system of equations for determining w_m, $m = 0, \ldots, d-1$. This system is of the form

$$\sum_{m=0}^{d-1} \frac{m!\,k!}{(q+m+k+1)!}\, w_m = \frac{k!\,d!}{(q+k+d+1)!}, \quad 0 \le k \le d$$

for $q \in \{0, 1\}$ and is solvable only if the matrix with elements

$$(T_q)_{m+1,k+1} = \frac{m!\,k!}{(q+m+k+1)!}, \quad 0 \le m, k \le d$$

is singular. However, Lemma 3.7 implies that the matrix T_q is invertible so the linear system is inconsistent. It is hence not possible to choose $K_a = d_a$ or $K_b = d_b$. □

Some remarks regarding the implications of Theorem 3.8 are in order. First, note that the bounds in the theorem are not proved to be sharp; however, Theorem 3.5 ensures that the lower bound is attainable for the dimension of V_b if a basis is given by the functions ϕ_k of (3.5). However, this result does not apply to the space V_a of the functions $a_r^j(z)$. For instance, in the case $p = 5$, one can prove that it is indeed possible to take $K_a = 2$, but V_a cannot be the span of ϕ_0 and ϕ_1. But an example of a feasible two-dimensional space is that with basis $\psi_0(z) = \phi_1(z)$ and $\psi_1(z) = \phi_1(\frac{3}{5}z)$. A particular scheme is given in Table 4, though the usefulness of the bounds are questionable in this particular example. Using say $V_b = \mathrm{span}\{\phi_0, \phi_1, \phi_2\}$ combined with the above choice of V_a requires the computation with a total of 4 basis functions, whereas only 3 are necessary if one instead chooses $V_a = V_b$.

Furthermore, we note that the minimum attainable value of the parameters K_a and K_b depend only on the order p of the underlying Runge–Kutta scheme and the choice of the basis functions ψ_k. Specifically, the coefficients of the underlying Runge–Kutta scheme do not influence the minimum values of K_a and K_b.

4. Examples of exponential integrators.

In this section we will present examples of exponential integrators. For fourth order methods, one will notice that some well-known schemes are obtained for particular choices of the free parameters, suggesting that a search on the entire space of parameters may result in schemes which in some sense may have better properties than the known methods. The scheme of order 5 presented at the end is only included as an illustration of the proposed procedure for solving the order conditions. It remains a subject of future research to establish to which extent higher order exponential integrators are useful for practical purposes.

The procedure we have used in constructing schemes may be summarized as follows:

1. Choose an underlying Runge–Kutta scheme. This determines $\alpha_r^{j,0}$ and $\beta^{r,0}$.

TABLE 3
Coefficient function for a fourth order ETD scheme with classical RK4 as underlying scheme. Basis functions given by (3.5).

$$
\begin{aligned}
a_2^1(z) &= -(\tfrac{1}{2} + \rho_1)\phi_0(z) + (2\rho_1 + 2)\phi_1(z)\\
a_3^1(z) &= (1 + \rho_1 - \tfrac{1}{4}(\rho_2 + \rho_3))\phi_0(z) + (-2 - 2\rho_1 + \tfrac{1}{2}(\rho_2 + \rho_3))\phi_1(z)\\
a_3^2(z) &= (-1 + \tfrac{1}{4}(\rho_2 + \rho_3))\phi_0(z) + (3 - \tfrac{1}{2}(\rho_2 + \rho_3)\phi_1(z)\\
a_4^1(z) &= \tfrac{1}{2}(\rho_2 + \rho_3)\phi_0(z) - (\rho_2 + \rho_3)\phi_1(z)\\
a_4^2(z) &= -\frac{\rho_2}{2}\phi_0(z) + \rho_2\phi_1(z)\\
a_4^3(z) &= (1 - \tfrac{1}{2}\rho_3)\phi_0(z) + \rho_3\phi_1(z)
\end{aligned}
$$

$$
\begin{aligned}
b^1(z) &= (1 + \gamma_2)\,\phi_0(z) + (-3 - 6\gamma_2)\,\phi_1(z) + (6\gamma_2 + 2)\,\phi_2(z)\\
b^2(z) &= (-\gamma_1 - 2\gamma_2)\,\phi_0(z) + (6\gamma_1 + 12\gamma_2 + 2)\phi_1(z) + (-6\gamma_1 - 12\gamma_2 - 2)\,\phi_2(z)\\
b^3(z) &= \gamma_1\,\phi_0(z) + (-6\gamma_1 + 2)\,\phi_1(z) + (6\gamma_1 - 2)\,\phi_2(z)\\
b^4(z) &= \gamma_2\,\phi_0(z) + (-6\gamma_2 - 1)\,\phi_1(z) + (6\gamma_2 + 2)\,\phi_2(z)
\end{aligned}
$$

2. Choose basis functions $\psi_k(z)$ for the coefficient functions and determine K_a and K_b.
3. Use the order conditions for the trees of the form $W_+^m(\tau_C)$, where τ_C is a tree with only black nodes, and determine $\beta^{r,m}$, for $1 \leq m \leq K_b - 1$; see also (3.2).
4. Identify order conditions which are linear in $c'_r = \sum_{j=1}^s \alpha_r^{j,1}$ and which otherwise depend only on $\beta_r^{j,m}$, $0 \leq m \leq K_b - 1$ and $\alpha_r^{j,0}$, and solve for c'_r.
5. Identify remaining conditions which depend linearly on $\alpha_r^{j,1}$. Solve for $\alpha_r^{j,1}$ together with $c'_r = \sum_{j=1}^s \alpha_r^{j,1}$. Repeat this procedure to solve for $\alpha_r^{j,m}$, $2 \leq m \leq K_a - 1$.
6. $\beta^{r,m}$ are now uniquely determined for $m \geq K_b$ and $\alpha_r^{j,m}$ for $m \geq K_a$ by (3.10). Verify all remaining order conditions for $\beta^{r,m}$, $K_b \leq m \leq p - 1$ and for $\alpha_r^{j,m}$, $K_a \leq m \leq p-2$. If inconsistencies appear, the basis functions are not feasible.
7. Verify all remaining order conditions.

In most cases we have considered, once $\alpha_r^{j,0}$ and $\beta^{r,0}$ have been chosen, one can find the remaining $\alpha_r^{j,m}$ independently of the $\beta^{r,m}$. Most of the exponential integrators we find in the literature are based on the classical fourth order scheme of Kutta, and it is typical that one can combine $a_r^j(z)$ from one scheme with $b^r(z)$ from another scheme and still get overall order four.

In the class of ETD schemes, proposed by Cox and Matthews in [5] and Krogstad in [14], the space V_b is spanned by the three functions ϕ_0, ϕ_1, and ϕ_2 of (3.5). However, in the former reference, $\dim V_a = 2$ with a basis $\{\phi_0(z/2), z\phi_0(z/2)^2\}$. This V_a coincides with the one used in [3] given in Table 1(a).

Another choice is to use $\phi_k(z)$ of (3.5) both for V_a and V_b. In Table 3 we characterise all resulting schemes with $K_a = 2$ and $K_b = 3$. It is interesting to note that Theorem 3.8 predicts $K_a \geq 2$ and $K_b \geq 2$, and indeed, by choosing $\gamma_1 = \frac{1}{3}$ and $\gamma_2 = -\frac{1}{3}$, we see that ϕ_2 disappears from the $b^r(z)$-functions. Choosing $\gamma_1 = \gamma_2 = 0$, we recover the $b^r(z)$-functions obtained in [5].

Letting V_b be spanned by $\psi_0(z) = \phi_0(z)$ and $\psi_1(z) = \phi_0(z/2)$, one obtains the

TABLE 4

Coefficient functions for a fifth order exponential integrator with Fehlberg's fifth order RK as the underlying scheme. Here $a_i^j(z) = a_i^j\phi_1(z) + \hat{a}_i^j\phi_1(\frac{3}{5}z)$.

$$\begin{array}{c|cccccc}
0 & & & & & & \\
\frac{2}{9} & a_2^1(z) & & & & & \\
\frac{1}{3} & a_3^1(z) & a_3^2(z) & & & & \\
\frac{3}{4} & a_4^1(z) & a_4^2(z) & a_4^3(z) & & & \\
1 & a_5^1(z) & a_5^2(z) & a_5^3(z) & a_5^4(z) & & \\
\frac{5}{6} & a_6^1(z) & a_6^2(z) & a_6^3(z) & a_6^4(z) & a_6^5(z) & \\
\hline
 & b^1(z) & b^2(z) & b^3(z) & b^4(z) & b^5(z) & b^6(z)
\end{array}
\quad , \quad
\begin{array}{l}
b^1(z) = \frac{47}{150}\phi_0 - \frac{188}{75}\phi_1 + \frac{47}{15}\phi_2 \\
b^2(z) = 0 \\
b^3(z) = -\frac{43}{25}\phi_0 + \frac{132}{5}\phi_1 - 33\phi_2 \\
b^4(z) = \frac{4124}{75}\phi_0 - \frac{6152}{15}\phi_1 + \frac{1352}{3}\phi_2 \\
b^5(z) = \frac{189}{10}\phi_0 - \frac{662}{5}\phi_1 + 142\phi_2 \\
b^6(z) = -\frac{1787}{25}\phi_0 + \frac{12966}{25}\phi_1 - \frac{2814}{5}\phi_2
\end{array}$$

(i,j)	$(2,1)$	$(3,1)$	$(3,2)$	$(4,1)$	$(4,2)$	$(4,3)$	$(5,1)$	$(5,2)$	$(5,3)$	$(5,4)$	$(6,1)$	$(6,2)$	$(6,3)$	$(6,4)$	$(6,5)$
a_i^j	$-\frac{2}{3}$	$\frac{569}{11544}$	$-\frac{831}{3848}$	$-\frac{77157}{61568}$	$\frac{587979}{61568}$	$-\frac{405}{64}$	$\frac{655263}{7696}$	$-\frac{1148769}{7696}$	$\frac{1593}{40}$	$\frac{144}{5}$	$-\frac{2212835}{277056}$	$\frac{477285}{30784}$	$-\frac{39}{16}$	$-\frac{4}{9}$	$-\frac{185}{96}$
$\hat{a}_i^j$	$\frac{10}{9}$	$\frac{1355}{11544}$	$\frac{2755}{3848}$	$\frac{143535}{61568}$	$-\frac{821745}{61568}$	$\frac{675}{64}$	$-\frac{2031205}{23088}$	$\frac{1252665}{7696}$	$-\frac{405}{8}$	$-\frac{80}{3}$	$\frac{6888625}{831168}$	$-\frac{496525}{30784}$	$\frac{65}{16}$	$\frac{20}{27}$	$\frac{575}{288}$

unique solution

$$\begin{aligned}
b^1(z) &= \tfrac{1}{2}\phi_0(z) - \tfrac{1}{3}\phi_0(z/2) \\
b^2(z) &= b^3(z) = \tfrac{1}{3}\phi_0(z) \qquad (4.1)\\
b^4(z) &= -\tfrac{1}{6}\phi_0(z) + \tfrac{1}{3}\phi_0(z/2).
\end{aligned}$$

These weights coincide with the ones derived in the fourth order scheme in [3] given in Table 1. Yet another choice is to let V_b consist of functions of the form $p(z)\phi_0(z)$, where $p(z)$ is a polynomial of degree 1, and we recover $b^r(z)$ as in Table 1(b).

Finally, we give an example of a fifth order exponential integrator based on a scheme of Fehlberg. As indicated in section 3.2, we take $\dim V_a = 2$ with basis $\psi_0(z) = \phi_1(z)$ and $\psi_1(z) = \phi_1(\frac{3}{5}z)$. For V_b we use the basis $\psi_k(z) = \phi_k(z)$ for $k = 0, 1, 2$. The resulting coefficient functions are given in Table 4.

In summary, this paper presents a complete order theory for exponential integrators of the form (1.2)–(1.3). From deriving order conditions by means of bicolored trees to proving bounds for the lowest possible number of basis functions, the results presented herein provide a general framework for constructing schemes of this type. A number of issues are, however, not addressed in the present paper. These include systematically choosing basis functions ψ_k, and how to construct schemes with low error constants.

Exponential integrators are interesting from the point of view of handling unbounded or stiff operators, yet the order theory does not say anything about what happens for large eigenmodes of L in (1.1). Determining conditions for favorable behavior in light of such operators should be an arena for future work.

Acknowledgments. Thanks to Christian Bower for advice regarding the enumeration of bicolored trees, and to an anonymous referee for suggesting a simpler proof of Lemma 3.7.

REFERENCES

[1] H. Berland, B. Owren, and B. Skaflestad, *B-series and order conditions for exponential integrators*, Technical report 5/04, The Norwegian University of Science and Technology, Trondheim, Norway, 2004, http://www.math.ntnu.no/preprint/.

[2] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, John Wiley & Sons, Chichester, UK, 2003.

[3] E. Celledoni, A. Marthinsen, and B. Owren, *Commutator-free Lie group methods*, FGCS, 19 (2003), pp. 341–352.

[4] J. Certaine, *The solution of ordinary differential equations with large time constants*, in Mathematical Methods for Digital Computers, Wiley, New York, 1960, pp. 128–132.

[5] S. M. Cox and P. C. Matthews, *Exponential Time Differencing for Stiff Systems*, J. Comput. Phys., 176 (2002), pp. 430–455.

[6] B. L. Ehle and J. D. Lawson, *Generalized Runge–Kutta processes for stiff initial-value problems*, J. Inst. Math. Appl., 16 (1975), pp. 11–21.

[7] A. Friedli, *Verallgemeinerte Runge–Kutta Verfahren zur Lösung steifer Differentialgleichungssysteme*, in Numerical Treatment of Differential Equations (Proc. Conf., Math. Forschungsinst., Oberwolfach, 1976), Lecture Notes in Math., Vol. 631, Springer, Berlin, 1978, pp. 35–50.

[8] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, Structure-preserving algorithms for ordinary differential equations, Springer Series in Computational Mathematics, 31, Springer-Verlag, Berlin, 2002.

[9] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving ordinary differential equations.* I. Nonstiff problems, Springer Series in Computational Mathematics, 8, 2nd ed., Springer-Verlag, Berlin, 1993.

[10] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations.* II. Stiff and differential-algebraic problems, Springer Series in Computational Mathematics, 14, 2nd ed., Springer-Verlag, Berlin, 1996.

[11] M. Hochbruck, C. Lubich, and H. Selhofer, *Exponential integrators for large systems of differential equations*, SIAM J. Sci. Comput., 19 (1998), pp. 1552–1574.

[12] M. Hochbruck and A. Ostermann, *Explicit exponential Runge–Kutta methods for semilinear parabolic problems*, SIAM J. Numer. Anal., (2005), to appear.

[13] A.-K. Kassam and L. N. Trefethen, *Fourth-order time stepping for stiff PDEs*, SIAM J. Sci. Comput., 26 (2005), pp. 1214–1233.

[14] S. Krogstad, *Generalized integrating factor methods for stiff PDEs*, J. Comput. Phys., 203 (2005), pp. 72–88.

[15] J. D. Lawson, *Generalized Runge–Kutta processes for stable systems with large Lipschitz constants*, SIAM J. Numer. Anal., 4 (1967), pp. 372–380.

[16] B. Minchev and W. M. Wright, *A review of exponential integrators for semilinear problems*, Technical report 2/05, The Norwegian University of Science and Technology, Trondheim, Norway, 2005, http://www.math.ntnu.no/preprint/.

[17] H. Munthe-Kaas, *High order Runge–Kutta methods on manifolds*, in Proceedings of the NSF/CBMS Regional Conference on Numerical Analysis of Hamiltonian Differential Equations (Golden, CO, 1997), Appl. Numer. Math., 29 (1999), pp. 115–127.

[18] H. Munthe-Kaas and B. Owren, *Computations in a free Lie algebra*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 957–981.

[19] A. Murua and J. M. Sanz-Serna, *Order conditions for numerical integrators obtained by composing simpler integrators*, R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci., 357 (1999), pp. 1079–1100.

[20] S. P. Nørsett, *An A-stable modification of the Adams–Bashforth methods*, in Conference on Numerical Solution of Differential Equations (Dundee, Scotland, 1969), Springer, Berlin, 1969, pp. 214–219.

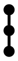
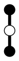
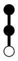
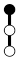
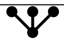
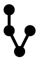
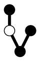
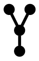
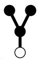
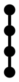
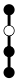
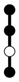
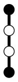
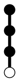
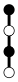
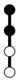
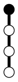
[21] T. Steihaug and A. Wolfbrandt, *An attempt to avoid exact Jacobian and nonlinear equations in the numerical solution of stiff differential equations*, Math. Comp., 33 (1979), pp. 521–534.

[22] M. Zennaro, *Natural continuous extensions of Runge–Kutta methods*, Math. Comp., 46 (1986), pp. 119–133.

CORRECTION TO "*B*-SERIES AND ORDER CONDITIONS FOR EXPONENTIAL INTEGRATORS"

There were errors in the trees in Table 2 in the published version of "***B***-Series and Order Conditions for Exponential Integrators." The correct table follows.

Table 2
Trees, elementary differentials and coefficients for $\tau \in T'$ with $|\tau| \leq 4$.

	$\lvert\tau\rvert$	Tree	$F(\tau)$	$\gamma(\tau)$	$u_1(\tau)$	$\sigma(\tau)$
1	1		N	1	$\sum_r \beta^{r,0}$	1
2	2		$N'N$	2	$\sum_r \beta^{r,0} c_r$	1
3	2		LN	2	$\sum_r \beta^{r,1}$	1
4	3		$N''(N,N)$	3	$\sum_r \beta^{r,0} c_r^2$	2
5	3		$N'N'N$	6	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,0} c_j$	1
6	3		$N'(LN)$	6	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,1}$	1
7	3		$LN'N$	6	$\sum_r \beta^{r,1} c_r$	1
8	3		L^2N	6	$\sum_r \beta^{r,2}$	1
9	4		$N'''(N,N,N)$	4	$\sum_r \beta^{r,0} c_r^3$	6
10	4		$N''(N'N,N)$	8	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,0} c_j c_r$	1
11	4		$N''(LN,N)$	8	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,1} c_r$	1
12	4		$N'N''(N,N)$	12	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,0} c_j^2$	2
13	4		$LN''(N,N)$	12	$\sum_r \beta^{r,1} c_r^2$	2
14	4		$N'N'N'N$	24	$\sum_{r,j,k} \beta^{r,0} \alpha_r^{j,0} \alpha_j^{k,0} c_k$	1
15	4		$N'N'(LN)$	24	$\sum_{r,j,k} \beta^{r,0} \alpha_r^{j,0} \alpha_j^{k,1}$	1
16	4		$N'(LN'N)$	24	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,1} c_j$	1
17	4		$N'(L^2N)$	24	$\sum_{r,j} \beta^{r,0} \alpha_r^{j,2}$	1
18	4		$LN'N'N$	24	$\sum_{r,j} \beta^{r,1} \alpha_r^{j,0} c_j$	1
19	4		$LN'(LN)$	24	$\sum_{r,j} \beta^{r,1} \alpha_r^{j,1}$	1
20	4		$L^2N'N$	24	$\sum_r \beta^{r,2} c_r$	1
21	4		L^3N	24	$\sum_r \beta^{r,3}$	1

SIAM J. Numer. Anal.
Vol. 43, No. 4, pp. 1728–1749

SUPERCONVERGENCE OF THE VELOCITY IN MIMETIC FINITE DIFFERENCE METHODS ON QUADRILATERALS[*]

M. BERNDT[†], K. LIPNIKOV[†], M. SHASHKOV[†], M. F. WHEELER[‡], AND I. YOTOV[§]

Abstract. Superconvergence of the velocity is established for mimetic finite difference approximations of second-order elliptic problems over h^2-uniform quadrilateral meshes. The superconvergence result holds for a full tensor coefficient. The analysis exploits the relation between mimetic finite differences and mixed finite element methods via a special quadrature rule for computing the scalar product in the velocity space. The theoretical results are confirmed by numerical experiments.

Key words. mixed finite element, mimetic finite difference, tensor coefficient, superconvergence

AMS subject classifications. 65N06, 65N12, 65N15, 65N22, 65N30

DOI. 10.1137/040606831

1. Introduction.

We consider the numerical approximation of a linear second-order elliptic problem. In porous medium applications, this equation models single phase Darcy flow and is usually written as a first-order system for the fluid pressure p and velocity $\mathbf{u}$:

$$\begin{aligned} \mathbf{u} &= -\mathbf{K}\,\mathrm{grad}\,p && \text{in} \quad \Omega, \\ \mathrm{div}\,\mathbf{u} &= f && \text{in} \quad \Omega, \\ \mathbf{u}\cdot\mathbf{n} &= g && \text{on} \quad \partial\Omega, \end{aligned} \tag{1.1}$$

where $\Omega \subset \Re^2$, $\mathbf{n}$ is the outward unit normal to $\partial\Omega$, and $\mathbf{K} \in \Re^{2\times 2}$ is a symmetric uniformly positive definite full tensor representing the rock permeability divided by the fluid viscosity. We assume that system (1.1) satisfies the compatibility condition

$$\int_\Omega f\,d\mathbf{x} + \int_{\partial\Omega} g\,ds = 0.$$

In this paper, we analyze the convergence of a mimetic finite difference (MFD) method on quadrilateral meshes. The method uses discrete operators that preserve certain critical properties of the original continuum differential operators. Conservation laws, solution symmetries, and the fundamental identities and theorems of vector

[*] Received by the editors April 15, 2004; accepted for publication (in revised form) February 7, 2005; published electronically November 22, 2005. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sinum/43-4/60683.html

[†] Mathematical Modeling and Analysis Group, Theoretical Division, Mail Stop B284, Los Alamos National Laboratory, Los Alamos, NM 87545 (berndt@lanl.gov, lipnikov@lanl.gov, shashkov@ lanl.gov). The research of these authors was supported by the U.S. Department of Energy under contract W-7405-ENG-36, LA-UR-03-7904.

[‡] ICES, Department of Aerospace Engineering and Engineering Mechanics, Department of Petroleum and Geosystems Engineering, and Department of Mathematics, The University of Texas at Austin, Austin, TX 78712 (mfw@ices.utexas.edu). The work of this author was partially supported by NSF grant EIA-0121523 and by NPACI grant UCSD 10181410.

[§] Department of Mathematics, 301 Thackeray Hall, University of Pittsburgh, Pittsburgh, PA 15260 (yotov@math.pitt.edu). The work of this author was partially supported by NSF grants DMS 0107389, DMS 0112239, and DMS 0411694 and by DOE grant DE-FG02-04ER25618.

and tensor calculus are examples of such properties. This "mimetic" technique has been applied successfully to several applications including diffusion [22, 15, 18], magnetic diffusion and electromagnetics [14], continuum mechanics [17], and gas dynamics [8]. For problem (1.1), the mimetic technique uses discrete flux $\mathcal{G}$ and divergence $\mathcal{DIV}$ operators for the continuum operators $-\mathbf{K}\text{grad}$ and div, respectively, which are adjoint to each other, i.e., $\mathcal{G} = \mathcal{DIV}^*$. It is straightforward to extend the MFD method to locally refined meshes with hanging nodes [16], unstructured three-dimensional meshes composed of hexahedra, tetrahedra, and any cell type having three faces intersecting at each vertex.

A connection between the MFD method and the mixed finite element (MFE) method with Raviart–Thomas finite elements has been established in [4]. In particular, it was shown that the scalar product in the velocity space proposed in [15] for MFD methods can be viewed as a quadrature rule in the context of MFE methods. Another closely related method is the control-volume MFE method [7, 9].

MFE discretizations on quadrilateral grids have been studied in [25, 26, 2, 13]. These methods are based on the Piola transformation [25, 6], which preserves continuity of the normal component of the velocity $\mathbf{u}$ across mesh edges. Unfortunately, this results in the necessity to integrate rational functions over quadrilaterals. The task becomes even more complicated when the diffusion tensor is full and nonconstant. The results in [4] provide an efficient numerical quadrature rule with a minimal number of points. Moreover, the connection between the two methods allows for extensions of MFE methods to general polygons and polyhedra.

The aforementioned connection provides a suitable functional frame for rigorous analysis of convergence of mimetic discretizations. In [4], first-order convergence for the fluid pressure and velocity was shown. In this paper, we establish velocity superconvergence for MFD discretizations of (1.1) on h^2-uniform quadrilateral meshes (as defined in (2.2)–(2.3)). Precise calculation of the fluid velocity is important for porous media and other applications. The points or lines where the numerical solution is superclose to the exact solution may be used to improve the accuracy of the overall simulation. Various superconvergence results for MFE methods have been established for rectangular meshes [21, 19, 27, 10, 11, 12, 3, 1] and general quadrilateral meshes [2, 13].

In [13], velocity superconvergence is established for the MFE discretization of (1.1) on h^2-uniform quadrilateral grids. In this paper, we exploit the relation between MFD methods and MFE methods with the quadrature rule (3.10) to establish superconvergence for velocities in MFD discretizations. In particular, we show that the computed normal velocities are superclose to the true normal velocities at the midpoints of the edges. In [18], an alternative quadrature is introduced, which preserves symmetry of the exact solution on polar grids. This symmetry preservation is important for problems of radiation transport in the asymptotic diffusion limit. The analysis of superconvergence for symmetry-preserving quadratures is left for future investigation.

The paper outline is as follows. In section 2, we describe the MFE method for (1.1). In section 3, the MFD method is presented and related to the MFE method with a quadrature rule. The main superconvergence results are presented in section 4. Superconvergence of the normal velocities at the midpoints of the edges is established in section 5. In section 6, numerical experiments are given that confirm the theoretical results.

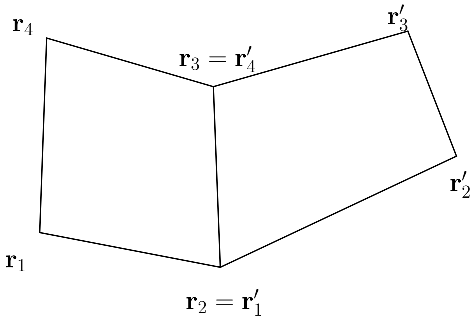

$\mathbf{r}_4$
$\mathbf{r}_3'$
$\mathbf{r}_3 = \mathbf{r}_4'$
$\mathbf{r}_2'$
$\mathbf{r}_1$
$\mathbf{r}_2 = \mathbf{r}_1'$

FIG. 2.1. *h^2-uniform quadrilateral grid.*

2. The MFE method. To simplify the exposition, we assume without loss of generality that $g = 0$, i.e., homogeneous Neumann boundary conditions are imposed on $\partial\Omega$.

Throughout this paper, we shall use the notation $\|\cdot\|_{k,D}$, $\|\cdot\|_{\mathrm{div},D}$, and $\|\cdot\|_D$ for the norms on the Hilbert spaces $H^k(D)$, $H(\mathrm{div}; D)$, and $L_2(D)$, respectively, where $D \subset \Omega$. In addition, $|\cdot|_{k,D}$ will denote the seminorm on $H^k(D)$. To simplify notation, we shall omit the subscript D when $D = \Omega$. Finally, we denote by $(\cdot,\cdot)$ the L^2-inner product on Ω of either scalar or vector functions. Let

$$\mathbf{V} = \{\mathbf{v} \in H(\mathrm{div};\Omega) : \mathbf{v}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\} \quad \text{and} \quad W = \left\{ w \in L_2(\Omega) : \int_\Omega w\, d\mathbf{x} = 0 \right\}.$$

The variational formulation of (1.1) is as follows: find a pair $(\mathbf{u}, p) \in \mathbf{V} \times W$ such that

$$\begin{aligned} (\mathbf{K}^{-1}\mathbf{u}, \mathbf{v}) - (p, \mathrm{div}\, \mathbf{v}) &= 0, \\ (\mathrm{div}\, \mathbf{u}, w) &= (f, w) \qquad \forall\, (\mathbf{v}, w) \in \mathbf{V} \times W. \end{aligned} \tag{2.1}$$

For the discretization of (2.1), denote by $\mathcal{T}_h$ a *shape-regular* partition (see [5, Remark 2.2, p. 113]) of $\bar{\Omega}$ into convex quadrilateral elements of diameter not greater than h. For two examples of shape-regular grids, see Figure 6.1. We assume that the grid is h^2-uniform. Following [13], the quadrilateral partition $\mathcal{T}_h$ is called h^2-uniform if each element is an h^2-parallelogram, i.e.,

$$\|(\mathbf{r}_2 - \mathbf{r}_1) - (\mathbf{r}_3 - \mathbf{r}_4)\| \le Ch^2, \tag{2.2}$$

and any two adjacent quadrilaterals form an h^2-parallelogram, i.e.,

$$\|(\mathbf{r}_2 - \mathbf{r}_1) - (\mathbf{r}_2' - \mathbf{r}_1')\| \le Ch^2, \tag{2.3}$$

where $\mathbf{r}_1'$, $\mathbf{r}_2'$, $\mathbf{r}_3'$, and $\mathbf{r}_4'$ are the vertices of the adjacent element (see Figure 2.1).

For any convex quadrilateral e, there exists a bijection mapping $\mathbf{F}_e : \hat{e} \to e$, where $\hat{e}$ is the reference unit square with vertices $\hat{\mathbf{r}}_1 = (0, 0)^T$, $\hat{\mathbf{r}}_2 = (1, 0)^T$, $\hat{\mathbf{r}}_3 = (1, 1)^T$, and $\hat{\mathbf{r}}_4 = (0, 1)^T$. Denote by $\mathbf{r}_i = (x_i, y_i)^T$, $i = 1, 2, 3, 4$, the four corresponding vertices of element e as shown in Figure 2.2. Then, $\mathbf{F}_e$ is the bilinear mapping given by

$$\mathbf{F}_e(\hat{\mathbf{r}}) = \mathbf{r}_1 (1 - \hat{x})(1 - \hat{y}) + \mathbf{r}_2\, \hat{x}(1 - \hat{y}) + \mathbf{r}_3\, \hat{x}\hat{y} + \mathbf{r}_4 (1 - \hat{x})\hat{y}. \tag{2.4}$$

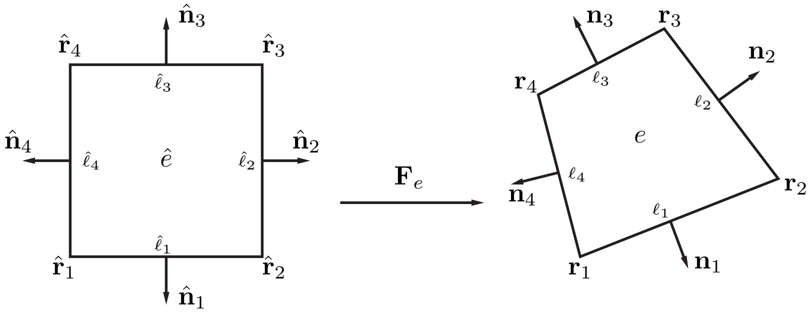


FIG. 2.2. *Bilinear mapping and orientation of normal vectors.*

Note that the Jacobi matrix $\mathbf{DF}_e$ and its Jacobian J_e are linear functions of $\hat{x}$ and $\hat{y}$. Indeed, straightforward computations yield

$$\mathbf{DF}_e = [(1-\hat{y})\,\mathbf{r}_{21} + \hat{y}\,\mathbf{r}_{34},\ (1-\hat{x})\,\mathbf{r}_{41} + \hat{x}\,\mathbf{r}_{32}] \tag{2.5}$$

and

$$J_e = 2|T_{124}| + 2(|T_{123}| - |T_{124}|)\hat{x} + 2(|T_{134}| - |T_{124}|)\hat{y}, \tag{2.6}$$

where $\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j$ and $|T_{ijk}|$ is the area of the triangle with vertices $\mathbf{r}_i$, $\mathbf{r}_j$, and $\mathbf{r}_k$. Since e is convex, the Jacobian J_e is always positive, i.e., $J_e > 0$.

Let ℓ_i and $\hat{\ell}_i$, $i = 1, 2, 3, 4$, be the edges of e and $\hat{e}$, respectively. Let $\mathbf{n}_i$ and $\hat{\mathbf{n}}_i$ be the unit outward normal vectors to ℓ_i and $\hat{\ell}_i$, respectively (see Figure 2.2). Similarly, let $\boldsymbol{\tau}_i$ and $\hat{\boldsymbol{\tau}}_i$ be the unit tangential vectors to ℓ_i and $\hat{\ell}_i$, respectively. It is easy to see from (2.5) that for any edge ℓ_i,

$$\mathbf{n}_i = \frac{1}{|\ell_i|} J_e \mathbf{DF}_e^{-T} \hat{\mathbf{n}}_i \qquad \text{and} \qquad \boldsymbol{\tau}_i = \frac{1}{|\ell_i|} \mathbf{DF}_e \hat{\boldsymbol{\tau}}_i. \tag{2.7}$$

The reader is referred to [6] for suitable choices for the pair of finite element spaces $\mathbf{V}^h \subset \mathbf{V}$ and $W^h \subset W$. In this paper, we consider the lowest-order Raviart–Thomas finite element spaces RT_0 [25, 20] defined on the reference element $\hat{e}$ as

$$\hat{\mathbf{V}}(\hat{e}) = P_{1,0}(\hat{e}) \times P_{0,1}(\hat{e}), \qquad \hat{W}(\hat{e}) = P_0(\hat{e}),$$

where $P_{1,0}$ (or $P_{0,1}$) denotes the space of polynomials linear in the $\hat{x}$ (or $\hat{y}$) variable and constant in the other variable, and P_0 denotes the space of constant functions. The velocity space on any convex quadrilateral e is defined through the Piola transformation [6]

$$\frac{1}{J_e}\mathbf{DF}_e :\ L_2(\hat{e}) \times L_2(\hat{e}) \to L_2(e) \times L_2(e) \qquad \forall e \in \mathcal{T}_h.$$

The RT_0 spaces on $\mathcal{T}_h$ are given by

$$\begin{aligned} \mathbf{V}^h &= \{\mathbf{v} \in \mathbf{V}\ :\ \mathbf{v}|_e = J_e^{-1}\mathbf{DF}_e \hat{\mathbf{v}} \circ \mathbf{F}_e^{-1},\ \hat{\mathbf{v}} \in \hat{\mathbf{V}}(\hat{e}) \quad \forall e \in \mathcal{T}_h\}, \\ W^h &= \{w \in W\ :\ w|_e = \hat{w} \circ \mathbf{F}_e^{-1},\ \hat{w} \in \hat{W}(\hat{e}) \quad \forall e \in \mathcal{T}_h\}. \end{aligned} \tag{2.8}$$

Two properties of Piola's transformation will be important in our analysis. For any $\hat{\mathbf{v}} \in \hat{\mathbf{V}}(\hat{e})$ and the related $\mathbf{v} = J_e^{-1}\mathbf{DF}_e \hat{\mathbf{v}} \circ \mathbf{F}_e^{-1}$,

$$J_e \mathrm{div}\ \mathbf{v} = \widehat{\mathrm{div}}\ \hat{\mathbf{v}} \qquad \text{and} \qquad |\ell_i|\,\mathbf{v}\cdot\mathbf{n}_i = \hat{\mathbf{v}}\cdot\hat{\mathbf{n}}_i. \tag{2.9}$$

Note that since $\mathbf{V}^h \subset H(\text{div};\Omega)$, any vector in $\mathbf{V}^h$ has continuous normal components on the edges. A function in W^h is uniquely determined by its values at the cell-centers and a vector in $\mathbf{V}^h$ is uniquely determined by its normal components on the edges. Therefore, $\dim W^h = N_p$ and $\dim \mathbf{V}^h = N_e$, where N_p is the number of elements and N_e is the number of interior edges. Let $\{\psi_i^h\}$, $i = 1, N_p$, be a basis for W^h such that

$$\psi_i^h(c_j) = \delta_{ij} \equiv \begin{cases} 1, & i = j, \\ 0, & i \neq j, \end{cases}$$

where c_j is the center of element e_j, $j = 1, N_p$. Similarly, let ϕ_i^h, $i = 1, N_e$, be a basis for $\mathbf{V}^h$ such that $\phi_i^h \cdot \mathbf{n}_j = \delta_{ij}$, where $\mathbf{n}_j$ is a fixed unit normal vector on edge ℓ_j, $j = 1, N_e$. In order to simplify notation, we use the same way for global and local indexing of mesh edges and corresponding normal vectors.

Given the finite element spaces $\mathbf{V}^h$ and W^h, we define the discrete problem: find $(\mathbf{u}^h,\, p^h) \in \mathbf{V}^h \times W^h$ such that

$$\begin{aligned} (\mathbf{K}^{-1}\mathbf{u}^h,\, \mathbf{v}^h)_h - (p^h,\, \text{div}\, \mathbf{v}^h) &= 0, \\ (\text{div}\, \mathbf{u}^h,\; w^h) &= (f,\, w^h) \qquad \forall\, (\mathbf{v}^h,\, w^h) \in \mathbf{V}^h \times W^h, \end{aligned} \tag{2.10}$$

where $(\cdot,\, \cdot)_h$ is a continuous bilinear form corresponding to the application of a numerical quadrature rule for computing $(\cdot,\, \cdot)$. A detailed discussion of this quadrature rule is given in section 3.

3. MFD discretizations.

In this section, we derive an MFD discretization of (1.1) and show its connection with the MFE method (2.10).

The *first* step in the mimetic technique is to specify discrete degrees of freedom for pressure and velocity. The discrete pressure unknowns are defined at the centers of the quadrilaterals, one unknown per mesh cell. The discrete velocities are defined at the midpoints of mesh edges as normal components. In other words, an edge-based unknown is a scalar and represents the orthogonal projection of a velocity vector onto the unit vector $\mathbf{n}_i$ normal to the mesh edge ℓ_i.

The *second* step in the mimetic technique is to equip the spaces of discrete pressures and velocities with scalar products. We denote the vector space of cell-centered pressures by Q^d. The dimension of Q^d equals the number of mesh cells N_p. The scalar product on the vector space Q^d is given by

$$[p^d,\, q^d]_{Q^d} = \sum_{i=1}^{N_p} |e_i|\, p_i^d\, q_i^d \qquad \forall p^d,\; q^d \in Q^d, \tag{3.1}$$

where $|e_i|$ denotes the area of cell e_i and p_i^d, q_i^d are cell-centered pressure components.

It is easy to see that the vector space Q^d is isometric to the MFE space W^h in (2.8). Indeed, for any $p^h \in W^h$, there exists a unique $p^d = (p_1^d, p_2^d, \ldots, p_{N_p}^d)^T \in Q^d$ such that $p^h = \sum_{i=1}^{N_p} p_i^d \psi_i^h$ and

$$(p^h,\, q^h) = [p^d,\, q^d]_{Q^d}.$$

Note that the discrete MFD pressure variable, p_i^d, corresponds to the value of the MFE pressure function at the cell-center, $p^h(c_i)$.

We denote the vector space of edge-based velocities by X^d. The dimension of X^d equals the number of interior mesh edges N_e. The scalar product on X^d is given by

$$[\mathbf{u}^d,\,\mathbf{v}^d]_{X^d} = \sum_{e\in\mathcal{T}_h} [\mathbf{u}^d,\,\mathbf{v}^d]_{X^d,e}, \tag{3.2}$$

where $[\mathbf{u}^d,\,\mathbf{v}^d]_{X^d,e}$ is a scalar product over cell e involving only the normal velocity components on cell edges. Recall that a velocity vector can be recovered from two orthogonal projections on any two noncollinear vectors. Since the mesh cell is convex, any pair of normal vectors to edges with a common point satisfies the above requirement. The orthogonal projections are exactly the degrees of freedom associated with cell edges. As shown in Figure 3.1, four recovered velocity vectors can be associated with the four vertices of the quadrilateral. For example, velocity $\mathbf{v}_1$ is recovered from its projections onto the normal vectors $\mathbf{n}_1$ and $\mathbf{n}_2$. For a general quadrilateral e, we denote by $\mathbf{v}^d(\mathbf{r}_j)$ the velocity recovered at jth vertex $\mathbf{r}_j$, $j = 1, 2, 3, 4$. Then, the cell-based scalar product is given by

$$[\mathbf{u}^d,\,\mathbf{v}^d]_{X^d,e} = \frac{1}{2}\sum_{j=1}^{4} |T_j|\,\mathbf{K}^{-1}(\mathbf{r}_j)\mathbf{u}^d(\mathbf{r}_j)\cdot\mathbf{v}^d(\mathbf{r}_j), \tag{3.3}$$

where $|T_j|$ is the area of the triangle with vertices $\mathbf{r}_{j-1}$, $\mathbf{r}_j$, and $\mathbf{r}_{j+1}$ (see Figures 2.2 and 3.1). For example, triangles T_1 and T_4 are the shaded triangles in Figure 3.1. Note that (3.3) is indeed an inner product, since $\mathbf{K}$ is a symmetric and positive definite tensor and

$$[\mathbf{v}^d,\mathbf{v}^d]_{X^d} \geq C|||\mathbf{v}^d|||^2, \tag{3.4}$$

where $|||\cdot|||$ is the Euclidean vector norm.

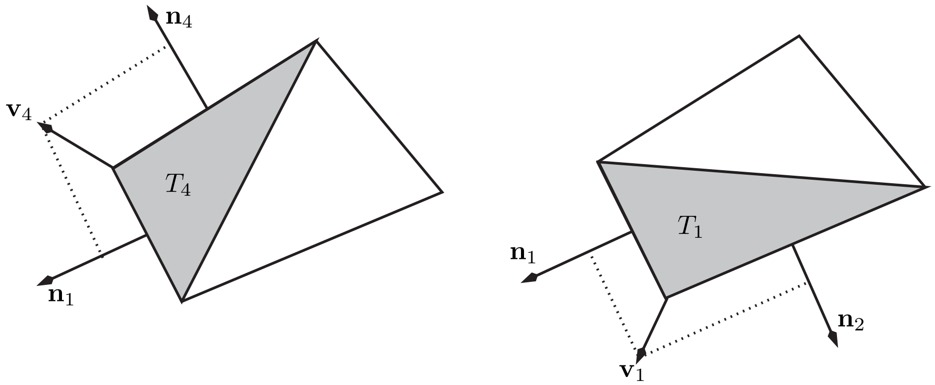


FIG. 3.1. *Recovered vectors* $\mathbf{v}_1$, $\mathbf{v}_4$ *and triangles* T_1, T_4.

The vector space X^d is isomorphic to the MFE space $\mathbf{V}^h$ in (2.8), since both spaces have the same definitions of degrees of freedom. In particular, for any $\mathbf{v}^h \in \mathbf{V}^h$, there exists a unique $\mathbf{v}^d = (v_1^d, v_2^d, \ldots, v_{N^e}^d)^T \in X^d$ such that $\mathbf{v}^h = \sum_{i=1}^{N_e} v_i^d \phi_i^h$. Note that the discrete MFD velocity variable, v_i^d, corresponds to the MFE normal velocity component, $\mathbf{v}^h\cdot\mathbf{n}_i$, on edge ℓ_i.

The *third* step in the mimetic technique is to derive a discrete approximation to the divergence operator, $\mathcal{DIV} : X^d \to Q^d$, which we shall refer to as the *prime* operator. For a cell e, the Gauss divergence theorem gives

$$\mathcal{DIV}\,\mathbf{u}^d|_e = \frac{1}{|e|}\left(u_1^d|\ell_1| + u_2^d|\ell_2| + u_3^d|\ell_3| + u_4^d|\ell_4|\right), \tag{3.5}$$

where $u_1^d, \dots, u_4^d$ are the normal velocity components on element e and the normal vectors are oriented as shown in Figure 2.2.

The *fourth* step in the mimetic technique is to derive a discrete flux operator $\mathcal{G}$ (for the continuous operator $-\mathbf{K}\mathrm{grad}$) adjoint to the discrete divergence operator $\mathcal{DIV}$ with respect to scalar products (3.1) and (3.2), i.e.,

$$[\mathcal{DIV}\mathbf{u}^d,\, p^d]_{Q^d} \equiv [\mathbf{u}^d,\, \mathcal{G}p^d]_{X^d} \qquad \forall \mathbf{u}^d \in X^d \qquad \forall p^d \in Q^d.$$

To derive the explicit formula for $\mathcal{G}$, we consider an auxiliary scalar product $\langle\cdot\,,\,\cdot\rangle$ and relate it to scalar products (3.1) and (3.2). Denote by $\langle\cdot\,,\,\cdot\rangle$ the standard vector dot product. Then

$$[p^d,\, q^d]_{Q^d} = \langle \mathcal{D}p^d,\, q^d\rangle \qquad \text{and} \qquad [\mathbf{u}^d,\, \mathbf{v}^d]_{X^d} = \langle \mathcal{M}\mathbf{u}^d,\, \mathbf{v}^d\rangle,$$

where $\mathcal{D}$ is a diagonal matrix, $\mathcal{D} = \mathrm{diag}\{|e_1|, \dots, |e_{N_p}|\}$, and $\mathcal{M}$ is a sparse symmetric mass matrix with a 5-point stencil. Restricted to a cell, this stencil connects edge-based unknowns if and only if the corresponding edges have a common point. Combining the last two formulae, we get

$$\begin{aligned}[\mathbf{u}^d,\, \mathcal{DIV}^*p^d]_{X^d} &= \langle \mathbf{u}^d,\, \mathcal{M}\,\mathcal{DIV}^*p^d\rangle = [\mathcal{DIV}\mathbf{u}^d,\, p^d]_{Q^d} \\ &= \langle \mathbf{u}^d,\, \mathcal{DIV}^t\,\mathcal{D}\,p^d\rangle \qquad \forall \mathbf{u}^d \in X^d \quad \forall p^d \in Q^d,\end{aligned}$$

where $\mathcal{DIV}^t$ is the adjoint of $\mathcal{DIV}$ with respect to the auxiliary scalar product. Therefore,

$$\mathcal{G} = \mathcal{M}^{-1}\,\mathcal{DIV}^t\,\mathcal{D}. \tag{3.6}$$

The MFD method approximating first-order system (1.1) may be summarized as follows:

$$\mathbf{u}^d = \mathcal{G}\,p^d, \qquad \mathcal{DIV}\,\mathbf{u}^d = f^d, \tag{3.7}$$

where $f^d = (f_1^d, \dots, f_{N_p}^d)^t$, and entry f_i^d is the integral average of f over cell e_i.

The basic tool for the error analysis of the discrete solution $(\mathbf{u}^d,\, p^d) \in X^d \times Q^d$ is based on the following transformation. Multiplying the first equation in (3.7) by $\mathcal{M}\mathbf{v}^d$ and the second one by $\mathcal{D}q^d$, we get

$$\begin{aligned}[\mathbf{u}^d,\, \mathbf{v}^d]_{X^d} - [p^d,\, \mathcal{DIV}\,\mathbf{v}^d]_{Q^d} &= 0, \\ [q^d,\, \mathcal{DIV}\,\mathbf{u}^d]_{Q^d} &= [f^d,\, q^d]_{Q^d} \qquad \forall (\mathbf{v}^d,\, q^d) \in X^d \times Q^d.\end{aligned} \tag{3.8}$$

Using the isomorphism between the finite element space $\mathbf{V}^h \times W^h$ and the vector space $X^d \times Q^d$, we define finite element functions p^h, q^h, f^h, $\mathbf{u}^h$, and $\mathbf{v}^h$ corresponding to vectors p^d, q^d, f^d, $\mathbf{u}^d$, and $\mathbf{v}^d$, respectively. Then

$$[p^d,\, \mathcal{DIV}\,\mathbf{v}^d]_{Q^d} = (p^h,\, \mathrm{div}\ \mathbf{v}^h) \qquad \text{and} \qquad [q^d,\, \mathcal{DIV}\,\mathbf{u}^d]_{Q^d} = (q^h,\, \mathrm{div}\ \mathbf{u}^h).$$

The definition of f^d implies that

$$[f^d,\, q^d]_{Q^d} = (f^h,\, q^h) = (f,\, q^h).$$

Finally, by introducing the quadrature rule

$$(\mathbf{K}^{-1}\mathbf{u}^h,\, \mathbf{v}^h)_h \equiv [\mathbf{u}^d,\, \mathbf{v}^d]_{X^d}, \tag{3.9}$$

we reduce problem (3.7) to the finite element problem (2.10).

The scalar product in the space of velocities given by (3.3) is obviously not unique. In the context of MFE methods, it is a quadrature rule for numerical integration of $(\mathbf{K}^{-1}\mathbf{u}^h,\, \mathbf{v}^h)$:

$$(\mathbf{K}^{-1}\mathbf{u}^h,\, \mathbf{v}^h)_{h,e} = \frac{1}{2}\sum_{j=1}^{4} |T_j|\, \mathbf{K}^{-1}(\mathbf{r}_j)\mathbf{u}^h(\mathbf{r}_j) \cdot \mathbf{v}^h(\mathbf{r}_j), \tag{3.10}$$

where $\mathbf{u}^h(\mathbf{r}_j)$ is the recovered velocity at vertex $\mathbf{r}_j$. In the context of MFE methods, we shall refer to (3.10) as the MFD quadrature rule. The global scalar product is obtained by summing over quadrilaterals, i.e.,

$$(\mathbf{K}^{-1}\mathbf{u}^h,\, \mathbf{v}^h)_h = \sum_{e\in\mathcal{T}_h} (\mathbf{K}^{-1}\mathbf{u}^h,\, \mathbf{v}^h)_{h,e}. \tag{3.11}$$

Note that (3.4) implies that there exists a constant $C_0 > 0$ such that

$$(\mathbf{K}^{-1}\mathbf{v}^h, \mathbf{v}^h)_h \geq C_0\|\mathbf{v}^h\|^2 \qquad \forall \mathbf{v}^h \in \mathbf{V}^h. \tag{3.12}$$

It was shown in [4] that the element quadrature rule (3.10) is exact for any constant vector $\mathbf{u}^h$, constant tensor $\mathbf{K}$, and $\mathbf{v}^h \in \mathbf{V}^h$.

4. Superconvergence estimates for the velocity. We begin by recalling the mixed projection operator $\Pi : H^1(\Omega) \times H^1(\Omega) \to \mathbf{V}^h$ satisfying

$$(\operatorname{div}(\Pi\,\mathbf{v} - \mathbf{v}),\, w) = 0 \qquad \forall w \in W^h. \tag{4.1}$$

The operator Π is defined locally on each element e by

$$\widehat{\Pi\mathbf{v}} = \hat{\Pi}\hat{\mathbf{v}},$$

where $\hat{\Pi} : H^1(\hat{e}) \times H^1(\hat{e}) \to \hat{\mathbf{V}}(\hat{e})$ is the reference element projection operator satisfying

$$\int_{\hat{\ell}_i} (\hat{\Pi}\hat{\mathbf{v}} - \hat{\mathbf{v}}) \cdot \hat{\mathbf{n}}_i = 0, \quad i = 1, 2, 3, 4. \tag{4.2}$$

The approximation properties of Π have been established in [25, 26]:

$$\|\Pi\mathbf{v}\|_{\mathrm{div}} \leq C\|\mathbf{v}\|_1, \tag{4.3}$$

$$\|\Pi\mathbf{v} - \mathbf{v}\| \leq Ch\|\mathbf{v}\|_1, \tag{4.4}$$

$$\|\operatorname{div}(\Pi\mathbf{v} - \mathbf{v})\| \leq Ch\|\mathbf{v}\|_2. \tag{4.5}$$

The following lemma gives several approximation properties of $\hat{\Pi}$ which will be used in the analysis.

LEMMA 4.1. *The operator* $\hat{\Pi}$ *defined in* (4.2) *satisfies, for any* $\hat{\mathbf{v}} = (\hat{v}_1,\, \hat{v}_2)$ *in* $H^1(\hat{e}) \times H^1(\hat{e})$, *the following:*

$$\int_{\hat{e}} \frac{\partial}{\partial \hat{x}}(\hat{\Pi}\hat{\mathbf{v}} - \hat{\mathbf{v}})_1\, d\hat{x}\, d\hat{y} = 0, \quad \int_{\hat{e}} \frac{\partial}{\partial \hat{y}}(\hat{\Pi}\hat{\mathbf{v}} - \hat{\mathbf{v}})_2\, d\hat{x}\, d\hat{y} = 0, \tag{4.6}$$

$$\left\|\frac{\partial}{\partial \hat{x}}(\hat{\Pi}\hat{\mathbf{v}})_1\right\|_{\hat{e}} \leq C\left\|\frac{\partial}{\partial \hat{x}}\hat{v}_1\right\|_{\hat{e}}, \quad \left\|\frac{\partial}{\partial \hat{y}}(\hat{\Pi}\hat{\mathbf{v}})_2\right\|_{\hat{e}} \leq C\left\|\frac{\partial}{\partial \hat{y}}\hat{v}_2\right\|_{\hat{e}}, \tag{4.7}$$

$$\|\hat{\Pi}\hat{\mathbf{v}}\|_{1,\hat{e}} \leq C\|\hat{\mathbf{v}}\|_{1,\hat{e}}. \tag{4.8}$$

Proof. The identities in (4.6) follow easily from definition (4.2). In particular, writing (4.2) for the two vertical edges gives

$$\int_0^1 (\hat{\Pi}\hat{\mathbf{v}} - \hat{\mathbf{v}})_1(0,\hat{y})\,d\hat{y} = 0, \qquad \int_0^1 (\hat{\Pi}\hat{\mathbf{v}} - \hat{\mathbf{v}})_1(1,\hat{y})\,d\hat{y} = 0.$$

Subtracting the above equations and applying the fundamental theorem of calculus implies the first identity in (4.6). The proof of the second identity is similar. Note that (4.6) means that $\frac{\partial}{\partial \hat{x}}(\hat{\Pi}\hat{\mathbf{v}})_1$ and $\frac{\partial}{\partial \hat{y}}(\hat{\Pi}\hat{\mathbf{v}})_2$ are the L^2-orthogonal projections of $\frac{\partial}{\partial \hat{x}}\hat{v}_1$ and $\frac{\partial}{\partial \hat{y}}\hat{v}_2$, respectively, onto the space of constants, which implies (4.7). Finally, it is easy to see that (4.2) implies

$$\|\hat{\Pi}\hat{\mathbf{v}}\|_{\hat{e}} \le C\|\hat{\mathbf{v}}\|_{1,\hat{e}},$$

which, combined with (4.7), gives (4.8). ☐

We also make use of the L^2-projection operator $P_h : W \to W^h$ such that for $p \in W$,

$$(P_h\, p - p,\, w) = 0 \qquad \forall w \in W^h. \tag{4.9}$$

Denote the quadrature error by

$$\sigma(\mathbf{q},\mathbf{v}) \equiv (\mathbf{q},\mathbf{v}) - (\mathbf{q},\mathbf{v})_h. \tag{4.10}$$

The variational formulation (2.1) and the discrete problem (2.10) give rise to the error equations

$$\begin{aligned} (\mathbf{K}^{-1}(\Pi\mathbf{u} - \mathbf{u}^h), \mathbf{v}^h)_h &= (P_h p - p^h, \operatorname{div}\mathbf{v}^h) \\ &\qquad + (\mathbf{K}^{-1}(\Pi\mathbf{u} - \mathbf{u}), \mathbf{v}^h) - \sigma(\mathbf{K}^{-1}\Pi\mathbf{u}, \mathbf{v}^h), \\ (\operatorname{div}(\Pi\mathbf{u} - \mathbf{u}^h), w^h) &= 0, \end{aligned} \tag{4.11}$$

where we used (4.9) and (4.1) in the first and second equations, respectively. We note that, using (2.9), the second equation in (4.11) gives

$$0 = (\operatorname{div}(\Pi\mathbf{u} - \mathbf{u}^h), w^h)_e = (\widehat{\operatorname{div}}(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}}^h), \hat{w}^h)_{\hat{e}} \qquad \forall w^h \in W_h.$$

Since $\widehat{\operatorname{div}}\,\hat{\mathbf{V}}^h = \hat{W}_h$, taking $\hat{w}^h = \widehat{\operatorname{div}}(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}}^h)$ implies that $\widehat{\operatorname{div}}(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}}^h) = 0$ and therefore, by (2.9),

$$\operatorname{div}(\Pi\mathbf{u} - \mathbf{u}^h) = 0. \tag{4.12}$$

Taking $\mathbf{v}^h = \Pi\mathbf{u} - \mathbf{u}^h \in \mathbf{V}^h$ and $w^h = P_h p - p^h$ in (4.11) gives

$$(\mathbf{K}^{-1}(\Pi\mathbf{u} - \mathbf{u}^h), \Pi\mathbf{u} - \mathbf{u}^h)_h = (\mathbf{K}^{-1}(\Pi\mathbf{u} - \mathbf{u}), \Pi\mathbf{u} - \mathbf{u}^h) - \sigma(\mathbf{K}^{-1}\Pi\mathbf{u}, \Pi\mathbf{u} - \mathbf{u}^h). \tag{4.13}$$

The estimate for the first term on the right-hand side of (4.13) follows from Theorem 5.1 in [13] and (4.12):

$$\begin{aligned} &(\mathbf{K}^{-1}(\Pi\mathbf{u} - \mathbf{u}), \Pi\mathbf{u} - \mathbf{u}^h) \\ &\qquad \le C\,h^2\left(\|\mathbf{u}\|_2\|\Pi\mathbf{u} - \mathbf{u}^h\| + \|\mathbf{u}\|_1\|\operatorname{div}(\Pi\mathbf{u} - \mathbf{u}^h)\|\right) \\ &\qquad = C\,h^2\|\mathbf{u}\|_2\,\|\Pi\mathbf{u} - \mathbf{u}^h\|. \end{aligned} \tag{4.14}$$

The second term on the right-hand side of (4.13) can be bounded using Lemma 4.3:

$$|\sigma(\mathbf{K}^{-1}\Pi\mathbf{u}, \Pi\mathbf{u} - \mathbf{u}^h)| \leq C\, h^2 \|\mathbf{u}\|_2 \|\Pi\mathbf{u} - \mathbf{u}^h\|. \tag{4.15}$$

Combining (4.14), (4.15), and (3.12), we obtain the following superconvergence result.

THEOREM 4.2. *Let* $\mathbf{K}^{-1} \in W^{2,\infty}(\Omega)$. *For the velocity* $\mathbf{u}^h$ *of the MFE method* (2.1), *on* h^2*-uniform quadrilateral grids, there exists a positive constant* C *independent of* h *such that*

$$\|\Pi\mathbf{u} - \mathbf{u}^h\| \leq C\, h^2 \|\mathbf{u}\|_2. \tag{4.16}$$

We now proceed to prove estimate (4.15).

LEMMA 4.3. *Let* $\mathbf{v} \in \mathbf{V}^h$, *and let* $\mathbf{K}^{-1} \in W^{2,\infty}(\Omega)$. *There exists a positive constant* C *independent of* h *such that*

$$|\sigma(\mathbf{K}^{-1}\Pi\mathbf{u}, \mathbf{v})| \leq C\, h^2 (\|\mathbf{u}\|_2 \, \|\mathbf{v}\| + \|\mathbf{u}\|_1 \|\mathrm{div}\, \mathbf{v}\|). \tag{4.17}$$

Proof. For an element $e \in \mathcal{T}^h$, we define the error

$$\sigma_e(\mathbf{K}^{-1}\Pi\mathbf{u}, \mathbf{v}) = \int_e \mathbf{K}^{-1}\Pi\mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} - (\mathbf{K}^{-1}\Pi\mathbf{u}, \mathbf{v})_{h,e}. \tag{4.18}$$

With (3.10), the second term on the right-hand side of (4.18) can be written as

(4.19)

$$\begin{aligned}
(\mathbf{K}^{-1}\Pi\mathbf{u}, \mathbf{v})_{h,e} &= \frac{1}{2}\sum_{j=1}^{4} |T_j| \mathbf{K}^{-1}(\mathbf{r}_j)\Pi\mathbf{u}(\mathbf{r}_j) \cdot \mathbf{v}(\mathbf{r}_j) \\
&= \frac{1}{2}\sum_{j=1}^{4} |T_j| \hat{\mathbf{K}}^{-1}(\hat{\mathbf{r}}_j) \left(\frac{1}{J_e}\mathbf{DF}_e\hat{\Pi}\hat{\mathbf{u}}\right)(\hat{\mathbf{r}}_j) \cdot \left(\frac{1}{J_e}\mathbf{DF}_e\hat{\mathbf{v}}\right)(\hat{\mathbf{r}}_j) \\
&= \frac{1}{2}\sum_{j=1}^{4} \frac{|T_j|}{J_e(\hat{\mathbf{r}}_j)} \frac{1}{J_e(\hat{\mathbf{r}}_j)} \mathbf{DF}_e^T(\hat{\mathbf{r}}_j)\hat{\mathbf{K}}^{-1}(\hat{\mathbf{r}}_j)\mathbf{DF}_e(\hat{\mathbf{r}}_j)\, \hat{\Pi}\hat{\mathbf{u}}(\hat{\mathbf{r}}_j) \cdot \hat{\mathbf{v}}(\hat{\mathbf{r}}_j) \\
&= \frac{1}{4}\sum_{j=1}^{4} \mathbf{B}_e(\hat{\mathbf{r}}_j)\, \hat{\Pi}\hat{\mathbf{u}}(\hat{\mathbf{r}}_j) \cdot \hat{\mathbf{v}}(\hat{\mathbf{r}}_j) \\
&\equiv (\mathbf{B}_e\hat{\Pi}\hat{\mathbf{u}}, \hat{\mathbf{v}})_T,
\end{aligned}$$

where the subscript T denotes the trapezoidal rule on element $\hat{e}$ and we define $\mathbf{B}_e = \frac{1}{J_e}\mathbf{DF}_e^T\hat{\mathbf{K}}^{-1}\mathbf{DF}_e$. Here we used (2.6) to conclude that $\frac{|T_j|}{J_e(\hat{\mathbf{r}}_j)} = \frac{1}{2}$. Considering the first term on the right-hand side of (4.18), we obtain

$$\begin{aligned}
\int_e \mathbf{K}^{-1}\Pi\mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} &= \int_{\hat{e}} \hat{\mathbf{K}}^{-1}\frac{1}{J_e}\mathbf{DF}_e\hat{\Pi}\hat{\mathbf{u}} \cdot \frac{1}{J_e}\mathbf{DF}_e\hat{\mathbf{v}} J_e \; d\hat{\mathbf{x}} \\
&= \int_{\hat{e}} \frac{1}{J_e}\mathbf{DF}_e^T\hat{\mathbf{K}}^{-1}\mathbf{DF}_e\hat{\Pi}\hat{\mathbf{u}} \cdot \hat{\mathbf{v}} \; d\hat{\mathbf{x}} \\
&= \int_{\hat{e}} \mathbf{B}_e\hat{\Pi}\hat{\mathbf{u}} \cdot \hat{\mathbf{v}} \; d\hat{\mathbf{x}}.
\end{aligned} \tag{4.20}$$

Substituting (4.19) and (4.20) into (4.18), we obtain

$$\sigma_e(\mathbf{K}^{-1}\Pi\mathbf{u}, \mathbf{v}) = \int_{\hat{e}} \mathbf{B}_e\hat{\Pi}\hat{\mathbf{u}} \cdot \hat{\mathbf{v}} \; d\hat{\mathbf{x}} - \big(\mathbf{B}_e\hat{\Pi}\hat{\mathbf{u}}, \hat{\mathbf{v}}\big)_T \equiv \sigma_{\hat{e}}\big(\mathbf{B}_e\hat{\Pi}\hat{\mathbf{u}}, \hat{\mathbf{v}}\big). \tag{4.21}$$

Hereafter we shall omit the subscripts e and $\hat{e}$. Let

$$E(f) \equiv \int_{\hat{e}} f(\hat{x},\hat{y})d\hat{x}\,d\hat{y} - (f)_T$$

be the error of the trapezoidal rule for integrating a function $f(\hat{x},\hat{y})$ on $\hat{e}$. Then,

$$\sigma(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}},\hat{\mathbf{v}}) = E\big((\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1\hat{v}_1\big) + E\big((\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_2\hat{v}_2\big). \tag{4.22}$$

We next bound the first term on the right-hand side in (4.22). The argument for the bound on the second term is similar.

Using the trapezoidal rule error representation from Lemma A.1 based on the Peano kernel theorem (see [23, Theorem 5.2-3, p. 142]), we write

$$\begin{aligned} E((\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1\hat{v}_1) &= \int_0^1\int_0^1 \phi(\hat{x})\frac{\partial^2}{\partial\hat{x}^2}\big((\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1\hat{v}_1\big)(\hat{x},0)\,d\hat{x}\,d\hat{y} \\ &\quad+ \int_0^1\int_0^1 \phi(\hat{y})\frac{\partial^2}{\partial\hat{y}^2}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(0,\hat{y})\hat{v}_1(0,\hat{y})\,d\hat{x}\,d\hat{y} \\ &\quad+ \int_0^1\int_0^1 \psi(\hat{x},\hat{y})\frac{\partial^2}{\partial\hat{x}\partial\hat{y}}\big((\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1\hat{v}_1\big)(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y} \\ &\equiv (I) + (II) + (III), \end{aligned} \tag{4.23}$$

where $\phi(t) = t(t-1)/2$ and $\psi(s,t) = (1-s)(1-t) - 1/4$. Denote by B_{11}, B_{12}, B_{21}, and B_{22} the components of the tensor $\mathbf{B}$. Since $\hat{v}_1(0,\hat{y})$ is constant in $\hat{y}$, the second term in (4.23) is

$$\begin{aligned} (II) &= \int_0^1\int_0^1 \phi(\hat{y})\frac{\partial^2}{\partial\hat{y}^2}B_{11}(0,\hat{y})(\hat{\Pi}\hat{\mathbf{u}})_1(0,\hat{y})\hat{v}_1(0,\hat{y})\,d\hat{x}\,d\hat{y} \\ &\quad+ \int_0^1\int_0^1 \phi(\hat{y})\frac{\partial^2}{\partial\hat{y}^2}B_{12}(0,\hat{y})(\hat{\Pi}\hat{\mathbf{u}})_2(0,\hat{y})\hat{v}_1(0,\hat{y})\,d\hat{x}\,d\hat{y} \\ &\quad+ 2\int_0^1\int_0^1 \phi(\hat{y})\frac{\partial}{\partial\hat{y}}B_{12}(0,\hat{y})\frac{\partial}{\partial\hat{y}}(\hat{\Pi}\hat{\mathbf{u}})_2(0,\hat{y})\hat{v}_1(0,\hat{y})\,d\hat{x}\,d\hat{y} \\ &\equiv (II)_1 + (II)_2 + (II)_3. \end{aligned} \tag{4.24}$$

Using (4.8), for the first two terms on the right-hand side, we have

$$|(II)_1| + |(II)_2| \le C|\mathbf{B}|_{2,\infty,\hat{e}}\|\hat{\mathbf{u}}\|_{1,\hat{e}}\|\hat{v}_1\|_{\hat{e}}.$$

Since $\frac{\partial}{\partial\hat{y}}(\hat{\Pi}\hat{\mathbf{u}})_2$ is a constant, we rewrite the last term in (4.24) as

$$\begin{aligned} (II)_3 &= 2\int_0^1\int_0^1 \phi(\hat{y})\frac{\partial}{\partial\hat{y}}B_{12}(0,\hat{y})\frac{\partial}{\partial\hat{y}}(\hat{\Pi}\hat{\mathbf{u}})_2(\hat{x},\hat{y})\hat{v}_1(0,\hat{y})\,d\hat{x}\,d\hat{y} \\ &\le C|\mathbf{B}|_{1,\infty,\hat{e}}\left\|\frac{\partial}{\partial\hat{y}}(\hat{\Pi}\hat{\mathbf{u}})_2\right\|_{\hat{e}}\|\hat{v}_1\|_{\hat{e}} \le C|\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}}\|\hat{v}_1\|_{\hat{e}}, \end{aligned}$$

using (4.7). A combination of the last two bounds implies that

$$|(II)| \le C\big(|\mathbf{B}|_{2,\infty,\hat{e}}\|\hat{\mathbf{u}}\|_{1,\hat{e}} + |\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}}\big)\|\hat{v}_1\|_{\hat{e}}. \tag{4.25}$$

For the last term in (4.23), since $\hat{v}_1(\hat{x},\hat{y})$ is constant in $\hat{y}$, we have

$$(4.26)\quad (III) = \int_0^1\int_0^1 \psi(\hat{x},\hat{y})\frac{\partial^2}{\partial\hat{x}\partial\hat{y}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},\hat{y})\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y} + \int_0^1\int_0^1 \psi(\hat{x},\hat{y})\frac{\partial}{\partial\hat{y}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y} \equiv (III)_1 + (III)_2.$$

Using (4.7),

$$(4.27)\qquad |(III)_1| \le C(|\mathbf{B}|_{2,\infty,\hat{e}}\|\hat{\mathbf{u}}\|_{1,\hat{e}} + |\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}})\|\hat{v}_1\|_{\hat{e}}.$$

To bound $(III)_2$ we note that $\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},\hat{y})$ is a constant and $\int_0^1\int_0^1 \psi(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y} = 0$. Therefore, by the Bramble–Hilbert lemma [5], and using (4.7),

$$(4.28)\quad |(III)_2| \le C\left|\frac{\partial}{\partial\hat{y}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1\right|_{1,\hat{e}}\|\hat{v}_1\|_{\hat{e}} \le C(|\mathbf{B}|_{2,\infty,\hat{e}}\|\hat{\mathbf{u}}\|_{1,\hat{e}} + |\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}})\|\hat{v}_1\|_{\hat{e}}.$$

The first term in the error representation (4.23) is

$$(4.29)\qquad (I) = \int_0^1\int_0^1 \phi(\hat{x})\frac{\partial^2}{\partial\hat{x}^2}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},0)\,\hat{v}_1(\hat{x},0)\,d\hat{x}\,d\hat{y} + 2\int_0^1\int_0^1 \phi(\hat{x})\frac{\partial}{\partial\hat{x}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},0)\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},0)\,d\hat{x}\,d\hat{y} = (I)_1 + (I)_2.$$

The first term on the right-hand side can be bounded in a way similar to (II):

$$(4.30)\qquad |(I)_1| \le C(|\mathbf{B}|_{2,\infty,\hat{e}}\|\hat{\mathbf{u}}\|_{1,\hat{e}} + |\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}})\|\hat{v}_1\|_{\hat{e}}.$$

We rewrite the second term on the right-hand side in (4.29) as

$$(4.31)\quad \frac{1}{2}(I)_2 = \int_0^1\int_0^1 \phi(\hat{x})\left(\frac{\partial}{\partial\hat{x}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},0) - \frac{\partial}{\partial\hat{x}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},\hat{y})\right)\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},0)\,d\hat{x}\,d\hat{y} + \int_0^1\int_0^1 \phi(\hat{x})\frac{\partial}{\partial\hat{x}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},0)\,d\hat{x}\,d\hat{y} \equiv (I)_{2,1} + (I)_{2,2}.$$

To estimate the first term in (4.31), we write

$$\frac{\partial}{\partial\hat{x}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},\hat{y}) - \frac{\partial}{\partial\hat{x}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},0) = \int_0^{\hat{y}} \frac{\partial^2}{\partial\hat{x}\partial\hat{y}}(\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1(\hat{x},\hat{t})\,d\hat{t}.$$

This allows us to bound the first term in (4.31) in a way similar to bounds (4.25) and (4.30):

$$(4.32)\qquad |(I)_{2,1}| \le C(|\mathbf{B}|_{2,\infty,\hat{e}}\|\hat{\mathbf{u}}\|_{1,\hat{e}} + |\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}})\|\hat{v}_1\|_{\hat{e}}.$$

The second term on the right-hand side in (4.31) can be rewritten as

$$(4.33)\quad (I)_{2,2} = \int_0^1\int_0^1 \phi(\hat{x})\frac{\partial}{\partial\hat{x}}(\mathbf{B}(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}}))_1(\hat{x},\hat{y})\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y} + \int_0^1\int_0^1 \phi(\hat{x})\frac{\partial}{\partial\hat{x}}(\mathbf{B}\hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y} \equiv (I)_{2,2,1} + (I)_{2,2,2},$$

where we used that $\frac{\partial}{\partial \hat{x}}\hat{v}_1(\hat{x},0) = \frac{\partial}{\partial \hat{x}}\hat{v}_1(\hat{x},\hat{y})$ on e, since $\hat{v}_1$ is a constant in $\hat{y}$.

To estimate the second term in (4.33), we use the identity

$$\frac{\partial}{\partial \hat{x}}\hat{v}_1 = -\frac{\partial}{\partial \hat{y}}\hat{v}_2 + \widehat{\operatorname{div}}\,\hat{\mathbf{v}}. \tag{4.34}$$

We rewrite $(I)_{2,2,2}$ as

$$\begin{aligned}(I)_{2,2,2} &= -\int_0^1\int_0^1 \phi(\hat{x})\,\frac{\partial}{\partial \hat{x}}(\mathbf{B}\hat{\mathbf{u}})_1\,\frac{\partial}{\partial \hat{y}}\hat{v}_2\,d\hat{x}\,d\hat{y} + \int_0^1\int_0^1 \phi(\hat{x})\,\frac{\partial}{\partial \hat{x}}(\mathbf{B}\hat{\mathbf{u}})_1\,\widehat{\operatorname{div}}\,\hat{\mathbf{v}}\,d\hat{x}\,d\hat{y}\\
&= \int_{\hat{\ell}_1} - \int_{\hat{\ell}_3}\phi(\hat{x})\,\frac{\partial}{\partial \hat{x}}(\mathbf{B}\hat{\mathbf{u}})_1\,\hat{v}_2\,d\hat{x} + \int_0^1\int_0^1 \phi(\hat{x})\,\frac{\partial^2}{\partial \hat{x}\partial \hat{y}}(\mathbf{B}\hat{\mathbf{u}})_1\,\hat{v}_2\,d\hat{x}\,d\hat{y}\\
&\quad + \int_0^1\int_0^1 \phi(\hat{x})\,\frac{\partial}{\partial \hat{x}}(\mathbf{B}\hat{\mathbf{u}})_1\,\widehat{\operatorname{div}}\,\hat{\mathbf{v}}\,d\hat{x}\,d\hat{y}.\end{aligned} \tag{4.35}$$

Clearly, the last two terms can be bounded by

$$C(|(\mathbf{B}\hat{\mathbf{u}})_1|_{2,\hat{e}}\|\hat{v}_2\|_{\hat{e}} + |(\mathbf{B}\hat{\mathbf{u}})_1|_{1,\hat{e}}\|\widehat{\operatorname{div}}\,\hat{\mathbf{v}}\|_{\hat{e}}). \tag{4.36}$$

We postpone the estimate of the edge integrals in (4.35) for later.

To bound the first term on the right-hand side in (4.33), we have

$$\begin{aligned}(I)_{2,2,1} &= \int_0^1\int_0^1 \phi(\hat{x})\frac{\partial}{\partial \hat{x}}B_{11}(\hat{x},\hat{y})(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial \hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y}\\
&\quad + \int_0^1\int_0^1 \phi(\hat{x})B_{11}(\hat{x},\hat{y})\frac{\partial}{\partial \hat{x}}(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial \hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y}\\
&\quad + \int_0^1\int_0^1 \phi(\hat{x})\frac{\partial}{\partial \hat{x}}B_{12}(\hat{x},\hat{y})(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}})_2(\hat{x},\hat{y})\frac{\partial}{\partial \hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y}\\
&\quad + \int_0^1\int_0^1 \phi(\hat{x})B_{12}(\hat{x},\hat{y})\frac{\partial}{\partial \hat{x}}(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}})_2(\hat{x},\hat{y})\frac{\partial}{\partial \hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y}\\
&\equiv (I)_{2,2,1,1} + (I)_{2,2,1,2} + (I)_{2,2,1,3} + (I)_{2,2,1,4}.\end{aligned} \tag{4.37}$$

Since $\hat{\Pi}\hat{\mathbf{u}}$ is exact for constants, using the Bramble–Hilbert lemma and the inverse inequality, we can bound the first and the third terms in (4.37) as

$$|(I)_{2,2,1,1}| + |(I)_{2,2,1,3}| \le C|\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}}\|\hat{v}_1\|_{\hat{e}}. \tag{4.38}$$

For the second term in (4.37), a Taylor expansion of B_{11} about any fixed point $(\hat{x}_0,\hat{y}_0) \in \hat{e}$ gives

$$(I)_{2,2,1,2} = \int_0^1\int_0^1 \phi(\hat{x})B_{11}(\hat{x}_0,\hat{y}_0)\frac{\partial}{\partial \hat{x}}(\hat{\Pi}\hat{\mathbf{u}} - \hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial \hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y} + R, \tag{4.39}$$

where

$$|R| \le C|\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}}\|\hat{v}_1\|_{\hat{e}}, \tag{4.40}$$

using (4.7) for the last inequality. To bound the first term on the right-hand side in (4.39), we note that

$$(\phi^2)''(\hat{x}) = 6\phi(\hat{x}) + \frac{1}{2}, \quad (\phi^2)'(0) = (\phi^2)'(1) = 0.$$

Therefore, using (4.6), we have

$$\begin{aligned}
&\int_0^1\int_0^1 \phi(\hat{x})B_{11}(\hat{x}_0,\hat{y}_0)\frac{\partial}{\partial\hat{x}}(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y}\\
&\quad=\frac{1}{6}\int_0^1\int_0^1 \frac{\partial^2}{\partial\hat{x}^2}(\phi^2)(\hat{x})B_{11}(\hat{x}_0,\hat{y}_0)\frac{\partial}{\partial\hat{x}}(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y}\\
&\quad=-\frac{1}{6}\int_0^1\int_0^1 \frac{\partial}{\partial\hat{x}}(\phi^2)(\hat{x})B_{11}(\hat{x}_0,\hat{y}_0)\frac{\partial^2}{\partial\hat{x}^2}(\hat{\Pi}\hat{\mathbf{u}}-\hat{\mathbf{u}})_1(\hat{x},\hat{y})\frac{\partial}{\partial\hat{x}}\hat{v}_1(\hat{x},\hat{y})\,d\hat{x}\,d\hat{y}\\
&\quad\le C|\mathbf{B}|_{\infty,\hat{e}}|\hat{\mathbf{u}}|_{2,\hat{e}}\|\hat{v}_1\|_{\hat{e}}.
\end{aligned}\tag{4.41}$$

A combination of (4.39)–(4.41) gives

$$|(I)_{2,2,1,2}|\le C(|\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}}+|\mathbf{B}|_{\infty,\hat{e}}|\hat{\mathbf{u}}|_{2,\hat{e}})\|\hat{v}_1\|_{\hat{e}}.\tag{4.42}$$

To complete the estimate of $(I)_{2,2,1}$, it remains to bound $(I)_{2,2,1,4}$. Using that $\frac{\partial}{\partial\hat{x}}(\hat{\Pi}\hat{\mathbf{u}})_2=0$ and (4.34), we have

(4.43)

$$\begin{aligned}
(I)_{2,2,1,4}&=\int_0^1\int_0^1\phi(\hat{x})B_{12}\frac{\partial}{\partial\hat{x}}\hat{u}_2\frac{\partial}{\partial\hat{y}}\hat{v}_2\,d\hat{x}\,d\hat{y}-\int_0^1\int_0^1\phi(\hat{x})B_{12}\frac{\partial}{\partial\hat{x}}\hat{u}_2\,\widehat{\operatorname{div}\mathbf{v}}\,d\hat{x}\,d\hat{y}\\
&=\int_{\hat{\ell}_3}-\int_{\hat{\ell}_1}\phi(\hat{x})\,B_{12}\frac{\partial}{\partial\hat{x}}\hat{u}_2\,\hat{v}_2\,d\hat{x}-\int_0^1\int_0^1\phi(\hat{x})\frac{\partial}{\partial\hat{y}}\left(B_{12}\frac{\partial}{\partial\hat{x}}\hat{u}_2\right)\hat{v}_2\,d\hat{x}\,d\hat{y}\\
&\quad-\int_0^1\int_0^1\phi(\hat{x})B_{12}\frac{\partial}{\partial\hat{x}}\hat{u}_2\,\widehat{\operatorname{div}\mathbf{v}}\,d\hat{x}\,d\hat{y}.
\end{aligned}$$

The last two terms above are bounded by

$$C[(|\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}}+|\mathbf{B}|_{\infty,\hat{e}}|\hat{\mathbf{u}}|_{2,\hat{e}})\|\hat{v}_2\|_{\hat{e}}+|\mathbf{B}|_{\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}}\|\widehat{\operatorname{div}\mathbf{v}}\|_{\hat{e}}].\tag{4.44}$$

Combining (4.23)–(4.44), we obtain

$$E\big((\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1\hat{v}_1\big)=T_1+T_2+T_3,\tag{4.45}$$

where

$$\begin{aligned}
|T_1|\le C[&(|\mathbf{B}|_{2,\infty,\hat{e}}\|\hat{\mathbf{u}}\|_{1,\hat{e}}+|\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}}+|\mathbf{B}|_{\infty,\hat{e}}|\hat{\mathbf{u}}|_{2,\hat{e}})\|\hat{\mathbf{v}}\|_{\hat{e}}\\
&+(|\mathbf{B}|_{1,\infty,\hat{e}}|\hat{\mathbf{u}}|_{\hat{e}}+|\mathbf{B}|_{\infty,\hat{e}}|\hat{\mathbf{u}}|_{1,\hat{e}})\|\widehat{\operatorname{div}\mathbf{v}}\|_{\hat{e}}],
\end{aligned}\tag{4.46}$$

$$T_2=\int_{\hat{\ell}_1}-\int_{\hat{\ell}_3}\phi(\hat{x})\,\frac{\partial}{\partial\hat{x}}(\mathbf{B}\hat{\mathbf{u}})_1(\hat{x},\hat{y})\,\hat{v}_2(\hat{x},\hat{y})\,d\hat{x},\tag{4.47}$$

and

$$T_3=\int_{\hat{\ell}_3}-\int_{\hat{\ell}_1}\phi(\hat{x})\,B_{12}(\hat{x},\hat{y})\frac{\partial}{\partial\hat{x}}\hat{u}_2(\hat{x},\hat{y})\hat{v}_2(\hat{x},\hat{y})\,d\hat{x}.$$

Using Lemma 4.4 below, T_1 can be bounded as follows:

(4.48)

$$\begin{aligned}
|T_1|\le C\,&\big[\big(h^2\|\mathbf{K}^{-1}\|_{2,\infty,e}\|\mathbf{u}\|_{1,e}+h\|\mathbf{K}^{-1}\|_{1,\infty,e}\,h|\mathbf{u}|_{1,e}+\|\mathbf{K}^{-1}\|_{\infty,e}\,h^2|\mathbf{u}|_{2,e}\big)\,\|\hat{\mathbf{v}}\|_{\hat{e}}\\
&\quad+\big(h\|\mathbf{K}^{-1}\|_{1,\infty,e}\|\mathbf{u}\|_{e}+\|\mathbf{K}^{-1}\|_{\infty,e}h\|\mathbf{u}\|_{1,e}\big)\,h\|\operatorname{div}\mathbf{v}\|_{e}\big]\\
&\le C\,h^2\,(\|\mathbf{K}^{-1}\|_{2,\infty,e}\|\mathbf{u}\|_{2,e}\|\mathbf{v}\|_{e}+\|\mathbf{K}^{-1}\|_{1,\infty,e}\|\mathbf{u}\|_{1,e}\|\operatorname{div}\mathbf{v}\|_{e}),
\end{aligned}$$

using the fact that $|\hat{\mathbf{u}}|_{j,\hat{e}} \le Ch^j \|\mathbf{u}\|_{j,e}$ and $\|\widehat{\operatorname{div} \hat{\mathbf{v}}}\|_{\hat{e}} \le Ch\|\operatorname{div} \mathbf{v}\|_e$ (see [13, Lemma 5.5]).

The term $\sum_e T_2$ is treated in Lemma 4.5 below.

Finally, term T_3 in (4.45) can be rewritten as

$$T_3 = \int_{\hat{\ell}_3} - \int_{\hat{\ell}_1} \phi(\hat{s})\, B_{12}(\hat{s},\hat{y}) \frac{\partial}{\partial \hat{s}} (\hat{\mathbf{u}} \cdot \hat{\mathbf{n}}_k) \hat{\mathbf{v}} \cdot \hat{\mathbf{n}}_k \, d\hat{s}.$$

A similar term appears in the proof of Theorem 5.1 in [13]. Following the argument there, it can be shown that

$$\left| \sum_e T_3 \right| \le C \sum_e h^2 \|\mathbf{u}\|_{2,e} \|\mathbf{v}\|_e. \tag{4.49}$$

A combination of estimates (4.45), (4.48), (4.51), and (4.49) implies that

$$\sum_e |E\big((\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_1 \hat{v}_1\big)| \le Ch^2 (\|\mathbf{u}\|_2 \|\mathbf{v}\| + \|\mathbf{u}\|_1 \|\operatorname{div} \mathbf{v}\|).$$

The argument for $E((\mathbf{B}\hat{\Pi}\hat{\mathbf{u}})_2 \hat{v}_2)$ is analogous. This completes the proof of the lemma. □

We next give the proofs of the two auxiliary lemmas used in the above argument.

Lemma 4.4. *If $\mathbf{K}^{-1} \in W^{2,\infty}(\Omega)$, then for all $e \in \mathcal{T}_h$ there exists a positive constant C independent of h such that*

$$|\mathbf{B}|_{s,\infty,\hat{e}} \le C\, h^s \, \|\mathbf{K}^{-1}\|_{s,\infty,e}, \qquad s = 0,1,2.$$

Proof. First, for a quasi-uniform mesh, we have

$$c_1\, h \le \|\mathbf{DF}\|_{\infty,\hat{e}} \le c_2\, h, \quad c_3\, h^2 \le \|J\|_{\infty,\hat{e}} \le c_4\, h^2$$

with some positive constants c_1–c_4. This implies that

$$\|\mathbf{B}\|_{\infty,\hat{e}} \le C\, \|\hat{\mathbf{K}}^{-1}\|_{\infty,\hat{e}}. \tag{4.50}$$

Second, for an h^2-uniform mesh, we have additional estimates. Let $\alpha = (\alpha_1, \alpha_2)$, $\alpha_i \ge 0$, be a double index, and let $|\alpha| = \alpha_1 + \alpha_2$. In the case $|\alpha| = 1$, the definition of the bilinear mapping (2.4)–(2.6) and (2.2) imply that

$$\|\hat{\partial}^\alpha \mathbf{DF}\|_{\infty,\hat{e}} \le C\, h^2 \qquad \text{and} \qquad \left\| \hat{\partial}^\alpha \frac{1}{J} \mathbf{DF} \right\|_{\infty,\hat{e}} \le C.$$

In the case $|\alpha| = 2$, we have the estimates

$$\|\hat{\partial}^\alpha \mathbf{DF}\|_{\infty,\hat{e}} = 0 \qquad \text{and} \qquad \left\| \hat{\partial}^\alpha \frac{1}{J} \mathbf{DF} \right\|_{\infty,\hat{e}} \le C\, h.$$

As a result, we get

$$\|\hat{\partial}^\alpha \mathbf{B}\|_{\infty,\hat{e}} \le C\, (h \|\hat{\mathbf{K}}^{-1}\|_{\infty,\hat{e}} + \|\hat{\partial}^\alpha \hat{\mathbf{K}}^{-1}\|_{\infty,\hat{e}})$$

for $|\alpha| = 1$ and

$$\|\hat{\partial}^\alpha \mathbf{B}\|_{\infty,\hat{e}} \le C\, (h^2\, \|\hat{\mathbf{K}}^{-1}\|_{\infty,\hat{e}} + h\, \|\hat{\partial}^{\alpha-1} \hat{\mathbf{K}}^{-1}\|_{\infty,\hat{e}} + \|\hat{\partial}^\alpha \hat{\mathbf{K}}^{-1}\|_{\infty,\hat{e}})$$

for $|\alpha| = 2$. Since $\hat{\mathbf{K}}^{-1} = \mathbf{K}^{-1} \circ \mathbf{F}$, using the chain rule and $\|\hat{\partial}^\alpha \mathbf{F}\|_{\infty,\hat{e}} \le C\, h^{|\alpha|}$ for $|\alpha| \le 2$, we obtain

$$\|\hat{\partial}^\alpha \hat{\mathbf{K}}^{-1}\|_{\infty,\hat{e}} \le C\, h^{|\alpha|}\, \|\mathbf{K}^{-1}\|_{|\alpha|,\infty,e}, \quad |\alpha| = 0, 1, 2,$$

which implies

$$\|\hat{\partial}^\alpha \mathbf{B}\|_{\infty,\hat{e}} \le C\, h^{|\alpha|} \|\mathbf{K}^{-1}\|_{|\alpha|,\infty,e}, \quad |\alpha| = 0, 1, 2,$$

completing the proof. □

LEMMA 4.5. *If* $\mathbf{K}^{-1} \in W^{2,\infty}(\Omega)$*, then*

$$\sum_e T_2 = 0, \tag{4.51}$$

where T_2 *is defined in* (4.47).

Proof. Summing over all elements in (4.47), we have

$$\sum_e T_2 = \sum_e \sum_{k=1,3} \int_{\hat{\ell}_k} \phi(\hat{s})\, \frac{\partial}{\partial \hat{s}}((\mathbf{B}\hat{\mathbf{u}}) \cdot \hat{\boldsymbol{\tau}}_k)\hat{\mathbf{v}} \cdot \hat{\mathbf{n}}_k \, d\hat{s}. \tag{4.52}$$

Using (2.7), we have that for any edge ℓ,

$$(\mathbf{B}\hat{\mathbf{u}}) \cdot \hat{\boldsymbol{\tau}} = \frac{1}{J}\mathbf{DF}^T \hat{\mathbf{K}}^{-1}\mathbf{DF}\hat{\mathbf{u}} \cdot |\ell|\mathbf{DF}^{-1}\boldsymbol{\tau} = |\ell|(\mathbf{K}^{-1}\mathbf{u}) \cdot \boldsymbol{\tau}.$$

Therefore, using (2.9), the sum in (4.52) becomes

$$\sum_e T_2 = \sum_e \sum_{k=1,3} |\ell_k|^2 \int_{\ell_k} \phi(s)\, \frac{\partial}{\partial s}\big((\mathbf{K}^{-1}\mathbf{u}) \cdot \boldsymbol{\tau}_k\big)\mathbf{v} \cdot \mathbf{n}_k \, ds. \tag{4.53}$$

Since $\mathbf{v} \in \mathbf{V}^h$, $\mathbf{v} \cdot \mathbf{n} = 0$ on exterior edges and $\mathbf{v} \cdot \mathbf{n}$ is continuous across interior edges. The assumed regularity for $\mathbf{K}$ and $\mathbf{u}$ implies that $\mathbf{K}^{-1}\mathbf{u}$ and $\frac{\partial}{\partial s}(\mathbf{K}^{-1}\mathbf{u})$ are continuous across interior edges. Note that each interior edge ℓ appears twice in the sum in (4.53), which now can be rewritten as a sum of interior edge integrals

$$\sum_e T_2 = \sum_\ell |\ell|^2 \int_\ell \phi(s)\, \frac{\partial}{\partial s}((\mathbf{K}^{-1}\mathbf{u}) \cdot \boldsymbol{\tau})[\mathbf{v} \cdot \mathbf{n}] \, ds = 0,$$

where $[\mathbf{v} \cdot \mathbf{n}]$ denotes the jump in the normal component of $\mathbf{v}$. □

5. Superconvergence to the average edge fluxes and at the edge midpoints. We now discuss how the superconvergence result from section 4 can be applied to obtain superconvergence for the computed velocity to the average edge fluxes and at the midpoints of the edges. Define, for any $\mathbf{v} \in (H^1(\Omega))^2$,

$$\forall e \in \mathcal{T}_h, \quad |||\mathbf{v}|||_e^2 = \sum_{k=1}^{4} \left(\int_{\ell_k} \mathbf{v} \cdot \mathbf{n}_k \, ds \right)^2, \tag{5.1}$$

$$|||\mathbf{v}|||^2 = \sum_{e \in \mathcal{T}_h} |||\mathbf{v}|||_e^2. \tag{5.2}$$

Using the well-known property of the Piola transformation [6],

$$\int_\ell \mathbf{v}\cdot\mathbf{n}\,ds = \int_{\hat\ell} \hat{\mathbf{v}}\cdot\hat{\mathbf{n}}\,d\hat{s} \quad \forall\,\mathbf{v}\in(H^1(\Omega))^2, \tag{5.3}$$

and transforming to the reference element and back, it is easy to see that $|||\cdot|||$ is a norm on $\mathbf{V}^h$ and there exist constants c_1 and c_2 independent of h such that

$$c_1\|\mathbf{v}\| \le |||\mathbf{v}||| \le c_2\|\mathbf{v}\| \quad \forall\,\mathbf{v}\in\mathbf{V}^h.$$

It is clear from (4.2) and (5.3) that $|||\Pi\mathbf{v}-\mathbf{v}|||=0$ for any $\mathbf{v}\in(H^1(\Omega))^2$. Therefore,

$$|||\mathbf{u}-\mathbf{u}^h||| \le |||\Pi\mathbf{u}-\mathbf{u}^h||| \le c_2\|\Pi\mathbf{u}-\mathbf{u}^h\| \le Ch^2\,\|\mathbf{u}\|_2, \tag{5.4}$$

using Theorem 4.2. This implies edgewise superconvergence of the computed velocity $\mathbf{u}^h\cdot\mathbf{n}$ to $\frac{1}{|\ell|}\int_\ell \mathbf{u}\cdot\mathbf{n}\,ds$ in a discrete L^2-sense.

Remark 5.1. *The superconvergence result* (5.4) *implies similar superconvergence for* $|||\mathbf{u}-\mathbf{u}^h|||_M$ *with*

$$|||\mathbf{v}|||_M^2 = \sum_{e\in\mathcal{T}_h}\sum_{k=1}^{4} |\ell_k|^2(\mathbf{v}\cdot\mathbf{n}_k)^2(m_k),$$

where m_k *is the midpoint of* ℓ_k. *Our choice of reporting the results in* $|||\cdot|||$ *is motivated by the fact that average fluxes are easier to measure than pointwise values and therefore are of greater practical interest.*

6. Numerical experiments. In this section, we present the details of the numerical implementation. Instead of solving saddle point problem (2.10), we reduce it to an equivalent system with a symmetric positive definite matrix using the standard hybridization technique.

Let $\mathbf{V}_e^h$ be the restriction of $\mathbf{V}^h$ to quadrilateral e and Λ_ℓ^h be the space of constant functions over edge ℓ. Define

$$\tilde{\mathbf{V}}^h = \prod_e \mathbf{V}_e^h \quad\text{and}\quad \Lambda^h = \prod_\ell \Lambda_\ell^h.$$

Note that the normal component of $\mathbf{v}^h\in\mathbf{V}^h$ is continuous across interior mesh edges and $\mathbf{v}^h\cdot\mathbf{n}=0$ on exterior edges. Therefore,

$$\mathbf{V}^h = \left\{\tilde{\mathbf{v}}^h\in\tilde{\mathbf{V}}^h:\ \sum_e(\mu^h,\,\tilde{\mathbf{v}}^h\cdot\mathbf{n}_e)_{\partial e}=0 \quad \forall\mu^h\in\Lambda^h\right\},$$

where $\mathbf{n}_e$ is the outward normal vector for quadrilateral e.

It has been shown by many authors (see, e.g., [6]) that the original formulation (2.10) is equivalent to the mixed-hybrid formulation: find $(\tilde{\mathbf{u}}^h,\,p^h,\,\lambda^h)\in\tilde{\mathbf{V}}^h\times W^h\times\Lambda^h$ such that

$$\begin{aligned}(\mathbf{K}^{-1}\tilde{\mathbf{u}}^h,\,\tilde{\mathbf{v}}^h)_{h,e}-(p^h,\,\mathrm{div}\,\tilde{\mathbf{v}}^h)_e+(\lambda^h,\,\tilde{\mathbf{v}}^h\cdot\mathbf{n}_e)_{\partial e} &= 0 && \forall\tilde{\mathbf{v}}^h\in\tilde{\mathbf{V}}^h,\\ (\mathrm{div}\,\tilde{\mathbf{u}}^h,\ w^h)_e &= (f,\,w^h)_e && \forall w^h\in W^h,\\ \sum_e(\mu^h,\,\tilde{\mathbf{u}}^h\cdot\mathbf{n}_e)_{\partial e} &= 0 && \forall\mu^h\in\Lambda^h.\end{aligned} \tag{6.1}$$

System (6.1) can be written in the matrix form as

$$\begin{pmatrix} M & B^T & C^T \\ B & 0 & 0 \\ C & 0 & 0 \end{pmatrix} \begin{pmatrix} u \\ p \\ \lambda \end{pmatrix} = \begin{pmatrix} 0 \\ f \\ 0 \end{pmatrix}, \tag{6.2}$$

where

$$D = \begin{pmatrix} M & B^T \\ B & 0 \end{pmatrix}$$

is a block-diagonal matrix (after a permutation of columns and rows) with as many blocks as mesh elements. Each block is a 5×5 matrix. Therefore, vectors u and p can be explicitly eliminated from (6.2) resulting in a system

$$S\,\lambda = b, \tag{6.3}$$

where S is a sparse symmetric positive definite matrix. For logically rectangular meshes, S has at most seven nonzero elements in each row and column. Its nonzero entries represent connections between edge-based unknowns belonging to the same cell.

Problem (6.3) was solved with the preconditioned conjugate gradient (PCG) method. In the numerical experiments, we used one V-cycle of the algebraic multigrid method [24] as a preconditioner. The stopping criterion for the PCG method was the relative decrease in the norm of the residual by a factor of 10^{-12}.

We solved the boundary problem (1.1) with a known analytic solution

$$p(x,\,y) = x^3\,y^2 + x\,\cos(xy)\,\sin(x)$$

and tensor coefficient

$$\mathbf{K}(x,\,y) = \begin{pmatrix} (x+1)^2 + y^2 & -xy \\ -xy & (x+1)^2 \end{pmatrix}.$$

It is pertinent to note here that the superconvergence result established in the previous section for the homogeneous Neumann boundary condition can be extended to the case of general Neumann boundary value problem.

In example 1, the computational domain Ω is the unit square. The computational grid is constructed from a uniform rectangular grid via the mapping

$$x(\xi,\eta) = \xi + 0.06\,\sin(2\pi\eta)\,\sin(2\pi\xi), \qquad y(\xi,\eta) = \eta + 0.06\,\sin(2\pi\eta)\,\sin(2\pi\xi),$$

where $0 < \eta, \xi < 1$, and subsequent random distortion of mesh node positions (see Figure 6.1). The maximum value of the distortion is proportional to the square of the local mesh size; i.e., the resulting grid satisfies assumptions (2.2) and (2.3). We test both Neumann and Dirichlet boundary conditions. The results for the Neumann problem are shown in Table 6.1. The convergence rates were computed using the linear regression for the data in the rows for $1/h = 32, 64, 128, 256$. In addition to norm (5.2), we show the convergence rate in the discrete L_∞-norm:

$$|||\mathbf{u} - \mathbf{u}^h|||_\infty = \max_{\ell_k} \left| \frac{1}{|\ell_k|} \int_{\ell_k} \mathbf{u}\cdot\mathbf{n}_k\,ds \;\, - \mathbf{u}^h\cdot\mathbf{n}_k \right|,$$

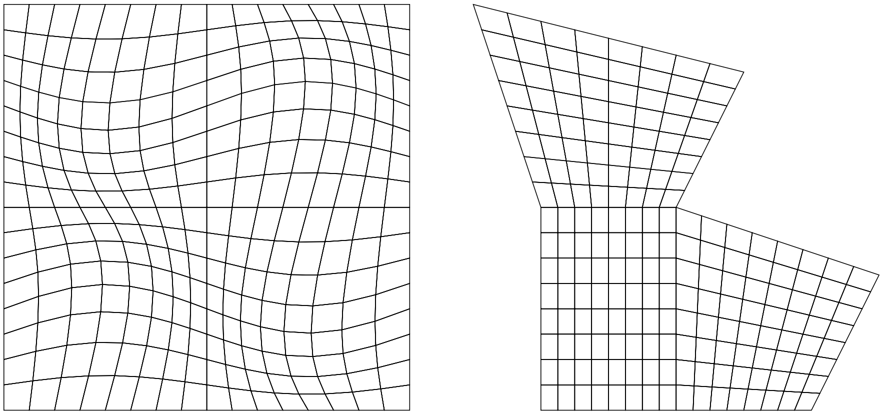


Fig. 6.1. *Examples of meshes used in numerical experiments.*

Table 6.1
Convergence rates for example 1: Neumann boundary conditions.

$1/h$	$\|\|\|\mathbf{u}-\mathbf{u}^h\|\|\|_\infty$	$\|\|\|\mathbf{u}-\mathbf{u}^h\|\|\|$	$\|\|\|p-p^h\|\|\|_\infty$	$\|\|\|p-p^h\|\|\|$
8	8.32e-2	5.47e-2	4.75e-3	1.45e-3
16	2.84e-2	1.69e-2	1.57e-3	3.99e-4
32	8.84e-3	4.49e-3	4.40e-4	1.03e-4
64	2.42e-3	1.14e-3	1.16e-4	2.59e-5
128	6.32e-4	2.87e-4	2.96e-5	6.48e-6
256	1.61e-4	7.17e-5	7.49e-6	1.62e-6
Rate	1.93	1.99	1.96	2.00

where the maximum is taken over all mesh edges. The convergence rates for the pressure variable are shown in the following discrete norms:

$$|||p-p^h|||^2 = \sum_{e_i \in \mathcal{T}_h} |p(c_i) - p^h(c_i)|^2\, |e_i|$$

and

$$|||p-p^h|||_\infty = \max_{e_i \in \mathcal{T}_h} |p(c_i) - p^h(c_i)|,$$

where c_i is the geometric center of element e_i. The use of the geometric center instead of the mass center is due to the following property of the MFD method. The method is exact for linear solutions when the pressure variable, $p(c_i)$, is evaluated at the geometric center c_i [15]. The second-order convergence rate is observed for both the pressure and velocity variables in the discrete L_2- and L_∞-norms.

In the case of Dirichlet boundary conditions, a loss of one half order in the convergence rate for the velocity in the L_2-norm is expected (see, e.g., [12, 3]). The convergence rates are shown in Table 6.2. Note that the velocity convergence rate in the L_2-norm is larger than the theoretical bound of $O(h^{1.5})$. However, the convergence rate in the L_∞-norm is only $O(h)$.

In example 2, the computational domain Ω consists of three quadrilaterals (see Figure 6.1). A sequence of grids is obtained by uniform refinement of these quadrilaterals. The left bottom corner of the domain is located at the point $(1, 0)$. The results

Table 6.2
Convergence rates for example 1: Dirichlet boundary conditions.

$1/h$	$\lvert\lvert\lvert\mathbf{u}-\mathbf{u}^h\rvert\rvert\rvert_\infty$	$\lvert\lvert\lvert\mathbf{u}-\mathbf{u}^h\rvert\rvert\rvert$	$\lvert\lvert\lvert p-p^h\rvert\rvert\rvert_\infty$	$\lvert\lvert\lvert p-p^h\rvert\rvert\rvert$
8	1.50e-1	8.58e-2	5.08e-3	2.08e-3
16	7.20e-2	2.59e-2	1.64e-3	5.53e-3
32	4.24e-2	6.97e-3	4.71e-4	1.42e-4
64	2.39e-2	1.81e-3	1.26e-4	3.57e-5
128	1.27e-2	4.65e-4	3.26e-5	8.95e-6
256	6.55e-3	1.19e-4	8.26e-6	2.24e-6
Rate	0.90	1.96	1.95	2.00

Table 6.3
Convergence rates for example 2: Neumann boundary conditions.

$1/h$	$\lvert\lvert\lvert\mathbf{u}-\mathbf{u}^h\rvert\rvert\rvert_\infty$	$\lvert\lvert\lvert\mathbf{u}-\mathbf{u}^h\rvert\rvert\rvert$	$\lvert\lvert\lvert p-p^h\rvert\rvert\rvert_\infty$	$\lvert\lvert\lvert p-p^h\rvert\rvert\rvert$
8	1.59e-1	1.08e-1	8.84e-3	5.05e-3
16	5.23e-2	2.79e-2	2.74e-3	1.21e-3
32	1.72e-2	7.07e-3	8.33e-4	2.95e-4
64	5.65e-3	1.78e-3	2.26e-4	7.30e-5
128	1.85e-3	4.45e-4	5.84e-5	1.82e-5
256	6.06e-4	1.11e-4	1.48e-5	4.53e-6
Rate	1.61	2.00	1.94	2.01

of our numerical experiments are shown in Table 6.3. We realize that the grid is only locally h^2-uniform. However, the second-order convergence rate for the velocity variable in the L_2 norm is attained.

7. Conclusion. We have proved the superconvergence estimate for the velocity variable on h^2-uniform quadrilateral grids when the exact integration of velocities is replaced by a novel 4-point quadrature rule. The theoretical results for the full diffusion tensor have been confirmed with numerical experiments.

Appendix. Representation of the trapezoidal rule error.

Lemma A.1. *Let $f(x, y)$ be a function defined on a rectangular domain $[a, b] \times [c, d]$. The trapezoidal rule error*

$$E(f) \equiv \int_a^b \int_c^d f(x,y)\,dx\,dy - (f)_T$$

can be represented as

$$\begin{aligned} E(f) = (d-c)\int_a^b \frac{(x-a)(x-b)}{2}\frac{\partial^2}{\partial x^2}f(x,c)\,dx \\ + (b-a)\int_c^d \frac{(y-c)(y-d)}{2}\frac{\partial^2}{\partial y^2}f(a,y)\,dy \\ + \int_a^b\int_c^d \left((x-b)(y-d) - \frac{(b-a)(d-c)}{4}\right)\frac{\partial^2}{\partial x\partial y}f(x,y)\,dx\,dy. \end{aligned}$$

Proof. Define a function

$$g^k(x,s) \equiv (x-s)_+^k \equiv \begin{cases} (x-s)^k, & x \geq s, \\ 0, & x < s, \end{cases}$$

where $k \geq 0$. The Peano kernel theorem (see [23, Theorem 5.2-3, p. 142]) states that the error of the trapezoidal rule is given by

$$\begin{aligned} E(f) = &\int_a^b A_{2,0}(s) f^{(2,0)}(s,c)\,ds \\ &+ \int_c^d A_{0,2}(t) f^{(0,2)}(a,t)\,dt \\ &+ \int_a^b \int_c^d A_{1,1}(s,t) f^{(1,1)}(s,t)\,ds\,dt, \end{aligned} \tag{A.1}$$

where $f^{(i,j)}(x,y) = \frac{\partial^{i+j}}{\partial x^i\,\partial y^j} f(x,y)$ for $i, j \geq 0$ and

$$A_{2,0}(s) = E(g^1(x,s)), \quad A_{0,2}(t) = E(g^1(y,t)), \quad A_{1,1}(s,t) = E(g^0(x,s)g^0(y,t)).$$

Straightforward calculations give

$$\begin{aligned} A_{2,0}(s) &= \int_a^b \int_c^d g^1(s,x)\,dx\,dy - \frac{(b-a)(d-c)}{4} \sum_{j=1}^4 g(x_j, s) \\ &= (d-c)\left(\int_s^b (x-s)\,dx - \frac{b-a}{2}\,(g(a,s) - g(b,s))\right) \\ &= (d-c)\frac{(s-a)(s-b)}{2}. \end{aligned} \tag{A.2}$$

Similarly, we get

$$A_{0,2}(t) = (b-a)\frac{(t-c)(t-d)}{2} \quad \text{and} \quad A_{1,1}(s,t) = (s-b)(t-d) - \frac{(b-a)(d-c)}{4}. \tag{A.3}$$

A substitution of (A.2) and (A.3) into (A.1) completes the proof. □

REFERENCES

[1] T. Arbogast, L. C. Cowsar, M. F. Wheeler, and I. Yotov, *Mixed finite element methods on nonmatching multiblock grids,* SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.

[2] T. Arbogast, C. N. Dawson, P. T. Keenan, M. F. Wheeler, and I. Yotov, *Enhanced cell-centered finite differences for elliptic equations on general geometry,* SIAM J. Sci. Comput., 19 (1998), pp. 404–425.

[3] T. Arbogast, M. F. Wheeler, and I. Yotov, *Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences,* SIAM J. Numer. Anal., 34 (1997), pp. 828–852.

[4] M. Berndt, K. Lipnikov, J. D. Moulton, and M. Shashkov, *Convergence of mimetic finite difference discretizations of the diffusion equation,* J. Numer. Math., 9 (2001), pp. 253–284.

[5] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics,* Cambridge University Press, Cambridge, UK, 1997.

[6] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods,* Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[7] Z. Cai, J. E. Jones, S. F. McCormick, and T. F. Russell, *Control-volume mixed finite element methods,* Comput. Geosci., 1 (1997), pp. 289–315.

[8] J. Campbell and M. Shashkov, *A tensor artificial viscosity using a mimetic finite difference algorithm,* J. Comput. Phys., 172 (2001), pp. 739–765.

[9] S.-H. Chou, D. Y. Kwak, and K. Y. Kim, *A general framework for constructing and analyzing mixed finite volume methods on quadrilateral grids: The overlapping covolume case,* SIAM J. Numer. Anal., 39 (2001), pp. 1170–1196.

[10] J. Douglas, Jr. and J. Wang, *Superconvergence for mixed finite element methods on rectangular domains,* Calcolo, 26 (1989), pp. 121–134.

[11] R. Durán, *Superconvergence for rectangular mixed finite element methods,* Numer. Math., 58 (1990), pp. 287–298.

[12] R. E. Ewing, R. D. Lazarov, and J. Wang, *Superconvergence of the velocity along the Gauss lines in mixed finite element methods,* SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.

[13] R. E. Ewing, M. Liu, and J. Wang, *Superconvergence of mixed finite element approximations over quadrilaterals,* SIAM J. Numer. Anal., 36 (1999), pp. 772–787.

[14] J. M. Hyman and M. Shashkov, *Mimetic discretizations for Maxwell's equations and the equations of magnetic diffusion,* Progr. Electromagn. Res., 32 (2001), pp. 89–121.

[15] J. M. Hyman, M. Shashkov, and S. Steinberg, *The numerical solution of diffusion problems in strongly heterogeneous non-isotropic materials,* J. Comput. Phys., 132 (1997), pp. 130–148.

[16] K. Lipnikov, J. Morel, and M. Shashkov, *Mimetic finite difference methods for diffusion equations on non-orthogonal AMR meshes,* J. Comput. Phys., 199 (2004), pp. 589–597.

[17] L. Margolin, M. Shashkov, and P. Smolarkiewicz, *A discrete operator calculus for finite difference approximations,* Comput. Methods Appl. Mech. Engrg., 187 (2000), pp. 365–383.

[18] J. E. Morel, R. M. Roberts, and M. Shashkov, *A local support-operators diffusion discretization scheme for quadrilateral $r-z$ meshes,* J. Comput. Phys., 144 (1998), pp. 17–51.

[19] M. Nakata, A. Weiser, and M. F. Wheeler, *Some superconvergence results for mixed finite element methods for elliptic problems on rectangular domains,* in The Mathematics of Finite Elements and Applications, V. J. Whiteman, ed., Academic Press, London, 1985.

[20] R. A. Raviart and J. M. Thomas, *A mixed finite element method for 2nd order elliptic problems,* in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.

[21] T. F. Russell and M. F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media,* in The Mathematics of Reservoir Simulation, Frontiers Appl. Math. 1, R. E. Ewing, ed., SIAM, Philadelphia, 1983, pp. 35–106.

[22] M. Shashkov and S. Steinberg, *Solving diffusion equations with rough coefficients in rough grids,* J. Comput. Phys., 129 (1996), pp. 383–405.

[23] A. H. Stroud, *Approximate Calculation of Multiple Integrals,* Prentice–Hall, Englewood Cliffs, NJ, 1971.

[24] K. Stüben, *Algebraic multigrid (AMG): Experiences and comparisons,* Appl. Math. Comput., 13 (1983), pp. 419–452.

[25] J. M. Thomas, *Sur l'analyse numérique des méthods d'éléments finis hybrides et mixtes,* Ph.D. thesis, Université Pierre et Marie Curie, Paris, 1977.

[26] J. Wang and T. P. Mathew, *Mixed finite element method over quadrilaterals,* in Conference on Adv. Numer. Methods and Appl., I. T. Dimov, B. Sendov, and P. Vassilevski, eds., World Scientific, River Edge, NJ, 1994, pp. 203–214.

[27] A. Weiser and M. F. Wheeler, *On convergence of block-centered finite-differences for elliptic problems,* SIAM J. Numer. Anal., 25 (1988), pp. 351–375.

SIAM J. Numer. Anal.
Vol. 43, No. 4, pp. 1750–1765

A C^2 TRIVARIATE MACROELEMENT BASED ON THE WORSEY–FARIN SPLIT OF A TETRAHEDRON*

PETER ALFELD† AND LARRY L. SCHUMAKER‡

Abstract. A C^2 trivariate macroelement is constructed based on the Worsey–Farin split of a tetrahedron into twelve subtetrahedra. The element uses supersplines of degree 9 and provides optimal order approximation of smooth functions.

Key words. tetrahedral splines, macroelements, Bernstein–Bézier methods

AMS subject classifications. 41A63, 41A15, 65D07

DOI. 10.1137/040612609

1. Introduction.

This paper is a companion to our recent paper [5] in which we constructed a C^2 trivariate macroelement based on Clough–Tocher splits of a tetrahedron using polynomials of degree 13 on the subtetrahedra. The purpose of this paper is to describe an alternative C^2 macroelement which works with polynomials of degree 9 instead. To be able to use the lower-degree polynomials, we have to work with a more complicated split. Here we choose the Worsey–Farin split [23]. It divides a tetrahedron into twelve subtetrahedra, as compared with the four subtetrahedra involved in a Clough–Tocher split.

We recall [5] that a trivariate *macroelement* defined on a tetrahedron T consists of a pair $(\mathcal{S}, \Lambda)$, where $\mathcal{S}$ is a space of splines (piecewise polynomial functions) defined on a partition of T into subtetrahedra, and $\Lambda := \{\lambda_i\}_{i=1}^n$ is a set of linear functionals which define values and derivatives of a spline s at certain points in T in such a way that for any given values z_i, there is a unique spline $s \in \mathcal{S}$ with $\lambda_i s = z_i$ for $i = 1, \ldots, n$. These functionals are called the *nodal degrees of freedom* of the element. A macroelement has *smoothness* C^r provided that if the element is used to construct an interpolating spline locally on each tetrahedron of a tetrahedral partition $\triangle$, then the resulting piecewise function is C^r continuous globally. Our aim here is to construct a C^2 macroelement.

The paper is organized as follows. In section 2 we present some background material and notation. The construction of our macroelement for a single tetrahedron is presented in section 3, where we also give a minimal determining set for the space and calculate its dimension. In section 4 we collect several lemmas concerning bivariate spline spaces which are used in our construction. The macroelement space for a Worsey–Farin refinement of an arbitrary tetrahedral partition is discussed in section 5, where again we give a dimension statement and an explicit minimal determining set. Section 6 is devoted to the construction of a nodal determining set for our macroelement space and an associated Hermite interpolation operator along with an error bound for it. We conclude the paper with a number of remarks.

*Received by the editors July 30, 2004; accepted for publication (in revised form) April 12, 2005; published electronically November 22, 2005.

http://www.siam.org/journals/sinum/43-4/61260.html

†Department of Mathematics, University of Utah, 155 South 1400 East, JWB 233, Salt Lake City, UT 84112-0090 (pa@math.utah.edu).

‡Department of Mathematics, Vanderbilt University, Nashville, TN 37240 (s@mars.cas.vanderbilt.edu).

2. Preliminaries. Throughout the paper, we write $\mathcal{P}_d^j$ for the $\binom{d+j}{j}$-dimensional linear space of polynomials of degree d in j variables. Given a tetrahedral partition $\triangle$ of a polyhedral domain Ω, we define

$$\mathcal{S}_d^r(\triangle) := \{s \in C^r(\Omega) : s|_T \in \mathcal{P}_d^3 \quad \text{for all } T \in \triangle\}.$$

In dealing with polynomials and splines, we will use the well-known Bernstein–Bézier methods as used, for example, in [1, 2, 3, 4, 5, 6, 7, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 23, 24]. As usual, given a tetrahedron $T := \langle v_1, v_2, v_3, v_4\rangle$ and a polynomial p of degree d, we denote the B-coefficients of p by $c_{ijkl}^{T,d}$ and associate them with the domain points $\xi_{ijkl}^{T,d} := \frac{(iv_1+jv_2+kv_3+lv_4)}{d}$, where $i+j+k+l = d$. We write $\mathcal{D}_{T,d}$ for the set of all domain points associated with T. We say that the domain point $\xi_{ijkl}^{T,d}$ has *distance $d-i$ from the vertex v_1*, with similar definitions for the other vertices. We say that $\xi_{ijkl}^{T,d}$ is at a *distance $i+j$ from the edge $e := \langle v_3, v_4\rangle$*, with similar definitions for the other edges of T. If $\triangle$ is a tetrahedral partition of a set Ω, we write $\mathcal{D}_{\triangle,d}$ for the collection of all domain points associated with tetrahedra in $\triangle$, where common points in neighboring tetrahedra are not repeated. Given $\xi \in \mathcal{D}_{T,d}$, we denote the associated Bernstein polynomial by $B_\xi^{T,d}$.

Given $\rho > 0$, we refer to the set $D_\rho(v)$ of all domain points which are within a distance ρ from v as the *ball of radius ρ around v*. Similarly, we refer to the set $R_\rho(v)$ of all domain points which are at a distance ρ from v as the *shell of radius ρ around v*. If e is an edge of $\triangle$, we define the *tube of radius ρ around e* to be the set of domain points whose distance to e is at most ρ.

If F is a face of a tetrahedron T, then the domain points in $\mathcal{D}_{T,d}$ which lie on F associated with a trivariate polynomial on T can be considered to be the domain points of a bivariate polynomial of degree d defined on the triangle F. If $F := \langle v_1, v_2, v_3\rangle$ is such a face, we write $\mathcal{D}_{F,d}$ for this set of domain points. As usual, we call the set of points $D_\rho(v_1)$ in $\mathcal{D}_{F,d}$ within a distance ρ from v_1 the *disk of radius ρ around v_1*. Similarly, the set of points $R_\rho(v_1)$ in $\mathcal{D}_{F,d}$ at a distance ρ from v_1 is called the *ring of radius ρ around v_1*. We use the same notation for disks/balls and shells/rings, but the meaning will be clear from the context.

Suppose $\mathcal{S}$ is a linear subspace of $\mathcal{S}_d^0(\triangle)$, and suppose $\mathcal{M}$ is a subset of $\mathcal{D}_{\triangle,d}$. Then $\mathcal{M}$ is said to be a *determining set for $\mathcal{S}$* provided that if $s \in \mathcal{S}$ and its B-coefficients satisfy $c_\xi = 0$ for all $\xi \in \mathcal{M}$, then $s \equiv 0$. It is called a *minimal determining set* (MDS) *for $\mathcal{S}$* provided there is no smaller determining set. It is well known that $\mathcal{M}$ is an MDS for $\mathcal{S}$ if and only if setting the coefficients $\{c_\xi\}_{\xi\in\mathcal{M}}$ of a spline in $\mathcal{S}$ uniquely determines all coefficients of s. It is also known that the cardinality of any MDS for $\mathcal{S}$ equals the dimension of $\mathcal{S}$.

Now suppose $\mathcal{N}$ is a collection of linear functionals λ, where λs is defined by a combination of values or derivatives of s at a point η_λ in Ω. Then $\mathcal{N}$ is said to be a *nodal determining set* (NDS) *for $\mathcal{S}$* provided that if $s \in \mathcal{S}$ and $\lambda s = 0$ for all $\lambda \in \mathcal{N}$, then $s \equiv 0$. It is called a *nodal minimal determining set* (NMDS) *for $\mathcal{S}$* provided that there is no smaller NDS or, equivalently, for each set of real numbers $\{z_\lambda\}_{\lambda\in\mathcal{N}}$, there exists a unique $s \in \mathcal{S}$ such that $\lambda s = z_\lambda$ for all $\lambda \in \mathcal{N}$.

3. The basic macroelement on one tetrahedron. Given a tetrahedron $T := \langle v_1, v_2, v_3, v_4\rangle$, let v_T be a point in the interior of T. In this section we take v_T to be an arbitrary point in T, but to obtain a C^2 macroelement space on a general tetrahedral partition, we need to be more careful in the selection of v_T; see section 5 below. In addition, for each face F of T, let v_F be a point in the interior of F. For tetrahedral partitions with more than one tetrahedron, we will also have to choose these points in a special way. Suppose now that we connect v_T to each vertex v of T and to each point v_F, and we connect each v_F to the vertices of the face in which it lies. Then T is split into 12 subtetrahedra. This split was used in [23] to construct a C^1 piecewise cubic trivariate macroelement. We refer to it as the *Worsey–Farin split* and denote it by T_{WF}. We write $\mathcal{V}_T$, $\mathcal{E}_T$, and $\mathcal{F}_T$ for the sets of vertices, edges, and faces of T. Let $\mathcal{E}_T^c$ be the set of four edges connecting v_T to the face points v_F, and for each $F := \langle v_1, v_2, v_3\rangle \in \mathcal{F}_T$, let $\mathcal{E}_F$ be the set of three oriented edges $\langle v_i, v_F\rangle$, $i = 1, 2, 3$. We write $\mathcal{F}_T^0$ for the set of 12 faces of $\triangle_{WF}$ of the form $\langle v_T, v_F, v\rangle$, where $v \in \mathcal{V}_T$.

We need some additional notation before introducing our basic macroelement. Suppose $t := \langle v_T, v_F, v_1, v_2\rangle$ and $\tilde{t} := \langle v_T, v_F, v_2, v_3\rangle$ are two tetrahedra in T_{WF} which share the face $F := \langle v_T, v_F, v_2\rangle \in \mathcal{F}_T^0$. Let c_{ijkl} and $\tilde{c}_{ijkl}$ be the coefficients of the B-representations of $s|_t$ and $s|_{\tilde{t}}$, respectively. Then we define the linear functionals ν_F and μ_F by

$$\begin{aligned} \nu_F s &:= \tilde{c}_{0,1,3,5} - \sum_{i+j+k=5} c_{0,i+1,j,k+3} B^{t,5}_{ijk}(v_3), \\ \mu_F s &:= \tilde{c}_{1,0,3,5} - \sum_{i+j+k=5} c_{1,i,j,k+3} B^{t,5}_{ijk}(v_3), \end{aligned} \tag{3.1}$$

where $B^{t,5}_{ijk}$ are the Bernstein polynomials of degree 5 with respect to the triangle $\langle v_F, v_1, v_2\rangle$. Note that $\nu_F s$ involves coefficients of s on the shell $R_9(v_T)$, while $\mu_F s$ involves coefficients of s on the shell $R_8(v_T)$.

We now introduce our basic macroelement space as the following space of supersplines defined on T_{WF}:

$$\begin{aligned} \mathcal{S}_2(T_{WF}) := \{ s \in C^2(T) : s|_t \in \mathcal{P}_9^3 &\text{ for all } t \in T_{WF}, \\ s \in C^3(e) &\text{ for all } e \in \mathcal{E}_T, \\ s \in C^7(e) &\text{ for all } e \in \mathcal{E}_T^c, \\ \nu_F s = \mu_F s = 0 &\text{ for all } F \in \mathcal{F}_T^0, \\ s \in C^4(v) &\text{ for all } v \in \mathcal{V}_T, \\ s \in C^7(v_T) \}. \end{aligned} \tag{3.2}$$

As usual, if v is a vertex of T_{WF}, then $s \in C^\rho(v)$ means that all polynomial pieces of s defined on tetrahedra sharing the vertex v have common derivatives up to order ρ at v. If e is an edge of T_{WF}, then $s \in C^\mu(e)$ means that all subpolynomials of s defined on tetrahedra sharing the edge e have common derivatives up to order μ on e.

Before proceeding, we first make some remarks about our fairly complicated definition of $\mathcal{S}_2(T_{WF})$. The construction is the result of a considerable amount of experimentation with the first author's Java code for working with trivariate splines; see Remark 6. In creating $\mathcal{S}_2(T_{WF})$, we had two aims in mind: to create a macroelement which will be globally C^2 smooth, and to minimize the complexity and number of degrees of freedom. First, we observe that we are forced to impose the C^4 supersmoothness at the vertices of T, since otherwise we could not make macroelements on adjoining tetrahedra join with C^2 smoothness; see Remark 4. Since derivatives up to order 4 at the vertices are not allowed to interfere (or, equivalently, balls of radius 4 around the vertices are not allowed to overlap), this forces us to use polynomials of degree (at least) 9. The additional supersmoothness in the definition of $\mathcal{S}_2(T_{WF})$ has been imposed in order to remove unnecessary degrees of freedom from our macroelement. While other choices are possible, we found that this choice is the most symmetric, while at the same time providing stable computations.

For each vertex v of T, let T_v be one of the tetrahedra in T_{WF} attached to v. For each edge $e := \langle u, v\rangle$ of T, let T_e be one of the two tetrahedra containing e, and let $E_3(e)$ denote the set of domain points in the tube of radius 3 around e which do not lie in the balls $D_4(u)$ or $D_4(v)$. Finally, for each face $F := \langle v_1, v_2, v_3\rangle$ of T, let $T_{F,i} := \langle v_{_T}, v_{_F}, v_i, v_{i+1}\rangle$, $i = 1, 2, 3$, where we set $v_4 := v_1$.

Theorem 3.1. *The space $\mathcal{S}_2(T_{WF})$ has dimension* 292. *Moreover,*

$$\mathcal{M} := \bigcup_{v\in\mathcal{V}_T} \mathcal{M}_v \cup \bigcup_{e\in\mathcal{E}_T} \mathcal{M}_e \cup \bigcup_{F\in\mathcal{F}_T} \mathcal{M}_F \cup \mathcal{M}_T \tag{3.3}$$

is an MDS for $\mathcal{S}_2(T_{WF})$, where

(1) $\mathcal{M}_v := D_4(v) \cap T_v$,

(2) $\mathcal{M}_e := E_3(e) \cap T_e$,

(3) $\mathcal{M}_F := \{\xi^{T_{F,1}}_{2430}, \xi^{T_{F,2}}_{2430}, \xi^{T_{F,3}}_{2430}\}$,

(4) $\mathcal{M}_T := D_3(v_{_T}) \cap T_{v_{_T}}$.

Proof. We shall show that $\mathcal{M}$ is an MDS for $\mathcal{S}_2(T_{WF})$, which in turn implies that the dimension of $\mathcal{S}_2(T_{WF})$ is just the cardinality of $\mathcal{M}$. The cardinalities of the sets $\mathcal{M}_v$, $\mathcal{M}_e$, $\mathcal{M}_F$, $\mathcal{M}_T$ are 35, 20, 3, and 20, respectively. Since T has four vertices, six edges, and four faces, it follows that the dimension of $\mathcal{S}_2(T_{WF})$ is $4 \times 35 + 6 \times 20 + 4 \times 3 + 20 = 292$.

To show that $\mathcal{M}$ is an MDS for $\mathcal{S}_2(T_{WF})$, we need to show that setting the coefficients $\{c_\xi\}_{\xi\in\mathcal{M}}$ of a spline $s \in \mathcal{S}_2(T_{WF})$ *consistently* determines all other coefficients of s. First, for each vertex $v \in \mathcal{V}_T$, the C^4 smoothness at v implies that all coefficients corresponding to domain points in $D_4(v)$ are consistently determined. Moreover, for each edge $e \in \mathcal{E}_T$, the C^3 smoothness around e implies that the coefficients of s in the tube of radius 3 around e are consistently determined.

We now examine the coefficients corresponding to domain points on the shell $R_9(v_{_T})$, i.e., on the outer faces of T_{WF}. Let $F := \langle v_1, v_2, v_3\rangle$ be a face of this shell. We can consider the coefficients of s corresponding to the domain points on F (see Figure 1 (left)) as the coefficients of a bivariate spline $g := s|_F$ in the space $\widetilde{\mathcal{S}}^2_9(F_{CT})$ defined in (4.2) below, where F_{CT} is the Clough–Tocher split of F into three subtriangles. By the above, it is clear that all coefficients of g corresponding to the domain points marked with dots or triangles in Figure 1 (left) are already determined. But then, by Lemma 4.1 below, all other coefficients of g are determined. Repeating this argument for each face of $R_9(v_{_T})$, we conclude that the coefficients of s are determined for all domain points on the shell $R_9(v_{_T})$.

Now consider the coefficients of s corresponding to domain points lying on the shell $R_8(v_{_T})$. For each face $F := \langle v_1, v_2, v_3\rangle$ of this shell, we can consider the B-coefficients of s corresponding to domain points on F (see Figure 1 (right)) to be the coefficients of a bivariate spline g in the space $\widetilde{\mathcal{S}}^2_8(F_{CT})$ defined in (4.5) below. It is clear from the above that all coefficients of g corresponding to domain points marked with dots or triangles in Figure 1 (right) are already determined. But then by Lemma 4.2 below, all other coefficients of g are determined. Repeating this argument for each face of $R_8(v_{_T})$, we conclude that the coefficients of s are determined for all domain points on the shell $R_8(v_{_T})$.

Next we consider the shell $R_7(v_{_T})$. Let F be a face of this shell. We can consider the coefficients of s corresponding to domain points on F to be the coefficients of a bivariate spline g in $\mathcal{S}^2_7(F_{CT}) \cap C^7(v_{_F})$, which means that g is actually a polynomial

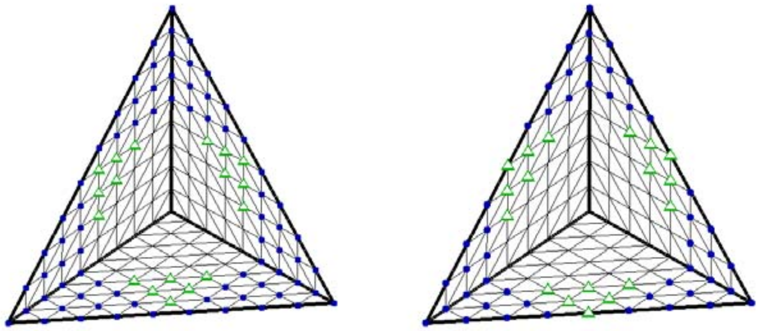


FIG. 1. *Domain points of $\mathcal{S}_2(T_{WF})$ on faces of $R_9(v_T)$ and $R_8(v_T)$, respectively.*

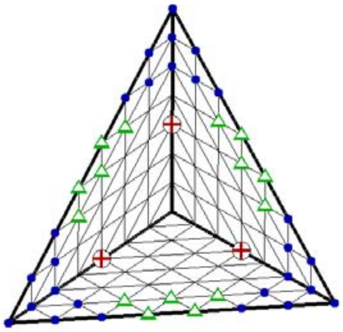


FIG. 2. *Domain points of $\mathcal{S}_2(T_{WF})$ on a face of $R_7(v_T)$.*

of degree 7. All coefficients corresponding to domain points marked with dots or triangles in Figure 2 are already determined. In addition, those corresponding to $\oplus$ are also determined as those points are in $\mathcal{M}$. It now follows from Lemma 4.3 below that all other coefficients of g are determined. Repeating this argument for each face of $R_7(v_T)$, we conclude that all coefficients of s corresponding to domain points on the shell $R_7(v_T)$ are determined.

To show that the coefficients of s corresponding to the remaining domain points in T_{WF} are determined, we note that by the C^7 smoothness at v_T, we may consider the B-coefficients of s corresponding to domain points in the ball $D_7(v_T)$ as those of a trivariate polynomial g of degree 7 considered as a spline in $\mathcal{S}_7^2(t_{WF})$, where t_{WF} is the Worsey–Farin split of the tetrahedron t whose vertices are the vertices of $D_7(v_T)$. By the above, g is determined on the faces of t. Now setting the coefficients $\{c_\xi\}_{\xi\in\mathcal{M}_T}$ is equivalent to setting the derivatives of g up to order 3 at the point v_T. But then Lemma 3.2 below shows that this combined information determines g. □

LEMMA 3.2. *Suppose $g \in \mathcal{P}_7^3$ and let $\{c_\xi\}_{\xi\in\mathcal{D}_{t,7}}$ be its set of B-coefficients relative to a tetrahedron t. Suppose we are given values for the coefficients corresponding to the set of domain points lying on the faces of t. Let w be any point in the interior of t. Then the remaining coefficients of g are uniquely determined by the values of $\{D^\alpha g(w)\}_{|\alpha|\le 3}$.*

Proof. The set of domain points in $\mathcal{D}_{t,7}$ which do not lie on the faces of t is

$\Gamma := \{\xi^{t,7}_{ijkl} : i, j, k, l \geq 1\}$. The cardinality of this set is 20. Now the equations

$$D^\alpha g(w) = z_\alpha, \quad |\alpha| \leq 3,$$

provide a linear system of 20 equations for the $\{c_\xi\}_{\xi\in\Gamma}$. We claim that this system is nonsingular. To see this, we show that if g is zero on the faces of t and $z_\alpha = 0$ for all $|\alpha| \leq 3$, then $g \equiv 0$. The fact that g vanishes on faces implies that it can be written as $g = \ell_1\ell_2\ell_3\ell_4 q$, where $q \in \mathcal{P}^3_3$, and where for $i = 1, 2, 3$, ℓ_i is a nontrivial linear polynomial which vanishes on the ith face of t. But now the condition $D^\alpha g(w) = 0$ for $|\alpha| \leq 3$ implies $D^\alpha q(w) = 0$ for $|\alpha| \leq 3$, which implies $q \equiv 0$, which in turn implies that $g \equiv 0$. □

4. Some bivariate lemmas. In this section we establish some properties of certain bivariate spline spaces defined on the well-known Clough–Tocher split of a triangle $F := \langle v_1, v_2, v_3\rangle$ in $\mathbb{R}^2$. Given $v_{\scriptscriptstyle F}$ in the interior of F, we connect it to all three vertices of F to split it into three subtriangles $F_i := \langle v_{\scriptscriptstyle F}, v_i, v_{i+1}\rangle$. Let $e_i := \langle v_i, v_{i+1}\rangle$ and $\tilde{e}_i := \langle v_i, v_{\scriptscriptstyle F}\rangle$, $i = 1, 2, 3$, where $v_4 := v_1$. Note that in this section we do not make any special assumptions about the location of $v_{\scriptscriptstyle F}$, just that it be in the interior of F. For $d \geq 2$, let

$$\mathcal{S}^2_d(F_{CT}) := \{s \in C^2(F) : s|_{F_i} \in \mathcal{P}^2_d,\ i = 1, 2, 3\}.$$

Given $1 \leq l \leq 3$, suppose $\{c_{ijk}\}$ and $\{\tilde{c}_{ijk}\}$ are the coefficients of $s \in \mathcal{S}^2_d(F_{CT})$ relative to F_{l-1} and F_l, respectively, where we identify $v_4 = v_1$. Then we define the linear functional $\tau^n_{l,m,d}$ by

$$\tau^n_{l,m,d}s := \tilde{c}_{m-n,d-m,n} - \sum_{i+j+k=n} c_{i+m-n,j,k+d-m} B^{l,n}_{ijk}(v_{l+1}), \tag{4.1}$$

where $B^{l,n}_{ijk}$ are the Bernstein polynomials of degree n relative the triangle F_{l-1}. Note that $\tau^n_{l,m,d}$ describes an individual C^n smoothness condition involving the coefficients on ring $R_m(v_l)$.

LEMMA 4.1. *Let*

$$\widetilde{\mathcal{S}}^2_9(F_{CT}) := \{s \in \mathcal{S}^2_9(F_{CT}) \cap C^7(v_{\scriptscriptstyle F}) : s \in C^4(v_{\scriptscriptstyle l}) \text{ and } \tau^5_{l,6,9}s = 0,\ l = 1, 2, 3\}. \tag{4.2}$$

Then $\dim \widetilde{\mathcal{S}}^2_9(F_{CT}) = 63$, *and the set*

$$\mathcal{M}_9 := \bigcup_{i=1}^{3} (\mathcal{M}_{v_i} \cup \mathcal{M}_{e_i})$$

is an MDS for $\widetilde{\mathcal{S}}^2_9(F_{CT})$, *where*

(1) $\mathcal{M}_v := D_4(v) \cap t_v$, *where* t_v *is some triangle of* F_{CT} *attached to* v*;*

(2) $\mathcal{M}_e$ *is the set of domain points whose distance to* $e := \langle u, v\rangle$ *is at most* 3 *and which do not lie in the disks* $D_4(u)$ *or* $D_4(v)$.

Proof. Points in the sets $\mathcal{M}_e$ are marked with small triangles in Figure 1 (left), while points in the disks $D_4(v)$ are marked with dots. By Theorem 2.2 in [19], $\dim \mathcal{S}^2_9(F_{CT}) \cap C^7(v_{\scriptscriptstyle F}) = 75$. To get the subspace $\widetilde{\mathcal{S}}^2_9(F_{CT})$, for each $l = 1, 2, 3$ we have to enforce three extra smoothness conditions at the vertex v_l to get $C^4(v_l)$ as well as the special smoothness condition corresponding to $\tau^5_{l,6,9}$. It follows that

$\dim \mathcal{S}_9^2(F_{CT}) \geq 63$. Since the cardinality of $\mathcal{M}_9$ is 63, to show that $\mathcal{M}_9$ is an MDS for $\widetilde{\mathcal{S}}_9^2(F_{CT})$ and $\dim \widetilde{\mathcal{S}}_9^2(F_{CT}) = 63$, it suffices to show that if s is a spline in $\widetilde{\mathcal{S}}_9^2(F_{CT})$ whose coefficients satisfy $c_\xi = 0$ for all $\xi \in \mathcal{M}_9$, then $s \equiv 0$. By the definition of $\mathcal{M}_9$, it is clear that all coefficients marked with dots or triangles in Figure 1 (left) are zero. We now examine the coefficients corresponding to the remaining domain points.

First consider the ring $R_5(v_1)$. All coefficients corresponding to domain points on this ring are already zero except for the three corresponding to domain points within a distance 1 of the edge $\tilde{e}_1$. To compute these three coefficients, we proceed as in Lemma 3.3 of [9] and Lemma 2.1 of [4]. The C^7 smoothness at v_F implies that s satisfies individual C^1, C^2, and C^3 continuity conditions on ring $R_5(v_1)$, i.e., $\tau_{1,5,9}^n s = 0$ for $n = 1, 2, 3$. This leads to a linear system of equations with matrix

$$M_3 := \begin{pmatrix} a_2 & a_1 & -1 \\ 2a_2a_1 & a_1^2 & 0 \\ 3a_2a_1^2 & a_1^3 & 0 \end{pmatrix}, \tag{4.3}$$

where (a_1, a_2, a_3) are the barycentric coordinates of v_3 relative to the triangle F_1. This matrix is nonsingular since its determinant is $-a_2a_1^4$ and a_1, a_2 are both nonzero. Coefficients on the rings $R_5(v_2)$ and $R_5(v_3)$ can be computed in a similar way.

Now consider the ring $R_6(v_1)$. At this point, all coefficients corresponding to domain points on the ring $R_6(v_1)$ are determined to be zero except for the five corresponding to domain points within a distance 2 of $\tilde{e}_1$. Now the C^7 smoothness at v_F implies that s satisfies individual C^1 through C^4 smoothness conditions on ring $R_6(v_1)$. Coupling this with the special smoothness condition $\tau_{1,6,9}^5 s = 0$, we are led to the system of equations $\tau_{1,6,9}^n s = 0$ for $n = 1, \ldots, 5$. The matrix of this system is

$$M_5 := \begin{pmatrix} 0 & a_2 & a_1 & -1 & 0 \\ a_2^2 & 2a_2a_1 & a_1^2 & 0 & -1 \\ 3a_2^2a_1 & 3a_2a_1^2 & a_1^3 & 0 & 0 \\ 6a_2^2a_1^2 & 4a_2a_1^3 & a_1^4 & 0 & 0 \\ 10a_2^2a_1^3 & 5a_2a_1^4 & a_1^5 & 0 & 0 \end{pmatrix}. \tag{4.4}$$

This is a nonsingular matrix since its determinant is equal to $-a_2^3a_1^9$. Coefficients on the rings $R_6(v_2)$ and $R_6(v_3)$ can be computed in a similar way. Now all remaining coefficients of s can be computed from the smoothness conditions by solving similar nonsingular 5×5 systems. We conclude that all coefficients of s must be zero, which completes the proof of the lemma. ◻

Lemma 4.2. *Let*

$$\widetilde{\mathcal{S}}_8^2(F_{CT}) := \{s \in \mathcal{S}_8^2(F_{CT}) \cap C^7(v_F) : s \in C^3(v_l) \text{ and } \tau_{l,5,8}^5 s = 0,\ l = 1, 2, 3\}. \tag{4.5}$$

Then $\dim \widetilde{\mathcal{S}}_8^2(F_{CT}) = 48$, *and the set*

$$\mathcal{M}_8 := \bigcup_{i=1}^{3} (\mathcal{M}_{v_i} \cup \mathcal{M}_{e_i})$$

is an MDS for $\widetilde{\mathcal{S}}_8^2(F_{CT})$, *where*

(1) $\mathcal{M}_v := D_3(v) \cap t_v$, *where* t_v *is some triangle of* F_{CT} *attached to* v*;*

(2) $\mathcal{M}_e$ *is the set of domain points whose distance to* $e := \langle u, v\rangle$ *is at most* 2 *and which do not lie in the disks* $D_3(u)$ *or* $D_3(v)$.

Proof. The proof is very similar to proof of Lemma 4.1, so we can be brief. By Theorem 2.2 in [19], $\dim \mathcal{S}_8^2(F_{CT}) \cap C^7(v_{\scriptscriptstyle F}) = 54$. To get the subspace $\widetilde{\mathcal{S}}_8^2(F_{CT})$, for each $l = 1, 2, 3$, we have to enforce one extra smoothness condition at the vertex v_l to get $C^3(v_l)$ along with the special smoothness condition corresponding to $\tau_{l,5,8}^5$. It follows that $\dim \widetilde{\mathcal{S}}_8^2(F_{CT}) \geq 48$. Since the cardinality of $\mathcal{M}_8$ is 48, to show that it is an MDS for $\widetilde{\mathcal{S}}_8^2(F_{CT})$ and $\dim \widetilde{\mathcal{S}}_8^2(F_{CT}) = 48$, it suffices to show that if $c_\xi = 0$ for all $\xi \in \mathcal{M}_8$, then $s \equiv 0$. We already know that all coefficients of s corresponding to domain points marked with dots or triangles in Figure 1 (right) are zero. But then the remaining coefficients can be computed from the same linear systems as in Lemma 4.1. □

Lemma 4.3. *The set*

$$\mathcal{M}_7 := \mathcal{M}_F \cap \bigcup_{i=1}^{3} (\mathcal{M}_{v_i} \cup \mathcal{M}_{e_i})$$

is an MDS for $\mathcal{P}^2 = \mathcal{S}_7^2(F_{CT}) \cap C^7(v_{\scriptscriptstyle F})$, *where*

(1) $\mathcal{M}_v := D_2(v) \cap t_v$, *where* t_v *is some triangle of* F_{CT} *attached to* v*;*

(2) $\mathcal{M}_e$ *is the set of domain points whose distance to* $e := \langle u, v\rangle$ *is at most* 1 *and which do not lie in the disks* $D_2(u)$ *or* $D_2(v)$*;*

(3) $\mathcal{M}_F := \{\xi_{430}^{F_1}, \xi_{430}^{F_2}, \xi_{430}^{F_3}\}$.

Proof. Points in $\mathcal{M}_F$ are marked with $\oplus$ in Figure 2, while points in $\mathcal{M}_e$ are marked with small triangles. Points in the disks $D_2(v)$ are marked with dots. The dimension of $\mathcal{P}_7^2$ is 36 and the cardinality of $\mathcal{M}$ is also 36. Thus, it suffices to show that $\mathcal{M}$ is a determining set. Suppose $s \in \mathcal{P}_7^2$, and $c_\xi = 0$ for all $\xi \in \mathcal{M}$. This means that the B-coefficients of s corresponding to all marked domain points in Figure 2 (left) are zero. First, we note that the coefficients corresponding to the three remaining domain points on $R_3(v_1)$ can be computed from a nonsingular 3×3 linear system with the matrix M_3 given in (4.3). The same holds for the rings $R_3(v_2)$ and $R_3(v_3)$. Now consider $R_4(v_1)$. There are four unknown coefficients corresponding to the unmarked points on this ring, and they can be computed from a system of four equations with the matrix

$$M_4 := \begin{pmatrix} 0 & a_2 & -1 & 0 \\ a_2^2 & 2a_2a_1 & 0 & -1 \\ 3a_2^2a_1 & 3a_2a_1^2 & 0 & 0 \\ 6a_2^2a_1^2 & 4a_2a_1^3 & 0 & 0 \end{pmatrix}.$$

The determinant of this matrix is $-6a_1^4a_2^3 \neq 0$. We can repeat this for the other two vertices v_2, v_3. The remaining coefficients of s are then determined exactly as in Lemmas 4.1 and 4.2. □

5. The macroelement space $\boldsymbol{\mathcal{S}_2(\triangle_{WF})}$.

We now show that the construction of the previous section can be used to define a C^2 macroelement space defined on a general tetrahedral partition, provided that the split points $v_{\scriptscriptstyle T}$ and $v_{\scriptscriptstyle F}$ are chosen appropriately. Suppose $\triangle$ is an arbitrary tetrahedral partition of a polyhedral domain Ω, and that the points $v_{\scriptscriptstyle T}$ are chosen so that for any pair of tetrahedra sharing a common face F, the line connecting the center points passes through the interior of

F. This can be insured, for example, by taking v_T to be the centers of the inscribed balls in each tetrahedron T; see [23]. We now take $\triangle_{WF}$ to be the refined partition obtained by applying the Worsey–Farin split to each tetrahedron in $\triangle$, where, for every face F shared by two tetrahedra T and $\widetilde{T}$, the split point v_F on F is taken to be the intersection of F with the line connecting v_T and $v_{\tilde{T}}$.

Let $\mathcal{V}$, $\mathcal{E}$, and $\mathcal{F}$ be the sets of vertices, edges, and faces of $\triangle$, respectively. Let V, E, F be the cardinalities of these sets, and denote the number of tetrahedra in $\triangle$ by N_T. We write $\mathcal{F}^0 = \bigcup_{T\in\triangle} \mathcal{F}_T^0$, where $\mathcal{F}_T^0$ is defined in section 3. Let $\mathcal{E}^c := \bigcup_{T\in\triangle} \mathcal{E}_T^c$, where $\mathcal{E}_T^c$ is also defined in section 3. We now define the following *macroelement space*:

$$
\begin{aligned}
\mathcal{S}_2(\triangle_{WF}) := \{ & s \in C^2(\Omega) : s|_t \in \mathcal{P}_9^3 \ \text{ for all } t \in \triangle_{WF}, \\
& s \in C^3(e) \ \text{ for all } e \in \mathcal{E}, \\
& s \in C^7(e) \ \text{ for all } e \in \mathcal{E}^c, \\
& \nu_F s = \mu_F s = 0 \ \text{ for all } F \in \mathcal{F}^0, \\
& s \in C^4(v) \ \text{ for all } v \in \mathcal{V}, \\
& s \in C^7(v_T) \ \text{ for all } T \in \triangle \}.
\end{aligned}
\tag{5.1}
$$

To define an MDS for $\mathcal{S}_2(\triangle_{WF})$ we need some more notation. For each vertex v of $\triangle$, let T_v be one of the tetrahedra in $\triangle_{WF}$ attached to v. For each edge $e := \langle u, v\rangle$ of $\triangle$, let T_e be one of the tetrahedra containing e, and let $E_3(e)$ denote the set of domain points in the tube of radius 3 around e which do not lie in the balls $D_4(u)$ or $D_4(v)$. Finally, for each face $F := \langle v_1, v_2, v_3\rangle$ of $\triangle$, let $T_{F,i} := \langle v_T, v_F, v_i, v_{i+1}\rangle$, $i = 1, 2, 3$, where v_T is the split point of some tetrahedron in $\triangle$ containing F (if F is a boundary face, there is just one such tetrahedron—otherwise, there are two).

THEOREM 5.1. *The space* $\mathcal{S}_2(\triangle_{WF})$ *has dimension* $35V + 20E + 3F + 20N_T$. *Moreover, the set*

$$
\mathcal{M} := \bigcup_{v\in\mathcal{V}} \mathcal{M}_v \cup \bigcup_{e\in\mathcal{E}} \mathcal{M}_e \cup \bigcup_{F\in\mathcal{F}} \mathcal{M}_F \cup \bigcup_{T\in\triangle} \mathcal{M}_T
\tag{5.2}
$$

is an MDS for $\mathcal{S}_2(\triangle_{WF})$, *where*

(1) $\mathcal{M}_v := D_4(v) \cap T_v$,
(2) $\mathcal{M}_e := E_3(e) \cap T_e$,
(3) $\mathcal{M}_F := \{\xi_{2430}^{T_{F,1}}, \xi_{2430}^{T_{F,2}}, \xi_{2430}^{T_{F,3}}\}$,
(4) $\mathcal{M}_T := D_3(v_T) \cap T_{v_T}$.

Proof. We shall show that $\mathcal{M}$ is an MDS for $\mathcal{S}_2(\triangle_{WF})$. This implies that the dimension of $\mathcal{S}_2(\triangle_{WF})$ is just the cardinality of $\mathcal{M}$, which is easily seen to be equal to the given formula.

To show that $\mathcal{M}$ is an MDS for $\mathcal{S}_2(T_{WF})$, we need to show that if $s \in \mathcal{S}_2(T_{WF})$, then we can set the coefficients $\{c_\xi\}_{\xi\in\mathcal{M}}$ to arbitrary values, and all other coefficients will be *consistently* determined. First, since the balls $D_4(v)$ do not overlap, it is clear that we can set all of the coefficients corresponding to the sets $\mathcal{M}_v$ to arbitrary values, and then by the C^4 smoothness at vertices, all other coefficients corresponding to domain points in balls $D_4(v)$ will be consistently determined. Similarly, since the sets $E_3(e)$ do not overlap each other or any of the balls $D_4(v)$, we can set all of the coefficients corresponding to the sets $\mathcal{M}_e$ to arbitrary values, and then by the C^3 smoothness around edges, all other coefficients corresponding to domain points in the sets $E_3(e)$ will be consistently determined.

Now we can use Lemma 4.1 to compute coefficients corresponding to the remaining domain points on the faces of the shells $R_9(v_T)$ for all T. For interior faces F, this means computing the same coefficients twice, once for each tetrahedron sharing F. But we will get the same values since these coefficients are computed in the same way using only known coefficients associated with domain points on F.

We can now use Lemma 4.2 to uniquely compute coefficients corresponding to the remaining domain points on the faces of the shells $R_8(v_{_T})$ for all T. But now we have to check that if $T := \langle v_{_T}, v_1, v_2, v_3\rangle$ and $\widetilde{T} := \langle v_{_{\tilde{T}}}, v_1, v_2, v_3\rangle$ are two tetrahedra sharing a face $F := \langle v_1, v_2, v_3\rangle$, then these computed coefficients satisfy all C^1 smoothness conditions across F. Note that the split point $v_{_F}$ lies on the line from $v_{_T}$ to $v_{_{\tilde{T}}}$. Let $g := s|_T$ and $\tilde{g} := s|_{\widetilde{T}}$. Consider the typical subtriangle $f := \langle v_{_F}, v_1, v_2\rangle$ of F_{CT}. By the geometry, each of the C^1 smoothness conditions involving coefficients associated with domain points in f reduces to a relationship of the form

$$b = sc + rd,$$

where $(r, s, 0, 0)$ are the barycentric coordinates of $v_{_T}$ with respect to the tetrahedron $\langle v_{_{\tilde{T}}}, v_{_F}, v_1, v_2\rangle$. Here b is a coefficient of g corresponding to a domain point ξ_b in t which lies at a distance 1 from F, i.e., in $F_8 := R_8(v_{_T}) \cap F$; see Figure 1 (right). Similarly, d is a coefficient of $\tilde{g}$ corresponding to a domain point ξ_d in $\tilde{t}$ which lies at a distance 1 from F, i.e., in $\widetilde{F}_8 := R_8(v_{_{\tilde{T}}}) \cap F$. The coefficient c is a coefficient of g corresponding to the domain point on F which lies on the straight line between ξ_b and ξ_d. Let Γ_8 be the set of $n := 66$ domain points in Figure 1 (right) marked with either a dot or a triangle. Let $\{b_i\}_{i=1}^n$ and $\{d_i\}_{i=1}^n$ be the corresponding coefficients of g and $\tilde{g}$, respectively, and let $\{c_i\}_{i=1}^n$ be the coefficients of g corresponding to the associated domain points on F; see Figure 1 (left). Then by the smoothness of s at vertices and around edges, it is clear that all C^1 continuity conditions with tips at points in Γ_8 are satisfied, i.e.,

$$b_i = sc_i + rd_i, \quad i = 1, \dots, n. \tag{5.3}$$

Now let ξ be any other domain point in Figure 1 (right), and let b, c, d be the coefficients entering into the C^1 smoothness condition with a tip at ξ. Then in view of Lemma 4.2, b can be computed as a linear combination of the $b_1, \dots, b_n$; i.e., there exist $\{\alpha_i\}_{i=1}^n$ such that

$$b = \sum_{i=1}^n \alpha_i b_i. \tag{5.4}$$

Since F_8, and $\tilde{F}_8$ are just scaled versions of $F_9 := F$, it follows that (5.4) also holds with b's replaced by either c's or d's. But then using (5.3), we have

$$[1,\, -s,\, -r] \begin{pmatrix} b \\ c \\ d \end{pmatrix} = [1,\, -s,\, -r] \begin{pmatrix} b_1 \cdots b_n \\ c_1 \cdots c_n \\ d_1 \cdots d_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = 0,$$

which shows that the C^1 smoothness condition with tip at ξ and involving b, c, d is also satisfied.

Now for each face of F, we set the coefficients corresponding to $\mathcal{M}_F$. These sets are clearly separated from each other and from the sets $D_4(v)$ and $E_3(e)$. If F is an interior face of $\triangle$, then there are two tetrahedra T and $\widetilde{T}$ sharing the face F, and $\mathcal{M}_F$ lies in just one of them, say T. Next we use the C^2 smoothness conditions to uniquely determine the coefficients for the corresponding points in $\widetilde{T}$. Now we can use Lemma 4.3 to compute the coefficients of s corresponding to the remaining domain points on faces F of the shells $R_7(v_T)$ for all T. We now check that these computed coefficients satisfy all C^2 smoothness conditions across F. Each domain point in F_7 (see Figure 2) is the tip of a C^2 smoothness condition. Assuming a, b, c, d, e are the coefficients on $F_7, F_8, F_9, \tilde{F}_8, \tilde{F}_7$, the typical condition has the form

$$a = s^2 c + 2rsd + r^2 e,$$

where r, s are as before. By construction, these smoothness conditions are satisfied for all points ξ marked with dots, triangles, or $\oplus$ in Figure 2. There are $n = 45$ such points. Writing $\{a_i, b_i, c_i, d_i, e_i\}_{i=1}^n$ for the associated coefficients, we have

$$a_i = s^2 c_i + 2rsd_i + r^2 e_i, \quad i = 1, \ldots, n.$$

Now if ξ is any other point in F_7, then by Lemmas 4.1–4.3, there are α_i such that

$$[1,\ -s^2,\ -2rs,\ -r^2] \begin{pmatrix} a \\ c \\ d \\ e \end{pmatrix} = [1,\ -s^2,\ -2rs,\ -r^2] \begin{pmatrix} a_1 \cdots a_n \\ c_1 \cdots c_n \\ d_1 \cdots d_n \\ e_1 \cdots e_n \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = 0,$$

which shows that the C^2 smoothness condition with tip at ξ and involving a, c, d, e is also satisfied.

To complete the proof, we now apply Lemma 3.2 to uniquely compute the coefficients of s corresponding to the remaining domain points in the balls $D_7(v_T)$ for all T. □

6. An NMDS and Hermite interpolation.

In this section we show how to construct an NMDS for the macroelement space of the previous section, and then use it to solve a certain Hermite interpolation problem. First we need some additional notation.

Given any multi-index $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, we write D^α for the partial derivative $D_x^{\alpha_1} D_y^{\alpha_2} D_z^{\alpha_3}$. For each edge $e := \langle u, v\rangle$ of a tetrahedron $T \in \triangle$, suppose X_e is the plane perpendicular to e at the point u. We endow X_e with Cartesian coordinate axes whose origin lies at the point u. Then for any multi-index $\beta = (\beta_1, \beta_2)$, we define D_e^β to be the corresponding derivative. It corresponds to a directional derivative of order $|\beta| := \beta_1 + \beta_2$ in a direction lying in X_e. Associated with e we also need notation for the following sets of equally spaced points in the interior of e:

$$\eta_{e,j}^i := \frac{(i-j+1)u + jv}{i+1}, \quad j = 1, \ldots, i, \tag{6.1}$$

for all $i > 0$.

For each face $F := \langle v_1, v_2, v_3\rangle$ of $\triangle$, let D_F be the directional derivative associated with a unit normal vector to F, and let $D_{F,i}$ be the directional derivatives associated with the vectors $\langle v_i, v_F\rangle$ for $i = 1, 2, 3$, where as before v_F is the split point in the face F.

If η is a point in $\mathbb{R}^3$, we write ε_η for the point-evaluation functional associated with η, so that for any trivariate function, $\varepsilon_\eta f := f(\eta)$.

Theorem 6.1. *The set*

$$\mathcal{N} := \bigcup_{v\in\mathcal{V}} \mathcal{N}_v \cup \bigcup_{e\in\mathcal{E}} \mathcal{N}_e \cup \bigcup_{F\in\mathcal{F}} \mathcal{N}_F \cup \bigcup_{T\in\triangle} \mathcal{N}_T \tag{6.2}$$

is an NMDS for $\mathcal{S}_2(\triangle_{WF})$, where

(1) $\mathcal{N}_v := \{\varepsilon_v D^\alpha\}_{|\alpha|\le 4}$,
(2) $\mathcal{N}_e := \bigcup_{i=1}^{3} \bigcup_{j=1}^{i} \{\varepsilon_{\eta_{e,j}^i} D_e^\beta\}_{|\beta|=i}$,
(3) $\mathcal{N}_F := \{\varepsilon_{v_i} D_F^2 D_{F,i}^4\}_{i=1}^3$,
(4) $\mathcal{N}_T := \{\varepsilon_{v_T} D^\alpha\}_{|\alpha|\le 3}$.

Proof. It is easy to see that the cardinality of the set $\mathcal{N}$ matches the dimension of $\mathcal{S}_2(\triangle_{WF})$ as given in Theorem 5.1. We already know that the set $\mathcal{M}$ defined in that theorem is an MDS for $\mathcal{S}_2(\triangle_{WF})$. Thus, to show that $\mathcal{N}$ is an NMDS, it suffices to show that if $s \in \mathcal{S}_2(\triangle_{WF})$, then setting the values $\{\lambda s\}_{\lambda\in\mathcal{N}}$ determines all coefficients in the set $\{c_\xi\}_{\xi\in\mathcal{M}}$.

For each $v \in \mathcal{V}$, we can compute the coefficients in $\mathcal{M}_v$ from the values of the derivatives $D^\alpha s(v)$ corresponding to $\mathcal{N}_v$. Then for each edge $e \in \mathcal{E}$, the coefficients in $\mathcal{M}_e$ can be computed from the derivatives of s corresponding to $\mathcal{N}_e$. We now use Lemmas 4.1 and 4.2 as in the proof of Theorem 3.1 to compute all remaining coefficients corresponding to domain points on the shells $R_9(v_T)$ and $R_8(v_T)$ of tetrahedra in $\triangle$.

Now fix $F \in \mathcal{F}$, and consider the set $\mathcal{M}_F$. It consists of the three domain points $\{\xi_{2430}^{T_{F,1}}, \xi_{2430}^{T_{F,2}}, \xi_{2430}^{T_{F,3}}\}$, where $T_{F,i}$ are three tetrahedra in $\triangle_{WF}$ lying on one side of F and sharing the face F. These domain points are marked with $\oplus$ in Figure 2 (left). To compute the coefficient corresponding to $\xi_{2430}^{T_{F,1}}$, we first solve a 3×3 system of equations with associated matrix M_3 as in (4.3) to get the coefficients corresponding to the unmarked domain points on $R_3(v_1)$ in Figure 2 (left). Then the coefficient corresponding to $\xi_{2430}^{T_{F,1}}$ can be computed from the value of the derivative $D_F^2 D_{F,i}^4 s(v_1)$. The coefficients corresponding to the other two points in $\mathcal{M}_F$ can be computed in a similar way. Now we can use Lemma 4.3 to compute the coefficients of s corresponding to the remaining domain points on shells $R_7(v_T)$. Finally, as shown in the proof of Theorem 3.1, for each tetrahedron T in $\triangle$, we can use the values $\{\lambda s\}_{\lambda\in\mathcal{N}_T}$ to compute the coefficients c_ξ of s for $\xi \in \mathcal{M}_T$. □

Theorem 6.1 shows that for any function $f \in C^6(\Omega)$, there is a unique spline $s \in \mathcal{S}_2(\triangle_{WF})$ solving the Hermite interpolation problem

$$\lambda s = \lambda f \quad \text{for all } \lambda \in \mathcal{N},$$

or, equivalently,

(1) $D^\alpha s(v) = D^\alpha f(v)$ for all $|\alpha| \le 4$ and all $v \in \mathcal{V}$;
(2) $D_e^\beta s(\eta_{e,j}^i) = D_e^\beta f(\eta_{e,j}^i)$ for all $|\beta| = i$ with $1 \le j \le i$ and $1 \le i \le 3$, and for all edges e of $\triangle$;
(3) $D_F^2 D_{F,i}^4 s(v_i) = D_F^2 D_{F,i}^4 f(v_i)$, $i = 1, 2, 3$, for each face $F := \langle v_1, v_2, v_3\rangle$ of $\triangle$;
(4) $D^\alpha s(v_T) = D^\alpha f(v_T)$ for all $|\alpha| \le 3$ and all tetrahedra $T \in \triangle$.

The nodal functionals described in (6.2) involve some derivatives of order higher than 2, even though s is only C^2 globally. However, s is in $C^4(v)$ at vertices and in $C^3(e)$ around edges, and so the third and fourth derivatives appearing in $\mathcal{N}_e$ and $\mathcal{N}_v$ are well defined. But it is not in $C^6(v)$ at a vertex v, and so if F is an interior face, then the derivatives in $\mathcal{N}_F$ are applied to just one of the polynomial pieces of s which share F.

The mapping which takes functions $f \in C^6(\Omega)$ to this Hermite interpolating spline defines a linear operator $\mathcal{I}_{WF} : C^6(\Omega) \to \mathcal{S}_2(\triangle_{WF})$. The construction guarantees that $\mathcal{I}_{WF} s = s$ for every spline $s \in \mathcal{S}_2(\triangle_{WF})$, and in particular for all trivariate polynomials of degree 9. We now discuss error bounds for this interpolation process, which in turn provides an estimate for the approximation power of the space $\mathcal{S}_2(\triangle_{WF})$.

It is well known that the key to getting error bounds for these types of spline interpolation operators is to show that the construction of the interpolating spline is both *local* and *stable*. The localness of the operator is clear from the way in which the B-coefficients of the interpolating spline s are computed. More precisely, for every domain point ξ, the corresponding coefficient c_ξ of s depends only on values of f and its derivatives at points in $\mathrm{star}(T_\xi)$, where $T_\xi \in \triangle$ is a tetrahedron containing ξ. Concerning stability, we have the following.

LEMMA 6.2. *Given a tetrahedral partition $\triangle$, let $\triangle_{WF}$ be a corresponding Worsey–Farin partition, and let $\theta_{_{WF}}$ be the smallest angle between any two edges in $\triangle_{WF}$ sharing a vertex. Then*

$$|c_\xi| \le C \sum_{i=0}^{6} |\Omega_T|^i |s|_{i,\Omega_T} \quad \textit{for all } \xi \in \mathcal{D}_{\triangle_{WF},9}, \tag{6.3}$$

where Ω_T is the union of the tetrahedra in $\mathrm{star}(T)$*, $|\Omega_T|$ is its diameter, and C is a constant depending only on $\theta_{_{WF}}$.*

Proof. To see that (6.3) holds, we review the computation of the coefficients of s as described in the proof of Theorem 6.1. For domain points in balls of the form $D_4(v)$, where v is a vertex of $\triangle$, (6.3) follows from the well-known connection between B-coefficients in such a ball and derivatives at v. Then in the next step we compute coefficients in the sets $\mathcal{M}_e$ from the derivatives corresponding to $\mathcal{N}_e$. This involves solving some systems of equations whose stability depends on $\theta_{_{WF}}$. Now Lemmas 4.1 and 4.2 are used to compute coefficients corresponding to domain points on shells $R_9(v_{_T})$ and $R_8(v_{_T})$. This involves solving linear systems with matrices M_3 and M_5 whose inverses are bounded by a constant depending on $\theta_{_{WF}}$. Next we go to the shells $R_7(v_{_T})$. After solving 3×3 systems for the coefficients on the 3-rings around the vertices of a face F of such a shell, we compute the coefficients in $\mathcal{M}_F$ from the derivatives of s associated with $\mathcal{N}_F$ (this is where the sixth derivatives come in). The bound (6.3) also holds for these coefficients. Now the coefficients corresponding to the remaining coefficients on the shells $R_7(v_{_T})$ are computed from Lemma 4.3, which involves solving systems with matrices M_4 and M_5. Next, we use Lemma 3.2 to solve for the remaining 20 coefficients of $s|_{D_7(v_{_T})}$ (written as single polynomial). The matrix M_{20} of this system depends only on the barycentric coordinates (a_1, a_2, a_3, a_4) of $v_{_T}$, which are all bounded away from zero by a constant depending on $\theta_{_{WF}}$. This insures that the inverse of M_{20} is also bounded by a constant depending on $\theta_{_{WF}}$. These coefficients are then converted to the final coefficients of s on $D_7(v_{_T})$ by subdivision about the point $v_{_T}$, which is known to be stable. □

Given a tetrahedral partition $\triangle$, we write $|\triangle|$ for the diameter of the largest tetrahedron in $\triangle$.

THEOREM 6.3. *There exists a constant K depending only on $\theta_{_{WF}}$ such that for every $f \in C^{m+1}(\Omega)$ with $5 \le m \le 9$,*

$$\|D^\alpha (f - \mathcal{I}_{_{WF}} f)\|_\Omega \le K |\triangle|^{m+1-|\alpha|} |f|_{m+1,\Omega} \tag{6.4}$$

for all $|\alpha| \le m$.

Proof. Since the proof is similar to the proof of Theorem 3.3 in [5] and Theorem 6.2 in [20] (see also [17, 18] for similar arguments in the bivariate case), we can be brief. Fix $T \in \triangle$, and let $f \in C^{m+1}(\Omega)$. By Lemma 4.3.8 of [8], there exists a polynomial $q := q_{f,T} \in \mathcal{P}_9^3$ such that

$$\|D^\beta (f - q)\|_{\Omega_T} \le |(f-q)|_{|\beta|,\Omega_T} \le K_1 |\Omega_T|^{m+1-|\beta|} |f|_{m+1,\Omega_T} \tag{6.5}$$

for all $|\beta| \le m$, where Ω_T is the union of the tetrahedra in star(T). Now fix α with $|\alpha| \le m$. Then since $\mathcal{I}_{WF} q = q$,

$$\|D^\alpha(f - \mathcal{I}_{WF} f)\|_T \le \|D^\alpha(f-q)\|_T + \|D^\alpha \mathcal{I}_{WF}(f-q)\|_T.$$

It suffices to estimate the second quantity. Applying the Markov inequality [22] to each of the polynomials $\mathcal{I}_{WF}(f-q)|_{T_j}$, where $T_1, \ldots, T_{12}$ are the tetrahedra in the Worsey–Farin split of T, we have

$$\|D^\alpha \mathcal{I}_{WF}(f-q)\|_{T_j} \le K_2 |\triangle|^{-|\alpha|} \|\mathcal{I}_{WF}(f-q)\|_{T_j},$$

where K_2 is a constant depending only on θ_{WF}. Let c_ξ be the B-coefficients of the polynomial $\mathcal{I}_{WF}(f-q)|_{T_j}$ relative to the tetrahedron T_j. Then combining (6.3) with the fact that the Bernstein basis polynomials form a partition of unity, it is easy to see that

$$\|\mathcal{I}_{WF}(f-q)\|_{T_j} \le K_3 \max_{\xi \in \mathcal{D}_{T_j,d}} |c_\xi| \le K_4 \sum_{i=0}^{6} |\Omega_T|^i |f-q|_{i,\Omega_T}.$$

Taking the maximum over j and combining this with (6.5) gives

$$\|\mathcal{I}_{WF}(f-q)\|_T \le K_5 |\triangle|^{m+1} |f|_{m+1,\Omega_T},$$

which gives

$$\|D^\alpha(f - \mathcal{I}_{WF} f)\|_T \le K_6 |\triangle|^{m+1-|\alpha|} |f|_{m+1,\Omega_T}.$$

Finally, we take the maximum over all tetrahedra T in $\triangle$ to get (6.4). □

7. Remarks.

Remark 1. In the bivariate setting, C^r macroelements on various splits have been studied by several authors; see, e.g., [3, 4, 13, 14] and the references therein.

Remark 2. A C^2 trivariate polynomial macroelement defined on nonsplit tetrahedra was constructed in [16] using polynomials of degree 17. For C^r trivariate polynomial macroelements using polynomials of degree $8r+1$, see [15].

Remark 3. C^1 trivariate macroelements were constructed on the Worsey–Farin split using splines of degree 3 in [23]. Stability issues and the approximation power were not addressed. For other C^1 trivariate macroelements, see [1, 24].

Remark 4. By examining slices through T_{WF}, it can be shown that it is not possible to construct C^2 macroelements on the Worsey–Farin split using splines with smoothness less than 3 around the edges or smoothness 4 at the vertices. This in turn implies that the minimal degree possible is 9.

Remark 5. In section 5 we have shown that our local construction of a macroelement on a single tetrahedron given in Theorem 3.1 leads to a C^2 macroelement space for general tetrahedral partitions, provided, for each interior face F, we choose the split point v_F on F to lie on the line connecting the interior split points v_T and $v_{\tilde{T}}$ of the two tetrahedra T and $\widetilde{T}$ which share the face F. This geometry causes the smoothness conditions across F to be essentially univariate in nature. Tests using the Java program have shown that without this condition, we do not get C^2 continuity.

Remark 6. The Java code of the first author for examining piecewise polynomial functions on tetrahedral partitions was a key tool in developing the macroelements described in this paper. The code uses residual arithmetic to compute the dimension of trivariate spline spaces, find MDSs, and solve the smoothness equations. It can be downloaded from http://www.math.utah.edu/∼pa/3DMDS, along with associated documentation.

Remark 7. We have also used the Java code to explore the possibility of imposing additional smoothness conditions on our superspline space $\mathcal{S}_2(T_{WF})$ to get a space of dimension 272 which is uniquely determined by the domain points of Theorem 3.1, minus the set $\mathcal{M}_T$. This would give us a C^2 macroelement which is defined by *natural degrees of freedom* only, i.e., information on the boundary of the tetrahedron T that is necessary to insure the global smoothness and local construction. However, we have not been able to find a symmetric way to do this, and expect that if it can be done at all, it would require imposing various individual smoothness conditions of the form (4.1). A similar approach was successful in the bivariate case; see [3, 4], where we used it to get natural degrees of freedom for bivariate macroelement spaces.

Remark 8. We can remove the special smoothness conditions involving ν and μ in the definition (3.2) of the space $\mathcal{S}_2(T_{WF})$ to get an alternative macroelement space which has 9 degrees of freedom per face rather than 3, and thus has a total of 316 degrees of freedom rather than 292. The proof that this alternative element is C^2 proceeds along the same lines as the proof of Theorem 5.1, and the global space has dimension $35V + 20E + 9F + 20N_T$. The corresponding nodal basis (and associated Hermite interpolation operator) requires derivatives up to order 4 only, rather than the order 6 required for the element described here.

Remark 9. It is possible to create macroelements with fewer degrees of freedom by the process of *condensation*. This amounts to further restricting the spline space by forcing cross-derivatives along edges or through faces of the tetrahedron T to be of reduced degree. The main problem with this strategy is that it produces elements which no longer have the capability of reproducing the full polynomial space, and thus have reduced approximation power.

Remark 10. In this paper we have given error bounds for Hermite interpolation with our macroelement in the uniform norm. Analogous results hold for the p-norms and can be proved using appropriate quasi-interpolation operators; see section 10 of [12] for the bivariate case.

Remark 11. Using the Java code mentioned in Remark 6, one can easily check that there is a similar C^3 macroelement on the Worsey–Farin split of a tetrahedron which uses splines of degree 13 which are C^6 around the vertices, C^5 around the edges, C^9 at the centroid v_T, and C^9 along edges connecting v_T to points v_F. This space has dimension 984, with 916 natural degrees of freedom; see Remark 7.

Remark 12. We have recently learned [10] that Ming-Jun Lai and Alain Le Méhauté have independently studied C^r macroelements based on the Worsey–Farin split.

Remark 13. Using the Java software, we have also designed C^2 macroelements based on a trivariate analog of the double Clough–Tocher split of a tetrahedron which is obtained by first applying the Clough–Tocher split, and then applying it again to each of the resulting four subtetrahedra. We report on this element in [6].

Remark 14. It has recently been shown (see [15]) that if incenters are used to construct the bivariate Powell–Sabin element, then the stability of the element depends only on the smallest angle in the original triangulation before applying the Powell–Sabin splits. We conjecture that the analogous statement holds here—namely, that the stability of our element depends only on the smallest angle in the original tetrahedral partition $\triangle$ rather than on the smallest angle θ_{WF} in $\triangle_{WF}$. This is an important distinction, since even though we are using incenters, theoretically the angles in the Clough–Tocher splits of the faces could be arbitrarily small. We are still working on this conjecture.

Acknowledgment. We would like to thank Ming-Jun Lai for useful discussions.

REFERENCES

[1] P. Alfeld, *A trivariate Clough-Tocher scheme for tetrahedral data*, Comput. Aided Geom. Design, 1 (1984), pp. 169–181.

[2] P. Alfeld, *Bivariate splines and minimal determining sets*, J. Comput. Appl. Math., 119 (2000), pp. 13–27.

[3] P. Alfeld and L. L. Schumaker, *Smooth macro-elements based on Powell-Sabin triangle splits*, Adv. Comput. Math., 16 (2002), pp. 29–46.

[4] P. Alfeld and L. L. Schumaker, *Smooth macro-elements based on Clough-Tocher triangle splits*, Numer. Math., 90 (2002), pp. 597–616.

[5] P. Alfeld and L. L. Schumaker, *A C^2 trivariate macro-element based on the Clough-Tocher split of a tetrahedron*, Comput. Aided Geom. Design, 22 (2005), pp. 710–721.

[6] P. Alfeld and L. L. Schumaker, *A C^2 trivariate double-Clough-Tocher macro-element*, in Approximation Theory XI: Gatlinburg 2004, C. Chui, M. Neamtu, and L. L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2005, pp. 1–14.

[7] P. Alfeld, L. L. Schumaker, and W. Whiteley, *The generic dimension of the space of C^1 splines of degree $d \geq 8$ on tetrahedral decompositions*, SIAM J. Numer. Anal., 30 (1993), pp. 889–920.

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

[9] A. Ibrahim and L. L. Schumaker, *Super spline spaces of smoothness r and degree $d \geq 3r+2$*, Constr. Approx., 7 (1991), pp. 401–423.

[10] M. J. Lai, *private communication*.

[11] M.-J. Lai and A. LeMéhauté, *A new kind of trivariate C^1 spline*, Adv. Comput. Math., 21 (2004), pp. 273–292.

[12] M.-J. Lai and L. L. Schumaker, *On the approximation power of bivariate splines*, Adv. Comput. Math., 9 (1998), pp. 251–279.

[13] M. J. Lai and L. L. Schumaker, *Macro-elements and stable local bases for splines on Clough-Tocher triangulations*, Numer. Math., 88 (2001), pp. 105–119.

[14] M. J. Lai and L. L. Schumaker, *Macro-elements and stable local bases for splines on Powell-Sabin triangulations*, Math. Comp., 72 (2003), pp. 335–354.

[15] M. J. Lai and L. L. Schumaker, *Splines on Triangulations*, monograph in preparation.

[16] A. Le Méhauté, *Interpolation et approximation par des fonctions polynomiales par morceaux dans $\mathbb{R}^n$*, dissertation, Université de Rennes, France, 1984.

[17] G. Nürnberger, V. Rayevskaya, L. L. Schumaker, and F. Zeilfelder, *Local Lagrange interpolation with C^2 splines of degree seven on triangulations*, in Advances in Constructive Approximation, M. Neamtu and E. Saff, eds., Nashboro Press, Brentwood, TN, 2004, pp. 345–370.

[18] G. Nürnberger, V. Rayevskaya, L. L. Schumaker, and F. Zeilfelder, *Local Lagrange interpolation with bivariate splines of arbitrary smoothness*, Constr. Approx., June, 2005.

[19] L. L. Schumaker, *Dual bases for spline spaces on cells*, Comput. Aided Geom. Design, 5 (1988), pp. 277–284.

[20] L. L. Schumaker and T. Sorokina, *C^1 quintic splines on type-4 tetrahedral partitions*, Adv. Comput. Math., 21 (2004), pp. 421–444.

[21] L. L. Schumaker and T. Sorokina, *A trivariate box macro-element*, Constr. Approx., 21 (2005), pp. 413–431.

[22] D. R. Wilhelmsen, *A Markov inequality in several dimensions*, J. Approx. Theory, 11 (1974), pp. 216–220.

[23] A. J. Worsey and G. Farin, *An n-dimensional Clough-Tocher interpolant*, Constr. Approx., 3 (1987), pp. 99–110.

[24] A. J. Worsey and B. Piper, *A trivariate Powell-Sabin interpolant*, Comput. Aided Geom. Design, 5 (1988), pp. 177–186.

[25] A. Ženišek, *Polynomial approximation on tetrahedrons in the finite element method*, J. Approx. Theory, 7 (1973), pp. 334–351.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1766–1782

ROBUST A POSTERIORI ERROR ESTIMATES FOR STATIONARY CONVECTION-DIFFUSION EQUATIONS[*]

R. VERFÜRTH[†]

Abstract. We analyze a posteriori error estimators for finite element discretizations of convection-dominated stationary convection-diffusion equations using locally refined, isotropic meshes. The estimators are based on either the evaluation of local residuals or the solution of discrete local problems with Dirichlet or Neumann boundary conditions. All estimators yield global upper and lower bounds for the error measured in a norm that incorporates the standard energy norm and a dual norm of the convective derivative. They are fully robust in the sense that the ratio of the upper and lower bounds is uniformly bounded with respect to the size of the convection. The estimates are also uniform with respect to the size of the zero-order reaction term and also hold for the limit case of vanishing reaction.

Key words. a posteriori error estimates, convection-diffusion equations

AMS subject classifications. 65N30, 65N15, 65J15

DOI. 10.1137/040604261

1. Introduction.

We consider stationary convection-diffusion equations

$$\begin{aligned} -\varepsilon\Delta u + \underline{a}\cdot\nabla u + bu &= f \quad \text{in } \Omega,\\ u &= 0 \quad \text{on } \Gamma_D,\\ \varepsilon\frac{\partial u}{\partial n} &= g \quad \text{on } \Gamma_N \end{aligned} \tag{1.1}$$

in a polygonal domain Ω in $\mathbb{R}^n$, $n \geq 2$, with Lipschitz boundary Γ consisting of two disjoint components Γ_D and Γ_N. The data have to satisfy the following conditions:

(A1) $0 < \varepsilon \ll 1$.

(A2) $\underline{a} \in W^{1,\infty}(\Omega)^n$, $b \in L^\infty(\Omega)$.

(A3) There are two constants $\beta \geq 0$ and $c_b \geq 0$, which do not depend on ε, such that $-\frac{1}{2}\operatorname{div}\underline{a} + b \geq \beta$ and $\|b\|_{L^\infty} \leq c_b\beta$.

(A4) The Dirichlet boundary Γ_D has positive $(n-1)$-dimensional Lebesgue measure and includes the inflow boundary $\{x \in \Gamma : \underline{a}(x)\cdot\underline{n}(x) < 0\}$.

Assumption (A3) allows us to handle simultaneously the case of a nonvanishing zero-order reaction term and that of absent reaction, the latter corresponding to $\beta = 0$. In the case $\beta = 0$ we set $c_b = 0$. Assumption (A1) of course means that we are interested in the convection-dominated regime.

We analyze three a posteriori error estimators for finite element discretizations (standard Galerkin or SUPG) of problem (1.1). One estimator is based on the evaluation of local residuals, and the other two are based on the solution of auxiliary local discrete convection-diffusion problems with Dirichlet or Neumann boundary conditions. All estimators yield global upper and lower bounds on the error of the finite element discretization measured in a norm that incorporates the standard energy norm of problem (1.1) and a dual norm of the convective derivative (cf. section 2 for

[*] Received by the editors February 18, 2004; accepted for publication (in revised form) April 22, 2005; published electronically November 22, 2005.
http://www.siam.org/journals/sinum/43-4/60426.html

[†] Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany (rv@num1.ruhr-uni-bochum.de).

the definition of the norms). All estimates are fully robust in the sense that the ratio of upper and lower bounds is uniformly bounded with respect to the mesh-size, to the viscosity ε, and to the parameter β.

Our analysis is restricted to shape-regular meshes (cf. section 2). This includes locally refined meshes but excludes anisotropic elements with large aspect ratios. One could try to extend our analysis to anisotropic meshes following the lines of [6]. One then has to establish Lemmas 3.3 and 3.6 for anisotropic elements too. This is partially done in [7]. It is to be expected that the left-hand side of the upper error bound (4.7) in Theorem 4.1 will then include a factor which measures the alignment of the mesh with the error (cf. [7, Theorem 7.1]).

The present results complement and improve the results of [10] in several respects. There the analysis was restricted to the case $\beta = 1$; here we also treat the limiting case of vanishing reaction. Up to obvious modifications due to the presence of the parameter β, the residual estimator and the auxiliary local Dirichlet problems are the same here and in [10]. But here the solution of the local problem is evaluated with respect to a norm that also incorporates a mesh-dependent norm of the convective derivative. Most important, the present estimates are fully robust, whereas in [10] the ratios of the upper and lower bounds depend on the mesh-Péclet number. The present auxiliary local Neumann problems differ from their analogue in [10] by a new approximation of the convection and reaction terms. Moreover, the estimator considered here is based on a mesh-dependent norm that takes into account the convective derivative.

A comparison of the present results and of those in [10] leads to the following conclusions:

- A large ratio of the error estimators to the energy norm of the error may be attributed to a large convective derivative of the error and thus to insufficiently resolved interior and boundary layers.
- The full equivalence of the error estimators with parameter-independent constants is due to the incorporation of the convection into the norms used for evaluating the solutions of the auxiliary problems.

Our results should also be compared with those of [8]. There the residual-free bubbles method is applied to problem (1.1) with $b = 0$ and $\operatorname{div} \underline{a} = 0$, i.e., $\beta = 0$. The meshes are assumed to be shape-regular. The error estimator is a residual one and is proved to be robust. Our results for the case $\beta = 0$ and those of [8] differ in the scaling of the norm that measures the error and of the weights used in the error estimator. Such a rescaling, however, is not possible in the general case $\beta > 0$.

We have implemented the residual error estimator of section 4 for some examples. The Java applet and a user guide are available at www.ruhr-uni-bochum.de/num1. The estimators of sections 5 and 6 are much more complex both with respect to implementational and computational work. Up to now we did not have the time to implement them.

The article is organized as follows. In section 2 we present the variational formulation of problem (1.1), its finite element discretization, and the relevant norms. In section 3 we collect some auxiliary results which are needed for deriving the error bounds. In sections 4–6 we introduce the error estimators and prove their robustness. In what follows all constants are independent of the mesh-size, of the viscosity ε, and of the parameter β.

2. Variational formulation and finite element discretization. For any bounded open subset ω of Ω with Lipschitz boundary γ, we denote by $H^k(\omega)$, $k \in \mathbb{N}$,

$L^2(\omega) = H^0(\omega)$, and $L^2(\gamma)$ the usual Sobolev and Lebesgue spaces equipped with the standard norms $\|.\|_{k;\omega} = \|.\|_{H^k(\omega)}$ and $\|.\|_{0;\gamma} = \|.\|_{L^2(\gamma)}$ (cf. [1]). Similarly, $(.,.)_\omega$ and $(.,.)_\gamma$ denote the L^2-scalar products on ω and γ, respectively. If $\omega = \Omega$, we will omit the index Ω.

Set

$$H^1_D(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\} \tag{2.1}$$

and define the bilinear form B on $H^1(\Omega) \times H^1(\Omega)$ by

$$B(u, v) = \varepsilon(\nabla u, \nabla v) + (\underline{a} \cdot \nabla u, v) + (bu, v). \tag{2.2}$$

Then the standard variational formulation of problem (1.1) is to find $u \in H^1_D(\Omega)$ such that

$$B(u, v) = (f, v) + (g, v)_{\Gamma_N} \quad \forall v \in H^1_D(\Omega). \tag{2.3}$$

Due to assumption (A3), the natural energy norm for problem (2.3) is given by

$$|||v||| = \left\{\varepsilon\|\nabla u\|_0^2 + \beta\|u\|_0^2\right\}^{1/2}. \tag{2.4}$$

The dual space of $H^1_D(\Omega)$ is denoted by $H^1_D(\Omega)^*$ and is equipped with the dual norm

$$|||\varphi|||_* = \sup_{v \in H^1_D(\Omega)\backslash\{0\}} \frac{\langle \varphi, v\rangle}{|||v|||}, \tag{2.5}$$

where $\langle .,.\rangle$ denotes the corresponding duality pairing. This norm will be used for bounding the convective derivative.

For the finite element discretization we consider a family $\mathcal{T}_h$, $h > 0$, of partitions of Ω into simplices or parallelepipeds that satisfies the following properties.

Affine equivalence: Each element can be mapped by an invertible affine mapping onto the unit n-simplex or the unit n-cube.

Admissibility: Any two elements are either disjoint or share a complete ℓ-dimensional face $(0 \le \ell \le n-1)$.

Shape regularity: The ratio of the diameter h_K of any element K to the diameter ρ_K of the largest ball inscribed into this element is uniformly bounded.

We fix a natural number $k \ge 1$ and denote for any element K by $R_k(K)$ the space of all polynomials of total degree at most k if K is a simplex or of maximal degree at most k if K is a parallelepiped. With this definition the finite element space is given by

$$X_h = \left\{v \in C(\Omega) : v_{|K} \in R_k(K)\ \forall K \in \mathcal{T}_h, v = 0 \text{ on } \Gamma_D\right\}. \tag{2.6}$$

Next we define a bilinear form B_δ on $X_h \times X_h$ and a linear form ℓ_δ on X_h by

$$B_\delta(u_h, v_h) = B(u_h, v_h) + \sum_{K \in \mathcal{T}_h} \delta_K(-\varepsilon\Delta u_h + \underline{a} \cdot \nabla u_h + bu_h, \underline{a} \cdot \nabla v_h)_K \tag{2.7}$$

and

$$\ell_\delta(v_h) = (f, v_h) + (g, v_h)_{\Gamma_N} + \sum_{K \in \mathcal{T}_h} \delta_K(f, \underline{a} \cdot \nabla v_h)_K. \tag{2.8}$$

The δ_K are nonnegative stabilization parameters. We will always assume that

$$\delta_K \|\underline{a}\|_{L^\infty(K)} \leq c h_K \quad \forall K \in \mathcal{T}_h. \tag{2.9}$$

With these definitions, the finite element discretization of problem (1.1) consists in finding $u_h \in X_h$ such that

$$B_\delta(u_h, v_h) = \ell_\delta(v_h) \quad \forall v_h \in X_h. \tag{2.10}$$

The choice $\delta_K = 0$ for all K yields the standard Galerkin discretization; the choice $\delta_K > 0$ for all K corresponds to the SUPG-discretizations (cf., e.g., [4], [5]). Condition (2.9) is satisfied for all choices of δ_K used in practice. Assumptions (A3), (A4), and (2.9) and standard arguments for SUPG-discretizations imply that problem (2.10) admits a unique solution.

3. Auxiliary results. In this section we collect some auxiliary results and notation that will be helpful for the estimates of the subsequent sections. We start with a stability result for the bilinear form (2.2).

LEMMA 3.1. *The bilinear form* (2.2) *satisfies the upper bound*

$$B(u, v) \leq \max\{c_b, 1\} \{|||u||| + |||\underline{a} \cdot \nabla u|||_*\} |||v||| \quad \forall u, v \in H^1_D(\Omega) \tag{3.1}$$

and the inf-sup *condition*

$$\inf_{u \in H^1_D(\Omega)\backslash\{0\}} \sup_{v \in H^1_D(\Omega)\backslash\{0\}} \frac{B(u, v)}{\{|||u||| + |||\underline{a} \cdot \nabla u|||_*\} |||v|||} \geq \frac{1}{2 + \max\{c_b, 1\}}. \tag{3.2}$$

The constant c_b is the one of assumption (A3).

Proof. The upper bound (3.1) follows from assumption (A3), the definition (2.4) of the energy norm $|||.|||$, and the definition (2.5) of the dual norm $|||.|||_*$.

To prove the inf-sup condition (3.2), we fix an arbitrary function $u \in H^1_D(\Omega)$ and choose a real number θ greater than 0 and less than 1. Due to the definition (2.5) of the dual norm there is a function $v_\theta \in H^1_D(\Omega)$ with

$$|||v_\theta||| = 1 \quad \text{and} \quad (\underline{a} \cdot \nabla u, v_\theta) \geq \theta |||\underline{a} \cdot \nabla u|||_*.$$

Set $w_\theta = u + \frac{1}{1+\max\{c_b,1\}} |||u||| v_\theta$. The bilinearity of B then yields

$$B(u, w_\theta) = B(u, u) + \frac{1}{1 + \max\{c_b, 1\}} |||u||| B(u, v_\theta).$$

Integration by parts and assumptions (A3) and (A4) imply the coercivity of B, i.e.,

$$B(u, u) \geq |||u|||^2.$$

The definition of v_θ and assumption (A3) on the other hand give

$$B(u, v_\theta) = (\underline{a} \cdot \nabla u, v_\theta) + \varepsilon(\nabla u, \nabla v_\theta) + (bu, v_\theta) \geq \theta |||\underline{a} \cdot \nabla u|||_* - \max\{c_b, 1\} |||u|||.$$

Since

$$|||w_\theta||| \le \frac{2+\max\{c_b,1\}}{1+\max\{c_b,1\}}|||u|||$$

these estimates yield

$$\begin{aligned}\sup_{v\in H^1_D(\Omega)\backslash\{0\}} \frac{B(u,v)}{|||v|||} &\ge \frac{B(u,w_\theta)}{|||w_\theta|||}\\ &\ge \frac{1}{2+\max\{c_b,1\}}\{|||u|||+\theta|||\underline{a}\cdot\nabla u|||_*\}.\end{aligned}$$

Since $0<\theta<1$ and $u\in H^1_D(\Omega)$ were arbitrary, this proves the inf-sup condition (3.2). □

Remark 3.2. A similar result is established in [9]. There, however, the solution u and the test-function v are measured with respect to the same norm which is an interpolation norm between $|||.|||$ and $|||.|||+|||\underline{a}\cdot\nabla.|||_*$. The present result is better suited for our purposes, since, in what follows, we thus have to estimate Clément-type interpolants and bubble functions with respect to the energy norm and can do this by invoking standard results from the literature. It is worth noting that by the same argument a similar stability result can be established for the bilinear form B restricted to $X_h\times X_h$. One only has to replace $|||.|||_*$ by its discrete analogue for which the supremum is taken with respect to X_h instead of $H^1_D(\Omega)$. In section 5 we will use a similar result for analyzing the local auxiliary problems. Lemma 3.1 should also be compared with the results in [2]. There the norm $\sup_v B(u,v)/\|\nabla v\|_0$ is used to derive suboptimal, i.e., ε-dependent, a posteriori error bounds.

Next we introduce some notation that will be needed for the error estimates. We denote by $\mathcal{N}_h$ the set of all element vertices that do not lie on the Dirichlet boundary Γ_D and by $\mathcal{E}_h$ the set of all $(n-1)$-dimensional element faces that are not contained in Γ_D.

With each $E\in\mathcal{E}_h$ we associate a unit vector $\underline{n}_E$ that is orthogonal to E and that points to the outside of Ω if E is part of the boundary Γ. For any interior face E in Ω we denote by $[.]_E$ the jump across E in direction $\underline{n}_E$. The jump $[.]_E$ of course depends on the orientation of $\underline{n}_E$. But quantities of the form $[\underline{n}_E\cdot .]_E$ are independent thereof.

With every element K we associate two sets ω_K and $\tilde{\omega}_K$ which consist of the union of all elements that share an $(n-1)$-dimensional face with K and of the union of all elements that share at least one point with K, respectively. For a face $E\in\mathcal{E}_h$ the sets ω_E and $\tilde{\omega}_E$ are defined analogously.

For every vertex $x\in\mathcal{N}_h$ we denote by λ_x the nodal bases function which is uniquely defined by the properties

$$\lambda_{x|K}\in R_1(K)\quad \forall K\in\mathcal{T}_h,\quad \lambda_x(y)=0\quad \forall y\in\mathcal{N}_h\backslash\{x\},\quad \lambda_x(x)=1.$$

The support of a nodal bases function λ_x is denoted by ω_x and consists of all elements that share the vertex x. With this notation we can define a Clément-type interpolation operator $I_h : L^1(\Omega)\longrightarrow\{\varphi\in C(\Omega):\varphi_{|K}\in R_1(K)\text{ for all }K\in\mathcal{T}_h,\varphi=0\text{ on }\Gamma_D\}$ by (cf. [11])

$$I_h v=\sum_{x\in\mathcal{N}_h}\left\{\frac{1}{|\omega_x|}\int_{\omega_x}v\right\}\lambda_x. \tag{3.3}$$

Here $|\omega_x|$ denotes the n-dimensional Lebesgue measure of ω_x.

Lemma 3.3. *For every $S \in \mathcal{T}_h \cup \mathcal{E}_h$ denote by h_S its diameter and set*

$$\alpha_S = \min\{h_S \varepsilon^{-1/2}, \beta^{-1/2}\}. \tag{3.4}$$

Then the following estimates hold for all elements K, all faces E, and all functions $v \in H^1_D(\Omega)$:

$$\begin{aligned} \|v - I_h v\|_{0;K} &\le c_1 \alpha_K |||v|||_{\tilde{\omega}_K}, \\ \|v - I_h v\|_{0;E} &\le c_2 \varepsilon^{-1/4} \alpha_E^{1/2} |||v|||_{\tilde{\omega}_E}, \\ |||I_h v|||_K &\le c_3 |||v|||_{\tilde{\omega}_K}. \end{aligned}$$

Here $|||.|||_A$ denotes the restriction of $|||.|||$ to the measurable set A.

Proof. The proof of Lemma 3.3 follows from Lemma 3.1 in [10] and Proposition 2.1 in [11] with the arguments used in the proof of Lemma 3.2 in [10]. □

Remark 3.4. In the case $\beta = 0$ the minimum in (3.4) of course yields $\alpha_S = \varepsilon^{-1/2} h_S$ for all S.

Remark 3.5. Proposition 2.1 of [11] is only proved for simplicial elements. A close inspection of the arguments, however, reveals that they immediately carry over to parallelepipeds that are the affine image of the unit cube. The crucial point here is that the Jacobian of the transformation is constant.

Next we define element and face bubble functions that will be used in deriving lower error bounds. For every element K we denote by $\mathcal{N}_K$ the set of its vertices and set

$$\psi_K = \gamma_K \prod_{x \in \mathcal{N}_K} \lambda_x, \tag{3.5}$$

where the constant γ_K is chosen such that ψ_K equals 1 at the barycenter of K. Note that the support of ψ_K is contained in K and that $\|\psi_K\|_{L^\infty(K)} = 1$.

For every face E we set

$$\theta_E = \min\{\varepsilon^{1/2} \beta^{-1/2} h_E^{-1}, 1\} \tag{3.6}$$

and denote by $\mathcal{N}_E$ the set of its vertices. (Note that $\theta_E = 1$ in the case $\beta = 0$.) Consider first a face E that is not contained in the boundary. It is shared by exactly two elements $K_{E,1}$ and $K_{E,2}$. For $i = 1, 2$ we define an affine transformation $F_i : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ as follows: We first map $K_{E,i}$ onto the reference element such that the image of E is contained in the hyperplane $\{x_n = 0\}$; then we apply the transformation $(x_1, \dots, x_{n-1}, x_n) \longrightarrow (x_1, \dots, x_{n-1}, \theta_E x_n)$; and finally we transform back using the inverse of the affine transformation of the first step. With this definition we set

$$\psi_E = \gamma_E \prod_{x \in \mathcal{N}_E} \lambda_x \circ F_i^{-1} \quad \text{on } K_{E,i}\, , i = 1, 2, \tag{3.7}$$

where the constant γ_E is chosen such that ψ_E equals 1 at the barycenter of E. Note that the support of ψ_E is contained in $F_1(K_{E,1}) \cup F_2(K_{E,2}) \subset K_{E,1} \cup K_{E,2} = \omega_E$ and that $\|\psi_E\|_{L^\infty(E)} = 1$.

If a face E is contained in the Neumann boundary Γ_N, the definition of ψ_E is modified in the obvious way taking into account that now E is the face of exactly one element K_E.

LEMMA 3.6. *The following estimates hold for all elements K, all polynomials $v \in R_k(K)$, all faces E, and all polynomials $\sigma \in R_k(E)$:*

$$\begin{aligned}
(v, \psi_K v)_K &\geq c_4 \|v\|_{0;K}^2, \\
|||\psi_K v|||_K &\leq c_5 \alpha_K^{-1} \|v\|_{0;K}, \\
(\sigma, \psi_E \sigma)_E &\geq c_6 \|\sigma\|_{0;E}^2, \\
|||\psi_E \sigma|||_{\omega_E} &\leq c_7 \varepsilon^{1/4} \alpha_E^{-1/2} \|\sigma\|_{0;E}, \\
\|\psi_E \sigma\|_{0;\omega_E} &\leq c_8 \varepsilon^{1/4} \alpha_E^{1/2} \|\sigma\|_{0;E}.
\end{aligned}$$

Here, a polynomial σ defined on a face E is continued in the canonical way to a polynomial defined on $\mathbb{R}^n$. The constants $c_4, \ldots, c_8$ depend only on the polynomial degree k and on the ratios h_K/ρ_K.

Proof. The estimates are proven with the same arguments as in the proof of Lemma 3.3 in [10]. For parallelepipeds one only has to take into account that the transformation to the unit cube is affine and thus has a constant Jacobian. □

4. A residual error estimator.

Denote by f_h, g_h, $\underline{a}_h$, and b_h the L^2-projections of the data f, g, $\underline{a}$, and b onto the space of piecewise constant functions corresponding to $\mathcal{T}_h$. For abbreviation we define element residuals R_K by

$$R_K = f_h + \varepsilon \Delta u_h - \underline{a}_h \cdot \nabla u_h - b_h u_h, \tag{4.1}$$

face residuals R_E by

$$R_E = \begin{cases} -[\varepsilon \underline{n}_E \cdot \nabla u_h]_E & \text{if } E \not\subset \Gamma, \\ g_h - \varepsilon \underline{n}_E \cdot \nabla u_h & \text{if } E \subset \Gamma_N, \\ 0 & \text{if } E \subset \Gamma_D, \end{cases} \tag{4.2}$$

elementwise data errors D_K by

$$D_K = \{f - f_h + (\underline{a}_h - \underline{a}) \cdot \nabla u_h + (b_h - b) u_h\}_{|K}, \tag{4.3}$$

and edgewise data errors D_E, $E \in \mathcal{E}_h \cap \Gamma_N$, by

$$D_E = g - g_h. \tag{4.4}$$

Here, of course, u_h denotes the solution of the discrete problem (2.10). Note that, as usual, R_K is defined elementwise. In particular, the term Δu_h has to be interpreted as the Laplacian applied to the restriction of u_h to the element K.

THEOREM 4.1. *For every element K define the error indicator η_K by*

$$\eta_K = \left\{ \alpha_K^2 \|R_K\|_{0;K}^2 + \sum_{E \in \mathcal{E}_h; E \subset \partial K} \varepsilon^{-1/2} \alpha_E \|R_E\|_{0;E}^2 \right\}^{1/2} \tag{4.5}$$

and the data error indicator Θ_K *by*

$$\Theta_K = \left\{ \alpha_K^2 \left[\|f - f_h\|_{0;K}^2 + \|(\underline{a} - \underline{a}_h) \cdot u_h\|_{0;K}^2 + \|(b - b_h)u_h\|_{0;K}^2 \right] \right. \\ \left. + \sum_{E \in \mathcal{E}_h; E \subset \partial K \cap \Gamma_N} \varepsilon^{-1/2} \alpha_E \|g - g_h\|_{0;E}^2 \right\}^{1/2}. \tag{4.6}$$

Then the error between the solutions u *and* u_h *of problems* (2.3) *and* (2.10) *is bounded from above by*

$$|||u - u_h||| + |||\underline{a} \cdot \nabla(u - u_h)|||_* \le c^* \left\{ \sum_{K \in \mathcal{T}_h} \left[\eta_K^2 + \Theta_K^2 \right] \right\}^{1/2} \tag{4.7}$$

and from below by

$$\left\{ \sum_{K \in \mathcal{T}_h} \eta_K^2 \right\}^{1/2} \le c_* \left[|||u - u_h||| + |||\underline{a} \cdot \nabla(u - u_h)|||_* + \left\{ \sum_{K \in \mathcal{T}_h} \Theta_K^2 \right\}^{1/2} \right]. \tag{4.8}$$

The constant c^* *depends only on the constants* $c_1, \dots, c_3$ *of Lemma* 3.3 *and on the ratios* h_K/ρ_K*; the constant* c_* *depends only on the constants* $c_4, \dots, c_8$ *of Lemma* 3.6 *and on the ratios* h_K/ρ_K*.*

Proof. As usual, we define the residual $R(u_h)$ of the discrete solution u_h by

$$\langle R(u_h), v \rangle = (f, v) + (g, v)_{\Gamma_N} - B(u_h, v) \quad \forall v \in H_D^1(\Omega). \tag{4.9}$$

Since $\langle R(u_h), v \rangle = B(u - u_h, v)$ for all $v \in H_D^1(\Omega)$, Lemma 3.1 implies

$$\frac{1}{1 + \max\{c_b, 1\}} |||R(u_h)|||_* \le |||u - u_h||| + |||\underline{a} \cdot \nabla(u - u_h)|||_* \\ \le \{2 + \max\{c_b, 1\}\} \, |||R(u_h)|||_*. \tag{4.10}$$

Integration by parts elementwise yields the following L^2-representation of the residual:

$$\begin{aligned} \langle R(u_h), v \rangle &= \sum_{K \in \mathcal{T}_h} (f + \varepsilon \Delta u_h - \underline{a} \cdot \nabla u_h - b u_h, v)_K \\ &\quad - \sum_{E \in \mathcal{E}_h \cap \Omega} ([\varepsilon \underline{n}_E \cdot \nabla u_h]_E, v)_E + \sum_{E \in \mathcal{E}_h \cap \Gamma_n} (g - \varepsilon \underline{n}_E \cdot \nabla u_h, v)_E \\ &= \sum_{K \in \mathcal{T}_h} (R_K, v)_K + \sum_{E \in \mathcal{E}_h} (R_E, v)_E \\ &\quad + \sum_{K \in \mathcal{T}_h} (D_K, v)_K + \sum_{E \in \mathcal{E}_h \cap \Gamma_n} (D_E, v)_E. \end{aligned} \tag{4.11}$$

Lemma 3.3 and the Cauchy–Schwarz inequality therefore imply for all $v \in H_D^1(\Omega)$

$$\begin{aligned}&\langle R(u_h),v - I_hv\rangle \\ &\qquad \le c|||v|||\Bigg\{\sum_{K\in\mathcal{T}_h}\alpha_K^2\|R_K\|_{0;K}^2 + \sum_{E\in\mathcal{E}_h}\varepsilon^{-1/2}\alpha_E\|R_E\|_{0;E}^2 \\ &\qquad\qquad + \sum_{K\in\mathcal{T}_h}\alpha_K^2\|D_K\|_{0;K}^2 + \sum_{E\in\mathcal{E}_h\cap\Gamma_n}\varepsilon^{-1/2}\alpha_E\|D_E\|_{0;E}^2\Bigg\}^{1/2}.\end{aligned} \tag{4.12}$$

The constant c depends only on the constants c_1 and c_2 of Lemma 3.3 and on the ratios h_K/ρ_K.

From the definition of problems (2.3) and (2.10) we conclude that

$$\langle R(u_h), I_hv\rangle = \sum_{K\in\mathcal{T}_h}\delta_K\{(R_K, \underline{a}\cdot\nabla I_hv)_K + (D_K, \underline{a}\cdot\nabla I_hv)_K\}.$$

Lemma 3.3, condition (2.9), and the Cauchy–Schwarz inequality therefore imply

$$\langle R(u_h), I_hv\rangle \le c|||v|||\Bigg\{\sum_{K\in\mathcal{T}_h}\alpha_K^2\big\{\|R_K\|_{0;K}^2 + \|D_K\|_{0;K}^2\big\}\Bigg\}^{1/2}. \tag{4.13}$$

Estimates (4.10), (4.12), and (4.13) and the triangle inequality for the D_K prove the upper bound (4.7).

For the proof of the lower bound (4.8) we proceed as in the proof of [12, Lemma 5.1] and define a function w_h by

$$w_h = \gamma_1\sum_{K\in\mathcal{T}_h}\alpha_K^2\psi_KR_K + \gamma_2\sum_{E\in\mathcal{E}_h}\varepsilon^{-1/2}\alpha_E\psi_ER_E. \tag{4.14}$$

The constants γ_1 and γ_2 are arbitrary at present and will be determined below. The subsequent arguments are based on the following observation:

- the supports of the ψ_K are mutually disjoint;
- the support of a ψ_K intersects the support of at most $2n$ different ψ_E's;
- the support of a ψ_E intersects the support of at most two ψ_K's;
- the support of a ψ_E intersects the support of at most $2n-2$ other ψ_E's.

Lemma 3.6 therefore yields

$$\begin{aligned}|||w_h|||^2 \le{}& \gamma_1^2\sum_{K\in\mathcal{T}_h}\alpha_K^4|||\psi_KR_K|||_K^2 \\ &+ 2\gamma_1\gamma_2\sum_{K\in\mathcal{T}_h}\Bigg\{\sum_{E;\omega_E\cap K\neq\emptyset}\alpha_K^2\varepsilon^{-1/2}\alpha_E|||\psi_KR_K|||_K|||\psi_ER_E|||_K\Bigg\} \\ &+ \gamma_2^2\sum_{E\in\mathcal{E}_h}\Bigg\{\sum_{E';\omega_E\cap\omega_{E'}\neq\emptyset}\varepsilon^{-1}\alpha_E\alpha_{E'}|||\psi_ER_E|||_{\omega_E}|||\psi_{E'}R_{E'}|||_{\omega_{E'}}\Bigg\} \\ \le{}& (2n+1)\max\{\gamma_1^2,\gamma_2^2\}\max\{c_5,c_7\}\sum_{K\in\mathcal{T}_h}\eta_K^2.\end{aligned} \tag{4.15}$$

Since $h_E \le h_K$ for all faces E of any element K, Lemma 3.6 also implies that

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h}(R_K, w_h)_K + \sum_{E\in\mathcal{E}_h}(R_E, w_h)_E \\
&\quad = \gamma_1 \sum_{K\in\mathcal{T}_h} \alpha_K^2 (R_K, \psi_K R_K)_K + \gamma_2 \sum_{E\in\mathcal{E}_h} \varepsilon^{-1/2}\alpha_E (R_E, \psi_E R_E)_E \\
&\qquad + \gamma_2 \sum_{E\in\mathcal{E}_h} \left\{ \sum_{K;K\cap\omega_E\neq\emptyset} \varepsilon^{-1/2}\alpha_E (R_K, \psi_E R_E)_K \right\} \\
&\quad \ge \gamma_1 \sum_{K\in\mathcal{T}_h} c_4 \alpha_K^2 \|R_K\|_{0;K}^2 + \gamma_2 \sum_{E\in\mathcal{E}_h} c_6 \varepsilon^{-1/2}\alpha_E \|R_E\|_{0;E}^2 \\
&\qquad - \gamma_2 \sum_{E\in\mathcal{E}_h} \left\{ \sum_{K;K\cap\omega_E\neq\emptyset} c_8 \varepsilon^{-1/4}\alpha_E^{1/2}\alpha_K \|R_K\|_{0;K} \|\psi_E R_E\|_{0;E} \right\} \\
&\quad \ge (\gamma_1 c_4 - 2n\gamma_2 c_8^2 c_6^{-1}) \sum_{K\in\mathcal{T}_h} \alpha_K^2 \|R_K\|_{0;K}^2 + \frac{1}{2}\gamma_2 c_6 \sum_{E\in\mathcal{E}_h} \varepsilon^{-1/2}\alpha_E \|R_E\|_{0;E}^2 \\
&\quad \ge \min\left\{ \gamma_1 c_4 - 2n\gamma_2 c_8^2 c_6^{-1}, \frac{1}{2}\gamma_2 c_6 \right\} \sum_{K\in\mathcal{T}_h} \eta_K^2.
\end{aligned} \tag{4.16}$$

From Lemma 3.6 we also obtain

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h}(D_K, w_h)_K + \sum_{E\in\mathcal{E}_h\cap\Gamma_n}(D_E, w_h)_E \\
&\quad = \gamma_1 \sum_{K\in\mathcal{T}_h} \alpha_K^2 (D_K, \psi_K R_K)_K \\
&\qquad + \gamma_2 \sum_{K\in\mathcal{T}_h} \left\{ \sum_{E;E\subset\partial K} \varepsilon^{-1/2}\alpha_E (D_K, \psi_E R_E)_K \right\} \\
&\qquad + \gamma_2 \sum_{E\in\mathcal{E}_h\cap\Gamma_n} \varepsilon^{-1/2}\alpha_E (D_E, \psi_E R_E)_E \\
&\quad \le \gamma_1 \sum_{K\in\mathcal{T}_h} \alpha_K^2 \|R_K\|_{0;K} \|D_K\|_{0;K} \\
&\qquad + \gamma_2 \sum_{K\in\mathcal{T}_h} \left\{ \sum_{E;E\subset\partial K} c_8 \varepsilon^{-1/4}\alpha_E^{3/2} \|R_E\|_{0;E} \|D_K\|_{0;K} \right\} \\
&\qquad + \gamma_2 \sum_{E\in\mathcal{E}_h\cap\Gamma_n} \varepsilon^{-1/2}\alpha_E \|R_E\|_{0;E} \|D_E\|_{0;E} \\
&\quad \le 2n \max\{\gamma_1, \gamma_2\} \max\{1, c_8\} \left\{ \sum_{K\in\mathcal{T}_h} \Theta_K^2 \right\}^{1/2} \left\{ \sum_{K\in\mathcal{T}_h} \eta_K^2 \right\}^{1/2}.
\end{aligned} \tag{4.17}$$

Now we choose

$$\gamma_2 = \frac{2}{c_6} \quad \text{and} \quad \gamma_1 = \frac{1}{c_4}\left(1 + \frac{4nc_8^2}{c_6^2}\right).$$

This choice gives

$$\min\left\{\gamma_1 c_4 - 2n\gamma_2 c_8^2 c_6^{-1}, \frac{1}{2}\gamma_2 c_6\right\} = 1.$$

Estimates (4.16), (4.17), and (4.15), equation (4.11), and the triangle inequality for the D_K therefore imply

$$\sum_{K\in\mathcal{T}_h} \eta_K^2 \le c\left\{\sum_{K\in\mathcal{T}_h} \eta_K^2\right\}^{1/2}\left[|||R(u_h)|||_* + \left\{\sum_{K\in\mathcal{T}_h} \Theta_K^2\right\}^{1/2}\right].$$

In combination with Lemma 3.1 this establishes the lower bound (4.8). □

5. An error estimator based on the solution of local Dirichlet problems.

In this section we present an error estimator that is based on the solution of auxiliary local discrete problems with Dirichlet boundary conditions. To this end we fix an arbitrary element K and set

$$V_K = \operatorname{span}\{\psi_{K'}v, \psi_E\sigma : K' \subset \omega_K, E \subset \partial K\backslash\Gamma_D, v \in R_k(K'), \sigma \in R_k(E)\}.$$

Then we consider the following problem: Find $v \in V_K$ such that

$$\begin{aligned}\varepsilon(\nabla v, \nabla w)_{\omega_K} &+ (\underline{a}\cdot\nabla v, w)_{\omega_K} + (bv, w)_{\omega_K}\\ &= (f_h, w)_{\omega_K} + (g_h, w)_{\partial K\cap\Gamma_N} - \varepsilon(\nabla u_h, \nabla w)_{\omega_K}\\ &\quad - (\underline{a}_h\cdot\nabla u_h, w)_{\omega_K} - (b_h u_h, w)_{\omega_K} \quad \forall w \in V_K.\end{aligned}\tag{5.1}$$

The following lemma is a discrete analogue of Lemma 3.1. It in particular implies the unique solvability of problem (5.1).

LEMMA 5.1. *Denote by π_{V_K} the L^2-projection onto V_K. Then the following estimates are valid:*

$$\begin{aligned}\sup_{v\in V_K\backslash\{0\}}\ \sup_{w\in V_K\backslash\{0\}} &\frac{\varepsilon(\nabla v, \nabla w)_{\omega_K} + (\underline{a}\cdot\nabla v, w)_{\omega_K} + (bv, w)_{\omega_K}}{\left\{|||v|||_{\omega_K}^2 + \alpha_K^2\|\pi_{V_K}(\underline{a}\cdot\nabla v)\|_{0;\omega_K}^2\right\}^{1/2}|||w|||_{\omega_K}}\\ &\le \sqrt{2}\max\{c_a, c_b, 1\}\end{aligned}\tag{5.2}$$

and

$$\begin{aligned}\inf_{v\in V_K\backslash\{0\}}\ \sup_{w\in V_K\backslash\{0\}} &\frac{\varepsilon(\nabla v, \nabla w)_{\omega_K} + (\underline{a}\cdot\nabla v, w)_{\omega_K} + (bv, w)_{\omega_K}}{\left\{|||v|||_{\omega_K}^2 + \alpha_K^2\|\pi_{V_K}(\underline{a}\cdot\nabla v)\|_{0;\omega_K}^2\right\}^{1/2}|||w|||_{\omega_K}}\\ &\ge \frac{1}{1 + 3c_a^2\max\{c_b, 1\}^2}.\end{aligned}\tag{5.3}$$

The constant c_b is that of assumption (A3). *The constant c_a depends only on the polynomial degree k and on the ratios h_K/ρ_K (cf.* (5.5) *below).*

Proof. The definition of π_{V_K} implies

$$(\underline{a}\cdot\nabla v, w)_{\omega_K} = (\pi_{V_K}(\underline{a}\cdot\nabla v), w)_{\omega_K} \quad \forall v, w \in V_K.$$

This identity, assumption (A3), and the definition (2.4) of the energy norm yield for all $v, w \in V_K$

$$\begin{aligned}\varepsilon(\nabla v, \nabla w)_{\omega_K} &+ (\underline{a}\cdot\nabla v, w)_{\omega_K} + (bv, w)_{\omega_K}\\ &\le \max\{c_b, 1\}|||v|||_{\omega_K}|||w|||_{\omega_K} + \|\pi_{V_K}(\underline{a}\cdot\nabla v)\|_{0;\omega_K}\|w\|_{0;\omega_K}.\end{aligned}\tag{5.4}$$

Since the functions in V_K vanish at the vertices of K, the norms $h_K\|\nabla.\|_{0;\omega_K}$ and $\| .\|_{0;\omega_K}$ are equivalent on V_K. Therefore $\alpha_K|||.|||_{\omega_K}$ and $\| .\|_{0;\omega_K}$ are also equivalent norms on V_K; i.e., there is a constant $c_a \geq 1$ which depends only on the polynomial degree k and on the ratios h_K/ρ_K such that

$$\frac{1}{c_a}\alpha_K|||w|||_{\omega_K} \leq \|w\|_{0;\omega_K} \leq c_a\alpha_K|||w|||_{\omega_K} \tag{5.5}$$

holds for all $w \in V_K$. Inequalities (5.4) and (5.5) prove the upper bound (5.2).

For the proof of the lower bound (5.3) we proceed as in the proof of estimate (3.2). We consider an arbitrary function $v \in V_K$ and set $w_\gamma = v + \gamma\alpha_K^2\pi_{V_K}(\underline{a}\cdot\nabla v)$. The constant γ is arbitrary at present and will be determined below. The norm equivalence (5.5) implies

$$\begin{aligned}|||w_\gamma|||_{\omega_K} &\leq |||v|||_{\omega_K} + c_a\gamma\alpha_K\|\pi_{V_K}(\underline{a}\cdot\nabla v)\|_{0;\omega_K}\\ &\leq \left\{1 + c_a^2\gamma^2\right\}^{1/2}\left\{|||v|||_{\omega_K}^2 + \alpha_K^2\|\pi_{V_K}(\underline{a}\cdot\nabla v)\|_{0;\omega_K}^2\right\}^{1/2}.\end{aligned}$$

Assumptions (A3) and (A4), integration by parts, and the definition (2.4) of the energy norm on the other hand imply that the bilinear form on the left-hand side of problem (5.1) is coercive on V_K with constant 1. Inserting w_γ as a test-function in this bilinear form therefore yields

$$\begin{aligned}&\varepsilon(\nabla v, \nabla w_\gamma)_{\omega_K} + (\underline{a}\cdot\nabla v, w_\gamma)_{\omega_K} + (bv, w_\gamma)_{\omega_K}\\ &\quad= \varepsilon(\nabla v, \nabla v)_{\omega_K} + (\underline{a}\cdot\nabla v, v)_{\omega_K} + (bv, v)_{\omega_K}\\ &\qquad+ \gamma\alpha_K^2\{\varepsilon(\nabla v, \nabla\pi_{V_K}(\underline{a}\cdot\nabla v))_{\omega_K} + (\underline{a}\cdot\nabla v, \pi_{V_K}(\underline{a}\cdot\nabla v))_{\omega_K} + (bv, \pi_{V_K}(\underline{a}\cdot\nabla v))_{\omega_K}\}\\ &\quad\geq |||v|||_{\omega_K}^2 + \gamma\alpha_K^2\|\pi_{V_K}(\underline{a}\cdot\nabla v)\|_{0;\omega_K}^2 - c_a\max\{c_b, 1\}\gamma\alpha_K|||v|||_{\omega_K}\|\pi_{V_K}(\underline{a}\cdot\nabla v)\|_{0;\omega_K}\\ &\quad\geq \left(1 - \frac{1}{2}c_a^2\max\{c_b, 1\}^2\gamma\right)|||v|||_{\omega_K}^2 + \frac{1}{2}\gamma\alpha_K^2\|\pi_{V_K}(\underline{a}\cdot\nabla v)\|_{0;\omega_K}^2.\end{aligned}$$

Now we choose $\gamma = 2/(1 + c_a^2\max\{c_b, 1\}^2)$. Since

$$\begin{aligned}&(1 + c_a^2\max\{c_b, 1\}^2)\left\{1 + \frac{4c_a^2}{(1 + c_a^2\max\{c_b, 1\}^2)^2}\right\}^{1/2}\\ &\quad= \left\{(1 + c_a^2\max\{c_b, 1\}^2)^2 + 4c_a^2\right\}^{1/2}\\ &\quad\leq 1 + 3c_a^2\max\{c_b, 1\}^2\end{aligned}$$

this proves estimate (5.3). □

We denote by $v_K \in V_K$ the unique solution of problem (5.1) and define the error indicator $\eta_{D,K}$ by

$$\eta_{D,K} = \left\{|||v_k|||_{\omega_K}^2 + \alpha_K^2\|\pi_{V_K}(\underline{a}\cdot\nabla v_K)\|_{0;\omega_K}^2\right\}^{1/2}. \tag{5.6}$$

Remark 5.2. The function $u_h + v_K$ is a finite element approximation to the solution u_K of the local convection-diffusion problem

$$\begin{aligned}-\varepsilon\Delta u_K + \underline{a}\nabla\cdot u_K + bu_K &= f_h \quad \text{in } \omega_K,\\ u_K &= u_h \quad \text{on } \partial\omega_K\backslash(\partial K \cap \Gamma_N),\\ \varepsilon\underline{n}_K\cdot\nabla u_K &= g_h \quad \text{on } \partial K \cap \Gamma_N.\end{aligned}$$

Problem (5.1) is the same as problem (5.1) in [10]. But, contrary to [10], the present error indicator also takes into account the convective derivative of v_K. This modification is crucial for obtaining fully robust error estimates.

THEOREM 5.3. *There are two constants $c_\dagger$ and $c^\dagger$ which depend only on the polynomial degree k and on the ratios h_K/ρ_K such that the estimate*

$$\frac{1}{c_\dagger}\eta_K \le \eta_{D,K} \le c^\dagger \left\{ \sum_{K'\subset\omega_K} \eta_{K'}^2 \right\}^{1/2} \tag{5.7}$$

holds for all elements K. Moreover $\eta_{D,K}$ yields the upper error bound

$$|||u-u_h||| + |||\underline{a}\cdot\nabla(u-u_h)|||_* \le \hat{c}^* \left\{ \sum_{K\in\mathcal{T}_h} \left[\eta_{D,K}^2 + \Theta_K^2\right] \right\}^{1/2} \tag{5.8}$$

and the lower error bound

$$\left\{ \sum_{K\in\mathcal{T}_h} \eta_{D,K}^2 \right\}^{1/2} \le \hat{c}_* \left[|||u-u_h||| + |||\underline{a}\cdot\nabla(u-u_h)|||_* + \left\{ \sum_{K\in\mathcal{T}_h} \Theta_K^2 \right\}^{1/2} \right]. \tag{5.9}$$

The constants $\hat{c}_*$ and $\hat{c}^*$ depend only on the polynomial degree k and on the ratios h_K/ρ_K.

Proof. In view of Theorem 4.1 we only have to prove estimate (5.7). Integration by parts of the right-hand side of problem (5.1) yields for all $w \in V_K$

$$\begin{aligned}(f_h, w)_{\omega_K} &+ (g_h, w)_{\partial K\cap\Gamma_N} - \varepsilon(\nabla u_h, \nabla w)_{\omega_K} - (\underline{a}_h\cdot\nabla u_h, w)_{\omega_K} - (b_h u_h, w)_{\omega_K} \\ &= \sum_{K'\subset\omega_K} (R_{K'}, w)_{K'} + \sum_{E\subset\partial K\backslash\Gamma_D} (R_E, w)_E.\end{aligned}$$

Hence we have

$$\begin{aligned}&\sup_{w\in V_K} \frac{1}{|||w|||_{\omega_K}} \left\{\varepsilon(\nabla v_K, \nabla w)_{\omega_K} + (\underline{a}\cdot\nabla v_K, w)_{\omega_K} + (b v_K, w)_{\omega_K}\right\} \\ &\qquad = \sup_{w\in V_K} \frac{1}{|||w|||_{\omega_K}} \left\{ \sum_{K'\subset\omega_K} (R_{K'}, w)_{K'} + \sum_{E\subset\partial K\backslash\Gamma_D} (R_E, w)_E \right\}.\end{aligned} \tag{5.10}$$

Lemma 5.1 implies that the left-hand side of (5.10) is bounded from above and from below by constant multiples of $\{|||v_K|||_{\omega_K}^2 + \alpha_K^2\|\pi_{V_K}(\underline{a}\cdot\nabla v_K)\|_{0;\omega_K}^2\}^{1/2}$. The norm equivalence (5.5) yields that the right-hand side of (5.10) is bounded from above by a constant multiple of $\left\{\sum_{K'\subset\omega_K}\eta_{K'}^2\right\}^{1/2}$. The arguments in establishing estimates (4.15) and (4.16) on the other hand show that, taking a suitable combination of functions $\psi_{K'}R_{K'}$, $K' \subset \omega_K$, and $\psi_E R_E$, $E \subset \partial K\backslash\Gamma_N$, allows us to bound the right-hand side of (5.10) from below by a constant multiple of $\{\sum_{K'\subset\omega_K}\alpha_{K'}^2\|R_{K'}\|_{0;K'}^2 + \sum_{E\subset\partial K\backslash\Gamma_D}\varepsilon^{-1/2}\alpha_E\|R_E\|_{0;E}^2\}^{1/2}$, which in turn is an upper bound for η_K. □

6. An error estimator based on the solution of local Neumann problems. In this section we present an error estimator that is based on the solution of auxiliary local discrete problems with Neumann boundary conditions. The main difficulty in constructing the discrete local problem now is to ensure the coercivity of the corresponding bilinear form. To achieve this we have to approximate the reaction b and the convection $\underline{a}$ by discrete quantities such that assumption (A3) remains valid for this approximation and such that the normal component of the discrete convection is piecewise constant on the edges, respectively, faces, of $\mathcal{T}_h$.

To this end we recall that b_h is the L^2-projection of b onto the space of piecewise constant functions corresponding to $\mathcal{T}_h$. For the approximation of the convection we denote by $\underline{a}_{\mathrm{RT}_h}$ the approximation of $\underline{a}$ in the lowest-order Raviart–Thomas space RT_0 corresponding to $\mathcal{T}_h$ which is defined by (cf. [3, sections III.3.1 and III.3.2]):

$$\underline{a}_{\mathrm{RT}_h|K} \in R_0(K)^n + \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} R_0(K) \quad \forall K \in \mathcal{T}_h,$$

$$\int_E \underline{n}_E \cdot \underline{a}_{\mathrm{RT}_h} = \int_E \underline{n}_E \cdot \underline{a} \quad \forall E \in \mathcal{E}_h.$$

Note that $\underline{n}_E \cdot \underline{a}_{\mathrm{RT}_h}$ is piecewise constant on the edges, respectively, faces. Therefore we can associate with each element K the collection of its outflow edges, respectively, faces, by setting

$$\mathcal{E}_K^+ = \{E \in \mathcal{E}_h \cap \partial K : \underline{n}_K \cdot \underline{a}_{\mathrm{RT}_h} \geq 0\},$$

where $\underline{n}_K$ denotes the outward normal to K.

With these definitions we set

$$\widetilde{V}_K = \mathrm{span}\,\{\psi_K v, \psi_E \sigma : v \in R_k(K), E \in \mathcal{E}_K^+, \sigma \in R_k(E)\}$$

and consider the following problem: Find $\widetilde{v} \in \widetilde{V}_K$ such that

$$\begin{aligned} &\varepsilon(\nabla\widetilde{v}, \nabla w)_K + (\underline{a}_{\mathrm{RT}_h} \cdot \nabla\widetilde{v}, w)_K + (b_h\widetilde{v}, w)_K \\ &\qquad = (R_K, w)_K + \sum_{E\in\mathcal{E}_K^+} (R_E, w)_E \quad \forall w \in \widetilde{V}_K. \end{aligned} \tag{6.1}$$

The following lemma is an analogue of Lemma 5.1. It in particular implies the unique solvability of problem (6.1).

Lemma 6.1. *The following estimates hold for all elements K:*

$$\begin{aligned} &\sup_{v\in\widetilde{V}_K\setminus\{0\}} \sup_{w\in\widetilde{V}_K\setminus\{0\}} \frac{\varepsilon(\nabla v, \nabla w)_K + (\underline{a}_{\mathrm{RT}_h} \cdot \nabla v, w)_K + (b_h v, w)_K}{\{|||v|||_K^2 + \alpha_K^2 \|\underline{a}_{\mathrm{RT}_h} \cdot \nabla v\|_{0;K}^2\}^{1/2} |||w|||_K} \\ &\qquad \leq \sqrt{2}\max\{\widetilde{c}_a, c_b, 1\} \end{aligned} \tag{6.2}$$

and

$$\begin{aligned} &\inf_{v\in\widetilde{V}_K\setminus\{0\}} \sup_{w\in\widetilde{V}_K\setminus\{0\}} \frac{\varepsilon(\nabla v, \nabla w)_K + (\underline{a}_{\mathrm{RT}_h} \cdot \nabla v, w)_K + (b_h v, w)_K}{\{|||v|||_K^2 + \alpha_K^2 \|\underline{a}_{\mathrm{RT}_h} \cdot \nabla v\|_{0;K}^2\}^{1/2} |||w|||_K} \\ &\qquad \geq \frac{1}{1 + 3\widetilde{c}_a^2 \max\{c_b, 1\}^2}. \end{aligned} \tag{6.3}$$

The constant c_b is that of assumption (A3). *The constant $\widetilde{c}_a$ depends only on the polynomial degree k and on the ratios h_K/ρ_K (cf.* (6.6) *below).*

Proof. Choose an arbitrary element K and keep it fixed in what follows.

Since b_h is constant on K we conclude from assumption (A3) that

$$\|b_h\|_{L^\infty(K)} = \left|\frac{1}{|K|}\int_K b_h\right| = \left|\frac{1}{|K|}\int_K b\right| \le c_b\beta.$$

From [3, sections III.3.1 and III.3.2] we know that $\operatorname{div}\underline{a}_{\mathrm{RT}_h}$ is constant on K and satisfies $\int_K \operatorname{div}\underline{a}_{\mathrm{RT}_h} = \int_K \operatorname{div}\underline{a}$. Hence we get from assumption (A3)

$$b_h - \frac{1}{2}\operatorname{div}\underline{a}_{\mathrm{RT}_h} = \frac{1}{|K|}\int_K \left\{b_h - \frac{1}{2}\operatorname{div}\underline{a}_{\mathrm{RT}_h}\right\} = \frac{1}{|K|}\int_K \left\{b - \frac{1}{2}\operatorname{div}\underline{a}\right\} \ge \beta. \tag{6.4}$$

This shows that b_h and $\underline{a}_{\mathrm{RT}_h}$ satisfy assumption (A3) with the same constants β and c_b. Hence we obtain the following analogue of estimate (5.4) for all $v, w \in \widetilde{V}_K$:

$$\begin{aligned}\varepsilon(\nabla v,\nabla w)_K &+ (\underline{a}_{\mathrm{RT}_h}\cdot\nabla v, w)_K + (b_h v, w)_K \\ &\le \max\{c_b,1\}|||v|||_K|||w|||_K + \|\underline{a}_{\mathrm{RT}_h}\cdot\nabla v\|_{0;K}\|w\|_{0;K}.\end{aligned} \tag{6.5}$$

The same arguments as in the proof of estimate (5.5) imply that there is a constant $\widetilde{c}_a \ge 1$ which depends only on the polynomial degree k and on the ratios h_K/ρ_K such that

$$\frac{1}{\widetilde{c}_a}\alpha_K|||w|||_{\omega_K} \le \|w\|_{0;\omega_K} \le \widetilde{c}_a\alpha_K|||w|||_{\omega_K} \tag{6.6}$$

holds for all $w \in \widetilde{V}_K$. Inequalities (6.5) and (6.6) prove the upper bound (6.2).

For the proof of the lower bound (6.3) we only have to check the coercivity of the bilinear form on the left-hand side of problem (6.1). Once this is done, the inf-sup condition (6.3) is established with the same arguments as in the proof of Lemma 6.1. For every $w \in \widetilde{V}_K$ we have

$$\begin{aligned}\varepsilon(\nabla w,\nabla w)_K &+ (\underline{a}_{\mathrm{RT}_h}\cdot\nabla w, w)_K + (b_h w, w)_K \\ &= \varepsilon\|\nabla w\|^2_{0;K} + \int_K \frac{1}{2}\operatorname{div}(\underline{a}_{\mathrm{RT}_h}w^2) + \int_K\left\{b_h - \frac{1}{2}\operatorname{div}\underline{a}_{\mathrm{RT}_h}\right\}w^2 \\ &= \varepsilon\|\nabla w\|^2_{0;K} + \int_{\partial K}\frac{1}{2}\underline{n}_K\cdot\underline{a}_{\mathrm{RT}_h}w^2 + \int_K\left\{b_h - \frac{1}{2}\operatorname{div}\underline{a}_{\mathrm{RT}_h}\right\}w^2 \\ &\ge \varepsilon\|\nabla w\|^2_{0;K} + \beta\|w\|^2_{0;K}.\end{aligned}$$

In the last step we have used estimate (6.4) and the definition of $\widetilde{V}_K$, which implies that $\int_{\partial K}\frac{1}{2}\underline{n}_K\cdot\underline{a}_{\mathrm{RT}_h}w^2 \ge 0$. □

We denote by $\widetilde{v}_K \in \widetilde{V}_K$ the unique solution of problem (6.1) and define the error indicator $\eta_{N,K}$ by

$$\eta_{N,K} = \left\{|||\widetilde{v}_k|||^2_K + \alpha_K^2\|\underline{a}_{\mathrm{RT}_h}\cdot\nabla\widetilde{v}_K\|^2_{0;K}\right\}^{1/2}. \tag{6.7}$$

Remark 6.2. The function $\widetilde{v}_K$ is a finite element approximation to the solution $\widetilde{u}_K$ of the local convection-diffusion problem

$$\begin{aligned}-\varepsilon\Delta\widetilde{u}_K + \underline{a}\nabla\cdot\widetilde{u}_K + b\widetilde{u}_K &= R_K \quad \text{in } K,\\ \varepsilon\underline{n}_K\cdot\nabla\widetilde{u}_K &= R_E \quad \text{on } E\ ,\ E\in\mathcal{E}_K^+,\\ \widetilde{u}_K &= 0 \quad \text{on } \partial K\backslash\cup_{E\in\mathcal{E}_K^+}E.\end{aligned}$$

Due to the approximation of the convection and reaction terms, problem (6.1) is different from problem (5.7) in [10]. Moreover the present error indicator also takes into account the convective derivative of $\widetilde{v}_K$. These modifications are crucial for obtaining fully robust error estimates.

THEOREM 6.3. *There are two constants $\widetilde{c}_\dagger$ and $\widetilde{c}^\dagger$ which depend only on the polynomial degree k and on the ratios h_K/ρ_K such that the estimate*

$$\frac{1}{\widetilde{c}_\dagger}\left\{\alpha_K^2\|R_K\|_{0;K}^2+\sum_{E\in\mathcal{E}_K^+}\varepsilon^{-1/2}\alpha_E\|R_E\|_{0;E}^2\right\}^{1/2}\le\eta_{N,K}\le\widetilde{c}^\dagger\eta_K \tag{6.8}$$

holds for all elements K. Moreover $\eta_{N,K}$ yields the upper error bound

$$|||u-u_h|||+|||\underline{a}\cdot\nabla(u-u_h)|||_*\le\widetilde{c}^*\left\{\sum_{K\in\mathcal{T}_h}\left[\eta_{N,K}^2+\Theta_K^2\right]\right\}^{1/2} \tag{6.9}$$

and the lower error bound

$$\left\{\sum_{K\in\mathcal{T}_h}\eta_{N,K}^2\right\}^{1/2}\le\widetilde{c}_*\left[|||u-u_h|||+|||\underline{a}\cdot\nabla(u-u_h)|||_*+\left\{\sum_{K\in\mathcal{T}_h}\Theta_K^2\right\}^{1/2}\right]. \tag{6.10}$$

The constants $\widetilde{c}_*$ and $\widetilde{c}^*$ depend only on the polynomial degree k and on the ratios h_K/ρ_K.

Proof. Estimate (6.8) is proven with the same arguments as estimate (5.7). The error bounds (6.9) and (6.10) follow from estimate (6.8), Theorem 4.1, and the observation that for every edge, respectively, face, E, which is not part of the Dirichlet boundary Γ_D, there is at least one element K_E with $E\in\mathcal{E}_{K_E}^+$. □

Acknowledgments. The main part of this work was carried out during a stay at the Department of Mathematics of the Polytechnico di Torino. We want to thank Prof. Claudio Canuto and Dr. Stefano Berrone for their kind hospitality and the many fruitful discussions.

REFERENCES

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

[2] L. ANGERMANN, *Balanced a posteriori error estimates for finite volume type discretizations of convection dominated elliptic problems*, Computing, 55 (1995), pp. 305–323.

[3] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.

[4] L. P. FRANCA, S. L. FREY, AND T. J. R. HUGHES, *Stabilized finite element methods* I: *Application to the advective-diffusive model*, Comput. Methods Appl. Mech. Engrg., 95 (1992), pp. 253–276.

[5] T. J. R. HUGHES AND A. BROOKS, *Streamline upwind/Petrov Galerkin formulations for the convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 54 (1982), pp. 199–259.

[6] G. KUNERT, *An a posteriori residual error estimator for the finite element method on anisotropic tetrahedral meshes*, Numer. Math., 86 (2000), pp. 471–490.

[7] G. KUNERT AND R. VERFÜRTH, *Edge residuals dominate a posteriori error estimates for linear finite element methods on anisotropic triangular and tetrahedral meshes*, Numer. Math., 86 (2000), pp. 283–303.

[8] G. SANGALLI, *A robust a posteriori error estimator for the residual-free bubbles method applied to advection-diffusion problems*, Numer. Math., 89 (2001), pp. 379–399.

[9] G. Sangalli, *A uniform analysis of nonsymmetric and coercive linear operators*, SIAM J. Math. Anal., 36 (2005), pp. 2033–2048.

[10] R. Verfürth, *A posteriori error estimators for convection-diffusion equations*, Numer. Math., 80 (1998), pp. 641–663.

[11] R. Verfürth, *Error estimates for some quasi-interpolation operators*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 695–713.

[12] R. Verfürth, *A posteriori error estimates for finite element discretizations of the heat equation*, Calcolo, 40 (2003), pp. 195–212.

SIAM J. NUMER. ANAL.
Vol. 43, No. 4, pp. 1783–1802

ROBUST A POSTERIORI ERROR ESTIMATES FOR NONSTATIONARY CONVECTION-DIFFUSION EQUATIONS[*]

R. VERFÜRTH[†]

Abstract. We consider discretizations of convection dominated nonstationary convection-diffusion equations by A-stable θ-schemes in time and conforming finite elements in space on locally refined, isotropic meshes. For these discretizations we derive a residual a posteriori error estimator. The estimator yields upper bounds on the error which are global in space and time and lower bounds that are global in space and local in time. The error estimates are fully robust in the sense that the ratio between upper and lower bounds is uniformly bounded in time, does not depend on any step-size in space or time nor on any relation between these both, and is uniformly bounded with respect to the size of the convection. Moreover, the estimates are uniform with respect to the size of the zero-order reaction term and also hold for the limit case of vanishing reaction.

Key words. a posteriori error estimates, convection dominated convection-diffusion equations, θ-scheme

AMS subject classifications. 65N30, 65N15, 65J15

DOI. 10.1137/040604273

1. Introduction.

We consider nonstationary convection-diffusion equations

$$\begin{aligned} \frac{\partial u}{\partial t} - \varepsilon\Delta u + \underline{a}\cdot\nabla u + bu &= f \qquad \text{in } \Omega\times(0,T], \\ u &= 0 \qquad \text{on } \Gamma_D\times(0,T], \\ \varepsilon\frac{\partial u}{\partial n} &= g \qquad \text{on } \Gamma_N\times(0,T], \\ u &= u_0 \quad \text{in } \Omega \end{aligned} \tag{1.1}$$

in a bounded space-time cylinder with a polygonal cross-section $\Omega\subset\mathbb{R}^d$, $d\geq 2$, having a Lipschitz boundary Γ consisting of two disjoint parts Γ_D and Γ_N. The final time T is arbitrary, but kept fixed in what follows. We assume that the data satisfy the following conditions:

(A1) $f\in C(0,T;L^2(\Omega))$, $g\in C(0,T;L^2(\Gamma_N))$, $\underline{a}\in C(0,T;W^{1,\infty}(\Omega)^d)$, $b\in C(0,T;L^\infty(\Omega))$.

(A2) $0<\varepsilon\ll 1$.

(A3) There are two constants $\beta\geq 0$ and $c_b\geq 0$, which do not depend on ε, such that $-\frac{1}{2}\mathrm{div}\underline{a}+b\geq\beta$ and $\|b\|_{L^\infty(\Omega)}\leq c_b\beta$ in $(0,T]$.

(A4) The Dirichlet boundary Γ_D has positive $(d-1)$-dimensional measure and includes the inflow boundary $\{x\in\Gamma : \underline{a}(x)\cdot\underline{n}(x)<0\}$.

Assumption (A3) allows us to handle simultaneously the case of a nonvanishing zero-order reaction term and the one of absent reaction, the latter one corresponding to $\beta=0$. In the case $\beta=0$ we set $c_b=0$. Assumption (A2) of course means that we are interested in the convection-dominated regime. Assumption (A1) can be replaced

[*] Received by the editors February 18, 2004; accepted for publication (in revised form) April 22, 2005; published electronically November 22, 2005.
http://www.siam.org/journals/sinum/43-4/60427.html

[†] Fakultät für Mathematik, Ruhr-Universität Bochum, D-44780 Bochum, Germany (rv@num1.ruhr-uni-bochum.de).

by weaker conditions concerning the temporal smoothness. Its present form, however, simplifies the analysis.

We use the A-stable θ-schemes for the time discretization of problem (1.1). The spatial discretization is based on standard conforming finite element spaces using the standard Galerkin formulation or a stabilized SUPG-scheme. The spatial meshes must be shape-regular (cf. section 2). This includes locally refined meshes but excludes anisotropic elements with large aspect ratios. For this space-time discretization we analyze a residual error estimator and establish upper and lower bounds for the error. The upper bounds are global with respect to space and time; the lower bounds are global with respect to space and local with respect to time. The ratio of upper and lower bounds is uniformly bounded with respect to any mesh-size, to the final time, to the parameter β, and—most important—to the viscosity ε. Thus the error estimates are fully robust. Contrary to standard residual error estimates, the present estimator requires the solution of an auxiliary discrete stationary reaction-diffusion problem at each time-level. This is the price that we must pay for the ε-independent bounds. The computational effort for evaluating the error estimator is thus comparable to an additional time-step for each time-level and similar to the extra work required by the now popular estimators that are based on the solution of suitable discrete adjoint problems [3].

The article is organized as follows. In section 2 we introduce some function spaces and norms. Section 3 is devoted to the finite element discretization. Using energy estimates we prove in section 4 that the error is equivalent to a residual which is defined in a suitable dual space. This residual is split into three parts: one corresponding to the approximation of the data, a contribution corresponding to a spatial error, and a part corresponding to a temporal error. The latter can be further decomposed into a diffusive and a convective part. In section 5 we derive upper and lower bounds for the spatial part of the residual. The temporal part is treated in section 6. Combining these results we obtain in section 7 a first error estimator. This estimator yields upper and lower bounds on the error and is fully robust in the sense described above. However, it is not suited for practical computations since it incorporates a dual norm of the convective derivative of the finite element solution. This contribution is due to the convective part of the temporal residual. Standard approaches bound this contribution by inverse estimates and therefore lead to estimates that are no longer robust. The results of section 7 show that sharp upper and lower bounds with parameter-independent constants for this term are mandatory for obtaining a robust and computable a posteriori error estimator. In section 8 we finally bound the critical dual norm by computable quantities based on the solution of a discrete stationary reaction-diffusion problem at each time-level. This yields our final error estimates which are stated in Theorem 8.2.

2. Function spaces. For any bounded open subset ω of Ω with Lipschitz boundary γ, we denote by $H^k(\omega)$, $k \in \mathbb{N}$, $L^2(\omega) = H^0(\omega)$, and $L^2(\gamma)$ the usual Sobolev and Lebesgue spaces equipped with the standard norms $\|.\|_{k;\omega} = \|.\|_{H^k(\omega)}$ and $\|.\|_{0;\gamma} = \|.\|_{L^2(\gamma)}$ (cf. [1]). Similarly, $(.,.)_\omega$ and $(.,.)_\gamma$ denote the scalar products of $L^2(\omega)$ and $L^2(\gamma)$, respectively. If $\omega = \Omega$, we will omit the index Ω.

Set

$$H^1_D(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}. \tag{2.1}$$

We equip $H^1_D(\Omega)$ with the norm

$$|||v||| = \left\{\varepsilon\|\nabla v\|_0^2 + \beta\|v\|_0^2\right\}^{1/2}. \tag{2.2}$$

Due to assumptions (A3) and (A4) this is the natural energy norm of problem (1.1). The dual space of $H_D^1(\Omega)$ is denoted by $H_D^1(\Omega)^*$ and equipped with the norm

$$|||\varphi|||_* = \sup_{v\in H_D^1(\Omega)\setminus\{0\}} \frac{\langle\varphi, v\rangle}{|||v|||}, \tag{2.3}$$

where $\langle .,.\rangle$ denotes the corresponding duality pairing.

$H^{1/2}(\Gamma_N)$ denotes the space of Γ_N-traces of H^1-functions and is equipped with the trace norm induced by the energy norm, i.e.,

$$\|\varphi\|_{H^{1/2}(\Gamma_N)} = \inf\left\{|||v||| : v \in H_D^1(\Omega)\ ,\ v = \varphi \text{ on } \Gamma_N\right\}.$$

$H^{-1/2}(\Gamma_N)$ denotes the dual space of $H^{1/2}(\Gamma_N)$ and is equipped with the corresponding dual norm. Thus the norms of $H^{1/2}(\Gamma_N)$ and $H^{-1/2}(\Gamma_N)$ depend on the energy norm and consequently on the parameters ε and β.

For any separable Banach space V and any two numbers $a < b$ we denote by $L^2(a, b; V)$ and $L^\infty(a, b; V)$ the spaces of measurable functions u defined on (a, b) with values in V such that the function $t \to \|u(., t)\|_V$ is square integrable, respectively, essentially bounded. These are Banach spaces equipped with the norms

$$\begin{aligned}\|u\|_{L^2(a,b;V)} &= \left\{\int_a^b \|u(., t)\|_V^2 dt\right\}^{1/2},\\ \|u\|_{L^\infty(a,b;V)} &= \operatorname*{ess.sup}_{a<t<b} \|u(., t)\|_V\end{aligned}$$

(cf. [4, Vol. 5, Chap. XVIII, sect. 1]). For abbreviation we introduce the space

$$\begin{aligned}X(a, b) = \big\{u \in L^2(a, b; H_D^1(\Omega)) \cap L^\infty(a, b; L^2(\Omega)) :&\\ \partial_t u + \underline{a}\cdot\nabla u \in L^2(a, b; H_D^1(\Omega)^*)\big\}&\end{aligned} \tag{2.4}$$

and equip it with its graph norm

$$\begin{aligned}\|u\|_{X(a,b)} = \Bigg\{\operatorname*{ess.sup}_{a<t<b} \|u(., t)\|_0^2 + \int_a^b |||u(., t)|||^2 dt&\\ + \int_a^b |||(\partial_t u + \underline{a}\cdot\nabla u)(., t)|||_*^2 dt\Bigg\}^{1/2}&.\end{aligned} \tag{2.5}$$

Here the derivative $\partial_t u$ has to be understood in the distributional sense [4, Vol. 5, Chap. 18, Sect. 1].

The weak form of problem (1.1) consists in finding $u \in L^2(0, T; H_D^1(\Omega))$ such that $\partial_t u \in L^2(0, T; H_D^1(\Omega)^*)$, $u(., 0) = u_0$ in $H_D^1(\Omega)^*$, and for almost every $t \in (0, T)$ and all $v \in H_D^1(\Omega)$

$$(\partial_t u, v) + \varepsilon(\nabla u, \nabla v) + (\underline{a}\cdot\nabla u, v) + (bu, v) = (f, v) + (g, v)_{\Gamma_N}. \tag{2.6}$$

Assumptions (A1)–(A4) imply that problem (2.3) admits a unique solution [2], [4].

For later use we note that integration by parts and assumptions (A3) and (A4) imply

$$\varepsilon(\nabla v, \nabla v) + (\underline{a}\cdot\nabla v, v) + (bv, v) \geq |||v|||^2 \quad \forall v \in H_D^1(\Omega). \tag{2.7}$$

Similarly, assumption (A3) and definition (2.2) imply

$$\varepsilon(\nabla v, \nabla w) + (bv, w) \le \max\{c_b, 1\} |||v|||\, |||w||| \quad \forall v, w \in H_D^1(\Omega). \tag{2.8}$$

3. Finite element discretization. For the discretization we choose an integer $N \ge 1$ and intermediate times $0 = t_0 < t_1 < \cdots < t_N = T$ and set $\tau_n = t_n - t_{n-1}$, $1 \le n \le N$. With each intermediate time t_n, $0 \le n \le N$, we associate a partition $\mathcal{T}_{h,n}$ of Ω and a corresponding finite element space $X_{h,n}$. These have to satisfy the following conditions:

(1) *Affine equivalence:* every element $K \in \mathcal{T}_{h,n}$ can be mapped by an invertible affine mapping onto the standard reference d-simplex or the standard unit cube in $\mathbb{R}^d$.

(2) *Admissibility:* any two elements are either disjoint or share a vertex, or a complete edge, or (if $d = 3$) a complete face.

(3) *Shape-regularity:* for any element K the ratio of its diameter h_K to the diameter ρ_K of the largest inscribed ball is bounded uniformly with respect to all partitions $\mathcal{T}_{h,n}$ and to N.

(4) *Transition condition:* for $1 \le n \le N$ there is an affinely equivalent, admissible, and shape-regular partition $\widetilde{\mathcal{T}}_{h,n}$ such that it is a refinement of both $\mathcal{T}_{h,n}$ and $\mathcal{T}_{h,n-1}$ and such that $\sup_{1\le n\le N} \sup_{K\in\widetilde{\mathcal{T}}_{h,n}} \sup_{K'\in\mathcal{T}_{h,n}; K\subset K'} \frac{h_{K'}}{h_K} < \infty$.

(5) Each $X_{h,n}$ is a subset of $H_D^1(\Omega)$ and consists of continuous functions which are piecewise polynomials, the degrees being bounded uniformly with respect to all partitions $\mathcal{T}_{h,n}$ and to N.

(6) Each $X_{h,n}$ contains the space of continuous, piecewise linear finite elements corresponding to $\mathcal{T}_{h,n}$.

Condition (1) restricts quadrilateral elements to parallelograms and cubic elements to parallelepipeds. In two dimensions, triangular and quadrilateral elements may be mixed. In three dimensions this is also possible if one adds prismatic elements.

Condition (2) excludes hanging nodes.

Condition (3) is a standard one and allows for locally refined meshes. However, it excludes anisotropic elements with large aspect ratios.

Condition (4) is due to the simultaneous presence of finite element functions defined on different grids. In practice the partition $\mathcal{T}_{h,n}$ is usually obtained from $\mathcal{T}_{h,n-1}$ by a combination of refinement and of coarsening. In this case condition (4) restricts only the coarsening. It must not be too abrupt nor too strong.

We choose a parameter $\theta \in [\frac{1}{2}, 1]$ and keep it fixed in what follows. For every time-level $n \ge 1$ we introduce the abbreviations

$$\begin{aligned} f^{n\theta} &= \theta f(., t_n) + (1-\theta) f(., t_{n-1}), \\ g^{n\theta} &= \theta g(., t_n) + (1-\theta) g(., t_{n-1}), \\ \underline{a}^{n\theta} &= \theta \underline{a}(., t_n) + (1-\theta) \underline{a}(., t_{n-1}), \\ b^{n\theta} &= \theta b(., t_n) + (1-\theta) b(., t_{n-1}). \end{aligned}$$

Furthermore we denote by π_0 the L^2-projection onto $X_{h,0}$.

Then the space-time discretization of problem (1.1) consists in finding $u_h^n \in X_{h,n}$, $0 \le n \le N$, such that

$$u_h^0 = \pi_0 u_0 \tag{3.1}$$

and, for $n = 1, \dots, N$, and all $v_h \in X_{h,n}$

$$
\begin{aligned}
\left(\frac{u_h^n - u_h^{n-1}}{\tau_n}, v_h\right) &+ \varepsilon(\theta\nabla u_h^n + (1-\theta)\nabla u_h^{n-1}, \nabla v_h) \\
&+ (\underline{a}^{n\theta} \cdot \nabla(\theta u_h^n + (1-\theta)u_h^{n-1}), v_h) \\
&+ (b^{n\theta}(\theta u_h^n + (1-\theta)u_h^{n-1}), v_h) \\
&+ \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \delta_K \bigg(\frac{u_h^n - u_h^{n-1}}{\tau_n} - \varepsilon\Delta(\theta u_h^n + (1-\theta)u_h^{n-1}) \\
&\qquad + \underline{a}^{n\theta} \cdot \nabla(\theta u_h^n + (1-\theta)u_h^{n-1}) \\
&\qquad + b^{n\theta}(\theta u_h^n + (1-\theta)u_h^{n-1})\ ,\ \underline{a}^{n\theta} \cdot \nabla v_h\bigg)_K \\
&= (f^{n\theta}, v_h) + (g^{n\theta}, v_h)_{\Gamma_N} + \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \delta_K (f^{n\theta}, \underline{a}^{n\theta} \cdot \nabla v_h)_K.
\end{aligned}
\tag{3.2}
$$

The δ_K are nonnegative stabilization parameters. The choice $\delta_K = 0$ for all K yields the standard Galerkin discretization; the choice $\delta_K > 0$ for all K corresponds to the SUPG-discretizations (cf., e.g., [5], [6]). In what follows we will always assume that

$$
\delta_K \|\underline{a}^{n\theta}\|_{L^\infty(K)} \le h_K \quad \forall K \in \widetilde{\mathcal{T}}_{h,n},\ 0 \le n \le N.
\tag{3.3}
$$

This condition is satisfied for all choices of δ_K used in practice.

Assumptions (A3), (A4), and (3.3) and standard arguments for SUPG-discretizations (cf., e.g., [5], [6]) imply that problems (3.1), (3.2) admit a unique solution $(u_h^n)_{0\le n\le N}$. With this sequence we associate the function $u_{h,\tau}$ which is *piecewise affine* on the time-intervals $[t_{n-1}, t_n]$, $1 \le n \le N$, and which equals u_h^n at time t_n, $0 \le n \le N$. Since the function $t \to u_{h,\tau}(.,t)$ is continuous and piecewise affine with values in $H_D^1(\Omega)$, it is differentiable in the distributional sense [4, Vol. 5, Chap. XVIII, sect. 1] and its weak derivative satisfies

$$
\partial_t u_{h,\tau} = \frac{u_h^n - u_h^{n-1}}{\tau_n} \quad \text{on } (t_{n-1}, t_n).
\tag{3.4}
$$

4. The equivalence of error and residual.

With the function $u_{h,\tau}$ defined by the solution of problems (3.1), (3.2) we associate the residual $R(u_{h,\tau}) \in L^2(0,T; H_D^1(\Omega)^*)$ via

$$
\begin{aligned}
\langle R(u_{h,\tau}), v\rangle = (f, v) + (g, v)_{\Gamma_N} &- (\partial_t u_{h,\tau}, v) - \varepsilon(\nabla u_{h,\tau}, \nabla v) \\
&- (\underline{a} \cdot \nabla u_{h,\tau}, v) - (b u_{h,\tau}, v)
\end{aligned}
\tag{4.1}
$$

for all $v \in H_D^1(\Omega)$. The following lemma shows that this residual and the error $u - u_{h,\tau}$ are equivalent. Its proof is based on standard energy estimates. Recall that $H_D^1(\Omega)$ and its dual space $H_D^1(\Omega)^*$ are equipped with the energy norm $|||.|||$ and the dual norm $|||.|||_*$, respectively.

LEMMA 4.1. *For all $w \in L^2(0,T; H_D^1(\Omega))$ the following lower bound on the error holds:*

$$
\int_0^T \langle R(u_{h,\tau}), w\rangle dt \le \sqrt{2}\max\{1, c_b\}\|u - u_{h,\tau}\|_{X(0,T)} \|w\|_{L^2(0,T;H_D^1(\Omega))},
\tag{4.2}
$$

where c_b is the constant of assumption (A3). *Conversely, for all n between* 1 *and N, the error can be bounded from above by*

$$\begin{aligned}\|u-u_{h,\tau}\|_{X(0,t_n)} \leq \Big\{&2(1+\max\{1,c_b\}^2)\|u_0-\pi_0 u_0\|_0^2 \\ &+2(2+\max\{1,c_b\}^2)\|R(u_{h,\tau})\|^2_{L^2(0,t_n;H^1_D(\Omega)^*)}\Big\}^{1/2}.\end{aligned}\tag{4.3}$$

Proof. Equations (2.6) and (4.1) imply for all $v \in H^1_D(\Omega)$ that

$$\begin{aligned}&(\partial_t(u-u_{h,\tau}),v)+\varepsilon(\nabla(u-u_{h,\tau}),\nabla v) \\ &\qquad+(\underline{a}\cdot\nabla(u-u_{h,\tau}),v)+(b(u-u_{h,\tau}),v)=\langle R(u_{h,\tau}),v\rangle.\end{aligned}\tag{4.4}$$

This identity, definitions (2.2) and (2.3) of the norms $|||.|||$ and $|||.|||_*$, and inequality (2.8) yield for all $0<t<T$ and all $v\in H^1_D(\Omega)$ the estimate

$$\begin{aligned}\langle R(u_{h,\tau}),v\rangle \leq\; &|||(\partial_t(u-u_{h,\tau})+\underline{a}\cdot\nabla(u-u_{h,\tau}))(.,t)|||_*|||v||| \\ &+\max\{1,c_b\}|||(u-u_{h,\tau})(.,t)|||\;|||v|||.\end{aligned}$$

Taking into account the definitions (2.4), (2.5) of $X(0;T)$ and of its norm, this estimate proves the bound (4.2).

To prove estimate (4.3) we choose an integer n between 1 and N and a time t between 0 and t_n and insert $v=(u-u_{h,\tau})(.,t)$ in (4.4). Taking into account (2.7), this gives

$$\begin{aligned}&\frac{1}{2}\frac{d}{dt}\|(u-u_{h,\tau})(.,t)\|_0^2+|||(u-u_{h,\tau})(.,t)|||^2 \\ &\quad\leq(\partial_t(u-u_{h,\tau})(.,t),(u-u_{h,\tau})(.,t))+\varepsilon(\nabla(u-u_{h,\tau})(.,t),\nabla(u-u_{h,\tau})(.,t)) \\ &\qquad+(\underline{a}\cdot\nabla(u-u_{h,\tau})(.,t),(u-u_{h,\tau})(.,t))+(b(u-u_{h,\tau})(.,t),(u-u_{h,\tau})(.,t)) \\ &\quad=\langle R(u_{h,\tau})(.,t),(u-u_{h,\tau})(.,t)\rangle \\ &\quad\leq|||R(u_{h,\tau})(.,t)|||_*|||(u-u_{h,\tau})(.,t)||| \\ &\quad\leq\frac{1}{2}|||R(u_{h,\tau})(.,t)|||_*^2+\frac{1}{2}|||(u-u_{h,\tau})(.,t)|||^2\end{aligned}$$

and thus

$$\frac{d}{dt}\|(u-u_{h,\tau})(.,t)\|_0^2+|||(u-u_{h,\tau})(.,t)|||^2\leq|||R(u_{h,\tau})(.,t)|||_*^2.$$

Integrating this estimate from 0 to t implies

$$\begin{aligned}\|(u-u_{h,\tau})(.,t)\|_0^2-\|u_0-\pi_0u_0\|_0^2+\int_0^t|||(u-u_{h,\tau})(.,s)|||^2\,ds \\ \leq\|R(u_{h,\tau})\|^2_{L^2(0,t;H^1_D(\Omega)^*)} \\ \leq\|R(u_{h,\tau})\|^2_{L^2(0,t_n;H^1_D(\Omega)^*)}.\end{aligned}$$

Since $t\in(0,t_n]$ was arbitrary, this yields

$$\|u-u_{h,\tau}\|^2_{L^\infty(0,t_n;L^2(\Omega))}\leq\|u_0-\pi_0u_0\|_0^2+\|R(u_{h,\tau})\|^2_{L^2(0,t_n;H^1_D(\Omega)^*)}\tag{4.5}$$

and

$$\|u-u_{h,\tau}\|^2_{L^2(0,t_n;H^1_D(\Omega))}\leq\|u_0-\pi_0u_0\|_0^2+\|R(u_{h,\tau})\|^2_{L^2(0,t_n;H^1_D(\Omega)^*)}.\tag{4.6}$$

Equation (4.4) and estimate (2.8), on the other hand, imply

$$|||\partial_t(u-u_{h,\tau}) + \underline{a}\cdot\nabla(u-u_{h,\tau})|||_* \le |||R(u_{h,\tau})|||_* + \max\{1,c_b\}|||u-u_{h,\tau}|||.$$

Taking the square of this inequality, integrating from 0 to t_n, and inserting estimate (4.6) we arrive at

$$\begin{aligned}
&\|\partial_t(u-u_{h,\tau}) + \underline{a}\cdot\nabla(u-u_{h,\tau})\|^2_{L^2(0,t_n;H^1_D(\Omega)^*)}\\
&\quad\le 2\|R(u_{h,\tau})\|^2_{L^2(0,t_n;H^1_D(\Omega)^*)} + 2\max\{1,c_b\}^2\|u-u_{h,\tau}\|^2_{L^2(0,t_n;H^1_D(\Omega))}\\
&\quad\le 2\max\{1,c_b\}^2\|u_0-\pi_0u_0\|^2_0\\
&\qquad + 2(1+\max\{1,c_b\}^2)\|R(u_{h,\tau})\|^2_{L^2(0,t_n;H^1_D(\Omega)^*)}.
\end{aligned}\tag{4.7}$$

Combining estimates (4.5)–(4.7) proves the bound (4.3). □

The subsequent analysis relies on an appropriate decomposition of the residual $R(u_{h,\tau})$. To this end we define a temporal residual $R_\tau(u_{h,\tau}) \in L^2(0,T;H^1_D(\Omega)^*)$ and a spatial residual $R_h(u_{h,\tau}) \in L^2(0,T;H^1_D(\Omega)^*)$ by setting, for all $v \in H^1_D(\Omega)$ and all $1 \le n \le N$,

$$\begin{aligned}
\langle R_\tau(u_{h,\tau}),v\rangle &= \varepsilon(\nabla[\theta u_h^n + (1-\theta)u_h^{n-1} - u_{h,\tau}],\nabla v)\\
&\quad + (\underline{a}^{n\theta}\cdot\nabla[\theta u_h^n + (1-\theta)u_h^{n-1} - u_{h,\tau}],v)\\
&\quad + (b^{n\theta}[\theta u_h^n + (1-\theta)u_h^{n-1} - u_{h,\tau}],v) \qquad \text{on } (t_{n-1},t_n]
\end{aligned}\tag{4.8}$$

and

$$\begin{aligned}
\langle R_h(u_{h,\tau}),v\rangle &= (f^{n\theta},v) + (g^{n\theta},v)_{\Gamma_N} - \left(\frac{u_h^n - u_h^{n-1}}{\tau_n},v\right)\\
&\quad - \varepsilon(\theta\nabla u_h^n + (1-\theta)\nabla u_h^{n-1},\nabla v)\\
&\quad - (\underline{a}^{n\theta}\cdot\nabla[\theta u_h^n + (1-\theta)u_h^{n-1}],v)\\
&\quad - (b^{n\theta}[\theta u_h^n + (1-\theta)u_h^{n-1}],v) \qquad \text{on } (t_{n-1},t_n].
\end{aligned}\tag{4.9}$$

The time discretization of the data is taken into account by a data-residual $R_D(u_{h,\tau}) \in L^2(0,T;H^1_D(\Omega)^*)$ which is defined by

$$\begin{aligned}
\langle R_D(u_{h,\tau}),v\rangle &= (f-f^{n\theta},v) + (g-g^{n\theta},v)_{\Gamma_N}\\
&\quad + ((\underline{a}^{n\theta}-\underline{a})\cdot\nabla u_{h.\tau},v) + ((b^{n\theta}-b)u_{h,\tau},v) \quad \text{on } (t_{n-1},t_n].
\end{aligned}\tag{4.10}$$

From (3.4) we obtain the decomposition

$$R(u_{h,\tau}) = R_D(u_{h,\tau}) + R_\tau(u_{h,\tau}) + R_h(u_{h,\tau}).\tag{4.11}$$

5. Estimation of the spatial residual. The techniques required for the estimation of the spatial residual $R_h(u_{h,\tau})$ are similar to those used in the stationary case [10, sect. 4]. But it should be stressed that we are not interested in estimating the error between $u_{h,\tau}$ and the solution of the variational problem obtained from the temporal semidiscretization of (1.1). Moreover, we have to pay particular attention to the fact that $u_{h,\tau}$ is the linear interpolant of the functions $(u_h^n)_{0\le n\le N}$ that live on different spatial meshes.

We denote by $\widetilde{\mathcal{E}}_{h,n}$, $1 \le n \le N$, the set of all edges (if $d = 2$), respectively, faces (if $d = 3$), of $\widetilde{\mathcal{T}}_{h,n}$. With each edge or face $E \in \widetilde{\mathcal{E}}_{h,n}$ we associate a unit vector $\underline{n}_E$

orthogonal to E such that it points to the outward of Ω if E lies on the boundary. For every edge or face E that is not contained in the boundary Γ we denote by $[.]_E$ the jump across E in direction $\underline{n}_E$. The quantity $[.]_E$ of course depends on the orientation of $\underline{n}_E$, but quantities of the form $[\underline{n}_E \cdot .]_E$ are independent thereof. With each edge, respectively, face, we associate the set ω_E which is the union of the elements that share E.

We denote by $f_{h,\tau}$, $g_{h,\tau}$, $\underline{a}_{h,\tau}$, and $b_{h,\tau}$ functions which are *piecewise constant* on the time-intervals and which, on each interval $(t_{n-1}, t_n]$, equal the L^2-projection of $f^{n\theta}$, $g^{n\theta}$, $\underline{a}^{n\theta}$, and $b^{n\theta}$ respectively onto the space of piecewise constant functions corresponding to $\mathcal{T}_{h,n}$. With this notation we define element residuals R_K, $K \in \widetilde{\mathcal{T}}_{h,n}$, $1 \le n \le N$, by

$$\begin{aligned} R_K = f_{h,\tau} &- \frac{u_h^n - u_h^{n-1}}{\tau_n} + \varepsilon\Delta(\theta u_h^n + (1-\theta)u_h^{n-1}) \\ &- \underline{a}_{h,\tau} \cdot \nabla(\theta u_h^n + (1-\theta)u_h^{n-1}) - b_{h,\tau}(\theta u_h^n + (1-\theta)u_h^{n-1}), \end{aligned} \tag{5.1}$$

edge, respectively, face, residuals R_E, $E \in \widetilde{\mathcal{E}}_{h,n}$, $1 \le n \le N$, by

$$R_E = \begin{cases} -\left[\varepsilon \underline{n}_E \cdot \nabla(\theta u_h^n + (1-\theta)u_h^{n-1})\right]_E & \text{if } E \not\subset \Gamma, \\ g_{h,\tau} - \varepsilon \underline{n}_E \cdot \nabla(\theta u_h^n + (1-\theta)u_h^{n-1}) & \text{if } E \subset \Gamma_N, \\ 0 & \text{if } E \subset \Gamma_D, \end{cases} \tag{5.2}$$

elementwise data errors D_K, $K \in \widetilde{\mathcal{T}}_{h,n}$, $1 \le n \le N$, by

$$\begin{aligned} D_K = \big\{ f^{n\theta} &- f_{h,\tau} + (\underline{a}_{h,\tau} - \underline{a}^{n\theta})\nabla \cdot (\theta u_h^n + (1-\theta)u_h^{n-1}) \\ &+ (b_{h,\tau} - b^{n\theta})(\theta u_h^n + (1-\theta)u_h^{n-1}) \big\}_{|K}, \end{aligned} \tag{5.3}$$

and edge-, respectively, facewise, data errors D_E, $E \in \widetilde{\mathcal{E}}_{h,n} \cap \Gamma_N$, $1 \le n \le N$, by

$$D_E = g^{n\theta} - g_{h,\tau}. \tag{5.4}$$

Here, of course, $(u_h^n)_{0 \le n \le N}$ denotes the solution of problems (3.1) and (3.2).

Note that, as usual, the residuals R_K are defined elementwise. In particular Δu_h^n and Δu_h^{n-1} must be interpreted as the Laplacian applied to the restriction to K of the corresponding functions. Here, we need the transition condition that $\widetilde{\mathcal{T}}_{h,n}$ is a common refinement of $\mathcal{T}_{h,n}$ and $\mathcal{T}_{h,n-1}$.

For every n between 1 and N we denote by $\mathcal{N}_{h,n}$ the set of all element vertices in $\mathcal{T}_{h,n}$ that do not lie on the Dirichlet boundary Γ_D. With every vertex $x \in \mathcal{N}_{h,n}$ we associate the nodal bases function λ_x which is uniquely defined by the properties

$$\lambda_{x|K} \in R_1(K) \quad \forall K \in \mathcal{T}_{h,n}, \quad \lambda_x(y) = 0 \quad \forall y \in \mathcal{N}_{h,n} \backslash \{x\}, \quad \lambda_x(x) = 1.$$

Here, as usual, $R_k(K)$ denotes the set of all polynomials of total degree k, if K is a simplex, and of maximal degree k, if K is a parallelepiped. The support of a nodal basis function λ_x is denoted by ω_x and consists of all elements in $\mathcal{T}_{h,n}$ that share the vertex x. With this notation we can define a Clément-type interpolation operator $I_{h,n} : L^1(\Omega) \longrightarrow \{\varphi \in C(\Omega) : \varphi_{|K} \in R_1(K) \text{ for all } K \in \mathcal{T}_{h,n}, \varphi = 0 \text{ on } \Gamma_D\}$ by (cf. [9])

$$I_{h,n} v = \sum_{x \in \mathcal{N}_{h,n}} \left\{ \frac{1}{|\omega_x|} \int_{\omega_x} v \right\} \lambda_x. \tag{5.5}$$

Here $|\omega_x|$ denotes the d-dimensional Lebesgue-measure of ω_x. Due to condition (6) of section 3 the image of $I_{h,n}$ is contained in $X_{h,n}$.

LEMMA 5.1. *For every $S \in \widetilde{\mathcal{T}}_{h,n} \cup \widetilde{\mathcal{E}}_{h,n}$, $1 \le n \le N$, denote by h_S its diameter and set*

$$\alpha_S = \min\{h_S \varepsilon^{-1/2}, \beta^{-1/2}\}. \tag{5.6}$$

Then the following estimates hold for all n between 1 and N, all elements $K \in \widetilde{\mathcal{T}}_{h,n}$, all edges, respectively, faces, E of K, and all functions $v \in H^1_D(\Omega)$:

$$\begin{aligned}\|v - I_h v\|_{0;K} &\le c_1 \alpha_K |||v|||_{\tilde{\omega}_K},\\ \|v - I_h v\|_{0;E} &\le c_2 \varepsilon^{-1/4} \alpha_E^{1/2} |||v|||_{\tilde{\omega}_K},\\ |||I_h v|||_K &\le c_3 |||v|||_{\tilde{\omega}_K}.\end{aligned}$$

Here, $\tilde{\omega}_K$ is the union of all elements in $\mathcal{T}_{h,n}$ that share at least one vertex with the element $K' \in \mathcal{T}_{h,n}$ that contains K and $|||.|||_A$ denotes the restriction of $|||.|||$ to the measurable set A.

Proof. The proof of Lemma 5.1 follows from Lemma 3.1 in [7] and Proposition 2.1 in [8] with the arguments used in the proof of Lemma 3.2 in [7]. □

Remark 5.2. In the case $\beta = 0$ the minimum in (5.6) of course yields $\alpha_S = \varepsilon^{-1/2} h_S$ for all S.

For every element $K \in \widetilde{\mathcal{T}}_{h,n}$, $1 \le n \le N$, we denote by $\mathcal{N}_K$ the set of its vertices and set

$$\psi_K = \gamma_K \prod_{x \in \mathcal{N}_K} \lambda_x, \tag{5.7}$$

where the constant γ_K is chosen such that ψ_K equals 1 at the barycenter of K. Note that the support of ψ_K is contained in K and that $\|\psi_K\|_{L^\infty(K)} = 1$.

For every edge, respectively, face, $E \in \widetilde{\mathcal{E}}_{h,n}$, $1 \le n \le N$, we set

$$\theta_E = \min\{\varepsilon^{1/2} \beta^{-1/2} h_E^{-1}, 1\} \tag{5.8}$$

and denote by $\mathcal{N}_E$ the set of its vertices. (Note that $\theta_E = 1$ in the case $\beta = 0$.) Consider first a face E that is not contained in the boundary. It is shared by exactly two elements $K_{E,1}$ and $K_{E,2}$. For $i = 1, 2$ we define an affine transformation $F_i : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ as follows: We first map $K_{E,i}$ onto the reference element such that the image of E is contained in the hyperplane $\{x_d = 0\}$; then we apply the transformation $(x_1, \dots, x_{d-1}, x_d) \longrightarrow (x_1, \dots, x_{d-1}, \theta_E x_d)$; and finally we transform back using the inverse of the affine transformation of the first step. With this definition we set

$$\psi_E = \gamma_E \prod_{x \in \mathcal{N}_E} \lambda_x \circ F_i^{-1} \quad \text{on } K_{E,i},\ i = 1, 2, \tag{5.9}$$

where the constant γ_E is chosen such that ψ_E equals 1 at the barycenter of E. Note that the support of ψ_E is contained in $F_1(K_{E,1}) \cup F_2(K_{E,2}) \subset K_{E,1} \cup K_{E,2} = \omega_E$ and that $\|\psi_E\|_{L^\infty(E)} = 1$.

If an edge, respectively, face, E is contained in the Neumann boundary Γ_N, the definition of ψ_E is modified in the obvious way, taking into account that now E is the face of exactly one element K_E.

LEMMA 5.3. *The following estimates hold for all n between 1 and N, all elements $K \in \widetilde{\mathcal{T}}_{h,n}$, all polynomials $v \in R_k(K)$, all edges, respectively, faces, $E \in \widetilde{\mathcal{E}}_{h,n}$, and all polynomials $\sigma \in R_k(E)$:*

$$\begin{aligned}(v, \psi_K v)_K &\geq c_4 \|v\|_{0;K}^2,\\ |||\psi_K v|||_K &\leq c_5 \alpha_K^{-1} \|v\|_{0;K},\\ (\sigma, \psi_E \sigma)_E &\geq c_6 \|\sigma\|_{0;E}^2,\\ |||\psi_E \sigma|||_{\omega_E} &\leq c_7 \varepsilon^{1/4} \alpha_E^{-1/2} \|\sigma\|_{0;E},\\ \|\psi_E \sigma\|_{0;\omega_E} &\leq c_8 \varepsilon^{1/4} \alpha_E^{1/2} \|\sigma\|_{0;E}.\end{aligned}$$

Here, a polynomial σ defined on an edge, respectively, face, E is continued in the canonical way to a polynomial defined on $\mathbb{R}^d$. The constants $c_4, \dots, c_8$ depend only on the polynomial degree k in condition (5) *of section* 3 *and on the ratios h_K/ρ_K in condition* (3).

Proof. The estimates are proven with the same arguments as in the proof of Lemma 3.3 in [7]. For parallelepipeds one only has to take into account that the transformation to the unit cube is affine and thus has a constant Jacobian. □

With these preparations we are now ready to bound the spatial residual.

LEMMA 5.4. *For every n between 1 and N define a spatial error indicator η_h^n by*

$$\eta_h^n = \left\{ \sum_{K \in \widetilde{\mathcal{T}}_{h,n}} \alpha_K^2 \|R_K\|_{L^2(K)}^2 + \sum_{E \in \widetilde{\mathcal{E}}_{h,n}} \varepsilon^{-1/2} \alpha_E \|R_E\|_{L^2(E)}^2 \right\}^{1/2} \tag{5.10}$$

and a spatial data error indicator Θ_h^n by

$$\Theta_h^n = \left\{ \sum_{K \in \widetilde{\mathcal{T}}_{h,n}} \alpha_K^2 \|D_K\|_{L^2(K)}^2 + \sum_{E \in \widetilde{\mathcal{E}}_{h,n} \cap \Gamma_N} \varepsilon^{-1/2} \alpha_E \|D_E\|_{L^2(E)}^2 \right\}^{1/2}. \tag{5.11}$$

Then there are functions $w_n \in H_D^1(\Omega)$, $1 \leq n \leq N$, and constants $c^\dagger$ and $c_\dagger$ such that on each interval $(t_{n-1}, t_n]$, $1 \leq n \leq N$, the following estimates hold:

$$|||R_h(u_{h,\tau})|||_* \leq c^\dagger \{\eta_h^n + \Theta_h^n\} \tag{5.12}$$

and

$$\begin{aligned}(\eta_h^n)^2 &\leq \langle R_h(u_{h,\tau}), w_n \rangle + \Theta_h^n\, \eta_h^n,\\ |||w_n||| &\leq c_\dagger \eta_h^n.\end{aligned} \tag{5.13}$$

The constants $c^\dagger$ and $c_\dagger$ depend on the ratios h_K/ρ_K in condition (3) *of section* 3*. The constant $c^\dagger$ in addition depends on the ratios $h_{K'}/h_K$ in condition* (4)*. The constant $c_\dagger$ in addition depends on the maximum of the polynomial degrees of the finite element functions.*

Proof. Choose an integer n between 1 and N and keep it fixed in what follows.

Integration by parts on the elements in $\widetilde{\mathcal{T}}_{h,n}$ yields the following L^2-representation of the spatial residual:

$$\begin{aligned}\langle R_h(u_{h,\tau}), v \rangle = &\sum_{K \in \widetilde{\mathcal{T}}_{h,n}} (R_K, v)_K + \sum_{E \in \widetilde{\mathcal{E}}_{h,n}} (R_E, v)_E\\ &+ \sum_{K \in \widetilde{\mathcal{T}}_{h,n}} (D_K, v)_K + \sum_{E \in \widetilde{\mathcal{E}}_{h,n} \cap \Gamma_n} (D_E, v)_E.\end{aligned} \tag{5.14}$$

Lemma 5.1 and the Cauchy–Schwarz inequality therefore imply, for all $v \in H^1_D(\Omega)$,

$$\begin{aligned} \langle R_h(u_{h,\tau}), & v - I_{h,n}v\rangle \\ &\le c|||v|||\Bigg\{ \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2 \|R_K\|^2_{0;K} + \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} \varepsilon^{-1/2}\alpha_E\|R_E\|^2_{0;E} \\ &\qquad + \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2\|D_K\|^2_{0;K} + \sum_{E\in\widetilde{\mathcal{E}}_{h,n}\cap\Gamma_n} \varepsilon^{-1/2}\alpha_E\|D_E\|^2_{0;E}\Bigg\}^{1/2}. \end{aligned} \tag{5.15}$$

The constant c depends only on the constants c_1 and c_2 of Lemma 5.1 and on the ratios h_K/ρ_K.

From the definition of problem (3.2) and definition (4.9) of the spatial residual we conclude that

$$\langle R_h(u_{h,\tau}), I_{h,n}v\rangle = \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \delta_K\Big\{(R_K, \underline{a}^{n\theta}\cdot\nabla I_{h.n}v)_K + (D_K, \underline{a}^{n\theta}\cdot\nabla I_{h,n}v)_K\Big\}.$$

Lemma 5.1, condition (3.3), and the Cauchy–Schwarz inequality therefore imply

$$\langle R_h(u_{h,\tau}), I_{h,n}v\rangle \le c|||v|||\Bigg\{\sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2\Big\{\|R_K\|^2_{0;K} + \|D_K\|^2_{0;K}\Big\}\Bigg\}^{1/2}. \tag{5.16}$$

Equation (5.14) and estimates (5.15) and (5.16) prove the upper bound (5.12).

For the proof of the lower bound (5.13) we proceed as in the proof of [9, Lem. 5.1] and define the function w_n by

$$w_n = \gamma_1 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2\psi_K R_K + \gamma_2 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} \varepsilon^{-1/2}\alpha_E\psi_E R_E. \tag{5.17}$$

The constants γ_1 and γ_2 are arbitrary at present and will be determined below. The subsequent arguments are based on the following observations:

- The supports of the ψ_K are mutually disjoint.
- The support of a ψ_K intersects the support of at most $2d$ different ψ_E's.
- The support of a ψ_E intersects the support of at most two ψ_K's.
- The support of a ψ_E intersects the support of at most $2d-2$ other ψ_E's.

Lemma 5.3 therefore yields

$$\begin{aligned} |||w_h|||^2 \le\; & \gamma_1^2 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^4 |||\psi_K R_K|||^2_K \\ &+ 2\gamma_1\gamma_2 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \Bigg\{\sum_{E;\omega_E\cap K\neq\emptyset} \alpha_K^2\varepsilon^{-1/2}\alpha_E |||\psi_K R_K|||_K |||\psi_E R_E|||_K\Bigg\} \\ &+ \gamma_2^2 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} \Bigg\{\sum_{E';\omega_E\cap\omega_{E'}\neq\emptyset} \varepsilon^{-1}\alpha_E\alpha_{E'} |||\psi_E R_E|||_{\omega_E} |||\psi_{E'}R_{E'}|||_{\omega_{E'}}\Bigg\} \\ \le\; & (2d+1)\max\{\gamma_1^2,\gamma_2^2\}\max\{c_5,c_7\}\left(\eta_h^n\right)^2. \end{aligned} \tag{5.18}$$

Since $h_E \le h_K$ for all edges, respectively, faces, E of any element K, Lemma 5.3 also implies that

$$\begin{aligned}
&\sum_{K\in\widetilde{\mathcal{T}}_{h,n}} (R_K,w_n)_K + \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} (R_E,w_n)_E \\
&\qquad = \gamma_1 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2 (R_K,\psi_K R_K)_K + \gamma_2 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} \varepsilon^{-1/2}\alpha_E (R_E,\psi_E R_E)_E \\
&\qquad\quad + \gamma_2 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} \left\{ \sum_{K;K\cap\omega_E\neq\emptyset} \varepsilon^{-1/2}\alpha_E (R_K,\psi_E R_E)_K \right\} \\
&\qquad \ge \gamma_1 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} c_4 \alpha_K^2 \|R_K\|_{0;K}^2 + \gamma_2 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} c_6 \varepsilon^{-1/2}\alpha_E \|R_E\|_{0;E}^2 \\
&\qquad\quad - \gamma_2 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} \left\{ \sum_{K;K\cap\omega_E\neq\emptyset} c_8 \varepsilon^{-1/4}\alpha_E^{1/2}\alpha_K \|R_K\|_{0;K} \|\psi_E R_E\|_{0;E} \right\} \\
&\qquad \ge (\gamma_1 c_4 - 2d\gamma_2 c_8^2 c_6^{-1}) \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2 \|R_K\|_{0;K}^2 \\
&\qquad\quad + \frac{1}{2}\gamma_2 c_6 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}} \varepsilon^{-1/2}\alpha_E \|R_E\|_{0;E}^2 \\
&\qquad \ge \min\left\{ \gamma_1 c_4 - 2d\gamma_2 c_8^2 c_6^{-1}, \frac{1}{2}\gamma_2 c_6 \right\} (\eta_h^n)^2 .
\end{aligned} \tag{5.19}$$

From Lemma 5.3 we also obtain

$$\begin{aligned}
&\sum_{K\in\widetilde{\mathcal{T}}_{h,n}} (D_K,w_h)_K + \sum_{E\in\widetilde{\mathcal{E}}_{h,n}\cap\Gamma_n} (D_E,w_h)_E \\
&\qquad = \gamma_1 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2 (D_K,\psi_K R_K)_K \\
&\qquad\quad + \gamma_2 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \left\{ \sum_{E;E\subset\partial K} \varepsilon^{-1/2}\alpha_E (D_K,\psi_E R_E)_K \right\} \\
&\qquad\quad + \gamma_2 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}\cap\Gamma_n} \varepsilon^{-1/2}\alpha_E (D_E,\psi_E R_E)_E \\
&\qquad \le \gamma_1 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2 \|R_K\|_{0;K} \|D_K\|_{0;K} \\
&\qquad\quad + \gamma_2 \sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \left\{ \sum_{E;E\subset\partial K} c_8 \varepsilon^{-1/4}\alpha_E^{3/2} \|R_E\|_{0;E} \|D_K\|_{0;K} \right\} \\
&\qquad\quad + \gamma_2 \sum_{E\in\widetilde{\mathcal{E}}_{h,n}\cap\Gamma_n} \varepsilon^{-1/2}\alpha_E \|R_E\|_{0;E} \|D_E\|_{0;E} \\
&\qquad \le 2d \max\{\gamma_1,\gamma_2\} \max\{1,c_8\} \Theta_h^n \eta_h^n .
\end{aligned} \tag{5.20}$$

Now we choose

$$\gamma_2 = \frac{2}{c_6} \quad \text{and} \quad \gamma_1 = \frac{1}{c_4}\left(1 + \frac{4dc_8^2}{c_6^2}\right).$$

This choice gives

$$\min\left\{\gamma_1 c_4 - 2d\gamma_2 c_8^2 c_6^{-1}, \frac{1}{2}\gamma_2 c_6\right\} = 1.$$

Estimates (5.19), (5.20), and (5.18) and equation (5.14) now imply the lower bound (5.13). □

6. Estimation of the temporal residual. The following lemma provides us with sharp upper and lower bounds for the temporal residual.

LEMMA 6.1. *For every integer n between 1 and N and every real number δ larger than 0 and less than 1 there is a function $z_{n,\delta} \in L^2(0,T;H_D^1(\Omega))$ such that the following estimates hold on each interval $(t_{n-1}, t_n]$:*

$$\tag{6.1} \begin{aligned} &\left\{\int_{t_{n-1}}^{t_n} |||R_\tau(u_{h,\tau})(.,s)|||_*^2 ds\right\}^{1/2} \\ &\qquad \leq \sqrt{\frac{2}{3}}\max\{c_b, 1\}\tau_n^{1/2}\left\{|||u_h^n - u_h^{n-1}|||^2 + |||\underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1})|||_*^2\right\}^{1/2} \end{aligned}$$

and

$$\tag{6.2} \begin{aligned} &\int_{t_{n-1}}^{t_n} \langle R_\tau(u_{h,\tau})(.,s), z_{n,\delta}(.,s)\rangle ds \\ &\quad \geq \frac{\delta}{12(\delta + \max\{c_b, 1\}^2)}\tau_n\left\{|||u_h^n - u_h^{n-1}|||^2 + \delta|||\underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1})|||_*^2\right\}, \\ &\left\{\int_{t_{n-1}}^{t_n} |||z_{n,\delta}(.,s)|||^2 ds\right\}^{1/2} \\ &\quad \leq \sqrt{\frac{8}{3}}\tau_n^{1/2}\left\{|||u_h^n - u_h^{n-1}|||^2 + |||\underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1})|||_*^2\right\}^{1/2}. \end{aligned}$$

Proof. Since the function $t \to u_{h,\tau}(.,t)$ is continuous and piecewise affine with values in $H_D^1(\Omega)$, we have on each time interval $[t_{m-1}, t_m]$

$$\theta u_h^m + (1-\theta)u_h^{m-1} - u_{h,\tau} = \left[\theta - \frac{t - t_{m-1}}{\tau_m}\right](u_h^m - u_h^{m-1}).$$

For abbreviation we define for each m between 1 and N the quantity $r_m \in H_D^1(\Omega)^*$ by

$$\begin{aligned} \langle r_m, v\rangle = {}& \varepsilon(\nabla(u_h^m - u_h^{m-1}), \nabla v) + (\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1}), v) \\ &+ (b^{m\theta}(u_h^m - u_h^{m-1}), v) \qquad \forall v \in H_D^1(\Omega). \end{aligned}$$

Then we obtain the following representation of the temporal residual

$$\tag{6.3} R_\tau(u_{h,\tau}) = \left[\theta - \frac{t - t_{m-1}}{\tau_m}\right] r_m \qquad \text{on } (t_{m-1}, t_m],\ 1 \leq m \leq N.$$

A straightforward calculation gives

$$\int_{t_{m-1}}^{t_m} \left[\theta - \frac{t-t_{m-1}}{\tau_m}\right]^2 dt = \tau_m \frac{1}{3}\left[\theta^3 + (1-\theta)^3\right]$$

and consequently

$$\frac{1}{12}\tau_m \le \int_{t_{m-1}}^{t_m} \left[\theta - \frac{t-t_{m-1}}{\tau_m}\right]^2 dt \le \frac{1}{3}\tau_m. \tag{6.4}$$

From (2.8) we conclude that

$$\begin{aligned} |||r_m|||_* &\le \max\{c_b,1\}|||u_h^m - u_h^{m-1}||| + |||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_* \\ &\le \sqrt{2}\max\{c_b,1\}\left\{|||u_h^m - u_h^{m-1}|||^2 + |||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_*^2\right\}^{1/2}. \end{aligned} \tag{6.5}$$

Inequalities (6.4) and (6.5) prove the upper bound (6.1).

Due to the definition (2.3) of $|||.|||_*$ there is for each $\delta \in (0,1)$ a function $\varphi_{m,\delta} \in H_D^1(\Omega)$ with

$$\begin{aligned} |||\varphi_{m,\delta}||| &= |||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_*, \\ (\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1}), \varphi_{m,\delta}) &\ge \delta|||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_*^2. \end{aligned}$$

We set

$$\zeta_{m,\delta} = (u_h^m - u_h^{m-1}) + \gamma\varphi_{m,\delta}, \tag{6.6}$$

where γ is a constant that will be fixed below. Obviously we have

$$|||\zeta_{m,\delta}||| \le \max\{1,\gamma\}\left\{|||u_h^m - u_h^{m-1}||| + |||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_*\right\}.$$

Inequalities (2.7) and (2.8) on the other hand yield

$$\begin{aligned} \langle r_m, \zeta_{m,\delta}\rangle &\ge |||u_h^m - u_h^{m-1}|||^2 + \gamma\delta|||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_*^2 \\ &\quad - \gamma\max\{c_b,1\}|||u_h^m - u_h^{m-1}|||\ |||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_* \\ &\ge \left\{1 - \frac{1}{2}\gamma\delta^{-1}\max\{c_b,1\}^2\right\}|||u_h^m - u_h^{m-1}|||^2 \\ &\quad + \frac{1}{2}\gamma\delta|||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_*^2. \end{aligned}$$

Now we choose

$$\gamma = \frac{2\delta}{\delta + \max\{c_b,1\}^2}$$

and obtain

$$\begin{aligned} |||\zeta_{m,\delta}||| &\le 2\left\{|||u_h^m - u_h^{m-1}||| + |||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_*\right\}, \\ \langle r_m, \zeta_{m,\delta}\rangle &\ge \frac{\delta}{\delta + \max\{c_b,1\}^2}\Big\{|||u_h^m - u_h^{m-1}|||^2 \\ &\qquad + \delta|||\underline{a}^{m\theta}\cdot\nabla(u_h^m - u_h^{m-1})|||_*^2\Big\}. \end{aligned} \tag{6.7}$$

Equation (6.3) and estimates (6.4), (6.7) show that the function

$$z_{m,\delta} = \left[\theta - \frac{t-t_{m-1}}{\tau_m}\right]\zeta_{m,\delta}$$

yields the lower bounds (6.2). □

7. A preliminary a posteriori error estimate. The following lemma provides us with a posteriori error bounds which are robust in the sense described in the introduction. However, they are not suited for practical computations since they involve terms of the form $|||\underline{a}^{n\theta} \cdot \nabla(u_h^n - u_h^{n-1})|||_*$. In the next section we will bound these terms by computable quantities. Recall that $H_D^1(\Omega)^*$ is equipped with $|||.|||_*$.

LEMMA 7.1. *The error between the solution u of problem* (2.6) *and the solution $u_{h,\tau}$ of problems* (3.1), (3.2) *is bounded from above by*

$$\begin{aligned}
&\|u - u_{h,\tau}\|_{X(0,T)} \\
&\quad \leq c^* \Bigg\{ \|u_0 - \pi_0 u_0\|_0^2 \\
&\qquad + \sum_{n=1}^{N} \tau_n \Big[(\eta_h^n)^2 + |||u_h^n - u_h^{n-1}|||^2 + |||\underline{a}^{n\theta} \cdot \nabla(u_h^n - u_h^{n-1})|||_*^2 \Big] \\
&\qquad + \sum_{n=1}^{N} \tau_n \, (\Theta_h^n)^2 \\
&\qquad + \big\| f - f^{n\theta} - (\underline{a} - \underline{a}^{n\theta}) \cdot \nabla u_{h,\tau} - (b - b^{n\theta}) u_{h,\tau} \big\|_{L^2(0,T;H_D^1(\Omega)^*)}^2 \\
&\qquad + \big\| g - g_{h,\tau} \big\|_{L^2(0,T;H^{-1/2}(\Gamma_N))}^2 \Bigg\}^{1/2}
\end{aligned} \tag{7.1}$$

and on each interval $(t_{n-1}, t_n]$, $1 \leq n \leq N$, from below by

$$\begin{aligned}
&\tau_n^{1/2} \Big\{ (\eta_h^n)^2 + |||u_h^n - u_h^{n-1}|||^2 + |||\underline{a}^{n\theta} \cdot \nabla(u_h^n - u_h^{n-1})|||_*^2 \Big\}^{1/2} \\
&\leq c_* \Bigg\{ \|u - u_{h,\tau}\|_{X(t_{n-1},t_n)}^2 \\
&\qquad + \tau_n \, (\Theta_h^n)^2 \\
&\qquad + \big\| f - f^{n\theta} - (\underline{a} - \underline{a}^{n\theta}) \cdot \nabla u_{h,\tau} - (b - b^{n\theta}) u_{h,\tau} \big\|_{L^2(t_{n-1},t_n;H_D^1(\Omega)^*)}^2 \\
&\qquad + \big\| g - g_{h,\tau} \big\|_{L^2(t_{n-1},t_n;H^{-1/2}(\Gamma_N))}^2 \Bigg\}^{1/2}.
\end{aligned} \tag{7.2}$$

The quantities η_h^n and Θ_h^n are defined in (5.10) *and* (5.11), *respectively. The constants c^* and c_* depend on the ratios h_K/ρ_K. The constant c^* in addition depends on the ratios $h_{K'}/h_K$. The constant c_* in addition depends on the maximum of the polynomial degrees of the finite element functions. All constants are independent of the final time T, the viscosity ε, and the parameter β.*

Proof. The upper bound (7.1) follows from estimates (4.3), (5.12), and (6.1) and the decomposition (4.11) of the residual.

For the proof of the lower bound (7.2) we choose an integer n between 1 and N and a real number δ larger than 0 and less than 1.

First we insert the function $z_{n,\delta}$ of Lemma 6.1 into the representation (4.11) of the residual. Estimates (6.2), (5.12), and (4.2) then imply

$$\begin{aligned}
&\frac{\delta}{12(\delta+\max\{c_b,1\}^2)}\tau_n\Big\{|||u_h^n-u_h^{n-1}|||^2+\delta|||\underline{a}^{n\theta}\cdot\nabla(u_h^n-u_h^{n-1})|||_*^2\Big\}\\
&\quad\le\int_{t_{n-1}}^{t_n}\langle R_\tau(u_{h,\tau})(.,s),z_{n,\delta}(.,s)\rangle ds\\
&\quad=\int_{t_{n-1}}^{t_n}\langle R(u_{h,\tau})(.,s)-R_D(u_{h,\tau})(.,s)-R_h(u_{h,\tau})(.,s),z_{n,\delta}(.,s)\rangle ds
\end{aligned}$$

and

$$\begin{aligned}
&\int_{t_{n-1}}^{t_n}\langle R(u_{h,\tau})(.,s)-R_D(u_{h,\tau})(.,s)-R_h(u_{h,\tau})(.,s),z_{n,\delta}(.,s)\rangle ds\\
&\quad\le\sqrt{\frac{8}{3}}\tau_n^{1/2}\left\{|||u_h^n-u_h^{n-1}|||^2+|||\underline{a}^{n\theta}\cdot\nabla(u_h^n-u_h^{n-1})|||_*^2\right\}^{1/2}\\
&\quad\cdot\bigg\{2\max\{c_b,1\}^2\|u-u_{h,\tau}\|^2_{X(t_{n-1},t_n)}\\
&\qquad+\left\|f-f^{n\theta}-(\underline{a}-\underline{a}^{n\theta})\cdot\nabla u_{h,\tau}-(b-b^{n\theta})u_{h,\tau}\right\|^2_{L^2(t_{n-1},t_n;H^1_D(\Omega)^*)}\\
&\qquad+\left\|g-g_{h,\tau}\right\|^2_{L^2(t_{n-1},t_n;H^{-1/2}(\Gamma_N))}\\
&\qquad+c^\dagger\tau_n\left(\Theta_h^n\right)^2+c^\dagger\tau_n\left(\eta_h^n\right)^2\bigg\}^{1/2}.
\end{aligned}$$

Since $\delta\in(0,1)$ was arbitrary and since $\sqrt{\frac{8}{3}}\le 2$ this yields the estimate

$$\begin{aligned}
&\tau_n^{1/2}\Big\{|||u_h^n-u_h^{n-1}|||^2+|||\underline{a}^{n\theta}\cdot\nabla(u_h^n-u_h^{n-1})|||_*^2\Big\}^{1/2}\\
&\quad\le c'\Big\{2\max\{c_b,1\}^2\|u-u_{h,\tau}\|^2_{X(t_{n-1},t_n)}+c^\dagger\tau_n\left(\Theta_h^n\right)^2\\
&\qquad+\left\|f-f^{n\theta}-(\underline{a}-\underline{a}^{n\theta})\cdot\nabla u_{h,\tau}-(b-b^{n\theta})u_{h,\tau}\right\|^2_{L^2(t_{n-1},t_n;H^1_D(\Omega)^*)}\\
&\qquad+\left\|g-g_{h,\tau}\right\|^2_{L^2(t_{n-1},t_n;H^{-1/2}(\Gamma_N))}\\
&\qquad+c^\dagger\tau_n\left(\eta_h^n\right)^2\Big\}^{1/2}
\end{aligned}\tag{7.3}$$

with $c'=24(1+\max\{c_b,1\}^2)$.

Next we insert the function $(\alpha+1)(\frac{t-t_{n-1}}{\tau_n})^\alpha w_n$ into the representation (4.11) of the residual. Here w_n is the function of Lemma 5.4 and α denotes a nonnegative constant that will be determined below. Estimate (5.13) and the decomposition (4.11) of the residual then yield

$$\begin{aligned}
\tau_n\left(\eta_h^n\right)^2&\le\int_{t_{n-1}}^{t_n}(\alpha+1)\left(\frac{t-t_{n-1}}{\tau_n}\right)^\alpha\langle R_h(u_{h,\tau}),w_n\rangle dt+\tau_n\Theta_h^n\,\eta_h^n\\
&=\int_{t_{n-1}}^{t_n}(\alpha+1)\left(\frac{t-t_{n-1}}{\tau_n}\right)^\alpha\langle R(u_{h,\tau})-R_D(u_{h,\tau})-R_\tau(u_{h,\tau}),w_n\rangle dt\\
&\quad+\tau_n\Theta_h^n\,\eta_h^n.
\end{aligned}$$

Since

$$\int_{t_{n-1}}^{t_n} (\alpha+1)^2 \left(\frac{t-t_{n-1}}{\tau_n}\right)^{2\alpha} dt = \frac{(\alpha+1)^2}{2\alpha+1}\tau_n \le (2\alpha+1)\tau_n$$

and $\sqrt{\frac{2}{3}} \le 1$, estimates (4.2), (5.13), and (6.1) imply that

$$\begin{aligned}
&\int_{t_{n-1}}^{t_n} (\alpha+1) \left(\frac{t-t_{n-1}}{\tau_n}\right)^{\alpha} \langle R(u_{h,\tau}) - R_D(u_{h,\tau}), w_n\rangle dt \\
&\quad \le \sqrt{2\alpha+1}\, c_\dagger \tau_n^{1/2} \eta_h^n \Big\{ \sqrt{2}\max\{c_b, 1\} \|u - u_{h,\tau}\|_{X(t_{n-1},t_n)} \\
&\qquad + \left\| f - f^{n\theta} - (\underline{a} - \underline{a}^{n\theta})\cdot\nabla u_{h,\tau} - (b - b^{n\theta})u_{h,\tau}\right\|_{L^2(t_{n-1},t_n;H_D^1(\Omega)^*)} \\
&\qquad + \left\| g - g_{h,\tau}\right\|^2_{L^2(t_{n-1},t_n;H^{-1/2}(\Gamma_N))} \Big\}.
\end{aligned}$$

Since

$$\int_{t_{n-1}}^{t_n} (\alpha+1) \left(\frac{t-t_{n-1}}{\tau_n}\right)^{\alpha} \left[\theta - \frac{t-t_{n-1}}{\tau_n}\right] dt = \left(\theta - \frac{\alpha+1}{\alpha+2}\right)\tau_n$$

and $\sqrt{\frac{2}{3}} \le 1$, we conclude from estimates (5.13) and (6.1) that

$$\begin{aligned}
&\int_{t_{n-1}}^{t_n} (\alpha+1) \left(\frac{t-t_{n-1}}{\tau_n}\right)^{\alpha} \langle R_\tau(u_{h,\tau}), w_n\rangle dt \\
&\quad \le \left|\theta - \frac{\alpha+1}{\alpha+2}\right| \max\{c_b, 1\} c_\dagger \eta_h^n \tau_n \Big\{ |\!|\!| u_h^n - u_h^{n-1} |\!|\!| + |\!|\!| \underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1}) |\!|\!|_* \Big\}.
\end{aligned}$$

Combining these estimates and inserting (7.3) we arrive at the estimate

$$\begin{aligned}
&\tau_n \left(\eta_h^n\right)^2 \\
&\quad \le \left|\theta - \frac{\alpha+1}{\alpha+2}\right| c_\dagger c'' \tau_n \left(\eta_h^n\right)^2 \\
&\qquad + \tau_n^{1/2} \eta_h^n c_\dagger c''' \left[\sqrt{2\alpha+1} + \left|\theta - \frac{\alpha+1}{\alpha+2}\right|\right] \\
&\qquad\quad \cdot \Big\{ \|u - u_{h,\tau}\|_{X(t_{n-1},t_n)} + \tau_n^{1/2}\Theta_h^n \\
&\qquad\qquad + \left\| f - f^{n\theta} - (\underline{a} - \underline{a}^{n\theta})\cdot\nabla u_{h,\tau} - (b - b^{n\theta})u_{h,\tau}\right\|_{L^2(t_{n-1},t_n;H_D^1(\Omega)^*)} \\
&\qquad\qquad + \left\| g - g_{h,\tau}\right\|^2_{L^2(t_{n-1},t_n;H^{-1/2}(\Gamma_N))} \Big\}
\end{aligned} \tag{7.4}$$

with constants c'' and c''' that depend only on the constant c_b of assumption (A3).

Now we choose the parameter α such that the first term on the right-hand side of (7.4) is balanced by the term on the left-hand side. In case of the Crank–Nicolson scheme, i.e., $\theta = \frac{1}{2}$, this is obvious: We have to choose $\alpha = 0$. In the remaining cases $\frac{1}{2} < \theta \le 1$ we set

$$\alpha = \frac{2c_\dagger c''(2\theta - 1)}{2c_\dagger c''(1-\theta) + 1}.$$

Since we may assume that $c_\dagger c'' \geq 1$ this implies

$$\frac{\alpha+1}{\alpha+2} \leq \theta \quad \text{and} \quad \left|\theta - \frac{\alpha+1}{\alpha+2}\right| c_\dagger c'' \leq \frac{1}{2}.$$

Estimate (7.4) therefore takes the form

$$\begin{aligned} &\tau_n^{1/2}\eta_h^n \\ &\quad \leq c\Big\{ \|u - u_{h,\tau}\|_{X(t_{n-1},t_n)} + \tau_n^{1/2}\Theta_h^n \\ &\qquad + \big\|f - f^{n\theta} - (\underline{a} - \underline{a}^{n\theta})\cdot\nabla u_{h,\tau} - (b - b^{n\theta})u_{h,\tau}\big\|_{L^2(t_{n-1},t_n;H_D^1(\Omega)^*)} \\ &\qquad + \big\|g - g_{h,\tau}\big\|^2_{L^2(t_{n-1},t_n;H^{-1/2}(\Gamma_N))}\Big\} \end{aligned} \tag{7.5}$$

with a constant c that depends only on the constants c_b and $c_\dagger$. Estimates (7.3) and (7.5) obviously imply the lower bound (7.2). □

8. A robust a posteriori error estimator. In this section we derive computable and robust bounds for the terms $|||\underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1})|||_*$ in Lemma 7.1. Standard approaches bound this term by inverse estimates eventually combined with integration by parts. These approaches, however, lead to estimates which incorporate a factor $\varepsilon^{-1/2}$ and which are not robust.

The idea which leads to robust estimates is as follows: Due to the definition of the dual norm, these quantities equal the energy norm of the weak solutions of suitable stationary reaction-diffusion equations. These solutions are approximated by suitable finite element functions. The error of the approximations is estimated by robust error estimators for reaction-diffusion equations.

LEMMA 8.1. *For every integer n between 1 and N set*

$$\widetilde{X}_{h,n} = \{v \in C(\Omega) : v_{|K} \in R_1(K)\ \forall K \in \widetilde{\mathcal{T}}_{h,n},\ v = 0 \text{ on } \Gamma_D\}$$

and denote by $\widetilde{u}_h^n \in \widetilde{X}_{h,n}$ the unique solution of the discrete reaction-diffusion problem

$$\varepsilon(\nabla\widetilde{u}_h^n, \nabla v_h) + \beta(\widetilde{u}_h^n, v_h) = (\underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1}), v_h) \quad \forall v_h \in \widetilde{X}_{h,n}. \tag{8.1}$$

Define the error indicator $\widetilde{\eta}_h^n$ by

$$\begin{aligned} \widetilde{\eta}_h^n = \Bigg\{ &\sum_{K\in\widetilde{\mathcal{T}}_{h,n}} \alpha_K^2 \|\underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1}) + \varepsilon\Delta\widetilde{u}_h^n - \beta\widetilde{u}_h^n\|_{0;K}^2 \\ &+ \sum_{E\in\widetilde{\mathcal{E}}_{h,n}\setminus\Gamma_D} \varepsilon^{-1/2}\alpha_E \|[\underline{n}_E\cdot\nabla\widetilde{u}_h^n]_E\|_{0;E}^2 \Bigg\}^{1/2}. \end{aligned} \tag{8.2}$$

Then there are two constants $\widetilde{c}_\dagger$ and $\widetilde{c}^\dagger$ which depend only on the ratios h_K/ρ_K such that the following estimates are valid

$$\widetilde{c}_\dagger\{|||\widetilde{u}_h^n||| + \widetilde{\eta}_h^n\} \leq |||\underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1})|||_* \leq \widetilde{c}^\dagger\{|||\widetilde{u}_h^n||| + \widetilde{\eta}_h^n\}. \tag{8.3}$$

Proof. We choose an integer n between 1 and N and keep it fixed in what follows. Denote by $\widetilde{U}^n \in H_D^1(\Omega)$ the unique solution of the stationary reaction-diffusion equation

$$\varepsilon(\nabla\widetilde{U}^n, \nabla v) + \beta(\widetilde{U}^n, v) = (\underline{a}^{n\theta}\cdot\nabla(u_h^n - u_h^{n-1}), v) \quad \forall v \in H_D^1(\Omega).$$

The definitions (2.2) and (2.3) of the energy norm $|||.|||$ and of the dual norm $|||.|||_*$, respectively, imply that

$$|||\widetilde{U}^n||| = |||\underline{a}^{n\theta} \cdot \nabla(u_h^n - u_h^{n-1})|||_*.$$

Inserting $v_h = \widetilde{u}_h^n$ as a test function in the discrete problem (8.1) we obtain

$$|||\widetilde{u}_h^n||| \le |||\underline{a}^{n\theta} \cdot \nabla(u_h^n - u_h^{n-1})|||_*.$$

The triangle inequality therefore yields

$$\frac{1}{3}\Big\{|||\widetilde{u}_h^n||| + |||\widetilde{U}^n - \widetilde{u}_h^n|||\Big\} \le |||\underline{a}^{n\theta} \cdot \nabla(u_h^n - u_h^{n-1})|||_* \le \Big\{|||\widetilde{u}_h^n||| + |||\widetilde{U}^n - \widetilde{u}_h^n|||\Big\}.$$

Since $\underline{a}^{n\theta} \cdot \nabla(u_h^n - u_h^{n-1})$ is a piecewise polynomial we know from [7] that $\widetilde{\eta}_h^n$ yields upper and lower bounds for $|||\widetilde{U}^n - \widetilde{u}_h^n|||$ with multiplicative constants that depend only on the ratios h_K/ρ_K. This proves estimate (8.3). □

Combining Lemmas 7.1 and 8.1 we obtain our final result.

THEOREM 8.2. *The error between the solution* u *of problem* (2.6) *and the solution* $u_{h,\tau}$ *of problems* (3.1), (3.2) *is bounded from above by*

(8.4)
$$\begin{aligned}
&\|u - u_{h,\tau}\|_{X(0,T)} \\
&\quad\le \widetilde{c}^*\Bigg\{\|u_0 - \pi_0 u_0\|_0^2 + \sum_{n=1}^{N}\tau_n\Big[(\eta_h^n)^2 + |||u_h^n - u_h^{n-1}|||^2 + (\widetilde{\eta}_h^n)^2 + |||\widetilde{u}_h^n|||^2\Big] \\
&\qquad + \sum_{n=1}^{N}\tau_n\,(\Theta_h^n)^2 + \big\|f - f^{n\theta} - (\underline{a} - \underline{a}^{n\theta})\cdot\nabla u_{h,\tau} - (b - b^{n\theta})u_{h,\tau}\big\|^2_{L^2(0,T;H_D^1(\Omega)^*)} \\
&\qquad + \big\|g - g_{h,\tau}\big\|^2_{L^2(0,T;H^{-1/2}(\Gamma_N))}\Bigg\}^{1/2}
\end{aligned}$$

and on each interval $(t_{n-1}, t_n]$, $1 \le n \le N$, *from below by*

$$\begin{aligned}
&\tau_n^{1/2}\Big\{(\eta_h^n)^2 + |||u_h^n - u_h^{n-1}|||^2 + (\widetilde{\eta}_h^n)^2 + |||\widetilde{u}_h^n|||^2\Big\}^{1/2} \\
&\quad\le \widetilde{c}_*\Bigg\{\|u - u_{h,\tau}\|^2_{X(t_{n-1},t_n)} + \tau_n\,(\Theta_h^n)^2 \\
&\qquad + \big\|f - f^{n\theta} - (\underline{a} - \underline{a}^{n\theta})\cdot\nabla u_{h,\tau} - (b - b^{n\theta})u_{h,\tau}\big\|^2_{L^2(t_{n-1},t_n;H_D^1(\Omega)^*)} \\
&\qquad + \big\|g - g_{h,\tau}\big\|^2_{L^2(t_{n-1},t_n;H^{-1/2}(\Gamma_N))}\Bigg\}^{1/2}.
\end{aligned} \tag{8.5}$$

The quantities η_h^n *and* Θ_h^n *are defined in* (5.10) *and* (5.11), *respectively. The constants* $\widetilde{c}^*$ *and* $\widetilde{c}_*$ *depend on the ratios* h_K/ρ_K. *The constant* $\widetilde{c}^*$ *in addition depends on the ratios* $h_{K'}/h_K$. *The constant* $\widetilde{c}_*$ *in addition depends on the maximum of the polynomial degrees of the finite element functions. All constants are independent of the final time* T, *the viscosity* ε, *and the parameter* β.

Remark 8.3. Theorem 8.2 shows that the quantity $\tau_n^{1/2}\{(\eta_h^n)^2 + |||u_h^n - u_h^{n-1}|||^2 + (\widetilde{\eta}_h^n)^2 + |||\widetilde{u}_h^n|||^2\}^{1/2}$ is a robust error indicator in the sense described in the introduction.

The remaining terms on the right-hand side of estimate (8.4) and the second and third term on the right-hand side of estimate (8.5) are data errors. They can be bounded a priori by computable norms involving the data f, g, $\underline{a}$, and b. The term $\tau_n^{1/2}\eta_h^n$ can be interpreted as a spatial error indicator. The terms $\tau_n^{1/2}\{|||u_h^n - u_h^{n-1}|||^2 + (\widetilde{\eta}_h^n)^2 + |||\widetilde{u}_h^n|||^2\}^{1/2}$ on the other hand can be viewed as temporal error indicators.

Acknowledgments. Part of this work was carried out during a stay at the Department of Mathematics of the Polytechnico di Torino. We want to thank Prof. Claudio Canuto and Dr. Stefano Berrone for their kind hospitality and the many fruitful discussions.

REFERENCES

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

[2] H. Amann, *Linear and Quasilinear Parabolic Problems, Volume* I: *Abstract Linear Theory*, Birkhäuser Boston, Boston, 1995.

[3] G. Bangerth and R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Birkhäuser Verlag, Basel, 2003.

[4] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Springer-Verlag, Berlin, Heidelberg, New York, 1992.

[5] L. P. Franca, S. L. Frey, and T. J. R. Hughes, *Stabilized finite element methods* I: *Application to the advective-diffusive model*, Comput. Methods Appl. Mech. Engrg., 95 (1992), pp. 253–276.

[6] T. J. R. Hughes and A. Brooks, *Streamline upwind/Petrov Galerkin formulations for the convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 54 (1982), pp. 199–259.

[7] R. Verfürth, *Robust a posteriori error estimators for a singularly perturbed reaction-diffusion equation*, Numer. Math., 78 (1998), pp. 479–493.

[8] R. Verfürth, *Error estimates for some quasi-interpolation operators*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 695–713.

[9] R. Verfürth, *A posteriori error estimates for finite element discretizations of the heat equation*, Calcolo, 40 (2003), pp. 195–212.

[10] R. Verfürth, *Robust A Posteriori Error Estimates for Stationary Convection-Diffusion Equations*, Report, Ruhr-Universität Bochum, Bochum, Germany, 2004.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 1803–1827

CONVERGENCE OF ADAPTIVE FINITE ELEMENT METHODS FOR GENERAL SECOND ORDER LINEAR ELLIPTIC PDEs*

KHAMRON MEKCHAY[†] AND RICARDO H. NOCHETTO[‡]

Abstract. We prove convergence of adaptive finite element methods (AFEMs) for general (nonsymmetric) second order linear elliptic PDEs, thereby extending the result of Morin, Nochetto, and Siebert [*SIAM J. Numer. Anal.*, 38 (2000), pp. 466–488; *SIAM Rev.*, 44 (2002), pp. 631–658]. The proof relies on quasi-orthogonality, which accounts for the bilinear form not being a scalar product, together with novel error and oscillation reduction estimates, which now do not decouple. We show that AFEMs are a contraction for the sum of energy error plus oscillation. Numerical experiments, including oscillatory coefficients and both coercive and noncoercive convection-diffusion PDE, illustrate the theory and yield optimal meshes.

Key words. a posteriori error estimators, quasi-orthogonality, adaptive mesh refinement, error and oscillation reduction estimates, optimal meshes

AMS subject classifications. 65N12, 65N15, 65N30, 65N50, 65Y20

DOI. 10.1137/04060929X

1. Introduction and main result.

Let Ω be a polyhedral bounded domain in $\mathbb{R}^d$ $(d = 2, 3)$. We consider a homogeneous Dirichlet boundary value problem for a general second order elliptic partial differential equation (PDE):

$$\mathcal{L}u = -\nabla\cdot(\mathbf{A}\nabla u) + \mathbf{b}\cdot\nabla u + c\,u = f \quad \text{in } \Omega, \tag{1.1}$$

$$u = 0 \quad \text{on } \partial\Omega. \tag{1.2}$$

The choice of boundary condition is made for ease of presentation, since similar results are valid for other boundary conditions. We also assume

- $\mathbf{A} : \Omega \mapsto \mathbb{R}^{d\times d}$ is Lipschitz and symmetric positive definite with smallest eigenvalue a_- and largest eigenvalue a_+, i.e.,

$$a_-(x)\,|\xi|^2 \le \mathbf{A}(x)\xi\cdot\xi \le a_+(x)\,|\xi|^2 \quad \forall \xi \in \mathbb{R}^d, \qquad x \in \Omega; \tag{1.3}$$

- $\mathbf{b} \in [L^\infty(\Omega)]^d$ is divergence free ($\nabla\cdot\mathbf{b} = 0$ in Ω);
- $c \in L^\infty(\Omega)$ is nonnegative ($c \ge 0$ in Ω);
- $f \in L^2(\Omega)$.

The purpose of this paper is to prove the following convergence results for adaptive finite element methods (AFEMs) for (1.1)–(1.2) and document their performance computationally.

THEOREM 1.1 (convergence of AFEMs). *Let $\{u_k\}_{k\in\mathbb{N}_0}$ be a sequence of finite element solutions corresponding to a sequence of nested finite element spaces $\{\mathbb{V}_k\}_{k\in\mathbb{N}_0}$ produced by the AFEM of section* 3.5*, which involves loops of the form*

[*] Received by the editors June 6, 2004; accepted for publication (in revised form) April 25, 2005; published electronically November 22, 2005.
http://www.siam.org/journals/sinum/43-5/60929.html

[†] Department of Mathematics, University of Maryland, College Park, MD 20742 (kxm@math.umd.edu). The work of this author was partially supported by NSF grant DMS-0204670.

[‡] Department of Mathematics and Institute for Physical Science and Technology, University of Maryland, College Park, MD 20742 (rhn@math.umd.edu). The work of this author was partially supported by NSF grant DMS-0204670 and the General Research Board of the University of Maryland.

$$\mathsf{SOLVE} \rightarrow \mathsf{ESTIMATE} \rightarrow \mathsf{MARK} \rightarrow \mathsf{REFINE}.$$

There exist constants $\sigma, \gamma > 0$ and $0 < \xi < 1$, depending solely on the shape regularity of meshes, the data, the parameters used by AFEM, and a number $0 < s \leq 1$ dictated by the interior angles of $\partial\Omega$, such that if the initial meshsize h_0 satisfies $h_0^s \|\mathbf{b}\|_{L^\infty} < \sigma$, then for any two consecutive iterations k and $k+1$, we have

$$\left|\!\left|\!\left| u - u_{k+1} \right|\!\right|\!\right|^2 + \gamma \ \mathsf{osc}_{k+1}\Omega^2 \leq \xi^2 \left(\left|\!\left|\!\left| u - u_k \right|\!\right|\!\right|^2 + \gamma \ \mathsf{osc}_k \Omega^2 \right). \tag{1.4}$$

Therefore, AFEM converges with a linear rate ξ, namely,

$$\left|\!\left|\!\left| u - u_k \right|\!\right|\!\right|^2 + \gamma \ \mathsf{osc}_k \Omega^2 \leq C_0 \ \xi^{2k},$$

where $C_0 := \left|\!\left|\!\left| u - u_0 \right|\!\right|\!\right|^2 + \gamma \ \mathsf{osc}_0(\Omega)^2$.

Hereafter, $|\!|\!| \cdot |\!|\!|$ denotes the energy norm induced by the operator $\mathcal{L}$ and $\mathsf{osc}(\Omega)$, the oscillation term, stands for the information missed by the averaging process associated with FEM. This convergence result extends those of Morin, Nochetto, and Siebert [7, 8] in several ways:

- We deal with a full second order linear elliptic PDE with variable coefficients $\mathbf{A}, \mathbf{b}$, and c, whereas in [7, 8] $\mathbf{A}$ is assumed to be piecewise constant and $\mathbf{b}$ and c to vanish.
- The underlying bilinear form $\mathcal{B}$ is nonsymmetric due to the first order term $\mathbf{b} \cdot \nabla u$. Since $\mathcal{B}$ is no longer a scalar product as in [7, 8], the Pythagoras equality relating u, u_k, and u_{k+1} fails; we prove a *quasi-orthogonality* property instead.
- The oscillation terms depend on discrete solutions in addition to data. Therefore, oscillation and error cannot be reduced separately as in [7, 8].
- The oscillation terms do not involve the oscillation of the jump residuals. This is achieved by exploiting positivity and continuity of $\mathbf{A}$.
- Since error and oscillation are now coupled, in order to prove convergence we need to handle them together. This leads to a novel argument and result, the contraction property (1.4), according to which both error and oscillation decrease together.

This paper is organized as follows. In section 2, we introduce the bilinear form, the energy norm, recall existence and uniqueness of solutions, and state the quasi-orthogonality property. In section 3, we describe the procedures used in AFEM, namely, SOLVE, ESTIMATE, MARK, and REFINE; state new error and oscillation reduction estimates; present the adaptive algorithm AFEM; and prove its convergence. In section 4, we prove the quasi-orthogonality property of section 2 and the error and oscillation reduction estimates of section 3. In section 5, we present three numerical experiments to illustrate properties of AFEM. We conclude in section 6 with extensions to $\mathbf{A}$ piecewise Lipschitz with discontinuities aligned with the initial mesh, as well as noncoercive bilinear form $\mathcal{B}$ due to $\nabla \cdot \mathbf{b} \neq 0$ and a numerical experiment.

2. Discrete solution and quasi-orthogonality.

For an open set $G \subset \mathbb{R}^d$ we denote by $H^1(G)$ the usual Sobolev space of functions in $L^2(G)$ whose first derivatives are also in $L^2(G)$, endowed with the norm

$$\|u\|_{H^1(G)} := \left(\|u\|_{L^2(G)} + \|\nabla u\|_{L^2(G)} \right)^{1/2}.$$

We use the symbols $\|\cdot\|_{H^1}$ and $\|\cdot\|_{L^2}$ when $G = \Omega$. Moreover, we denote by $H_0^1(G)$ the space of functions in $H^1(G)$ that vanish on the boundary in the trace sense.

A weak solution of (1.1) and (1.2) is a function u satisfying

$$u \in H^1_0(\Omega) \ : \ \mathcal{B}[u,v] = \langle f, v\rangle \qquad \forall\, v \in H^1_0(\Omega), \tag{2.1}$$

where $\langle u, v\rangle := \int_\Omega uv$ for any $u, v \in L^2(\Omega)$, and the bilinear form is defined on $H^1_0(\Omega) \times H^1_0(\Omega)$ as

$$\mathcal{B}[u,v] := \langle \mathbf{A}\nabla u, \nabla v\rangle + \langle \mathbf{b}\cdot\nabla u + c\, u, v\rangle. \tag{2.2}$$

By the Cauchy–Schwarz inequality one can easily show the *continuity* of the bilinear form

$$|\mathcal{B}[u,v]| \le C_B \, \|u\|_{H^1} \, \|v\|_{H^1},$$

where C_B depends only on the data. Combining Poincaré inequality with the divergence free condition $\nabla\cdot\mathbf{b}=0$, one has *coercivity* in $H^1_0(\Omega)$

$$\mathcal{B}[v,v] \ge \int_\Omega a_- \, |\nabla v|^2 + cv^2 \ge c_B \, \|v\|^2_{H^1},$$

where c_B depends only on the data. Existence and uniqueness of (2.1) thus follows from the Lax–Milgram theorem [5].

We define the energy norm on $H^1_0(\Omega)$ by $|\!|\!| v |\!|\!|^2 := \mathcal{B}[v,v]$, which is equivalent to $H^1_0(\Omega)$-norm $\|\cdot\|_{H^1}$. In fact we have

$$c_B \, \|v\|^2_{H^1} \le |\!|\!| v |\!|\!|^2 \le C_B \, \|v\|^2_{H^1} \qquad \forall\, v \in H^1_0(\Omega). \tag{2.3}$$

2.1. Discrete solutions on nested meshes. Let $\{\mathcal{T}_H\}$ be a shape regular family of nested conforming meshes over Ω, that is, there exists a constant γ^* such that

$$\frac{H_T}{\rho_T} \le \gamma^* \qquad \forall\, T \in \bigcup_H \mathcal{T}_H, \tag{2.4}$$

where, for each $T \in \mathcal{T}_H$, H_T is the diameter of T and ρ_T is the diameter of the biggest ball contained in T; the global meshsize is $h_H := \max_{T\in\mathcal{T}_H} H_T$.

Let $\{\mathbb{V}_H\}$ be a corresponding family of nested finite element spaces consisting of continuous piecewise polynomials over $\mathcal{T}_H$ of fixed degree $n \ge 1$ that vanish on the boundary. Let u_H be a discrete solution of (2.1) satisfying

$$u_H \in \mathbb{V}_H \ : \ \mathcal{B}[u_H, v_H] = \langle f, v_H\rangle \qquad \forall\, v_H \in \mathbb{V}_H. \tag{2.5}$$

The effect of quadrature is not considered in this paper. Existence and uniqueness of this problem follows from the Lax–Milgram theorem, since $\mathbb{V}_H \subset H^1_0(\Omega)$.

2.2. Quasi-orthogonality. Consider two consecutive nested meshes $\mathcal{T}_H \subset \mathcal{T}_h$, i.e., $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$. For the corresponding spaces $\mathbb{V}_H \subset \mathbb{V}_h \subset H^1_0(\Omega)$, let $u_h \in \mathbb{V}_h$ and $u_H \in \mathbb{V}_H$ be the discrete solutions. Since the bilinear form is nonsymmetric, it is not a scalar product and the orthogonality relation between $u - u_H$ and $u_h - u_H$, the so-called Pythagoras equality, fails to hold. We have instead a perturbation result referred to as quasi-orthogonality provided that the initial mesh is fine enough. This result is stated below and the proof is given in section 4.

LEMMA 2.1 (quasi-orthogonality). *Let $f \in L^2(\Omega)$. There exists a constant $C^* > 0$, solely depending on the shape regularity constant γ^*, the data $\mathbf{A}, \mathbf{b}$, and c, and a* *number $0 < s \le 1$ dictated only by the interior angles of $\partial\Omega$, such that if the meshsize h_0 of the initial mesh satisfies $C^* h_0^s \|\mathbf{b}\|_{L^\infty} < 1$, then*

$$\|\!|u - u_h|\!\|^2 \le \Lambda_0 \|\!|u - u_H|\!\|^2 - \|\!|u_h - u_H|\!\|^2, \tag{2.6}$$

where $\Lambda_0 := (1 - C^* h_0^s \|\mathbf{b}\|_{L^\infty})^{-1}$. The equality holds provided $\mathbf{b} = 0$ in Ω.

3. Adaptive algorithm.

The adaptive procedure consists of loops of the form

$$\mathsf{SOLVE} \to \mathsf{ESTIMATE} \to \mathsf{MARK} \to \mathsf{REFINE}.$$

The procedure SOLVE solves (2.5) for the discrete solution u_H. The procedure ESTIMATE determines the element indicators $\eta_H(T)$ and oscillation $\mathsf{osc}_H(T)$ for all elements $T \in \mathcal{T}_H$. Depending on their relative sizes, these quantities are later used by the procedure MARK to mark elements T and thereby create a subset $\widehat{\mathcal{T}}_H$ of $\mathcal{T}_H$ of elements to be refined. Finally, procedure REFINE partitions those elements in $\widehat{\mathcal{T}}_H$ and a few more to maintain mesh conformity. These procedures are discussed in more detail below.

3.1. Procedure SOLVE: Linear solver.

We employ linear solvers, either direct or iterative methods, such as preconditioned GMRES, CG, and BICG, to solve linear system (2.5). In other words, given a mesh $\mathcal{T}_k$, an initial guess u_{k-1} for the solution, and the data $\mathbf{A}, \mathbf{b}, c, f$, SOLVE computes the discrete solution

$$u_k := \mathsf{SOLVE}(\mathcal{T}_k, u_{k-1}, \mathbf{A}, \mathbf{b}, c, f).$$

3.2. Procedure ESTIMATE: A posteriori error estimate.

Since we assume exact numerical integration, subtracting (2.5) from (2.1) yields the Galerkin orthogonality

$$\mathcal{B}[u - u_H, v_H] = 0 \qquad \forall\, v_H \in \mathbb{V}_H. \tag{3.1}$$

In addition to $\mathcal{T}_H$, let $\mathcal{S}_H$ denote the set of interior faces (edges or sides) of the mesh (triangulation) $\mathcal{T}_H$. We consider the *residual* $\mathcal{R}(u_H) \in H^{-1}(\Omega)$ defined by

$$\mathcal{R}(u_H) := f + \nabla\cdot(\mathbf{A}\nabla u_H) - \mathbf{b}\cdot\nabla u_H - c\, u_H$$

and its relation to the error $\mathcal{L}(u - u_H) = \mathcal{R}(u_H)$. It is then clear that to estimate $\|\!|u - u_H|\!\|$ we can equivalently deal with $\|\mathcal{R}(u_H)\|_{H^{-1}(\Omega)}$. To this end, we integrate by parts elementwise the bilinear form $\mathcal{B}[u - u_H, v]$ to obtain the *error representation formula*

$$\mathcal{B}[u - u_H, v] = \sum_{T\in\mathcal{T}_H} \int_T R_T(u_H) v + \sum_{S\in\mathcal{S}_H} \int_S J_S(u_H) v \qquad \forall\, v \in H_0^1(\Omega), \tag{3.2}$$

where the *element residual* $R_T(u_H)$ and the *jump residual* $J_S(u_H)$ are defined as

$$R_T(u_H) := f + \nabla\cdot(\mathbf{A}\nabla u_H) - \mathbf{b}\cdot\nabla u_H - c\, u_H \qquad \text{in } T \in \mathcal{T}_H, \tag{3.3}$$

$$J_S(u_H) := -\mathbf{A}\nabla u_H^+ \cdot \nu^+ - \mathbf{A}\nabla u_H^- \cdot \nu^- := [\![\mathbf{A}\nabla u_H]\!]_S \cdot \nu_S \qquad \text{on } S \in \mathcal{S}_H, \tag{3.4}$$

where S is the common side of elements T^+ and T^- with unit outward normals ν^+ and ν^-, respectively, and $\nu_S = \nu^-$. Whenever convenient, we will use the abbreviations $R_T = R_T(u_H)$ and $J_S = J_S(u_H)$.

3.2.1. Upper bound. For $T \in \mathcal{T}_H$ and $S \in \mathcal{S}_h$ an interior face, we define the *local error indicator* $\eta_H(T)$ by

$$\eta_H(T)^2 := H_T^2 \left\| R_T(u_H) \right\|_{L^2(T)}^2 + \sum_{S \subset \partial T} H_S \left\| J_S(u_H) \right\|_{L^2(S)}^2 . \tag{3.5}$$

Given a subset $\omega \subset \Omega$, we define the *error estimator* $\eta_H(\omega)$ by

$$\eta_H(\omega)^2 := \sum_{T \in \mathcal{T}_H,\, T \subset \omega} \eta_H(T)^2.$$

Hence, $\eta_H(\Omega)$ is the error estimator of Ω with respect to the mesh $\mathcal{T}_H$. Using (3.1), (3.2), and properties of the Clément interpolation, as shown in [1, 3, 13], we obtain the upper bound of the error in terms of the estimator,

$$|\!|\!| u - u_H |\!|\!|^2 \leq C_1 \eta_H(\Omega)^2, \tag{3.6}$$

where the constant $C_1 > 0$ depends only on the shape regularity γ^*, coercivity constant c_B, and continuity constant C_B of the bilinear form.

3.2.2. Lower bound. Using the explicit construction of Verfürth [1, 13] via bubble functions and positivity and continuity of A, we can get a local lower bound of the error in terms of local indicators and oscillation. That is, there exist constants $C_2, C_3 > 0$, depending only on the shape regularity γ^*, C_B , and c_B, such that

$$C_2\, \eta_H(T)^2 - C_3 \sum_{T \subset \omega_T} H_T^2 \left\| R_T - \overline{R_T} \right\|_{L^2(T)}^2 \leq \left\| u - u_H \right\|_{H^1(\omega_T)}^2 , \tag{3.7}$$

where the domain ω_T consists of all elements sharing at least a side with T, and $\overline{R_T}$ is any polynomial approximation of R_T on T. However, for the purpose of proving Lemmas 3.1 and 3.2, we will assume that $\overline{R_T} \in \mathbb{P}_{n-1}(T)$ is the L^2-projection of R_T. We define the *oscillation* on the elements $T \in \mathcal{T}_H$ by

$$\mathsf{osc}_H(T)^2 := H_T^2 \left\| R_T - \overline{R_T} \right\|_{L^2(T)}^2 , \tag{3.8}$$

and for a subset $\omega \subset \Omega$, we define

$$\mathsf{osc}_H(\omega)^2 := \sum_{T \in \mathcal{T}_H,\, T \subset \omega} \mathsf{osc}_H(T)^2 .$$

Remark 3.1. We see from (3.7) that if the oscillation $\mathsf{osc}_H(\omega_T)$ is small compared to the indicator $\eta_H(T)$, then a large $\eta_H(T)$ implies a large local error $\| u - u_H \|_{H^1(\omega_T)}$. This explains why refining elements with large indicators usually tend to equidistribute the errors, which is an ultimate goal of adaptivity. This idea is employed by the procedure MARK of section 3.3.

Remark 3.2. The oscillation $\mathsf{osc}_H(T)$ does not involve oscillation of the jump residual $J_S(u_H)$ as is customary [1, 13]. This result follows from the positivity and continuity of $\mathbf{A}$, and is explained in section 4.2.

Remark 3.3. The oscillation $\mathsf{osc}_H(T)$ depends on $R_T = R_T(u_H)$, which in turn depends on the discrete solution u_H. This is a fundamental difference with Morin, Nochetto, and Siebert [7, 8], where the oscillation is purely a data oscillation. It is not clear now that the oscillation will decrease when the mesh $\mathcal{T}_H$ will be refined because

u_H will also change. Controlling the decay of $\mathsf{osc}_H(T)$ is thus a major challenge addressed in this work; see sections 3.3 and 3.4. It is not possible to show that the oscillation will always decrease as the mesh gets refined as in [7, 8].

For a given mesh $\mathcal{T}_H$ and discrete solution u_H, along with the input data $\mathbf{A}, \mathbf{b}, c$, and f, the procedure $\mathsf{ESTIMATE}$ computes indicators $\eta_H(T)$ and oscillations $\mathsf{osc}_H(T)$ for all elements $T \in \mathcal{T}_H$ according to (3.5) and (3.8):

$$\{\eta_H(T), \mathsf{osc}_H(T)\}_{T\in\mathcal{T}_H} = \mathsf{ESTIMATE}(\mathcal{T}_H, u_H, \mathbf{A}, \mathbf{b}, c, f).$$

3.3. Procedure MARK.

Our goal is to devise a marking procedure, namely, to identify a subset $\widehat{\mathcal{T}}_H$ of the mesh $\mathcal{T}_H$ such that, after refining, both error and oscillation will be reduced. We use two strategies for this: Marking Strategy E deals with the error estimator and Marking Strategy O does so with the oscillation.

3.3.1. Marking Strategy E: Error reduction.

This strategy was introduced by Dörfler [4] to enforce error reduction.

Marking Strategy E. *Given a parameter $0 < \theta < 1$, construct a subset $\widehat{\mathcal{T}}_H$ of $\mathcal{T}_H$ such that*

$$\sum_{T\in\widehat{\mathcal{T}}_H} \eta_H(T)^2 \geq \theta^2 \eta_H(\Omega)^2, \tag{3.9}$$

and mark all elements in $\widehat{\mathcal{T}}_H$ for refinement.

We will see later that Marking Strategy E guarantees error reduction in the absence of oscillation terms. Since the latter account for information missed by the averaging process associated with the finite element method, we need a separate procedure to guarantee oscillation reduction.

3.3.2. Marking Strategy O: Oscillation reduction.

This procedure was introduced by Morin, Nochetto, and Siebert [7, 8] as a separate means for reducing oscillation.

Marking Strategy O. *Given a parameter $0 < \hat{\theta} < 1$ and the subset $\widehat{\mathcal{T}}_H \subset \mathcal{T}_H$ produced by Marking Strategy E, enlarge $\widehat{\mathcal{T}}_H$ such that*

$$\sum_{T\in\widehat{\mathcal{T}}_H} \mathsf{osc}_H(T)^2 \geq \hat{\theta}^2 \mathsf{osc}_H(\Omega)^2\,, \tag{3.10}$$

and mark all elements in $\widehat{\mathcal{T}}_H$ for refinement.

Given a mesh $\mathcal{T}_H$ and all information about the local error indicators $\eta_H(T)$ and oscillation $\mathsf{osc}_H(T)$, together with user parameters θ and $\hat{\theta}$, MARK generates a subset $\widehat{\mathcal{T}}_H$ of $\mathcal{T}_H$,

$$\widehat{\mathcal{T}}_H = \mathsf{MARK}(\theta, \hat{\theta}\ ;\ \mathcal{T}_H, \{\eta_H(T), \mathsf{osc}_H(T)\}_{T\in\mathcal{T}_H}).$$

3.4. Procedure REFINE.

The following interior node property, due to Morin, Nochetto, and Siebert [7, 8], is known to be necessary for error and oscillation reduction.

Interior Node Property. *Refine each marked element $T \in \widehat{\mathcal{T}}_H$ to obtain a new mesh $\mathcal{T}_h$ compatible with $\mathcal{T}_H$ such that*

> *T and the $d+1$ adjacent elements $T' \in \mathcal{T}_H$ of T, as well as their common sides, contain a node of the finer mesh $\mathcal{T}_h$ in their interior.*

In addition to the interior node property, we assume that the refinement is done in such a way that the new mesh $\mathcal{T}_h$ is conforming, which guarantees that both $\mathcal{T}_H$

and $\mathcal{T}_h$ are nested. With this property, we have a reduction factor $\gamma_0 < 1$ of element size, i.e., if $T \in \mathcal{T}_h$ is obtained by refining $T' \in \widehat{\mathcal{T}}_H$, then $h_T \le \gamma_0 H_{T'}$. For example, when $d = 2$ with triangular elements, to have the interior node property we can use the three newest bisections for each single refinement step, whence $\gamma_0 \le 1/2$.

Given a mesh $\mathcal{T}_H$ and a marked set $\widehat{\mathcal{T}}_H$, REFINE constructs the refinement $\mathcal{T}_h$ satisfying the interior node property:

$$\mathcal{T}_h = \mathsf{REFINE}(\mathcal{T}_H, \widehat{\mathcal{T}}_H).$$

Combining the marking strategies of section 3.3 with the interior node property, we obtain the following two crucial results whose proofs are given in section 4.

LEMMA 3.1 (error reduction). *There exist constants C_4 and C_5, depending only on the shape regularity constant γ^* and θ, such that*

$$\eta_H(T)^2 \le C_4 \left\| u_h - u_H \right\|^2_{H^1(\omega_T)} + C_5 \mathsf{osc}_H \omega_T{}^2 \qquad \forall\, T \in \widehat{\mathcal{T}}_H. \tag{3.11}$$

We realize that the local energy error between consecutive discrete solutions is bounded below by the local indicators for elements in the marked set $\widehat{\mathcal{T}}_H$, provided the oscillation term is sufficiently small relative to the energy error.

LEMMA 3.2 (oscillation reduction). *There exist constants $0 < \rho_1 < 1$ and $0 < \rho_2$, depending only on γ^* and $\hat{\theta}$, such that*

$$\mathsf{osc}_h \Omega^2 \le \rho_1 \mathsf{osc}_H \Omega^2 + \rho_2 \left|\!\left|\!\left| u_h - u_H \right|\!\right|\!\right|^2 . \tag{3.12}$$

We have that the oscillation reduces with a factor $\rho_1 < 1$ provided the energy error between consecutive discrete solutions is relatively small.

Remark 3.4 (coupling of error and oscillation). Lemmas 3.1 and 3.2 seem to lead to conflicting demands on the relative sizes of error and oscillation. These two concepts are indeed coupled, which contrasts with [7, 8], where the oscillation depends only on the data and reduces separately from the error. This suggests that we must handle them together, this being the main contribution of this paper. We make this assertion explicit in Theorem 1.1.

3.5. Adaptive algorithm AFEM. The adaptive algorithm consists of the loops of procedures SOLVE, $\mathsf{ESTIMATE}$, MARK, and REFINE, consecutively, given that the parameters θ and $\hat{\theta}$ are chosen according to Marking Strategies E and O.

ALGORITHM AFEM.

```
Choose parameters $0 < \theta, \hat{\theta} < 1$.
 1. Pick an initial mesh $\mathcal{T}_0$, initial guess $u_{-1} = 0$, and set $k = 0$.
 2. $u_k = \mathsf{SOLVE}(\mathcal{T}_k, u_{k-1}, \mathbf{A}, \mathbf{b}, c, f)$.
 3. $\{\eta_k(T), \mathsf{osc}_k(T)\}_{T \in \mathcal{T}_k} = \mathsf{ESTIMATE}(\mathcal{T}_k, u_k, \mathbf{A}, \mathbf{b}, c, f)$.
 4. $\widehat{\mathcal{T}_k} = \mathsf{MARK}(\theta, \hat{\theta}\ ;\ \mathcal{T}_k, \{\eta_k(T), \mathsf{osc}_k(T)\}_{T \in \mathcal{T}_k})$.
 5. $\mathcal{T}_{k+1} = \mathsf{REFINE}(\mathcal{T}_k, \widehat{\mathcal{T}_k})$.
 6. Set $k = k + 1$ and go to step 2.
```

THEOREM 1.1 (convergence of AFEM). *Let $\{u_k\}_{k \in \mathbb{N}_0}$ be a sequence of finite element solutions corresponding to a sequence of nested finite element spaces $\{\mathbb{V}^k\}_{k \in \mathbb{N}_0}$ produced by AFEM. There exist constants $\sigma, \gamma > 0$ and $0 < \xi < 1$, depending solely on the mesh regularity constant γ^*, data, parameters θ and $\hat{\theta}$, and a number $0 < s \le 1$ dictated by interior angles of $\partial\Omega$, such that if the initial meshsize h_0 satisfies $h_0^s \|\mathbf{b}\|_{L^\infty} < \sigma$, then for any two consecutive iterations k and $k+1$, we have*

$$\left|\!\left|\!\left| u - u_{k+1} \right|\!\right|\!\right|^2 + \gamma\ \mathsf{osc}_{k+1} \Omega^2 \le \xi^2 \left(\left|\!\left|\!\left| u - u_k \right|\!\right|\!\right|^2 + \gamma\ \mathsf{osc}_k \Omega^2 \right). \tag{3.13}$$

Therefore, AFEM converges with a linear rate ξ, namely,

$$|||u-u_k|||^2+\gamma\ \mathsf{osc}_k\Omega^2\le C_0\ \xi^{2k},$$

where $C_0:=|||u-u_0|||^2+\gamma\ \mathsf{osc}_0\Omega^2$.

Proof. We just prove the contraction property (3.13), which obviously implies the decay estimate. For convenience, we introduce the notation

$$e_k:=|||u-u_k|||\,,\qquad \varepsilon_k:=|||u_{k+1}-u_k|||\,,\qquad \mathsf{osc}_k:=\mathsf{osc}_k(\Omega)\,.$$

The idea is to use the quasi-orthogonality (2.6) and replace the term $|||u_{k+1}-u_k|||^2$ using new results of error and oscillation reduction estimates (3.11) and (3.12). We proceed in three steps as follows.

Step 1. We first get a lower bound for ε_k in terms of e_k. To this end, we use Marking Strategy E and the upper bound (3.6) to write

$$\theta^2e_k^2\le C_1\theta^2\eta_k(\Omega)^2\le C_1\sum_{T\in\widehat{\mathcal{T}_k}}\eta_k(T)^2.$$

Adding (3.11) of Lemma 3.1 over all marked elements $T\in\widehat{\mathcal{T}_k}$, and observing that each element can be counted at most $D:=d+2$ times due to overlap of the sets ω_T, together with $\|v\|_{H^1}^2\le c_B^{-1}\,|||v|||^2$ for all $v\in H_0^1(\Omega)$, we arrive at

$$\theta^2e_k^2\le\frac{DC_1C_4}{c_B}\ \varepsilon_k^2+DC_1C_5\ \mathsf{osc}_k^2.$$

If $\Lambda_1:=\frac{\theta^2c_B}{DC_1C_4},\Lambda_2:=\frac{C_5c_B}{C_4}$, then this implies the lower bound for ε_k^2,

$$\varepsilon_k^2\ge\Lambda_1e_k^2-\Lambda_2\mathsf{osc}_k^2. \tag{3.14}$$

Step 2. If h_0 is sufficiently small so that the quasi-orthogonality (2.6) of Lemma 2.1 holds with $\Lambda_0=(1-C^*h_0^s\,\|\mathbf{b}\|_{L^\infty})^{-1}$, then

$$e_{k+1}^2\le\Lambda_0e_k^2-\varepsilon_k^2.$$

Replacing the fraction $\beta\varepsilon_k^2$ of ε_k^2 via (3.14) we obtain

$$e_{k+1}^2\le(\Lambda_0-\beta\Lambda_1)e_k^2+\beta\Lambda_2\mathsf{osc}_k^2-(1-\beta)\varepsilon_k^2,$$

where $0<\beta<1$ is a constant to be chosen suitably. We now assert that it is possible to choose h_0 compatible with Lemma 2.1 and also that

$$0<\alpha:=\Lambda_0-\beta\Lambda_1<1.$$

A simple calculation shows that this is the case provided

$$C^*h_0^s\,\|\mathbf{b}\|_{L^\infty}<\frac{\beta\Lambda_1}{(1+\beta\Lambda_1)}<1,$$

i.e., $h_0^s\,\|\mathbf{b}\|_{L^\infty}<\sigma$ with $\sigma:=\frac{\beta\Lambda_1}{C^*(1+\beta\Lambda_1)}$. Consequently,

$$e_{k+1}^2\le\alpha e_k^2+\beta\Lambda_2\mathsf{osc}_k^2-(1-\beta)\varepsilon_k^2. \tag{3.15}$$

Step 3. To remove the last term of (3.15) we resort to the oscillation reduction estimate of Lemma 3.2

$$\mathsf{osc}_{k+1}^2 \le \rho_1 \mathsf{osc}_k^2 + \rho_2 \varepsilon_k^2.$$

We multiply it by $(1-\beta)/\rho_2$ and add it to (3.15) to deduce

$$e_{k+1}^2 + \frac{1-\beta}{\rho_2}\mathsf{osc}_{k+1}^2 \le \alpha\, e_k^2 + \left(\beta\Lambda_2 + \frac{\rho_1}{\rho_2}(1-\beta)\right)\mathsf{osc}_k^2.$$

If $\gamma := \frac{1-\beta}{\rho_2}$, then we would like to choose $\beta < 1$ in such a way that

$$\beta\Lambda_2 + \rho_1\gamma = \mu\gamma$$

for some $\mu < 1$. A simple calculation yields

$$\beta = \frac{\frac{\mu-\rho_1}{\rho_2}}{\Lambda_2 + \frac{\mu-\rho_1}{\rho_2}},$$

and shows that $\rho_1 < \mu < 1$ guarantees that $0 < \beta < 1$. Therefore,

$$e_{k+1}^2 + \gamma\ \mathsf{osc}_{k+1}^2 \le \alpha\ e_k^2 + \mu\gamma\ \mathsf{osc}_k^2$$

and the asserted estimate (3.13) follows upon taking $\xi = \max(\alpha, \mu) < 1$. □

Remark 3.5 (comparison with [7, 8]). In [7, 8] the oscillation is independent of discrete solutions, i.e., $\rho_2 = 0$, and is reduced by the factor $\rho_1 < 1$ in (3.12). Consequently, Step 3 is avoided by setting $\beta = 1$, and the decay of e_k and osc_k is monitored separately. Since this is no longer possible, e_k and osc_k are now combined and decreased together.

Remark 3.6 (splitting of ε_k). The idea of splitting ε_k is already used by Chen and Jia [2] in examining one time step for the heat equation. This is because a mass (zero order) term naturally occurs, which did not take place in [7, 8]. The elliptic operator is just the Laplacian in [2].

Remark 3.7 (effect of convection). Assuming that $h_0^s \|\mathbf{b}\|_{L^\infty} < \sigma$ implies that the local Péclet number is sufficiently small for the Galerkin method not to exhibit oscillations. This appears to be essential for u_0 to contain relevant information and guide correctly the adaptive process. This restriction is difficult to verify in practice because it involves unknown constants. However, starting from coarser meshes than needed in theory does not seem to be a problem in our examples (see section 5.3) where we carefully express the constant σ in terms of data.

Remark 3.8 (vanishing convection). If $\mathbf{b} = 0$, then Theorem 1.1 has no restriction on the initial mesh. This thus extends the convergent result of Morin, Nochetto, and Siebert [7, 8] to variable diffusion coefficient and zero order terms.

Remark 3.9 (optimal β). The choice of β can be optimized. In fact, we can easily see that

$$\alpha = \Lambda_0 - \beta\Lambda_1, \qquad \mu = \rho_1 + \frac{\beta}{1-\beta}\rho_2\Lambda_2$$

yield a unique value $0 < \beta_* < 1$ for which $\alpha = \mu$ and the contraction constant ξ of Theorem 1.1 is minimal. This β_* depends on the geometric constants Λ_0, Λ_1, and Λ_2 as well on $\theta, \hat{\theta}$, and h_0, but it is not computable.

4. Proofs of lemmas. Let $\widehat{\mathcal{T}}_H \subset \mathcal{T}_H$ be a set of marked elements obtained from procedure MARK. Let $\mathcal{T}_h$ be a refined mesh obtained from procedure REFINE, and let $\mathbb{V}_H \subset \mathbb{V}_h$ be nested spaces corresponding to compatible meshes $\mathcal{T}_H$ and $\mathcal{T}_h$, respectively. For convenience, set

$$e_h := u - u_h, \qquad e_H := u - u_H, \qquad \varepsilon_H := u_h - u_H.$$

4.1. Proof of Lemma 2.1: Quasi-orthogonality. In view of the Galerkin orthogonality (3.1), namely, $\mathcal{B}[e_h, v_h] = 0$, $v_h \in \mathbb{V}_h$, we have

$$|\!|\!| e_H |\!|\!|^2 = |\!|\!| e_h |\!|\!|^2 + |\!|\!| \varepsilon_H |\!|\!|^2 + \mathcal{B}[\varepsilon_H, e_h].$$

If $\mathbf{b} = 0$, then $\mathcal{B}$ is symmetric and $\mathcal{B}[\varepsilon_H, e_h] = \mathcal{B}[e_h, \varepsilon_H] = 0$. For $\mathbf{b} \neq 0$, instead, $\mathcal{B}[\varepsilon_H, e_h] \neq 0$, and we must account for this term. It is easy to see that $\nabla \cdot \mathbf{b} = 0$ and that integration by parts yields

$$\mathcal{B}[\varepsilon_H, e_h] = \mathcal{B}[e_h, \varepsilon_H] + \langle \mathbf{b} \cdot \nabla \varepsilon_H, e_h \rangle - \langle \mathbf{b} \cdot \nabla e_h, \varepsilon_H \rangle = 2 \langle \mathbf{b} \cdot \nabla \varepsilon_H, e_h \rangle.$$

Hence,

$$|\!|\!| e_h |\!|\!|^2 = |\!|\!| e_H |\!|\!|^2 - |\!|\!| \varepsilon_H |\!|\!|^2 - 2 \langle \mathbf{b} \cdot \nabla \varepsilon_H, e_h \rangle .$$

Using the Cauchy–Schwarz inequality and replacing the $H^1(\Omega)$-norm by the energy norm, we have for any $\delta > 0$ to be chosen later

$$-2 \langle \mathbf{b} \cdot \nabla \varepsilon_H, e_h \rangle \leq \delta \, \|e_h\|_{L^2}^2 + \frac{\|\mathbf{b}\|_{L^\infty}^2}{\delta c_B} |\!|\!| \varepsilon_H |\!|\!|^2 .$$

We then realize the need to relate $L^2(\Omega)$ and energy norms to replace $\|e_h\|_{L^2}$ by $|\!|\!| e_h |\!|\!|$. This requires a standard duality argument whose proof is reported in Ciarlet [3].

LEMMA 4.1 (duality). *Let $f \in L^2(\Omega)$ and $u \in H^{1+s}(\Omega)$ for some $0 < s \leq 1$ be the solution of* (2.1)*, where s depends on the interior angles of $\partial\Omega$ ($s = 1$ if Ω is convex). Then, there exists a constant C_D, depending only on the shape regularity constant γ^* and the data of* (1.1)*, such that*

$$\|e_h\|_{L^2} \leq C_D h^s \, \|e_h\|_{H^1} . \tag{4.1}$$

Inserting this estimate in the preceding two bounds, and using $h \leq h_0$, the mesh-size of the initial mesh, in conjunction with (2.3), we deduce

$$\left(1 - \delta {C_D}^2 c_B^{-1} h_0^{2s}\right) |\!|\!| e_h |\!|\!|^2 \leq |\!|\!| e_H |\!|\!|^2 - \left(1 - \|\mathbf{b}\|_{L^\infty}^2 (\delta c_B)^{-1}\right) |\!|\!| \varepsilon_H |\!|\!|^2 .$$

We now choose $\delta = \frac{\|\mathbf{b}\|_{L^\infty}}{C_D h_0^s}$ to equate both parentheses, as well as h_0 sufficiently small for $\delta {C_D}^2 h_0^{2s} c_B^{-1} = C^* h_0^s \, \|\mathbf{b}\|_{L^\infty} < 1$ with $C^* := C_D / c_B$. We end up with

$$|\!|\!| e_h |\!|\!|^2 \leq \frac{1}{1 - C^* h_0^s \, \|\mathbf{b}\|_{L^\infty}} |\!|\!| e_H |\!|\!|^2 - |\!|\!| \varepsilon_H |\!|\!|^2 .$$

This implies (2.6) and concludes the proof. $\square$

4.2. Proof of Lemma 3.1: Error reduction. Upon restricting the test function v in (3.2) to $\mathbb{V}_h \supset \mathbb{V}_H$, we obtain the error representation

$$\mathcal{B}[\varepsilon_H, v_h] = \sum_{T\in\mathcal{T}_H} \int_T \overline{R_T} v_h + \int_T (R_T - \overline{R_T}) v_h + \sum_{S\in\mathcal{S}_H} \int_S J_S\, v_h \qquad \forall\, v_h \in \mathbb{V}_h, \tag{4.2}$$

where we use the abbreviations $R_T = R_T(u_H)$ and $J_S = J_S(u_H)$, and $\overline{R_T} = \Pi_T^{n-1} R_T$ denotes the L_2-projection of R_T onto the space of polynomials $\mathbb{P}_{n-1}(T)$ over the element $T \in \mathcal{T}_H$. Except for avoiding the oscillation terms of the jump residual J_S, the proof goes back to [4, 7, 8]. We proceed in three steps.

Step 1. *Interior residual.* Let $T \in \mathcal{T}_H$, and let x_T be an interior node of T generated by the procedure REFINE. Let $\psi_T \in \mathbb{V}_h$ be a bubble function which satisfies $\psi_T(x_T) = 1$, vanishes on ∂T, and $0 \le \psi_T \le 1$; hence supp $\psi_T \subset T$. Since $\overline{R_T} \in \mathbb{P}_{n-1}(T)$ and $\psi_T > 0$ in a polyhedron of measure comparable with that of T, we have

$$C \left\|\overline{R_T}\right\|_{L^2(T)}^2 \le \int_T \psi_T\, \overline{R_T}^2 = \int_T \overline{R_T}(\psi_T\, \overline{R_T}).$$

Since $\psi_T \overline{R_T}$ is a piecewise polynomial of degree $\le n$ over $\mathcal{T}_h$, it is an admissible test function in (4.2) which vanishes outside T (and in particular on all $S \in \mathcal{S}_H$). Therefore,

$$\begin{aligned} C \left\|\overline{R_T}\right\|_{L^2(T)}^2 &\le \mathcal{B}[\varepsilon_H, \psi_T \overline{R_T}] + \int_T (\overline{R_T} - R_T)\psi_T \overline{R_T} \\ &\le C \left(H_T^{-1} \left\|\varepsilon_H\right\|_{H^1(T)} + \left\|R_T - \overline{R_T}\right\|_{L^2(T)} \right) \left\|\overline{R_T}\right\|_{L^2(T)}, \end{aligned}$$

because of an inverse inequality for $\psi_T \overline{R_T}$. This, together with the triangle inequality, yields the desired estimate for $H_T^2 \left\|R_T\right\|_{L^2(T)}^2$,

$$H_T^2 \left\|R_T\right\|_{L^2(T)}^2 \le C \left(\left\|\varepsilon_H\right\|_{H^1(T)}^2 + H_T^2 \left\|R_T - \overline{R_T}\right\|_{L^2(T)}^2 \right). \tag{4.3}$$

Step 2. *Jump residual.* Let $S \in \mathcal{S}_H$ be an interior side of $T_1 \in \widehat{\mathcal{T}}_H$, and let $T_2 \in \mathcal{T}_H$ be the other element sharing S. Let x_S be an interior node of S created by procedure REFINE. Let $\psi_S \in \mathbb{V}_h$ be a bubble function in $\omega_S := T_1 \cup T_2$ such that $\psi_S(x_S) = 1$, ψ_S vanishes on $\partial\omega_S$, and $0 \le \psi_S \le 1$; hence supp $\psi_S \subset \omega_S$.

Since u_H is continuous, $[\![\nabla u_H]\!]_S$ is parallel to ν_S, i.e., $[\![\nabla u_H]\!]_S = j_S\, \nu_S$. Moreover, the coefficient matrix $\mathbf{A}(x)$ being continuous implies

$$J_S = \mathbf{A}(x)\, [\![\nabla u_H]\!]_S \cdot \nu_S = j_S\, \mathbf{A}(x) \nu_S \cdot \nu_S = a(x)\, j_S,$$

where $a(x) := \mathbf{A}(x)\nu_S \cdot \nu_S$ satisfies $0 < \underline{a}_S \le a(x) \le \overline{a}_S$ with $\underline{a}_S, \overline{a}_S$ the smallest and largest eigenvalues of $\mathbf{A}(x)$ on S. Consequently,

$$\left\|J_S\right\|_{L^2(S)}^2 \le \overline{a}_S^2 \int_S j_S^2 \le C \overline{a}_S^2 \int_S j_S^2 \psi_S \le C \frac{\overline{a}_S^2}{\underline{a}_S} \int_S (j_S\, \psi_S) J_S, \tag{4.4}$$

where the second inequality follows from j_S being a polynomial and $\psi_S > 0$ in a polygon of measure comparable with that of S.

We now extend j_S to ω_S by first mapping to the reference element, next extending constantly along the normal to $\hat{S}$, and finally mapping back to ω_S. The resulting extension $\mathsf{E}_h(j_S)$ is a piecewise polynomial of degree $\le n-1$ in ω_S so that $\psi_S \mathsf{E}_h(j_S) \in \mathbb{V}_h$, and satisfies $\|\psi_S \mathsf{E}_h(j_S)\|_{L^2(\omega_S)} \leq C H_S^{1/2} \|j_S\|_{L^2(S)}$. Since $v_h = \psi_S \mathsf{E}_h(j_S)$ is an admissible test function in (4.2) which vanishes on all sides of $\mathcal{S}_H$ but S, we arrive at

$$\begin{aligned}\int_S J_S(j_S\psi_S) &= \mathcal{B}[\varepsilon_H, v_h] - \int_{T_1} R_{T_1}\, v_h - \int_{T_2} R_{T_2}\, v_h \\ &\leq C\left(H_S^{-1/2}\|\varepsilon_H\|_{H^1_S(\omega_S)} + H_S^{1/2}\sum_{i=1}^{2}\|R_{T_i}\|_{L^2(T_i)}\right)\|j_S\|_{L^2(S)}.\end{aligned} \tag{4.5}$$

Therefore,

$$H_S\|J_S\|^2_{L^2(S)} \leq C\left(\|\varepsilon_H\|^2_{H^1(\omega_S)} + \sum_{i=1}^{2} H^2_{T_i}\|R_{T_i}\|^2_{L^2(T_i)}\right). \tag{4.6}$$

Step 3. *Final estimate.* To remove the interior residual from the right-hand side of (4.6) we observe that both T_1 and T_2 contain an interior node according to procedure REFINE. Hence, (4.3) implies

$$H_S\|J_S\|^2_{L^2(S)} \leq C\left(\|\varepsilon_H\|^2_{H^1(\omega_S)} + \sum_{i=1}^{2} H^2_{T_i}\left\|R_{T_i} - \overline{R_{T_i}}\right\|^2_{L^2(T_i)}\right). \tag{4.7}$$

The asserted estimate for $\eta_H(T)^2$ is thus obtained by adding this bound to (4.3). The constant C depends on the shape regularity constant γ^* and the ratio $\overline{a}_S^2/\underline{a}_S$ of largest and smallest eigenvalues of $\mathbf{A}(x)$ for $x \in S$. □

Remark 4.1 (positivity). The use of $\mathbf{A}(x)$ being positive definite in (4.4) avoids having oscillation terms on S. This comes at the expense of a constant depending on $\overline{a}_S^2/\underline{a}_S$. If we were to proceed in the usual manner, as in [1, 9, 13], we would end up with an oscillation of the form

$$\begin{aligned}H_S^{1/2}\|(\mathbf{A}-\mathbf{A}(x_S))\,[\![\nabla u_H]\!]_S\cdot\nu_S\|_{L^2(S)} &= H_S^{1/2}\|(a-a(x_S))j_S\|_{L^2(S)} \\ &\leq C H_S^{3/2}\|\mathbf{A}\|_{W^1_\infty(S)}\|j_S\|_{L^2(S)} \\ &\leq C H_S\left\|H_S^{1/2} J_S\right\|_{L^2(S)},\end{aligned}$$

where $C > 0$ also depends on the ratio $\overline{a}_S/\underline{a}_S$ dictated by the variation of $a(x)$ on S. This oscillation can be absorbed into the term $H_S^{1/2}\|J_S\|_{L^2(S)}$ provided that the meshsize H_S is sufficiently small; see [9]. We do not need this assumption in our present discussion.

Remark 4.2 (continuity of $\mathbf{A}$). The continuity of $\mathbf{A}$ is instrumental in avoiding jump oscillations, which in turn makes computations simpler. However, jump oscillations cannot be avoided when $\mathbf{A}$ exhibits discontinuities across interelement boundaries of the initial mesh. We get instead of (4.7)

$$CH_S\|J_S\|^2_{L^2(S)} \leq \|\varepsilon_H\|^2_{H^1(\omega_S)} + \sum_{i=1}^{2} H^2_{T_i}\left\|R_{T_i} - \overline{R_{T_i}}\right\|^2_{L^2(T_i)} + H_S\left\|J_S - \overline{J_S}\right\|^2_{L^2(S)}, \tag{4.8}$$

where $\overline{J_S}$ is the best L^2-projection of J_S onto $\mathbb{P}_{n-1}(S)$. To obtain estimate (4.8) we proceed as follows. Starting from a polynomial $\overline{J_S}$, we get an estimate similar to that of (4.4),

$$C\left\|\overline{J_S}\right\|^2_{L^2(S)} \leq \int_S \psi_S \overline{J_S}^2 = \int_S J_S(\psi_S\overline{J_S}) + \int_S (\overline{J_S} - J_S)(\psi_S\overline{J_S}). \tag{4.9}$$

In contrast to (4.4), we see that the oscillation term $(\overline{J_S} - J_S)$ cannot be avoided when $\mathbf{A}$ has a discontinuity across S. We estimate the first term on the right-hand side of (4.9) exactly as we have argued with (4.5) and thereby arrive at

$$\int_S J_S(\overline{J_S}\psi_S) \le C\left(H_S^{-1/2}\,\|\varepsilon_H\|_{H^1_S(\omega_S)} + H_S^{1/2}\sum_{i=1}^{2}\|R_{T_i}\|_{L^2(T_i)}\right)\left\|\overline{J_S}\right\|_{L^2(S)}.$$

This and a further estimate of the second term on the right-hand side of (4.9) yield

$$H_S\left\|\overline{J_S}\right\|^2_{L^2(S)} \le C\left(\|\varepsilon_H\|^2_{H^1(\omega_S)} + \sum_{i=1}^{2}H_{T_i}^2\,\|R_{T_i}\|^2_{L^2(T_i)} + H_S\left\|J_S - \overline{J_S}\right\|^2_{L^2(S)}\right),$$

whence assertion (4.8) follows using the triangle inequality for $\|J_S\|_{L^2(S)}$. Combining with (4.3), we deduce an estimate for $\eta_H(T)$ similar to (3.11), namely,

$$\eta_H(T)^2 \le C\left(\|\varepsilon_H\|^2_{H^1(\omega_T)} + \mathsf{osc}_H(\omega_T)^2\right)$$

with the new oscillation term involving jumps on interior sides

$$\mathsf{osc}_H(T)^2 := H_T^2\left\|R_T - \overline{R_T}\right\|^2_{L^2(T)} + \sum_{S\subset\partial T} H_S\left\|J_S - \overline{J_S}\right\|^2_{L^2(S)}. \tag{4.10}$$

In section 6.1, we discuss the case of a discontinuous $\mathbf{A}$. We show an oscillation reduction property of $\mathsf{osc}_H(T)$, defined by (4.10), similar to Lemma 3.2.

4.3. Proof of Lemma 3.2: Oscillation reduction. The proof hinges on the Marking Strategy O and the interior node property. We point out that if $T \in \mathcal{T}_h$ is contained in $T' \in \widehat{\mathcal{T}}_H$, then REFINE gives a reduction factor $\gamma_0 < 1$ of element size

$$h_T \le \gamma_0 H_{T'}. \tag{4.11}$$

The proof proceeds in three steps as follows.

Step 1. *Relation between oscillations.* We would like to relate $\mathsf{osc}_h(T')$ and $\mathsf{osc}_H(T')$ for any $T' \in \mathcal{T}_H$. To this end, we note that for all $T \in \mathcal{T}_h$ contained in T', we can write

$$R_T(u_h) = R_T(u_H) - \mathcal{L}_T(\varepsilon_H) \qquad \text{in } T,$$

where $\varepsilon_H = u_h - u_H$ as before and

$$\mathcal{L}_T(\varepsilon_H) := -\nabla\cdot(\mathbf{A}\nabla\varepsilon_H) + \mathbf{b}\cdot\nabla\varepsilon_H + c\,\varepsilon_H \qquad \text{in } T.$$

By Young's inequality, we have for all $\delta > 0$

$$\begin{aligned}\mathsf{osc}_h(T)^2 &= h_T^2\left\|R_T(u_h) - \overline{R_T(u_h)}\right\|^2_{L^2(T)}\\ &\le (1+\delta)h_T^2\left\|R_T(u_H) - \overline{R_T(u_H)}\right\|^2_{L^2(T)} + (1+\delta^{-1})h_T^2\left\|\mathcal{L}_T(\varepsilon_H) - \overline{\mathcal{L}_T(\varepsilon_H)}\right\|^2_{L^2(T)},\end{aligned}$$

where $\overline{R_T(u_h)}$, $\overline{R_T(u_H)}$, and $\overline{\mathcal{L}_T(\varepsilon_H)}$ are L^2-projections of $R_T(u_h)$, $R_T(u_H)$, and $\mathcal{L}_T(\varepsilon_H)$ onto polynomials of degree $\le n-1$ on T. We next observe that

$$\left\|\mathcal{L}_T(\varepsilon_H) - \overline{\mathcal{L}_T(\varepsilon_H)}\right\|_{L^2(T)} \le \|\mathcal{L}_T(\varepsilon_H)\|_{L^2(T)}$$

and that, according to (4.11),

$$h_T \le \gamma_{T'} H_{T'}$$

provided $\gamma_{T'} = \gamma_0$ if $T' \in \widehat{\mathcal{T}}_H$ and $\gamma_{T'} = 1$ otherwise. Therefore, if $\mathcal{T}_h(T')$ denotes all $T \in \mathcal{T}_h$ contained in T',

$$\begin{aligned} \mathsf{osc}_h(T')^2 &= \sum_{T\in\mathcal{T}_h(T')} \mathsf{osc}_h(T)^2 \\ &\le (1+\delta)\gamma_{T'}^2 \mathsf{osc}_H(T')^2 + (1+\delta^{-1}) \sum_{T\in\mathcal{T}_h(T')} h_T^2 \left\|\mathcal{L}_T(\varepsilon_H)\right\|_{L^2(T)}^2, \end{aligned} \tag{4.12}$$

since $R_T(u_H) = R_{T'}(u_H)$ and $\overline{R_T(u_H)}$ is the best L^2-approximation of $R_{T'}(u_H)$ in T.

Step 2. *Estimate of* $\mathcal{L}_T(\varepsilon_H)$. In order to estimate $\|\mathcal{L}_T(\varepsilon_H)\|_{L^2(T)}$ in terms of $\|\varepsilon_H\|_{H^1(T)}$, we first split it as follows:

$$\|\mathcal{L}_T(\varepsilon_H)\|_{L^2(T)} \le \|\nabla\cdot(\mathbf{A}\nabla\varepsilon_H)\|_{L^2(T)} + \|\mathbf{b}\cdot\nabla\varepsilon_H\|_{L^2(T)} + \|c\,\varepsilon_H\|_{L^2(T)}.$$

We denote these terms N_A, N_B, and N_C, respectively. Since

$$N_A \le \|(\nabla\cdot\mathbf{A})\cdot\nabla\varepsilon_H\|_{L^2(T)} + \|\mathbf{A} : H(\varepsilon_H)\|_{L^2(T)},$$

where $H(\varepsilon_H)$ is the Hessian of ε_H in T, invoking the Lipschitz continuity of $\mathbf{A}$ together with an inverse estimate in T, we infer that

$$N_A \le C_A \Big(\|\nabla\varepsilon_H\|_{L^2(T)} + h_T^{-1}\|\nabla\varepsilon_H\|_{L^2(T)}\Big),$$

where C_A depends on $\mathbf{A}$ and the shape regularity constant γ^*. Besides, we readily have

$$N_B \le C_B \|\nabla\varepsilon_H\|_{L^2(T)}, \qquad N_C \le C_C \|\varepsilon_H\|_{L^2(T)},$$

where C_B and C_C depend on $\mathbf{b}$ and c. Combining these estimates, we arrive at

$$h_T^2 \|\mathcal{L}_T(\varepsilon_H)\|_{L^2(T)}^2 \le C_* \|\varepsilon_H\|_{H^1(T)}^2. \tag{4.13}$$

Step 3. *Choice of* δ. We insert (4.13) into (4.12) and add over $T' \in \mathcal{T}_H$. Recalling the definition of $\gamma_{T'}$ and utilizing (3.10), we deduce

$$\begin{aligned} \sum_{T'\in\mathcal{T}_H} \gamma_{T'}^2 \mathsf{osc}_H(T')^2 &= \gamma_0^2 \sum_{T'\in\widehat{\mathcal{T}}_H} \mathsf{osc}_H(T')^2 + \sum_{T'\in\mathcal{T}_H\setminus\widehat{\mathcal{T}}_H} \mathsf{osc}_H(T')^2 \\ &= \mathsf{osc}_H(\Omega)^2 - (1-\gamma_0^2) \sum_{T'\in\widehat{\mathcal{T}}_H} \mathsf{osc}_H(T')^2 \\ &\le (1-(1-\gamma_0^2)\hat\theta^2)\mathsf{osc}_H(\Omega)^2, \end{aligned}$$

where $\hat\theta$ is the user's parameter in (3.10). Moreover, since $C_* \|\varepsilon_H\|_{H^1}^2 \le C_o \|\!|\varepsilon_H|\!\|^2$ with $C_o = C_* c_B^{-1}$ in light of (2.3), we end up with

$$\mathsf{osc}_h(\Omega)^2 \le (1+\delta)\big(1-(1-\gamma_0^2)\hat\theta^2\big)\mathsf{osc}_H(\Omega)^2 + (1+\delta^{-1})C_o \|\!|\varepsilon_H|\!\|^2.$$

To complete the proof, we finally choose δ sufficiently small so that

$$\rho_1 = (1+\delta)\big(1-(1-\gamma_0^2)\hat\theta^2\big) < 1, \qquad \rho_2 = (1+\delta^{-1})C_o. \qquad \square$$

5. Numerical experiments. We test the performance of the adaptive algorithm AFEM with several examples. We are thus able to study how meshes adapt to various effects from lack of regularity of solutions and convexity of domains to data smoothness, boundary layers, changing boundary conditions, etc. For simplicity, we stick to the case of piecewise linear finite element on polygonal domains in $\mathbb{R}^2$. The implementation is done using the ALBERT toolbox of Schmidt and Siebert [11, 12].

5.1. Implementation. We employ the four main procedures as given by Morin, Nochetto, and Siebert [7, 8]: SOLVE, ESTIMATE, MARK, and REFINE. We slightly modified the built-in adaptive solver for elliptic problems of ALBERT toolbox [11] to make it work for the general PDE (1.1) and mixed boundary conditions, as follows:

- SOLVE: We used built-in solvers provided by ALBERT, such as GMRES and CG.
- ESTIMATE: We modified ALBERT for computing the estimator so that it works for (1.1), and added procedures for computing oscillations which are not provided.
- MARK: We employed Marking Strategies E and O to find a marked set $\widehat{\mathcal{T}}_H$.
- REFINE: We employed the three newest bisections for each refinement step to enforce the interior node property.

Remark 5.1 (quadrature). Computations of integrals involving nonconstant functions $f, \mathbf{A}, \mathbf{b}, c, g$, and the exact solution u, use a quadrature rule of order 5. Our experiments indicate that increasing the quadrature order does not change the results. We refer to [3, 11, 12] for details on quadrature.

For convenience of presentation, we introduce the following notation:

- $\mathsf{DOF_k}$:= number of elements in $\mathcal{T}_k$;
- $\mathsf{EOC_e} := \frac{\log(\mathsf{e_{k-1}}/\mathsf{e_k})}{\log(\mathsf{DOF_k}/\mathsf{DOF_{k-1}})}$, experimental order of convergence, $e_k := |\!|\!| u - u_k |\!|\!|$;
- $\mathsf{EOC}_\eta := \frac{\log(\eta_{\mathsf{k}-1}/\eta_{\mathsf{k}})}{\log(\mathsf{DOF_k}/\mathsf{DOF_{k-1}})}$, experimental order of convergence of $\eta_k := \eta_k(\Omega)$;
- $\mathsf{RF_E} := \frac{\mathsf{e_k}}{\mathsf{e_{k-1}}}$ and $\mathsf{RF_O} := \frac{\mathsf{osc_k}}{\mathsf{osc_{k-1}}}$, reduction factors of the error and the oscillation;
- $\mathsf{Eff} := \eta_\mathsf{k}/\mathsf{e_k}$, effectivity index, i.e., the ratio between the estimator and the error;
- $\mathsf{M_E}$ and $\mathsf{M_O}$ are the number of marked elements due to Marking Strategy E and the additional marked elements due to Marking Strategy O, respectively.

The experimental order of convergence $\mathsf{EOC_e}$ measures how the error e_k decreases as $\mathsf{DOF_k}$ increases. In fact we have $e_k \approx C\ \mathsf{DOF_k}^{-\mathsf{EOC_e}}$.

5.2. Experiment 1: Oscillatory coefficients and nonconvex domain. We consider PDE (1.1) with the Dirichlet boundary condition $u = g$ on the nonconvex L-shape domain $\Omega := (-1,1)^2 \backslash [0,1] \times [-1,0]$. We also take the exact solution

$$u(r) = r^{\frac{2}{3}} \sin\left(\frac{2}{3}\theta\right),$$

where $r^2 := x^2 + y^2$ and $\theta := \tan^{-1}(y/x) \in [0, 2\pi)$. We deal with variable coefficients $\mathbf{A}(x,y) = a(x,y)\mathbf{I},\ \mathbf{b}(x,y) = \mathbf{0}$, and $c(x,y)$ defined by

$$a(x,y) = \frac{1}{4 + P(\sin(\frac{2\pi x}{\epsilon}) + \sin(\frac{2\pi y}{\epsilon}))}, \tag{5.1}$$

$$c(x,y) = A_c(\cos^2(lx) + \cos^2(lx)), \tag{5.2}$$

Table 5.1

Experiment 1 (*oscillatory coefficients and nonconvex domain*): *The parameters of AFEM are* $\theta = \hat{\theta} = 0.5$, *and those controlling the oscillatory coefficients are* $P = 1.8, \epsilon = 0.4, A_c = 4.0, l = 1.0$, *as described in* (5.1) *and* (5.2). *The experimental order of convergence* $\mathsf{EOC_e}$ *is close to the optimal rate of 0.5, which indicates quasi-optimal meshes. The oscillation reduction factor* $\mathsf{RF_O}$ *is smaller than the error reduction factor* $\mathsf{RF_E}$, *which confirms that oscillation decreases faster than error. The effectivity index* Eff *is approximately* 2.0. *There are no additional marked elements from oscillation for this* $\theta = 0.5$, *i.e.,* $\mathsf{M_O} = 0$. *However, this is not the case if* $\theta < 0.3$; *see section* 5.3.

k	$\mathsf{DOF_k}$	$\lVert u - u_k \rVert$	$\mathsf{EOC_e}$	$\mathsf{RF_E}$	$\mathsf{RF_O}$	Eff	$\mathsf{M_E}$	$\mathsf{M_O}$
–	24	2.181e-01	–	–	–	4.504	3	0
1	65	1.481e-01	0.388	0.679	0.446	2.994	10	0
2	229	1.056e-01	0.268	0.713	0.558	2.475	11	0
3	423	8.812e-02	0.295	0.834	0.652	2.222	13	0
4	651	5.083e-02	1.276	0.577	0.314	2.053	37	0
5	1156	3.305e-02	0.750	0.650	0.444	2.028	89	0
6	2299	2.206e-02	0.588	0.668	0.408	1.980	253	0
7	5148	1.445e-02	0.525	0.655	0.658	1.965	771	0
8	12678	7.991e-03	0.657	0.553	0.175	1.957	1833	0
9	29979	4.911e-03	0.566	0.615	0.426	2.032	–	–

Table 5.2

Experiment 1 (*oscillatory coefficients and nonconvex domain*): *Standard uniform refinement is performed using the same values for parameters* P, ϵ, A_c, *and* l *as that of AFEM given in Table* 5.1 *above.* $\mathsf{EOC_e}$ *is now suboptimal and close to the expected value* 1/3. *The effectivity index* Eff *is around* 2, *which is about the same as AFEM. We need about* 10^5 *DOFs to get the error around* 10^{-2}, *whereas for AFEM we need only* 10^4 *DOFs.*

$\mathsf{DOF_k}$	$\lVert u - u_k \rVert$	$\mathsf{EOC_e}$	$\mathsf{RF_E}$	$\mathsf{RF_O}$	Eff
384	1.005e-01	0.400	0.574	0.300	2.398
1536	4.809e-02	0.532	0.478	0.195	2.127
6144	2.597e-02	0.444	0.540	0.182	1.984
24576	1.551e-02	0.372	0.597	0.242	1.845
98304	9.585e-03	0.347	0.618	0.264	1.745

where P, ϵ, A_c, and l are parameters. The functions f in (1.1) and g are defined accordingly. The results are shown in Tables 5.1 and 5.2 and Figure 5.1. The observations and conclusions of this experiment are as follows:

- AFEM gives an optimal rate of convergence of order ≈ 0.5, while standard uniform refinement achieves the suboptimal rate of 0.3 as expected from theory.
- Both AFEM and FEM with uniform refinement perform with the effectivity index Eff ≈ 2.0, which gives the estimate of constant $C_1 \approx 0.5$ for upper bound (3.6); no weights have been used in (3.5). For AFEM, the reduction factors of error and oscillation are approximately 0.7 and 0.5 as DOF increases (Table 5.1). The oscillation thus decreases faster than the error and becomes insignificant asymptotically for k large. Additionally, AFEM outperforms FEM in terms of the CPU time vs. energy error.
- Figure 5.1 depicts the effect of a corner singularity and rapid variation of diffusion coefficient $a(x, y)$ in mesh grading; c does not play much of a role.
- The number of additional marked elements $\mathsf{M_O}$ due to Marking Strategy O depends on parameters θ and $\hat{\theta}$. For this example, $\mathsf{M_O} = 0$ because the parameter θ is sufficiently big; hence the condition for Marking Strategy O is automatically satisfied. Similar experiments for $\theta < 0.3$ and $\hat{\theta} = 0.5$ yield $\mathsf{M_O} \neq 0$, and $\mathsf{M_O}$ becomes even dominant for $\theta = 0.1$; see Experiment 2 for more details.

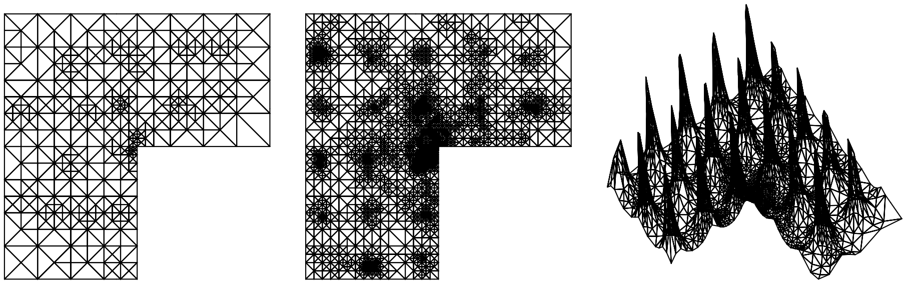


FIG. 5.1. *Experiment* 1 (*oscillatory coefficients and nonconvex domain*)*: Parameters of AFEM are* $\theta = \hat{\theta} = 0.5$, *and those of oscillatory coefficients are* $P = 1.8, \epsilon = 0.4, A_c = 1.0, l = 1.0$. *The sequence of graded meshes after* 4 *and* 7 *iterations shows that mesh refinement is dictated by geometric* (*corner*) *singularities as well as periodic variations of the diffusion coefficient but not much from the zero order term. Also on the right, a* 3*D plot of the diffusion coefficient* $a(x, y)$ *of* (5.1) *interpolated onto the mesh of iteration* 7 *is shown. This shows the combined effect of rapidly varying* $a(x, y)$ *and the exact solution* $u = r^{\frac{2}{3}} \sin(\frac{2}{3}\theta)$*: meshes are refined more where* $a(x, y)$ *has a large gradient.*

5.3. Experiment 2: Convection-dominated diffusion. We consider the convection-dominated diffusion elliptic model problem (1.1) with the Dirichlet boundary condition $u = g$ on the convex domain $\Omega := (0, 1)^2$, with the isotropic diffusion coefficient $\mathbf{A} = \epsilon \mathbf{I}$, $\epsilon = 10^{-3}$, convection velocity $\mathbf{b} = (y, \frac{1}{2} - x)$, and $c = f = 0$; note that $\nabla \cdot \mathbf{b} = 0$. The Dirichlet boundary condition $g(x, y)$ on $\partial\Omega$, a pulse, is the continuous piecewise linear function given by

$$g(x,y) = \begin{cases} 1, & \{.2 + \tau \leq x \leq .5 - \tau;\ y = 0\}, \\ 0, & \partial\Omega \setminus \{.2 \leq x \leq .5;\ y = 0\}, \\ \text{linear}, & \{(.2 \leq x \leq .2 + \tau) \text{ or } (.5 - \tau \leq x \leq .5); y = 0\}, \end{cases} \tag{5.3}$$

where τ is a parameter. This problem models the transport of a pulse from $\partial\Omega$ inside Ω and back to $\partial\Omega$. Results are reported in Table 5.3 and Figures 5.2 and 5.3 for parameters $\theta = 0.3, \hat{\theta} = 0.6, \tau = 0.005$, starting from a coarser mesh than what we would need in theory. To see whether oscillation plays any role in AFEM, Table 5.4 shows results of AFEM without using Marking Strategy O. Observations and conclusions follow:

- Tables 5.3 and 5.4 document the role of oscillation in AFEM. Without marking due to the oscillation $\mathsf{M_O} = \mathsf{0}$, the estimator $\eta(\Omega)$ still reduces at an optimal rate but the oscillation reduction $\mathsf{RF_O}$ is not stable. The factor $\mathsf{RF_O}$ approximates ρ_1 of Lemma 3.2 and thus controls the oscillation decay between consecutive iterations. In fact Table 5.4 indicates that lack of control of $\mathsf{RF_O}$ leads to more iterations for the same estimator. Tables 5.3 and 5.4 illustrate the need of Marking Strategy O to control the reduction rate of oscillations and confirm the convergence theory of AFEM. Our experiments show that the ratio $\mathsf{M_E}/\mathsf{M_O}$ depends inversely on the ratio $\theta/\hat{\theta}$. If $\theta = \hat{\theta}$, then $\mathsf{M_E}$ dominates $\mathsf{M_O}$.
- Comparison of computational costs is measured using the CPU time used by each procedure. On average, about 80% of the total CPU time is used by SOLVE; the other procedures ESTIMATE, MARK, and REFINE use about 5–10%.

Table 5.3

Experiment 2: AFEM with parameters $\theta = 0.3, \hat{\theta} = 0.6$, and $\tau = 0.005$. The optimal decay ≈ 0.5 of the estimator $\eta(\Omega)$ is computational evidence of optimal meshes. The reduction factor of oscillation $\mathsf{RF_O} := \mathsf{osc}_k/\mathsf{osc}_{k-1}$ gives an estimate of constant $\rho_1 \approx 0.5$ in Lemma 3.2. In contrast to Experiment 1, the additional marking $\mathsf{M_O}$ due to oscillation dominates $\mathsf{M_E}$ from Marking Strategy E. This controls $\mathsf{RF_O}$, the decay of oscillations, which decrease together with the error according to Theorem 1.1.

$\mathsf{DOF_k}$	$\eta_k(\Omega)$	EOC_η	$\mathsf{RF_O}$	$\mathsf{M_E}$	$\mathsf{M_O}$
64	1.74e-1	–	–	2	5
147	9.48e-2	0.73	0.27	8	7
360	2.35e-2	1.55	0.33	4	9
500	1.68e-2	1.02	0.50	5	15
762	1.12e-2	0.95	0.43	10	23
1170	8.58e-3	0.62	0.52	15	70
2173	6.10e-3	0.55	0.48	22	137
3862	4.75e-3	0.43	0.48	30	298
7149	3.45e-3	0.51	0.50	80	600
13981	2.60e-3	0.42	0.51	–	–

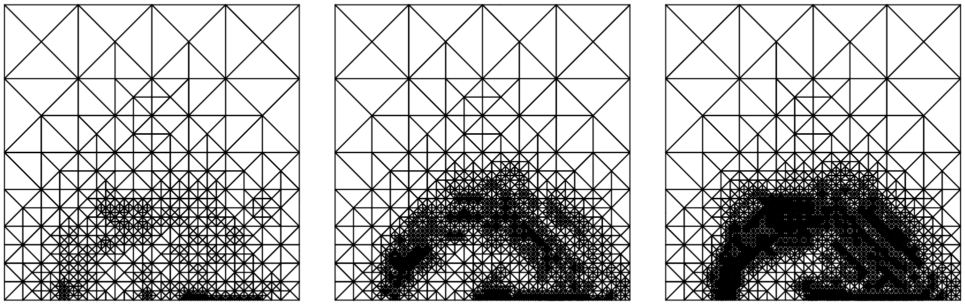


Fig. 5.2. *Experiment 2 (convection-dominated diffusion with $\epsilon = 10^{-3}, \mathbf{b} = (y, \frac{1}{2} - x)$): Adaptively refined meshes after 5, 7, and 8 iterations corresponding to Table 5.3 starting from a uniform mesh coarser than required in theory. After a few iterations, AFEM detects the region of rapid variation (circular transport of a pulse) and boundary layer in the outflow, whereas the rest of the mesh remains unchanged. Refinement in the smooth region is caused by early oscillations.*

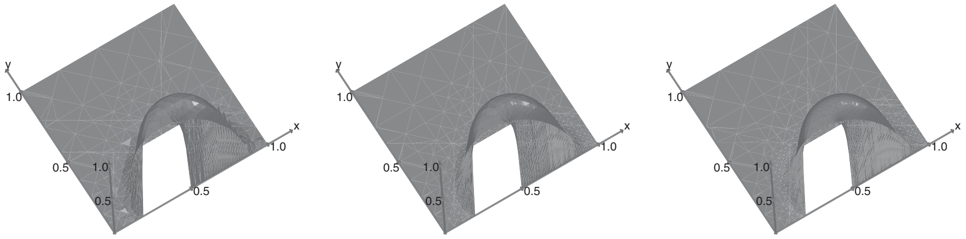


Fig. 5.3. *Experiment 2 (convection-dominated diffusion with $\epsilon = 10^{-3}, \mathbf{b} = (y, \frac{1}{2} - x)$): Plots of solutions after 5, 7, and 8 iterations. No oscillations (of Galerkin solutions) are detected after a few iterations even though AFEM is not stabilized.*

- In theory, the initial meshsize h_0 must satisfy

$$C^* B h_0 < \frac{\beta \Lambda_1}{1 + \beta \Lambda_1} = \beta_0,$$

where $B = \|\mathbf{b}\|_{L^\infty}$, $\beta_0 = O(1)$, and C^* is the constant from Lemma 2.1. In this particular case, we can express C^* in terms of ϵ and B quite explicitly.

Table 5.4

Experiment 2: AFEM performance without Marking Strategy O, using the same parameters as for Table 5.3. *The reduction factor of oscillation* $\mathsf{RF_O}$ *is not as stable as our AFEM shown in Table* 5.3. *The estimator still reduces at the optimal rate but requires a few more iterations to reach the same level as that of our AFEM.*

$\mathsf{DOF_k}$	$\eta_k(\Omega)$	EOC_η	$\mathsf{RF_O}$
64	1.74e-1	–	–
95	1.02e-1	1.34	0.59
244	3.81e-2	1.31	0.86
414	1.75e-2	4.09	0.62
654	9.42e-3	1.18	0.70
834	9.05e-3	0.16	0.59
1577	5.43e-3	0.89	0.93
2970	3.56e-3	0.51	0.92
4250	2.84e-3	0.62	0.82
6502	2.15e-3	0.65	0.59
10209	1.66e-3	0.57	0.62

We first observe that the H^2-regularity theory gives [5]

$$\begin{cases} \mathcal{L}\varphi = \zeta & \text{in } \Omega \\ \quad \varphi = 0 & \text{on } \partial\Omega \end{cases} \implies \|\varphi\|_{H^2(\Omega)} \leq C_R B^{1/2} \epsilon^{-3/2} \|\zeta\|_{L^2(\Omega)}$$

with $C_R > 0$ independent of data. We also note that C_D of Lemma 4.1 satisfies

$$C_I C_R \left(\frac{B}{\epsilon}\right)^{\frac{3}{2}} h_0 \leq \frac{1}{2} \implies C_D = 2 C_I C_R \left(\frac{B}{\epsilon}\right)^{\frac{1}{2}},$$

where C_I is an interpolation constant solely dependent on shape regularity. This results from the usual duality argument and the fact that $\nabla \cdot \mathbf{b} = 0$, namely,

$$|\langle e_h, \zeta \rangle| = |\mathcal{B}[e_h, \varphi]| \leq C_I h_0 \left(\epsilon \|\nabla e_h\|_{L^2} + B \|e_h\|_{L^2}\right) \|\varphi\|_{H^2}.$$

We finally recall that $C^* = C_D/\epsilon$ (see section 4.1), to arrive at

$$h_0 < \frac{\beta_0}{2 C_I C_R} \left(\frac{\epsilon}{B}\right)^{3/2},$$

which is consistent with the previous restriction on h_0. We stress that this implies $h_0 \approx 10^{-4}$ in theory, whereas $h_0 \approx 10^{-1}$ works in our examples; see Figures 5.2 and 5.3.

- The local Péclet number $P_e = \frac{h_0 B}{\epsilon}$ is about 10^2 at the beginning. Since $P_e > 1$, and the Galerkin method is not stabilized, oscillations are observed in the first few iterations but cured later by AFEM via local refinement; see Figure 5.3, which displays solutions without oscillations for iterations 7 and 8. Figure 5.2 depicts several graded meshes and confirms that mesh refinement is localized around the pulse location and outflow boundary layer. Minor refinement in the smooth region is caused by early oscillations.

Table 5.5

Experiment 3 *(drift-diffusion model): Performance of AFEM with the parameters $\theta = 0.6$, $\hat{\theta} = 0.75$ and model parameters $\chi = 10$, $r_1 = 0.75$, and $\alpha = 0.04$. The optimal decay ≈ 0.5 of the estimator $\eta(\Omega)$ is computational evidence of quasi-optimal meshes. AFEM outperforms uniform refinement (compare with Table* 5.6*).*

$\mathsf{DOF_k}$	$\eta_k(\Omega)$	EOC_η	$\mathsf{RF_O}$
1154	6.645	1.880	0.267
1546	3.824	1.888	0.252
2448	2.144	1.259	0.206
4032	1.455	0.776	0.285
6790	1.086	0.560	0.340
12188	0.737	0.663	0.253
23386	0.518	0.540	0.287
45728	0.363	0.529	0.261

5.4. Experiment 3: Drift-diffusion model. We consider a model problem that comes from a mathematical model in semiconductors and chemotaxis:

$$\begin{aligned} -\nabla\cdot(\nabla u + \chi u \nabla\psi) &= 0 \qquad && \text{in } \Omega := (0,1)^2, \\ u &= g && \text{on } \Gamma \subset \partial\Omega, \\ \partial_\nu u &= 0 && \text{on } \partial\Omega \setminus \Gamma, \end{aligned}$$

where χ is a constant. The radial function ψ is defined in Ω by

$$\psi(x,y) := \begin{cases} 1, & \{\sqrt{x^2+y^2} \le r_1\}, \\ \alpha, & \{\sqrt{x^2+y^2} \ge r_1 + \alpha\}, \\ \text{linear}, & \{r_1 < \sqrt{x^2+y^2} < r_1 + \alpha\}, \end{cases}$$

where α is a small parameter and $r_1 < 1$ is a constant. The Dirichlet boundary condition on Γ is assumed to be

$$g(x,y) = \begin{cases} 1, & \{x = 0; 0 \le y \le 0.5\} \bigcup \{y = 0; 0 \le x \le 0.5\}, \\ 0, & \{x = 1; 0.5 \le y \le 1\} \bigcup \{y = 1; 0.5 \le x \le 1\}. \end{cases}$$

We resort to the following transformation (exponential fitting) to symmetrize the problem:

$$\rho := \exp(\chi\psi)u \implies -\nabla\cdot(\exp(-\chi\psi)\nabla\rho) = 0,$$

which gives a simpler form of the model problem with a variable scalar coefficient $a = \exp(-\chi\psi)$. We apply AFEM to solve for ρ and obtain solution u via $u = \exp(-\chi\psi)\rho$. The experiment is performed using the parameters $\chi = 10.0, r_1 = 0.75$, and $\alpha = 0.04$ for the model problem, and parameters $\theta = 0.6$, $\hat{\theta} = 0.75$ for AFEM. Results are reported in Tables 5.5 and 5.6, and Figure 5.4. Conclusions and observations follow:

- From Tables 5.5 and 5.6 we see again that AFEM outperforms FEM with standard uniform refinement. Since the decay of the estimator $\eta(\Omega)$ is optimal, we have computational evidence of optimal meshes.
- Figure 5.4 displays a discrete solution u_8 and graded meshes after 8 and 10 iterations; note the drastic variation of u_8 across the annulus $r_1 < r < r_1 + \alpha$. Meshes adapt well to lack of smoothness, namely, refinement concentrates in the transition layer, where $\nabla\psi$ does not vanish, and at the midpoints of boundary sides, where boundary conditions change.

Table 5.6

Experiment 3 *(drift-diffusion model): Performance of FEM with uniform refinement and the same parameters* χ, r_1, *and* α *as for AFEM given in Table* 5.5. *To have the estimator around* 0.9, *uniform refinement needs about* 65,000 *DOFs, whereas AFEM needs only around* 10,000 *DOFs.*

$\mathsf{DOF_k}$	$\eta_k(\Omega)$	EOC_η	$\mathsf{RF_O}$
1024	179.831	3.186	0.009
2048	30.769	2.547	0.026
4096	11.031	1.479	0.096
8192	3.983	1.469	0.106
16384	2.173	0.874	0.188
32768	1.296	0.745	0.216
65536	0.874	0.567	0.250

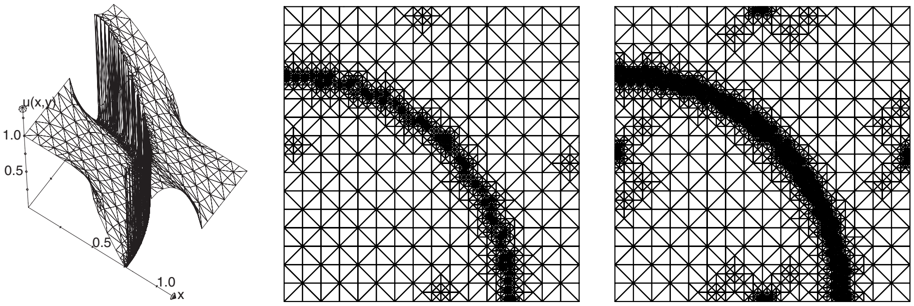


Fig. 5.4. *Experiment* 3 *(drift-diffusion model): Discrete solution* u_8 *and refined meshes after* 8 *and* 10 *iterations. Mesh grading is quite pronounced in the internal layer where* $\nabla\psi$ *does not vanish, and at the midpoints of the boundary sides, where boundary conditions change. The solution* $u(x,y)$ *has a thin transition layer where* $\nabla\psi \neq 0$, *and meshes are highly refined there.*

6. Extensions. We extend the model problem (1.1) by considering now $\mathbf{A}$ with discontinuities aligned with the initial mesh and a nondivergence-free $\mathbf{b}$. Note that if $\nabla\cdot\mathbf{b} \neq 0$, then the bilinear form $\mathcal{B}$ may be noncoercive if $c - \frac{1}{2}\nabla\cdot\mathbf{b} \not\geq 0$.

6.1. Discontinuous A. We first observe that Lemma 4.1, and thus Lemma 2.1, still holds because the regularity H^{1+s} required in the duality argument is valid; see [6] for example. The continuity of $\mathbf{A}$ is used instead for obtaining error and oscillation reduction estimates (Lemmas 3.1 and 3.2) in that the element oscillation $\mathsf{osc}_H(T)$ does not involve oscillation of the jump residual on ∂T. Remark 4.2 shows that when $\mathbf{A}$ has discontinuities across element faces, we still obtain the error reduction estimate (3.11) of Lemma 3.1, but this time the oscillation is defined by (4.10) and involves oscillation of the jump residual. To prove convergence it suffices to show the oscillation reduction estimate (3.12), for the new concept of element oscillation, namely, $\mathsf{osc}_H(T)^2 = \mathsf{osc}_{R,H}(T)^2 + \sum_{S\subset\partial T}\mathsf{osc}_{J,H}(S)^2$ with

$$\mathsf{osc}_{R,H}(T)^2 := H_T^2 \left\| R_T(u_H) - \overline{R_T(u_H)} \right\|_{L^2(T)}^2 \qquad \forall\, T \in \mathcal{T}_H,$$

$$\mathsf{osc}_{J,H}(S)^2 := H_S \left\| J_S(u_H) - \overline{J_S(u_H)} \right\|_{L^2(S)}^2 \qquad \forall\, S \in \mathcal{S}_H.$$

We proceed in three steps as follows:

Step 1. *Oscillation of interior residual.* Invoking the same arguments as in the proof of Lemma 3.2 in section 4.3, we obtain an oscillation reduction estimate for the interior residual

$$\mathsf{osc}_{R,h}(T')^2 \leq (1+\delta)\gamma_{T'}^2 \mathsf{osc}_{R,H}(T')^2 + C_*(1+\delta^{-1}) \|\varepsilon_H\|_{H^1(T')}^2 \qquad \forall\, T' \in \mathcal{T}_H,$$

where $\mathsf{osc}_{R,h}(T')$ is defined to be $\mathsf{osc}_h(T')$ in (4.12).

Step 2. *Oscillation of jump residual.* To obtain an estimate for $\mathsf{osc}_{J,h}(S)$ we write

$$J_S(u_h) = \gamma_S \left[\!\left[\mathbf{A}\nabla u_H\right]\!\right]_S \cdot \nu_S + \left[\!\left[\mathbf{A}\nabla \varepsilon_H\right]\!\right]_S \cdot \nu_s = \gamma_S J_S(u_H) + J_S(\varepsilon_H),$$

where $\gamma_S = 1$ if $S \subset S' \in \mathcal{S}_H$ and $\gamma_S = 0$ otherwise, since $\mathbf{A}\nabla u_H$ is continuous on S in the second case. Using Young's inequality, we have for all $\delta > 0$

$$\begin{aligned}\mathsf{osc}_{J,h}(S)^2 \leq (1+\delta)\gamma_S h_S \left\| J_S(u_H) - \overline{J_S(u_H)} \right\|_{L^2(S)}^2 \\ + (1+\delta^{-1}) h_S \left\| J_S(\varepsilon_H) - \overline{J_S(\varepsilon_H)} \right\|_{L^2(S)}^2,\end{aligned}$$

where $\overline{J_S(u_H)}$ and $\overline{J_S(\varepsilon_H)}$ are L^2-projections of $J_S(u_H)$ and $J_S(\varepsilon_H)$ onto $\mathbb{P}_{n-1}(S)$. For the second term we observe that

$$\begin{aligned}\left\| J_S(\varepsilon_H) - \overline{J_S(\varepsilon_H)} \right\|_{L^2(S)} &\leq \|J_S(\varepsilon_H)\|_{L^2(S)} = \left\| \left[\!\left[\mathbf{A}\nabla\varepsilon_H\right]\!\right]_S \cdot \nu_S \right\|_{L^2(S)} \\ &\leq \left\|\mathbf{A}^+ \nabla \varepsilon_H^+ \cdot \nu_S\right\|_{L^2(S)} + \left\|\mathbf{A}^- \nabla \varepsilon_H^- \cdot \nu_S\right\|_{L^2(S)} \\ &\leq \|\mathbf{A}\|_{L^\infty(\omega_S)} \left(\left\|\nabla\varepsilon_H^+\right\|_{L^2(S)} + \left\|\nabla\varepsilon_H^-\right\|_{L^2(S)} \right) \\ &\leq C_A h_S^{-1/2} \|\varepsilon_H\|_{H^1(\omega_S)},\end{aligned}$$

where C_A depends on $\mathbf{A}$ and the shape regularity constant γ^*. For simplicity, let $\mathcal{S}_h(T')$ denote all $S \in \mathcal{S}_h$ contained in $T' \in \mathcal{T}_H$; hence

$$\begin{aligned}\mathsf{osc}_{J,h}(T')^2 &= \sum_{S \in \mathcal{S}_h(T')} \mathsf{osc}_{J,h}(S)^2 \\ &\leq (1+\delta) \sum_{S \in \mathcal{S}_h(T')} \gamma_S h_S \left\| J_S(u_H) - \overline{J_S(u_H)} \right\|_{L^2(S)}^2 + (1+\delta^{-1}) C_A \|\varepsilon_H\|_{H^1(\omega_{T'})}^2.\end{aligned}$$

In light of the reduction factor of the element size $h_S \leq \gamma_{T'} H_{S'}$, and definitions of γ_S and $\gamma_{T'}$, we obtain

$$\begin{aligned}\sum_{S \in \mathcal{S}_h(T')} \gamma_S h_S \left\| J_S(u_H) - \overline{J_S(u_H)} \right\|_{L^2(S)}^2 &\leq \gamma_{T'} \sum_{S' \in \mathcal{S}_H(T')} H_{S'} \left\| J_{S'}(u_H) - \overline{J_{S'}(u_H)} \right\|_{L^2(S')}^2 \\ &= \gamma_{T'} \mathsf{osc}_{J,H}(T')^2,\end{aligned}$$

because for $S \subset S' \subset \partial T'$, we have $J_S(u_H) = J_{S'}(u_H)$ and $\overline{J_S(u_H)}$ is the best L^2-approximation of $J_S(u_H)$ on S. Therefore,

$$\mathsf{osc}_{J,h}(T')^2 \leq (1+\delta)\gamma_{T'} \mathsf{osc}_{J,H}(T')^2 + (1+\delta^{-1}) C_A \|\varepsilon_H\|_{H^1(\omega_{T'})}^2 \qquad \forall\, T' \in \mathcal{T}_H.$$

Step 3. *Choice of* δ. Combining results from Steps 1 and 2 above using $\gamma_{T'} \leq 1$, $C_{**} = \max\{C_*, C_A\}$, and the definition of $\mathsf{osc}_h(T)$, we arrive at

$$\mathsf{osc}_h(T')^2 \leq (1+\delta)\gamma_{T'} \mathsf{osc}_H(T')^2 + C_{**}(1+\delta^{-1}) \|\varepsilon_H\|_{H^1(\omega_{T'})}^2.$$

Proceeding as in Step 3 of the proof of Lemma 3.2, this time with Marking Strategy O performed according to the new definition of $\mathsf{osc}_H(T)$, we arrive at

$$\mathsf{osc}_h(\Omega)^2 \le (1+\delta)(1-(1-\gamma_0)\hat\theta^2)\mathsf{osc}_H(\Omega)^2 + C_o(1+\delta^{-1}) \left\|\kern-0.25ex\left|\varepsilon_H\right|\kern-0.25ex\right\|^2$$

with $C_o = C_{**}c_B^{-1}$. The assertion thus follows by choosing δ sufficiently small so that

$$\rho_1 := (1+\delta)(1-(1-\gamma_0)\hat\theta^2) < 1, \qquad \rho_2 := C_o(1+\delta^{-1}).$$

6.2. Noncoercive $\boldsymbol{\mathcal{B}}$. In this section, we prove convergence of AFEM for the case $c - \frac{1}{2}\nabla\cdot\mathbf{b} \not\geq 0, c \ge 0$; the case $c < 0$ can be treated as well. According to what we have so far, the assumption of $\nabla\cdot\mathbf{b} = 0$ is used for proving quasi-orthogonality and for having equivalence between energy norm $|||v|||^2 := \mathcal{B}[v,v]$ and H^1-norm as in (2.3), where $\mathcal{B}$ is coercive. Since now $\mathcal{B}$ may be noncoercive, we cannot define the energy norm in this manner. We instead define the energy norm by $|||v|||^2 := \int_\Omega \mathbf{A}\nabla v\cdot\nabla v + c\, v^2$, and we have equivalence of norms

$$c_E \left\|v\right\|^2_{H^1(\Omega)} \le |||v|||^2 \le C_E \left\|v\right\|^2_{H^1(\Omega)}, \tag{6.1}$$

where constants c_E and C_E depend only on the data $\mathbf{A}$, c, and Ω. The lack of coercivity is now replaced by Gårding's inequality

$$|||v|||^2 - \gamma_G \left\|v\right\|^2_{L^2(\Omega)} \le \mathcal{B}[v,v] \qquad \forall\, v \in H^1_0(\Omega), \tag{6.2}$$

where $\gamma_G = \left\|\nabla\cdot\mathbf{b}\right\|_\infty/2$. To see this we integrate by parts the middle term of $\mathcal{B}[v,v]$,

$$\int_\Omega \mathbf{b}\cdot\nabla v\, v = \frac{1}{2}\int_\Omega \mathbf{b}\cdot\nabla(v^2) = -\int_\Omega \frac{\nabla\cdot\mathbf{b}}{2} v^2 \qquad \forall\, v \in H^1_0(\Omega).$$

The same calculation leads to the sharp upper bound for $\mathcal{B}[v,v]$:

$$\mathcal{B}[v,v] \le |||v|||^2 + \gamma_G \left\|v\right\|^2_{L^2(\Omega)} \qquad \forall\, v \in H^1_0(\Omega). \tag{6.3}$$

Existence and uniqueness of weak solutions follows from the maximum principle for $c \ge 0$ [5]. Schatz showed in [10] that the discrete problem (2.5) has a unique solution if the meshsize h is sufficiently small, i.e., $h \le h^*$ for some constant h^* depending on the shape regularity and data but not computable; the results in [10] are also valid for graded meshes. Assuming $h_0 \le h^*$, to prove convergence of AFEM it thus suffices to prove quasi-orthogonality. We follow the steps of Lemma 2.1.

Using the same notation as in section 4 for e_h, e_H, and ε_H, expanding $\mathcal{B}[e_H, e_H]$, and noticing that $e_H = e_h + \varepsilon_H$ and $\mathcal{B}[e_h, \varepsilon_H] = 0$, we arrive at

$$\mathcal{B}[e_h, e_h] = \mathcal{B}[e_H, e_H] - \mathcal{B}[\varepsilon_H, \varepsilon_H] - \mathcal{B}[\varepsilon_H, e_h], \tag{6.4}$$

where this time integration by parts yields

$$\begin{aligned}\mathcal{B}[\varepsilon_H, e_h] &= \mathcal{B}[e_h, \varepsilon_H] + \langle \mathbf{b}\cdot\nabla\varepsilon_H, e_h\rangle - \langle \mathbf{b}\cdot\nabla e_h, \varepsilon_H\rangle \\ &= 2\langle \mathbf{b}\cdot\nabla\varepsilon_H, e_h\rangle + \langle \nabla\cdot\mathbf{b}\, e_h, \varepsilon_H\rangle.\end{aligned}$$

Consequently, using the Cauchy–Schwarz inequality and (6.1), we have for all $\delta > 0$

$$|\mathcal{B}[\varepsilon_H, e_h]| \le (2\left\|\mathbf{b}\right\|_\infty \left\|\nabla\varepsilon_H\right\|_{L^2} + \left\|\nabla\cdot\mathbf{b}\right\|_\infty \left\|\varepsilon_H\right\|_{L^2}) \left\|e_h\right\|_{L^2} \le C_b^2\delta\, |||\varepsilon_H|||^2 + \delta^{-1}\left\|e_h\right\|^2_{L^2},$$

where constant $C_b = \max\{2\|\mathbf{b}\|_\infty, \|\nabla\cdot\mathbf{b}\|_\infty\} c_E^{-1}/2$.

Using (6.2) and (6.3) to estimate terms $\mathcal{B}[e_h, e_h]$, $\mathcal{B}[e_H, e_H]$, and $\mathcal{B}[\varepsilon_H, \varepsilon_H]$ in (6.4), and combining with the previous estimate, we infer that

$$\|\!|e_h|\!\|^2 - (\gamma_G + \delta^{-1})\|e_h\|_{L^2}^2 \le \|\!|e_H|\!\|^2 + \gamma_G\|e_H\|_{L^2}^2 - (1 - C_b^2\delta)\|\!|\varepsilon_H|\!\|^2 + \gamma_G\|\varepsilon_H\|_{L^2}^2 .$$

Since $\|\varepsilon_H\|_{L^2}^2 \le 2\|e_h\|_{L^2}^2 + 2\|e_H\|_{L^2}^2$, estimates for $\|e_h\|_{L^2}$ and $\|e_H\|_{L^2}$ of the form (4.1), obtained via duality, with $C_6 := \frac{C_D}{\sqrt{c_E}}$ imply

$$\Lambda_h \ \|\!|e_h|\!\|^2 \le \Lambda_H \|\!|e_H|\!\|^2 - \Lambda_\varepsilon \|\!|\varepsilon_H|\!\|^2 , \tag{6.5}$$

where $\Lambda_h = 1 - C_6^2 h_0^{2s}(3\gamma_G + \delta^{-1})$, $\Lambda_H = 1 + 3\gamma_G C_6^2 h_0^{2s}$, and $\Lambda_\varepsilon = 1 - C_b^2\delta$.

Consequently, to get $\Lambda_h = \Lambda_\varepsilon$, we choose δ depending on h_0 so that

$$\delta(h_0) = \frac{C_G h_0^{2s} + \sqrt{C_G^2 h_0^{4s} + 4C_b^2 C_6^2 h_0^{2s}}}{2C_b^2} > 0,$$

where $C_G = 3\gamma_G C_6^2$. We further choose h_0 sufficiently small so that $C_b^2\delta(h_0) < 1$, whence $\Lambda_h = \Lambda_\varepsilon > 0$. This can be achieved for $h_0^s \le \min\left\{C_6 C_b C_G^{-1}, (3C_6C_b)^{-1}\right\}$ because

$$\begin{aligned} C_b^2\delta(h_0) &= \frac{C_G}{2}h_0^{2s} + C_b C_6 h_0^s \sqrt{1 + h_0^{2s} C_G^2 (4C_b^2 C_6^2)^{-1}} \\ &\le 2C_b C_6 h_0^s \big(1 + h_0^s C_G (4C_b C_6)^{-1}\big) < 3C_b C_6 h_0^s \le 1. \end{aligned}$$

We conclude that if the meshsize h_0 of the initial mesh satisfies

$$h_0^s \le \min\left\{C_6 C_b C_G^{-1}, (3C_6C_b)^{-1}, (h^*)^s\right\}, \tag{6.6}$$

then quasi-orthogonality holds, i.e., for $\Lambda_0 := \Lambda_H/\Lambda_h$,

$$\|\!|e_h|\!\|^2 \le \Lambda_0 \|\!|e_H|\!\|^2 - \|\!|\varepsilon_H|\!\|^2 , \tag{6.7}$$

and Λ_0 can be made arbitrarily close to 1 by decreasing h_0. Convergence of AFEM finally follows as in Theorem 1.1.

6.3. Experiment 4: Noncoercive $\boldsymbol{\mathcal{B}}$. We repeat Experiment 2 in section 5.3 with $\mathbf{b} = (x-1, y+1)$, and thus $\mathcal{B}$ is noncoercive because $c - \frac{1}{2}\nabla\cdot\mathbf{b} = -1$. For a better view of solutions we change the boundary condition $g(x,y)$ to be 1 on the x-axis from $(.4+\tau)$ to $(.8-\tau)$, with τ defined as in (5.3). Results of AFEM with $\theta = \hat{\theta} = 0.5, \tau = 0.005$ are reported in Figure 6.1. Observations and conclusions follow:

- Figure 6.1 shows oscillations of the Galerkin solution near internal and boundary layers after 4 iterations. AFEM detects this effect and corrects it after 6 iterations by selective local refinement which does not spread in regions of smoothness.
- The resulting graded meshes are optimal and capture internal layers (diffuse boundary of pulse g being transported) and the outflow boundary layer, even though the initial uniform mesh is far coarser than required by theory; see (6.6) which is a restriction similar to that discussed in section 5.3 (Experiment 2). Moreover, the performance of AFEM as to the estimator decay and oscillation control is analogous to section 5.3 (Experiment 2).

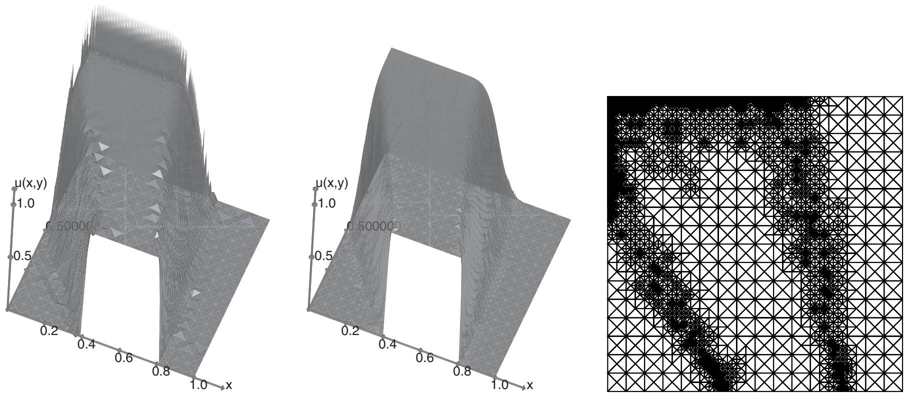


FIG. 6.1. *Experiment* 4 *(noncoercive $\mathcal{B}$ with $\epsilon = 10^{-3}, \mathbf{b} = (x-1, y+1)$): Three-dimensional plots of solutions after* 4 *and* 6 *iterations and graded mesh after* 6 *iterations. Oscillations of Galerkin solutions are observed near internal and boundary layers in a first few iterations but AFEM eliminates them after* 6 *iterations.*

Acknowledgments. We would like to thank the four referees for their careful reading and constructive comments and suggestions.

REFERENCES

[1] M. Ainsworth and J.T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, New York, 2000.

[2] Z. Chen and F. Jia, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp., 73 (2004), pp. 1163–1197.

[3] Ph. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978; reprinted, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[4] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

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

[6] R.B. Kellogg, *On the Poisson equation with intersecting interfaces*, Appl. Anal., 4 (1975), pp. 101–129.

[7] P. Morin, R.H. Nochetto, and K.G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[8] P. Morin, R.H. Nochetto, and K.G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.

[9] R.H. Nochetto, *Removing the saturation assumption in a posteriori error analysis*, Istit. Lombardo Accad. Sci. Lett. Rend. A, 127 (1993), pp. 67–82.

[10] A.H. Schatz, *An observation concerning Ritz–Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.

[11] A. Schmidt and K.G. Siebert, *ALBERT: An Adaptive Hierarchical Finite Element Toolbox*, Documentation, Preprint 06/2000, Universität Freiburg, Freiburg, Germany, 2000.

[12] A. Schmidt and K.G. Siebert, *ALBERT—Software for scientific computations and applications*, Acta Math. Univ. Comenian., 70 (2001), pp. 105–122.

[13] R. Verfürth, *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Technique*, Wiley–Teubner, Chichester, UK, 1996.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 1828–1849

ENTROPIC DISCRETIZATION OF A QUANTUM DRIFT-DIFFUSION MODEL*

SAMY GALLEGO[†] AND FLORIAN MÉHATS[†]

Abstract. This paper is devoted to the discretization and numerical simulation of a new quantum drift-diffusion model that was recently derived. In a first step, we introduce an implicit semidiscretization in time which possesses some interesting properties: this system is well-posed, it preserves the positivity of the density, the total charge is conserved, and it is entropic (a free energy is dissipated). Then, after a discretization of the space variable, we define a numerical scheme which has the same properties and is equivalent to a convex minimization problem. These results are illustrated by some numerical simulations.

Key words. quantum drift-diffusion, Schrödinger–Poisson, entropic scheme, convex minimization

AMS subject classifications. 65K10, 65M12, 65N25, 76Y05, 82C10, 82D37

DOI. 10.1137/040610556

1. Introduction.

Recently, Degond and Ringhofer [15, 16] explored a new direction for quantum hydrodynamic (QHD) models by extending Levermore's moment approach [33] to the context of quantum mechanics. Their strategy consists in defining a notion of "local" quantum equilibrium as the minimizer of an entropy functional under local moment constraints. Such equilibria are defined, thanks to a relation between the thermodynamic quantities (such as the chemical potential or the temperature) and the extensive quantities (the densities), in a nonlocal way. In [15], QHD models were derived from quantum kinetic equations by moment expansions closed by these quantum equilibria. In this reference, Degond and Ringhofer also sketched an important program related to these QHD models, including, namely, the setting up of a rigorous framework for this formal modeling, the inclusion of other quantum effects (the Pauli exclusion principle, spin effects, etc.), and the numerical discretization and simulation. Following the same approach, these authors then introduced in [17] a family of ad hoc collision operators which decrease the quantum entropy and relax to the equilibria. Afterwards, this strategy was applied in [13] in order to derive quantum diffusive models: a quantum drift-diffusion (QDD) model and a quantum energy-transport (QET) model. In a work in progress [8], other diffusive models of the type of the spherical harmonic expansion (SHE) model are also constructed in the quantum framework.

All these fluid models are written as conservation laws coupled to constitutive equations. The quantum character of these models lies in these constitutive equations, which are nonlocal in space and make these systems difficult to analyze (papers [15, 13] remained at a formal level). However, an interesting property of these models is that—at least formally—a fluid entropy functional is dissipated. This feature gives an indication of the well-posedness of these systems; in addition, it is interesting to

*Received by the editors June 25, 2004; accepted for publication (in revised form) April 14, 2005; published electronically November 22, 2005. This research was supported by the European network HYKE, funded by the EC as contract HPRN-CT-2002-00282.

http://www.siam.org/journals/sinum/43-5/61055.html

[†]MIP, Laboratoire CNRS (UMR 5640), Université Paul Sabatier, 118, route de Narbonne, 31062 Toulouse Cedex 04, France (gallego@mip.ups-tlse.fr, mehats@mip.ups-tlse.fr).

recall that the entropic property is obtained as a by-product of the strategy of entropy minimization.

In this paper, we are interested in the QDD model with two objectives. First, the present work is a first step in the rigorous analysis of this system coupled to the Poisson equation. Second, we study the discretization of this system and its numerical simulation.

Let us now describe the main results of this paper. The QDD system is given by (2.8)–(2.10). Actually, we are not yet able to answer the question of the well-posedness of this system. Nevertheless, we introduce, instead, and analyze rigorously a semidiscretized (in time) version of this model, defined by (3.1)–(3.3), and which presents the same entropy dissipation property as the QDD system. This first set of results is given in Theorem 3.1. Next, concerning the second objective of the paper, the implicit numerical scheme (4.1)–(4.3) is defined. This scheme is well-posed and equivalent to a problem of convex minimization. Then, we show that this scheme is stable in the sense of a discrete entropy. These results concerning the numerical scheme are stated in Theorem 4.1.

We end this introduction with bibliographical notes on quantum transport modeling. The QDD system applies to the modeling of nanoscale semiconductor devices. In the semiconductor industry, the classical drift-diffusion model has been a valuable tool for many years [11, 28, 35, 37, 48]. Currently, the ongoing miniaturization of electronic devices to the nanometer scale has created the need of models which take into account quantum effects. To this aim, two strategies can be followed.

The first approach, with a radical change in the level of description, consists of choosing full quantum models such as the Schrödinger equation, the von Neumann equation, or the Wigner equation [4, 9, 12, 18, 19, 32, 38, 45, 46]. These models are well-fitted for very small devices but lead to the resolution of huge numerical systems at the intermediate scale, which is currently considered by electronic engineers. Another reason why this approach is limited to very small devices is that the question of describing collisions in quantum transport models is extremely difficult and has not yet received a completely satisfactory answer. Therefore, full quantum models are still mainly reserved to ballistic transport in small devices.

The opposite strategy consists of introducing quantum correction terms in the classical drift-diffusion model. The most common quantum correction involves the Bohm potential, which naturally appears in QHD, thanks to an analogy between the Schrödinger equation and the pressureless Euler system corrected with the Bohm potential. This analogy can be seen, thanks to the Madelung transformation [34, 50], by considering the equations satisfied by the amplitude and the phase of a wavefunction solving the Schrödinger equation (see, e.g., [13] for more details). Next, assuming that adding this Bohm potential enables us to model quantum effects in classical macroscopic systems, several models with corrective terms have been written. In a fluid context, hydrodynamics models with quantum corrections have been studied in [22, 23, 24, 25, 26, 27, 29, 44, 51]. In a diffusive context, and closest to the QDD model studied in this paper, one can find the drift-diffusion model, corrected with the Bohm potential, called the density-gradient model (it is also sometimes called the QDD model, but in this paper we shall refer it as the density-gradient model in order to avoid any confusion with the QDD model presented here). This model was introduced in [1, 2], then mathematically and numerically studied in [3, 7, 29, 30, 41, 42]. One advantage of such an approach is that it takes into account collisions, at least heuristically. Another strength is that, as this method is based on an evolution of the classical drift-diffusion model, the numerical codes currently employed in the semiconductor industry can be adapted by following this evolution. Nevertheless, one has to insist on the fact that the justification of these models is far from obvious in the case of statistical mixtures (several attempts were made to address this issue; see, for instance, [22, 23, 24, 27]). Moreover, quantum corrections involving the Bohm potential produce high order terms in these systems and make their resolution difficult, from the mathematical and numerical points of view. To conclude this description, one can also cite two other recent attempts to model quantum effects in diffusive models [6, 43]. The models presented in these works are different, but both take the form of a drift-diffusion equation, coupled to the Poisson equation, and where the quantum phenomena are taken into account by a modification of the link between the density and the quasi-Fermi potential, via the resolution of a quasi-static Schrödinger equation.

As a compromise, the QDD model studied in this paper tries to reconcile these two approaches: this model is really quantum and nonlocal, while the length scales are macroscopic and collisions are modeled. Indeed, as is shown in section 2.3, the steady states of the QDD model solve the Schrödinger–Poisson system studied in [31, 39, 40], which shows the quantum character of this model. In addition, it has been shown in [13] that, at least formally, the limit of the QDD model as $\hbar$ goes to zero is the classical drift-diffusion model, while the leading order correction term in an $\hbar$ expansion is the Bohm potential, which shows a clear link between the QDD model and the density-gradient model described above.

The paper is organized as follows. In section 2, we write a formulation of the QDD model in a bounded domain and give some of its properties. Then, in section 3, we define the semidiscretization in time of the QDD system and show that this new system is well-posed and entropic. In section 4, the numerical scheme is constructed and we analyze its properties (well-posedness, stability). Finally, in section 5, we illustrate these properties by some numerical simulations.

2. The QDD model.

This section is devoted to the presentation of the QDD model. It is not clear which precise functional framework would be adapted to a rigorous analysis of this system. Nevertheless, we can still state some properties satisfied by any smooth solution of this system. This enables us to put into perspective the results of section 3. Indeed, we shall see in section 3 that similar properties are satisfied by the solutions of the semidiscretized QDD system (3.1)–(3.3), whereas their existence can be rigorously proved.

2.1. Notation: The QDD model on a bounded domain.

Let us first give a formulation of the QDD model in the case of bounded domains. This model, which describes the evolution of a quantum system of electrons, was derived in [13] and the most convenient equivalent form of this model was written in the review paper [14]. The first equation is the equation of mass conservation:

$$\partial_t n + \mathrm{div} j = 0. \tag{2.1}$$

The second equation of the model is the constitutive equation which gives the expression of the current

$$j = n\nabla(A - V). \tag{2.2}$$

In this equation, $V(t,x)$ is the self-consistent potential (modeling the interactions between the electrons) and $A(t,x)$ is the *quantum chemical potential*, linked to the density by a relation which is nonlocal in space and which is the key of this quantum model. In order to make this relation explicit, let us introduce the operator

$$H[A] = -h^2 \Delta + A + V^{ext},$$

whose domain $D(H)$ will be defined below and where h is the dimensionless Planck constant:

$$h = \frac{\hbar}{(2m^* L^2\, kT)^{1/2}},$$

m^* being the effective mass, L a characteristic length of the device, and T the temperature. Here, $V^{ext}(x)$ is an external potential applied to the system (assumed independent of time for simplicity). In the QDD model, the electron system is at any time in a *local quantum equilibrium* (see [15, 13]) and its density matrix is

$$\varrho = \exp(-H[A]), \tag{2.3}$$

where exp denotes the exponential of the operator. Notice that when the chemical potential A differs from the electrical potential, the operator $H[A]$ is not the Hamiltonian and ϱ is not the density matrix of a global quantum equilibria as usually defined [5]. A consequence of this formula (2.3) is the relation between the density and the chemical potential, given in a weak sense by

$$\forall \phi \in L^\infty \qquad \int n\phi \, dx = \mathrm{tr}(\exp(-H[A])\, \phi). \tag{2.4}$$

Here we used the usual convention where, for any test function ϕ, $\mathrm{tr}(\exp(-H[A])\, \phi)$ denotes the trace of the composition of the exponential of the operator $-H[A]$ with the operator of multiplication by ϕ. Finally, the last equation of the model is the Poisson equation, which links the density and the self-consistent potential:

$$-\alpha \Delta V = n. \tag{2.5}$$

In this equation, α is a positive dimensionless parameter proportional to the square of the Debye length of the system; more precisely, if ε_0 and ε_r denote the vacuum permittivity and the relative permittivity of the material, if $\mathcal{N}$ denotes a characteristic density, and if e denotes the elementary charge, we have

$$\alpha = \frac{\varepsilon_0\, \varepsilon_r\, kT}{e^2\, L^2\, \mathcal{N}}.$$

A given background charge density may be taken into account in this model, for instance, by a modification of the external potential V^{ext} and a shift of the chemical potential A.

Let $\Omega \subset \mathbb{R}^d$ be a regular bounded domain ($d \leq 3$). Its boundary is denoted by $\partial\Omega$ and $\nu(x)$ is the outward unit normal vector at $x \in \partial\Omega$. All the unknowns of the system $n(t,x)$, $j(t,x)$, $A(t,x)$, and $V(t,x)$ are defined for $t \geq 0$ and $x \in \Omega$. Now, we need to make the boundary conditions for this system precise. The most simple ones, which will be studied in this paper, prescribe a vanishing current at the boundary. This no-flux boundary condition takes the form of the Neumann condition

$$\nabla(A - V) \cdot \nu = 0 \quad \text{on } \partial\Omega.$$

(Recall that we assume A and V smooth enough to give sense to this Neumann condition; for the semidiscretized model analyzed in section 3, the $W^{2,p}$ regularity obtained in Theorem 3.1 is enough.) For the self-consistent potential, we consider a Dirichlet boundary condition

$$V = 0 \quad \text{on } \partial\Omega.$$

It remains to fix the domain of the Hamiltonian $H[A]$. In [21], the QDD model was written with Dirichlet boundary conditions for the wavefunctions, as well as its discrete version. Here, for technical reasons which will be explained further (we need to ensure the positivity of the density on $\overline{\Omega}$: see the beginning of the proof of Theorem 3.1), Neumann boundary conditions are chosen:

$$D(H) = \{\phi \in H^2(\Omega) \,:\, \nabla\phi \cdot \nu = 0 \quad \text{on } \partial\Omega\}. \tag{2.6}$$

Hence, if A belongs to, say, $L^2(\Omega)$, then the operator $H[A]$ is bounded from below and has a compact resolvent. Let us denote by $(\chi_p[A])_{p=1,\dots,\infty}$ an orthogonal basis of eigenfunctions, associated with the eigenvalues $\lambda_1[A] \le \lambda_2[A] \le \cdots \le \lambda_p[A] \le \cdots$. The nonlocal relation (2.4) between n and A takes a more explicit form

$$n[A] = \sum_{p\ge 1} e^{-\lambda_p[A]} \, |\chi_p[A]|^2. \tag{2.7}$$

To summarize this part, one can write the QDD model including self-consistent effects as follows:

$$\partial_t n + \operatorname{div}(n\nabla(A - V)) = 0, \tag{2.8}$$

$$-\alpha\Delta V = n, \tag{2.9}$$

$$n = \sum_{p} e^{-\lambda_p[A]} \, |\chi_p[A]|^2, \tag{2.10}$$

where $(\lambda_p[A], \chi_p[A])_p$ denote the eigenvalues and the eigenfunctions of the Hamiltonian $H[A] = -h^2\Delta + A + V^{ext}$ whose domain is $D(H) = \{\psi \in H^2(\Omega) \,:\, \partial_\nu\psi = 0\}$. The unknowns of this system are subject to the following no-flux boundary conditions on $\partial\Omega$:

$$V = 0, \quad \partial_\nu(A - V) = 0 \quad (\partial\Omega), \tag{2.11}$$

and to a Cauchy datum $n^0(x)$.

In this paper, the following assumptions on the data will be made.

Assumption 2.1. The initial datum n^0 is continuous and positive on $\overline{\Omega}$.

Assumption 2.2. The external potential V^{ext} is nonnegative and belongs to $L^\infty(\Omega)$.

2.2. Technical lemmas: The relation between *n* and *A*. In this subsection, we gather some technical lemmas that are used in this paper. The first lemma, which is given without proof, is directly adapted from [40] (the only difference lies in the domain $D(H)$; in [40], a Dirichlet boundary condition was considered instead of our Neumann boundary condition).

LEMMA 2.3. *Let $A \in H^1(\Omega)$ and let $n[A]$ be defined by*

$$n[A] = \sum_{p\ge 1} e^{-\lambda_p[A]} \, |\chi_p[A]|^2,$$

where $\lambda_p[A]$ and $\chi_p[A]$ are the spectral elements of

$$H[A] = -h^2\Delta + A + V^{ext},$$

whose domain $D(H)$ is defined by (2.6). *Then $n[A]$ is a continuous function on $\overline{\Omega}$. Moreover, the map F defined by*

$$A \in H^1(\Omega) \mapsto F[A] := \operatorname{tr}\big(e^{-H[A]}\big) = \int n[A]\, dx \tag{2.12}$$

is well defined, Fréchet C^∞, and strictly convex. Its first derivative in the direction $\phi \in H^1(\Omega)$ reads

$$d_A F \cdot \phi = -\operatorname{tr}\big(e^{-H[A]}\,\phi\big) = -\int n[A]\phi\, dx, \tag{2.13}$$

and its second derivative reads

$$d_A^2 F \cdot \phi \cdot \phi = -\sum_{p=1}^{\infty}\sum_{q=1}^{\infty} \frac{e^{-\lambda_p[A]} - e^{-\lambda_q[A]}}{\lambda_p[A] - \lambda_q[A]} \left| \int \phi\, \chi_p\, \overline{\chi_q}\, dx \right|^2, \tag{2.14}$$

where $\frac{e^{-\lambda_p[A]} - e^{-\lambda_q[A]}}{\lambda_p[A] - \lambda_q[A]}$ conventionally equals $-e^{-\lambda_p[A]}$ if $\lambda_p[A] = \lambda_q[A]$.

Notice that this lemma gives a sense to formula (2.7) as soon as A belongs to $H^1(\Omega)$.

LEMMA 2.4. *Let A and $\widetilde{A}$ belong to $H^1(\Omega)$ and, using the notation of Lemma* 2.3, *let*

$$n = n[A] = \sum_{p\geq 1} e^{-\lambda_p[A]}\, |\chi_p[A]|^2, \quad \widetilde{n} = n[\widetilde{A}] = \sum_{p\geq 1} e^{-\lambda_p[\widetilde{A}]}\, |\chi_p[\widetilde{A}]|^2.$$

Then we have

$$\int \big(n(A - \widetilde{A}) + n - \widetilde{n}\big)\, dx \leq 0. \tag{2.15}$$

Proof. The functional $F[A]$ defined in Lemma 2.3 is convex; thus we have the inequality

$$F[\widetilde{A}] - F[A] \geq d_A F \cdot (\widetilde{A} - A).$$

The desired result is a consequence of the expression (2.13) of $d_A F$. □

2.3. Steady states and entropy dissipation. The steady states of the QDD system are well known: these are the solutions of the Schrödinger–Poisson system studied by Nier in [40]. Following this reference, the following proposition can be proved (its proof is left to the reader).

PROPOSITION 2.5. *Let $N > 0$ and let (n, A, V) be a steady state of* (2.8)–(2.10) *such that $\int n(x)\, dx = N$. Assume that n is continuous and positive on $\overline{\Omega}$. Then there exists a constant ϵ_F such that $A = V - \epsilon_F$ and (n, V, ϵ_F) is the unique solution of the Schrödinger–Poisson system under a constraint of the total charge:*

$$\begin{cases} -h^2\Delta\chi_p + (V + V^{ext})\,\chi_p = \lambda_p\,\chi_p \quad (p = 1, \ldots, \infty), \\ \chi_p \in D(H), \qquad \displaystyle\int \chi_p\, \overline{\chi_q} = \delta_{pq}, \end{cases} \tag{2.16}$$

$$-\alpha\Delta V = n = \sum_p e^{\epsilon_F - \lambda_p}\, |\chi_p|^2, \quad V \in H_0^1(\Omega), \tag{2.17}$$

$$\int n(x)\, dx = N. \tag{2.18}$$

Next, the following formal result shows that the QDD system coupled with the Poisson equation is entropic.

Proposition 2.6. *Let* (n, A, V) *be a smooth solution of* (2.8)–(2.10). *Then the following properties hold.*

(i) *The following free energy* $S(t)$ *is a decreasing function of time and is bounded from below* (*by a negative constant depending only on* Ω *and* h)*:*

$$S(t) = -\int n\,(A+1)\, dx + \frac{\alpha}{2}\int |\nabla V|^2\, dx.$$

(ii) *If* $(n^\infty, A^\infty, V^\infty)$ *is the solution of* (2.16)–(2.18) *corresponding to* $N = \int n(0,x)\, dx$, *then the following relative entropy* $\Sigma(t)$ *is the sum of two nonnegative terms and is a decreasing function of time:*

$$\Sigma(t) = -\int (n\,(A - A^\infty) + n - n^\infty)\, dx + \frac{\alpha}{2}\int |\nabla(V - V^\infty)|^2 dx.$$

Proof. By applying (2.15) with $\widetilde{A} \equiv 0$, we get

$$-\int n\,(A+1) \geq -\int n[0]\, dx.$$

Assumption 2.2 gives $V^{ext} \geq 0$. Hence, by the min-max formula, the eigenvalues $\lambda_p[0]$ of $H[0] = -h^2\Delta + V^{ext}$ satisfy $\lambda_p[0] \geq \lambda_p^\Delta$, where λ_p^Δ are the eigenvalues of $-h^2\Delta$ with Neumann boundary conditions on $\partial\Omega$. Thus, we have

$$\int n[0]\, dx \leq \sum_p e^{-\lambda_p^\Delta}$$

and S is bounded from below by a constant which depends only on Ω and h.

Let us now remark that, due to the no-flux boundary conditions (2.11), an integration of the first equation of (2.8)–(2.10) yields the conservation of the total charge:

$$\forall t \geq 0 \quad \int n(t,x)\, dx = \int n(0,x)\, dx. \tag{2.19}$$

Independently, by differentiating with respect to time the functional $F[A]$ defined by (2.12), and recalling that V^{ext} is independent of time, we get

$$\frac{d}{dt}\int n(t,x)\, dx = \frac{d}{dt}F[A(t)] = d_A F \cdot \partial_t A = -\int n(t,x)\, \partial_t A(t,x)\, dx;$$

thus, we have

$$\frac{d}{dt}\int n\,(A+1)\, dx = \int (\partial_t n) A\, dx.$$

To prove item (i), it remains to remark that the Poisson equation with Dirichlet boundary conditions yields

$$\frac{d}{dt}\frac{\alpha}{2}\int |\nabla V|^2\,dx = \int (\partial_t n)\,V\,dx.$$

Consequently, we obtain

$$\frac{d}{dt}S(t) = -\int (\partial_t n)(A-V)\,dx = -\int n\,|\nabla(A-V)|^2\,dx \le 0, \tag{2.20}$$

which proves (i). Let us now prove (ii). The fact that the first term of $\Sigma(t)$ is nonnegative stems from (2.15). In addition, since we have $A^\infty = V^\infty - \epsilon_F$, we deduce the equivalent expression

$$\begin{aligned}\Sigma(t) &= -\int (n\,(A+\epsilon_F) + n - n^\infty)\,dx \\ &\quad + \alpha\int \nabla V\cdot\nabla V^\infty\,dx + \frac{\alpha}{2}\int |\nabla(V-V^\infty)|^2 \\ &= S(t) - \epsilon_F\int n\,dx + \int n^\infty\,dx + \frac{\alpha}{2}\int |\nabla V^\infty|^2\,dx,\end{aligned}$$

where we used the Poisson equation $-\alpha\Delta V = n$. Therefore, by using (2.19), we deduce

$$\frac{d}{dt}\Sigma(t) = \frac{d}{dt}S(t) \le 0. \qquad \square$$

Remark 2.7. *Equation* (2.20) *gives the expression of the entropy dissipation. This term indicates that, as time goes to infinity, $A-V$ should converge towards a constant. Thus any transient solution of the QDD model should converge to the (unique) corresponding steady state. In order to prove rigorously this convergence, we need to control n from below, which is an open problem.*

3. Semidiscretization in time.

This section is devoted to the study of a semidiscrete version of (2.8)–(2.10), which appears as a first step towards the numerical scheme that is presented in section 4. Let $\Delta t > 0$ be the time step. For $k \in \mathbb{N}$, the semidiscretized model is written as

$$\frac{n^{k+1}-n^k}{\Delta t} + \operatorname{div}(n^k\nabla(A^{k+1}-V^{k+1})) = 0, \tag{3.1}$$

$$-\alpha\Delta V^{k+1} = n^{k+1}, \tag{3.2}$$

$$n^{k+1} = \sum_p e^{-\lambda_p[A^{k+1}]}\,|\chi_p[A^{k+1}]|^2, \tag{3.3}$$

subject to the boundary conditions

$$V^{k+1} = 0, \quad \partial_\nu(A^{k+1}-V^{k+1}) = 0. \tag{3.4}$$

Recall that, in this system, $\lambda_p[\cdot]$ and $\chi_p[\cdot]$ denote the whole sequence of eigenvalues and eigenfunctions of the operator $H[\cdot]$ defined in section 2.1 by

$$H[A] = -h^2\Delta + A + V^{ext}.$$

The unknowns are the density $n^k(x)$, the quantum chemical potential $A^k(x)$, and the self-consistent potential $V^k(x)$ for $k \in \mathbb{N}^*$. For $k = 0$, the density n^0 is given satisfying Assumption 2.1. Then, the Poisson equation enables us to define V^0. Concerning the initial chemical potential A^0, since it is not clear whether (2.7) can be inverted, we choose to let A^0 be undetermined. Notice that A^0 is not required in this model to compute (n^k, A^k, V^k) for $k \geq 1$. An alternative choice for the initial conditions would be to take an initial datum A^0, then to deduce n^0 by (2.7) and V^0 by the Poisson equation. However, it seems more interesting, for physical reasons, to start from an initial density n^0.

The main result of this section is as follows.

Theorem 3.1. *Under Assumptions* 2.1 *and* 2.2, *we have the following properties.*

(i) *The semidiscretized model* (3.1)–(3.3) *is well-posed. For all $k \in \mathbb{N}^*$, the functions $A^k \in W^{2,p}(\Omega)$, $V^k \in W^{2,p}(\Omega)$ (for any $p < \infty$), and $n^k \in C(\overline{\Omega})$ are uniquely defined and, for all k, we have $n^k > 0$ on $\overline{\Omega}$.*

(ii) *The total charge is conserved,*

$$\forall k \in \mathbb{N} \quad \int n^k \, dx = \int n^0 \, dx, \tag{3.5}$$

and the following free energy S^k, defined for $k \geq 1$, is bounded from below and decreases as k increases:

$$S^k = -\int n^k \left(A^k + 1\right) dx + \frac{\alpha}{2} \int |\nabla V^k|^2 \, dx.$$

(iii) *If $(n^\infty, A^\infty, V^\infty)$ is the solution of the Schrödinger–Poisson system* (2.16)–(2.18) *corresponding to $N = \int n^0 \, dx$, then the following relative entropy Σ^k is the sum of two nonnegative terms and decreases as k increases:*

$$\Sigma^k = -\int \left(n^k \left(A^k - A^\infty\right) + n^k - n^\infty\right) dx + \frac{\alpha}{2} \int |\nabla(V^k - V^\infty)|^2.$$

Proof. (i) Let us first give the outline of this proof. We shall proceed by induction, for any function n^k, positive and continuous on $\overline{\Omega}$, we will show that there exists a unique pair $(A^{k+1}, V^{k+1}) \in H^1(\Omega) \times H_0^1(\Omega)$ satisfying

$$\frac{n[A^{k+1}] - n^k}{\Delta t} + \operatorname{div}(n^k \nabla(A^{k+1} - V^{k+1})) = 0 \tag{3.6}$$

and

$$-\alpha \Delta V^{k+1} = n[A^{k+1}], \tag{3.7}$$

with the boundary condition (3.4), where we recall the notation

$$n[A^{k+1}] = \sum_p e^{-\lambda_p[A^{k+1}]} |\chi_p[A^{k+1}]|^2.$$

Then, as soon as (A^{k+1}, V^{k+1}) is defined, it suffices to set $n^{k+1} = n[A^{k+1}]$ and (3.1)–(3.3) is satisfied. Moreover, the first part of Lemma 2.3 shows that n^{k+1} is continuous on $\overline{\Omega}$. Hence, (3.1)–(3.3) and standard elliptic regularity estimates imply that for any $p < \infty$ we have $V^{k+1} \in W^{2,p}(\Omega)$ and $A^{k+1} \in W^{2,p}(\Omega)$. By Sobolev embeddings, we deduce that $A^{k+1} \in L^\infty(\Omega)$, which is enough to apply Krein–Rutman's theorem

[10]; the choice of Neumann boundary conditions for the eigenfunction χ_p (see (2.6)) ensures the fact that χ_1^{k+1} does not vanish on the closed domain $\overline{\Omega}$. Consequently, n^{k+1} is itself positive and continuous on $\overline{\Omega}$ and can be used to initiate the next step of the induction; we are then able to construct $(A^{k+2}, V^{k+2}, n^{k+2})$. Finally, thanks to Assumption 2.1 on the initial density n^0, all of the sequence $(A^k, V^k, n^k)_{k\geq 1}$ can be constructed by induction.

Let us now prove the claim: for any given positive and continuous function n^k, one can construct a unique corresponding (A^{k+1}, V^{k+1}) satisfying (3.6) and (3.7). This proof, inspired by [39, 40], is based on a variational argument. We introduce the following functional, defined for $A \in H^1(\Omega)$ and $V \in H_0^1(\Omega)$:

$$J(A,V) = \frac{\Delta t}{2} \int n^k \, |\nabla(A-V)|^2 \, dx + \frac{\alpha}{2} \int |\nabla V|^2 \, dx + F[A] + \int n^k \, (A-V) \, dx,$$

where $F[A]$ is defined by

$$F[A] = \mathrm{tr} e^{-H[A]} = \sum_{p\geq 1} e^{-\lambda[A]}.$$

Note that this functional J depends on n^k. By Lemma 2.3, this functional is continuous, Fréchet differentiable, and its derivative is given by

$$\begin{aligned} d_{A,V} J \cdot (\delta A, \delta V) = \Delta t & \int n^k \, \nabla(A-V) \cdot \nabla(\delta A - \delta V) \, dx \\ & + \alpha \int \nabla V \cdot \nabla \delta V \, dx \\ & - \int n[A] \, \delta A \, dx + \int n^k \, (\delta A - \delta V) \, dx, \end{aligned}$$

where $\delta A \in H^1(\Omega)$, $\delta V \in H_0^1(\Omega)$, and we recall the notation

$$n[A] = \sum_p e^{-\lambda_p[A]} \, |\chi_p[A]|^2.$$

Therefore, it is readily seen that the critical points of J satisfy (3.1)–(3.4). To prove the existence and uniqueness of A^{k+1} and V^{k+1}, it suffices to show that J is strictly convex and coercive, since its unique minimizer will be (A^{k+1}, V^{k+1}). The strict convexity is a consequence of Lemma 2.3 (which states that F is strictly convex) of the strict convexity of the functional

$$V \in H_0^1(\Omega) \longmapsto \int |\nabla V|^2 \, dx$$

and of the convexity of the functional

$$(A,V) \in H^1(\Omega) \times H_0^1(\Omega) \longmapsto \int n^k \, |\nabla(A-V)|^2 \, dx.$$

It remains to prove the coercivity with respect to $A \in H^1(\Omega)$ and $V \in H_0^1(\Omega)$. Let $(A^\varepsilon, V^\varepsilon)$ be a sequence in $H^1(\Omega) \times H_0^1(\Omega)$, parametrized by $\varepsilon > 0$, such that $J(A^\varepsilon, V^\varepsilon)$ has an upper bound independent of ε. To prove the coercivity of J, it suffices to show that $\|A^\varepsilon\|_{H^1} + \|V^\varepsilon\|_{H^1}$ can be bounded independently of ε.

Setting $a^\varepsilon = \frac{1}{|\Omega|}\int A^\varepsilon\,dx$ (where $|\Omega|$ denotes the measure of Ω), we introduce the function $B^\varepsilon = A^\varepsilon - a^\varepsilon$. We have

$$J(A^\varepsilon, V^\varepsilon) = \frac{\Delta t}{2}\int n^k\,|\nabla(B^\varepsilon - V^\varepsilon)|^2\,dx + \frac{\alpha}{2}\int |\nabla V^\varepsilon|^2\,dx$$
$$+\,e^{-a^\varepsilon}\sum_p e^{-\lambda_p[B^\varepsilon]} + \int n^k\,(B^\varepsilon - V^\varepsilon)\,dx + a^\varepsilon \int n^k\,dx \le C,$$

where C does not depend on ε. We recall that there exist two constants $\underline{n} > 0$ and $\overline{n} > 0$, independent of ε, such that

$$\underline{n} \le n^k(x) \le \overline{n} \quad \text{on } \overline{\Omega}.$$

Hence, the Cauchy–Schwarz inequality gives

(3.8)
$$\frac{\Delta t}{2}\,\underline{n}\int |\nabla(B^\varepsilon - V^\varepsilon)|^2\,dx + \frac{\alpha}{2}\int |\nabla V^\varepsilon|^2\,dx - \overline{n}\,|\Omega|^{1/2}\big(\|B^\varepsilon\|_{L^2(\Omega)} + \|V^\varepsilon\|_{L^2(\Omega)}\big)$$
$$+\,e^{-a^\varepsilon}\sum_p e^{-\lambda_p[B^\varepsilon]} + a^\varepsilon\int n^k\,dx \le J(A^\varepsilon, V^\varepsilon) \le C.$$

In addition, denoting by $\widetilde{H^1}(\Omega)$ the space of $H^1(\Omega)$ functions which have a vanishing integral on Ω, a classical compactness argument shows that, for any $a_1 > 0$ and $a_2 > 0$, the norm

$$(B, V) \in \widetilde{H^1}(\Omega) \times H^1_0(\Omega) \quad \longmapsto \quad \big(a_1\|\nabla(B - V)\|^2_{L^2(\Omega)} + a_2\|\nabla V\|^2_{L^2(\Omega)}\big)^{1/2}$$

is equivalent on this space $\widetilde{H^1}(\Omega) \times H^1_0(\Omega)$ to the standard $H^1(\Omega) \times H^1(\Omega)$ norm. Hence, there exist two constants $C_0 > 0$ and $C_1 > 0$, independent of ε, such that

$$\frac{\Delta t}{2}\,\underline{n}\int |\nabla(B^\varepsilon - V^\varepsilon)|^2\,dx + \frac{\alpha}{2}\int |\nabla V^\varepsilon|^2\,dx - \overline{n}\,|\Omega|^{1/2}\left(\|B^\varepsilon\|_{L^2(\Omega)} + \|V^\varepsilon\|_{L^2(\Omega)}\right)$$
$$\ge C_0\|B^\varepsilon\|^2_{H^1(\Omega)} + C_0\|V^\varepsilon\|^2_{H^1(\Omega)} - C_1;$$

thus (3.8) gives

$$C_0\|B^\varepsilon\|^2_{H^1(\Omega)} + C_0\|V^\varepsilon\|^2_{H^1(\Omega)} + e^{-a^\varepsilon}\sum_p e^{-\lambda_p[B^\varepsilon]} + a^\varepsilon\int n^k\,dx \le C. \tag{3.9}$$

Let us now recall that the first eigenvalue of $H[B^\varepsilon]$ is defined by

$$\lambda_1[B^\varepsilon] = \min_{\substack{\phi\,\in\,H^1(\Omega)\\ \|\phi\|_{L^2(\Omega)} = 1}} \left(h^2\int |\nabla\phi|^2\,dx + \int (B^\varepsilon + V^{ext})\,\phi^2\,dx\right).$$

By choosing the test function $\phi(x) \equiv 1/\sqrt{|\Omega|}$ in this formula, we deduce from $\int B^\varepsilon\,dx = 0$ that

$$\lambda_1[B^\varepsilon] \le \frac{1}{|\Omega|}\int V^{ext}\,dx.$$

There exists, consequently, a constant $C_2 > 0$, independent of ε, such that

$$e^{-a^\varepsilon} \sum_p e^{-\lambda_p[B^\varepsilon]} \geq C_2\, e^{-a^\varepsilon},$$

and (3.9) implies

$$C_0 \|B^\varepsilon\|^2_{H^1(\Omega)} + C_0 \|V^\varepsilon\|^2_{H^1(\Omega)} + C_2 e^{-a^\varepsilon} + a^\varepsilon \int n^k \, dx \leq C.$$

Since $\int n^k \, dx > 0$, it is clear then that $\|B^\varepsilon\|_{H^1(\Omega)}$, $\|V^\varepsilon\|_{H^1(\Omega)}$, and $|a^\varepsilon|$ are bounded independently of ε. Thus $\|A^\varepsilon\|_{H^1(\Omega)}$ is bounded, which completes the proof of coercivity.

(ii) The conservation of mass (3.5) can be easily shown by an integration of (3.1) on Ω, which gives, thanks to the boundary conditions (3.4),

$$\int n^{k+1} \, dx = \int n^k \, dx. \tag{3.10}$$

To prove the decay of the free energy, let us adapt to the semidiscrete case the proof of Proposition 2.6. By using Lemma 2.4, we have

$$\int (n^k (A^k - A^{k+1}) + n^k - n^{k+1}) \, dx \leq 0;$$

thus

$$\begin{aligned} -\int (n^{k+1}\, A^{k+1} - n^k\, A^k + n^{k+1} - n^k) \, dx & \\ = -\int (n^{k+1} - n^k)\, A^{k+1} \, dx + \int (n^k (A^k - A^{k+1}) + n^k - n^{k+1}) \, dx & \\ \leq -\int (n^{k+1} - n^k)\, A^{k+1} \, dx. & \end{aligned} \tag{3.11}$$

In addition, by using the Poisson equation (3.2), we obtain

$$\begin{aligned} \frac{\alpha}{2} \int (|\nabla V^{k+1}|^2 - |\nabla V^k|^2) \, dx &= \frac{1}{2} \int (n^{k+1}\, V^{k+1} - n^k\, V^k) \, dx \\ &= \frac{1}{2} \int (n^{k+1} - n^k)\, V^{k+1} \, dx + \frac{1}{2} \int n^k (V^{k+1} - V^k) \, dx \\ &= \frac{1}{2} \int (n^{k+1} - n^k)\, V^{k+1} \, dx + \frac{1}{2} \int V^k (n^{k+1} - n^k) \, dx. \end{aligned}$$

By remarking that

$$0 \leq \alpha \int |\nabla (V^{k+1} - V^k)|^2 dx = \int (n^{k+1} - n^k)(V^{k+1} - V^k),$$

we deduce that

$$\frac{1}{2} \int V^k (n^{k+1} - n^k) \, dx \leq \frac{1}{2} \int V^{k+1} (n^{k+1} - n^k) \, dx$$

and get

$$\frac{\alpha}{2}\int \left(|\nabla V^{k+1}|^2 - |\nabla V^k|^2\right) dx \le \int V^{k+1}(n^{k+1} - n^k)\, dx.$$

By combining this inequality and (3.11), we obtain

$$\begin{aligned} S^{k+1} - S^k &\le -\int (n^{k+1} - n^k)(A^{k+1} - V^{k+1})\, dx \\ &= \Delta t \int (A^{k+1} - V^{k+1})\, \mathrm{div}(n^k \nabla(A^{k+1} - V^{k+1}))\, dx, \end{aligned}$$

thanks to (3.1). An integration by parts, using (3.4), finally gives

$$S^{k+1} - S^k \le -\Delta t \int n^k\, |\nabla(A^{k+1} - V^{k+1})|^2\, dx \le 0.$$

This proves (ii). Finally, to prove (iii), it suffices to remark as for Proposition 2.6 that

$$\Sigma^{k+1} - \Sigma^k = S^{k+1} - S^k \le 0. \qquad \square$$

4. The fully discretized system: Construction and analysis. We complete the construction of a numerical scheme for the QDD model (2.8)–(2.10) by now discretizing system (3.1)–(3.3) with respect to the space variable. In the following section, we construct the scheme and give in Theorem 4.1 its main properties: well-posedness, charge conservation, and entropy dissipation. These properties are proved in section 4.2. Section 4.3 is devoted to the particular question of the initial step: it is shown in Proposition 4.4 that, at the discrete level, there exists a unique chemical potential A corresponding to each positive density n.

4.1. Notation and main results. For simplicity, the space dimension is now $d = 1$. The domain is $\Omega = (0, 1)$ and the space gridstep is $\Delta x = 1/(N + 1)$. The grid is composed of the points $x_i = i\Delta x$ for $i = 0, \dots, N + 1$, where $N \in \mathbb{N}$. In order to write the fully discretized finite difference numerical scheme, let us introduce the following $N \times N$ matrices of discrete derivative:

$$D^- = \frac{1}{\Delta x}\begin{pmatrix} 0 & 0 & \cdots & \\ -1 & 1 & 0 & \cdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & 0 & -1 & 1 \end{pmatrix}, \quad D^+ = \frac{1}{\Delta x}\begin{pmatrix} -1 & 1 & 0 & \dots \\ 0 & -1 & 1 & \cdots \\ 0 & \ddots & \ddots & 1 \\ \vdots & \cdots & 0 & 0 \end{pmatrix},$$

$$\widetilde{D^-} = \frac{1}{\Delta x}\begin{pmatrix} 1 & 0 & \cdots & \\ -1 & 1 & 0 & \cdots \\ 0 & \ddots & \ddots & 0 \\ \vdots & 0 & -1 & 1 \end{pmatrix}, \quad \widetilde{D^+} = \frac{1}{\Delta x}\begin{pmatrix} -1 & 1 & 0 & \dots \\ 0 & -1 & 1 & \cdots \\ 0 & \ddots & \ddots & 1 \\ \vdots & \cdots & 0 & 1 \end{pmatrix},$$

$$\Delta_{Dir} = \frac{1}{\Delta x^2}\begin{pmatrix} -2 & 1 & 0 & \dots \\ 1 & -2 & \ddots & 0 \\ 0 & \ddots & \ddots & 1 \\ \vdots & \cdots & 1 & -2 \end{pmatrix}, \quad \Delta_{Neu} = \frac{1}{\Delta x^2}\begin{pmatrix} -1 & 1 & 0 & \dots \\ 1 & -2 & \ddots & 0 \\ 0 & \ddots & \ddots & 1 \\ \vdots & \cdots & 1 & -1 \end{pmatrix}.$$

Note that $\Delta_{Neu} = \widetilde{D^-}D^+ = \widetilde{D^+}D^-$. The unknowns are the following sequences of vectors in $\mathbb{R}^N$: $n^k = (n_i^k)_{1\leq i\leq N}$, $A^k = (A_i^k)_{1\leq i\leq N}$, $V^k = (V_i^k)_{1\leq i\leq N}$ and the scheme is written as

$$\frac{n^{k+1}-n^k}{\Delta t} + \frac{1}{2}\widetilde{D^-}(n^k\, D^+(A^{k+1}-V^{k+1})) + \frac{1}{2}\widetilde{D^+}(n^k\, D^-(A^{k+1}-V^{k+1})) = 0, \tag{4.1}$$

$$-\,\alpha\,\Delta_{Dir}V^k = n^k, \tag{4.2}$$

$$n^k = \sum_p \exp(-\ell_p[A^k])(X_p[A^k])^2 \tag{4.3}$$

for $k \in \mathbb{N}$ (here and in what follows, for any $(X,Y) \in \mathbb{R}^N \times \mathbb{R}^N$, XY denotes the direct product $(X_iY_i)_{1\leq i\leq N}$). In this discretized system, the definitions of $\ell_p[A]$ and $X_p[A]$ are the discrete analogue of those of $\lambda_p[A]$, $\chi_p[A]$ for the continuous problem. These quantities are the eigenvalues and the normalized eigenvectors of the discretized Hamiltonian with Neumann boundary conditions

$$M[A] = -h^2\Delta_{Neu} + \mathrm{Diag}(A + V^{ext}),$$

where $\mathrm{Diag}(A)$ denotes the diagonal matrix of coefficients $(A_i)_{1\leq i\leq N}$, and where the components of the vector V^{ext} are $V_i^{ext} = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} V^{ext}(x)\,dx$. Of course, the index p of the eigenvalues and eigenvectors belongs now to $\{1,\dots,N\}$. Moreover, the eigenvectors are normalized with respect to the euclidean norm $\|\cdot\|_N$ associated with the scalar product on $\mathbb{R}^N$:

$$(U,V)_N = \Delta x \sum_{i=1}^N U_i\, V_i.$$

Notice that the boundary conditions are already taken into account in this scheme, the values of the unknowns for $i = 0$ or $i = N+1$ being implicitly defined. To complete (4.1)–(4.3), it suffices to add an initial condition. If Cauchy data for the continuous problem n^0 are given, the vector $n^0 \in \mathbb{R}^N$ is chosen as follows:

$$n_i^0 = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} n^0(x)\,dx \quad \text{for } i = 1,\dots,N. \tag{4.4}$$

The numerical scheme (4.1)–(4.3) is clearly consistent with the QDD system (2.8)–(2.11). Its properties are listed in the following theorem, whose proof is developed in the three next subsections.

THEOREM 4.1. *If Assumptions* 2.1 *and* 2.2 *are satisfied, the numerical scheme* (4.1)*–*(4.4) *is consistent with* (2.8)*–*(2.11) *and has the following properties.*

(i) Well-posedness. *For all $k \in \mathbb{N}$, its numerical solution (n^k, A^k, V^k) is uniquely defined. Moreover, for all $k \in \mathbb{N}$, (A^{k+1}, V^{k+1}) is the unique minimizer of the strictly convex and coercive functional*

$$\begin{aligned}\widehat{J}(A,V) = {} & \frac{\Delta t\,\Delta x}{4}\sum_{i=1}^N n_i^k\,(D^+(A-V))_i^2 + \frac{\Delta t\,\Delta x}{4}\sum_{i=1}^N n_i^k\,(D^-(A-V))_i^2 \\ & + \frac{\alpha\,\Delta x}{2}\sum_{i=1}^N (D^+V)_i^2 + \frac{\alpha}{2\,\Delta x}(V_1)^2 + \frac{\alpha}{2\,\Delta x}(V_N)^2 \\ & + \sum_{p=1}^N \exp(-\ell_p[A]) + \Delta x \sum_{i=1}^N n_i^k\,(A_i - V_i).\end{aligned} \tag{4.5}$$

(ii) Charge conservation. *For all k and for all i we have $n_i^k > 0$ and the (discrete) total charge is conserved:*

$$\forall k \in \mathbb{N} \quad \Delta x \sum_{i=1}^{N} n_i^k = \Delta x \sum_{i=1}^{N} n_i^0. \tag{4.6}$$

(iii) Entropy dissipation. *The sequence of (discrete) free energies defined by*

(4.7)

$$S^k = -\Delta x \sum_{i=1}^{N} n_i^k \left(A_i^k + 1\right) + \frac{\alpha \Delta x}{2} \sum_{i=1}^{N} (D^+ V^k)_i^2 + \frac{\alpha}{2\,\Delta x}\left(V_1^k\right)^2 + \frac{\alpha}{2\,\Delta x}\left(V_N^k\right)^2$$

is decreasing and belongs to ℓ^∞. Moreover, there exists a constant $C > 0$ (depending only on Ω and h) such that, for any $K \in \mathbb{N}$, we have

$$\begin{aligned} -C \le S^K &+ \frac{\Delta t \Delta x}{2} \sum_{k=1}^{K} \sum_{i=1}^{N} n_i^{k-1} \left(D^+(A^k - V^k)\right)_i^2 \\ &+ \frac{\Delta t \Delta x}{2} \sum_{k=1}^{K} \sum_{i=1}^{N} n_i^{k-1} \left(D^-(A^k - V^k)\right)_i^2 \le S^0. \end{aligned} \tag{4.8}$$

4.2. Proof of well-posedness and entropy dissipation. For the sake of conciseness, we shall only sketch the proof of Theorem 4.1. Indeed, it suffices to adapt to the discrete case the proof of Theorem 3.1. These results are based on formulas of discrete integration by parts and on technical results concerning matrix analysis which are the discrete equivalents of the technical results stated in section 2.2, and that we have listed in Lemma 4.2 below.

It is worthwhile to mention that the similarity between the functional $J(A,V)$, introduced in the proof of Theorem 3.1, and the functional $\widehat{J}(A,V)$ of Theorem 4.1 is due to two useful formulas of discrete integration by parts: for any pair of vectors $(U,V) \in \mathbb{R}^N \times \mathbb{R}^N$, we have

$$\begin{aligned} -(\Delta_{Neu} U, V)_N &= -(\widetilde{D^-} D^+ U, V)_N = \left(D^+ U, D^+ V\right)_N \\ &= -(\widetilde{D^+} D^- U, V)_N = \left(D^- U, D^- V\right)_N \end{aligned} \tag{4.9}$$

and

$$-(\Delta_{Dir} U, V)_N = (D^+ U, D^+ V)_N + \frac{U_1 V_1 + U_N V_N}{\Delta x}. \tag{4.10}$$

Next, we gather in the following lemma some classical but useful technical results on matrices.

LEMMA 4.2. *Let $A \in \mathbb{R}^N$. Then the eigenvalues $\ell_p[A]$ of the matrix $M[A] = -h^2 \Delta_{Neu} + \mathrm{Diag}(A + V^{ext})$ are simple. (Up to a multiplication by -1) its first eigenvector $X_1[A]$ has positive components. The derivatives of the eigenvalues and eigenvectors of $M[A]$ with respect to A, in the direction δA, are given by*

$$\begin{aligned} d\ell_p[A] \cdot \delta A &= (\delta A \, X_p[A], X_p[A])_N, \\ dX_p[A] \cdot \delta A &= \sum_{q \ne p} \frac{1}{\ell_p[A] - \ell_q[A]} (\delta A \, X_p[A], X_q[A])_N \, X_q[A]. \end{aligned}$$

Proof. The simplicity of the eigenvalues of $M[A]$ is a general classical result for Hessenberg matrices [49], i.e., matrices $M = (m_{i,j})_{1\leq i,j\leq N}$ such that

$$m_{i,j} = 0 \text{ for } j < i-1 \quad \text{and} \quad m_{i,i-1} \neq 0 \text{ for } 2 \leq i \leq N.$$

This simplicity enables us to differentiate ℓ_p and $X_p[A]$ by using the classical perturbation theory.

Let $\lambda = 1 + \max_i |A_i|$. Then it is clear that the matrix $M[A] + \lambda I$ is invertible and satisfies the discrete maximum principle:

$$\forall Y \in \mathbb{R}^N\backslash\{0\} \quad Y \geq 0 \Longrightarrow (M[A] + \lambda I)^{-1}\, Y > 0,$$

where, for any vector $X \in \mathbb{R}^N$, the notation $X \geq 0$ (resp., $X > 0$) stands for $X_i \geq 0$ (resp., $X_i > 0$) for all $i = 1, \dots, N$. Hence the Perron–Frobenius theorem (see [49]) applies to the matrix $(M[A] + \lambda I)^{-1}$, the spectral radius of this matrix is an eigenvalue and, up to a multiplication by -1, the corresponding eigenvector has positive components. This vector is the ground state $X_1[A]$ of $M[A]$. □

REMARK 4.3. *Special care has to be taken for the initial step of the scheme. In the semidiscrete case of system* (3.1)–(3.3)*, the question of the initial step was left unsolved: for a given initial density $n^0(x)$, can we define a unique corresponding chemical potential A^0 such that* (3.3) *holds? In the fully discrete case, this question finds a positive answer, as stated in Theorem* 4.1(i). *Section* 4.3 *is devoted to this particular point of the theorem.*

4.3. Initialization of the chemical potential. As noted in Remark 4.3, one question has not been addressed yet concerning the numerical scheme (4.1)–(4.4): the computation of the initial chemical potential A^0 corresponding to the initial data n^0. While, in the continuous problem, we do not know whether (or in which functional framework) the nonlocal relation (2.7) linking n to A is invertible, this operation is possible with its discrete analogous (4.3). The aim of this section is to establish this property: we show that this problem is again equivalent to a convex minimization problem. Notice that this enables us to deduce a practical method to numerically solve this problem, by writing an algorithm for this optimization problem (see [20] for details). Note also that the possibility of inverting the constitutive relation $A \mapsto n[A]$, interesting for itself, is not mandatory for the other steps of the scheme (see Theorem 4.1(i)): the minimization of J for the computation of (A^{k+1}, V^{k+1}) does not require the knowledge of A^k. The following proposition is the main result of this subsection.

PROPOSITION 4.4. *Let $n \in (\mathbb{R}_+^*)^N$. Then there exists a unique $A \in \mathbb{R}^N$ such that*

$$n = \sum_{p=1}^{N} \exp(-\ell_p[A])\,(X_p[A])^2, \tag{4.11}$$

where $\ell_p[A]$ and $X_p[A]$ are the eigenvalues and the eigenvectors of the discrete Hamiltonian $M[A] = -h^2\Delta_{Neu} + \mathrm{Diag}(A + V^{ext})$.

Proof. Consider the functional

$$\Phi[A] = \sum_p \exp\left(-\ell_p[A]\right) + (n, A)_N. \tag{4.12}$$

Straightforward calculations using Lemma 4.2 lead to the expression of its first and second derivatives:

$$d\Phi_A \cdot \delta A = \left(n - \sum_p \exp(-\ell_p[A])\,(X_p[A])^2, \delta A\right)_N$$

and

$$d^2\Phi_A \cdot \delta A \cdot \delta A = \sum_{p=1}^{N} \exp(-\ell_p[A])(\delta A\, X_p[A],\, X_p[A])_N^2$$
$$- \sum_p \sum_{q\neq p} \frac{\exp(-\ell_p[A]) - \exp(-\ell_q[A])}{\ell_p[A]-\ell_q[A]} (\delta A\, X_p[A],\, X_q[A])_N^2.$$

It is then clear that this functional Φ is strictly convex and that its unique minimizer satisfies (4.11). To prove the existence of a solution to the problem, the major task is to prove the coercivity of this functional.

Recall that

$$\ell_1[A] = \min_{\|\phi\|_N=1}((-h^2\Delta_{Neu}\phi,\phi)_N + (\mathrm{Diag}(A+V^{ext})\phi,\phi)_N). \tag{4.13}$$

Let $i_0 \in \{1,\dots,N\}$ (arbitrary). By choosing the i_0th normalized basis vector as ϕ in (4.13) (i.e., $\phi_i = \delta_{i,i_0}/\sqrt{\Delta x}$), we obtain

$$\ell_1[A] \le A_{i_0} + \frac{2h^2}{\Delta x^2} + V_{i_0}^{ext}. \tag{4.14}$$

Hence, there exists a constant $C > 0$ depending only on Δx, h, and V^{ext} such that

$$\Phi[A] \ge C\sum_i \exp(-A_i) + \Delta x \sum_i n_i\, A_i. \tag{4.15}$$

Since for all i we have $n_i > 0$, it is clear that

$$\lim_{\|A\|\to\infty} \Phi[A] = +\infty.$$

This proves the coercivity of Φ. □

5. Numerical results. In order to simulate the QDD model, the numerical scheme (4.1)–(4.3) has been implemented by minimizing the functional $\widehat{J}$ defined by (4.5). Each strictly convex unconstrained minimization problem is solved by a Newton method (note that the Hessian matrix is explicit and always positive definite). The computation of the eigenelements of the discrete Hamiltonian $M[A]$ is performed by using the MATLAB function `eigs` [36]. For details concerning the practical implementation of the scheme, one can refer to [20].

The external potential is a discontinuous function playing the role of a double barrier structure potential and the initial density n^0 is concentrated on the left of the double barrier (see Figure 5.1). The initial step involves the inversion of the formula (4.11), i.e., the computation of the initial chemical potential A^0 corresponding to n^0. The calculation of A^0 is done by minimizing the strictly convex functional Φ defined in (4.12). Recall that A^0 is not used in the text following the algorithm.

In Figures 5.1, 5.2, 5.3, 5.4, and 5.5, we have represented, as functions of x, the density n, the total potential $V + V^{ext}$, and the electrochemical potential $A - V$ at the initial step and at different time steps: k = 3, 20, 100, 500. The parameters of these computations are the following:

Δx	Δt	h^2	α
0.01	0.005	0.02	0.1

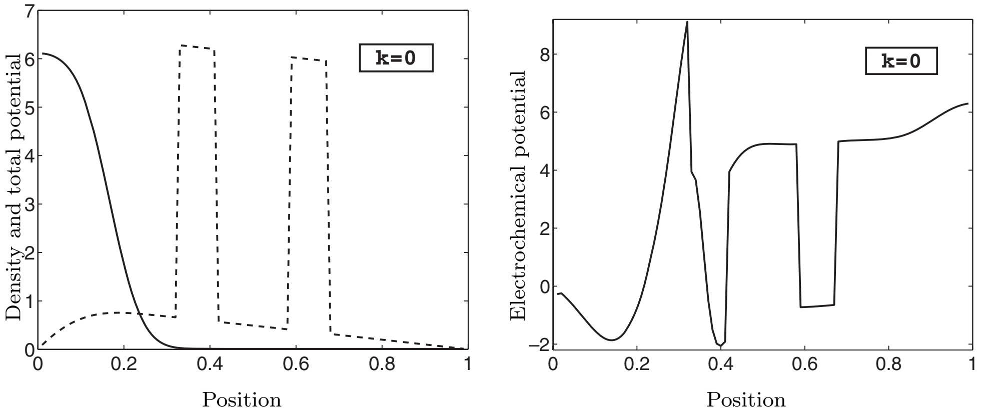


FIG. 5.1. *Numerical solution of the QDD model: Initial step. Left: The density $n(x)$ (solid line) and the total potential $(V + V^{ext})(x)$(dashed line) as functions of the position x. Right: The electrochemical potential $(A - V)(x)$.*

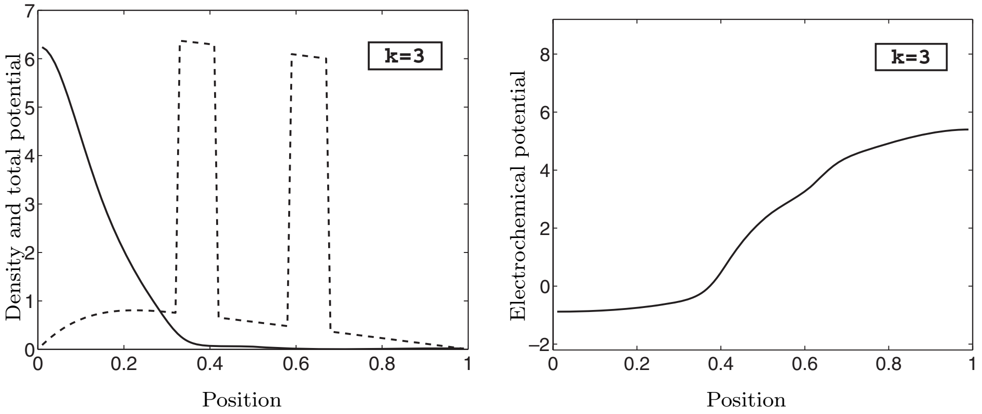


FIG. 5.2. *Numerical solution of the QDD model, after* 3 *iterations. The same quantities as in Figure* 5.1 *are represented.*

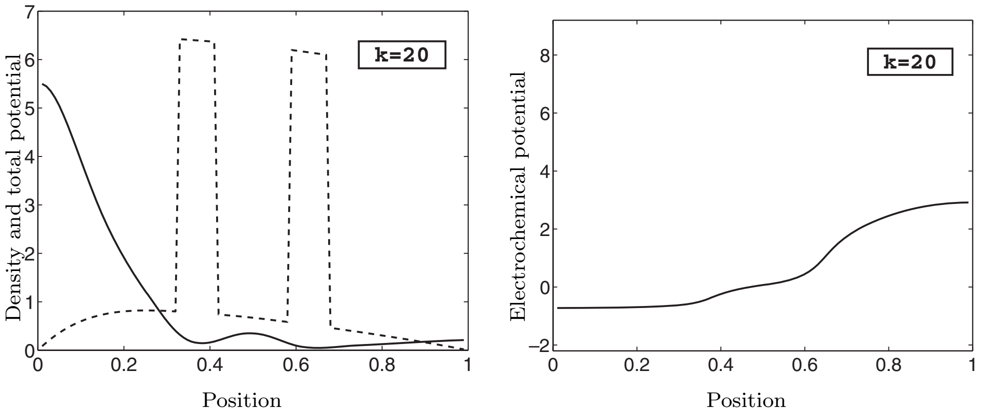


FIG. 5.3. *Numerical solution of the QDD model, after* 20 *iterations. The same quantities as in Figure* 5.1 *are represented.*

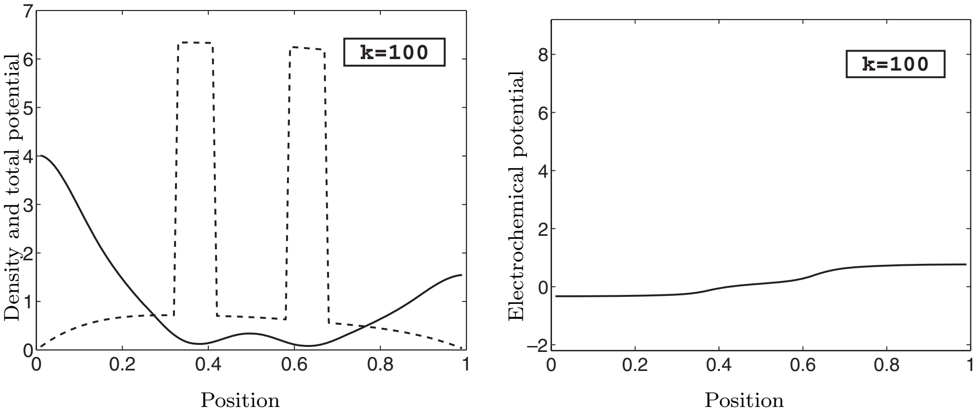


Fig. 5.4. *Numerical solution of the QDD model, after* 100 *iterations. The same quantities as in Figure* 5.1 *are represented.*

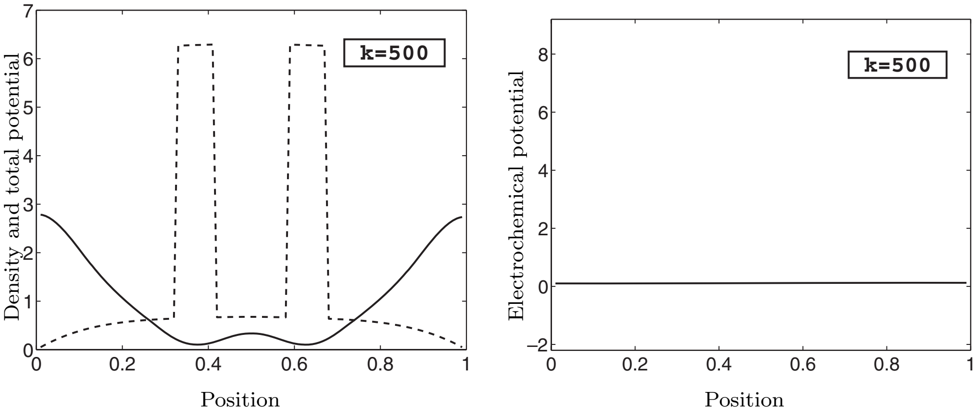


Fig. 5.5. *Numerical solution of the QDD model, after* 500 *iterations. The same quantities as in Figure* 5.1 *are represented.*

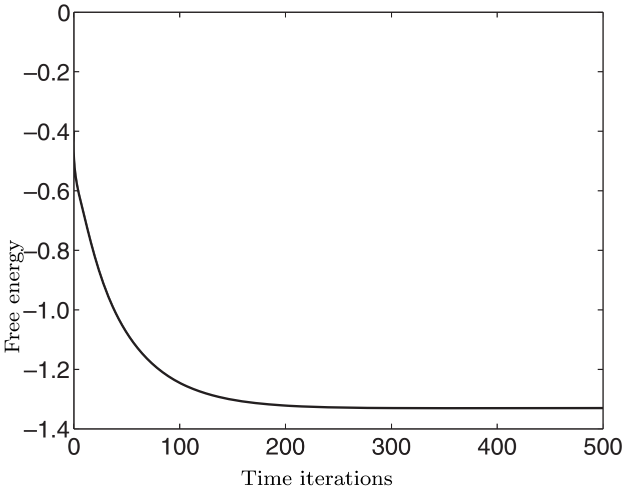


Fig. 5.6. *Free energy S^k as a function of the time step k.*

On the right-hand side of these figures, one can check that the electrochemical potential converges to a constant: at time $t = 500\Delta t$, one can consider that the system has converged to a steady state, which solves a discrete Schrödinger–Poisson system. In Figure 5.6, we show the evolution of the free energy S^k defined by (4.7) and check that it is a decreasing function, converging to a constant. In these simulations, the initial total charge is equal to 1 and this quantity is conserved during the evolution, up to a relative error of $10^{-4}\,\%$.

6. Conclusion. We have introduced a semidiscrete (in time) version (3.1)–(3.3) of the QDD model (2.8)–(2.10). We have proved that this system is well-posed and that its resolution amounts to minimizing a convex functional. Moreover, this semidiscrete model has the following interesting properties: it preserves the total charge and the positivity of the density and it dissipates the free energy. Then we have defined the numerical scheme (4.1)–(4.3) by discretizing the space variable in this system. As a consequence, this scheme possesses the same properties as the semidiscrete model. Finally, we have given some results of numerical simulations which have been performed with this scheme.

A lot of open questions arise naturally. Let us list a few of them. By passing formally to the limit in the semidiscrete model as Δt goes to zero, one obtains a solution of the initial QDD model. To make this statement rigorous, one of the most difficult points to be solved seems to be to find a bound from below for the density. Studying the long-time behavior of the semidiscrete model or the continuous model is also an interesting challenge: do their solutions converge to the solution of the Schrödinger–Poisson system studied in [39, 40]? Another important question is concerned with boundary conditions. We have chosen no-flux boundary conditions, but for practical use it is necessary to enable a current flow through the boundary. This issue will be investigated in a future work.

Acknowledgments. The authors wish to express their gratitude to Pierre Degond for initiating this work and for many helpful discussions. They would also like to thank Naoufel Ben Abdallah and Joachim Rehberg for fruitful discussions.

REFERENCES

[1] M. G. Ancona, *Diffusion-drift modeling of strong inversion layers*, COMPEL, 6 (1987), pp. 11–18.

[2] M. G. Ancona and G. J. Iafrate, *Quantum correction of the equation of state of an electron gas in a semiconductor*, Phys. Rev. B, 39 (1989), pp. 9536–9540.

[3] M. G. Ancona, Z. Yu, R. W. Dutton, P. J. Voorde, M. Cao, and D. Vook, *Density-gradient analysis of MOS tunneling*, IEEE Trans. Electron. Dev., 47 (2000), pp. 2310–2319.

[4] A. Arnold, J. L. Lopez, P. A. Markowich, and J. Soler, *An analysis of quantum Fokker–Planck models: A Wigner function approach*, Rev. Mat. Iberoamericana, 20 (2004), pp. 771–814.

[5] R. Balian, *From Microphysics to Macrophysics*, Springer, Berlin, 1982.

[6] N. Ben Abdallah, F. Méhats, and N. Vauchelet, *Analysis of a drift-diffusion-Schrödinger–Poisson system*, C. R. Acad. Sci. Paris Ser. I Math., 335 (2002), pp. 1007–1012.

[7] N. Ben Abdallah and A. Unterreiter, *On the stationary quantum drift-diffusion model*, Z. Angew. Math. Phys., 49 (1998), pp. 251–275.

[8] J.-P. Bourgade, F. Méhats, and C. Ringhofer, *Phonon Collision Operators Consistent with Quantum Entropy Relaxation and Quantum Spherical Harmonics Expansion Models*, submitted.

[9] S. Datta, *Nanoscale device modeling: The Green's function method*, Superlattices and Microstructures, 28 (2000), pp. 253–278.

[10] R. Dautray and J.-L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, *Vol.* 5, *Spectre des opérateurs*, INSTN: Collection Enseignement, Masson, Paris, 1988.

[11] P. Degond, *Mathematical modelling of microelectronics semiconductor devices*, in Proceedings of the Morningside Mathematical Center, Beijing, AMS/IP Stud. Adv. Math., AMS, Providence, RI, 2000, pp. 77–109.

[12] P. Degond and A. El Ayyadi, *A coupled Schrödinger drift-diffusion model for quantum semiconductor device simulations*, J. Comput. Phys., 181 (2002), pp. 222–259.

[13] P. Degond, F. Méhats, and C. Ringhofer, *Quantum energy-transport and drift-diffusion models*, J. Statist. Phys., 118 (2005), pp. 625–665.

[14] P. Degond, F. Méhats, and C. Ringhofer, *Quantum hydrodynamic models derived from entropy principle*, in Nonlinear Partial Differential Equations and Related Analysis, Contemp. Math. 371, AMS, Providence, RI, 2005, pp. 107–131.

[15] P. Degond and C. Ringhofer, *Quantum moment hydrodynamics and the entropy principle*, J. Statist. Phys., 112 (2003), pp. 587–628.

[16] P. Degond and C. Ringhofer, *A note on quantum moment hydrodynamics and the entropy principle*, C. R. Math. Acad. Sci. Paris, 335 (2002), pp. 967–972.

[17] P. Degond and C. Ringhofer, *Binary quantum collision operators conserving mass momentum and energy*, C. R. Math. Acad. Sci. Paris, 336 (2003), pp. 785–790.

[18] M. V. Fischetti, *Theory of electron transport in small semiconductor devices using the Pauli Master equation*, J. Appl. Phys., 83 (1998), pp. 270–291.

[19] W. R. Frensley, *Boundary conditions for open quantum systems driven far from equilibrium*, Rev. Modern Phys., 62 (1990), pp. 745–791.

[20] S. Gallego, *Étude théorique et numérique du modèle de dérive-diffusion quantique*, Rapport de stage de DEA, Toulouse, 2004.

[21] S. Gallego and F. Méhats, *Numerical approximation of a quantum drift-diffusion model*, C. R. Acad. Sci. Paris Ser. I Math., 339 (2004), pp. 519–524.

[22] C. Gardner, *The quantum hydrodynamic model for semiconductor devices*, SIAM J. Appl. Math., 54 (1994), pp. 409–427.

[23] C. Gardner and C. Ringhofer, *The smooth quantum potential for the hydrodynamic model*, Phys. Rev. E, 53 (1996), pp. 157–167.

[24] C. Gardner and C. Ringhofer, *The Chapman–Enskog expansion and the quantum hydrodynamic model for semiconductor devices*, VLSI Design, 10 (2000), pp. 415–435.

[25] I. Gasser and A. Jüngel, *The quantum hydrodynamic model for semiconductors in thermal equilibrium*, Z. Angew. Math. Phys., 48 (1997), pp. 45–59.

[26] I. Gasser and P. A. Markowich, *Quantum hydrodynamics, Wigner transforms and the classical limit*, Asymptot. Anal., 14 (1997), pp. 97–116.

[27] I. Gasser, P. Markowich, and C. Ringhofer, *Closure conditions for classical and quantum moment hierarchies in the small temperature limit*, Transport Theory Statist. Phys., 25 (1996), pp. 409–423.

[28] J. Jerome, *Analysis of Charge Transport. A Mathematical Study of Semiconductor Devices*, Springer, Berlin, 1996.

[29] A. Jüngel, *Quasi-hydrodynamic Semiconductor Equations*, Progr. Nonlinear Differential Equations Appl. 41, Birkhäuser, Boston, 2001.

[30] A. Jüngel and R. Pinnau, *A positivity preserving numerical scheme for a fourth-order parabolic equation*, SIAM J. Numer. Anal., 39 (2001), pp. 385–406.

[31] H.-C. Kaiser and J. Rehberg, *About a stationary Schrödinger–Poisson system with Kohn–Sham potential in a bounded two- or three-dimensional domain*, Nonlinear Anal., 41 (2000), pp. 33–72.

[32] N. C. Kluksdahl, A. M. Kriman, D. K. Ferry, and C. Ringhofer, *Self-consistent study of the resonant-tunneling diode*, Phys. Rev. B, 39 (1989), pp. 7720–7735.

[33] C. D. Levermore, *Moment closure hierarchies for kinetic theories*, J. Statist. Phys., 83 (1996), pp. 1021–1065.

[34] E. Madelung, *Quantentheorie in hydrodynamischer form*, Z. Phys., 40 (1926), pp. 322–326.

[35] P. A. Markowich, C. Ringhofer, and C. Schmeiser, *Semiconductor Equations*, Springer, Berlin, 1990.

[36] *MATLAB Documentation: The eigs Function*, The MathWorks, 1994–2004; available online from http://www.mathworks.com/access/helpdesk/help/techdoc/ref/eigs.html.

[37] M. Mock, *Analysis of Mathematical Models of Semiconductor Devices*, Book Press, Dublin, 1983.

[38] P. Mounaix, O. Vanbésien, and D. Lippens, *Effect of cathode spacer layer on the current voltage characteristics of resonant tunneling diodes*, Appl. Phys. Lett., 57 (1990), pp. 1517–1519.

[39] F. Nier, *A stationary Schrödinger–Poisson system arising from the modelling of electronic devices*, Forum Math., 2 (1990), pp. 489–510.

[40] F. Nier, *A variational formulation of Schrödinger–Poisson systems in dimension* $d \leq 3$, Comm. Partial Differential Equations, 18 (1993), pp. 1125–1147.

[41] R. Pinnau and A. Unterreiter, *The stationary current-voltage characteristics of the quantum drift diffusion model*, SIAM J. Numer. Anal., 37 (1999), pp. 211–245.

[42] R. Pinnau, *The linearized transient quantum drift diffusion model: Stability of stationary states*, ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 327–344.

[43] A. Pirovano, A. Lacaita, and A. Spinelli, *Two dimensional quantum effects in nanoscale MOSFETs*, IEEE Trans. Electron. Dev., 49 (2002), pp. 25–31.

[44] C. Pohl, *On the Numerical Treatment of Dispersive Equations*, Ph.D. thesis, TU, Berlin, 1998.

[45] E. Polizzi, *Modélisation et simulations numériques du transport quantique balistique dans les nanostructures semi-conductrices*, Ph.D. thesis, INSA, Toulouse, 2001.

[46] E. Polizzi and N. Ben Abdallah, *Self-consistent three dimensional models for quantum ballistic transport in open systems*, Phys. Rev. B, 66 (2002), pp. 245–301.

[47] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, *Vol.* 4. *Analysis of Operators*, Academic Press, New York, 1978.

[48] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*, Springer, Berlin, 1984.

[49] D. Serre, *Matrices. Theory and Applications*, translated from the 2001 French original, Grad. Texts in Math. 216, Springer, New York, 2002.

[50] E. Wigner, *On the quantum correction for thermodynamic equilibrium*, Phys. Rev., 40 (1932), pp. 749–759.

[51] J. R. Zhou and D. Ferry, *Modeling of quantum effects in ultrasmall HEMT devices*, IEEE Trans. Electron. Dev., 40 (1993), pp. 421–427.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 1850–1871

ON CONVERGENCE OF THE ADDITIVE SCHWARZ PRECONDITIONED INEXACT NEWTON METHOD*

HENG-BIN AN†

Abstract. The additive Schwarz preconditioned inexact Newton (ASPIN) method was recently introduced [X.-C. Cai and D. E. Keyes, *SIAM J. Sci. Comput.*, 24 (2002), pp. 183–200] to solve the systems of nonlinear equations with nonbalanced nonlinearities. Although the ASPIN method has successfully been used to solve some difficult nonlinear equations, its convergence property has not been studied since it was proposed. In this paper, the convergence property of the ASPIN method is studied, and the obtained result shows that this method is locally convergent. Furthermore, the convergence rate for the ASPIN method is discussed and the obtained result is similar to that of the inexact Newton method.

Key words. nonlinear systems, inexact Newton methods, nonlinear preconditioning, additive Schwarz, local convergence, convergence rate

AMS subject classifications. 65H10, 65N12, 65N22

DOI. 10.1137/040611653

1. Introduction.

Consider the nonlinear system of equations

$$F(u) = 0, \tag{1.1}$$

where $F : R^n \to R^n$ is a continuously differentiable function. For convenience of discussion, let $F = (F_1, F_2, \dots, F_n)^T$, $u = (u_1, u_2, \dots, u_n)^T$, and $J(u) = F'(u)$. A numerical solution for (1.1) is often required in many scientific and engineering computing areas such as the discretization of nonlinear partial differential equations; see [7, 16]. The inexact Newton method [8] is one of the most important and effective tools for solving such systems, in particular, when the problem is large and sparse. In applications, some global strategies, such as linesearch or trust region techniques, are often needed because the inexact Newton method is locally convergent [1, 2, 3, 4, 12]. In particular, if the linesearch backtracking technique is augmented in the inexact Newton method, then the inexact Newton with backtracking (INB) method is obtained [12, 13, 19]. This method is more robust and it can be briefly described here. Suppose $u^{(0)}$ is a given initial guess and let $u^{(k)}$ be the current approximate solution; the next approximate solution $u^{(k+1)}$ can be obtained through the following steps.

ALGORITHM 1.1 (INB [12]).

1. Inexactly solve the system

$$J\big(u^{(k)}\big)p = -F\big(u^{(k)}\big), \tag{1.2}$$

and obtain an inexact Newton direction $p^{(k)}$ such that

$$\|F\big(u^{(k)}\big) + J\big(u^{(k)}\big)p^{(k)}\| \le \eta_k \|F\big(u^{(k)}\big)\|. \tag{1.3}$$

2. Compute the new approximate solution

$$u^{(k+1)} = u^{(k)} + \lambda_k p^{(k)}.$$

*Received by the editors June 24, 2004; accepted for publication (in revised form) February 11, 2005; published electronically November 22, 2005.
http://www.siam.org/journals/sinum/43-5/61165.html

†State Key Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, P.O. Box 8009, Beijing 100088, People's Republic of China (anhengbin@sina.com).

Here $\eta_k \in [0,1)$ is the forcing term that controls how accurately system (1.2) should be solved, and $p^{(k)}$ is the inexact Newton direction of F at $u^{(k)}$. Step 2 in Algorithm 1.1 is a linesearch procedure that is used to find a satisfied step factor $\lambda_k \in (0,1]$ and then form the next approximate solution.

Usually, we use linear iterative methods, such as the classical splitting method or the modern Krylov subspace method, to inexactly solve system (1.2). Thus, the inexact Newton method is an inner-outer iterative method. In particular, when the Krylov subspace method is used in an inner iteration, we obtain the Newton–Krylov subspace method, which has been used successfully in many areas [1, 2, 3, 4, 16].

Although the inexact Newton method works very well for most nonlinear equations, this may often fail when it is used to solve some difficult problems. Many numerical experiments show that most failed cases in the inexact Newton method result from stagnation, particularly when it is used to solve some problems with non-balanced nonlinearities [1, 6]. Usually, the stagnation phenomenon is caused by the lack of a good initial guess and/or problematic regions such as boundary layers, singularities in the domain, and/or multiphysics domain, etc. See [17]. Considering this, Cai and Keyes [6] recently proposed a nonlinearly preconditioned inexact Newton algorithm: first convert system (1.1) into another nonlinear system $\mathcal{F}(u) = 0$ such that the two systems have the same solution $u^* \in R^n$; then use Algorithm 1.1 to solve $\mathcal{F}(u) = 0$.

$\mathcal{F}$ and F may have completely different forms, but they must have the same solution. Usually, $\mathcal{F}$ has more uniform nonlinearities, so it is relatively easy to solve. In [6], an especially preconditioned case, where $\mathcal{F}$ is obtained by the single-level nonlinear additive Schwarz method, is discussed in detail. The corresponding method is the additive Schwarz preconditioned inexact Newton (ASPIN) method. Numerical results in [6] show that the ASPIN method can solve some difficult problems where the traditional inexact Newton method fails.

Although the ASPIN method has better numerical results than the traditional inexact Newton method, it is unfortunate that until now the convergence property for the ASPIN method has not been given much importance except for some preliminary convergence analysis in the context of semilinear PDEs in paper [17]. In this paper, we show that the ASPIN method is locally convergent; thus we give theoretical support for the ASPIN method. Moreover, we will discuss the convergence rate of the ASPIN method.

The rest of the paper is organized as follows. In section 2, we briefly discuss the ASPIN method, and some of its properties are listed. In section 3, we show that the ASPIN method is locally convergent, and its convergence rate is discussed in section 4. Finally, in section 5, some brief conclusions are given.

2. The ASPIN method.

Assume that $F(u^*) = 0$ and $J(u^*)$ is invertible. To find the solution u^* of system (1.1), the ASPIN method solves another nonlinear system $\mathcal{F}(u) = 0$, which is obtained from (1.1) through the additive Schwarz preconditioning technique. Specifically, the ASPIN method can be described as follows.

Let

$$S = \{1, 2, \ldots, n\}$$

be an index set, i.e., one integer corresponds to each u_i and F_i. Assume that S has a partition $\{S_1, S_2, \ldots, S_N\}$ such that

$$\bigcup_{i=1}^{N} S_i = S \quad \text{and} \quad S_i \subset S.$$

Here the subsets may overlap. Let $n_i = |S_i|$ be the dimension of S_i; then

$$\sum_{i=1}^{N} n_i \geq n.$$

Assume that

$$S_i = \{i_1, i_2, \ldots, i_{n_i}\},$$

where $i_1 < i_2 < \cdots < i_{n_i}$. For $i = 1, 2, \ldots, N$, define matrices $E_i \in R^{n_i \times n}$ by

$$(E_i)_{k,l} = \begin{cases} 1, & l = i_k, \\ 0, & l \neq i_k. \end{cases}$$

Let

$$P_i = E_i^T E_i$$

and

$$V_i = P_i R^n, \quad F_{S_i} = P_i F.$$

It is easy to see that P_i is the orthogonal projection from R^n onto V_i.

For each $u \in R^n$, we define $T_i(u) \in V_i$ such that

$$F_{S_i}(u - T_i(u)) = 0, \quad i = 1, 2, \ldots, N, \tag{2.1}$$

and let

$$\mathcal{F}(u) = \sum_{i=1}^{N} T_i(u),$$

which is referred to as the additive Schwarz preconditioned nonlinear function. The ASPIN method tries to find the solution u^* of (1.1) by solving the nonlinear system

$$\mathcal{F}(u) = 0 \tag{2.2}$$

with the inexact Newton method.

About the solvability of (2.1), we have the following proposition.

Proposition 2.1. *If $E_i J(u^*) E_i^T$ is invertible for each i, then there exists a neighborhood U of u^* and a unique continuously differentiable function $T_i : R^n \to V_i$ for each i such that* (2.1) *holds for each $u \in U$, and also $T_i(u^*) = 0$. Moreover,*

$$T_i'(u) = E_i^T \left[E_i J(u - T_i(u)) E_i^T\right]^{-1} E_i J(u - T_i(u)). \tag{2.3}$$

Proof. Theorem 1.1 in [10] shows that there exist a neighborhood U_1 of u^* and a unique continuous function $T_i : R^n \to V_i$ for each i such that (2.1) holds for each $u \in U_1$, and $T_i(u^*) = 0$. Also, we know from [10] that T_i satisfies

$$T_i(v) - T_i(u) = DT_i(v, u)(v - u), \quad v, u \in U_1, \tag{2.4}$$

where

$$DT_i(v, u) = E_i^T \left[E_i DF(v - T_i(v), u - T_i(u)) E_i^T\right]^{-1} E_i DF(v - T_i(v), u - T_i(u)),$$

while

$$DF(v - T_i(v), u - T_i(u)) = \int_0^1 J([v - T_i(v)] + t[(u - T_i(u)) - (v - T_i(v))])\, dt.$$

Since $J(u)$ is continuous and $E_i J(u^*) E_i^T$ is invertible for each i, Lemma 2.3.3 in [20] shows that there exists a neighborhood U_2 of u^* such that $E_i J(u) E_i^T$ is invertible for each $u \in U_2$. Now let $U \subset U_1 \cap U_2$ be a neighborhood of u^* such that $u - T_i(u) \in U_2$ for each $u \in U$ and for each i.

For $u \in U$, let

$$A(u) = E_i^T [E_i J(u - T_i(u)) E_i^T]^{-1} E_i J(u - T_i(u));$$

then, according to (2.4) and Lemma 2.3.3 in [20], it is easy to verify that

$$\lim_{\|h\| \to 0} \frac{\|T_i(u + h) - T_i(u) - A(u)h\|}{\|h\|} = 0.$$

Thus, $T_i(u)$ is Fréchet-differentiable, and

$$T_i'(u) = A(u) = E_i^T [E_i J(u - T_i(u)) E_i^T]^{-1} E_i J(u - T_i(u)).$$

In addition, it is easy to see from Lemma 2.3.3 in [20] that $T_i'(u)$ is continuous. □

It should be pointed out that formula (2.3) has been given in [6], but it has been obtained in a different way. In addition, it should be noted that the condition of Proposition 2.1 is satisfied for any partition of S if $J(u^*)$ is positive definite.

Since the inexact Newton method concerns the Jacobian of the system, an analysis of the basic property of the Jacobian $\mathcal{F}'(u)$ is necessary. Because $T_i(u)$ is continuous and $T_i(u^*) = 0$, we know that when u is sufficiently close to u^*, $T_i(u)$ will be sufficiently close to 0, and as a result, $u - T_i(u)$ will be close to u. Since $J(u)$ is continuous, we may replace $J(u - T_i(u))$ by $J(u)$; therefore,

$$\begin{aligned} T_i'(u) &= E_i^T \left(E_i J(u - T_i(u)) E_i^T\right)^{-1} E_i J(u - T_i(u)) \\ &\approx E_i^T \left(E_i J(u) E_i^T\right)^{-1} E_i J(u) \\ &\equiv R_i(u). \end{aligned}$$

Let

$$\mathcal{J}(u) = \mathcal{F}'(u);$$

then

$$\mathcal{J}(u) = \sum_{i=1}^N T_i'(u) \approx \sum_{i=1}^N R_i(u) \equiv B(u).$$

In implementation of the ASPIN method, the Jacobian $\mathcal{J}(u)$ is replaced by $B(u)$, since the latter is easier to use.

Remark 2.1. *From the proof of Proposition* 2.1, *we know that if $E_i J(u^*) E_i^T$ is invertible for each i, then $E_i J(u - T_i(u)) E_i^T$ and $E_i J(u) E_i^T$ are invertible in the neighborhood U of u^*. In addition, $(E_i J(u - T_i(u)) E_i^T)^{-1}$ and $(E_i J(u) E_i^T)^{-1}$ are continuous in U, so $\mathcal{J}(u)$ and $B(u)$ are all continuous in U. Furthermore, it is easy to see that*

$$\lim_{u \to u^*} \mathcal{J}(u) = \lim_{u \to u^*} B(u) = \sum_{i=1}^N E_i^T \left(E_i J(u^*) E_i^T\right)^{-1} E_i J(u^*) = \mathcal{J}(u^*). \tag{2.5}$$

From [6] and [10], we can obtain the following result.

PROPOSITION 2.2. *If $E_i J(u^*) E_i^T$ is invertible for each i, then there exists a neighborhood $D \subset U$ of u^* with U determined in Proposition* 2.1 *such that*

(i) *$\mathcal{J}(u)$ and $B(u)$ are nonsingular in D;*

(ii) *the nonlinear systems* (1.1) *and* (2.2) *are equivalent in the sense that they have the same solution in D.*

REMARK 2.2. *Remark* 2.1 *points out that $\mathcal{J}(u)$ and $B(u)$ are all continuous in $D \subset U$; thus Lemma* 2.3.3 *in* [20] *and Proposition* 2.2 *show that $\mathcal{J}(u)^{-1}$ and $B(u)^{-1}$ are continuous in D.*

For convenience of discussion, we describe the ASPIN algorithm here. Let

$$f(u) = \frac{1}{2}\mathcal{F}(u)^T \mathcal{F}(u),$$

which will be used in linesearch in the inexact Newton method. Assume that $u^{(0)}$ is a given initial guess and $u^{(k)}$ is the current approximate solution; the next approximate solution $u^{(k+1)}$ for system (2.2) can be computed through the following steps.

ALGORITHM 2.1 (ASPIN [6]).

```
0. Let $\eta_{max} \in (0,1)$, $\alpha \in (0,1)$, $0 < \theta_{min} < \theta_{max} < 1$ be given.
1. Compute the nonlinear residual $g^{(k)} = \mathcal{F}(u^{(k)})$ through the following steps.
  1.1. Find $g_i^{(k)} = T_i(u^{(k)})$ by solving the local subdomain nonlinear systems
         $F_{S_i}\big(u^{(k)} - g_i^{(k)}\big) = 0, \quad i = 1, 2, \ldots, N,$
       with the initial point $g_i^{(k)} = 0$.
  1.2. Form the global residual
         $g^{(k)} = \sum_{i=1}^{N} g_i^{(k)}.$
  1.3. Check the stopping conditions on $g^{(k)}$.
2. Find the approximate inexact Newton direction $p^{(k)}$ by solving the system
         $B(u^{(k)})p = -\mathcal{F}\big(u^{(k)}\big)$
   such that
         $\big\|\mathcal{F}\big(u^{(k)}\big) + B\big(u^{(k)}\big)p^{(k)}\big\| \le \eta_k \big\|\mathcal{F}\big(u^{(k)}\big)\big\|,$
   where $\eta_k \in [0, \eta_{max}]$ is the forcing term.
3. Perform linesearch along $p^{(k)}$:
  3.1. Let $\lambda_k = 1$.
  3.2. While $f(u^{(k)} + \lambda_k p^{(k)}) > f(u^{(k)}) + \alpha\lambda_k \mathcal{F}(u^{(k)})^T B(u^{(k)})p^{(k)}$, do
         • choose $\theta \in [\theta_{min}, \theta_{max}]$,
         • let $\lambda_k = \theta\lambda_k$.
  3.3. Let $u^{(k+1)} = u^{(k)} + \lambda_k p^{(k)}$.
```

In step 1.1 of Algorithm 2.1, N subdomain nonlinear systems have to be solved in order to evaluate the preconditioned function $\mathcal{F}$ at a given point. Step 3 of Algorithm 2.1 is the linesearch procedure to find a satisfied step. For more details about the ASPIN method, see [6].

We point out that Algorithm 2.1 can be implemented in parallel. For details about implementation, see [6, 7].

3. Local convergence of the ASPIN method. We will prove in this section that the ASPIN method is locally convergent. Note that by Propositions 2.1 and 2.2, the ASPIN method is based on the local property of $F(u)$ at the solution u^* of system (1.1), so it seems impossible to obtain a global convergence result for this method.

In this section and the following, $\|\cdot\|$ always denotes the Euclidean norm for both vectors and matrices, and $N(u,\rho) = \{v \mid \|v-u\| < \rho\}$ represents the open ball with center u and radius ρ.

Since the analysis of secondary iteration would complicate the discussion without gaining more insight into the method, we assume that

(A$_1$) the value of $\mathcal{F}$ at each iterative point is evaluated exactly, i.e., (2.1) holds with u replaced by $u^{(k)} + \lambda_k p^{(k)}$ for each k; moreover, from now on, we assume that

(A$_2$) $E_i J(u^*) E_i^T$ is invertible for each i;

(A$_3$) D represents the neighborhood determined in Proposition 2.2; and

(A$_4$) $\delta > 0$ is a fixed small number such that $N(u^*, \delta) \subset D$; in addition, the following inequalities hold for any $u \in N(u^*, \delta)$:

(I$_1$) $\|B(u)\| \le 2M$;

(I$_2$) $\|B(u)^{-1}\| \le 2M$;

(I$_3$) $\|\mathcal{J}(u)\| \le 2M$;

(I$_4$) $\|\mathcal{J}(u)^{-1}\| \le 2M$;

(I$_5$) $\|\mathcal{J}(u) - B(u)\| \le \frac{1-\eta_{max}}{4M(1+\eta_{max})}$;

(I$_6$) $\|\mathcal{F}(u) - \mathcal{F}(u^*) - \mathcal{J}(u^*)(u-u^*)\| \le \frac{1}{2M}\|u-u^*\|$,

where

$$M := \max\{\|\mathcal{J}(u^*)\|, \|\mathcal{J}(u^*)^{-1}\|\}.$$

It is easy to see from Remarks 2.1 and 2.2 that inequalities (I$_1$)–(I$_5$) may hold with δ small enough. The last inequality may hold by Lemma 3.2.10 in [20]. In addition, we assume that the parameter α in Algorithm 2.1 is small enough so that

$$64\alpha M^4 \le \frac{1-\eta_{max}}{3+\eta_{max}}. \tag{3.1}$$

It should be pointed out that the above assumptions are not so strict; see the appendix, where an example is given.

Now we show that the inexact Newton direction computed in Algorithm 2.1 is also a regular inexact Newton direction for $\mathcal{F}$ in the sense that (1.3) holds.

Proposition 3.1. *Assume that* $u \in N(u^*, \delta)$. *If*

$$\|\mathcal{F}(u) + B(u)p\| \le \eta \|\mathcal{F}(u)\|, \quad \eta \in [0, \eta_{max}], \tag{3.2}$$

then

$$\|\mathcal{F}(u) + \mathcal{J}(u)p\| \le \frac{1+\eta}{2}\|\mathcal{F}(u)\|. \tag{3.3}$$

Proof. By (I$_2$) and (3.2),

$$\begin{aligned}\|p\| &= \|B(u)^{-1}[B(u)p + \mathcal{F}(u) - \mathcal{F}(u)]\| \\ &\le (1+\eta)\|B(u)^{-1}\|\|\mathcal{F}(u)\| \\ &\le 2M(1+\eta)\|\mathcal{F}(u)\|.\end{aligned}$$

Thus, according to (I_5) and (3.2),

$$\begin{aligned}\|\mathcal{J}(u)p+\mathcal{F}(u)\| &= \|[\mathcal{J}(u)-B(u)]p+B(u)p+\mathcal{F}(u)\|\\ &\le \|\mathcal{J}(u)-B(u)\|\|p\|+\|B(u)p+\mathcal{F}(u)\|\\ &\le \frac{1-\eta_{max}}{4M(1+\eta_{max})}\cdot 2M(1+\eta)\|\mathcal{F}(u)\|+\eta\|\mathcal{F}(u)\|\\ &\le \frac{1-\eta}{2}\|\mathcal{F}(u)\|+\eta\|\mathcal{F}(u)\|\\ &\le \frac{1+\eta}{2}\|\mathcal{F}(u)\|.\end{aligned}$$

Thus we obtain the required inequality. □

REMARK 3.1. *If $u \in N(u^*,\delta)$ and (3.2) holds, then we have*

$$\begin{aligned}\mathcal{F}(u)^T B(u)p &= \mathcal{F}(u)^T[B(u)p+\mathcal{F}(u)-\mathcal{F}(u)]\\ &\le -\|\mathcal{F}(u)\|^2+\|\mathcal{F}(u)\|\|B(u)p+\mathcal{F}(u)\|\\ &\le -(1-\eta)\|\mathcal{F}(u)\|^2\\ &< 0.\end{aligned}$$

In the same way, (3.3) *shows that*

$$\mathcal{F}(u)^T\mathcal{J}(u)p<0. \tag{3.4}$$

In particular, (3.4) *shows that if $u \in N(u^*,\delta)$ and p is computed by the ASPIN method, then p is a descent direction for the function $f(u)=\frac{1}{2}\|\mathcal{F}(u)\|^2$ at point u.*[1]

The following lemma is needed in our analysis.

LEMMA 3.2 (see [4, Lemma 3.4]). *Let $u \in R^n$ and $H : R^n \to R^n$ be continuously differentiable in a neighborhood of u. Assume that $H(u) \neq 0$ and $H'(u)$ is nonsingular. If $p \in R^n$ such that*

$$\|H(u)+H'(u)p\| \le \eta\|H(u)\|, \quad \eta\in[0,1),$$

then

$$\frac{|\nabla h(u)^T p|}{\|p\|} \ge \frac{1-\eta}{(1+\eta)\kappa(H'(u))}\|\nabla h(u)\|>0,$$

where $\kappa(H'(u))$ is the condition number for $H'(u)$ and $h(u)=\frac{1}{2}H(u)^T H(u)$.

Based on Lemma 3.2, we have the following result.

LEMMA 3.3. *Assume that $u \in N(u^*,\delta)$. If*

$$\|\mathcal{F}(u)+B(u)p\| \le \eta\|\mathcal{F}(u)\|, \quad \eta\in[0,\eta_{max}],$$

then it holds that

$$|\mathcal{F}(u)^T\mathcal{J}(u)p| \ge 4\alpha|\mathcal{F}(u)^T B(u)p|.$$

Proof. According to Proposition 3.1, we have

$$\|\mathcal{F}(u)+\mathcal{J}(u)p\| \le \frac{1+\eta}{2}\|\mathcal{F}(u)\|;$$

[1] p is a descent direction for $f(u)$ if $\nabla f(u)^T p<0$. In addition, note that $\nabla f(u)=\mathcal{J}(u)^T\mathcal{F}(u)$.

therefore, by Lemma 3.2, (I_3), and (I_4),

$$\begin{aligned}|\mathcal{F}(u)^T\mathcal{J}(u)p| &\geq \frac{1-\eta}{(3+\eta)\kappa(\mathcal{J}(u))}\|\mathcal{F}(u)^T\mathcal{J}(u)\|\|p\| \\ &\geq \frac{1-\eta}{(3+\eta)\kappa(\mathcal{J}(u))\|\mathcal{J}(u)^{-1}\|}\|\mathcal{F}(u)\|\|p\| \\ &\geq \frac{1-\eta}{8M^3(3+\eta)}\|\mathcal{F}(u)\|\|p\|. \qquad (3.5)\end{aligned}$$

Thus, by (I_1), (3.5), and (3.1),

$$\begin{aligned}4\alpha|\mathcal{F}(u)^TB(u)p| &\leq 4\alpha\|B(u)\|\|\mathcal{F}(u)\|\|p\| \\ &\leq 8\alpha M\|\mathcal{F}(u)\|\|p\| \\ &\leq 8\alpha M\frac{8M^3(3+\eta)}{1-\eta}|\mathcal{F}(u)^T\mathcal{J}(u)p| \\ &\leq 64\alpha M^4\frac{3+\eta_{max}}{1-\eta_{max}}|\mathcal{F}(u)^T\mathcal{J}(u)p| \\ &\leq |\mathcal{F}(u)^T\mathcal{J}(u)p|.\end{aligned}$$

This concludes the proof. □

The following lemma shows that if u is sufficiently close to u^*, then the direction p obtained in the ASPIN method will not be too long.

LEMMA 3.4. *Assume that* $u \in N(u^*, \frac{\delta}{2})$ *with* $\|\mathcal{F}(u)\| \leq \frac{\delta}{8M}$*. If* $p \in R^n$ *such that*

$$\|\mathcal{F}(u) + B(u)p\| \leq \eta\|\mathcal{F}(u)\|, \quad \eta \in [0, \eta_{max}],$$

then $[u, u+p] \subset N(u^*, \delta)$*, where* $[u, u+p]$ *represents the line segment between* u *and* $u+p$*.*

Proof. Since $\|\mathcal{F}(u)\| \leq \frac{\delta}{8M}$, by ($\mathrm{I}_2$), we have

$$\begin{aligned}\|p\| &= \|B(u)^{-1}[B(u)p + \mathcal{F}(u) - \mathcal{F}(u)]\| \\ &\leq \|B(u)^{-1}\| \cdot [\|B(u)p + \mathcal{F}(u)\| + \|\mathcal{F}(u)\|] \\ &\leq 2M(1+\eta)\|\mathcal{F}(u)\| \\ &\leq 2M(1+\eta_{max})\|\mathcal{F}(u)\| \\ &\leq \frac{\delta(1+\eta_{max})}{4} \\ &< \frac{\delta}{2}.\end{aligned}$$

Therefore, by $u \in N(u^*, \frac{\delta}{2})$,

$$\|(u+p) - u^*\| \leq \|u - u^*\| + \|p\| < \frac{\delta}{2} + \frac{\delta}{2} = \delta,$$

that is, $u + p \in N(u^*, \delta)$. Since $u \in N(u^*, \delta)$ and $N(u^*, \delta)$ is a convex set, we have $[u, u+p] \subset N(u^*, \delta)$. □

We can now show the following theorem, which shows that the linesearch procedure along p will succeed with a nonzero step factor λ.

THEOREM 3.5. *Assume that* $u \in N(u^*, \frac{\delta}{2})$ *with*

$$\mathcal{F}(u) \neq 0, \quad \|\mathcal{F}(u)\| \leq \frac{\delta}{8M}.$$

In addition, assume that there exists $\gamma > 0$ *such that*

$$\|\nabla f(v) - \nabla f(w)\| \leq \gamma \|v - w\| \quad \forall\, v, w \in N(u^*, \delta). \tag{3.6}$$

If $p \in R^n$ *such that*

$$\|\mathcal{F}(u) + B(u)p\| \leq \eta \|\mathcal{F}(u)\|, \quad \eta \in [0, \eta_{max}],$$

then the linesearch procedure along p *in Algorithm* 2.1 *will terminate in finite iterations and the obtained* λ *satisfies*

$$\lambda \geq \min\left\{1, \frac{\alpha\theta_{min}|\mathcal{F}(u)^T\mathcal{J}(u)p|}{\gamma\|p\|^2}\right\}.$$

Proof. Because $u \in N(u^*, \frac{\delta}{2})$, Lemma 3.3 shows that

$$|\mathcal{F}(u)^T\mathcal{J}(u)p| \geq 4\alpha|\mathcal{F}(u)^T B(u)p|. \tag{3.7}$$

Since $\mathcal{F}(u)^T\mathcal{J}(u)p < 0$, $\mathcal{F}(u)^T B(u)p < 0$, (3.7) shows that

$$\mathcal{F}(u)^T\mathcal{J}(u)p \leq 4\alpha\mathcal{F}(u)^T B(u)p,$$

so

$$\mathcal{F}(u)^T\mathcal{J}(u)p - \alpha\mathcal{F}(u)^T B(u)p \leq 3\alpha\mathcal{F}(u)^T B(u)p < 0. \tag{3.8}$$

Because $u \in N(u^*, \frac{\delta}{2})$ and $\|\mathcal{F}(u)\| \leq \frac{\delta}{8M}$, Lemma 3.4 shows that $[u, u+p] \subset N(u^*, \delta)$. Thus, by the mean value theorem, there exists $\xi \in [u, u+p]$ such that

$$f(u + \lambda p) = f(u) + \lambda \nabla f(\xi)^T p.$$

Therefore,

$$\begin{aligned} f(u+\lambda p) &= f(u) + \lambda\nabla f(\xi)^T p \\ &= f(u) + \alpha\lambda\mathcal{F}(u)^T B(u)p - \alpha\lambda\mathcal{F}(u)^T B(u)p + \lambda\nabla f(\xi)^T p \\ &= f(u) + \alpha\lambda\mathcal{F}(u)^T B(u)p + \lambda\{[\nabla f(\xi)^T p - \nabla f(u)^T p] \\ &\quad +[\nabla f(u)^T p - \alpha\mathcal{F}(u)^T B(u)p]\} \\ &= f(u) + \alpha\lambda\mathcal{F}(u)^T B(u)p + \lambda\{\lambda\zeta + [\mathcal{F}(u)^T\mathcal{J}(u)p - \alpha\mathcal{F}(u)^T B(u)p]\}, \end{aligned}$$

where

$$\zeta = \frac{\nabla f(\xi)^T p - \nabla f(u)^T p}{\lambda}.$$

By (3.6), we have

$$|\zeta| \leq \gamma\|p\|^2,$$

so

$$f(u+\lambda p) \leq f(u) + \alpha\lambda\mathcal{F}(u)^T B(u)p + \lambda\left\{\lambda\gamma\|p\|^2 + [\mathcal{F}(u)^T\mathcal{J}(u)p - \alpha\mathcal{F}(u)^T B(u)p]\right\}.$$

Thus, if

$$\lambda\gamma\|p\|^2 + [\mathcal{F}(u)^T\mathcal{J}(u)p - \alpha\mathcal{F}(u)^T B(u)p] \leq 0,$$

then λ is acceptable. Since λ is reduced by a factor $\theta \leq \theta_{max} < 1$ at each iteration of the while-loop, it follows from (3.8) that the while-loop will terminate in finite steps.

Let λ be the ultimate step factor. If $\lambda = 1$, then the needed conclusion trivially holds. Now suppose that the linesearch procedure is implemented at least once, and let λ^- be the penultimate value; then the above argument shows that

$$\lambda^- \gamma \|p\|^2 + [\mathcal{F}(u)^T \mathcal{J}(u)p - \alpha \mathcal{F}(u)^T B(u)p] > 0.$$

Consequently,

$$\lambda^- > \frac{|\mathcal{F}(u)^T \mathcal{J}(u)p| - \alpha |\mathcal{F}(u)^T B(u)p|}{\gamma \|p\|^2}.$$

Therefore, it follows from (3.7) that

$$\begin{aligned} \lambda &\geq \theta_{min} \lambda^- \\ &> \theta_{min} \frac{|\mathcal{F}(u)^T \mathcal{J}(u)p| - \alpha |\mathcal{F}(u)^T B(u)p|}{\gamma \|p\|^2} \\ &\geq \theta_{min} \frac{4\alpha |\mathcal{F}(u)^T B(u)p| - \alpha |\mathcal{F}(u)^T B(u)p|}{\gamma \|p\|^2} \\ &\geq \frac{\alpha \theta_{min} |\mathcal{F}(u)^T B(u)p|}{\gamma \|p\|^2}. \end{aligned}$$

Thus, we have obtained the required conclusion. □

LEMMA 3.6. *If* $u \in N(u^*, \frac{\delta}{2})$ *and*

$$\|\mathcal{F}(u)\| < \frac{\delta}{8M}, \quad \mathcal{F}(u) \neq 0,$$

then $u_+ \in N(u^*, \frac{\delta}{2})$, *where* $u_+ = u + s$ *and* s *is a step such that*

$$\begin{aligned} \|\mathcal{F}(u) + B(u)s\| &\leq \|\mathcal{F}(u)\|, \\ \|\mathcal{F}(u_+)\| &< \|\mathcal{F}(u)\|. \end{aligned}$$

Proof. Let $y \in N(u^*, \delta)$; then by (I_6),

$$\begin{aligned} \|\mathcal{F}(y)\| &\geq \|\mathcal{J}(u^*)(y - u^*)\| - \|\mathcal{F}(y) - \mathcal{F}(u^*) - \mathcal{J}(u^*)(y - u^*)\| \\ &\geq \frac{1}{\|\mathcal{J}(u^*)^{-1}\|} \|y - u^*\| - \frac{1}{2M} \|y - u^*\| \\ &\geq \frac{1}{M} \|y - u^*\| - \frac{1}{2M} \|y - u^*\| \\ &= \frac{1}{2M} \|y - u^*\|. \end{aligned}$$

So

$$\|y - u^*\| \leq 2M \|\mathcal{F}(y)\|$$

whenever $y \in N(u^*, \delta)$.

By (I_2) and the assumption that $\|\mathcal{F}(u)\| < \frac{\delta}{8M}$, we have

$$\begin{aligned}\|s\| &= \|B(u)^{-1}\{[B(u)s + \mathcal{F}(u)] - \mathcal{F}(u)\}\| \\ &\leq \|B(u)^{-1}\| \cdot [\|B(u)s + \mathcal{F}(u)\| + \|\mathcal{F}(u)\|] \\ &\leq 4M\|\mathcal{F}(u)\| \\ &< \frac{\delta}{2},\end{aligned}$$

so

$$\|u_+ - u^*\| \leq \|u - u^*\| + \|s\| < \delta.$$

Because

$$\|\mathcal{F}(u_+)\| < \|\mathcal{F}(u)\| < \frac{\delta}{8M},$$

we have

$$\|u_+ - u^*\| \leq 2M\|\mathcal{F}(u_+)\| < \frac{\delta}{4} < \frac{\delta}{2},$$

that is, $u_+ \in N(u^*, \frac{\delta}{2})$. □

The following theorem describes the local convergence property of the ASPIN method.

Theorem 3.7. *Assume that there exists $\gamma > 0$ such that*

$$\|\nabla f(v) - \nabla f(w)\| \leq \gamma\|v - w\| \quad \forall\, v, w \in N(u^*, \delta).$$

If $u^{(0)} \in N(u^*, \frac{\delta}{2})$ such that

$$\|\mathcal{F}(u^{(0)})\| < \frac{\delta}{8M}, \quad \mathcal{F}\big(u^{(0)}\big) \neq 0,$$

then the ASPIN method can generate a sequence $\{u^{(k)}\} \subset N(u^*, \frac{\delta}{2})$ and $u^{(k)} \to u^*$.

Proof. We first prove that the ASPIN method can generate a sequence $\{u^{(k)}\} \subset N(u^*, \frac{\delta}{2})$ by induction.

(i) Since $u^{(0)} \in N(u^*, \frac{\delta}{2}) \subset N(u^*, \delta)$, and

$$\big\|\mathcal{F}\big(u^{(0)}\big) + B\big(u^{(0)}\big)p^{(0)}\big\| \leq \eta_0\big\|\mathcal{F}\big(u^{(0)}\big)\big\|, \quad \eta_0 \in [0, \eta_{max}], \tag{3.9}$$

Theorem 3.5 guarantees that a point $u^{(1)} = u^{(0)} + \lambda_0 p^{(0)} \equiv u^{(0)} + s^{(0)}$ can be generated with $\lambda_0 \in (0, 1]$ and $\|\mathcal{F}(u^{(1)})\| < \|\mathcal{F}(u^{(0)})\|$. Furthermore, it follows from (3.9) that

$$\begin{aligned}\big\|\mathcal{F}\big(u^{(0)}\big) + B\big(u^{(0)}\big)s^{(0)}\big\| &= \big\|\mathcal{F}\big(u^{(0)}\big) + B\big(u^{(0)}\big)\big(\lambda_0 p^{(0)}\big)\big\| \\ &= \big\|\lambda_0\big[B\big(u^{(0)}\big)p^{(0)} + \mathcal{F}\big(u^{(0)}\big)\big] + (1 - \lambda_0)\mathcal{F}\big(u^{(0)}\big)\big\| \\ &\leq [\lambda_0\eta_0 + (1 - \lambda_0)]\big\|\mathcal{F}\big(u^{(0)}\big)\big\| \\ &\leq \big\|\mathcal{F}\big(u^{(0)}\big)\big\|.\end{aligned}$$

Thus, Lemma 3.6 shows that $u^{(1)} \in N(u^*, \frac{\delta}{2})$.

(ii) Assume that the ASPIN method has generated $\{u^{(1)}, u^{(2)}, \ldots, u^{(k)}\} \subset N(u^*, \frac{\delta}{2})$ such that

$$\|\mathcal{F}(u^{(1)})\| > \|\mathcal{F}(u^{(2)})\| > \cdots > \|\mathcal{F}(u^{(k)})\|,$$

and also a direction $p^{(k)} \in R^n$ has been computed such that

$$\|\mathcal{F}\big(u^{(k)}\big) + B\big(u^{(k)}\big)p^{(k)}\| \le \eta_k \|\mathcal{F}\big(u^{(k)}\big)\|, \quad \eta_k \in [0, \eta_{max}]. \tag{3.10}$$

Then in the same way as above, it is easy to prove that a point $u^{(k+1)} \in N(u^*, \frac{\delta}{2})$ can be produced, and

$$\|\mathcal{F}\big(u^{(k)}\big)\| > \|\mathcal{F}\big(u^{(k+1)}\big)\|.$$

Thus, by induction, the ASPIN method can generate a sequence $\{u^{(k)}\} \subset N(u^*, \frac{\delta}{2})$.

Next we show that $f\big(u^{(k)}\big) \to 0$. By (3.10),

$$\begin{aligned} \mathcal{F}\big(u^{(k)}\big)^T B\big(u^{(k)}\big)p^{(k)} &= \mathcal{F}\big(u^{(k)}\big)^T [B\big(u^{(k)}\big)p^{(k)} + \mathcal{F}\big(u^{(k)}\big) - \mathcal{F}\big(u^{(k)}\big)] \\ &\le -2(1-\eta_{max}) f\big(u^{(k)}\big), \end{aligned} \tag{3.11}$$

and it follows that

$$\begin{aligned} f(u^{(k)} + \lambda_k p^{(k)}) &\le f\big(u^{(k)}\big) + \alpha\lambda_k \mathcal{F}\big(u^{(k)}\big)^T B\big(u^{(k)}\big)p^{(k)} \\ &\le [1 - 2\alpha\lambda_k(1-\eta_{max})] f\big(u^{(k)}\big). \end{aligned}$$

Therefore, by Theorem 3.5, we have

$$\begin{aligned} f(u^{(k)} + \lambda_k p^{(k)}) &\le \left[1 - 2\alpha(1-\eta_{max}) \frac{\alpha\theta_{min} |\mathcal{F}\big(u^{(k)}\big)^T B\big(u^{(k)}\big)p^{(k)}|}{\gamma \|p^{(k)}\|^2}\right] f\big(u^{(k)}\big) \\ &\equiv (1 - ct_k) f\big(u^{(k)}\big), \end{aligned} \tag{3.12}$$

where we set

$$c = \frac{2\alpha^2\theta_{min}(1-\eta_{max})}{\gamma}$$

and

$$t_k = \frac{|\mathcal{F}\big(u^{(k)}\big)^T B\big(u^{(k)}\big)p^{(k)}|}{\|p^{(k)}\|^2}.$$

Because $\{f\big(u^{(k)}\big)\}$ is nonnegative and strictly decreased, $\lim_{k\to\infty} f\big(u^{(k)}\big)$ exists. If $\lim_{k\to\infty} f\big(u^{(k)}\big) > 0$, then (3.12) shows that

$$\lim_{k\to\infty} \frac{f\big(u^{(k+1)}\big)}{f\big(u^{(k)}\big)} \le \lim_{k\to\infty} (1 - ct_k) \le 1,$$

so

$$\lim_{k\to\infty} (1 - ct_k) = 1$$

or, equivalently,

$$\lim_{k\to\infty} t_k = 0.$$

Since (3.11) shows that

$$\frac{2(1-\eta_{max})f\big(u^{(k)}\big)}{\|p^{(k)}\|^2} \le \frac{|\mathcal{F}\big(u^{(k)}\big)^T B\big(u^{(k)}\big)p^{(k)}|}{\|p^{(k)}\|^2} = t_k,$$

we have

$$\lim_{k\to\infty} \frac{f\big(u^{(k)}\big)}{\|p^{(k)}\|^2} = 0.$$

Because $\lim_{k\to\infty} f\big(u^{(k)}\big) > 0$, it follows that

$$\|p^{(k)}\| \to \infty \ (k \to \infty). \tag{3.13}$$

But on the other hand, by (I_2) and (3.10), we have

$$\begin{aligned}\|p^{(k)}\| &= \|B\big(u^{(k)}\big)^{-1}[B\big(u^{(k)}\big)p^{(k)} + \mathcal{F}\big(u^{(k)}\big) - \mathcal{F}\big(u^{(k)}\big)]\|\\ &\le \|B\big(u^{(k)}\big)^{-1}\| \cdot [\|B\big(u^{(k)}\big)p^{(k)} + \mathcal{F}\big(u^{(k)}\big)\| + \|\mathcal{F}\big(u^{(k)}\big)\|]\\ &\le 2M(1+\eta_k)\|\mathcal{F}\big(u^{(k)}\big)\|\\ &\le 2M(1+\eta_{max})\|\mathcal{F}(u^{(0)})\|\\ &\le \frac{\delta(1+\eta_{max})}{4},\end{aligned}$$

which contradicts (3.13). Thus, we must have $f(u^{(k)}) \to 0$. Since $\{u^{(k)}\} \subset N(u^*, \delta)$, by ($\mathrm{I}_6$),

$$\begin{aligned}\|u^{(k)} - u^*\| &= 2\|u^{(k)} - u^*\| - \|u^{(k)} - u^*\|\\ &\le 2\|\mathcal{J}(u^*)^{-1}\|\|\mathcal{J}(u^*)(u^{(k)}-u^*)\| - 2M\|\mathcal{F}\big(u^{(k)}\big) - \mathcal{F}(u^*) - \mathcal{J}(u^*)(u^{(k)}-u^*)\|\\ &\le 2M[\|\mathcal{J}(u^*)(u^{(k)} - u^*)\| - \|\mathcal{F}\big(u^{(k)}\big) - \mathcal{F}(u^*) - \mathcal{J}(u^*)(u^{(k)} - u^*)\|]\\ &\le 2M\|\mathcal{F}\big(u^{(k)}\big)\|.\end{aligned}$$

Because $\|\mathcal{F}(u^{(k)})\| = \sqrt{2f(u^{(k)})} \to 0$, we have $u^{(k)} \to u^*$. □

Now, we complete the discussion for local convergence of the ASPIN method.

4. Convergence rate of the ASPIN method. In this section, we discuss the convergence rate of the ASPIN method. We will show that the ASPIN method is quadratically convergent under suitable conditions.

The following theorem shows that under suitable assumptions, $\lambda_k = 1$ is acceptable for all k sufficiently large.

THEOREM 4.1. *Assume that $f(u)$ is twice continuously differentiable in $N(u^*, \delta)$, and there exists $\gamma > 0$ such that for any $v, w \in N(u^*, \delta)$,*

$$\begin{aligned}\|\nabla f(v) - \nabla f(w)\| &\le \gamma\|v - w\|,\\ \|\nabla^2 f(v) - \nabla^2 f(w)\| &\le \gamma\|v - w\|.\end{aligned}$$

Let $\{u^{(k)}\}$ be the sequence generated by the ASPIN method such that $u^{(k)} \to u^*$. If $\eta_k \to 0$, then $u^{(k+1)} = u^{(k)} + p^{(k)}$ for all sufficiently large k.

Proof. Because $u^{(k)} \to u^*$, without loss of generality, we may assume that $\{u^{(k)}\} \subset N(u^*, \frac{\delta}{2})$. Thus, it follows from Proposition 3.1 that

$$\|\mathcal{F}\big(u^{(k)}\big) + \mathcal{J}\big(u^{(k)}\big)p^{(k)}\| \le \frac{1+\eta_{max}}{2}\|\mathcal{F}\big(u^{(k)}\big)\|. \tag{4.1}$$

Therefore, by (I_4) and (4.1),

$$\begin{aligned} \|p^{(k)}\| &= \|\mathcal{J}\big(u^{(k)}\big)^{-1}[\mathcal{J}\big(u^{(k)}\big)p^{(k)} + \mathcal{F}\big(u^{(k)}\big) - \mathcal{F}\big(u^{(k)}\big)]\| \\ &\le \|\mathcal{J}\big(u^{(k)}\big)^{-1}\| \cdot [\|\mathcal{J}\big(u^{(k)}\big)p^{(k)} + \mathcal{F}\big(u^{(k)}\big)\| + \|\mathcal{F}\big(u^{(k)}\big)\|] \\ &\le 2M\left(\frac{1+\eta_{max}}{2} + 1\right)\|\mathcal{F}\big(u^{(k)}\big)\| \\ &= M(3+\eta_{max})\|\mathcal{F}\big(u^{(k)}\big)\|. \end{aligned} \tag{4.2}$$

Because $\|\mathcal{F}(u^{(k)})\| \to 0$, the above inequality shows that

$$\|p^{(k)}\| \to 0 \ (k \to \infty).$$

By (I_4), we have

$$\|\nabla f\big(u^{(k)}\big)\| = \|\mathcal{J}\big(u^{(k)}\big)^T\mathcal{F}\big(u^{(k)}\big)\| \ge \|\mathcal{J}\big(u^{(k)}\big)^{-1}\|^{-1}\|\mathcal{F}\big(u^{(k)}\big)\| \ge \frac{1}{2M}\|\mathcal{F}\big(u^{(k)}\big)\|;$$

thus, Lemma 3.2 in connection with (I_3), (I_4), and (4.1) shows that

$$\begin{aligned} \frac{|\nabla f\big(u^{(k)}\big)^T p^{(k)}|}{\|p^{(k)}\|} &\ge \frac{1-\frac{1+\eta_{max}}{2}}{1+\frac{1+\eta_{max}}{2}} \cdot \frac{1}{\kappa(\mathcal{J}\big(u^{(k)}\big))} \cdot \|\nabla f\big(u^{(k)}\big)\| \\ &\ge \frac{1-\eta_{max}}{3+\eta_{max}} \cdot \frac{1}{4M^2} \cdot \frac{1}{2M}\|\mathcal{F}\big(u^{(k)}\big)\| \\ &= \frac{1-\eta_{max}}{8M^3(3+\eta_{max})}\|\mathcal{F}\big(u^{(k)}\big)\|, \end{aligned}$$

or we have

$$\begin{aligned} \|\mathcal{F}\big(u^{(k)}\big)\|\|p^{(k)}\| &\le \frac{8M^3(3+\eta_{max})}{1-\eta_{max}} \cdot |\nabla f\big(u^{(k)}\big)^T p^{(k)}| \\ &\equiv a|\nabla f\big(u^{(k)}\big)^T p^{(k)}|, \end{aligned} \tag{4.3}$$

where

$$a = \frac{8M^3(3+\eta_{max})}{1-\eta_{max}}.$$

From (4.2) and (4.3), we obtain

$$\begin{aligned} \|p^{(k)}\|^2 &\le M(3+\eta_{max})\|\mathcal{F}\big(u^{(k)}\big)\|\|p^{(k)}\| \\ &\le \frac{8M^4(3+\eta_{max})^2}{1-\eta_{max}} \cdot |\nabla f\big(u^{(k)}\big)^T p^{(k)}| \\ &\equiv b|\nabla f\big(u^{(k)}\big)^T p^{(k)}|, \end{aligned}$$

where

$$b = \frac{8M^4(3+\eta_{max})^2}{1-\eta_{max}}.$$

Next we show that $\lambda_k = 1$ is acceptable for all k sufficiently large. First note that if $\mathcal{F} = (\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n)^T$, then

$$\begin{aligned}\nabla^2 f(u) &= \mathcal{J}(u)^T \mathcal{J}(u) + \sum_{i=1}^{n} \mathcal{F}_i(u) \nabla^2 \mathcal{F}_i(u) \\ &\equiv \mathcal{J}(u)^T \mathcal{J}(u) + S(u).\end{aligned}$$

Since $u^{(k)} \to u^*$ and $\mathcal{F}(u^*) = 0$, it follows that $\|S(u^{(k)})\| \to 0$.

For each k, by the mean value theorem, there exists a $\xi^{(k)}$ on the line segment between $u^{(k)}$ and $u^{(k)} + p^{(k)}$ such that

$$f\big(u^{(k)} + p^{(k)}\big) - f\big(u^{(k)}\big) - \frac{1}{2}\nabla f\big(u^{(k)}\big)^T p^{(k)} = \frac{1}{2}\big(\nabla f\big(u^{(k)}\big) + \nabla^2 f\big(\xi^{(k)}\big)p^{(k)}\big)^T p^{(k)}.$$

This gives

$$\begin{aligned}&\left| f\big(u^{(k)} + p^{(k)}\big) - f\big(u^{(k)}\big) - \frac{1}{2}\nabla f\big(u^{(k)}\big)^T p^{(k)} \right| \\ &= \left| \frac{1}{2}\big(\nabla f\big(u^{(k)}\big) + \nabla^2 f(\xi^{(k)})p^{(k)}\big)^T p^{(k)} \right| \\ &= \frac{1}{2}\left| \big(\nabla f\big(u^{(k)}\big) + \nabla^2 f\big(u^{(k)}\big)p^{(k)}\big)^T p^{(k)} + (p^{(k)})^T\big(\nabla^2 f(\xi^{(k)}) - \nabla^2 f\big(u^{(k)}\big)\big)p^{(k)} \right| \\ &\le \frac{1}{2}\big[\big\|\mathcal{J}\big(u^{(k)}\big)^T\big(\mathcal{F}\big(u^{(k)}\big) + \mathcal{J}\big(u^{(k)}\big)p^{(k)}\big)\big\| \cdot \|p^{(k)}\| + \big(\big\|S\big(u^{(k)}\big)\big\| + \gamma\|p^{(k)}\|\big)\|p^{(k)}\|^2\big] \\ &\le \frac{1}{2}\big[\eta_k\big\|\mathcal{J}\big(u^{(k)}\big)\big\| \cdot \big\|\mathcal{F}\big(u^{(k)}\big)\big\| \cdot \|p^{(k)}\| + \big(\big\|S\big(u^{(k)}\big)\big\| + \gamma\|p^{(k)}\|\big)\|p^{(k)}\|^2\big] \\ &\le -\frac{1}{2}\big[2aM\eta_k + b\big(\big\|S\big(u^{(k)}\big)\big\| + \gamma\|p^{(k)}\|\big)\big]\nabla f\big(u^{(k)}\big)^T p^{(k)} \\ &\equiv -\frac{1}{2}\epsilon_k \nabla f\big(u^{(k)}\big)^T p^{(k)};\end{aligned}$$

therefore,

$$f(u^{(k)} + p^{(k)}) - f\big(u^{(k)}\big) \le \frac{1}{2}(1 - \epsilon_k)\nabla f\big(u^{(k)}\big)^T p^{(k)}. \tag{4.4}$$

Since η_k, $\|S(u^{(k)})\|$, and $\|p^{(k)}\|$ all converge to zero, it follows that $\epsilon_k \to 0$. Thus, for all k sufficiently large, we have

$$\epsilon_k < \frac{1}{2},$$

and consequently, for all k sufficiently large, it follows from (4.4) and Lemma 3.3 that

$$f(u^{(k)} + p^{(k)}) - f\big(u^{(k)}\big) \le \frac{1}{4}\nabla f\big(u^{(k)}\big)^T p^{(k)} \le \alpha \mathcal{F}\big(u^{(k)}\big)^T B\big(u^{(k)}\big)p^{(k)}.$$

Thus, λ_k is acceptable for all k sufficiently large. In other word, $u^{(k+1)} = u^{(k)} + p^{(k)}$ for all k sufficiently large. □

Theorem 4.1 shows that if the ASPIN iterative sequence converges to the solution u^* of system (1.1), then step 3 in Algorithm 2.1 will not be implemented for all k sufficiently large.

The following lemma shows that both $\mathcal{J}(u)$ and $B(u)$ are Lipschitz continuous near u^*.

LEMMA 4.2. *Assume that $F(u)$ is twice continuously differentiable. Then there exists a neighborhood $V \subset N(u^*, \frac{\delta}{2})$ of u^* such that both $\mathcal{J}(u)$ and $B(u)$ are Lipschitz continuous in V.*

Proof. Since $F''(w)$, $T_i'(u)$, $[E_i J(u) E_i^T]^{-1}$, and $E_i J(u)$ are continuous in $N(u^*, \frac{\delta}{2}) \subset U$, we define the constants

$$L_1 := \sup_{u \in N(u^*, \frac{\delta}{2})} \|F''(u)\|, \quad L_2 := \max_i \sup_{u \in N(u^*, \frac{\delta}{2})} \|T_i'(u)\|$$

and

$$L_3 := \max_i \sup_{u \in N(u^*, \frac{\delta}{2})} \|[E_i J(u) E_i^T]^{-1}\|, \quad L_4 := \max_i \sup_{u \in N(u^*, \frac{\delta}{2})} \|E_i J(u)\|.$$

Thus, for any $u, v \in N(u^*, \frac{\delta}{2})$, Lemma 3.3.5 in [20] shows that

$$\|J(u) - J(v)\| \leq \max_{t \in [0,1]} \|F''(u + t(v-u))\| \|u - v\| \leq L_1 \|u - v\|, \tag{4.5}$$

and Lemma 3.2.3 in [20] shows that

$$\|T_i(u) - T_i(v)\| \leq \max_{t \in [0,1]} \|T_i'(u + t(v-u))\| \|u - v\| \leq L_2 \|u - v\|. \tag{4.6}$$

Since, for each i, $u - T_i(u) \to u^*$ when $u \to u^*$, there exists a neighborhood $V \subset N(u^*, \frac{\delta}{2})$ of u^*, which is independent of i, such that $u - T_i(u) \in N(u^*, \frac{\delta}{2})$ whenever $u \in V$. Therefore, for any $u, v \in V$, it follows from (4.5) and (4.6) that

$$\begin{aligned} \|J(u - T_i(u)) - J(v - T_i(v))\| &\leq L_1 \|(u - v) - [T_i(u) - T_i(v)]\| \\ &\leq L_1 (1 + L_2) \|u - v\|. \end{aligned} \tag{4.7}$$

At the same time, for any $u \in V$ and for each i, by the definition of L_3, we have

$$\|[E_i J(u) E_i^T]^{-1}\| \leq L_3 \tag{4.8}$$

and

$$\|[E_i J(u - T_i(u)) E_i^T]^{-1}\| \leq L_3. \tag{4.9}$$

Now set

$$G(u) = E_i J(u - T_i(u)) E_i^T, \quad H(u) = E_i J(u - T_i(u)).$$

Then for any $u, v \in V$, by (4.7), (4.9), and the definition of L_4, we have

$$\begin{aligned} \|T_i'(u) - T_i'(v)\| &= \|E_i^T G(u)^{-1} H(u) - E_i^T G(v)^{-1} H(v)\| \\ &= \|E_i^T G(u)^{-1} H(u) - E_i^T G(u)^{-1} H(v) \\ &\quad + E_i^T G(u)^{-1} H(v) - E_i^T G(v)^{-1} H(v)\| \\ &\leq \|E_i^T G(u)^{-1} H(u) - E_i^T G(u)^{-1} H(v)\| \\ &\quad + \|E_i^T G(u)^{-1} H(v) - E_i^T G(v)^{-1} H(v)\| \\ &\leq \|G(u)^{-1}\| \|H(u) - H(v)\| + \|G(u)^{-1}\| \|G(u) - G(u)\| \|G(v)^{-1}\| \|H(v)\| \\ &\leq L_3 L_1 (1 + L_2) \|u - v\| + L_3^2 L_1 (1 + L_2) L_4 \|u - v\| \\ &= L_1 L_3 (1 + L_1)(1 + L_3 L_4) \|u - v\|. \end{aligned}$$

Thus,

$$\begin{aligned}\|\mathcal{J}(u)-\mathcal{J}(v)\| &= \left\| \sum_{i=1}^{N} T_i'(u) - \sum_{i=1}^{N} T_i'(v) \right\| \leq \sum_{i=1}^{N} \|T_i'(u)-T_i'(v)\| \\ &\leq NL_1L_3(1+L_1)(1+L_3L_4)\|u-v\| \\ &\equiv L_{\mathcal{J}}\|u-v\|\end{aligned}$$

for $u, v \in V$. That is, $\mathcal{J}(u)$ is Lipschitz continuous in V.

In a similar way, by using (4.5), (4.8), and the definition of L_4, one can prove that

$$\|R_i(u)-R_i(v)\| \leq L_1L_3(1+L_3L_4)\|u-v\|$$

for any $u, v \in V$, and each i. Thus,

$$\begin{aligned}\|B(u)-B(v)\| &= \left\| \sum_{i=1}^{N} R_i(u) - \sum_{i=1}^{N} R_i'(v) \right\| \leq \sum_{i=1}^{N} \|R_i(u)-R_i(v)\| \\ &\leq NL_1L_3(1+L_3L_4)\|u-v\| \\ &\equiv L_B\|u-v\|\end{aligned}$$

for all $u, v \in V$. Therefore, $B(u)$ is Lipschitz continuous in V. □

LEMMA 4.3. *For any $u \in N(u^*, \delta)$, it holds that*

$$\frac{1}{2M}\|u-u^*\| \leq \|\mathcal{F}(u)\| \leq \left(M + \frac{1}{2M}\right)\|u-u^*\|.$$

Proof. The first part of the inequality has been proved in the proof of Lemma 3.6, and the second part is easy. □

Based on the above preparations, we now have the following result.

THEOREM 4.4. *Assume that both $F(u)$ and $f(u)$ are twice continuously differentiable in $N(u^*, \delta)$, and there exists $\gamma > 0$ such that*

$$\begin{aligned}\|\nabla f(v) - \nabla f(w)\| &\leq \gamma\|v-w\|, \\ \|\nabla^2 f(v) - \nabla^2 f(w)\| &\leq \gamma\|v-w\|\end{aligned}$$

for any $v, w \in N(u^*, \delta)$. If $\{u^{(k)}\}$ is a sequence generated by the ASPIN method such that $u^{(k)} \to u^*$ and $u^{(k+1)} = u^{(k)} + p^{(k)}$ for all sufficiently large k, then

(i) *$u^{(k)} \to u^*$ superlinearly if and only if*

$$\|r_k\| = o(\|\mathcal{F}(u^{(k)})\|),$$

where

$$r_k = \mathcal{F}(u^{(k)}) + B(u^{(k)})p^{(k)};$$

(ii) *$u^{(k)} \to u^*$ quadratically if and only if*

$$\|r_k\| = \mathcal{O}(\|\mathcal{F}(u^{(k)})\|^2).$$

Proof. Since $u^{(k+1)} = u^{(k)} + p^{(k)}$ for all sufficiently large k, without loss of generality, we assume that $u^{(k+1)} = u^{(k)} + p^{(k)}$ for all k in the following argument.

Assume that $u^{(k)} \to u^*$ superlinearly. Since

$$\begin{aligned} r_k &= \mathcal{F}\big(u^{(k)}\big) + B\big(u^{(k)}\big)p^{(k)} \\ &= \big[\mathcal{F}\big(u^{(k)}\big) - \mathcal{F}(u^*) - \mathcal{J}(u^*)\big(u^{(k)} - u^*\big)\big] \\ &\quad - \big[B\big(u^{(k)}\big) - \mathcal{J}(u^*)\big]\big(u^{(k)} - u^*\big) + B\big(u^{(k)}\big)\big(u^{(k+1)} - u^*\big), \end{aligned} \tag{4.10}$$

by Lemma 3.2.10 in [20], continuity of $B(u)$ at u^* and the assumption that $u^{(k)} \to u^*$ superlinearly, we have

$$\|r_k\| \leq o(\|u^{(k)} - u^*\|) + o(1)\|u^{(k)} - u^*\| + o(\|u^{(k)} - u^*\|).$$

Therefore, by Lemma 4.3, it follows that

$$\|r_k\| = o(\|u^{(k)} - u^*\|) = o(\|\mathcal{F}\big(u^{(k)}\big)\|).$$

Conversely, assume that $\|r_k\| = o(\|\mathcal{F}\big(u^{(k)}\big)\|)$. Since

$$\begin{aligned} u^{(k+1)} - u^* &= (u^{(k)} - u^*) + p^{(k)} \\ &= (u^{(k)} - u^*) + B\big(u^{(k)}\big)^{-1}[r_k - \mathcal{F}\big(u^{(k)}\big)] \\ &= B\big(u^{(k)}\big)^{-1}\big\{\big[B\big(u^{(k)}\big) - \mathcal{J}(u^*)\big](u^{(k)} - u^*) \\ &\qquad + r_k - \big[\mathcal{F}\big(u^{(k)}\big) - \mathcal{F}(u^*) - \mathcal{J}(u^*)(u^{(k)} - u^*)\big]\big\}, \end{aligned} \tag{4.11}$$

thus, by (I$_2$), the continuity of $B(u)$ at u^*, the assumption that $\|r_k\| = o(\|\mathcal{F}\big(u^{(k)}\big)\|)$, and Lemma 3.2.10 in [20], we have

$$\|u^{(k+1)} - u^*\| \leq 2M[o(1)\|u^{(k)} - u^*\| + o(\|\mathcal{F}\big(u^{(k)}\big)\|) + o(\|u^{(k)} - u^*\|)].$$

Therefore, by Lemma 4.3,

$$\|u^{(k+1)} - u^*\| = o(\|u^{(k)} - u^*\|) + o(\|\mathcal{F}\big(u^{(k)}\big)\|) = o(\|u^{(k)} - u^*\|).$$

Since F is twice continuously differentiable in $N(u^*, \delta)$, Lemma 4.2 shows that both $\mathcal{J}(u)$ and $B(u)$ are Lipschitz continuous in a neighborhood $V \subset N(u^*, \frac{\delta}{2})$ of u^*. Thus, there exists $L > 0$ such that for any $u \in V$,

$$\|\mathcal{J}(u) - \mathcal{J}(u^*)\| \leq L\|u - u^*\| \tag{4.12}$$

and

$$\|B(u) - B(u^*)\| \leq L\|u - u^*\|. \tag{4.13}$$

By (4.12) and Lemma 3.2.12 in [20],

$$\|\mathcal{F}(u) - \mathcal{F}(u^*) - \mathcal{J}(u^*)(u - u^*)\| \leq \frac{L}{2}\|u - u^*\|^2 \quad \forall\, u \in V. \tag{4.14}$$

In a similar way, it follows from (4.10), (4.11), (4.13), (4.14), and Lemma 4.3 that $u^{(k)} \to u^*$ quadratically if and only if

$$\|r_k\| = \mathcal{O}(\|\mathcal{F}\big(u^{(k)}\big)\|^2).$$

This concludes the proof. □

From Theorems 4.1 and 4.4, we can obtain the following result.

COROLLARY 4.5. *Assume that both $F(u)$ and $f(u)$ are twice continuously differentiable in $N(u^*, \delta)$, and there exists $\gamma > 0$ such that for any $v, w \in N(u^*, \delta)$,*

$$\|\nabla f(v) - \nabla f(w)\| \leq \gamma \|v - w\|,$$
$$\|\nabla^2 f(v) - \nabla^2 f(w)\| \leq \gamma \|v - w\|.$$

Let $\{u^{(k)}\}$ be a sequence generated by the ASPIN method such that $u^{(k)} \to u^*$. Then

(i) *$u^{(k)} \to u^*$ superlinearly if $\eta_k \to 0$;*

(ii) *$u^{(k)} \to u^*$ quadratically if $\eta_k = \mathcal{O}(\|\mathcal{F}(u^{(k)})\|)$.*

Corollary 4.5 reflects how the forcing term influences the convergence rate of the ASPIN method. This result is similar to Corollary 3.5 in [8] for the inexact Newton method. In particular, by Corollary 4.5, we can determine the convergence rate of the ASPIN method by choosing proper forcing terms.

5. Conclusion.

The inexact Newton method is one of the effective tools for solving large sparse systems of nonlinear equations. By using nonlinear additive Schwarz preconditioning technique, Cai and Keyes [6] introduced the ASPIN method. This method is very effective for solving some nonlinear problems with strong nonbalanced nonlinearities. However, the convergence of the ASPIN method is not discussed by them or others.

In this paper, we discussed the convergence property of the ASPIN method and thus we provided a theoretical support for the ASPIN method. The convergence result is local since the design of the ASPIN only concerns the local properties of the original function. Furthermore, we discussed the convergence rate for the ASPIN method, and the result shows that the convergence rate of the ASPIN method is similar to that of the inexact Newton method. Thus, we can obtain the desired convergence rate by choosing proper forcing terms.

Appendix.

We give a simple example to show that our main assumptions, inequalities (I_1)–(I_6) in section 3, are not so strict. Consider the nonlinear equations

$$\begin{cases} 2u_1 - u_2 + \lambda e^{u_1} - \lambda = 0, \\ -u_1 + 2u_2 + \lambda e^{u_2} - \lambda = 0, \end{cases}$$

where $\lambda > 0$. It is obvious that $u^* = (0,0)^T$ is a solution of the system. Let

$$S = \{1, 2\}, \quad S_1 = \{1\}, \quad S_2 = \{2\}$$

and

$$T_1(u) = \begin{pmatrix} T_{11}(u) \\ 0 \end{pmatrix}, \quad T_2(u) = \begin{pmatrix} 0 \\ T_{22}(u) \end{pmatrix},$$

where T_{11}, $T_{22} : R^2 \to R$. By $F_{S_i}(u - T_i(u)) = 0$, we have

$$2(u_i - T_{ii}(u)) - u_{3-i} + \lambda e^{u_i - T_{ii}(u)} - \lambda = 0, \quad i = 1, 2. \tag{A.1}$$

It is easy to see that $T_{ii}(u)$ are continuous functions. Furthermore, we can obtain

$$T_1'(u) = \begin{pmatrix} 1 & -\left(2 + \lambda e^{u_1 - T_{11}(u)}\right)^{-1} \\ 0 & 0 \end{pmatrix}, \quad T_2'(u) = \begin{pmatrix} 0 & 0 \\ -\left(2 + \lambda e^{u_2 - T_{22}(u)}\right)^{-1} & 1 \end{pmatrix}.$$

Thus,

$$\mathcal{J}(u) = \begin{pmatrix} 1 & -\left(2+\lambda e^{u_1 - T_{11}(u)}\right)^{-1} \\ -\left(2+\lambda e^{u_2 - T_{22}(u)}\right)^{-1} & 1 \end{pmatrix}.$$

Besides, it is easy to see that

$$B(u) = \begin{pmatrix} 1 & -\left(2+\lambda e^{u_1}\right)^{-1} \\ -\left(2+\lambda e^{u_2}\right)^{-1} & 1 \end{pmatrix}.$$

Since

$$B(u^*) = \mathcal{J}(u^*) = \begin{pmatrix} 1 & -(2+\lambda)^{-1} \\ -(2+\lambda)^{-1} & 1 \end{pmatrix},$$

so

$$M = \max\{\|\mathcal{J}(u^*)\|, \|\mathcal{J}(u^*)^{-1}\|\} = \frac{2+\lambda}{1+\lambda}.$$

Now we will choose some proper δ such that inequalities (I_1)–(I_6) hold. Let $\delta \in (0, \frac{1}{2})$ and assume that $u = (u_1, u_2)^T \in N(u^*, \delta)$. Note that if

$$A = \begin{pmatrix} 1 & a \\ b & 1 \end{pmatrix},$$

then $\|A\| \le 1 + \max\{|a|, |b|\}$. Therefore,

$$\|B(u)\| \le 1 + \max_{1\le i\le 2} (2+\lambda e^{u_i})^{-1} < 1 + (2+\lambda e^{-\delta})^{-1} < 2M.$$

In addition, since

$$B(u)^{-1} = \frac{1}{1-(2+\lambda e^{u_1})^{-1}(2+\lambda e^{u_2})^{-1}} \begin{pmatrix} 1 & (2+\lambda e^{u_1})^{-1} \\ (2+\lambda e^{u_2})^{-1} & 1 \end{pmatrix},$$

one can easily verify that

$$\|B(u)^{-1}\| \le \frac{2+\lambda e^{-\delta}}{1+\lambda e^{-\delta}} < 2M.$$

Because $u \in N(u^*, \delta)$, by (A.1), we have

$$|u_i - T_{ii}(u)| < \frac{1}{2}|u_{3-i}| < \frac{1}{2}\delta, \quad i = 1, 2. \tag{A.2}$$

Thus, one may obtain

$$\|\mathcal{J}(u)\| \le 2M, \quad \|\mathcal{J}(u)^{-1}\| \le 2M.$$

By using (A.2) and the inequality $e^x < 1 + 2x$ $(0 < x < 1)$, we have

$$\begin{aligned} \|\mathcal{J}(u) - B(u)\| &\le \max_{1\le i\le 2} |(2+\lambda e^{u_i})^{-1} - (2+\lambda e^{u_i - T_{ii}(u)})^{-1}| \\ &\le \max_{1\le i\le 2} \frac{\lambda e^{\delta}(e^{|T_{ii}(u)|}-1)}{(2+\lambda e^{-\delta})^2} \\ &\le \frac{3\delta\lambda e^{\delta}}{4+4\lambda e^{-\delta}} \\ &\le \frac{3e\lambda\delta}{4+4\lambda}. \end{aligned}$$

This shows that inequality (I_5) holds when

$$\delta \leq \frac{(1+\lambda)(1-\eta_{max})}{3e\lambda M(1+\eta_{max})} = \frac{(1+\lambda)^2(1-\eta_{max})}{3e\lambda(2+\lambda)(1+\eta_{max})}.$$

By using (A.1), (A.2), and the fact that

$$e^x > 1+x, \quad x \in (0,1), \quad \text{and} \quad e^x < 1 + \frac{1}{2}x, \quad x \in (-1,0),$$

we have

$$|u_i - T_{ii}(u)| \leq \frac{2}{4+\lambda}|u_{3-i}|, \quad i = 1,2.$$

Therefore,

$$\begin{aligned}\|\mathcal{F}(u) - \mathcal{F}(u^*) - \mathcal{J}(u^*)(u-u^*)\| &= \left\| \begin{pmatrix} T_{11}(u) - u_1 + (2+\lambda)^{-1}u_2 \\ T_{22}(u) - u_2 + (2+\lambda)^{-1}u_1 \end{pmatrix} \right\| \\ &\leq \left\| \begin{pmatrix} |T_{11}(u) - u_1| + (2+\lambda)^{-1}|u_2| \\ |T_{22}(u) - u_2| + (2+\lambda)^{-1}|u_1| \end{pmatrix} \right\| \\ &\leq \frac{3}{2+\lambda}\|u\| < 2M\|u-u^*\|.\end{aligned}$$

Thus, inequality (I_6) holds.

Summing up the above discussion, we know that if

$$\delta \leq \min\left\{ \frac{1}{2}, \frac{(1+\lambda)^2(1-\eta_{max})}{3e\lambda(2+\lambda)(1+\eta_{max})} \right\},$$

then for any $u \in N(u^*, \delta)$, inequalities (I_1)–(I_6) hold.

For this example, other assumptions in the paper can also be checked to be true.

Acknowledgments. The author thanks the anonymous referees and Carol S. Woodward for many valuable suggestions.

REFERENCES

[1] S. Bellavia and B. Morini, *A globally convergent Newton-GMRES subspace method for systems of nonlinear equations*, SIAM J. Sci. Comput., 23 (2001), pp. 940–960.

[2] S. Bellavia, M. Macconi, and B. Morini, *A hybrid Newton-GMRES method for solving nonlinear equations*, in Numerical Analysis and Its Applications, Lecture Notes in Comput. Sci. 1988, L. Vulkov, J. Wasniewski, and P. Yalamov, eds., Springer-Verlag, Berlin, 2000, pp. 68–75.

[3] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 450–481.

[4] P. N. Brown and Y. Saad, *Convergence theory of nonlinear Newton-Krylov algorithms*, SIAM J. Optim., 4 (1994), pp. 297–330.

[5] X.-C. Cai, M. Dryja, and M. Sarkis, *Restricted additive Schwarz preconditioners with harmonic overlap for symmetric positive definite linear systems*, SIAM J. Numer. Anal., 41 (2003), pp. 1209–1231.

[6] X.-C. Cai and D. E. Keyes, *Nonlinearly preconditioned inexact Newton algorithms*, SIAM J. Sci. Comput., 24 (2002), pp. 183–200.

[7] X.-C. Cai, D. E. Keyes, and L. Marcinkowski, *Nonlinear additive Schwarz preconditioners and applications in computational fluid dynamics*, Internat. J. Numer. Methods Fluid Mech., 40 (2002), pp. 1463–1470.

[8] R. S. Dembo, S. C. Eisenstat, and T. Steihaug, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[9] J. E. Dennis and R. B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1983.

[10] M. Dryja and W. Hackbusch, *On the nonlinear domain decomposition method*, BIT, 37 (1997), pp. 296–311.

[11] M. Dryja and O. B. Widlund, *Domain decomposition algorithms with small overlap*, SIAM J. Sci. Comput., 15 (1994), pp. 604–620.

[12] S. C. Eisenstat and H. F. Walker, *Globally convergent inexact Newton methods*, SIAM J. Optim., 4 (1994), pp. 393–422.

[13] S. C. Eisenstat and H. F. Walker, *Choosing the forcing term in an inexact Newton method*, SIAM J. Sci. Comput., 17 (1996), pp. 16–32.

[14] D. R. Fokkema, G. L. G. Sleijpen, and H. A. van der Vorst, *Accelerated inexact Newton schemes for large systems of nonlinear equations*, SIAM J. Sci. Comput., 19 (1998), pp. 657–674.

[15] I. E. Kaporin and O. Axelsson, *On a class of nonlinear equation solvers based on the residual norm reduction over a sequence of affine subspaces*, SIAM J. Sci. Comput., 16 (1995), pp. 228–249.

[16] D. A. Knoll and D. E. Keyes, *Jacobian-free Newton-Krylov methods: A survey of approaches and applications*, J. Comput. Phys., 193 (2004), pp. 357–397.

[17] S. H. Lui, *Nonlinearly preconditioned Newton's method*, in Domain Decomposition Methods in Science and Engineering, Fourteenth International Conference on Domain Decomposition Methods, Cocoyoc, Mexico, I. Herrera, D. E. Keyes, O. B. Widlund, and R. Yates, eds., National Autonomous University of Mexico (UNAM), Mexico City, Mexico, 2003, pp. 95–105.

[18] L. Luksan, *Inexact trust region method for large sparse systems of nonlinear equations*, J. Optim. Theory Appl., 81 (1994), pp. 569–591.

[19] M. Pernice and H. F. Walker, *NITSOL: A Newton iterative solver for nonlinear systems*, SIAM J. Sci. Comput., 19 (1998), pp. 302–318.

[20] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Classics Appl. Math. 30, SIAM, Philadelphia, PA, 2000.

[21] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Comput., 7 (1986), pp. 856–869.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 1872–1896

CONVERGENCE OF THE MIMETIC FINITE DIFFERENCE METHOD FOR DIFFUSION PROBLEMS ON POLYHEDRAL MESHES[*]

FRANCO BREZZI[†], KONSTANTIN LIPNIKOV[‡], AND MIKHAIL SHASHKOV[‡]

Abstract. The stability and convergence properties of the mimetic finite difference method for diffusion-type problems on polyhedral meshes are analyzed. The optimal convergence rates for the scalar and vector variables in the mixed formulation of the problem are proved.

Key words. compatible discretizations, mimetic finite difference method, convergence, polyhedral meshes

AMS subject classifications. 65N06, 65N12, 65N15, 65N22, 65N30

DOI. 10.1137/040613950

1. Introduction.

The main goal of this paper is to establish convergence of *mimetic discretizations* of the first-order system that describes linear stationary diffusion on unstructured polyhedral meshes. The main idea of the mimetic finite difference (MFD) method is to mimic the underlying properties of the original continuum differential operators, e.g., conservation laws, solution symmetries, and the fundamental identities and theorems of vector and tensor calculus. For the linear diffusion problem, this means that the mimetic discretizations mimic the Gauss divergence theorem needed for the local mass conservation, the symmetry between the continuous gradient and divergence operators needed for proving symmetry and positivity of the resulting discrete operator, and the null spaces of the involved operators needed for stability of the discretizations.

The MFD method has been successfully employed for solving problems of continuum mechanics [19], electromagnetics [14], gas dynamics [8], and linear diffusion on simplicial and quadrilateral meshes in both the Cartesian and polar coordinates [15, 13, 20, 17]. Recent advances in extending the mimetic discretizations to general polygonal meshes [16] have inspired us to develop the rigorous convergence theory for unstructured polygonal and polyhedral meshes.

The polyhedral elements appear naturally in reservoir models simulating thinning or tapering out ("pinching out") of geological layers. The pinchouts are modeled with mixed types of mesh elements, pentahedrons, prisms, and tetrahedrons which

[*] Received by the editors August 26, 2004; accepted for publication (in revised form) January 31, 2005; published electronically November 22, 2005. Research was partially performed at the Los Alamos National Laboratory operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36, LAUR-04-5756. This work was performed by an employee of the U.S. Government or under U.S. Government contract. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sinum/43-5/61395.html

[†] Università di Pavia, Department of Mathematics, 27100 Pavia, Italy (brezzi@imati.cnr.it). The research of this author was supported by Italian MIVR projects FIRB 2001 and PRIN 2003.

[‡] Los Alamos National Laboratory, Theoretical Division, MS B284, Los Alamos, NM 87545 (lipnikov@lanl.gov, shashkov@lanl.gov). The research of these authors was partially supported by the DOE/ASCR Program in the Applied Mathematical Sciences.

are frequently obtained by collapsing some of the elements in a structured hexahedral or prismatic mesh.

Other sources of polyhedral meshes are the adaptive mesh refinement methods. A necessity to have a conformal mesh results in an abundant mesh refinement, e.g., in the methods using the red-green refinement strategy. However, the locally refined mesh may be considered as the conformal polyhedral mesh with degenerate elements (for instance, when the angle between two faces is 180°). If we know how to discretize a problem on a general polyhedral mesh, the superfluous mesh refinements can be avoided. A similar argument can be applied to nonmatching meshes which frequently may be treated as conformal polyhedral meshes with degenerate elements. This is the way followed, for instance, in [16] for two-dimensional (2D) meshes.

Allowing arbitrary shape for a mesh element provides greater flexibility in the mesh generation process, especially in the regions where the geometry is extremely complex. Even in the case of an unstructured hexahedral mesh, it may be beneficial to split the curvilinear faces into triangles in order to use more accurate discretization methods and to get a smaller number of unknowns relative to a tetrahedral partition. It is obvious that by splitting each face of a hexahedron into four triangles we get a 24-face polyhedron which is frequently nonconvex.

Some of the simulations in the fluid dynamics indicate that the polyhedral meshes may lead to superior convergence rates and accuracy relative to tetrahedral meshes. We refer readers to the CD-adapco group website (www.cd-adapco.com/news/18/newsdev.htm) for more detail. The polyhedral meshes are also used in a number of radiation–hydrodynamics applications [21, 22, 7]. For instance, one of the approaches to increase robustness of arbitrary Lagrangian–Eulerian simulations is to change the mesh connectivity which leads obviously to general polyhedral meshes.

The diffusion-type (elliptic) problems appear in many applications, for instance, the temperature equation in heat diffusion or the pressure equation in flow problems. The necessity to solve such problems arises in numerical methods for radiation transport coupled with hydrodynamics, mesh smoothing algorithms, etc. In this paper, we consider a diffusion problem formulated as a system of two first-order equations, which is suitable for deriving locally conservative discretizations.

The mimetic discretizations have demonstrated excellent robustness and accuracy in simulations; however, a rigorous convergence proof has always been lacking. The original approach to prove the convergence of these discretizations has been based on establishing the relationship between the MFD and mixed finite element methods [2, 3] which is certainly not enough for many interesting applications. In this paper, we developed a novel technique for proving convergence estimates which may be applied to the case of meshes consisting of arbitrary types of elements, e.g., tetrahedrons, pyramids, hexahedrons, degenerate polyhedrons, etc. The restrictions on a polyhedron shape imposed in section 2 still allow extremely complex elements which cover the majority of meshes used in applications. Note that the developed methodology can be applied to 2D diffusion problems on unstructured polygonal meshes with minor modifications.

The paper is organized as follows. In section 2, we describe the problem under consideration and the class of polyhedral meshes used in the convergence analysis. In section 3, we formulate the MFD method. In section 4, we prove the stability result. In section 5, we prove the convergence of mimetic discretizations. One of the key elements used in our technique, the lift property, is discussed in detail in the appendix.

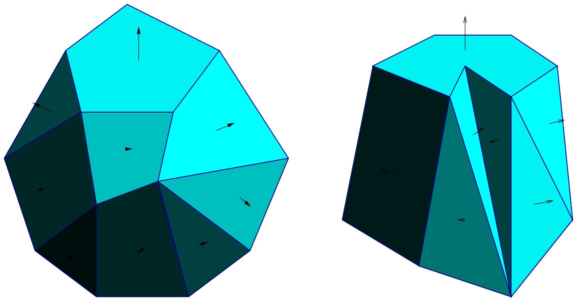


FIG. 2.1. *Two possible elements and the normal to their faces.*

2. The assumptions on the problem and on the mesh. Let us consider a model elliptic boundary value problem

$$\operatorname{div} \mathbf{F} = b, \tag{2.1}$$

$$\mathbf{F} = -\mathbf{K} \operatorname{grad} p. \tag{2.2}$$

Here p denotes a scalar function that we refer to as the pressure, $\mathbf{F}$ denotes a vector function that we refer to as the velocity, $\mathbf{K}$ denotes a full symmetric tensor, and b denotes a source function. The problem is posed in a bounded polyhedral domain $\Omega \subset \mathbb{R}^3$, and is subject to appropriate boundary conditions on $\partial\Omega$. For simplicity, we assume that the homogeneous Dirichlet boundary conditions are imposed on $\partial\Omega$. We also assume that $\mathbf{K}$ satisfies the following regularity and ellipticity property.

P1 (*regularity and ellipticity of* $\mathbf{K}$). Every component of $\mathbf{K}$ is in $W^1_\infty(\Omega)$ and $\mathbf{K}$ is strongly elliptic, meaning that there exist two positive constants κ_* and κ^* such that

$$\kappa_* \|\mathbf{v}\|^2 \leq \mathbf{v}^T \mathbf{K}(\mathbf{x})\, \mathbf{v} \leq \kappa^* \|\mathbf{v}\|^2 \quad \forall \mathbf{v} \in \mathbb{R}^3 \quad \forall \mathbf{x} \in \Omega. \tag{2.3}$$

Let $\mathcal{T}_h$ be a nonoverlapping conformal partition of Ω into polyhedral elements E. For every element E, we denote by $|E|$ its volume and by h_E its diameter. Similarly, for each face e we denote by $|e|$ its area and for every edge ℓ we denote by $|\ell|$ its length. Depending on context, we shall use ∂E either for the boundary of E or the union of element faces. We also set as usual

$$h = \sup_E h_E.$$

The elements E are assumed to be closed simply connected polyhedrons, rather general in shape (see, for instance, Figure 2.1). However, we need some basic assumptions of shape regularity. As we shall see, the assumptions are formally complicated sometimes, but they will hold for practically all partitions which are not totally unreasonable.

M1 (*assumptions on the domain* Ω). We assume that Ω is a polyhedron with a Lipschitz continuous boundary.

M2 (*number of faces and edges*). We assume that we have two positive integers N_e and N_ℓ such that every element E has at most N_e faces, and each face e has at most N_ℓ edges.

M3 (*volumes, areas, and lengths*). We assume that there exist three positive constants v_*, a_*, and l_* (for volume, area, and length, respectively) such that for

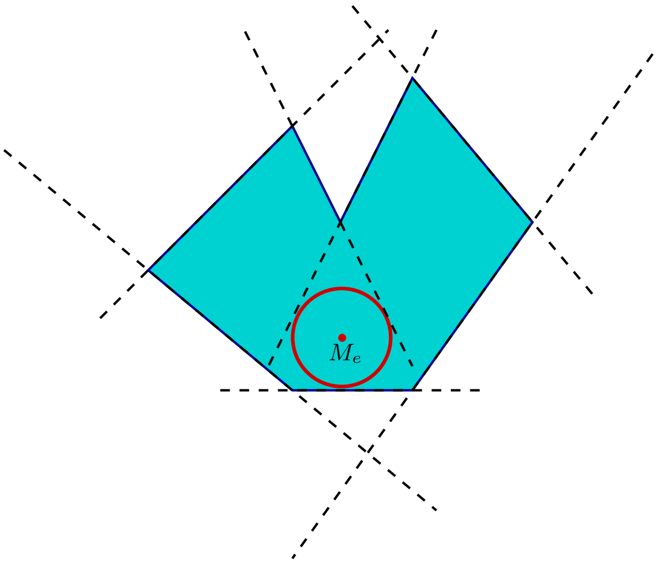


FIG. 2.2. *A star-shaped face with the circle of radius ρ_* centered at M_e.*

every element E we have

$$v_* \, h_E^3 \leq |E|, \quad a_* \, h_E^2 \leq |e|, \quad l_* \, h_E \leq |\ell| \tag{2.4}$$

for all faces e and edges ℓ of E.

M4 (*star-shaped faces*). We assume that the mesh faces are flat and that there exists a positive number γ_* such that for each element E and for each face $e \in \partial E$ there exists a point $M_e \in e$ such that e is star-shaped with respect to every point in the disk of radius $\gamma_* h_E$ centered at M_e.

We recall that e is star shaped with respect to a point $P \in e$ if every straight ray exiting from P (in the plane of e) intersects ∂e only once. In what follows we shall often use the notation

$$\rho_* = \gamma_* h_E, \tag{2.5}$$

which is illustrated in Figure 2.2.

M5 (*the pyramid property*). With the notation of Assumption M4, we further assume that for every $E \in \mathcal{T}_h$, and for every $e \in \partial E$, there exists a pyramid P_E^e *contained in* E such that its base equals e, its height equals $\gamma_* h_E$, and the projection of its vertex onto e is M_e.

M6 (*star-shaped elements*). We assume that there exists a positive number τ_* such that for each element E there exists a point $M_E \in E$ such that E is star shaped with respect to every point in the sphere of radius $\tau_* h_E$ centered at M_E.

As before, we say that E is star shaped with respect to a point $P \in E$ if every straight ray exiting from P intersects ∂E only once.

3. MFD method. Let us introduce an operator $\mathcal{G}$, $\mathcal{G}\, p = -\mathbf{K} \mathrm{grad}\, p$, which we refer to as the flux operator. Furthermore, we introduce the following scalar products:

$$(\mathbf{F}, \, \mathbf{G})_X = \int_\Omega \mathbf{F} \cdot \mathbf{K}^{-1} \mathbf{G} \, \mathrm{d}V \tag{3.1}$$

and

$$(p,\, q)_Q = \int_\Omega pq \,\mathrm{d}V \tag{3.2}$$

in the space X of velocities and in the space Q of pressures, respectively. Using the above notation, we may rewrite the Green's formula

$$\int_\Omega \mathbf{F} \cdot (\mathbf{K}^{-1} \mathcal{G}\, p) \,\mathrm{d}V = \int_\Omega p \,\mathrm{div}\, \mathbf{F} \,\mathrm{d}V \tag{3.3}$$

in the equivalent form

$$(\mathbf{F},\, \mathcal{G}\, p)_X = (p,\, \mathrm{div}\, \mathbf{F})_Q.$$

The last expression clearly states that the flux and divergence operators are adjoint to each other:

$$\mathcal{G} = \mathrm{div}^*.$$

The MFD method produces discretizations of these operators which are adjoint to each other with respect to scalar products in the discrete velocity and pressure spaces.

The *first* step of the MFD method is to specify the degrees of freedom for physical variables p and $\mathbf{F}$ and their location.

We consider the space Q^d of discrete pressures that are constant on each polyhedron E. For $\mathbf{q} \in Q^d$ we shall denote by q_E (or by $(\mathbf{q})_E$) its (constant) value on E. The dimension, N_Q, of Q^d is obviously equal to the number of polyhedrons in $\mathcal{T}_h$. In what follows, we shall denote by Q^d either the vector space $\mathbb{R}^{N_Q}$ or the space of piecewise constant functions depending on context. The identification will be obvious and no confusion should arise.

The definition of the space of discrete velocities requires some additional considerations. To every element E in $\mathcal{T}_h$ and to every face e of E we associate a number F_E^e and the vector field $F_E^e\, \mathbf{n}_E^e$, where $\mathbf{n}_E^e$ is the unit normal to e that points outside of E. We clearly make the *continuity* assumption that for each face e shared by two polyhedra E_1 and E_2, we have

$$F_{E_1}^e = -F_{E_2}^e. \tag{3.4}$$

We denote the vector space of face-based velocity unknowns by X^d. The number, N_X, of our discrete velocity unknowns is equal to the number of boundary faces plus *twice* the number of internal faces. In our theoretical discussion, we shall consider X^d as the subspace of $\mathbb{R}^{N_X}$ which verifies (3.4).

For a discrete velocity field $\mathbf{G}$ we will denote by $\mathbf{G}_E$ its restriction to the boundary of E, and by G_E^e (or by $(\mathbf{G}_E)^e$) the restriction of $\mathbf{G}_E \cdot \mathbf{n}_E$ to a face e belonging to the boundary of E. It will be convenient sometimes to use the notation

$$X_E^d := \{\text{restrictions of } X^d \text{ to the element } E\}. \tag{3.5}$$

It is clear that, in practice, condition (3.4) will make the number of *true independent unknowns* equal the total number of mesh faces. This means that, in a computer program, we shall prescribe one direction for the normal to each internal face e, and assign a single unknown G^e to each face, assuming that each of the two G_E^e coincides either with G^e (when the outward normal $\mathbf{n}_E$ on e coincides with the prescribed direction) or with $-G^e$ (otherwise).

To summarize, one pressure unknown is defined on each polyhedron and the discrete velocities are defined as face-based normal components. Once we get the degrees of freedom in Q^d and in X^d, we can define interpolation operators from the spaces of smooth enough scalar- and vector-valued functions to the discrete spaces Q^d and X^d, respectively. To every function q in $L^1(\Omega)$ we associate the element $\mathbf{q}^I \in Q^d$ defined by

$$(\mathbf{q}^I)_E := \frac{1}{|E|}\int_E q \,\mathrm{d}V \quad \forall E \in \mathcal{T}_h. \tag{3.6}$$

Similarly, for every vector-valued function $\mathbf{G} \in (L^s(\Omega))^3$, $s > 2$, with $\operatorname{div}\mathbf{G} \in L^2(\Omega)$, we define $\mathbf{G}^I \in X^d$ by

$$\left(\mathbf{G}^I_E\right)^e := \frac{1}{|e|}\int_e \mathbf{G}\cdot\mathbf{n}_E \,\mathrm{d}S \quad \forall\, E \in \mathcal{T}_h \quad \forall\, e \in \partial E. \tag{3.7}$$

In the next section, we shall prove that this interpolation operator is well defined and uniformly bounded. In what follows, we shall use bold capital letters either for vectors from X^d or for continuous vector functions depending on context and leaving no room for confusion.

The *second* step of the MFD method is to equip the spaces of discrete pressures and velocities with scalar products. The scalar product on the vector space Q^d is given by

$$[\mathbf{p},\, \mathbf{q}]_{Q^d} = \sum_{E\in\mathcal{T}_h} p_E\, q_E |E| \quad \forall\, \mathbf{p},\, \mathbf{q} \in Q^d. \tag{3.8}$$

In order to define the scalar product in X^d, we first define a scalar product $[\mathbf{F},\, \mathbf{G}]_E$ for every element $E \in \mathcal{T}_h$ in the following way. Let $e_1, e_2, \dots, e_{k_E}$ be a numbering of the faces of the element E (where k_E is clearly the total number of faces). We assume that we are given (for each E) a symmetric positive definite $k_E \times k_E$ matrix $M_E \equiv \{M_{E,i,j}\}$, and we set

$$[\mathbf{F},\, \mathbf{G}]_E = \sum_{i,j=1}^{k_E} M_{E,i,j}\, (\mathbf{F}_E)^{e_i}\, (\mathbf{G}_E)^{e_j} \quad \forall\, \mathbf{F},\, \mathbf{G} \in X^d \quad \forall E \in \mathcal{T}_h. \tag{3.9}$$

Some minimal approximation properties for the scalar product (3.9) are required. The construction of the matrix M_E is a nontrivial task for a polyhedral element. We shall return to this problem in section 5. For the time being, we just *assume* that the scalar product (3.9) has the following property.

S1 (*stability of* $[\cdot,\, \cdot]_E$). We assume that there exist two positive constants s_* and S^* independent of h and E such that, for every $\mathbf{G} \in X^d$ and for every $E \in \mathcal{T}_h$, one has

$$s_* \sum_{e\in\partial E} (G^e_E)^2\, |E| \le\, [\mathbf{G},\, \mathbf{G}]_E\, \le S^* \sum_{e\in\partial E} (G^e_E)^2\, |E|. \tag{3.10}$$

From (3.9) we can easily construct the scalar product in X^d by setting

$$[\mathbf{F},\, \mathbf{G}]_{X^d} = \sum_{E\in\mathcal{T}_h} [\mathbf{F},\, \mathbf{G}]_E \quad \forall\, \mathbf{F},\, \mathbf{G} \in X^d. \tag{3.11}$$

The *third* step of the MFD method is to derive an approximation to the divergence operator. The discrete divergence operator, $\mathcal{DIV}^d : X^d \to Q^d$, naturally arises from the Gauss divergence theorem as

$$(\mathcal{DIV}^d\, \mathbf{F})_E \overset{def}{=} \frac{1}{|E|} \sum_{e\in\partial E} F_E^e\, |e|. \tag{3.12}$$

We point out that our interpolation operators, in some sense, *commute* with the divergence operator. Indeed, for every vector field $\mathbf{G}$ smooth enough, we can use (3.12), (3.7), the Gauss divergence theorem, and (3.6) to obtain

$$(\mathcal{DIV}^d\, \mathbf{G}^I)_E = \frac{1}{|E|} \sum_{e\in\partial E} \left(\mathbf{G}_E^I\right)^e |e| = \frac{1}{|E|}\int_{\partial E} \mathbf{G}\cdot\mathbf{n}_E \,\mathrm{d}S = \frac{1}{|E|}\int_E \operatorname{div}\mathbf{G}\,\mathrm{d}V = (\operatorname{div}\mathbf{G})_E^I \tag{3.13}$$

for every element E in $\mathcal{T}_h$.

The *fourth* step of the MFD method is to define the discrete flux operator, $\mathcal{G}^d : Q^d \to X^d$, as the adjoint to the discrete divergence operator, $\mathcal{DIV}^d$, with respect to scalar products (3.8) and (3.11), i.e.,

$$[\mathbf{F},\, \mathcal{G}^d\, \mathbf{p}]_{X^d} = [\mathbf{p},\, \mathcal{DIV}^d\, \mathbf{F}]_{Q^d} \quad \forall \mathbf{p}\in Q^d \quad \forall \mathbf{F}\in X^d. \tag{3.14}$$

Using the discrete flux and divergence operators, the continuous problem (2.1), (2.2) is discretized as follows:

$$\mathcal{DIV}^d\;\; \mathbf{F}_d = \mathbf{b}, \tag{3.15}$$

$$\mathbf{F}_d = \mathcal{G}^d\, \mathbf{p}_d, \tag{3.16}$$

where $\mathbf{b} \equiv \mathbf{b}^I$ is the vector of mean values of the source function b.

4. Stability analysis. In this section we analyze the stability of the MFD discretization (3.15)–(3.16) following the well-established theory of saddle-point problems [5]. More precisely, we prove the coercivity condition (4.4) and the inf-sup condition (4.5).

Using the discrete Green's formula (3.14), we rewrite (3.15) and (3.16) in a form more suitable for analysis:

$$[\mathbf{F}_d,\, \mathbf{G}]_{X^d} - [\mathbf{p}_d,\, \mathcal{DIV}^d\, \mathbf{G}]_{Q^d} = 0 \quad \forall\, \mathbf{G}\in X^d, \tag{4.1}$$

$$[\mathcal{DIV}^d\, \mathbf{F}_d,\, \mathbf{q}]_{Q^d} = [\mathbf{b},\, \mathbf{q}]_{Q^d} \qquad \forall\, \mathbf{q}\in Q^d. \tag{4.2}$$

Let us introduce the following mesh norms on discrete spaces X^d and Q^d:

$$|||\mathbf{p}|||_{Q^d}^2 := [\mathbf{p},\, \mathbf{p}]_{Q^d}, \quad |||\mathbf{F}|||_{X^d}^2 := [\mathbf{F},\, \mathbf{F}]_{X^d},$$

and

$$|||\mathbf{F}|||_{div}^2 := |||\mathbf{F}|||_{X^d}^2 + \sum_{E\in\mathcal{T}_h} h_E^2\, \|\mathcal{DIV}^d\, \mathbf{F}\|_{L^2(E)}^2. \tag{4.3}$$

Let V^d be the space of divergence-free discrete fluxes:

$$V^d = \{\mathbf{F}\in X^d : \quad \mathcal{DIV}^d\, \mathbf{F} = 0\}.$$

We begin the stability analysis by noticing that the scalar product (3.11) is continuous. It is also obvious that the scalar product satisfies the V^d-ellipticity condition:

$$[\mathbf{F},\, \mathbf{F}]_{X^d} \geq |||\mathbf{F}|||_{div}^2 \quad \forall\, \mathbf{F} \in V^d. \tag{4.4}$$

The analysis of the inf-sup condition is more involved. Following [5], for every $\mathbf{q} \in Q^d$, we have to find a vector $\mathbf{G} \in X^d$ such that

$$[\mathcal{DIV}^d\, \mathbf{G},\, \mathbf{q}]_{Q^d} \geq \beta_* |||\mathbf{G}|||_{div}\ |||\mathbf{q}|||_{Q^d}, \tag{4.5}$$

where β_* is a positive constant independent of $\mathbf{q}$, $\mathbf{G}$, and $\mathcal{T}_h$. Let us denote by $q^h \in L^2(\Omega)$ the piecewise constant function on $\mathcal{T}_h$ with values given by the entries of the vector $\mathbf{q}$ (so that $(q^h)^I \equiv \mathbf{q}$). It is obvious that $\|q^h\|_{L^2(\Omega)} = |||\mathbf{q}|||_{Q^d}$. Let us consider the homogeneous Dirichlet boundary value problem

$$\Delta\psi = q^h \quad \text{in} \quad \Omega.$$

Since Ω has a Lipschitz-continuous boundary, there exist an $s > 2$ and a constant C^*_Ω such that

$$\|\psi\|_{W^1_s(\Omega)} \leq\, C^*_\Omega\, \|q^h\|_{L^2(\Omega)}. \tag{4.6}$$

Let $\mathbf{H} = \nabla\psi$, so that we have immediately

$$\operatorname{div} \mathbf{H} = q^h, \tag{4.7}$$

and from (4.6)

$$\|\mathbf{H}\|_{(L^s(\Omega))^3} + \left(\sum_{E \in \mathcal{T}_h} h_E^2\, \|\operatorname{div} \mathbf{H}\|^2_{L^2(E)} \right)^{1/2} \leq\, (C^*_\Omega + h)\|q^h\|_{L^2(\Omega)}. \tag{4.8}$$

We now set

$$\mathbf{G} := \mathbf{H}^I\ \equiv\ (\nabla\psi)^I, \tag{4.9}$$

where the interpolation operator is still the one defined in (3.7). Thanks to the commutative property (3.13) and to (4.7), we have

$$\mathcal{DIV}^d\ \mathbf{G} = (q^h)^I \equiv \mathbf{q}. \tag{4.10}$$

Thus, inequality (4.5) is reduced to

$$|||\mathbf{q}|||_{Q^d} \geq \beta_* |||\mathbf{G}|||_{div}. \tag{4.11}$$

At this point we need the following technical lemma.

Lemma 4.1. *Under Assumptions* M1–M6 *and* S1*, for every $s > 2$, there exists a positive constant β^*_s such that*

$$|||\mathbf{G}^I|||_{div} \leq \beta^*_s \left\{ \|\mathbf{G}\|_{(L^s(\Omega))^3} + \left(\sum_{E \in \mathcal{T}_h} h_E^2\, \|\operatorname{div} \mathbf{G}\|^2_{L^2(E)} \right)^{1/2} \right\} \tag{4.12}$$

for every $\mathbf{G} \in (L^s(\Omega))^3$ with $\operatorname{div} \mathbf{G} \in L^2(\Omega)$, and where $\mathbf{G}^I$ is defined in (3.7)*.*

Collecting (4.9) and (4.12), we get

$$|||\mathbf{G}|||_{div} = |||\mathbf{H}^I|||_{div} \leq \beta_s^* \left\{ \|\mathbf{H}\|_{(L^s(\Omega))^3} + \left(\sum_{E \in \mathcal{T}_h} h_E^2 \, \|\mathrm{div}\, \mathbf{H}\|^2_{L^2(E)} \right)^{1/2} \right\}.$$

This, together with (4.8), implies (4.11), and hence (4.5), with $\beta_* = \left(\beta_s^* \, (C_\Omega^* + h)\right)^{-1}$. Therefore, we have just to prove Lemma 4.1.

Proof of Lemma 4.1. From (3.13) we immediately have

$$|||\mathcal{DIV}^d \, \mathbf{G}^I|||_{Q^d} = |||(\mathrm{div}\, \mathbf{G})^I|||_{Q^d} \leq \|\mathrm{div}\, \mathbf{G}\|_{L^2(\Omega)}. \tag{4.13}$$

Therefore, in view of (4.3), it is sufficient to prove that there exists a constant $\widetilde{\beta_s^*}$ such that

$$|||\mathbf{G}^I|||_{X^d} \leq \widetilde{\beta_s^*} \left\{ \|\mathbf{G}\|_{(L^s(\Omega))^3} + \left(\sum_{E \in \mathcal{T}_h} h_E^2 \|\mathrm{div}\, \mathbf{G}\|^2_{L^2(E)} \right)^{1/2} \right\}. \tag{4.14}$$

The desired result (4.12) follows from (4.14) with $\beta_s^* = \widetilde{\beta_s^*} + 1$. In the following discussion, we shall make a wide use of the conjugate exponent t, depending on s through the usual formula

$$\frac{1}{s} + \frac{1}{t} = 1. \tag{4.15}$$

Assumption (3.10) implies clearly that

$$[\mathbf{G}^I, \, \mathbf{G}^I]_{X^d} \leq S^* \sum_{E \in \mathcal{T}_h} |E| \sum_{e \in \partial E} \left(G_E^e \right)^2, \tag{4.16}$$

so that we have to estimate the (G_E^e)'s in terms of $\mathbf{G}$, or, rather, in terms of the norm of $\mathbf{G}$ appearing in (4.12). Our basic instrument for that is called the *lift property*. The main difficulty, in various cases, will be to prove that the lift property holds true.

LP (lift property). For every $t < 2$ there exists a constant $\lambda^* = \lambda^*(t)$ such that for every $E \in \mathcal{T}_h$ and for every $e \in \partial E$ there exists a function φ_E^e from E to $\mathbb{R}$ that verifies

$$\varphi_E^e = 1 \quad \text{on} \quad e, \quad \varphi_E^e = 0 \quad \text{on} \quad \partial E \setminus e \tag{4.17}$$

and

$$\left\|\varphi_E^e\right\|_{L^2(E)} \leq \lambda^* h_E^{3/2}, \quad \left\|\nabla \varphi_E^e\right\|_{(L^t(E))^3} \leq \lambda^* h_E^{3/t-1}. \tag{4.18}$$

The lift property LP is proved in the appendix.

Up to an approximation of $\mathbf{G}$ by smooth functions, and passage to the limit, we have, using (3.7), (4.17), the Green's formula,

$$\begin{aligned} G_E^e = \frac{1}{|e|} \int_e \mathbf{G} \cdot \mathbf{n}_E \, \mathrm{d}S &= \frac{1}{|e|} \int_{\partial E} \varphi_E^e \mathbf{G} \cdot \mathbf{n}_E \, \mathrm{d}S \\ &= \frac{1}{|e|} \int_E \mathbf{G} \cdot \nabla \varphi_E^e \, \mathrm{d}V + \frac{1}{|e|} \int_E \varphi_E^e \, \mathrm{div}\, \mathbf{G} \, \mathrm{d}V. \end{aligned} \tag{4.19}$$

Using the Hölder inequality and (4.18) in (4.19), we then have

$$\begin{aligned}|e|\,G_E^e &\le \|\mathbf{G}\|_{L^s(E)}\left\|\nabla\varphi_E^e\right\|_{L^t(E)} + \|\mathrm{div}\,\mathbf{G}\|_{L^2(E)}\left\|\varphi_E^e\right\|_{L^2(E)}\\ &\le \lambda^*\left\{(h_E)^{3/t-1}\|\mathbf{G}\|_{L^s(E)} + (h_E)^{3/2}\,\|\mathrm{div}\,\mathbf{G}\|_{L^2(E)}\right\}.\end{aligned}$$

Taking the squares and remembering that $(a+b)^2 \le 2(a^2+b^2)$, we have

$$|e|^2\left(G_E^e\right)^2 \le 2\,(\lambda^*)^2\left\{(h_E)^{6/t-2}\|\mathbf{G}\|^2_{L^s(E)} + (h_E)^3\,\|\mathrm{div}\,\mathbf{G}\|^2_{L^2(E)}\right\}. \tag{4.20}$$

On the other hand, using conditions (2.4), we easily obtain

$$|E| \le h_E^3 = h_E^{-1}\left(h_E^2\right)^2 \le h_E^{-1}\,(a^*)^{-2}|e|^2. \tag{4.21}$$

We can now join (4.21) with (4.20) to deduce that

$$\begin{aligned}|E|\left(G_E^e\right)^2 &\le h_E^{-1}\,(a^*)^{-2}|e|^2\left(G_E^e\right)^2\\ &\le \sigma^*\left\{(h_E)^{6/t-3}\|\mathbf{G}\|^2_{L^s(E)} + (h_E)^2\,\|\mathrm{div}\,\mathbf{G}\|^2_{L^2(E)}\right\},\end{aligned} \tag{4.22}$$

where $\sigma^* = 2\,(\lambda^*)^2\,(a^*)^{-2}$. Now we can sum (4.22) over all faces e of E and then over all elements E of $\mathcal{T}_h$. We use (4.16) and Assumption M2 on the number of faces per element to get

$$\begin{aligned}|||\mathbf{G}^I|||^2_{X^d} &\le N_e\,S^*\,\sigma^*\left\{\sum_{E\in\mathcal{T}_h}(h_E)^{6/t-3}\|\mathbf{G}\|^2_{L^s(E)} + \sum_{E\in\mathcal{T}_h} h_E^2\,\|\mathrm{div}\,\mathbf{G}\|^2_{L^2(E)}\right\}\\ &\le N_e\,S^*\,\sigma^*\left\{\left(\sum_{E\in\mathcal{T}_h}\left\{(h_E)^{6/t-3}\right\}^r\right)^{1/r}\left(\sum_{E\in\mathcal{T}_h}\|\mathbf{G}\|^s_{L^s(E)}\right)^{2/s}\right.\\ &\qquad\left. + \sum_{E\in\mathcal{T}_h} h_E^2\|\mathrm{div}\,\mathbf{G}\|^2_{L^2(E)}\right\},\end{aligned} \tag{4.23}$$

where in the last step we applied the Hölder inequality with r, the conjugate exponent of $s/2$,

$$\frac{1}{r} + \frac{2}{s} = 1. \tag{4.24}$$

A simple algebraic manipulation using (4.15) and (4.24) gives

$$\sum_{E\in\mathcal{T}_h}\left\{(h_E)^{6/t-3}\right\}^r = \sum_{E\in\mathcal{T}_h} h_E^3 \le v_*^{-1}|\Omega|, \tag{4.25}$$

where we have also used (2.4) in the last step. Inserting (4.25) into (4.23), we finally get

$$|||\mathbf{G}^I|||_{X^d} \le \widetilde{\beta_s^*}\left\{\|\mathbf{G}\|_{(L^s(\Omega))^3} + \left(\sum_{E\in\mathcal{T}_h} h_E^2\|\mathrm{div}\,\mathbf{G}\|^2_{L^2(E)}\right)^{1/2}\right\}, \tag{4.26}$$

where $\widetilde{\beta_s^*}$ depends only on $\lambda^*(t)$, S^*, v_*, a_*, and N_e. This proves the assertion of the lemma. □

5. Convergence analysis.

5.1. Consistency assumption. In order to prove error estimates, we need some assumptions on the scalar product (3.11), and more precisely on the relationships between the continuous scalar product (3.1) and its discrete counterpart (3.11). Our basic assumption will be the following one.

S2 (*consistency of* $[\cdot,\cdot]_E$). For every element E, every linear function q^1 on E, and every $\mathbf{G} \in X^d$, we have

$$[(\tilde{\mathbf{K}}\,\nabla q^1)^I,\, \mathbf{G}]_E = \int_{\partial E} q^1\, \mathbf{G}_E \cdot \mathbf{n}_E \,\mathrm{d}S - \int_E q^1\,(\mathcal{DIV}^d\,\mathbf{G})_E \,\mathrm{d}V, \tag{5.1}$$

where $(\cdot)^I$ is the interpolation operator (3.7) and $\tilde{\mathbf{K}}$ is a constant tensor on E such that

$$\sup_{x\in E}\sup_{i,j} |\{\mathbf{K}(\mathbf{x})\}_{i,j} - \{\tilde{\mathbf{K}}\}_{i,j}| \le C_K^*\, h_E, \tag{5.2}$$

where C_K^* is a constant independent of E.

Note that $\tilde{\mathbf{K}}$ may be any reasonable piecewise constant approximation of $\mathbf{K}$. In practice, we use either the value of $\mathbf{K}$ at the polyhedron mass center or its mean value.

Condition (5.1) is rather new and requires some comments. First, we point out that for divergence-free vectors, $\mathbf{G} \in V^d$, it reads

$$[(\tilde{\mathbf{K}}\,\nabla q^1)^I,\, \mathbf{G}]_E = \int_{\partial E} q^1\, \mathbf{G}_E \cdot \mathbf{n}_E \,\mathrm{d}S \tag{5.3}$$

showing the remarkable property of using only *boundary integrals*. However, as $\mathcal{DIV}^d\,\mathbf{G}$ is constant in each E and q^1 is supposed to be linear, the volume integral appearing in (5.1) is not difficult to compute. Taking $\mathbf{G} = (\tilde{\mathbf{K}}\,\nabla\tilde{q}^1)^I$ (with $\tilde{q}^1$ another polynomial of degree ≤ 1) in (5.3), we conclude that Assumption S2 implies that *the scalar product* (3.11) *gives an exact value for the integral of two constant velocities.*

In the context of the local MFD method [13], and taking for simplicity $\tilde{\mathbf{K}} = \mathbf{I}$, condition (5.1) means that the discrete gradient operator is exact for linear functions, i.e., $\mathcal{G}^d\,(\mathbf{q}^1)^I$ is a constant vector whose entries are equal to ∇q^1. This property has been used in [18] to build a one-parameter family of symmetric positive definite matrices M_E for a triangle. As a particular case, this family includes the mass matrix appearing in the finite element discretizations with the Raviart–Thomas finite elements.

What is still remarkable in (5.1) is that *it does not require the construction of a lifting operator* from the values G_E^e on ∂E to the interior of E. It is not difficult to show, however, that *if* we have any reasonable lifting operator R_E, then the choice

$$[\mathbf{F},\, \mathbf{G}]_E := \int_E \tilde{\mathbf{K}}^{-1} R_E(\mathbf{F}_E) \cdot R_E(\mathbf{G}_E)\,\mathrm{d}V$$

will automatically satisfy (5.1) as well as (3.10). We have indeed the following theorem.

THEOREM 5.1. *Assume that for every element $E \in \mathcal{T}_h$ we have a lifting operator R_E acting on X_E^d (the restriction of X^d to E) and with values in $(L^2(E))^3$ such that*

$$\begin{aligned} R_E(\mathbf{G}_E)\cdot\mathbf{n}_E &\equiv \mathbf{G}_E\cdot\mathbf{n}_E \quad && \textit{on} \quad \partial E,\\ \operatorname{div} R_E(\mathbf{G}_E) &\equiv (\mathcal{DIV}^d\,\mathbf{G})_E \quad && \textit{in} \quad E \end{aligned} \tag{5.4}$$

for all $\mathbf{G} \in X^d$, *and*

$$R_E\big(\mathbf{G}_E^I\big) = \mathbf{G} \tag{5.5}$$

for all $\mathbf{G}$ *constant on* E. *Then the choices*

$$\{\tilde{\mathbf{K}}\}_{i,j} := \frac{1}{|E|}\int_E \{\mathbf{K}\}_{i,j}\,\mathrm{d}V \tag{5.6}$$

and

$$[\mathbf{F},\,\mathbf{G}]_E := \int_E \tilde{\mathbf{K}}^{-1} R_E(\mathbf{F}_E)\cdot R_E(\mathbf{G}_E)\,\mathrm{d}V \tag{5.7}$$

will automatically satisfy (5.2) *and* (5.1). *If, moreover, there exist two positive constants* c_R^* *and* C_R^*, *independent of* E *such that*

$$c_R^*\left(|E|\sum_{e\in\partial E}\big(G_E^e\big)^2\right)^{1/2} \le \|R_E(\mathbf{G})\|_{(L^2(E))^3} \le C_R^*\left(|E|\sum_{e\in\partial E}\big(G_E^e\big)^2\right)^{1/2} \tag{5.8}$$

for all $\mathbf{G} \in X^d$, *then* (3.10) *will also hold with constants* s_* *and* S^* *depending only on* c_R^*, C_R^*, *and on the constants* κ_*, κ^* *from* (2.3).

Proof. The validity of (5.2) is immediate. The validity of (5.1) is also easily checked:

$$\begin{aligned}
[(\tilde{\mathbf{K}}\,\nabla q^1)^I,\,\mathbf{G}]_E &= \int_E \tilde{\mathbf{K}}^{-1} R_E((\tilde{\mathbf{K}}\,\nabla q^1)^I_E)\cdot R_E(\mathbf{G}_E)\,\mathrm{d}V \quad \text{(use (5.5) and } \nabla q^1 = \text{const)}\\
&= \int_E \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{K}}\,\nabla q^1\cdot R_E(\mathbf{G}_E)\,\mathrm{d}V \qquad \text{(use } \tilde{\mathbf{K}}^{-1}\tilde{\mathbf{K}} = Id\text{)}\\
&= \int_E \nabla q^1\cdot R_E(\mathbf{G}_E)\,\mathrm{d}V \qquad \text{(integrate by parts)}\\
&= \int_{\partial E} q^1\, R_E(\mathbf{G}_E)\cdot\mathbf{n}_E\,\mathrm{d}S - \int_E q^1\,\mathrm{div}R_E(\mathbf{G}_E)\,\mathrm{d}V \quad \text{(use (5.4))}\\
&= \int_{\partial E} q^1\,\mathbf{G}_E\cdot\mathbf{n}_E\,\mathrm{d}S - \int_E q^1\,(\mathcal{DIV}^d\,\mathbf{G})_E\,\mathrm{d}V.
\end{aligned}$$

Finally, (3.10) follows immediately from (5.7), (2.3), and (5.8) after noting that (2.3) is equivalent to

$$(\kappa^*)^{-1}\|\mathbf{v}\|^2 \le \mathbf{v}^T\mathbf{K}^{-1}(\mathbf{x})\mathbf{v} \le (\kappa_*)^{-1}\|\mathbf{v}\|^2 \quad \forall\,\mathbf{v}\in\mathbb{R}^3 \quad \forall\,\mathbf{x}\in\Omega. \tag{5.9}$$

This ends the proof of the theorem. □

A possible way of getting (5.1) is, therefore, to construct a lifting operator R_E satisfying (5.4), (5.5), and (5.8), and then define M_E following (5.7). For instance, the way followed in [16] for polygonal domains can be interpreted as the construction of a lifting operator satisfying (5.4) and (5.5).

In general, we may consider assumption (5.1) as a system of linear equations where the unknowns are the coefficients of M_E, and use it, in each element E, to construct the matrix M_E. Since the matrix M_E should be symmetric and positive definite, this is a problem with nonlinear constraints. An analytical solution has been found only for triangular elements [18].

Let us see this in more detail. We consider an element E having k_E faces. Equation (5.1) should then hold for k_E different possible choices of $\mathbf{G}_E$ and three possible choices of q^1 corresponding to $q^1=x$, $q^1=y$, and $q^1=z$. Note that for $q^1=1$, (5.1) is automatically satisfied as it is reduced to our definition of the operator $\mathcal{DIV}^d$. We have, therefore, $3k_E$ equations. It can be shown that only $3k_E-3$ equations are linearly independent. Since $\tilde{\mathbf{K}}$, and hence M_E, is symmetric, the number of unknown coefficients of M_E is $k_E(k_E+1)/2$, that is, bigger than $3k_E-3$ as soon as $k_E\geq 4$. The system will always be compatible, since we could always define a lifting R_E first by solving, for each $\mathbf{G}_E$, the Neumann problem

$$\begin{aligned} \Delta\chi &= \mathcal{DIV}^d\,\mathbf{G}_E && \text{in} \quad E,\\ \partial\chi/\partial\mathbf{n}_E &= \mathbf{G}_E\cdot\mathbf{n}_E && \text{on} \quad \partial E, \end{aligned}$$

then by taking $R_E(\mathbf{G}_E) := \nabla\chi$, and finally by defining M_E through (5.7). This would be totally impractical but shows that at least a solution M_E of (5.1), symmetric and positive definite, exists (although, in general, the solution will not be unique).

A sparsity structure could be imposed on M_E in order to reduce the number of unknowns. For instance, we can require that each face interacts only with a few neighboring faces, reducing the number of unknowns to $3k_E-3$, which equals the number of equations and makes the linear system much easier to solve on the computer (see [6] for more detail).

An advantage of this approach is that it can be rather easily extended to faces that are not flat. This is a case in which the construction of an explicit lifting operator might prove to be very difficult. We shall consider meshes with curved faces in the future publications.

5.2. Error estimate for the vector variable. Using Assumption S2, we are going to prove error estimates for our discretization. Let $(p,\,\mathbf{F})$ be the exact solution of (2.1) and (2.2), let $(\mathbf{p}_d,\mathbf{F}_d)$ be the discrete solution (see (3.15) and (3.16)), and let $\mathbf{p}^I$ and $\mathbf{F}^I$ be the interpolants of the exact solution. Finally, for every element E, we denote by p^1_E a suitable polynomial of degree ≤ 1 that approximates p, and that will be decided later on. We notice first that from (2.1), (3.13), and (3.15), we easily have

$$\mathcal{DIV}^d\,(\mathbf{F}^I-\mathbf{F}_d) = \mathbf{b}-\mathbf{b} = 0. \tag{5.10}$$

Using (2.2) and (3.16), then (3.14), and finally (5.10), we get

$$\begin{aligned} [\mathbf{F}^I-\mathbf{F}_d,\,\mathbf{F}^I-\mathbf{F}_d]_{X^d} &= [(-\mathbf{K}\,\nabla p)^I,\,\mathbf{F}^I-\mathbf{F}_d]_{X^d} - [\mathcal{G}^{\,d}\mathbf{p}_d,\mathbf{F}^I-\mathbf{F}_d]_{X^d}\\ &= [(-\mathbf{K}\,\nabla p)^I,\,\mathbf{F}^I-\mathbf{F}_d]_{X^d} - [\mathbf{p}_d,\,\mathcal{DIV}^d\,(\mathbf{F}^I-\mathbf{F}_d)]_{Q^d}\\ &= [(-\mathbf{K}\,\nabla p)^I,\,\mathbf{F}^I-\mathbf{F}_d]_{X^d}. \end{aligned} \tag{5.11}$$

Then, adding and subtracting the terms, we have

$$\begin{aligned} |||\mathbf{F}^I-\mathbf{F}_d|||^2_{X^d} &= [(-\mathbf{K}\,\nabla p)^I+(\mathbf{K}\,\nabla p^1)^I,\,\mathbf{F}^I-\mathbf{F}_d]_{X^d} + [(-\mathbf{K}\nabla p^1)^I,\,\mathbf{F}^I-\mathbf{F}_d]_{X^d}\\ &= \mathbf{I}_1 + [(-\mathbf{K}\nabla p^1+\tilde{\mathbf{K}}\nabla p^1)^I,\,\mathbf{F}^I-\mathbf{F}_d]_{X^d} + [(-\tilde{\mathbf{K}}\nabla p^1)^I,\mathbf{F}^I-\mathbf{F}_d]_{X^d}\\ &= \mathbf{I}_1+\mathbf{I}_2+[(-\tilde{\mathbf{K}}\nabla p^1)^I,\,\mathbf{F}^I-\mathbf{F}_d]_{X^d}\\ &= \mathbf{I}_1+\mathbf{I}_2+\mathbf{I}_3. \end{aligned} \tag{5.12}$$

Using (5.1) and (5.10), the third term reads

$$\mathbf{I}_3 = \sum_{E\in\mathcal{T}_h} \left\{ \int_{\partial E} p^1_E\,(\mathbf{F}^I - \mathbf{F}_d)_E \cdot \mathbf{n}_E \,\mathrm{d}S - \int_E p^1_E\,(\mathcal{DIV}^d\,(\mathbf{F}^I - \mathbf{F}_d))_E \,\mathrm{d}V \right\}$$

$$= \sum_{E\in\mathcal{T}_h} \int_{\partial E} p^1_E\,(\mathbf{F}^I - \mathbf{F}_d)_E \cdot \mathbf{n}_E \,\mathrm{d}S. \tag{5.13}$$

We are, therefore, left with the problem of estimating $\mathbf{I}_1$, $\mathbf{I}_2$, and $\mathbf{I}_3$. A first estimate of $\mathbf{I}_2$ is trivial. From (5.2) we immediately have

$$\mathbf{I}_2 \equiv [(-\mathbf{K}\nabla p^1 + \tilde{\mathbf{K}}\nabla p^1)^I,\, \mathbf{F}^I - \mathbf{F}_d]_{X^d} \le C^*_K h\, |||(\nabla p^1)^I|||_{X^d}\, |||\mathbf{F}^I - \mathbf{F}_d|||_{X^d}, \tag{5.14}$$

where p^1 still has to be defined.

Let us recall some known properties of the approximation theory. For the sake of simplicity, we assume that our solution p is in $H^2(\Omega)$. Note that with a little additional effort we could use a weaker regularity and get a lower order of convergence.

We first recall that, under Assumption M6 (*star-shaped elements*), it is possible to find a constant C^*_{app}, depending only on τ_*, such that for every element E and for every $p \in H^2(E)$ there exist a constant p^0_E and a polynomial p^1_E of degree ≤ 1 such that

$$\left\|p - p^0_E\right\|_{L^2(E)} \le C^*_{app}\, h_E\, \|p\|_{H^1(E)}, \tag{5.15}$$

$$\left\|p - p^1_E\right\|_{L^2(E)} \le C^*_{app}\, h^2_E\, \|p\|_{H^2(E)}, \quad \left\|p - p^1_E\right\|_{H^1(E)} \le C^*_{app}\, h_E\, \|p\|_{H^2(E)} \tag{5.16}$$

(see [4, Lemma 4.3.8]). Concerning the error on faces, we can use a result due to Agmon made popular in the numerical analysis community by Arnold [1]. Applied to our case, it says that there exists a constant C^*_{agm}, depending only on the constant γ_* of Assumption M4, such that for every pyramid P^e_E (as described in Assumption M5), and for every function $\chi \in H^1(P^e_E)$, we have

$$\|\chi\|^2_{L^2(e)} \le C^*_{agm}\big(h^{-1}_E \|\chi\|^2_{L^2(P^e_E)} + h_E\, \|\chi\|^2_{H^1(P^e_E)}\big). \tag{5.17}$$

It is then immediate to derive from (5.17) that

$$\|\nabla\chi\|^2_{L^2(e)} \le C^*_{agm}\big(h^{-1}_E\, \|\chi\|^2_{H^1(P^e_E)} + h_E\, \|\chi\|^2_{H^2(P^e_E)}\big) \tag{5.18}$$

for every $\chi \in H^2(E)$. Applying this to the difference $p - p^1_E$, and using (5.16), we get

$$\left\|p - p^1_E\right\|^2_{L^2(e)} + h^2_E \left\|\nabla\big(p - p^1_E\big)\right\|^2_{L^2(e)} \le C^*_{face}\, h^3_E\, \|p\|^2_{H^2(E)}, \tag{5.19}$$

where C^*_{face} depends only on τ_* and γ_*.

Now, we can finish the estimate of $\mathbf{I}_2$. Note that ∇p^1 is a constant vector. Then, (5.16) and the triangle inequality give

$$|||\big(\nabla p^1_E\big)^I|||_{X^d} = \left\|\nabla p^1_E\right\|_{L^2(E)} \le \|\nabla p\|_{L^2(E)} + \left\|\nabla(p - p^1_E)\right\|_{L^2(E)} \le \big(1 + h_E C^*_{app}\big)\|p\|_{H^2(E)}.$$

Thus, we obtain immediately from (5.14) that

$$\mathbf{I}_2 \le C^*_{I_2}\, h\, \|p\|_{H^2(\Omega)}\, |||\mathbf{F}^I - \mathbf{F}_d|||_{X^d}, \tag{5.20}$$

where $C^*_{I_2}$ equals $(1 + h_E\, C^*_{app})C^*_K$ with C^*_K given in (5.2).

The estimate of $\mathbf{I}_1$ is obtained in the following lemma.

LEMMA 5.2. *Let $p \in H^2(\Omega)$ and let, in each $E \in \mathcal{T}_h$, p^1 be such that* (5.16) *holds. Let $(\cdot)^I$ be the interpolation operator defined in* (3.7)*, and let finally $\mathbf{G} \in X^d$. Then*

$$[(-\mathbf{K}\,\nabla p)^I + (\mathbf{K}\,\nabla p^1)^I,\, \mathbf{G}]_{X^d} \le C^*_{I_1}\, h\, \|p\|_{H^2(\Omega)}\, |||\mathbf{G}|||_{X^d}, \tag{5.21}$$

where the constant $C^*_{I_1}$ is independent of p, $\mathbf{G}$, and h.

Proof. The proof follows immediately from (3.10), the definition of the interpolation operator (3.7), the Cauchy–Schwarz inequality, and the approximation results quoted above. Indeed, we have

$$\begin{aligned} |||(-\mathbf{K}\,\nabla p)^I + (\mathbf{K}\,\nabla p^1)^I|||^2_{X^d} &\le S^* \sum_{E\in\mathcal{T}_h} \sum_{e\in\partial E} \big(((-\mathbf{K}\,\nabla p)^I + (\mathbf{K}\,\nabla p^1)^I)^e_E\big)^2\, |E| \\ &\le S^* \sum_{E\in\mathcal{T}_h} \sum_{e\in\partial E} \left(\frac{1}{|e|}\int_e \mathbf{K}\,\nabla\big(p - p^1_E\big)\cdot \mathbf{n}_E\, \mathrm{d}S\right)^2 |E| \\ &\le S^* \sum_{E\in\mathcal{T}_h} \sum_{e\in\partial E} \frac{1}{|e|}\, \big\|\mathbf{K}\,\nabla\big(p - p^1_E\big)\big\|^2_{L^2(e)} |E| \\ &\le C^*_{I_1}\, h^2\, \|p\|^2_{H^2(\Omega)}, \end{aligned}$$

where $C^*_{I_1}$ depends only on a_* given in (2.4), S^* given in (3.10), κ^* given in (2.3), N_e from Assumption M2, and C^*_{face} obtained in (5.19). □

The following lemma gives an estimate for $\mathbf{I}_3$.

LEMMA 5.3. *Let $p \in H^2(\Omega)$ and let, in each $E \in \mathcal{T}_h$, p^1 be such that* (5.16) *holds. Moreover, let $\mathbf{G} \in X^d$. Then*

$$\sum_{E\in\mathcal{T}_h} \int_{\partial E} p^1\, \mathbf{G}_E \cdot \mathbf{n}_E\, \mathrm{d}S \le C^*_{I_3}\, h\, \|p\|_{H^2(\Omega)}\, |||\mathbf{G}|||_{X^d}, \tag{5.22}$$

where the constant $C^*_{I_3}$ is independent of p, $\mathbf{G}$, and h.

Proof. The first (crucial) step of the proof uses the continuity of p and the fact that $\mathbf{G}_E \cdot \mathbf{n}_E$ takes opposite values for the two elements sharing a common internal face. Then, the result follows with usual instruments such as the Cauchy–Schwarz inequality and approximation results (5.16):

$$\begin{aligned} \sum_{E\in\mathcal{T}_h} \int_{\partial E} p^1_E\, \mathbf{G}_E \cdot \mathbf{n}_E\, \mathrm{d}S &= \sum_{E\in\mathcal{T}_h} \int_{\partial E} \big(p^1_E - p\big)\, \mathbf{G}_E \cdot \mathbf{n}_E\, \mathrm{d}S \\ &\le \sum_{E\in\mathcal{T}_h} \sum_{e\in\partial E} \big\|p - p^1_E\big\|_{L^2(e)}\, \big\|G^e_E\big\|_{L^2(e)} \\ &= \sum_{E\in\mathcal{T}_h} \sum_{e\in\partial E} \big\|p - p^1_E\big\|_{L^2(e)}\, \big|G^e_E\big|\, |e|^{1/2} \\ &\le v_*^{-1/2} (C^*_{face})^{1/2} \sum_{E\in\mathcal{T}_h} h_E \|p\|_{H_2(E)} \sum_{e\in\partial E} \big|G^e_E\big|\, |E|^{1/2} \\ &\le C^*_{I_3}\, h \|p\|_{H^2(\Omega)}\, |||\mathbf{G}|||_{X^d}, \end{aligned}$$

where $C^*_{I_3} = (v_*^{-1}\, s_*^{-1}\, C^*_{face})^{1/2} N_e$. This proves the assertion of the lemma. □

Combining (5.12) with (5.20), (5.21), and (5.22), we finally get the main convergence result.

THEOREM 5.4. *Under Assumptions* P1, M1–M6, *and* S1–S2, *let* $(p, \mathbf{F})$ *be the solution of* (2.1)–(2.2), *and let* $(p_d, \mathbf{F}_d)$ *be the discrete solution, given by* (3.15)–(3.16). *Moreover, let* $\mathbf{F}^I$ *be the interpolant of* $\mathbf{F}$, *introduced in* (3.7). *Then, we have*

$$|||\mathbf{F}^I - \mathbf{F}_d|||_{X^d} \le C^* \, h \, \|p\|_{H^2(\Omega)}, \tag{5.23}$$

where C^* *depends only upon the various constants appearing in Assumptions* P1, M1–M6, *and* S1–S2.

5.3. Error estimates for the scalar variable. In order to derive estimates on the scalar variable $\mathbf{p}_d$, we shall go back to the proof of inf-sup condition (4.5). For the sake of simplicity, we assume that Ω is convex. Let ψ be the solution of

$$\begin{aligned} -\mathrm{div}(\mathbf{K}\nabla\psi) &= \mathbf{p}^I - \mathbf{p}_d \quad && \text{in } \Omega, \\ \psi &= 0 && \text{on } \partial\Omega, \end{aligned}$$

where, for simplicity, we identified $\mathbf{p}_d - \mathbf{p}^I$ with the corresponding piecewise constant function. The convexity of Ω implies that there exists a constant C_Ω^*, depending only on Ω, such that

$$\|\psi\|_{H^2(\Omega)} \le C_\Omega^* \, |||\mathbf{p}_d - \mathbf{p}^I|||_{Q^d}. \tag{5.24}$$

We now set

$$\mathbf{H} = \mathbf{K}\nabla\psi \tag{5.25}$$

and define $\mathbf{G} \in X^d$ as $\mathbf{G} = \mathbf{H}^I$, so that

$$\mathcal{DIV}^d \, \mathbf{G} = \mathbf{p}_d - \mathbf{p}^I. \tag{5.26}$$

Finally, we denote by ψ^1 a piecewise linear approximation of ψ that satisfies (5.16) for each $E \in \mathcal{T}_h$. Using (5.26), then (4.1), then (3.6) and (3.13), then integrating by parts, and finally integrating once again by parts and using (2.1) and (2.2), we get

$$\begin{aligned} |||\mathbf{p}_d - \mathbf{p}^I|||_{Q^d}^2 &= [\mathcal{DIV}^d \, \mathbf{G}, \, \mathbf{p}_d - \mathbf{p}^I]_{Q^d} \\ &= [\mathbf{F}_d, \, \mathbf{G}]_{X^d} - [\mathcal{DIV}^d \, \mathbf{G}, \, \mathbf{p}^I]_{Q^d} = [\mathbf{F}_d, \, \mathbf{G}]_{X^d} - \int_\Omega p \, \mathrm{div}(\mathbf{K}\nabla\psi) \, \mathrm{d}V \\ &= [\mathbf{F}_d, \, \mathbf{G}]_{X^d} + \int_\Omega \mathbf{K}\nabla p \cdot \nabla\psi \, \mathrm{d}V \\ &= [\mathbf{F}_d, \, \mathbf{G}]_{X^d} + \int_\Omega b \, \psi \, \mathrm{d}V. \end{aligned}$$

Now, using the definition of $\mathbf{G}$ and adding and subtracting the terms, we have

$$\begin{aligned} |||\mathbf{p}_d - \mathbf{p}^I|||_{Q^d}^2 &= [\mathbf{F}_d, \, (\mathbf{K}\nabla\psi)^I - (\mathbf{K}\nabla\psi^1)^I]_{X^d} + [\mathbf{F}_d, \, (\mathbf{K}\nabla\psi^1)^I]_{X^d} + \int_\Omega b \, \psi \, \mathrm{d}V \\ &= J_1 + [\mathbf{F}_d, \, ((\mathbf{K} - \tilde{\mathbf{K}})\nabla\psi^1)^I]_{X^d} + [\mathbf{F}_d, \, (\tilde{\mathbf{K}}\nabla\psi^1)^I]_{X^d} + \int_\Omega b \, \psi \, \mathrm{d}V \\ &= J_1 + J_2 + [\mathbf{F}_d, \, (\tilde{\mathbf{K}}\nabla\psi^1)^I]_{X^d} + \int_\Omega b \, \psi \, \mathrm{d}V. \end{aligned} \tag{5.27}$$

Using (5.21), the term J_1 can be easily bounded by

$$J_1 \equiv [\mathbf{F}_d,\, (\mathbf{K}\nabla\psi)^I - (\mathbf{K}\nabla\psi^1)^I]_{X^d} \leq C^*_{I_1}\, h\, |||\mathbf{F}_d|||_{X^d}\, \|\psi\|_{H^2(\Omega)}. \tag{5.28}$$

The term J_2 is bounded as in (5.14), (5.20) by

$$J_2 \equiv [\mathbf{F}_d,\, ((\mathbf{K} - \tilde{\mathbf{K}})\nabla\psi^1)^I]_{X^d} \leq C^*_{I_2}\, h\, |||\mathbf{F}_d|||_{X^d}\, \|\psi\|_{H^2(\Omega)}. \tag{5.29}$$

For the third term in the last line of (5.27), we can use (5.1) to obtain

$$[\mathbf{F}_d,\, (\mathbf{K}\nabla\psi^1)^I]_{X^d} = \sum_{E\in\mathcal{T}_h} \int_{\partial E} \psi^1 (\mathbf{F}_d)_E \cdot \mathbf{n}_E \,\mathrm{d}S - \int_\Omega \mathbf{b}\,\psi^1 \,\mathrm{d}V. \tag{5.30}$$

With the help of (5.22), we then get

$$\left|[\mathbf{F}_d,\, (\tilde{\mathbf{K}}\nabla\psi^1)^I]_{X^d} + \int_\Omega b\,\psi \,\mathrm{d}V\right| \leq C^*_{I_3}\, h\, |||\mathbf{F}_d|||_{X^d}\, \|\psi\|_{H^2(\Omega)} + \left|\int_\Omega (b\,\psi - \mathbf{b}\psi^1)\,\mathrm{d}V\right|, \tag{5.31}$$

where the last term is easily bounded by $2\,C^*_{app}\, h\, \|b\|_{H^1(\Omega)}\, \|\psi\|_{H^1(\Omega)}$. Collecting inequalities (5.27)–(5.31), we obtain

$$|||\mathbf{p}_d - \mathbf{p}^I|||^2_{Q^d} \leq C^*\, h\, \left\{|||\mathbf{F}_d|||_{X^d} + \|b\|_{H^1(\Omega)}\right\} \|\psi\|_{H^2(\Omega)}, \tag{5.32}$$

which, combined with estimates (5.24), Theorem 5.4, and Lemma 4.1, gives the proof of the second convergence result.

Theorem 5.5. *Under assumptions of Theorem* 5.4*, plus the convexity of* Ω*, we have*

$$|||\mathbf{p}_d - \mathbf{p}^I|||_{Q^d} \leq C^*\, h\, (\|p\|_{H^2(\Omega)} + \|b\|_{H^1(\Omega)}), \tag{5.33}$$

where the constant C^* *depends only on the constants appearing in Assumptions* P1, M1–M6*, and* S1–S2*, on* C^*_Ω *appearing in* (5.24)*, and on* β^*_s *appearing in* (4.12).

It is interesting to note that, assuming that in each element E we had a suitable lifting R_E, a better estimate for the scalar variable could be obtained. We have indeed the following theorem.

Theorem 5.6. *Together with the assumptions of Theorem* 5.5*, assume, moreover, that for each element* E *we have a lifting operator* R_E *with properties* (5.4), (5.5)*, and* (5.8) *such that*

$$\|R_E(\mathbf{G}^I) - \mathbf{G}\|_{L^2(E)} \leq C^*_{Ra}\, h_E\, \|\mathbf{G}\|_{(H^1(E))^3} \quad \forall\, \mathbf{G} \in (H^1(E))^3 \quad \forall\, E \in \mathcal{T}_h, \tag{5.34}$$

where C^*_{Ra} *is a constant independent of* $\mathbf{G}$ *and* h_E*. Then, the choice*

$$[\mathbf{F},\, \mathbf{G}]_E := \int_E \mathbf{K}^{-1} R_E(\mathbf{F}_E) \cdot R_E(\mathbf{G}_E)\,\mathrm{d}V \tag{5.35}$$

will give

$$|||\mathbf{p}_d - \mathbf{p}^I|||_{Q^d} \leq C^*\, h^2\, \left(\|p\|_{H^2(\Omega)} + \|b\|_{H^1(\Omega)}\right), \tag{5.36}$$

where the constant C^* *depends only on the constants appearing in Assumptions* P1, M1–M6*, and* S1–S2*, on* C^*_Ω *appearing in* (5.24)*, on* β^*_s *appearing in* (4.12)*, and on* C^*_{Ra} *from* (5.34).

Proof. Let $R(\mathbf{G})$ be such that $R(\mathbf{G})|_E = R_E(\mathbf{G}_E)$. Following essentially [11] and using (5.26), then (4.1), (3.6), and (3.13) (as in the previous proof) with (5.4), then integrating by parts, and finally using (2.2) and (5.35), we get

$$\begin{aligned}|||\mathbf{p}_d - \mathbf{p}^I|||^2_{Q^d} &= [\mathcal{DIV}^d\,\mathbf{G},\, \mathbf{p}_d - \mathbf{p}^I]_{Q^d}\\ &= [\mathbf{F}_d,\, \mathbf{G}]_{X^d} - \int_\Omega p \operatorname{div} R(\mathbf{G})\,\mathrm{d}V\\ &= [\mathbf{F}_d,\, \mathbf{G}]_{X^d} + \int_\Omega \nabla p \cdot R(\mathbf{G})\,\mathrm{d}V = [\mathbf{F}_d,\, \mathbf{G}]_{X^d} + \int_\Omega \mathbf{K}^{-1}\mathbf{K}\nabla p \cdot R(\mathbf{G})\,\mathrm{d}V\\ &= \int_\Omega \mathbf{K}^{-1}(R(\mathbf{F}_d) - \mathbf{F})\, R(\mathbf{G})\,\mathrm{d}V.\end{aligned}$$

Adding and subtracting $\mathbf{H}$ defined in (5.25), we get

$$\begin{aligned}|||\mathbf{p}_d - \mathbf{p}^I|||^2_{Q^d} &= \int_\Omega \mathbf{K}^{-1}(R(\mathbf{F}_d) - \mathbf{F})\,(R(\mathbf{G}) - \mathbf{H})\,\mathrm{d}V + \int_\Omega \mathbf{K}^{-1}(R(\mathbf{F}_d) - \mathbf{F})\,\mathbf{H}\,\mathrm{d}V\\ &= J_3 + \int_\Omega (R(\mathbf{F}_d) - \mathbf{F})\,\nabla\psi\,\mathrm{d}V = J_3 - \int_\Omega \psi \operatorname{div}(R(\mathbf{F}_d) - \mathbf{F})\,\mathrm{d}V\\ &= J_3 - \int_\Omega (\mathbf{b}^I - b)\psi\,\mathrm{d}V\\ &= J_3 - \int_\Omega (\mathbf{b}^I - b)(\psi - \boldsymbol{\psi}^I)\,\mathrm{d}V = J_3 + J_4. \qquad (5.37)\end{aligned}$$

In their turn, J_3 and J_4 can be easily bounded using the previous estimates and the usual arguments. Indeed, the triangle inequality, then (3.10) and (5.8), and finally (5.23) and (5.34) imply that

$$\begin{aligned}\|R(\mathbf{F}_d) - \mathbf{F}\|_{(L^2(\Omega))^3} &\le \|R(\mathbf{F}_d - \mathbf{F}^I)\|_{(L^2(\Omega))^3} + \|R(\mathbf{F}^I) - \mathbf{F}\|_{(L^2(\Omega))^3}\\ &\le C^*_R s_*^{-1/2}|||\mathbf{F}_d - \mathbf{F}^I|||_{X^d} + \|R(\mathbf{F}^I) - \mathbf{F}\|_{(L^2(\Omega))^3}\\ &\le C\, h\, \|p\|_{H^2(\Omega)}. \qquad (5.38)\end{aligned}$$

Using assumption (5.34) and (5.24), we get

$$\|R(\mathbf{G}) - \mathbf{H}\|_{(L^2(\Omega))^3} = \|R(\mathbf{H}^I) - \mathbf{H}\|_{(L^2(\Omega))^3} \le C^*_{Ra} h \|\mathbf{H}\|_{(H^1(\Omega))^3} \le Ch|||\mathbf{p}_d - \mathbf{p}^I|||_{Q^d}. \qquad (5.39)$$

The approximation property (5.15) gives the following estimates:

$$|||\mathbf{b}^I - b\|_{L^2(\Omega)} \le C^*_{app}\, h\|b\|_{H^1(\Omega)} \qquad (5.40)$$

and

$$\|\psi - \boldsymbol{\psi}^I\|_{L^2(\Omega)} \le C^*_{app}\, h\|\psi\|_{H^1(\Omega)} \le C^*_{app} C^*_\Omega\, h\, |||\mathbf{p}_d - \mathbf{p}^I|||_{Q^d}. \qquad (5.41)$$

Inserting estimates (5.38)–(5.41) into (5.37), we immediately get the result. □

Remark 5.1. *It is very likely that our additional assumption* (5.34) *is not needed, as it should be possible to deduce it from* (5.4), (5.5), *possibly with minor additional assumptions on the geometry. However, in essentially all cases in which R_E can be explicitly built, it is easy to prove directly that* (5.34) *holds true. Therefore, we decided that it would be simpler to just assume it.*

6. Conclusion. In this paper, we have considered the MFD method for the mixed formulation of the diffusion problem on polyhedral meshes. We have proved the stability of the mimetic discretizations and the optimal convergence rates for the scalar and vector variables. The key elements of our methodology are the consistency Assumption S2 and the lift property LP. In future work, we plan to extend the convergence results to polyhedral meshes with curved faces.

Appendix. Proof of the list property. The purpose of this appendix is to prove the lift property (4.17)–(4.18), which we recall for convenience of the reader.

LP (lift property). For every $t < 2$ there exists a constant $\lambda^* = \lambda^*(t)$ such that for every $E \in \mathcal{T}_h$ and for every $e \in \partial E$ there exists a function φ_E^e from E to $\mathbb{R}$ that verifies

$$\varphi_E^e = 1 \quad \text{on} \quad e, \quad \varphi_E^e = 0 \quad \text{on} \quad \partial E \backslash e \tag{A.1}$$

and

$$\|\varphi_E^e\|_{L^2(E)} \leq \lambda^* h_E^{3/2}, \quad \|\nabla \varphi_E^e\|_{(L^t(E))^3} \leq \lambda^* h_E^{3/t-1}. \tag{A.2}$$

A traditional way would be to assume that there exist a finite number of reference elements $\hat{E}_1, \ldots, \hat{E}_1$ and a positive constant L^* such that for each $E \in \mathcal{T}_h$ there is an $\hat{E}_k$ and a bi-Lipschitz map Φ_k^E from $\hat{E}_k$ to E such that

$$|\Phi_k^E|_{W^1_\infty(\hat{E}_k)} \leq L^*, \qquad \|\Phi_k^E\|_{L_\infty(\hat{E}_k)} \leq L^*\, h_E \tag{A.3}$$

and

$$|(\Phi_k^E)^{-1}|_{W^1_\infty(E)} \leq L^*, \qquad \|(\Phi_k^E)^{-1}\|_{L_\infty(E)} \leq L^*\, h_E^{-1}. \tag{A.4}$$

Then, for each reference element $\hat{E}_k$ and for each face $\hat{e}$ of $\hat{E}_k$ we could construct the harmonic function $\hat{\varphi}_{\hat{E}_k}^{\hat{e}}$ with boundary value 1 on $\hat{e}$ and zero on the other faces, and verify that it belongs to $W_t^1(\hat{E}_k)$ for every $t < 2$. Finally each function φ_E^e could be constructed by combining one of the reference functions $\hat{\varphi}_{\hat{E}_k}^{\hat{e}}$ with the corresponding map Φ_k^E. This is surely feasible but will become rather cumbersome if we want to consider a big variety of possible shapes for our elements.

We decided here to follow a different path that requires only the fact that the faces are star shaped (M4) and the pyramid property (M5), which are possibly more difficult to explain but much easier to check and to enforce. The general idea is first to build a function $\hat{\varphi}_1$ on the unit cone $\mathcal{C}_1$; then, for every h, to build a function φ_h on a cone $\mathcal{C}_h$ obtained by scaling the unit cone; and finally, for each element E and for each face e, to map the cone $\mathcal{C}_{\gamma_* h_E}$ (where γ_* is given in Assumption M4) into the pyramid P_E^e described in Assumption M5 with a Lipschitz continuous mapping. This will give us a function $\varphi = \varphi_E^e$ on the pyramid, having the right norms. This function will finally be extended by zero to the whole element E, and still it will have the right norms. But let us look at the procedure in more detail.

For each element E and for each face e of E we want to build a function $\varphi = \varphi_E^e$ with the following properties.

- The support of φ is contained in the pyramid $P = P_E^e$ satisfying Assumption M5.
- $\varphi \equiv 1$ on e and $\varphi \equiv 0$ on the other faces of P_E^e.

- φ satisfies the following estimates:

$$\text{(A.5)} \qquad ||\varphi||_{L^2(P)} \leq \lambda^* \, h_E^{3/2} \quad \text{and} \quad ||\nabla\varphi||_{(L^t(P))^3} \leq \lambda^* \, h_E^{3/t-1},$$

where the constant λ^* is independent of E and e.

As we said before, we start our work on cones: for $\rho > 0$ we shall refer to the solid

$$\mathcal{C}_\rho \equiv \{(x, y, z)\colon \; 0 \leq z \leq \rho \quad \text{and} \quad x^2 + y^2 \leq (\rho - z)^2\}$$

as the *circular cone of radius* ρ.

LEMMA A.1. *Let $\mathcal{C}_1$ be the circular cone of radius* 1*, and let $\hat{\varphi}_1$ be the harmonic function that takes value* 1 *on the base and* 0 *on the lateral boundary. Then $\hat{\varphi}_1$ belongs to $L_\infty(\mathcal{C}_1)$ and $\nabla\hat{\varphi}_1$ belongs to $(L^t(\mathcal{C}_1))^3$ for all $t < 2$.*

Proof. The first part of the statement follows from the maximum principle, which gives $0 \leq \hat{\varphi}_1 \leq 1$. The second part of the statement follows immediately from the known results concerning domains with corners (see, e.g., [12] or [10]). □

In view of the previous lemma, we set

$$\text{(A.6)} \qquad \hat{C}_t := \|\nabla\varphi_1\|_{(L^t(\mathcal{C}_1))^3}.$$

It is clear that $\hat{C}_t$ depends on t and hence on s through (4.15).

LEMMA A.2. *For every positive real number h, let $\mathcal{C}_h$ be a circular cone of radius h. Then, there exists a function φ_h taking value* 1 *on the base, value zero on the lateral surface, and satisfying*

$$\text{(A.7)} \qquad ||\varphi_h||_{L^2(\mathcal{C}_h)} \leq |\mathcal{C}_h|^{1/2} \quad \text{and} \quad ||\nabla\varphi_h||_{(L^t(\mathcal{C}_h))^3} \leq h^{3/t-1}\hat{C}_t,$$

where $|\mathcal{C}_h|$ is the volume of $\mathcal{C}_h$.

Proof. The proof follows with the usual scaling arguments (see, e.g., [9, Theorem 3.1.2]). □

Consider now a face e of E. For convenience, we assume that (a) the face e lies in the plane $z = 0$, (b) M_e, defined in Assumption M4 (*star-shaped faces*), is the origin of the axes, and (c) the polyhedron E is locally in the half-space $z > 0$. By Assumptions M4 and M5 (*the pyramid property*), there exists a γ_* such that the circular cone $\mathcal{C}_h$ having its base on the face e (with center in M_e), and radius $h = \rho_* = \gamma_* \, h_E$, is strictly contained in the pyramid P_E^e having the same vertex and base equal to e. Hence, $\mathcal{C}_h$ is contained in E.

Let us show that Assumption M4 implies the existence of a radial mapping in the plane $z = 0$ which maps the disk D_{ρ_*} with center in M_e and radius ρ_* onto the face e, is one-to-one, Lipschitz continuous together with its inverse, and with W^1_∞ norms bounded in terms of γ_* and the number of edges of e.

LEMMA A.3. *Under Assumption* M4 *there exists a map Φ_2, mapping the disk D_{ρ_*} onto the face e, which is Lipschitz continuous together with the inverse map Φ_2^{-1}. Moreover,*

$$\text{(A.8)} \qquad \|\Phi_2\|_{W^1_\infty(D_{\rho_*})} \leq C_e^* \qquad \text{and} \qquad \|\Phi_2^{-1}\|_{W^1_\infty(e)} \leq C_e^*,$$

where C_e^* depends only on the constant γ_* from Assumption M4.

Proof. To show this, we note that the plane $z = 0$ can be split in a finite number of sectors by the vertices of e. Each sector corresponds to the straight rays coming

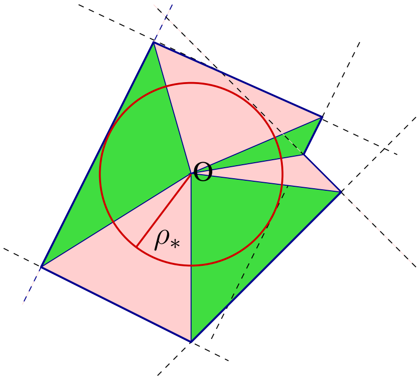


FIG. A.1. *The splitting of e in sectors.*

out of the origin M_e and intersecting the edge ℓ_k (see Figure A.1). For each point $\mathbf{P} \in D_{\rho_*}$, we first consider the ray emanating from the origin and passing through $\mathbf{P}$. This ray intersects ∂e at a point $\mathbf{V}(\mathbf{P})$. Our mapping is defined as follows:

$$\tilde{\mathbf{P}} \equiv \Phi_2(\mathbf{P}) := \frac{|\mathbf{V}(\mathbf{P})|}{\rho_*}\,\mathbf{P}. \tag{A.9}$$

It is clear that Φ_2 maps every point $\mathbf{P}$ onto a point $\tilde{\mathbf{P}}$ on the same ray so that

$$\mathbf{V}(\mathbf{P}) = \mathbf{V}(\tilde{\mathbf{P}}) \quad \forall\, \mathbf{P} \in D_{\rho^*}. \tag{A.10}$$

It is immediate to check that, on each ray, the map is continuous and monotone, and that it maps the points of the circumference of radius ρ_* onto the corresponding points of ∂e on the same ray. Hence it maps D_{ρ_*} onto e in a one-to-one way. It is also clear that the map is globally continuous, invertible, and the inverse map

$$\mathbf{P} \equiv \Phi_2^{-1}(\tilde{\mathbf{P}}) := \frac{\rho_*}{|\mathbf{V}(\mathbf{P})|}\,\tilde{\mathbf{P}} \equiv \frac{\rho_*}{|\mathbf{V}(\tilde{\mathbf{P}})|}\,\tilde{\mathbf{P}} \tag{A.11}$$

is also continuous and maps e onto D_{ρ_*}. Note that we used (A.10) in the last step.

In order to show the Lipschitz continuity, we have to bound the distance between the images $|\tilde{\mathbf{P}} - \tilde{\mathbf{Q}}|$ by a constant time the distance $|\mathbf{P} - \mathbf{Q}|$. For this, we note that Assumption M4 implies

$$1 \le \frac{|\mathbf{V}|}{\rho_*} \le \frac{h_E}{\gamma_*\, h_E} = \frac{1}{\gamma_*} \qquad \text{for every } \mathbf{V} \in \partial e. \tag{A.12}$$

As shown in Figure A.2, it also implies that for every point $\mathbf{V}$ on an edge ℓ of ∂e, the angle α_V between ℓ and the ray passing through $\mathbf{V}$ verifies

$$|\sin \alpha_V| = \frac{|\mathbf{H}_\ell|}{|\mathbf{V}|} \ge \frac{\rho_*}{|\mathbf{V}|} \ge \gamma_*, \tag{A.13}$$

where $\mathbf{H}_\ell$ is the orthogonal projection of the origin M_e on the line containing ℓ, and we used (A.12) in the last step.

The Lipschitz continuity is obvious when $\mathbf{P}$ and $\mathbf{Q}$ are on the same ray:

$$|\tilde{\mathbf{P}} - \tilde{\mathbf{Q}}| = \frac{|\mathbf{V}(\mathbf{P})|}{\rho_*}\,|\mathbf{P} - \mathbf{Q}| \le \frac{1}{\gamma_*}\,|\mathbf{P} - \mathbf{Q}|. \tag{A.14}$$

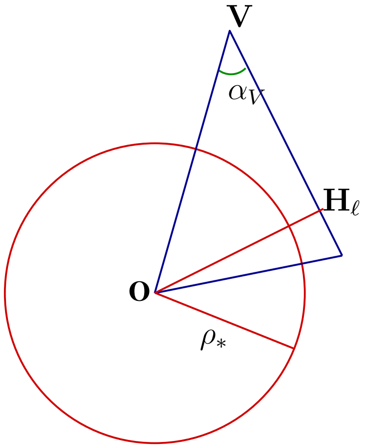


FIG. A.2. *Lower bound on* $|\sin \alpha_V|$.

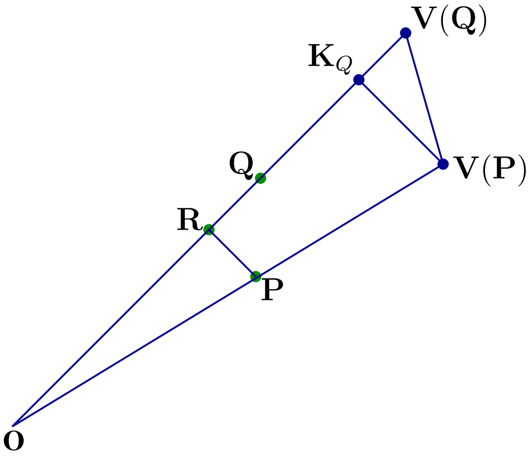


FIG. A.3. *Lipschitz continuity within a sector.*

If $\mathbf{P}$ and $\mathbf{Q}$ are on two different rays in the same sector, we first denote by $\mathbf{K}_Q$ and $\mathbf{R}$ (respectively) the orthogonal projections of $\mathbf{V}(\mathbf{P})$ (respectively, of $\mathbf{P}$) on the ray containing $\mathbf{Q}$ (see Figure A.3). Then, applying the Thaletes theorem, we get

$$\frac{|\mathbf{V}(\mathbf{P})-\mathbf{K}_Q|}{|\mathbf{V}(\mathbf{P})|}=\frac{|\mathbf{P}-\mathbf{R}|}{|\mathbf{P}|}\leq\frac{|\mathbf{P}-\mathbf{Q}|}{|\mathbf{P}|}. \tag{A.15}$$

Collecting (A.15), (A.13), and (A.12), we have

$$|\mathbf{V}(\mathbf{P})-\mathbf{V}(\mathbf{Q})|=\frac{|\mathbf{V}(\mathbf{P})-\mathbf{K}_Q|}{\left|\sin\left(\alpha_{V(Q)}\right)\right|}\leq\frac{|\mathbf{P}-\mathbf{Q}|}{\gamma_*|\mathbf{P}|}\,|\mathbf{V}(\mathbf{P})|\leq\frac{|\mathbf{P}-\mathbf{Q}|}{(\gamma_*)^2|\mathbf{P}|}\,\rho_*, \tag{A.16}$$

where obviously the role of **P** and **Q** can be interchanged. Finally, the triangle inequality together with (A.9) and (A.16) gives

$$\tag{A.17}\begin{aligned}|\tilde{\mathbf{P}}-\tilde{\mathbf{Q}}| &= \left|\frac{|\mathbf{V}(\mathbf{P})|\,\mathbf{P}-|\mathbf{V}(\mathbf{Q})|\,\mathbf{Q}}{\rho_*}\right| \le \frac{|\mathbf{V}(\mathbf{P})-\mathbf{V}(\mathbf{Q})|}{\rho_*}\,|\mathbf{P}| + \frac{|\mathbf{V}(\mathbf{Q})|}{\rho_*}\,|\mathbf{P}-\mathbf{Q}| \\ &\le \frac{|\mathbf{P}-\mathbf{Q}|}{(\gamma_*)^2} + \frac{1}{\gamma_*}|\mathbf{P}-\mathbf{Q}| = \frac{1+\gamma_*}{(\gamma_*)^2}|\mathbf{P}-\mathbf{Q}|.\end{aligned}$$

The case of **P** and **Q** belonging to different sectors can be easily deduced by inserting suitable intermediate points at the boundaries of the sectors and then using the triangle inequality.

In a similar way, we can show that the inverse mapping is also Lipschitz continuous. For instance, using (A.11) we get

$$\tag{A.18} |\mathbf{P}-\mathbf{Q}| = \left|\frac{\rho_*}{|\mathbf{V}(\tilde{\mathbf{P}})|}\tilde{\mathbf{P}} - \frac{\rho_*}{|\mathbf{V}(\tilde{\mathbf{Q}})|}\tilde{\mathbf{Q}}\right| = \frac{\rho_*}{|\mathbf{V}(\tilde{\mathbf{P}})|\,|\mathbf{V}(\tilde{\mathbf{Q}})|}\,||\mathbf{V}(\tilde{\mathbf{Q}})|\tilde{\mathbf{P}} - |\mathbf{V}(\tilde{\mathbf{P}})|\tilde{\mathbf{Q}}|.$$

Then, adding and subtracting $|\mathbf{V}(\mathbf{P})|\mathbf{P}$ and using the triangle inequality, we have

$$\tag{A.19} ||\mathbf{V}(\tilde{\mathbf{Q}})|\tilde{\mathbf{P}} - |\mathbf{V}(\tilde{\mathbf{P}})|\tilde{\mathbf{Q}}| \le |\mathbf{V}(\tilde{\mathbf{P}}) - \mathbf{V}(\tilde{\mathbf{Q}})|\,|\tilde{\mathbf{P}}| + |\mathbf{V}(\tilde{\mathbf{P}})|\,|\tilde{\mathbf{P}}-\tilde{\mathbf{Q}}|.$$

On the other hand, we can apply the argument of (A.16) to obtain

$$\tag{A.20} |\mathbf{V}(\tilde{\mathbf{P}}) - \mathbf{V}(\tilde{\mathbf{Q}})| \le \frac{|\tilde{\mathbf{P}}-\tilde{\mathbf{Q}}|}{(\gamma_*)^2|\tilde{\mathbf{P}}|}\,\rho_*.$$

Collecting (A.18), (A.19), and (A.20), and using (A.12) (this time as $\rho_*/|\mathbf{V}| \le 1$), we finally obtain

$$\tag{A.21} |\mathbf{P}-\mathbf{Q}| \le \frac{1}{(\gamma_*)^2}|\tilde{\mathbf{P}}-\tilde{\mathbf{Q}}| + |\tilde{\mathbf{P}}-\tilde{\mathbf{Q}}| = \frac{1+(\gamma_*)^2}{(\gamma_*)^2}\,|\tilde{\mathbf{P}}-\tilde{\mathbf{Q}}|.$$

This proves the assertion of the lemma. □

Now, we can construct a mapping Φ_3 from the cone $\mathcal{C}_h$ (having D_{ρ_*} as the base and with height equal to ρ_*) onto the pyramid P_E^e (having e as the base and with the same vertex as $\mathcal{C}_h$), which is Lipschitz continuous with its inverse, by taking

$$\tag{A.22} (\tilde{x},\, \tilde{y}) = \Phi_2(x,\, y), \quad \tilde{z} = z.$$

Again, the Lipschitz norms of the map Φ_3 and of its inverse depend only on γ_*. This proves the following lemma.

LEMMA A.4. *Under Assumption* M4 *there exists a map Φ_3, mapping the cone $\mathcal{C}_h$ onto the pyramid P_E^e, which is Lipschitz continuous together with the inverse map Φ_3^{-1}. Moreover,*

$$\tag{A.23} \|\Phi_3\|_{W^1_\infty(\mathcal{C}_h)} \le C^*_{pyr} \quad \textit{and} \quad \left\|\Phi_3^{-1}\right\|_{W^1_\infty(P_E^e)} \le C^*_{pyr},$$

where C^*_{pyr} depends only on the constant γ_* of Assumption M4.

The last step is to construct, for each element E and for each face $e \in \partial E$, the function φ_E^e satisfying (A.5) (with the right boundary conditions). Let

$$\varphi_E^e(x,\, y,\, z) = \varphi_h\big(\Phi_3^{-1}(x,\, y,\, z)\big),$$

where φ_h is the function from Lemma A.2 defined for the circular cone of radius $h = \rho_* = \gamma_* \, h_E$. It is clear that φ_E^e will be in $L^2(P_E^e)$, that $\nabla\varphi_E^e$ will be in $(L^t(P_E^e))^3$, and that their norms will be bounded by

$$\|\varphi_E^e\|_{L^2(P_E^e)} \leq C^*_{pyr} \, h_E^{3/2} \quad \text{and} \quad \|\nabla\varphi_E^e\|_{(L^t(P_E^e))^3} \leq \hat{C}_t \, C^*_{pyr} h_E^{3/t-1}, \tag{A.24}$$

where $\hat{C}_t$ is given in (A.6) and C^*_{pyr} depends only on γ_*. Hence φ_E^e satisfies (A.5) as required. Finally, we take the prolongation of φ_E^e (that we call again φ_E^e) by zero in $E \backslash P_E^e$.

This ends the proof of the lift property (A.1)–(A.2).

Acknowledgments. The authors thank organizers of the workshop "Compatible Spatial Discretizations for Partial Differential Equations" held at the Institute of Mathematics and Applications (IMA) in Minneapolis for drawing attention to the problem considered in this paper (the collaboration was actually started during the workshop).

REFERENCES

[1] D. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] M. Berndt, K. Lipnikov, J. D. Moulton, and M. Shashkov, *Convergence of mimetic finite difference discretizations of the diffusion equation*, East-West J. Numer. Math., 9 (2001), pp. 253–284.

[3] M. Berndt, K. Lipnikov, M. Shashkov, M. Wheeler, and I. Yotov, *Superconvergence of the velocity in mimetic finite difference methods on quadrilaterals*, SIAM J. Numer. Anal., 43 (2005), pp. 1728–1749.

[4] S. Brenner and L. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1994.

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

[6] F. Brezzi, K. Lipnikov, and V. Simoncini, *A family of mimetic finite difference methods on polygonal and polyhedral meshes*, Math. Models Methods Appl. Sci., 15 (2005), pp. 1533–1551.

[7] D. Burton, *Multidimensional Discretization of Conservation Laws for Unstructured Polyhedral Grids*, Technical report UCRL-JC-118306, Lawrence Livermore National Laboratory, Livermore, CA, 1994.

[8] J. Campbell and M. Shashkov, *A tensor artificial viscosity using a mimetic finite difference algorithm*, J. Comput. Phys., 172 (2001), pp. 739–765.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, New York, 1978.

[10] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains: Smoothness and Asymptotics of Solutions*, Springer-Verlag, Berlin, 1988.

[11] J. Douglas and J. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.

[12] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.

[13] J. Hyman, J. Morel, M. Shashkov, and S. Steinberg, *Mimetic finite difference methods for diffusion equations*, Comput. Geosci., 6 (2002), pp. 333–352.

[14] J. Hyman and M. Shashkov, *Mimetic discretizations for Maxwell's equations and the equations of magnetic diffusion*, Progr. Electromagn. Res., 32 (2001), pp. 89–121.

[15] J. Hyman, M. Shashkov, and S. Steinberg, *The numerical solution of diffusion problems in strongly heterogeneous non-isotropic materials*, J. Comput. Phys., 132 (1997), pp. 130–148.

[16] Y. Kuznetsov, K. Lipnikov, and M. Shashkov, *Mimetic finite difference method on polygonal meshes for diffusion-type problems*, Comput. Geosci., 8 (2004), pp. 301–324.

[17] K. Lipnikov, J. Morel, and M. Shashkov, *Mimetic finite difference methods for diffusion equations on non-orthogonal non-conformal meshes*, J. Comput. Phys., 199 (2004), pp. 589–597.

[18] R. Liska, M. Shashkov, and V. Ganza, *Analysis and optimization of inner products for mimetic finite difference methods on triangular grid*, Math. Comput. Simulation, 67 (2004), pp. 55–66.

[19] L. Margolin, M. Shashkov, and P. Smolarkiewicz, *A discrete operator calculus for finite difference approximations*, Comput. Methods Appl. Mech. Engrg., 187 (2000), pp. 365–383.

[20] J. Morel, R. Roberts, and M. Shashkov, *A local support-operators diffusion discretization scheme for quadrilateral r–z meshes*, J. Comput. Phys., 144 (1998), pp. 17–51.

[21] T. Palmer, M. Zika, and N. Madsen, *Unstructured polyhedral mesh thermal radiation diffusion*, Trans. Amer. Nuclear Soc., 83 (2000), pp. 248–249.

[22] K. Thompson and M. Adams, *A Spatial Discretization for Solving the Transport Equation on Unstructured Grids of Polyhedra*, Report LAUR-99-2121, Los Alamos National Laboratory, Los Alamos, NM, 1999.

SIAM J. Numer. Anal.
Vol. 43, No. 5, pp. 1897–1911

PIECEWISE POLYNOMIAL COLLOCATION FOR FREDHOLM INTEGRO-DIFFERENTIAL EQUATIONS WITH WEAKLY SINGULAR KERNELS[*]

INGA PARTS[†], ARVET PEDAS[†], AND ENN TAMME[†]

Abstract. In the first part of this paper we study the regularity properties of solutions of initial- or boundary-value problems of linear Fredholm integro-differential equations with weakly singular or other nonsmooth kernels. We then use these results in the analysis of a piecewise polynomial collocation method for solving such problems numerically. The main purpose of the paper is the derivation of optimal global convergence estimates and the analysis of the attainable order of convergence of numerical solutions for all values of the nonuniformity parameter of the underlying grid.

Key words. Fredholm integro-differential equation, weakly singular kernel, piecewise polynomial collocation method, graded grid, attainable order of convergence

AMS subject classifications. 65R20, 45J05

DOI. 10.1137/040612452

1. Introduction.

We present a study of the convergence behavior of a collocation method for the numerical solution of initial- or boundary-value problems of linear integro-differential equations of the form

$$\begin{aligned} &u'(t) = a(t)u(t) + f(t) + \int_0^b K(t,s)u(s)ds, \quad 0 \le t \le b, \\ &\alpha u(0) + \beta u(b) = \gamma, \end{aligned} \tag{1.1}$$

where $b, \alpha, \beta, \gamma \in \mathbf{R} = (-\infty, \infty)$, $b > 0$, and $\alpha + \beta \neq 0$. We assume that $a, f \in C^{m,\nu}[0,b]$, $K \in W^{m,\nu}(\Delta)$, $m \in \mathbf{N} = \{1, 2, \dots\}$, $\nu \in \mathbf{R}$, $\nu < 1$.

Here $C^{m,\nu}[0,b]$, $m \in \mathbf{N}$, $\nu < 1$, is defined as the collection of all continuous functions $u : [0,b] \to \mathbf{R}$, which are m times continuously differentiable in $(0,b)$ and such that the estimation

$$\left|u^{(i)}(t)\right| \le c \begin{cases} 1 & \text{if} \quad i < 1 - \nu, \\ 1 + |\log \varrho(t)| & \text{if} \quad i = 1 - \nu, \\ \varrho(t)^{1-\nu-i} & \text{if} \quad i > 1 - \nu \end{cases}$$

holds with $\varrho(t) = \min\{t, b - t\}$, $0 < t < b$, and with a constant $c = c(u)$ for all $t \in (0,b)$ and $i = 1, \dots, m$. Equipped with the norm

$$\|u\|_{m,\nu} = \max_{0 \le t \le b} |u(t)| + \sum_{i=1}^{m} \sup_{0<t<b} \left(w_{i+\nu-1}(t)\left|u^{(i)}(t)\right|\right), \quad u \in C^{m,\nu}[0,b],$$

$C^{m,\nu}[0,b]$ is a Banach space. Here

$$w_\lambda(t) = \begin{cases} 1 & \text{for} \quad \lambda < 0, \\ (1 + |\log \varrho(t)|)^{-1} & \text{for} \quad \lambda = 0, \\ \varrho(t)^\lambda & \text{for} \quad \lambda > 0, \end{cases}$$

[*] Received by the editors July 28, 2004; accepted for publication (in revised form) April 22, 2005; published electronically December 8, 2005. This work was supported by Estonian Science Foundation grant 5859.

http://www.siam.org/journals/sinum/43-5/61245.html

[†] Institute of Applied Mathematics, University of Tartu, Liivi 2, 50409 Tartu, Estonia (Inga.Parts@ut.ee, Arvet.Pedas@ut.ee, Enn.Tamme@ut.ee).

with $t \in (0,b)$. It is easy to see that if $\mu < \nu < 1$, then $C^{m,\mu}[0,b] \subset C^{m,\nu}[0,b]$ and $\|u\|_{m,\nu} \le c\|u\|_{m,\mu}$ for $u \in C^{m,\mu}[0,b]$, with a constant $c > 0$. Notice also that[1] $C^m[0,b] \subset C^{m,\nu}[0,b]$, $m \in \mathbf{N}$, $\nu < 1$.

The set $W^{m,\nu}(\Delta)$, with $m \in \mathbf{N}$, $\nu < 1$,

$$\Delta = \{(t,s) : \ 0 \le t \le b, \ 0 \le s \le b, \ t \ne s\},$$

consists of all m times continuously differentiable functions $K : \Delta \to \mathbf{R}$ satisfying

$$\left|\left(\frac{\partial}{\partial t}\right)^i \left(\frac{\partial}{\partial t} + \frac{\partial}{\partial s}\right)^j K(t,s)\right| \le c \begin{cases} 1 & \text{if} \quad \nu + i < 0, \\ 1 + |\log|t-s|| & \text{if} \quad \nu + i = 0, \\ |t-s|^{-\nu-i} & \text{if} \quad \nu + i > 0, \end{cases} \tag{1.2}$$

with a constant $c = c(K)$ for all $(t,s) \in \Delta$ and all nonnegative integers i and j such that $i + j \le m$.

Taking $i = j = 0$, the condition (1.2) yields that $K \in W^{m,\nu}(\Delta)$ may possess a weak singularity at $t = s$ for $0 \le \nu < 1$. If $\nu < 0$, then K itself is bounded on Δ, but its derivatives may be singular at $t = s$. Often the kernel K of problem (1.1) has the form $K = K_\nu(t,s) = \kappa(t,s)|t-s|^{-\nu}$, $0 < \nu < 1$, or $K = K_0(t,s) = \kappa(t,s)\log|t-s|$, where $\kappa \in C^m(\bar{\Delta})$, with $m \in \mathbf{N}$ and $\bar{\Delta} = [0,b] \times [0,b]$. Clearly, $K_\nu \in W^{m,\nu}(\Delta)$ for $0 < \nu < 1$ and $K_0 \in W^{m,0}(\Delta)$.

Note that Fredholm integro-differential equations are used by modeling various physical processes; see, e.g., [9]. A good source of information (including numerous additional references) on applications of integral and integro-differential equations is the monograph [4].

A special case of problem (1.1), with $\alpha = 1$, $\beta = 0$, and $K(t,s) = 0$ for $s > t$, is the initial-value problem for a Volterra integro-differential equation. Volterra integro-differential equations have been studied by many authors (see, e.g., [3, 4, 5, 6, 7, 12, 15, 18, 19, 20, 21]), but Fredholm-type equations have received less attention. There is some literature on the numerical solution of Fredholm integro-differential equations in case of smooth kernels; see, e.g., [11, 13, 16, 26]. To the authors' knowledge very little has been written on the numerical solution of Fredholm integro-differential equations with weakly singular kernels [25] (in contrast to weakly singular Fredholm integral equations; see, for example, [22, 14] and, especially, [8]). In order to fill this gap, the main purpose of the present paper is to generalize the corresponding results obtained in [6, 7, 15] for weakly singular Volterra integro-differential equations to a wide class of Fredholm integro-differential equations.

In the first part of this paper (section 2) we study the regularity properties of the solution of problem (1.1) with weakly singular or other nonsmooth kernels K. Moreover, we consider the case where the derivatives of the functions a and f in (1.1) may be unbounded on the interval $[0,b]$. We then use these results in the analysis of a piecewise polynomial collocation method for solving such equations numerically. Using special graded grids, we derive optimal global convergence estimates and analyze the attainable order of global and local convergence of numerical solutions for all values of the grading exponent of the underlying grid (sections 4 and 5). In section 3 we formulate some auxiliary results which we need in the analysis of proposed algorithms and in section 6 we present a numerical example to clarify the obtained theoretical results. The main results of the paper are formulated in Theorems 2.1, 4.1, and 5.1.

[1] By $C^k(\Omega)$ we denote the set of k times ($k \ge 0$) continuously differentiable functions on $\Omega \subset \mathbf{R}^n$, $C^0(\Omega) = C(\Omega)$; by c we denote positive constants, which may be different in different inequalities (in sections 3–5 they are independent of $N \in \mathbf{N}$).

Notice that similar results for integral equations may be found, for example, in [1, 2, 4, 5, 8, 10, 14, 17, 22, 24].

2. Smoothness of the solution. In what follows, for given Banach spaces E and F we denote by $\mathcal{L}(E, F)$ the Banach space of linear bounded operators $A : E \to F$ with the norm $\|A\| = \sup\{\|Au\|_F : u \in E, \|u\|_E \leq 1\}$. The regularity of the solution of problem (1.1) is described in the following theorem.

THEOREM 2.1. *Let $a, f \in C^{m,\nu}[0, b]$, $K \in W^{m,\nu}(\Delta)$, $m \in \mathbf{N}$, $\nu \in \mathbf{R}$, $\nu < 1$, $\alpha, \beta, \gamma \in \mathbf{R}$, $\alpha + \beta \neq 0$. Moreover, assume that the homogeneous problem*

$$u'(t) = a(t)u(t) + \int_0^b K(t, s)u(s)ds, \quad \alpha u(0) + \beta u(b) = 0, \tag{2.1}$$

corresponding to the problem (1.1)*, has in the set $\{u : u \in C[0, b], u' \in L^\infty(0, b)\}$ only the trivial solution $u = 0$.*

Then problem (1.1) *has a unique solution $u \in C^{m+1,\nu-1}[0, b] \subset C^{m,\nu}[0, b]$ and u', the derivative of the solution of* (1.1)*, belongs to $C^{m,\nu}[0, b]$.*

Proof. If $\alpha + \beta \neq 0$ and $v \in L^\infty(0, b)$, then the problem

$$u'(t) = v(t), \quad \alpha u(0) + \beta u(b) = \gamma$$

has a unique solution

$$u(t) = (Jv)(t) + \frac{\gamma}{\alpha + \beta}, \quad 0 \leq t \leq b, \tag{2.2}$$

where

$$(Jv)(t) = \int_0^t v(s)ds - \frac{\beta}{\alpha + \beta} \int_0^b v(s)ds = \int_0^b \kappa(t - s)v(s)ds, \quad 0 \leq t \leq b, \tag{2.3}$$

with

$$\kappa(\tau) = \begin{cases} -\dfrac{\beta}{\alpha + \beta} & \text{if} \quad -b \leq \tau < 0, \\[2ex] 1 - \dfrac{\beta}{\alpha + \beta} & \text{if} \quad 0 < \tau \leq b. \end{cases}$$

It follows from (2.3) and the expression of the norm $\|\cdot\|_{m,\nu}$ that $J \in \mathcal{L}(C^{m,\nu}[0, b], C^{m+1,\nu-1}[0, b])$. Since $\kappa^{(i)}(\tau) = 0$ for $\tau \neq 0$ and $i = 1, 2, \ldots$, then (see [23, 24]) J is compact as an operator from $C^{m,\nu}[0, b]$ to $C^{m,\nu}[0, b]$. Further, we can write (1.1) in the form

$$u'(t) = (Au)(t) + (Tu)(t) + f(t), \quad 0 \leq t \leq b, \quad \alpha u(0) + \beta u(b) = \gamma,$$

where

$$(Au)(t) = a(t)u(t), \quad (Tu)(t) = \int_0^b K(t, s)u(s)ds, \quad 0 \leq t \leq b. \tag{2.4}$$

Therefore, if u is a solution of problem (1.1), then it can be presented in the form (2.2), where v is the solution of equation

$$v = T_1 v + f_1, \tag{2.5}$$

with $T_1 = (A+T)J$ (see (2.3)–(2.4)) and

$$f_1(t) = f(t) + \frac{\gamma}{\alpha+\beta}a(t) + \frac{\gamma}{\alpha+\beta}\int_0^b K(t,s)ds, \quad 0 \le t \le b. \tag{2.6}$$

Next we show that T_1 is compact as an operator from $C^{m,\nu}[0,b]$ into $C^{m,\nu}[0,b]$ and $f_1 \in C^{m,\nu}[0,b]$. Indeed, since $K \in W^{m,\nu}(\Delta)$, then (see [22]) $T \in \mathcal{L}(C^{m,\nu}[0,b], C^{m,\nu}[0,b])$. If $a, v \in C^{m,\nu}[0,b]$, then (cf. [6])

$$\|av\|_{m,\nu} \le c\|a\|_{m,\nu}\|v\|_{m,\nu}.$$

Therefore, $A \in \mathcal{L}(C^{m,\nu}[0,b], C^{m,\nu}[0,b])$. This together with the compactness of $J \in \mathcal{L}(C^{m,\nu}[0,b], C^{m,\nu}[0,b])$ implies that $T_1 : C^{m,\nu}[0,b] \to C^{m,\nu}[0,b]$ is linear and compact. Since $1 \in C^{m,\nu}[0,b]$, then $T1 \in C^{m,\nu}[0,b]$. This together with $f, a \in C^{m,\nu}[0,b]$ yields $f_1 \in C^{m,\nu}[0,b]$.

Further, since problem (2.1) has only the trivial solution, then equation $v = T_1 v$ has only the trivial solution $v = 0$ in $L^\infty(0,b)$ and therefore also in $C^{m,\nu}[0,b] \subset L^\infty(0,b)$. Thus, by the Fredholm alternative, $I - T_1$ has a bounded inverse $(I-T_1)^{-1} : C^{m,\nu}[0,b] \to C^{m,\nu}[0,b]$ (here I is the identity mapping), and equation (2.5) has a unique solution $v = (I-T_1)^{-1}f_1 \in C^{m,\nu}[0,b]$. This, in turn, implies that problem (1.1) has a unique solution u and

$$u = Jv + \frac{\gamma}{\alpha+\beta} \in C^{m+1,\nu-1}[0,b]. \qquad \square$$

Remark 2.1. In [6, 7] it is shown that an initial-value problem for a linear Volterra integro-differential equation in the form

$$u'(t) = a(t)u(t) + f(t) + \int_0^t K(t,s)u(s)ds, \quad 0 \le t \le b, \quad u(0) = \gamma,$$

has a unique solution and thus the corresponding homogeneous problem cannot have nontrivial solutions. Such a situation does not take place for Fredholm integro-differential equations. For example, the homogeneous problem

$$u'(t) = 2\int_0^1 u(s)ds, \quad t \in [0,b], \ u(0) = 0,$$

has the nontrivial solution $u(t) = ct$, $c \ne 0$.

Remark 2.2. In Theorem 2.1, we have assumed that $\alpha+\beta \ne 0$. Actually, Theorem 2.1 holds also in the case $\alpha = -\beta \ne 0$, for example, in the case of a periodic boundary condition $u(0) = u(b)$. In order to prove this we can use the circumstance that for every $v \in L^\infty(0,b)$ and $\gamma \in \mathbf{R}$ a problem

$$u'(t) + u(t) = v(t), \quad u(0) - u(b) = \gamma$$

has a unique solution

$$u(t) = (J_1 v)(t) + \frac{\gamma e^{-t}}{1-e^{-b}},$$

where

$$(J_1 v)(t) = e^{-t}\int_0^t v(s)e^s ds + \frac{e^{-t-b}}{1-e^{-b}}\int_0^b v(s)e^s ds, \quad 0 \le t \le b.$$

It is easy to see that $J_1 \in \mathcal{L}(C^{m,\nu}[0,b], C^{m+1,\nu-1}[0,b])$ and J_1 is compact as an operator from $C^{m,\nu}[0,b]$ into $C^{m,\nu}[0,b]$. Using J_1 instead of J, on the basis of similar arguments as in the proof of Theorem 2.1, we can see that the statement of Theorem 2.1 holds also for $\alpha+\beta=0$, $\alpha=-\beta\neq 0$.

3. Piecewise polynomial interpolation. For $N \in \mathbf{N}$, $r \in \mathbf{R}$, $r \geq 1$, let $\Pi_N = \Pi_N^{(r)} = \{t_0, \ldots, t_{2N} : 0 = t_0 < t_1 < \cdots < t_{2N} = b\}$ be a partition (a graded grid) of the interval $[0,b]$ given by the grid points

$$t_j = \frac{b}{2}\left(\frac{j}{N}\right)^r, \quad j = 0, 1, \ldots, N, \qquad (3.1)$$

$$t_{N+j} = b - t_{N-j}, \quad j = 1, \ldots, N.$$

Here the real number $r \in [1,\infty)$ characterizes the nonuniformity of the grid Π_N: For $r > 1$ the points (3.1) are more densely clustered near the endpoints of the interval $[0,b]$. It is easy to see that $t_j - t_{j-1} \leq (rb/2)N^{-1}$, $j = 1, \ldots, 2N$.

For given integers $m \geq 0$ and $-1 \leq d \leq m-1$, let $S_m^{(d)}(\Pi_N)$ be the spline space of piecewise polynomial functions on the grid Π_N:

$$S_m^{(d)}(\Pi_N) = \left\{v \in C^d[0,b] : v\big|_{\sigma_j} \in \pi_m,\ j = 1, \ldots, 2N\right\},\ 0 \leq d \leq m-1,$$

$$S_m^{(-1)}(\Pi_N) = \left\{v : v\big|_{\sigma_j} \in \pi_m,\ j = 1, \ldots, 2N\right\}.$$

Here $v\big|_{\sigma_j}$ $(j = 1, \ldots, 2N)$ is the restriction of v onto the subinterval $\sigma_j = [t_{j-1}, t_j] \subset [0,b]$, and π_m denotes the set of polynomials of degree not exceeding m. Note that the elements of $S_m^{(-1)}(\Pi_N)$ may have jump discontinuities at the interior points $t_1, \ldots, t_{2N-1}$ of the grid Π_N.

In every interval $[t_{j-1}, t_j]$, $j = 1, \ldots, 2N$, we introduce $m \geq 1$ interpolation points:

$$t_{jk} = t_{j-1} + \eta_k(t_j - t_{j-1}), \quad k = 1, \ldots, m,\ j = 1, \ldots, 2N, \qquad (3.2)$$

where $\eta_1, \ldots, \eta_m$ are some fixed parameters which do not depend on j and N and satisfy the conditions

$$0 \leq \eta_1 < \cdots < \eta_m \leq 1. \qquad (3.3)$$

To a given continuous function $v : [0,b] \to \mathbf{R}$ we assign a piecewise polynomial interpolation function $\mathcal{P}_N v \in S_{m-1}^{(-1)}(\Pi_N)$, which interpolates v at the points (3.2):

$$(\mathcal{P}_N v)(t_{jk}) = v(t_{jk}), \quad k = 1, \ldots, m,\ j = 1, \ldots, 2N.$$

Thus, $(\mathcal{P}_N v)(t)$ is independently defined in every subinterval $[t_{j-1}, t_j]$, $j = 1, \ldots, 2N$, and may be discontinuous at the points $t = t_j$, $j = 1, \ldots, 2N-1$; we may treat $\mathcal{P}_N v$ as a two-valued function at these points. Note that in the case $\eta_1 = 0$, $\eta_m = 1$, $\mathcal{P}_N v$ is a continuous function on $[0,b]$.

We also introduce an interpolation operator $\mathcal{P}_N$ which assigns to every continuous function $v : [0,b] \to \mathbf{R}$ its piecewise polynomial interpolation function $\mathcal{P}_N v$.

From [22, pp. 115–119], we obtain Lemmas 3.1–3.3 (cf. also [6]).

LEMMA 3.1. *Let the interpolation nodes* (3.2) *with grid points* (3.1) *and parameters* (3.3) *be used. Then* $\mathcal{P}_N \in \mathcal{L}(C[0,b], L^\infty(0,b))$ *and* $\|\mathcal{P}_N\|_{\mathcal{L}(C[0,b],L^\infty(0,b))} \leq c, N \in \mathbf{N}$, *with a positive constant c which is independent of N.*

LEMMA 3.2. *Let $v \in C^{m,\nu}[0,b]$, $m \in \mathbf{N}$, $\nu \in \mathbf{R}$, $\nu < 1$, and let the interpolation nodes* (3.2) *with grid points* (3.1) *and parameters* (3.3) *be used. Then the following estimates hold:*

$$\big\|v - \mathcal{P}_N v\big\|_\infty \le c \begin{cases} N^{-r(1-\nu)} & \text{for} \quad 1 \le r < \dfrac{m}{1-\nu}, \\ N^{-m}(1+\log N) & \text{for} \quad r = \dfrac{m}{1-\nu} = 1, \\ N^{-m} & \text{for} \quad r = \dfrac{m}{1-\nu} > 1 \text{ or } r > \dfrac{m}{1-\nu}, r \ge 1, \end{cases}$$

$$\int_0^b \big|v(t) - (\mathcal{P}_N v)(t)\big| dt \le c \begin{cases} N^{-r(2-\nu)} & \text{for} \quad 1 \le r < \dfrac{m}{2-\nu}, \\ N^{-m}(1+\log N) & \text{for} \quad r = \dfrac{m}{2-\nu} \ge 1, \\ N^{-m} & \text{for} \quad r > \dfrac{m}{2-\nu}, r \ge 1. \end{cases}$$

Here c is a positive constant not depending on N and

$$\big\|v - \mathcal{P}_N v\big\|_\infty = \max_{1 \le j \le 2N} \sup_{t_{j-1} < t < t_j} \big|v(t) - (\mathcal{P}_N v)(t)\big|. \tag{3.4}$$

LEMMA 3.3. *Let the conditions of Lemma* 3.2 *be fulfilled. Then*

$$\sup_{t_{j-1} < s < t_j} |v(s) - (\mathcal{P}_N v)(s)| \le c\,(t_j - t_{j-1})^m \begin{cases} 1 & \text{if} \quad m < 1-\nu, \\ 1 + |\log t_j| & \text{if} \quad m = 1-\nu, \\ t_j^{1-\nu-m} & \text{if} \quad m > 1-\nu, \end{cases}$$

for $j = 1, \ldots, N$, and

$$\sup_{t_{j-1} < s < t_j} |v(s) - (\mathcal{P}_N v)(s)| \le c\,(t_j - t_{j-1})^m \begin{cases} 1 & \text{if} \quad m < 1-\nu, \\ 1 + |\log(b - t_{j-1})| & \text{if} \quad m = 1-\nu, \\ (b - t_{j-1})^{1-\nu-m} & \text{if} \quad m > 1-\nu, \end{cases}$$

for $j = N+1, \ldots, 2N$, with a positive constant c which is independent of j and N.

4. Collocation method. Problem (1.1) is equivalent to problem (2.2), (2.5). In order to solve problem (1.1) we construct a collocation method for the numerical solution of problem (2.2), (2.5).

We look for an approximate solution u_N of (1.1) in the form

$$u_N(t) = \int_0^t v_N(s)ds - \frac{\beta}{\alpha+\beta}\int_0^b v_N(s)ds + \frac{\gamma}{\alpha+\beta}, \quad t \in [0,b], \ N \in \mathbf{N}, \tag{4.1}$$

where v_N satisfies the following conditions:

(4.2)

$$\begin{aligned} & v_N \in S_{m-1}^{(-1)}(\Pi_N), \ m \in \mathbf{N}, \\ & v_N(t_{jk}) = a(t_{jk})\Big(\int_0^{t_{jk}} v_N(s)ds - \frac{\beta}{\alpha+\beta}\int_0^b v_N(s)ds\Big) \\ & \qquad + \int_0^b K(t_{jk},s)\Big(\int_0^s v_N(\tau)d\tau - \frac{\beta}{\alpha+\beta}\int_0^b v_N(\tau)d\tau\Big)ds + f_1(t_{jk}), \\ & \qquad k = 1, \ldots, m, \ j = 1, \ldots, 2N, \end{aligned}$$

with f_1 and $\{t_{jk}\}$ given by the formulas (2.6) and (3.2), respectively.

Remark 4.1. Since $v_N \in S_{m-1}^{(-1)}(\Pi_N)$, then u_N (see (4.1)) belongs to $S_m^{(0)}(\Pi_N) \subset C[0,b]$. If $\eta_1 = 0$ and $\eta_m = 1$ (see (3.3)), then $v_N \in S_{m-1}^{(0)}(\Pi_N) \subset C[0,b]$ and $u_N \in S_m^{(1)}(\Pi_N) \subset C^1[0,b]$.

Remark 4.2. The collocation conditions (4.2) form a system of equations whose exact form is determined by the choice of a basis in $S_{m-1}^{(-1)}(\Pi_N)$ (or in $S_{m-1}^{(0)}(\Pi_N)$ if $\eta_1 = 0$ and $\eta_m = 1$). For instance, in each subinterval $[t_{j-1}, t_j] \subset [0,b]$, $j = 1, \dots, 2N$, we may use the Lagrange fundamental polynomial representation

$$v_N(t) = \sum_{k=1}^{m} c_{jk}\varphi_k\left(\frac{t - t_{j-1}}{t_j - t_{j-1}}\right), \quad t \in [t_{j-1}, t_j],\ j = 1, \dots, 2N,$$

where $c_{jk} = v_N(t_{jk})$,

$$\varphi_k(\tau) = \prod_{q=1,\ q\neq k}^{m} \left(\frac{\tau - \eta_q}{\eta_k - \eta_q}\right), \quad \tau \in [0,1],\ k = 1, \dots, m.$$

The conditions (4.2) then lead to a system of linear algebraic equations for the coefficients c_{jk}, $k = 1, \dots, m$, $j = 1, \dots, 2N$.

Remark 4.3. The conditions (4.2) have the operator equation representation

$$v_N = \mathcal{P}_N T_1 v_N + \mathcal{P}_N f_1, \quad T_1 = (A + T)J, \tag{4.3}$$

with A, T, J and $\mathcal{P}_N$, determined in sections 2 and 3, respectively.

THEOREM 4.1. *Let the conditions of Theorem* 2.1 *be fulfilled and let the interpolation nodes* (3.2) *with grid points* (3.1) *and parameters* (3.3) *be used.*

Then there exists an $N_0 \in \mathbf{N}$ such that, for $N \geq N_0$, the settings (4.2) *determine a unique approximation $v_N \in S_{m-1}^{(-1)}(\Pi_N)$ to $v = u'$, where u is the exact solution of problem* (1.1). *Moreover, if $N \geq N_0$, then an approximation u_N for u is defined by the formula* (4.1), *and the following error estimates hold:*

$$\|u - u_N\|_\infty \leq c \begin{cases} N^{-r(2-\nu)} & \textit{for} \quad 1 \leq r < \dfrac{m}{2-\nu}, \\ N^{-m}(1 + \log N) & \textit{for} \quad r = \dfrac{m}{2-\nu} \geq 1, \\ N^{-m} & \textit{for} \quad r > \dfrac{m}{2-\nu}, r \geq 1, \end{cases} \tag{4.4}$$

$$\|u' - v_N\|_\infty \leq c \begin{cases} N^{-r(1-\nu)} & \textit{for}\ \ 1 \leq r < \dfrac{m}{1-\nu}, \\ N^{-m}(1 + \log N) & \textit{for}\ \ r = \dfrac{m}{1-\nu} = 1, \\ N^{-m} & \textit{for}\ \ r = \dfrac{m}{1-\nu} > 1 \ \textit{or}\ r > \dfrac{m}{1-\nu},\ r \geq 1. \end{cases} \tag{4.5}$$

Here c is a positive constant not depending on N, and the norm $\|\cdot\|_\infty$ is defined by the formula (3.4).

Proof. Due to the assumptions of Theorem 2.1, $f_1 \in C[0,b]$ and $T_1 = (A+T)J$ is compact as an operator from $L^\infty(0,b)$ to $C[0,b]$ and to $L^\infty(0,b)$, too. Since equation $v = T_1 v$ has in $L^\infty(0,b)$ only the trivial solution $v = 0$, then there exists an inverse operator $(I - T_1)^{-1} \in \mathcal{L}(L^\infty(0,b), L^\infty(0,b))$ and equation (2.5) has a unique solution $v = (I - T_1)^{-1} f_1 \in L^\infty(0,b)$. By Theorem 2.1, $v \in C^{m,\nu}[0,b]$. A standard discussion (cf. [6]) together with Lemmas 3.1 and 3.2 yields that there exists a number $N_0 \in \mathbf{N}$ such that for $N \geq N_0$ the operator $(I - \mathcal{P}_N T_1)$ is invertible in $L^\infty(0,b)$ and

$$\left\|(I - \mathcal{P}_N T_1)^{-1}\right\|_{\mathcal{L}(L^\infty(0,b), L^\infty(0,b))} \leq c, \quad N \geq N_0. \tag{4.6}$$

Thus, (4.3) possesses a unique solution $v_N \in S_{m-1}^{(-1)}(\Pi_N)$ for $N \geq N_0$ and

$$\|v - v_N\|_\infty \leq c\|v - \mathcal{P}_N v\|_\infty, \quad N \geq N_0,$$

where $v = u' \in C^{m,\nu}[0,b]$ is the solution of (2.5). This together with Lemma 3.2 yields the estimate (4.5).

Further, using v_N, we find for $N \geq N_0$ an approximation u_N for u in the form (4.1). It follows from (2.2), (2.3), and (4.1) that

$$\begin{aligned} u(t) - u_N(t) &= (J(v - v_N))(t) \\ &= \int_0^t [v(s) - v_N(s)]ds - \frac{\beta}{\alpha + \beta} \int_0^b [v(s) - v_N(s)]ds, \quad t \in [0,b]. \end{aligned} \tag{4.7}$$

Therefore

$$\max_{0 \leq t \leq b} |u(t) - u_N(t)| \leq \left(1 + \left|\frac{\beta}{\alpha + \beta}\right|\right) \int_0^b |v(s) - v_N(s)|ds. \tag{4.8}$$

Since $T_1 = (A + T)J$, $(I - \mathcal{P}_N T_1)(v - v_N) = v - \mathcal{P}_N v$, and

$$\left(I - \mathcal{P}_N T_1\right)^{-1} = I + \left(I - \mathcal{P}_N T_1\right)^{-1} \mathcal{P}_N T_1, \quad N \geq N_0,$$

we get from (4.6) and Lemma 3.1 the estimate

$$\begin{aligned} |v(s) - v_N(s)| \leq &\left|v(s) - (\mathcal{P}_N v)(s)\right| \\ &+ c \textstyle\int_0^b \left|v(t) - (\mathcal{P}_N v)(t)\right| dt, \quad s \in [0,b], \quad N \geq N_0. \end{aligned} \tag{4.9}$$

Now it follows from (4.8) and (4.9) that

$$\|u - u_N\|_\infty \leq c \int_0^b \left|v(t) - (\mathcal{P}_N v)(t)\right| dt, \quad N \geq N_0.$$

This together with $v \in C^{m,\nu}[0,b]$ and Lemma 3.2 yields the estimate (4.4). □

5. Superconvergence phenomenon. It follows from Theorem 4.1 that by using method (4.1), (4.2), one can reach a convergence order

$$\|u - u_N\|_\infty \leq cN^{-m}, \quad \|u' - v_N\|_\infty \leq cN^{-m} \tag{5.1}$$

for sufficiently large values of the grid parameter r and for every choice of collocation parameters $\eta_1, \ldots, \eta_m$ satisfying the condition (3.3). Since $u_N \in S_m^{(0)}(\Pi_N)$, the first estimate of (5.1) is not of optimal order. In the following we show that by a careful choice of the collocation parameters (3.3) it is possible, assuming a little more regularity of functions a, f, and K, to prove a superconvergence result for values of v_N at the collocation points and improve the convergence rate of u_N in the maximum norm. We refer also to the papers [15] and [20], where similar results for initial-value problems of Volterra integro-differential equations are given.

Theorem 5.1. *Let* $a, f \in C^{m+1,\nu}[0,b],\ K \in W^{m+1,\nu}(\Delta),\ m \in \mathbf{N},\ \nu \in \mathbf{R},\ \nu < 1$*;* $\alpha, \beta, \gamma \in \mathbf{R},\ \alpha+\beta \neq 0$*, and assume that problem* (2.1) *has in the set* $\{u \in C[0,b] : u' \in L^\infty(0,b)\}$ *only the trivial solution* $u = 0$*. Moreover, let the interpolation nodes* (3.2) *with grid points* (3.1) *and parameters* (3.3) *be used and let the parameters* $\eta_1, \dots, \eta_m$ *in* (3.3) *be chosen so that the quadrature approximation*

$$\int_0^1 g(s)ds \approx \sum_{k=1}^m w_k g(\eta_k), \quad 0 \le \eta_1 < \cdots < \eta_m \le 1, \tag{5.2}$$

with appropriate weights $w_k = w_k^{(m)},\ k = 1, \dots, m$*, is exact for all polynomials* g *of degree* m*.*

Then the statements of Theorem 4.1 *are valid. Moreover, for all* $N \ge N_0$ *the following error estimates hold:*

$$\max_{k=1,\dots,m,\ j=1,\dots,2N} |u'(t_{jk}) - v_N(t_{jk})| \le c\,\Theta_N(m,\nu,r) \tag{5.3}$$

and

$$\|u - u_N\|_\infty \le c\,\Theta_N(m,\nu,r). \tag{5.4}$$

Here u *is the exact solution of problem* (1.1)*,* u_N*, and* v_N *are determined by method* (4.1), (4.2)*,* c *is a positive constant not depending on* N*,* $\|\cdot\|_\infty$ *is defined by the formula* (3.4)*, and*

$$\Theta_N(m,\nu,r) = \begin{cases} N^{-r(2-\nu)} & \textit{for} \quad 1 \le r < \frac{m+1}{2-\nu}, \\ N^{-m-1}(1+\log N) & \textit{for} \quad r = \frac{m+1}{2-\nu} \ge 1, \\ N^{-m-1} & \textit{for} \quad r > \frac{m+1}{2-\nu},\ r \ge 1. \end{cases} \tag{5.5}$$

Proof. We know from the proof of Theorem 4.1 that (4.3) has a unique solution $v_N \in S_{m-1}^{(-1)}(\Pi_N)$ for $N \ge N_0$. We have for it and v, the solution of (2.5), that

$$(I - \mathcal{P}_N T_1)(v_N - \mathcal{P}_N v) = \mathcal{P}_N T_1(\mathcal{P}_N v - v). \tag{5.6}$$

As $I - \mathcal{P}_N T_1$ is invertible in $L^\infty(0,b)$ for $N \ge N_0$, we obtain from (4.6), (5.6), and Lemma 3.1 the estimate

$$\|\mathcal{P}_N v - v_N\|_\infty \le c\|T_1(v - \mathcal{P}_N v)\|_\infty, \quad N \ge N_0. \tag{5.7}$$

Since $T_1 = (A+T)J$ and $(\mathcal{P}_N v)(t_{jk}) = v(t_{jk}),\ k = 1, \dots, m,\ j = 1, \dots, 2N$, it follows from (2.3), (2.4), and (5.7) that

$$\begin{gathered} |v(t_{jk}) - v_N(t_{jk})| \le \|\mathcal{P}_N v - v_N\|_\infty \le c \max_{0 \le t \le b} \left| \int_0^t [v(s) - (P_N v)(s)]ds \right|, \\ k = 1, \dots, m,\ j = 1, \dots, 2N,\ N \ge N_0. \end{gathered} \tag{5.8}$$

It follows from Theorem 2.1 that $v \in C^{m+1,\nu}[0,b]$. Using this we can show that

$$\max_{0 \le t \le b} \left| \int_0^t [v(s) - (P_N v)(s)]ds \right| \le c\,\Theta_N(m,\nu,r), \tag{5.9}$$

where $\Theta_N(m,\nu,r)$ is given by the formula (5.5). This together with (5.8) and $v=u'$ yields (5.3).

In order to prove (5.9) we choose $m+1$ parameters $0 \le \tilde\eta_1 < \tilde\eta_2 < \cdots < \tilde\eta_{m+1} \le 1$ such that $\{\eta_1 \ldots, \eta_m\} \subset \{\tilde\eta_1, \ldots, \tilde\eta_{m+1}\}$ and set

$$\tilde t_{jk} = t_{j-1} + \tilde\eta_k(t_j - t_{j-1}), \quad k=1,\ldots,m+1,\ j=1,\ldots,2N,$$

where $\{t_j\}$ are given by the formulas (3.1). Moreover, we introduce an operator $\tilde{\mathcal{P}}_N$ which assigns to every continuous function $z : [0,b] \to \mathbf{R}$ its piecewise polynomial interpolation function $\tilde{\mathcal{P}}_N z \in S_m^{(-1)}(\Pi_N)$ such that

$$(\tilde{\mathcal{P}}_N z)(\tilde t_{jk}) = z(\tilde t_{jk}), \quad k=1,\ldots,m+1,\ j=1,\ldots,2N.$$

Due to Lemma 3.2,

$$\int_0^b |v(s) - (\tilde{\mathcal{P}}_N v)(s)| ds \le c\,\Theta_N(m,\nu,r). \tag{5.10}$$

Further, the quadrature approximation (5.2) is exact for all polynomials of degree not exceeding m. This yields that the equality

$$\int_{t_{j-1}}^{t_j} g(s)ds = (t_j - t_{j-1}) \sum_{k=1}^m w_k g(t_{jk}) \quad (j=1,\ldots,2N)$$

holds for all polynomials g of degree not exceeding m. Therefore for $j=1,\ldots,2N$

$$\left|\int_0^{t_j} [v(s) - (P_N v)(s)]ds\right| = \left|\int_0^{t_j} [v(s) - (\tilde P_N v)(s)]ds\right| \le \int_0^b |v(s) - (\tilde P_N v)(s)|ds.$$

This together with (5.10) yields

$$\max_{1\le j\le 2N} \left|\int_0^{t_j} [v(s) - (P_N v)(s)]ds\right| \le c\,\Theta_N(m,\nu,r). \tag{5.11}$$

Fix $t \in [0,b]$ and let $n \in \{1,\ldots,2N\}$ be such that $t \in [t_{n-1}, t_n]$. Actually, we consider only the case $n=1,\ldots,N$. For $n=N+1,\ldots,2N$ the argument is similar. It follows from Lemma 3.3 that

$$\left|\int_{t_{n-1}}^t [v(s) - (\mathcal{P}_N v)(s)]ds\right| \le c(t_n - t_{n-1})^{m+1} \begin{cases} 1 & \text{if} \quad m < 1-\nu, \\ 1 + |\log t_n| & \text{if} \quad m = 1-\nu, \\ t_n^{1-\nu-m} & \text{if} \quad m > 1-\nu. \end{cases}$$

Due to (3.1),

$$\begin{aligned}(t_n - t_{n-1})^{m+1} t_n^{1-\nu-m} &\le c n^{r(2-\nu)-m-1} N^{-r(2-\nu)} \\ &\le c \begin{cases} N^{-r(2-\nu)} & \text{if} \quad r(2-\nu) < m+1, \\ N^{-m-1} & \text{if} \quad r(2-\nu) \ge m+1. \end{cases}\end{aligned}$$

Therefore

$$\left|\int_{t_{n-1}}^t [v(s) - (\mathcal{P}_N v)(s)]ds\right| \le c\,\Theta_N(m,\nu,r), \quad t \in [t_{n-1}, t_n]. \tag{5.12}$$

This together with (5.11) yields (5.9), implying the estimate (5.3).

Let us prove the statement (5.4). Fix $t \in [0, b]$, let $n \in \{1, \dots, 2N\}$ be such that $t \in [t_{n-1}, t_n]$. Using (4.7) and $u' = v$ we obtain that

$$\begin{aligned}|u(t) - u_N(t)| \le \left| \int_0^{t_{n-1}} [v(s) - v_N(s)]ds \right| + \left| \int_{t_{n-1}}^t [v(s) - v_N(s)]ds \right| \\ + \left| \frac{\beta}{\alpha+\beta} \right| \left| \int_0^b [v(s) - v_N(s)]ds \right|, \quad t \in [t_{n-1}, t_n].\end{aligned} \tag{5.13}$$

Consider the first term on the right-hand side of (5.13). We have

$$\begin{aligned}\left| \int_0^{t_{n-1}} [v(s) - v_N(s)]ds \right| \le \left| \int_0^{t_{n-1}} [v(s) - (\tilde{\mathcal{P}}_N v)(s)]ds \right| + \left| \int_0^{t_{n-1}} [(\tilde{\mathcal{P}}_N v)(s) - v_N(s)]ds \right| \\ \le \int_0^b |v(s) - (\tilde{\mathcal{P}}_N v)(s)|ds + \sum_{j=1}^{n-1} (t_j - t_{j-1}) \sum_{k=1}^m |w_k| |v(t_{jk}) - v_N(t_{jk})|.\end{aligned}$$

This together with (5.3) and (5.10) yields

$$\left| \int_0^{t_{n-1}} [v(s) - v_N(s)]ds \right| \le c\,\Theta_N(m, \nu, r). \tag{5.14}$$

In a similar way we obtain that

$$\left| \int_0^b [v(s) - v_N(s)]ds \right| \le c\,\Theta_N(m, \nu, r). \tag{5.15}$$

It remains to estimate the second term on the right-hand side of (5.13). We have

$$\begin{aligned}\left| \int_{t_{n-1}}^t [v(s) - v_N(s)]ds \right| \le \left| \int_{t_{n-1}}^t [v(s) - (\mathcal{P}_N v)(s)]ds \right| \\ + \int_{t_{n-1}}^t |(\mathcal{P}_N v)(s) - v_N(s)|ds, \quad t \in [t_{n-1}, t_n].\end{aligned} \tag{5.16}$$

By (5.8) and (5.9),

$$\int_{t_{n-1}}^t |(\mathcal{P}_N v)(s) - v_N(s)|ds \le (t - t_{n-1}) \|\mathcal{P}_N v - v_N\|_\infty \le c\,\Theta_N(m, \nu, r)$$

for all $t \in [t_{n-1}, t_n]$. This together with the estimates (5.12)–(5.16) yields (5.4). □

Remark 5.1. Problem (1.1) can be rewritten also in the form

$$u = J(A + T)u + Jf + \frac{\gamma}{\alpha + \beta}, \tag{5.17}$$

where J and A, T are defined by the formulas (2.3) and (2.4), respectively. Using (5.17) one can construct another collocation method for the numerical solution of problem (1.1); cf. [6, 18]. This method will be discussed elsewhere.

6. Numerical experiments. Consider the following boundary-value problem:

$$u'(t) = u(t) + f(t) + \int_0^1 |t-s|^{-\frac{1}{2}} u(s)ds, \quad t \in [0,1], \quad u(0)+u(1) = 2. \tag{6.1}$$

The forcing function f is selected so that $u(t) = t^{\frac{3}{2}} + (1-t)^{\frac{3}{2}}$ is the exact solution. Actually, this is problem (1.1), where $a(t) = 1$,

$$f(t) = \tfrac{3}{2}\big(t^{\frac{1}{2}} - (1-t)^{\frac{1}{2}}\big) - \big(t^{\frac{3}{2}} + (1-t)^{\frac{3}{2}}\big) - \big(t^2 + (1-t)^2\big) \int_0^1 x^{-\frac{1}{2}} (1-x)^{\frac{3}{2}} dx$$
$$- t^{\frac{1}{2}} \int_0^1 x^{-\frac{1}{2}} (1-t-xt)^{\frac{3}{2}} dx - (1-t)^{\frac{1}{2}} \int_0^1 x^{-\frac{1}{2}} (t + (1-t)x)^{\frac{3}{2}} dx,$$

$K(t,s) = |t-s|^{-\frac{1}{2}}$, $\alpha = \beta = 1$, $\gamma = 2$, and $b = 1$. Moreover, it is easy to check that $a, f \in C^{m,\nu}[0,b]$ and $K \in W^{m,\nu}(\Delta)$ with $\nu = \frac{1}{2}$ and arbitrary $m \in \mathbf{N}$.

Problem (6.1) is solved numerically by method (4.1), (4.2), in the case $m = 2$. An approximation $v_N \in S_1^{(-1)}(\Pi_N)$ to $v = u'$, the derivative of the solution u of (6.1), is presented in the form (see Remark 4.2)

$$v_N(t) = c_{j1} \frac{t_{j2} - t}{t_{j2} - t_{j1}} + c_{j2} \frac{t - t_{j1}}{t_{j2} - t_{j1}}, \ t \in [t_{j-1}, t_j], \ j = 1, \ldots, 2N,$$

where t_{j1} and t_{j2} are defined by the formula (3.2), with $m = 2$ and $0 \le \eta_1 < \eta_2 \le 1$. If $\eta_1 > 0$ or $\eta_2 < 1$, then the collocation conditions (4.2) lead to a system of $4N$ linear algebraic equations for finding the coefficients $c_{j1} = v_N(t_{j1}), \ c_{j2} = v_N(t_{j2}), \ j = 1, \ldots, 2N$. If $\eta_1 = 0$ and $\eta_2 = 1$, then $t_{j2} = t_{j+1,1} = t_j, \ j = 1, \ldots, 2N-1$, and the collocation conditions (4.2) give us a system of $2N+1$ linear algebraic equations for finding the coefficients $c_{11} = v_N(t_0), \ c_{12} = c_{21} = v_N(t_1), \ldots, c_{2N-1,2} = c_{2N,1} = v_N(t_{2N-1}), \ c_{2N,2} = v_N(t_{2N})$.

TABLE 6.1
Results in the case $\eta_1 = \frac{1}{4}, \ \eta_2 = \frac{3}{4}$.

	$r=1$		$r=1.5$		$r=2$		$r=4.1$	
N	ε_N	ϱ_N (2.83)	ε_N	ϱ_N (4)	ε_N	ϱ_N (4)	ε_N	ϱ_N (4)
4	2.3E-3	2.53	1.0E-3	3.85	7.2E-4	4.09	1.5E-3	5.11
8	8.8E-4	2.64	2.7E-4	3.77	1.7E-4	4.30	2.7E-4	5.42
16	3.2E-4	2.71	7.1E-5	3.78	4.0E-5	4.17	5.8E-5	4.66
32	1.2E-4	2.75	1.9E-5	3.76	9.8E-6	4.08	1.3E-5	4.35
64	4.2E-5	2.78	5.0E-6	3.79	2.4E-6	4.04	3.2E-6	4.18
128	1.5E-5	2.79	1.3E-6	3.82	6.0E-7	4.02	7.8E-7	4.09
256	5.4E-6	2.80	3.4E-7	3.85	1.5E-7	4.01	1.9E-7	4.04
N	ε'_N	ϱ'_N (1.41)	ε'_N	ϱ'_N (1.68)	ε'_N	ϱ'_N (2)	ε'_N	ϱ'_N (4)
4	1.7E-1	1.42	1.2E-1	1.68	8.4E-2	2.00	3.4E-2	4.10
8	1.2E-1	1.42	7.1E-2	1.68	4.2E-2	2.00	8.1E-3	4.13
16	8.4E-2	1.42	4.2E-2	1.68	2.1E-2	2.00	2.0E-3	4.14
32	5.9E-2	1.41	2.5E-2	1.68	1.1E-2	2.00	4.8E-4	4.14
64	4.2E-2	1.41	1.5E-2	1.68	5.3E-3	2.00	1.1E-4	4.14
128	3.0E-2	1.41	8.8E-3	1.68	2.6E-3	2.00	2.8E-5	4.14
256	2.1E-2	1.41	5.3E-3	1.68	1.3E-3	2.00	6.7E-6	4.14

In Tables 6.1–6.4 some of the obtained numerical results for different values of the parameters N, r, η_1, and η_2 are presented. The quantities ε_N and ε'_N are the approximate values of the norms $\|u_N - u\|_\infty$ and $\|v_N - u'\|_\infty$, defined as follows:

$$\varepsilon_N = \{\max |u_N(\tau_{jk}) - u(\tau_{jk})| : k = 0, \ldots, 10, \ j = 1, \ldots, 2N\},$$
$$\varepsilon'_N = \{\max |v_N(\tau_{jk}) - u'(\tau_{jk})| : k = 0, \ldots, 10, \ j = 1, \ldots, 2N\},$$

where

$$\tau_{jk} = t_{j-1} + k\frac{t_j - t_{j-1}}{10}, \quad k = 0, \ldots, 10, \quad j = 1, \ldots, 2N.$$

The ratios $\varrho_N = \frac{\varepsilon_{N/2}}{\varepsilon_N}$ and $\varrho'_N = \frac{\varepsilon'_{N/2}}{\varepsilon'_N}$, characterizing the observed convergence rate, are also presented.

Table 6.2
Results in the case $\eta_1 = 0,\ \eta_2 = 1.$

	$r=1$		$r=1.5$		$r=2$		$r=4.1$	
N	ε_N	ϱ_N (2.83)	ε_N	ϱ_N (4)	ε_N	ϱ_N (4)	ε_N	ϱ_N (4)
4	1.8E-2	2.19	8.4E-3	3.34	5.0E-3	4.09	6.3E-3	4.49
8	7.5E-3	2.42	2.3E-3	3.66	1.2E-3	4.04	1.6E-3	3.92
16	2.9E-3	2.59	6.1E-4	3.72	3.1E-4	4.01	4.0E-4	4.07
32	1.1E-3	2.68	1.6E-4	3.78	7.7E-5	4.01	9.9E-5	4.04
64	4.0E-4	2.71	4.2E-5	3.82	1.9E-5	4.01	2.5E-5	4.02
128	1.5E-4	2.74	1.1E-5	3.85	4.8E-6	4.00	6.1E-6	4.01
256	5.3E-5	2.76	2.8E-6	3.88	1.2E-6	4.00	1.5E-6	4.00
N	ε'_N	ϱ'_N (1.41)	ε'_N	ϱ'_N (1.68)	ε'_N	ϱ'_N (2)	ε'_N	ϱ'_N (4)
4	1.3E-1	1.39	9.0E-2	1.64	6.3E-2	1.95	3.0E-2	3.96
8	9.1E-2	1.39	5.4E-2	1.66	3.2E-2	1.96	7.8E-3	3.89
16	6.5E-2	1.40	3.3E-2	1.67	1.6E-2	1.98	1.9E-3	4.00
32	4.6E-2	1.41	1.9E-2	1.68	8.2E-3	1.99	4.9E-4	4.00
64	3.3E-2	1.41	1.2E-2	1.68	4.1E-3	2.00	1.2E-4	4.00
128	2.3E-2	1.41	6.9E-3	1.68	2.1E-3	2.00	3.0E-5	4.00
256	1.6E-2	1.41	4.1E-3	1.68	1.0E-3	2.00	7.6E-6	4.00

In order to facilitate the comparison of numerical experiments with theoretical results we have used the notation $\varrho_N(\delta_r)$ and $\varrho'_N(\delta'_r)$ in the headings of Tables 6.1–6.2, where δ_r and δ'_r are the ratios (that are independent of N) corresponding to the error estimates (4.4) and (4.5) of Theorem 4.1 for $m = 2$, respectively. Since these error estimates do not depend on the values of the parameters η_1 and η_2, satisfying $0 \le \eta_1 < \eta_2 \le 1$, we get the same values for δ_r and for δ'_r in Tables 6.1–6.2.

Table 6.3
Results in the case $\eta_1 = (3 - \sqrt{3})/6,\ \eta_2 = (3 + \sqrt{3})/6.$

	$r=1$		$r=1.5$		$r=2$		$r=4.1$	
N	ε_N	ϱ_N (2.83)	ε_N	ϱ_N (4.76)	ε_N	ϱ_N (≈ 8)	ε_N	ϱ_N (8)
4	1.5E-3	2.72	5.8E-4	4.58	3.2E-4	5.82	8.0E-4	6.84
8	5.5E-4	2.78	1.2E-4	4.72	4.3E-5	7.42	8.6E-5	9.32
16	2.0E-4	2.81	2.6E-5	4.75	5.5E-6	7.70	9.9E-6	8.73
32	7.0E-5	2.82	5.4E-6	4.76	7.0E-7	7.87	1.2E-6	8.48
64	2.5E-5	2.82	1.1E-6	4.76	8.9E-8	7.94	1.4E-7	8.26
128	8.8E-6	2.83	2.4E-7	4.76	1.1E-8	7.95	1.7E-8	8.15
256	3.1E-6	2.83	5.0E-8	4.75	1.4E-9	7.86	2.1E-9	8.08
N	ε'_N	ϱ'_N (1.41)	ε'_N	ϱ'_N (1.68)	ε'_N	ϱ'_N (2)	ε'_N	ϱ'_N (4)
4	1.6E-1	1.42	1.1E-1	1.69	8.0E-2	2.01	3.2E-2	4.11
8	1.1E-1	1.42	6.8E-2	1.68	4.0E-2	2.00	7.6E-3	4.13
16	8.0E-2	1.42	4.0E-2	1.68	2.0E-2	2.00	1.8E-3	4.14
32	5.7E-2	1.42	2.4E-2	1.68	1.0E-2	2.00	4.5E-4	4.14
64	4.0E-2	1.41	1.4E-2	1.68	5.0E-3	2.00	1.1E-4	4.14
128	2.8E-2	1.41	8.4E-3	1.68	2.5E-3	2.00	2.6E-5	4.14
256	2.0E-2	1.41	5.0E-3	1.68	1.3E-3	2.00	6.3E-6	4.14

Tables 6.3 and 6.4 show the dependence of the convergence rates on the nonuniformity parameter r, when Gaussian parameters $\eta_1 = (3-\sqrt{3})/6, \eta_2 = (3+\sqrt{3})/6$ are used. In Table 6.3, the ratios δ_r correspond to the error estimate (5.4) of Theorem 5.1, and the ratios δ'_r correspond to the error estimate (4.5) of Theorem 4.1, as in Tables 6.1–6.2.

Table 6.4

Results in the case $\eta_1 = (3-\sqrt{3})/6,\ \eta_2 = (3+\sqrt{3})/6$ *at the collocation points.*

	$r=1$		$r=1.5$		$r=2$		$r=4.1$	
N	ξ'_N	σ'_N (2.83)	ξ'_N	σ'_N (4.76)	ξ'_N	σ'_N (≈ 8)	ξ'_N	σ'_N (8)
4	3.1E-3	2.91	1.1E-3	4.95	6.4E-4	6.28	1.7E-3	7.44
8	1.1E-3	2.93	2.2E-4	4.97	8.2E-5	7.81	1.6E-4	10.75
16	3.6E-4	2.92	4.6E-5	4.91	1.0E-5	7.99	1.5E-5	10.22
32	1.2E-4	2.90	9.4E-6	4.86	1.3E-6	8.03	1.6E-6	9.70
64	4.3E-5	2.88	2.0E-6	4.82	1.6E-7	8.03	1.7E-7	9.11
128	1.5E-5	2.87	4.1E-7	4.79	2.0E-8	7.98	2.0E-8	8.70
256	5.2E-6	2.86	8.5E-8	4.77	2.6E-9	7.61	2.4E-9	8.18

To illustrate the fact that the superconvergence of the approximate solution in the supremum norm is the result of the superconvergence of the derivative v_N at the collocation points, the errors of v_N at the collocation points, denoted by

$$\xi'_N = \{\max |v_N(t_{jk}) - u'(t_{jk})| : k = 1, \dots, m,\ j = 1, \dots, 2N\},$$

are presented in Table 6.4. Similarly to the previous analysis, we have computed the ratio $\sigma'_N = \frac{\xi'_{N/2}}{\xi'_N}$ and used the notation $\sigma'_N(\delta'_r)$, where δ'_r corresponds to the error estimate (5.3) of Theorem 5.1.

From Tables 6.1–6.4 we can see that the numerical results are in good accordance with the theoretical error estimates of Theorems 4.1 and 5.1.

Acknowledgment. The authors are grateful to both referees for their suggestions and remarks which helped to improve the presentation of the paper.

REFERENCES

[1] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge Monogr. Appl. Comput. Math. 4, Cambridge University Press, Cambridge, UK, 1997.

[2] C. T. H. Baker, *The Numerical Treatment of Integral Equations*, Monogr. Numer. Anal., Clarendon Press, Oxford, UK, 1977.

[3] H. Brunner, *Polynomial spline collocation methods for Volterra integro-differential equations with weakly singular kernels*, IMA J. Numer. Anal., 6 (1986), pp. 221–239.

[4] H. Brunner, *Collocation Methods for Volterra Integral and Related Functional Differential Equations*, Cambridge Monogr. Appl. Comput. Math. 15, Cambridge University Press, Cambridge, UK, 2004.

[5] H. Brunner and P. J. van der Houwen, *The Numerical Solution of Volterra Equations*, CWI Monogr. 3, North–Holland, Amsterdam, 1986.

[6] H. Brunner, A. Pedas, and G. Vainikko, *Piecewise polynomial collocation methods for linear Volterra integro-differential equations with weakly singular kernels*, SIAM J. Numer. Anal., 39 (2001), pp. 957–982.

[7] H. Brunner, A. Pedas, and G. Vainikko, *A spline collocation method for linear Volterra integro-differential equations with weakly singular kernels*, BIT, 41 (2001), pp. 891–900.

[8] Z. Chen, C. A. Micchelli, and Y. Xu, *Fast collocation methods for second kind integral equations*, SIAM J. Numer. Anal., 40 (2002), pp. 344–375.

[9] M. Ganesh and I. H. Sloan, *Optimal order spline methods for nonlinear differential and integro-differential equations*, Appl. Numer. Math., 29 (1999), pp. 445–478.

[10] W. Hackbusch, *Integral Equations. Theory and Numerical Treatment*, Internat. Ser. Numer. Math 120, Birkhäuser-Verlag, Basel, Switzerland, 1995.

[11] R. J. Hangelbroek, H. G. Kaper, and G. K. Leaf, *Collocation methods for integro-differential equations*, SIAM J. Numer. Anal., 14 (1977), pp. 377–390.

[12] Q. Hu, *Geometric meshes and their application to Volterra integro-differential equations with singularities*, IMA J. Numer. Anal., 18 (1998), pp. 151–164.

[13] Q. Hu, *Interpolation correction for collocation solution for Fredholm integro-differential equation*, Math. Comp., 67 (1998), pp. 987–999.

[14] H. Kaneko, R. Noren, and P. Padilla, *Superconvergence of the iterated collocation methods for Hammerstein equations*, J. Comput. Appl. Math., 80 (1997), pp. 335–349.

[15] R. Kangro and I. Parts, *Superconvergence in the maximum norm of a class of piecewise polynomial collocation methods for solving linear weakly singular Volterra integro-differential equations*, J. Integral Equations Appl., 15 (2003), pp. 403–427.

[16] A. Karamete and M. Sezer, *A Taylor collocation method for the solution of linear integro-differential equations*, Int. J. Comput. Math., 79 (2002), pp. 987–1000.

[17] R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, 1989.

[18] I. Parts and A. Pedas, *Spline collocation methods for weakly singular Volterra integro-differential equations*, in Numerical Mathematics and Advanced Applications, Enumath 2001, F. Brezzi, A. Buffa, S. Corsar, and A. Merli, eds., Springer-Verlag, Milano, 2003, pp. 919–928.

[19] I. Parts and A. Pedas, *Collocation approximations for weakly singular Volterra integro-differential equations*, Math. Model. Anal., 8 (2003), pp. 315–328.

[20] T. Tang, *Superconvergence of numerical solutions to weakly singular Volterra integro-differential equations*, Numer. Math., 61 (1992), pp. 373–382.

[21] T. Tang, *A note on collocation methods for Volterra integro-differential equations with weakly singular kernels*, IMA J. Numer. Anal., 13 (1993), pp. 93–99.

[22] G. Vainikko, *Multidimensional Weakly Singular Integral Equations*, Lecture Notes in Math. 1549, Springer-Verlag, Berlin, 1993.

[23] G. Vainikko and A. Pedas, *The properties of solutions of weakly singular integral equations*, J. Austral. Math. Soc. Ser. B, 22 (1981) pp. 419–430.

[24] G. Vainikko, A. Pedas, and P. Uba, *Methods for Solving Weakly Singular Integral Equations*, University of Tartu, Tartu, Estonia, 1984 (in Russian).

[25] W. Volk, *The numerical solution of linear integro-differential equations by projection methods*, J. Integral Equations, 9 (1985), pp. 171–190.

[26] W. Volk, *The iterated Galerkin methods for linear integro-differential equations*, J. Comput. Appl. Math., 21 (1988), pp. 63–74.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 1912–1933

SPECTRAL METHODS BASED ON PROLATE SPHEROIDAL WAVE FUNCTIONS FOR HYPERBOLIC PDEs[*]

Q.-Y. CHEN[†], D. GOTTLIEB[†], AND J. S. HESTHAVEN[†]

Abstract. We examine the merits of using prolate spheroidal wave functions (PSWFs) as basis functions when solving hyperbolic PDEs using pseudospectral methods.

The relevant approximation theory is reviewed and some new approximation results in Sobolev spaces are established. An optimal choice of the band-limit parameter for PSWFs is derived for single-mode functions.

Our conclusion is that one might gain from using the PSWFs over the traditional Chebyshev or Legendre methods in terms of accuracy and efficiency for marginally resolved broadband solutions.

Key words. prolate spheroidal wave functions, spectral methods, penalty methods

AMS subject classifications. 65M70, 41A30, 65D05

DOI. 10.1137/S0036142903432425

1. Introduction.

Pseudospectral methods for PDEs [6, 13] approximate the solution by classical polynomials (usually Chebyshev or Legendre) or trigonometric polynomials. The main reason for their success is the spectral accuracy, i.e., the convergence rate depends only on the smoothness of the functions being approximated. This comes at a price, however, as the norm of the differentiation matrix is proportional to the square of the number, N, of interpolation points (or the order of the polynomials), resulting in small time-steps ($\sim N^{-2}$) [14], when using an explicit schemes for time integration.

This stringent restriction on the time-step can be attributed to the basis functions being classical orthogonal polynomials, the roots of which cluster near the boundaries of the interval, e.g., the smallest distance between any two roots of a Chebyshev polynomial of degree N is $O(N^{-2})$. In [18], it was suggested to use a singular mapping to change the basis functions to overcome this restriction, and this technique has been successfully used by many people (e.g., [1, 2, 10, 16, 20, 21]). However, as shown in [16, 20] this mapping only allows for doubling the time-step for practical N. If N is large, however, the time-step can be increased to scale as $O(N^{-1})$ [18, 10] without sacrificing the accuracy as the impact of the singular mapping becomes dominated by the finite precision. The mapping destroys the quadrature properties of the roots of the classical polynomials, which may be a disadvantage in certain applications, e.g., when filtering is needed or if integrals must be computed as part of the solution, e.g., in spectral element methods.

In this paper we assess the performance of pseudospectral methods based on prolate spheroidal wave functions (PSWF – ψ_k^c) rather than on polynomials. In [25], the authors demonstrate the merits of using PSWFs for the interpolation, integration

[*] Received by the editors July 25, 2003; accepted for publication (in revised form) April 13, 2005; published electronically December 8, 2005.

http://www.siam.org/journals/sinum/43-5/43242.html

[†] Division of Applied Mathematics, Brown University, Box F, Providence, RI 02912 (cqy@cfm.brown.edu, dig@cfm.brown.edu, Jan.Hesthaven@brown.edu). The work of the second author was partly supported by the NSF under contract DMS-0207451 and by HyperComp under contract HPC/F33615-01-c-1866. The work of the third author was partly supported by the NSF under contract DMS-0074257 and through NSF Career Award DMS-0132967, by ARO under contract DAAD19-01-1-0631, and by the Alfred P. Sloan Foundation through a Sloan Research Fellowship.

(quadrature), and differentiation of band-limited functions. They show, among other things, that for a prescribed accuracy fewer grid points are required for interpolation and integration than with Chebyshev polynomials. Furthermore, the differentiation matrix has a smaller condition number, approaching $O(N^{3/2})$, which suggest the possibility of increasing the time-step significantly for large values of N.

These basic observations have led to a surge of recent activity in the development of methods based on PSWFs, although the topic itself remains in its infancy. In [4, 5], the author studied the feasibility of using PSWFs as the basis functions in spectral element methods. More recently, in [3] Beylkin and Sandberg developed a two-dimensional solver for the acoustic wave equation by using a basis of approximate PSWFs. However, even basic aspects of approximation and stability theory for methods based on PSWFs remain unknown.

In this work we consider some of these issues, in particular in the context of solving hyperbolic PDEs by constructing pseudospectral methods based on quadrature points and roots associated with the PSWFs. The first step in this direction is to review and expand the relevant approximation theory. We discuss basic approximation properties such as the number of points per wavelength required to recover a meaningful result and show that only two points per wavelength are needed. Thus, the PSWF expansion recovers the Nyquist limit from Fourier theory, although defined on a finite interval. This should be contrasted with polynomial expansions where asymptotic estimates show that at least π points per wavelength are needed [14]. We derive a new result that demonstrates the spectral accuracy of approximations of smooth functions by the PSWFs.

Several variants of pseudospectral PSWF methods based on different interpolation points are subsequently discussed, the main differences being in the definition of the interpolation points, e.g., we consider genuine Gauss-type quadrature points as well as Gauss–Lobatto like points defined as the roots of $(1-x^2)(\psi_N^{2c})'$, where ψ_N^{2c} is the Nth order PSWF with bandwidth $2c$—this approach is clearly inspired by results from classical polynomials although they are in this case not associated with a quadrature. The performance of these slightly different methods are essentially equivalent although the latter choice is more appropriate for solving initial-boundary value problems. We finally consider the performance of these methods for solving a scalar hyperbolic equation as well as hyperbolic systems.

The results of our study can be summarized as follows.

- A practical relation between the two parameters, c and N, is $N = c$ to allow convergence.
- With this choice one observes spectral accuracy. When the solution is broadband and marginally resolved, the PSWF-based method is more accurate than the Chebyshev method with the same number of terms, i.e., generally more efficient.
- Theoretically the time-step Δt can be taken as $O(N^{-\frac{3}{2}})$ if $N \simeq \frac{2}{\pi}c$. However, the accuracy deteriorates significantly in this case.

The remaining part of the paper is organized as follows. In section 2, we present some mathematical background and define the PSWFs. Section 3 contains some approximation results, while section 4 deals with the construction of pseudospectral methods based on PSWFs. We discuss their stability and solve scalar hyperbolic equations as well as hyperbolic systems. In the appendix, we give the details of the proof of the main approximation result.

2. Preliminaries.

In this section, we shall summarize the notation and some general results regarding the PSWFs.

2.1. Prolate spheroidal wave functions.

A function $f(x) : [-1,1] \to [-1,1]$ is band-limited if there exist a $c > 0$ and a function $\phi(t) \in L^2[-1,1]$ such that

$$f(x) = F_c(\phi)(x) = \int_{-1}^{1} e^{icxt}\, \phi(t)\, dt.$$

It is easy to see that F_c: $L^2[-1,1] \to L^2[-1,1]$ is a compact operator, i.e., that it has eigenvalues $\lambda_0, \lambda_1, \lambda_2, \dots$, with the property $|\lambda_{i-1}| \geq |\lambda_i|\ \forall i > 0$. We shall denote by $\psi_j^c(x)$ the eigenfunction corresponding to λ_j. Then

$$\lambda_j\ \psi_j^c(x) \ = \int_{-1}^{1} e^{icxt}\ \psi_j^c(t)\, dt, \qquad x \in [-1,1], \tag{2.1}$$

and the eigenfunctions, $\{\psi_j^c\}_{j=0}^{+\infty}$, are the PSWFs. We choose to normalize them so that $\|\psi_j^c\|_{L^2[-1,1]} = 1$.

One easily checks that the PSWFs also satisfy

$$\mu_j\ \psi_j^c(x) \ = \int_{-1}^{1} \frac{\sin(c(x-t))}{x-t}\ \psi_j^c(t)\, dt, \qquad x \in [-1,1],$$

where

$$\mu_j = \frac{c}{2\pi}|\lambda_j|^2.$$

The following theorem gives some properties of the PSWFs (see [22, 25] and the references therein).

Theorem 2.1. *For all $c \geq 0$,*

- *ψ_0^c, $\psi_1^c, \dots$ are real, orthonormal, smooth, and complete in $L^2[-1,1]$, and they form a Chebyshev system* [17] *on $[-1,1]$;*
- *the ψ_k^c with even k are even functions, and those with odd k are odd;*
- *$\lambda_j = i^j|\lambda_j| \neq 0$, where i is the complex unit;*
- *among $\{\mu_j\}_{j=0}^{\infty}$, about $2c/\pi$ are very close to 1; order* $\log(c)$ *decay exponentially from 1 to nearly 0; the remaining ones are very close to 0.*

Furthermore, there exists a strictly increasing positive sequence $\chi_0, \chi_1, \dots$, such that

$$\Big((1-x^2)(\psi_j^c(x))'\Big)' + \big(\chi_j - c^2x^2\big)\psi_j^c(x) = 0. \tag{2.2}$$

When $c = 0$, the above equation reduces to the classic singular Sturm–Liouville problem with $p(x) = 1 - x^2$, $q(x) = 0$, $\omega(x) = 1$, and $\chi_j = j(j+1)$, i.e., the PSWFs with $c = 0$ are the normalized Legendre polynomials [6, 13].

Following [25], one can evaluate ψ_j^c by expressing it as

$$\psi_j^c(x) = \sum_{k=0}^{\infty} \beta_k^j \overline{P}_k(x), \qquad j = 0, 1, 2, \dots, \tag{2.3}$$

where $\overline{P}_k$ is the normalized Legendre polynomial of degree k. Substituting (2.3) into (2.2) and using the properties of the Legendre polynomials one obtains an eigenvalue problem

$$(A - \chi_j \cdot I)\beta^j = 0. \tag{2.4}$$

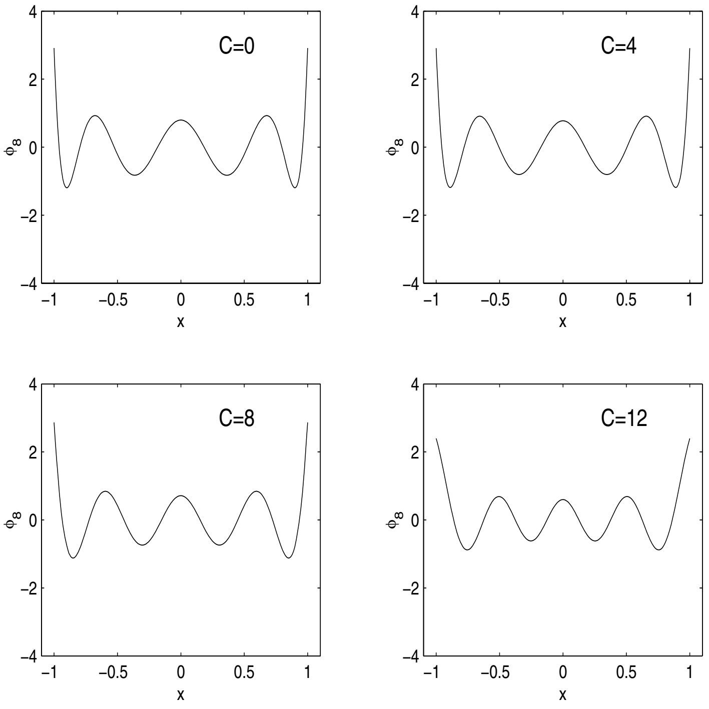


FIG. 2.1. *$\psi_8^c(x)$ for different values of c.*

Here A has the form [25]

$$\begin{cases} A_{k,k} = k(k+1) + \dfrac{2k(k+1)-1}{(2k+3)(2k-1)}\, c^2, \\ A_{k,k+2} = \dfrac{(k+2)(k+1)}{(2k+3)\sqrt{(2k+1)(2k+5)}}\, c^2, \\ A_{k+2,k} = A_{k,k+2} \end{cases}$$

for $k = 0, 1, 2, \dots$, where the remaining entries of A are zeros.

Since ψ_j^c is smooth, the coefficients β_k^j decay superalgebraically with respect to k. The following theorem [25] offers guidelines on where to truncate (2.3) to ensure a certain accuracy in the approximation of ψ_j^c.

THEOREM 2.2. *Assume ψ_m^c is the mth PSWF with band-limit c, and λ_m is the corresponding eigenvalue. If*

$$k \geq 2(\lfloor e \cdot c \rfloor + 1), \tag{2.5}$$

then $\forall c > 0$,

$$\left| \int_{-1}^{1} \psi_m^c(x) \overline{P_k(x)}\, dx \right| < \frac{1}{\lambda_m} \left(\frac{1}{2} \right)^{k-1}. \qquad \square$$

Solving (2.4) and using the corresponding eigenvector in the truncated version of (2.3) allows for the computation of one PSWF (Figure 2.1) for different values of the band-limit, c. In Figure 2.1, we note that the zeros of the PSWF move toward the center as c increases, approaching a uniform distribution. This observation suggests that by choosing a suitable $c > 0$ the PSWF method needs fewer points per wavelength to accurately resolve a wave problem as compared to approximations based on classical orthogonal polynomials. However, it also suggests that if one chooses c too large for a fixed N, the PSWF is unable to represent functions defined on the whole interval.

3. Approximation.

In this section, we consider in more detail the properties of approximations based on PSWFs. We first show that for the single wave $\cos(M\pi x)$, with the optimal $c = M\pi$, the continuous PSWF expansion converges exponentially fast when at least two PSWFs are retained per wavelength. Equivalently, two points per wavelength are required for exponential convergence of the discrete approximation. This should be contrasted with about π points per wavelength needed for methods using classical orthogonal polynomials.

The second result pertains to the approximation of a general smooth function with a finite series of PSWFs. Recall that, for an unknown function, the optimal choice of the bandwidth parameter, c, is unknown and the approximation depends on two parameters, c and N. A natural approach is assume that the parameters are related and our experiments show that $c = N$ is a good choice if we want to maintain the full accuracy (16 digits). We explain why we cannot use $c \geq (\pi/2)N$ and illustrate that there can be benefits in taking $c \simeq (\pi/2)N$, albeit at the price of a lower accuracy.

3.1. Approximation of waves-points per wavelength.

Let us consider the wave $u(x) = e^{iM\pi x}$. It follows directly from (2.1) and Theorem 2.1 that its PSWF expansion is

$$e^{iM\pi x} = \sum_{j=0}^{+\infty} \left(\lambda_j \psi_j^c(1)\right) \psi_j^c(x), \tag{3.1}$$

where $c = M\pi$.

Note that

$$|\lambda_j \psi_j^c(1)|^2 = |\lambda_j||\lambda_j \psi_j^c(1)^2|,$$

where the term $\lambda_j \psi_j^c(1)^2$ is the jth term in the expansion of $e^{i\pi M}$ (cf. (3.1)) and thus bounded—in fact it tends to zero with growing j. From [19], we know that $|\lambda_j|$ decays exponentially with j if $j > \frac{2c}{\pi} = 2M$. This establishes the result: *The accurate resolution of a wave requires two PSWFs per wave.* We recall here that expansions based on Chebyshev or Legendre polynomials require about π points per wave. Only mapped methods [20] may achieve similar resolution results for sufficiently high values of N.

In Figure 3.1, we plot the L^2-error of the truncated PSWF expansion of the function $\cos(M\pi x)$ versus $\frac{N}{M}$ (N is the number of terms in the expansion). It clearly confirms that when $N/M > 2$ the error decays exponentially.

In the above discussion we took $c = M\pi$, which is optimal. However, for general functions, we do not have a simple optimal c (see Figure 3.2) where we display the interpolation results with the PSWFs for two different functions. Clearly, the optimal c depends on the required accuracy and the function being approximated. This is due to the fact that an arbitrary function has many different modes and each mode has a distinct optimal c.

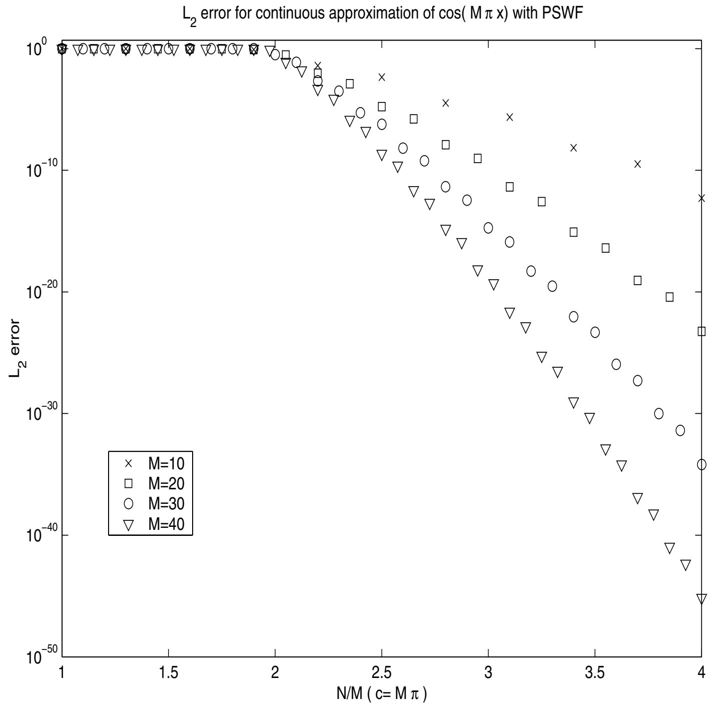


FIG. 3.1. *L^2-error of the PSWF expansion* (*truncated after N terms*) *of* $\cos(M\pi x)$ *versus* N/M. ×: $M = 10$; □: $M = 20$; ○: $M = 30$; ▽: $M = 40$.

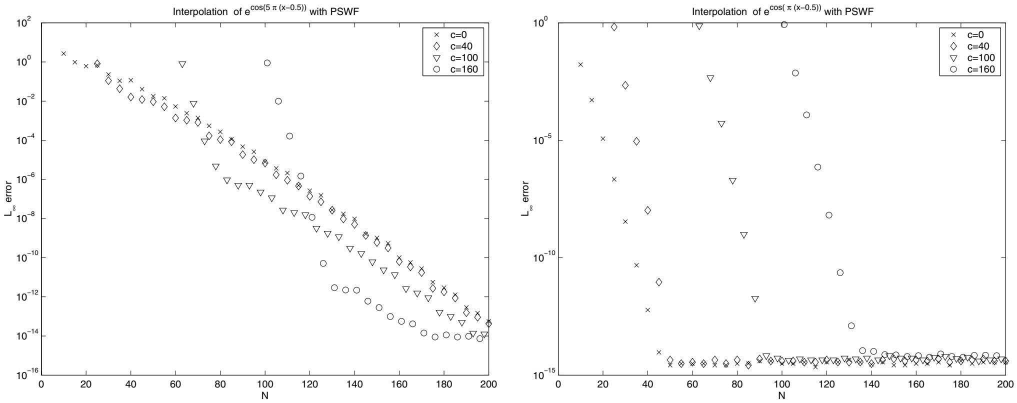


FIG. 3.2. *Demonstration that the optimal c, if it exists, depends on the specific problem and the required accuracy. Left: L_∞-error for the interpolation of* $e^{\cos(5\pi(x-0.5))}$. *For an error around* 10^{-6}, $c = 100$ *is the best choice. For an error as small as* 10^{-10}, $c = 160$ *is optimal. Right: L_∞-error for the interpolation of* $e^{\cos(\pi(x-0.5))}$. *Clearly,* $c = 0$ (*Legendre basis*) *is the best among the four choices.*

3.2. Error estimates. In this section, we consider the error estimates, in a Sobolev norm, of the PSWF expansion of a smooth function. Let $x \in [-1, 1]$, and consider the expansion $u(x) = \sum_{k=0}^{+\infty} \hat{u}_k \psi_k^c(x)$. The order of the convergence of the partial sum $u_N(x) = \sum_{k=0}^{N} \hat{u}_k \psi_k^c(x)$ is determined by

$$\|u - u_N\|_{L^2[-1,1]}^2 \leq \sum_{k=N+1}^{\infty} |\hat{u}_k|^2,$$

i.e., it depends solely on the decay rate of the coefficients $\{\hat{u}_k\}$.

Using the standard notation of $H^s[-1,1]$ for the Sobolev space of functions with distributional derivatives up to order s being square integrable in $L^2[-1,1]$, we prove in the appendix the following theorem.

Theorem 3.1. *Assume that $u \in H^s[-1,1]$ with the PSWF expansion $u(x) = \sum_{i=0}^{+\infty} \hat{u}_i \psi_i^c(x)$.*

If $q_N = \sqrt{\frac{c^2}{\chi_N}} < 1$, then

$$|\hat{u}_N| \leq D \left(N^{-\frac{2}{3}s} \|u\|_{H^s[-1,1]} + (q_N)^{\delta N} \ \|u\|_{L^2[-1,1]} \right), \tag{3.2}$$

where both δ and D are positive constants.

From (3.2) it is evident that the expansion coefficients, $\hat{u}_N$, may exhibit spectral convergence when $q_N < 1$. In [23], it is shown that if n grows with c as

$$n = \frac{2}{\pi} \left[c + b \log(2\sqrt{c}) \right]$$

for some b, then

$$\chi_n \sim c^2 + 2bc + O(1).$$

Thus

$$q_n < 1 \Leftrightarrow \chi_n > c^2 \Leftrightarrow b > 0 \Leftrightarrow n > \frac{2}{\pi} c.$$

Consequently, the finite PSWF expansion of a smooth function, $u \in C^\infty[-1,1]$,

$$\sum_{k=0}^{N} \hat{u}_k \psi_k^c(x)$$

is spectrally accurate if and only if

$$N > \frac{2}{\pi} c.$$

In Figure 3.3, we display the relationship between N and c ensuring that $q_N \leq 1$, obtained directly by solving the eigenvalue problem. This clearly confirms the above result. Figure 3.4 shows the loss of accuracy as N approaches $\frac{2}{\pi}c$. The loss of accuracy partially confirms Theorem 3.1. More precisely, the second term in (3.2) is dominant as N approaches $\frac{2}{\pi}c$, i.e., q_N approaches one. When q_N is very close to one, $(q_N)^{\delta N}$ cannot be small for any moderate N.

We notice that $c = N$ (which guarantees that q_N is bounded away from one) appears to be a good choice if one requires maximum accuracy, although larger values of c may also work if a reduced accuracy is acceptable. In section 4, we will further discuss the issue of choosing c when also considering the time-step and discrete stability.

Similar results are obtained when we use the PSWFs to interpolate a smooth function. In Figure 3.5, we compare interpolations based on PSWF and Chebyshev polynomials. Here we choose the number of grid points $N = c$. The results indicate that the PSWF interpolation is superior for functions with fine structures.

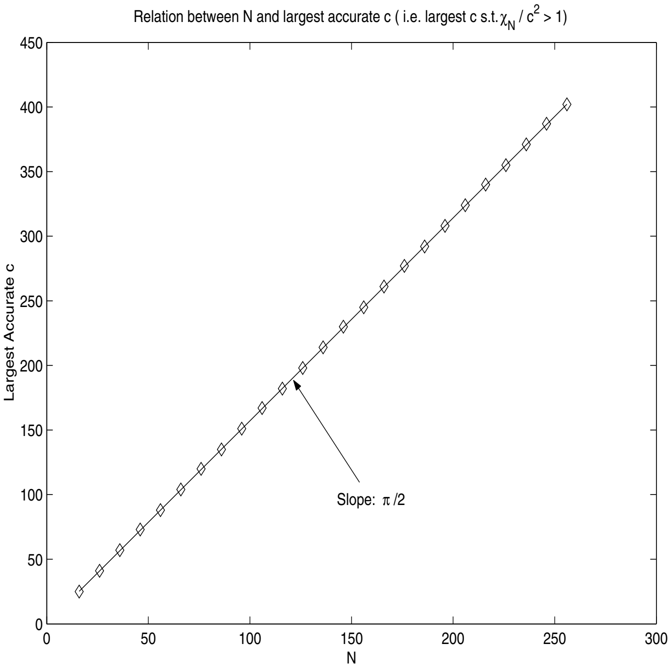


FIG. 3.3. *Computational validation of the relation between* N *and the biggest* c *which makes* $q_N \leq 1$. *The slope is* $\pi/2$, *as predicted in the text.*

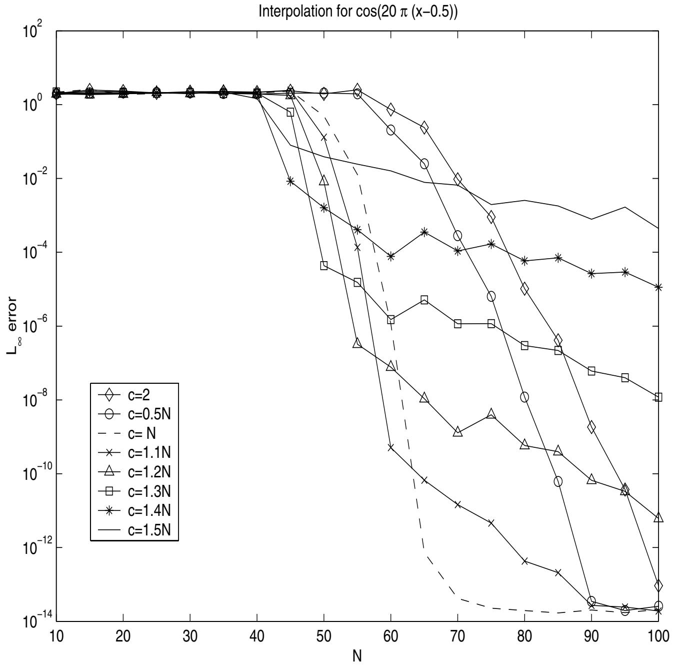


FIG. 3.4. *The loss of accuracy as* c *approaches* $\frac{\pi}{2}N$.

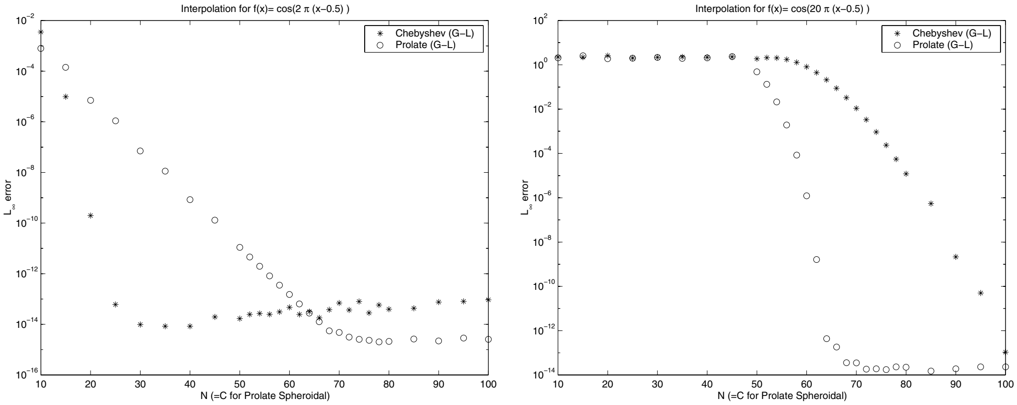


FIG. 3.5. *Interpolation results.* ○: *Prolate spheroidal wave basis;* ∗: *Chebyshev polynomials basis.* $c = N$ *for prolate spheroidal wave basis. Left:* $f(x) = \cos(2\pi(x - 0.5))$. *Right:* $f(x) = \cos(20\pi(x - 0.5))$.

4. Solving PDEs.

In the following we shall discuss the use of the PSWFs as a basis in spectral methods for solving wave problems. Particular attention shall be paid to issues of semidiscrete and fully discrete stability.

4.1. First order wave equation.

Consider the first order one-way wave equation

$$\begin{cases} u_t = u_x, & x \in [-1, 1], \\ u(1, t) = g(t), & \\ u(x, 0) = f(x) & \end{cases} \tag{4.1}$$

for which we shall seek a numerical solution.

Consider the interpolation points $\{x_0, \ldots, x_N\}$ which will be specified later. We define the Prolate–Lagrange function as $L_j(x) = \sum_{k=0}^{N} l_{jk}\psi_k^c(x)$ such that $L_j(x_k) = \delta_{jk}$. The existence of Prolate–Lagrange functions follows from the fact that the PSWFs form a Chebyshev system [17].

In a penalty Galerkin approximation we seek an approximation to the wave problem of the form

$$u_N(x, t) = \sum_{j=0}^{N} u_N(x_j, t) L_j(x)$$

such that the vector $\vec{U} = (u_N(x_0, t), \ldots, u_N(x_N, t))^T$ satisfies the equation

$$M \frac{d\vec{U}}{dt} = S\vec{U} - \tau(u_N(1, t) - g(t))\vec{e}_N. \tag{4.2}$$

Here, the boundary condition is imposed in a penalty way [7, 12, 15]. The matrices $M = (m_{jk})$ and $S = (s_{jk})$ are defined as

$$m_{jk} = \int_{-1}^{1} L_j(x) L_k(x)\, dx, \tag{4.3}$$

$$s_{jk} = \int_{-1}^{1} L_j(x) L_k'(x)\, dx, \tag{4.4}$$

and $\vec{e}_N = (0, \dots, 1)^T$.

THEOREM 4.1 (stability). *The semidiscrete method described in* (4.2) *is stable for* $\tau \geq 1/2$.

Proof. For the stability proof it suffices to assume that $g(t) = 0$. Multiplying (4.2) by $\vec{U}^T$, we get

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}\left(\vec{U}^T M \vec{U}\right) &= \sum_{jk} u_N(x_j,t) s_{kj} u_N(x_k,t) - \tau u_N(1,t)^2 \\
&= \sum_{jk} \int_{-1}^{1} u_N(x_j,t) u_N(x_k,t) L_j(x) L_k'(x)\, dx - \tau u_N(1,t)^2 \\
&= \int_{-1}^{1} u_N(x,t) \frac{\partial u_N(x,t)}{\partial x} - \tau u_N(1,t)^2 \\
&= \frac{1}{2}\left(u_N(1,t)^2 - u_N(-1,t)^2 - 2\tau u_N(1,t)^2\right).
\end{aligned}$$

Thus, if $\tau \geq \frac{1}{2}$, then

$$\frac{d}{dt} \sum_{jk} \int_{-1}^{1} u_N(x_j,t) u_N(x_k,t) L_j(x) L_k(x)\, dx \leq 0$$

or

$$\frac{d}{dt} \int_{-1}^{1} \left(u_N(x,t)\right)^2\, dx \leq 0.$$

This proves the theorem. □

One way to implement the pseudospectral (collocation) method is to replace the integrals in (4.3) and (4.4) by quadrature formulas based on the points $\{x_k\}$. Alternatively, one can substitute the approximation $u_N(x,t)$ for u into the PDE (4.1) and require that the obtained equation is satisfied at certain collocation points (in most cases $\{x_k\}$ are used as collocation points as well).

For the PSWF collocation method, we do not have a stability proof. The difficulty is caused by the fact that the product of any two of the first N PSWFs with band-limit c is not in the space spanned by the first $2N$ PSWFs with band-limit $2c$ for which the PSWF quadrature is exact. However, when using $\{x_k\}$ as the collocation points, we numerically verify that the eigenvalues of the differentiation matrix have negative real parts.

We shall consider two sets of grid points as $\{x_k\}$: the Gauss–Lobatto PSWF points (one way to compute them is given in [8]) and the zeros of $(1-x^2)(\psi_N^{2c})'$. Note that these points must be computed from PSWF with band-limit $2c$ (see [25]). As we find the performance of the methods based on these two sets of points to be almost equivalent, the latter will be used for the PSWF collocation method if not specified otherwise.

When using explicit time discretization, e.g., Runge–Kutta schemes, one faces a stability limit on the time-step Δt. A necessary condition for stability is that the product of Δt and the largest eigenvalue of the differential matrix, being $M^{-1}(S - \tau \vec{e}_N \vec{e}_N^T)$ in the current scheme, is inside the stability region of the time-stepping scheme.

In Figure 4.1, we observe that for fixed N the magnitude of the largest eigenvalue λ of the PSWF collocation method decreases when c/N increases. So without violating

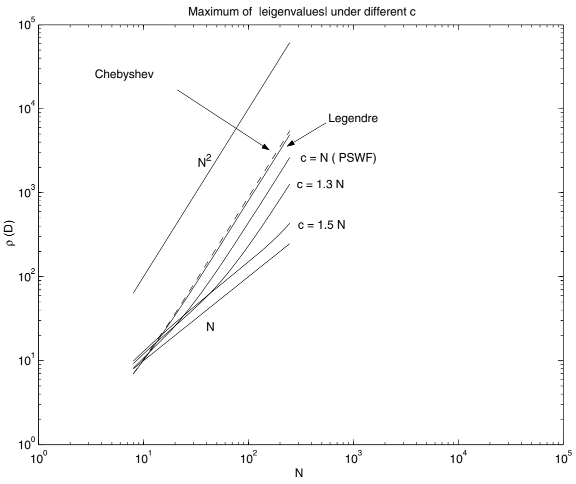


FIG. 4.1. *The largest absolute eigenvalue of the PSWF collocation method using different values of c, and the Chebyshev and Legendre collocation methods.*

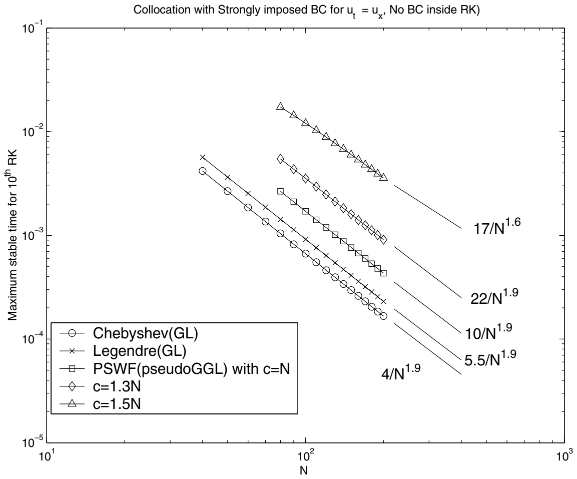


FIG. 4.2. *The largest stable time-step for the PSWF collocation method using different values of c, the Chebyshev and Legendre collocation methods. A* 10*th order explicit Runge–Kutta scheme was used.*

the stability condition, a larger c leads to larger Δt, as confirmed by Figure 4.2. When computing the largest stable time-step, we implemented a 10th order explicit Runge–Kutta scheme, the general form (mth order) for $u_t = Au$ with constant matrix A being given as [6]

$$\begin{aligned} u_1 &= u^n + \frac{\Delta t}{m}\, Au^n, \\ u_k &= u^n + \frac{\Delta t}{m+1-k}\, Au_{k-1}, \qquad k = 2, \dots, m-1, \\ u^{n+1} &= u^n + \Delta t\, Au_{m-1}. \end{aligned}$$

This ensures that the errors from the time integration are negligible.

In Figure 4.2, the largest stable time-step approaches a growth rate $O(N^{-\frac{3}{2}})$, when c goes to $(\pi/2)N$. This suggests that one can use a time-step of order $O(N^{-\frac{3}{2}})$ by letting $c = (\pi/2)N$. However, this choice of c causes a loss of accuracy, as demonstrated in Figure 3.4. In Table 4.1, we list the errors for the time-steps shown in Figure 4.2. It is evident that the accuracy is decreasing when c approaches $(\pi/2)N$. This is consistent with our analysis for the approximation using PSWFs.

Table 4.1

L_∞-error when solving $u_t = u_x$ for $u(x,t) = \cos(2\pi(x+t-0.5))$ with collocation methods. A 10th order explicit Runge–Kutta is used. For each N of each method, Δt is the largest stable time-step shown in Figure 4.2.

N	80	120	160	200
Chebyshev	3.453×10^{-14}	4.952×10^{-14}	1.521×10^{-13}	1.115×10^{-13}
Legendre	7.361×10^{-14}	1.117×10^{-12}	1.274×10^{-12}	1.592×10^{-12}
PSWF($c = N$)	9.770×10^{-15}	9.104×10^{-15}	2.081×10^{-14}	1.482×10^{-14}
PSWF($c = 1.3N$)	3.638×10^{-1}	2.860×10^{-9}	7.133×10^{-12}	9.137×10^{-14}
PSWF($c = 1.5N$)	5.022×10^{-2}	2.968×10^{-1}	8.051×10^{-2}	1.649×10^{-4}

The PSWF method offers a systematic way of balancing accuracy and stability. As a compromise, $c = N$ is used in all subsequent numerical tests. This yields a time-step which is twice the one obtained by a Legendre collocation method without sacrificing accuracy. Similar results can be obtained by using a mapping technique [16]. In some cases it may be beneficial to use a different value of c, e.g., in Figure 3.4, $c = 1.1N$ could be used if only about 10^{-9} accuracy was required. Similar improvements over the traditional Chebyshev collocation methods can also be achieved by the mapping technique which was first presented in [18], albeit at a loss of the quadrature. However, it will be impractical to use the PSWF collocation method if one wants to change c very often, as both the interpolation points and the differentiation matrix have to be recomputed when c is changed.

4.1.1. Numerical tests. The following numerical tests were carried out with a collocation method that determines a nodal approximation $u_N(x,t) = \sum_{j=0}^{N} u_N(x_j,t)L_j(x)$ such that the equation

$$\frac{\partial u_N}{\partial t} - \frac{\partial u_N}{\partial x} = 0 \tag{4.5}$$

is satisfied at the grid points $\{x_j\}$. The boundary condition is applied either strongly or by a penalty procedure as discussed above.

We considered three different initial conditions, listed in Table 4.2.

Table 4.2

Initial condition $f(x)$.

Smooth	Nonsmooth
$\cos(2\pi(x-0.5))$	
$\cos(20\pi(x-0.5))$	$\sin(20\pi(x-0.5)) + H(x-0.5)$

The Heaviside function $H(x)$ is defined as

$$H(x) = \begin{cases} 1 & \text{if } x \leq 0, \\ -1 & \text{otherwise.} \end{cases} \tag{4.6}$$

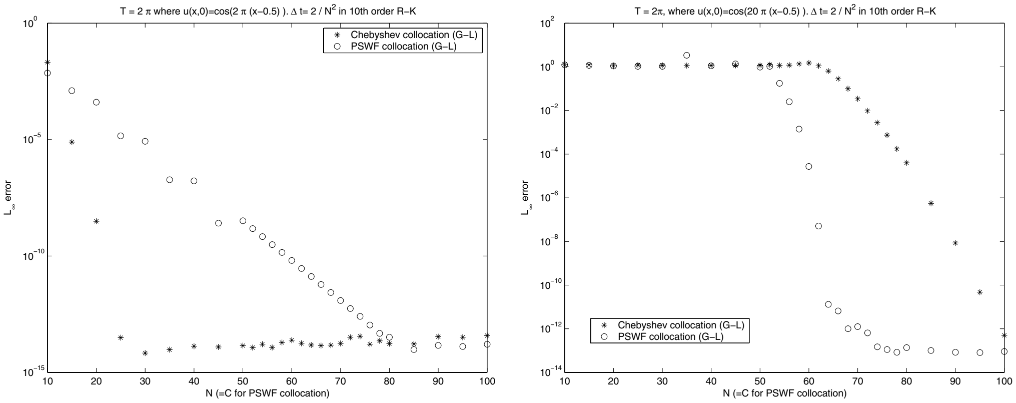


FIG. 4.3. *L_∞-error from solving $u_t = u_x$ with a collocation method and strongly imposed boundary condition. Final time: $T = 2\pi$. 10th order Runge–Kutta with $\Delta t = \frac{2}{N^2}$. Left: $u(x,t) = \cos(2\pi(x - 0.5 + t))$. Right: $u(x,t) = \cos(20\pi(x - 0.5 + t))$. ○: PSWF; ∗: Chebyshev.*

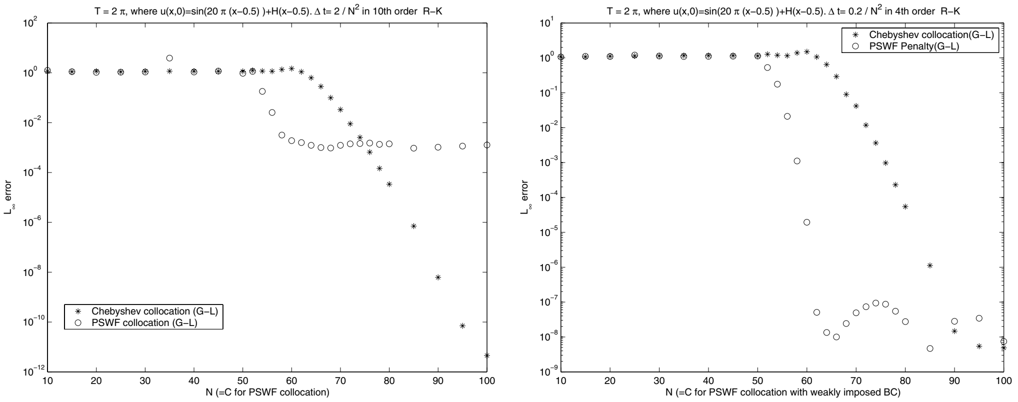


FIG. 4.4. *L_∞-error of solving $u_t = u_x$ by collocation methods. $u(x,0) = \sin(20\pi(x - 0.5)) + H(x - 0.5)$. Final time: $T = 2\pi$. Left: Chebyshev and PSWF collocation methods with strongly imposed boundary condition. Right: PSWF collocation method with weakly imposed boundary condition.*

In Figure 4.3, we show the errors from solving (4.5) with these smooth initial conditions. The Chebyshev method performs better for functions with small wave numbers, whereas the PSWF method is clearly better for functions with large wave numbers.

In Figure 4.4, we present the errors for the discontinuous initial condition. In this case the solution is discontinuous and the point of discontinuity propagates towards the boundary with a speed $a = 1$. We observe that the error does not decay below 10^{-4} when using a strongly imposed boundary condition.

When the boundary condition is imposed by a penalty procedure [7, 15, 12], the PSWF method is superior to the Chebyshev method (see the right part of Figure 4.4). We also applied the Legendre collocation method to solve the equation with discontinuous initial conditions. Similar to the PSWF collocation method, the weakly imposed boundary condition yields more accurate results than the strongly imposed boundary condition.

The improved performance with the weak imposition of the boundary condition

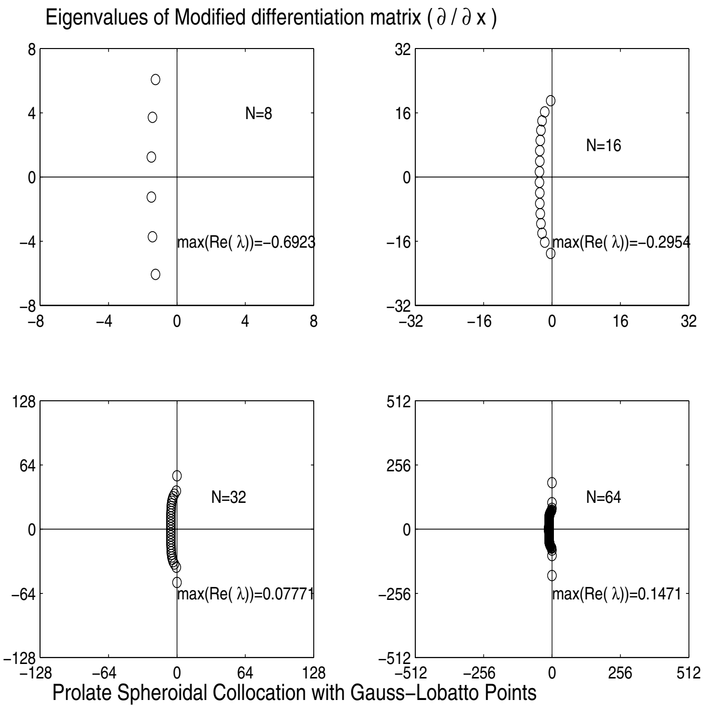


FIG. 4.5. *Eigenvalues of the differentiation matrix for the PSWF collocation method. Boundary condition is imposed strongly.*

can be linked to the behavior of the differentiation matrix. Figures 4.5 and 4.6 show the spectrum of the modified differentiation matrix for the PSWF collocation method with strongly and weakly imposed boundary conditions, respectively. We believe that the positive real parts of eigenvalues for $N = 32$ and 64 in Figure 4.5 are spurious and caused by round-off errors, as discussed in [24] for Chebyshev/Legendre spectral differentiation matrices. These results document the importance of imposing boundary conditions in a penalty way.

4.2. A cavity problem. In this section, we solve the one-dimensional Maxwell equations

$$\begin{cases} \epsilon \dfrac{\partial E}{\partial t} = \dfrac{\partial H}{\partial x}, \\ \mu \dfrac{\partial H}{\partial t} = \dfrac{\partial E}{\partial x}, \end{cases} \tag{4.7}$$

where $E(x,t)$ and $H(x,t)$ are the tangential electric and magnetic fields, and ϵ and μ are the relative permittivity and permeability of the materials.

We shall consider the test case of a one-dimensional cavity $[-1,1]$ filled with two dielectric media with a material interface at $x = 0$. Two perfectly conducting walls are located at $x = -1$ and $x = 1$. Denote by ϵ_1 and μ_1 the relative permittivity and permeability of the material at $[-1,0]$. Similarly, ϵ_2 and μ_2 are the relative permittivity and permeability of the material in $[0,1]$. The electric and magnetic fields in the two domains are denoted by E_1, H_1 and E_2, H_2.

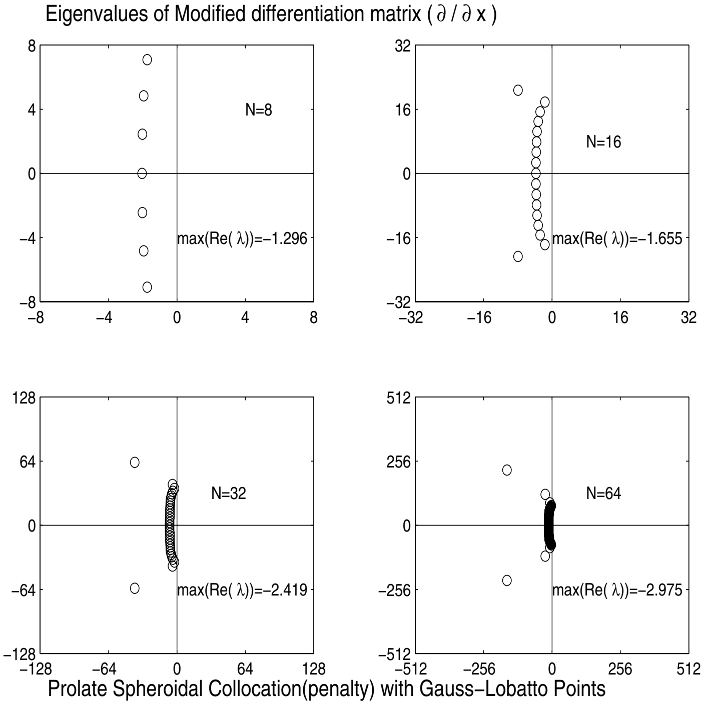


Fig. 4.6. *Eigenvalues of the differentiation matrix for the PSWF collocation method. Boundary condition is imposed weakly.*

Since the walls are perfectly conducting, the boundary conditions are

$$E_1(-1,t) = 0 \quad \text{or} \quad \frac{\partial H_1}{\partial x}|_{x=-1} = 0,$$

$$E_2(1,t) = 0 \quad \text{or} \quad \frac{\partial H_2}{\partial x}|_{x=1} = 0.$$

Denote $n_1 = \sqrt{\epsilon_1}$ and $n_2 = \sqrt{\epsilon_2}$, i.e., $\{n_i\}$ is the index of refraction. In all the following tests, we assume $\mu_1 = \mu_2 = 1.0$, $n_1 = 1$, and $n_2 = 10$.

In Figure 4.7, we display the solution at $t = 0$. (See [9] for the derivation of the exact solution.) When $n_1 \neq n_2$, the solution loses smoothness at the material interface. It is only globally C^0 in $[-1, 1]$. Thus without using domain decomposition, we can only get second order convergence with a Chebyshev or PSWF collocation method (see Figure 4.8). Because of this low order there is limited advantage to the use of the PSWF collocation method, although the PSWF method needs fewer points per wavelength to resolve the solution.

For the pointwise errors from both PSWF and Chebyshev collocations, there is a spike (Figure 4.9) propagating into the left-half domain and whose speed is the speed of a characteristic wave. It is caused by the initial condition being computed from the exact solution to the PDE, rather than an exact solution to the numerical scheme. One can remedy this by computing the initial conditions from the numerical scheme.

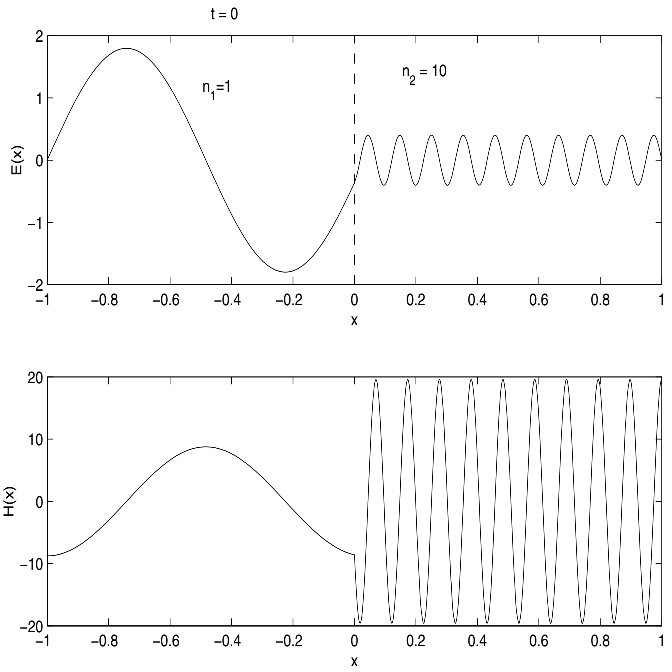


Fig. 4.7. *Exact solution at $t = 0$, $n_1 = 1.0$, $n_2 = 10.0$. Upper: electric field $E(x,t)$. Lower: magnetic field $H(x,t)$.*

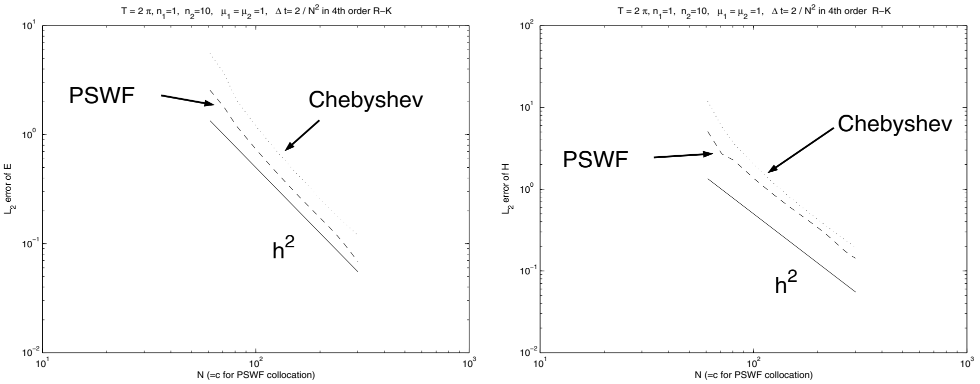


Fig. 4.8. *The discrete L_2-error at $t = 2\pi$. Strongly imposed boundary condition for the Chebyshev collocation method, weakly imposed boundary condition for the PSWF collocation method. Left: electric field $E(x,t)$. Right: magnetic field $H(x,t)$.*

Assume that the semidiscrete equation of (4.7) is

$$\begin{cases} \dfrac{d\vec{E}}{dt} = D_H\, \vec{H}, \\[2ex] \dfrac{d\vec{H}}{dt} = D_E\, \vec{E}. \end{cases} \tag{4.8}$$

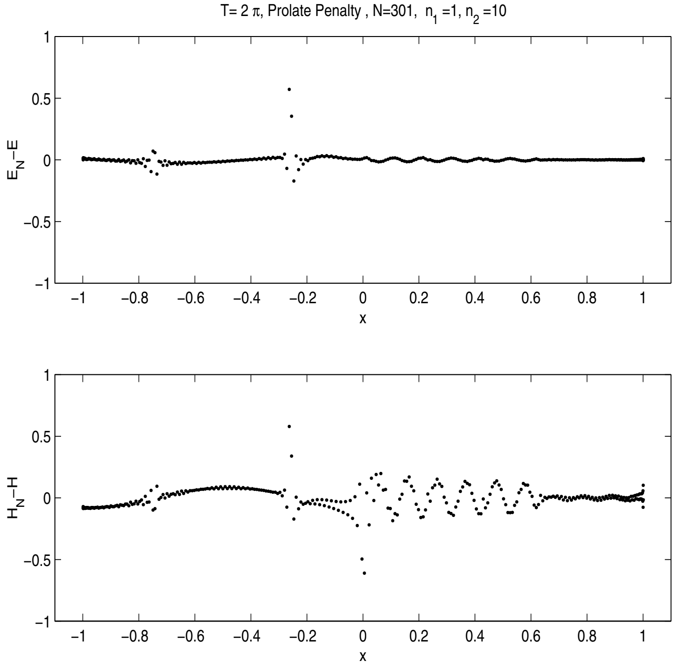


FIG. 4.9. *Pointwise error from the PSWF collocation method.* $c = N = 301$. $t = 2\pi$. *Upper: electric field. Lower: magnetic field.*

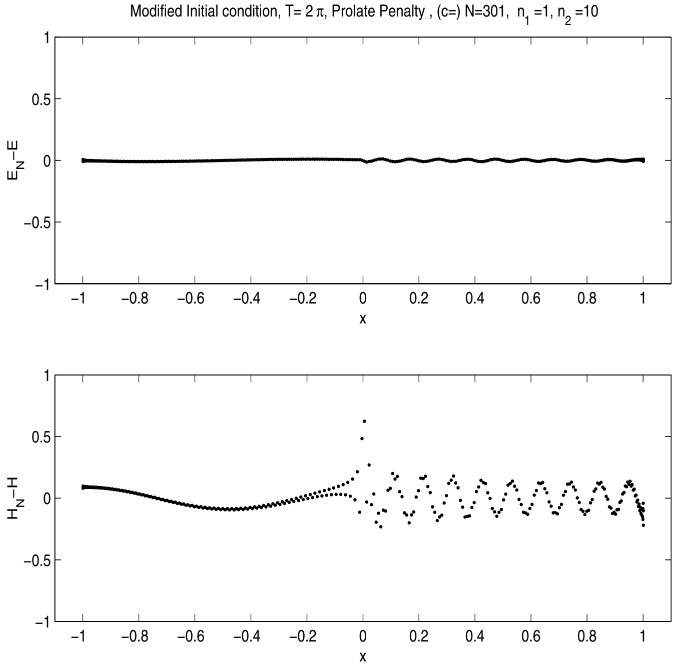


FIG. 4.10. *Pointwise errors with the initial conditions computed from the numerical scheme.*

Take the exact solution to the numerical scheme as $\vec{E} = \tilde{\vec{E}} e^{i\omega t}$ and $\vec{H} = \tilde{\vec{H}} e^{i\omega t}$, and introduce them into (4.8) to obtain an eigenvalue problem,

$$i\omega \begin{pmatrix} \tilde{\vec{E}} \\ \tilde{\vec{H}} \end{pmatrix} = \begin{pmatrix} 0 & D_H \\ D_E & 0 \end{pmatrix} \begin{pmatrix} \tilde{\vec{E}} \\ \tilde{\vec{H}} \end{pmatrix}.$$

The eigenvectors can be used as the initial conditions for the numerical scheme. The new results are shown in Figure 4.10, confirming this to be the source of the spike.

5. Conclusions. Our study of the applicability of PSWF-based methods to the numerical solution of time-dependent PDEs results in the following conclusions:

- The PSWF approximation requires two points per wavelength to resolve a single mode wave function ($\cos(m\pi x)$) if c is chosen as $c = m\pi$.
- Approximating a broadband function $u(x)$ by a finite expansion of the form $\sum_{n=0}^{N} \hat{u}_n \psi_n^c$, one obtains spectral accuracy for $N > \frac{2}{\pi} c$ with loss of accuracy when N approaches the limit. A robust choice is $N = c$.
- When solving the wave equation $u_t = u_x$ with explicit temporal schemes, the CFL bound on the time-step increases as $c \le (\pi/2)N$ increases. Asymptotically, $\Delta t = O(N^{-3/2})$ if c is very close to $(\pi/2)N$. However, this choice results in a deterioration of the accuracy. We found $c = N$ to be a good choice to ensure good accuracy and large stable time-step, the latter effectively increasing by a factor of 2 over methods based on classical orthogonal polynomials.
- For marginally resolved broadband problems, the PSWF-based method with a carefully chosen c is better than the Legendre/Chebyshev collocation methods. Fewer points are needed per wavelength for fast convergence and the allowable time-step is twice as large.
- The weak imposition of the boundary condition is necessary for the success of the method for problems with discontinuous initial conditions. By weakly applying the boundary condition, we improve the spectrum of the first order differentiation matrix of the PSWF collocation method, i.e., moving those eigenvalues with almost zero real parts a little distance away from the imaginary axis, thus introducing a small amount of dissipation.

Appendix. In this appendix, we prove Theorem 3.1.

Let $\beta_k = \beta_k^N$ be the coefficient in the expansion of ψ_N^c in terms of the normalized Legendre polynomials, i.e., $\psi_N^c(x) = \sum_{k=0}^{+\infty} \beta_k^N \overline{P}_k(x)$, where

$$\beta_k = \int_{-1}^{1} \overline{P}_k(x) \psi_N(x)\, dx.$$

The following recurrence relation for β_k is proven in [25]:

$$\begin{aligned} \frac{(k+2)(k+1)}{(2k+3)\sqrt{(2k+5)(2k+1)}}\,\beta_{k+2} &= \left(\frac{\Lambda - k(k+1)}{c^2} - \frac{2k(k+1)-1}{(2k+3)(2k-1)} \right) \beta_k \\ &\quad - \frac{k(k-1)}{(2k-1)\sqrt{(2k-3)(2k+1)}}\,\beta_{k-2}. \end{aligned} \tag{A.1}$$

Note that, from [23], $\Lambda = \chi_N = O(N^2)$. Let m be any integer satisfying

$$m = O(\Lambda^{1/3}) = O(N^{2/3}) \quad \text{and} \quad 2m(2m+1) < \frac{\ln 2}{2} \Lambda. \tag{A.2}$$

Then we have the following lemma.

LEMMA A.1. *Assume* $q = q_N = \sqrt{\frac{c^2}{\Lambda}} < 1$. *Then for any given* $k \leq 2m$, β_k *is bounded by*

$$|\beta_k| \leq \begin{cases} D\left(\frac{2}{q}\right)^k |\beta_0|, & k \text{ even}, \\ D\left(\frac{2}{q}\right)^k |\beta_1|, & k \text{ odd}, \end{cases} \tag{A.3}$$

where D *is a constant independent of* m.

Proof. We give the proof only for even k. The proof for odd k is similar.

Rewrite (A.1) as

$$\beta_{k+2} = \frac{1}{f(k+2)}\left(\frac{1}{q^2}\left(1 - \frac{k(k+1)}{\Lambda}\right) - g(k)\right)\,\beta_k - \frac{f(k)}{f(k+2)}\,\beta_{k-2}, \tag{A.4}$$

where $f(x) = \frac{x(x-1)}{(2x-1)\sqrt{(2x-3)(2x+1)}}$ and $g(x) = \frac{2x(x+1)-1}{(2x+3)(2x-1)}$. It is easy to verify that

$$1/4 \leq f(x) \leq 2\sqrt{5}/15, \qquad \frac{1}{2} \leq g(x) \leq \frac{11}{21} \qquad \text{for } x \geq 2.$$

Therefore, $f(x)/f(x+2) \leq 8\sqrt{5}/15$ when $x \geq 2$.

Since

$$k \leq 2m \Rightarrow \frac{1}{q^2}\left(1 - \frac{k(k+1)}{\Lambda}\right) \geq \frac{1}{q^2}\left(1 - \frac{\ln 2}{2}\right) > \frac{11}{21} \geq g(x) \qquad \text{for } x \geq 2,$$

the coefficient of β_k in (A.4) is positive. Hence, if we assume (A.3) is true for $k, k-2$, we can bound β_{k+2} as

$$\begin{aligned} |\beta_{k+2}| &\leq \frac{1}{f(k+2)}\left(\frac{1}{q^2}\left(1 - \frac{k(k+1)}{\Lambda}\right) - g(k)\right)\,|\beta_k| + \frac{f(k)}{f(k+2)}\,|\beta_{k-2}| \\ &\leq 4\frac{1}{q^2}\left(1 - \frac{\ln 2}{2}\right)\,D\left(\frac{2}{q}\right)^k |\beta_0| + \frac{8\sqrt{5}}{15}\,D\left(\frac{2}{q}\right)^{k-2} |\beta_0| \\ &\leq D\left(\frac{2}{q}\right)^{k+2}\left(1 - \frac{\ln 2}{2} + \frac{\sqrt{5}q^4}{30}\right)|\beta_0| \leq D\left(\frac{2}{q}\right)^{k+2} |\beta_0|. \end{aligned}$$

The last inequality follows from the facts that $q < 1$ and $1 - \frac{\ln 2}{2} + \frac{\sqrt{5}q^4}{30} < 1$. When $k = 0, 2$, (A.3) can be easily satisfied by modifying the constant D. This completes the proof. □

Define

$$A_k = \int_{-1}^{1} x^k \psi_N^c(x)\,dx. \tag{A.5}$$

One can check that $\sqrt{2}\beta_0 = A_0$ and $\sqrt{2/3}\beta_1 = A_1$.

LEMMA A.2. *Let* m *be an integer satisfying* (A.2). *Then*

$$|A_0| \leq Kq^{2m}\sqrt{\frac{2}{4m+1}}, \tag{A.6}$$

where K is a constant independent of m.

Proof. We first show that

$$|A_0| \le q^{2m}|A_{2m}| \prod_{l=1}^{m-1} \frac{1}{1-\frac{2l(2l+1)}{\Lambda}}. \tag{A.7}$$

Rewrite (2.2) as

$$\left((1-x^2)\psi'\right)' + \Lambda(1-q^2x^2)\psi = 0.$$

For $l \le m$, multiply the above equation by x^{2l}, then integrate on $[-1,1]$ to obtain

$$\begin{cases} 2l(2l-1)A_{2l-2} + \left(\Lambda - 2l(2l+1)\right)A_{2l} - \Lambda q^2 A_{2l+2} = 0, & l \ge 1, \\ A_0 - q^2 A_2 = 0, & l = 0. \end{cases}$$

Since $2m(2m+1) \le \Lambda$, all $A_0, A_2, \ldots, A_{2m+2}$ have the same sign. Thus

$$|A_{2l}| \le q^2\ |A_{2l+2}| \frac{\Lambda}{\Lambda - 2l(2l+1)} \le q^2\ |A_{2l+2}| \frac{1}{1-\frac{2l(2l+1)}{\Lambda}}.$$

Then (A.7) follows by induction.

To show (A.6), we note that $1 - x \ge e^{-2x}$ when $0 \le x \le \frac{\ln 2}{2}$. Therefore,

$$1 - \frac{2l(2l+1)}{\Lambda} \ge e^{-2\frac{2l(2l+1)}{\Lambda}} \qquad \text{if } l = 1, 2, \ldots, m-1,$$

which leads to

$$\prod_{l=1}^{m-1} \frac{1}{1-\frac{2l(2l+1)}{\Lambda}} \le e^{\frac{\sum_{l=1}^{m-1} 4l(2l+1)}{\Lambda}} \le e^{\frac{\frac{8}{3}m^3}{\Lambda}}.$$

From (A.2), $m = O(\Lambda^{1/3})$. So (A.5) yields

$$|A_{2m}| \le \|x^{2m}\|_{L^2[-1,1]}\ \|\psi\|_{L^2[-1,1]} \le \sqrt{\frac{2}{4m+1}},$$

which proves (A.6). □

In the same way, one can also show that $|A_1| \le Kq^{2m}\sqrt{\frac{2}{4m+3}}$ under the same conditions on m. We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Assume $u(x)$ has the Legendre expansion

$$u(x) = \sum_{k=0}^{+\infty} a_k P_k(x).$$

By definition,

$$\hat{u}_N = \int_{-1}^{1} u(x)\psi_N(x)\,dx = \int_{-1}^{1} \psi_N(x)\left(\sum_{k=0}^{+\infty} a_k P_k(x)\right) dx.$$

Let M be an integer such that

$$\frac{M+1}{2m} = \gamma \frac{\ln(1/q)}{\ln(2/q)}, \tag{A.8}$$

where m is defined in (A.2) and $0 < \gamma < 1$ is a constant. Denote by $u_M(x)$ the partial sum $u_M(x) = \sum_{k=0}^{M} a_k P_k(x)$. Then

$$\hat{u}_N = \int_{-1}^{1} u_M(x)\psi_N(x)\,dx \; + \; \int_{-1}^{1} (u(x) - u_M(x))\,\psi_N(x)\,dx.$$

We use I and II to represent the first and second terms, respectively. According to the error estimate of the Legendre approximation [11, 6],

$$|II| \le \|u - u_M\|_{L^2[-1,1]}\; \|\psi_N(x)\|_{L^2[-1,1]} \le DM^{-s}\|u\|_{H^s[-1,1]},$$

where D is a constant (in the following, D is used for different constants). Now,

$$\begin{aligned}
|I| &= \left|\sum_{k=0}^{M} a_k \int_{-1}^{1} P_k(x)\psi_N(x)\,dx\right| = \left|\sum_{k=0}^{M}\left(a_k\sqrt{\frac{2}{2k+1}}\right)\;\left(\int_{-1}^{1}\overline{P}_k(x)\psi_N(x)\,dx\right)\right| \\
&\le \left(\sum_{k=0}^{M}(a_k)^2\frac{2}{2k+1}\right)^{1/2}\;\left(\sum_{k=0}^{M}\left(\int_{-1}^{1}\overline{P}_k(x)\psi_N(x)\,dx\right)^2\right)^{1/2} \\
&\le \|u\|_{L^2[-1,1]}\;\left(\sum_{k=0}^{M}\beta_k^2\right)^{1/2} \\
&\le D\|u\|_{L^2[-1,1]}\;\left(\sum_{k=0}^{M}\left(\frac{2}{q}\right)^{2k}\right)^{1/2}\quad \max\,(|\beta_0|, |\beta_1|).
\end{aligned}$$

Here Lemma A.1 is used in the last reduction.

From Lemma A.2, $\beta_0 = \frac{1}{\sqrt{2}}A_0 < Kq^{2m}\sqrt{\frac{2}{4m+1}}$ and $\beta_1 = \sqrt{3/2}A_1 < Kq^{2m}\sqrt{\frac{2}{4m+3}}$, where K is a constant. Thus

$$\begin{aligned}
|I| &\le D\|u\|_{L^2[-1,1]}\;\left(\sum_{k=0}^{M}\left(\frac{2}{q}\right)^{2k}\right)^{1/2}\quad q^{2m}\sqrt{\frac{2}{4m+3}} \\
&\le D\|u\|_{L^2[-1,1]}\;\left(\frac{2}{q}\right)^{M+1}\quad q^{2m}\sqrt{\frac{2}{4m+3}} \\
&\le D\|u\|_{L^2[-1,1]}\;\left(q\left(\frac{2}{q}\right)^{\frac{M+1}{2m}}\right)^{2m}\;\sqrt{\frac{2}{4m+3}} \\
&\le D\|u\|_{L^2[-1,1]}\;p^{2m}\sqrt{\frac{2}{4m+3}},
\end{aligned}$$

where $p = q(\frac{2}{q})^{\frac{M+1}{2m}}$.

From (A.8), $p = q^{1-\gamma}$ and $M = O(m) = O(N^{2/3})$. Combining the bounds for I and II, we get

$$|\hat{u}_N| \le D\left(N^{-\frac{2}{3}s}\|u\|_{H^s[-1,1]} + (q_N)^{\frac{2}{3}(1-\gamma)N}\;\;\|u\|_{L^2[-1,1]}\right),$$

which proves Theorem 3.1 with $\delta = \frac{2}{3}(1-\gamma)$. □

Acknowledgments. We thank John Boyd for sharing his work [5] before publication. We would also like to thank the anonymous referees who helped improve the manuscript markedly.

REFERENCES

[1] R. Baltensperger and J.-P. Berrut, *The linear rational collocation method,* J. Comput. Appl. Math., 134 (2001), pp. 243–258.

[2] A. Bayliss and E. Turkel, *Mappings and accuracy for Chebyshev pseudo-spectral approximations,* J. Comput. Phys., 101 (1992), pp. 349–359.

[3] G. Beylkin and K. Sandberg, *Wave propagation using bases for bandlimited functions,* Wave Motion, 41 (2005), pp. 263–291.

[4] J. P. Boyd, *Approximation of an analytic function on a finite real interval by a bandlimited function and conjectures on properties of prolate spheroidal functions,* Appl. Comput. Harmon. Anal., 15 (2003), pp. 168–176.

[5] J. P. Boyd, *Prolate spheroidal wave functions as an alternative to Chebyshev and Legendre polynomials for spectral element and pseudospectral algorithms,* J. Comput. Phys., 199 (2004), pp. 688–716.

[6] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics,* Springer-Verlag, Berlin, 1988.

[7] M. H. Carpenter and D. Gottlieb, *Spectral methods on arbitrary grids,* J. Comput. Phys., 129 (1996), pp. 74–86.

[8] H. Cheng, V. Rokhlin, and N. Yarvin, *Nonlinear optimization, quadrature, and interpolation,* SIAM J. Optim., 9 (1999), pp. 901–923.

[9] A. Ditkowski, K. Dridi, and J. S. Hesthaven, *Convergent Cartesian grid methods for Maxwell's equations in complex geometries,* J. Comput. Phys., 170 (2001), pp. 39–80.

[10] W. S. Don and A. Solomonoff, *Accuracy enhancement for higher derivative using Chebyshev collocation and a mapping technique,* SIAM J. Sci. Comput., 18 (1997), pp. 1040–1055.

[11] D. Funaro, *Polynomial Approximation of Differential Equations,* Lecture Notes in Phys. 8, Springer-Verlag, Berlin, 1992.

[12] D. Funaro and D. Gottlieb, *A new method of imposing boundary conditions in pseudospectral approximations of hyperbolic equations,* Math. Comp., 51 (1988), pp. 599–613.

[13] D. Gottlieb, M. Y. Hussaini, and S. A. Orszag, *Theory and application of spectral methods,* in Spectral Methods for Partial Differential Equations, R. Voigt, D. Gottlieb, and M. Y. Hussaini, eds., SIAM, Philadelphia, PA, 1984, pp. 1–54.

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

[15] J. S. Hesthaven, *Spectral penalty methods,* Appl. Numer. Math., 33 (2000), pp. 23–41.

[16] J. S. Hesthaven, P. G. Dinsen, and J. P. Lynov, *Spectral collocation time-domain modeling of diffractive optical elements,* J. Comput. Phys., 155 (1999), pp. 287–306.

[17] S. Karlin and W. Studden, *Tchebysheff Systems with Applications in Analysis and Statistics,* Interscience, New York, 1966.

[18] D. Kosloff and H. Tal-Ezer, *A modified Chebyshev pseudospectral method with an $O(N^{-1})$ time step restriction,* J. Comput. Phys., 104 (1993), pp. 457–469.

[19] H. J. Landau and H. Widom, *Eigenvalue distribution of time and frequency limiting,* J. Math. Anal. Appl., 77 (1980), pp. 468–491.

[20] J. L. Mead and R. A. Renaut, *Accuracy, resolution, and stability properties of a modified Chebyshev method,* SIAM J. Sci. Comput., 24 (2002), pp. 143–160.

[21] S. C. Reddy, J. A. C. Weideman, and G. F. Norris, *On a Modified Chebyshev Pseudospectral Method,* Report, Oregon State University, 1999.

[22] D. Slepian and H. O. Pollak, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty.* I, Bell Syst. Tech. J., 40 (1961), pp. 43–64.

[23] D. Slepian, *Prolate spheroidal wave functions, Fourier analysis, and uncertainty.* IV: *Extensions to many dimensions, generalized prolate spheroidal wave functions,* Bell Syst. Tech. J., 43 (1964), pp. 3009–3058.

[24] L. N. Trefethen and M. R. Trummer, *An instability phenomenon in spectral methods,* SIAM J. Numer. Anal., 24 (1987), pp. 1008–1023.

[25] H. Xiao, V. Rokhlin, and N. Yarvin, *Prolate spheroidal wave functions, quadrature and interpolation,* Inverse Problems, 17 (2001), pp. 805–838.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 1934–1949

OPTIMAL SUPERCONVERGENCE ORDERS OF ITERATED COLLOCATION SOLUTIONS FOR VOLTERRA INTEGRAL EQUATIONS WITH VANISHING DELAYS*

HERMANN BRUNNER[†] AND QIYA HU[‡]

Abstract. In this paper we analyze the optimal convergence properties of collocation approximations to solutions of Volterra integral equations of the second kind with vanishing variable delays. The focus of the analysis is on the superconvergence of the (iterated) collocation approximation corresponding to collocation in the space of (discontinuous) piecewise polynomials of degree $m-1 \geq 0$. We show that on uniform meshes the iterated collocation solution possesses the global superconvergence order $m+1$, and that the solution's order of local superconvergence at the nodes of the underlying mesh cannot exceed $m+2$, in sharp contrast to problems with nonvanishing delays. Moreover, the optimal order $p^* = m+2$ can be attained only under suitable assumptions. This result also resolves a conjecture of 1997 regarding the attainable order of local superconvergence for Volterra integral equations containing a proportional delay qt with $q \in (0,1)$.

Key words. delay integral equation, proportional delays, collocation method, iterated collocation solution, optimal order of superconvergence

AMS subject classification. 65R20

DOI. 10.1137/040615705

1. Introduction.

In this paper we analyze the optimal orders of global and local superconvergence of iterated collocation solutions to Volterra integral equations with vanishing variable delays,

$$y(t) = f(t) + \int_0^t k_1(t,s,y(s))ds + \int_0^{\theta(t)} k_2(t,s,y(s))ds, \quad t \in J := [0,T], \tag{1.1}$$

where $\theta(t) := t - \tau(t) \geq 0$ is such that the (continuous) delay τ satisfies $\tau(0) = 0$. (Additional assumptions are stated in section 2.1.) Equation (1.1) includes the important special case where τ is the *proportional delay* $\tau(t) = (1-q)t$ with $0 < q < 1$, i.e.,

$$y(t) = f(t) + \int_0^t k_1(t,s,y(s))ds + \int_0^{qt} k_2(t,s,y(s))ds, \quad t \in J \tag{1.2}$$

(see [8, 4]).

It is well known that for classical Volterra integral equations (corresponding to $k_2 \equiv 0$ in (1.1)) the iterated collocation approximation based on collocation in the space of (discontinuous) piecewise polynomials of degree $m-1 \geq 0$ possesses the optimal superconvergence order $p^* = 2m$ at the nodes of the underlying (uniform)

*Received by the editors September 24, 2004; accepted for publication (in revised form) November 24, 2004; published electronically December 8, 2005.
http://www.siam.org/journals/sinum/43-5/61570.html

[†]Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, Newfoundland, Canada A1C 5S7 (hermann@math.mun.ca). The research of this author was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery grant 9406).

[‡]Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China (hqy@lsec.cc.ac.cn). The research of this author was supported by Special Funds for Major State Basic Research Projects of China G1999032804.

mesh, provided that the collocation parameters are the m Gauss points in $(0, 1)$. For Volterra integral equations with nonvanishing delays (e.g., constant delays), this property is preserved if the mesh is suitably constrained (see [2, 3, 15, 5]).

It was shown in [4] and [23] that these superconvergence properties on uniform meshes do not carry over to (1.1) with $k_2 \not\equiv 0$ and $\tau(t) = (1-q)t$ $(0 < q < 1)$. In fact, for this proportional delay the optimal (local) superconvergence order is at most $p^* = 2m-1$ when $q \neq \frac{1}{2}$; for $q = \frac{1}{2}$ and collocation at the Gauss points it was conjectured that $p^* = 2m$.

In the present paper we investigate the attainable order of the iterated collocation approximation for the general delay integral equation (1.1). We first show that, under natural assumptions, this approximation exhibits global superconvergence of order $m+1$. We then introduce a general condition characterizing meshes for which local superconvergence at the nodes can occur for a given vanishing delay $\tau(t)$ when the collocation points are the m Gauss points. In particular, this holds when the meshes are uniform, the delay has the form $\tau(t) = (1-q)t$ with $q = \frac{1}{2}$, and m is even. This means that the conjecture mentioned above is true only when $m = 2$. Furthermore, we explain why the result no longer holds when $m \geq 3$. A key technique for proving our main result is the use of an inductive (or recursive) method, which was developed in [7], [14], and [15].

The analysis of second-kind Volterra functional integral equations with proportional delays dates back to the work of Volterra [25, pp. 92–101] and Andreoli [1]. Among the more recent contributions to this subject are the papers by Morris, Feldstein, and Bowen [18] (see pp. 518–523), Chambers [8], and Mureşan [19]; see also Brunner [5, Chap. 5]. Related functional integral equations are studied in Esser [11], Iserles and Liu [17], Piila [21], Piila and Pitkäranta [22] (application in the asymptotic membrane theory of hyperbolic shells), and Denisov and Lorenzi [10] (see [10] also for references on applications, and compare with section 6).

Fox et al. [12] were the first to investigate the numerical solution of the so-called pantograph equation,

$$y'(t) = ay(t) + by(qt), \quad t \geq 0, \quad 0 < q < 1. \tag{1.3}$$

In particular, they analyzed the numerical treatment of its integrated form (a special case of (1.2)) by a global method, the τ-method [12, pp. 292–295], and discussed the behavior of the resulting error. The 1990s brought a renewed interest by numerical analysts in (1.3) and its more general versions; see Iserles [16] for an illuminating survey of the difficulties underlying the (stability) analysis of numerical methods for pantograph-type equations. The papers by Brunner [4], Takama, Muroya, and Ishiwata [23], and Brunner, Hu, and Lin [7] focus on the question of optimal local superconvergence in (iterated) collocation solutions for (1.2). It was shown that, on uniform meshes, collocation at the Gauss points will in general not yield the classical optimal order of (local) superconvergence at the mesh points. (See also Muroya, Ishiwata, and Brunner [20].) However, the complete analysis of this problem has remained open.

This paper was above all motivated by the fact that, in order to obtain insight into the superconvergence analysis of collocation approximations for Volterra functional integral equations with *general* vanishing delays or, eventually, with *state-dependent delays*, we first need to fully understand this analysis for the pantograph integral equation (1.2) and its more general version (1.1).

2. Preliminaries.

2.1. The iterated collocation approximation. For ease of exposition, we will consider the linear version of (1.1), namely

$$y(t) = f(t) + \int_0^t K_1(t,s)y(s)ds + \int_0^{\theta(t)} K_2(t,s)y(s)ds, \quad t \in J := [0,T], \tag{2.1}$$

with $\theta(t) := t - \tau(t) \geq 0$. More precisely, we will assume that the delay τ satisfies the following conditions:

(D1) $\tau \in C^\nu(J)$ for some $\nu \geq 1$, with $\tau(0) = 0$ and $\tau(t) > 0$ for $t > 0$;

(D2) $\min_{t\in J} \theta'(t) =: q_0 > 0$.

Regularity assumptions for the given functions τ, f and K_i $(i = 1, 2)$ will be stated in Theorem 4.1 (see also [8], [4], [7], and [5]).

For given $N \in \mathbb{N}$, let $J_N := \{t_n : \; 0 = t_0 < t_1 < \cdots < t_N = T\}$ denote a mesh for the given interval J, and set $e_n := (t_{n-1}, \; t_n], \; h_n := t_n - t_{n-1}$ $(n = 1, \ldots, N)$. In the following we shall be concerned with collocation solutions u lying in the finite-dimensional collocation space

$$S_{m-1}^{(-1)}(J_N) := \{v : \; v|_{e_n} =: v_n \in \mathcal{P}_{m-1} \; (n = 1, \ldots, N)\},$$

where $m \geq 1$ and $\mathcal{P}_{m-1}$ denotes the set of (real) polynomials of degree not exceeding $m - 1$. For a given set of collocation points, $X(N) := \bigcup_{n=1}^N X_n$, with

$$X_n := \{t_{nj} := t_{n-1} + c_j h_n, \; 0 < c_1 < \cdots < c_m \leq 1 \; (n = 1, \ldots, N)\},$$

we are looking for $u \in S_{m-1}^{(-1)}(J_N)$ satisfying the collocation equations

$$u(t) = f(t) + \int_0^t K_1(t,s)u(s)ds + \int_0^{\theta(t)} K_2(t,s)u(s)ds, \quad t \in X(N). \tag{2.2}$$

The collocation equation (2.2) defines a unique approximation $u \in S_{m-1}^{(-1)}(J_N)$ whenever the mesh diameter $h := \max_{(n)} h_n$ is sufficiently small. As for classical Volterra integral equations, this approximation u will be generated recursively by successive computation of its restrictions $u_1, \ldots, u_N$ to the subintervals $e_1, \ldots, e_N$ given by the mesh J_N (compare also [6] or [5]).

When the collocation solution u is known, we obtain the iterated collocation solution u^{it} corresponding to u by setting

$$u^{it}(t) := f(t) + \int_0^t K_1(t,s)u(s)ds + \int_0^{\theta(t)} K_2(t,s)u(s)ds, \qquad t \in J. \tag{2.3}$$

Note that $u^{it}(t) = u(t)$ whenever $t \in X(N)$. We shall see that u^{it} will exhibit a higher order of convergence than u itself if the set $\{c_i\}$ is chosen judiciously (compare sections 3 and 4). However, for $m > 2$, this gain is now no longer as large as in the case of classical Volterra integral equations (Theorem 4.2).

2.2. Notation. We set $Z_N := \{t_n : 1 \leq n \leq N\}$ and define the domains

$$\Omega_1 := \{(t,s) : \; 0 \leq s \leq t \leq T\} \quad \text{and} \quad \Omega_2 := \{(t,s) : \; 0 \leq s \leq \theta(t), \; t \in J\}.$$

We introduce the linear operator $\mathcal{K} : \; L^\infty(J) \to L^\infty(J)$ by setting

$$\mathcal{K}\phi(t) := \int_0^t K_1(t,s)\phi(s)ds + \int_0^{\theta(t)} K_2(t,s)\phi(s)ds, \; t \in J.$$

For a given nonnegative integer k we define the norm $\|\cdot\|_{k,\infty}$ by

$$\|v\|_{k,\infty} := \left(\sum_{n=1}^{N} \|v\|_{k,e_n,\infty}^2\right)^{\frac{1}{2}},$$

where

$$\|v\|_{k,e_n,\infty} := \max_{0\le j\le k}\left(\sup_{t\in e_n}\left|\frac{d^j}{dt^j}v(t)\right|\right).$$

For ease of notation, the norm $\|\cdot\|_{0,e_n,\infty}$ will often be abbreviated by writing $\|\cdot\|_{e_n,\infty}$.

Finally, let $\pi_h : C(J) \to S_{m-1}^{(-1)}(J_N)$ denote the linear interpolation operator such that $\pi_h v(t_{nj}) = v(t_{nj})$ $(n = 1,\dots,N;\ j = 1,\dots,m)$ for $v \in C(J)$. It is well known (see, e.g., [14], [7]) that

$$\|\pi_h v\|_{e_n,\infty} \le C\|v\|_{e_n,\infty} \quad \text{for } v \in C(J) \tag{2.4}$$

and

$$\|(\pi_h - \mathcal{I})v\|_{j,e_n,\infty} \le Ch^{k-j}\|v\|_{k,e_n,\infty}, \quad 0 \le j \le k \le m. \tag{2.5}$$

Here, $\mathcal{I}$ denotes the identity operator.

3. A global superconvergence result. There exist a number of papers dealing with the local superconvergence properties of the iterated collocation solution u^{it} for the delay Volterra integral equations (1.1) and (1.2) at the nodes Z_N (see, for example, [4], [23], [7], and [5]). In this section, we complement these results by one on the attainable order of *global* superconvergence of u^{it} on the entire interval J. Throughout the paper C will denote a generic positive constant that is independent of N but which will depend on the length T of the interval $J = [0,T]$ and on bounds for the given functions τ, f and K_i.

THEOREM 3.1. *Let the functions f and K_i in* (2.1) *satisfy $f \in C^\nu(J)$ and $K_i \in C^\nu(\Omega_i)$ $(i = 1,2)$, and let the delay τ in $\theta(t) = t - \tau(t)$ be subject to* (D1) *and* (D2) (section 2.1))*, with $\nu \ge m+1$. If the collocation parameters $\{c_i\}$ describing the collocation points $X(N)$ satisfy the orthogonality condition*

$$\int_0^1 \prod_{i=1}^{m}(s - c_i)\, ds = 0, \tag{3.1}$$

then we have

$$\max_{t\in J}|y(t) - u^{it}(t)| \le Ch^{m+1} \quad (\text{as } h \to 0^+). \tag{3.2}$$

In the proof of Theorem 3.1 we will have to resort to the following auxiliary results whose proofs can be found in [8], [4], [5] (Lemma 3.2), [7] (Lemma 3.3), and [9, section 6] or [13, pp. 212–213] (Lemma 3.4).

LEMMA 3.2. *Let $l \ge 1$ be a given integer. Assume that the functions f, τ, and K_i in* (2.1) *satisfy f, $\tau \in C^l(J)$ and $K_i \in C^l(\Omega_i)$ $(i = 1,2)$. Then* (2.1) *has a (unique) solution $y \in C^l(J)$.*

LEMMA 3.3. *Under the assumptions stated in Theorem* 3.1 *we have*

$$\|u - y\|_{e_n,\infty} \le Ch^m\|y\|_{e_n,\infty} \tag{3.3}$$

and

$$\|u\|_{j,\infty} \le C\|y\|_{m,\infty},\ 0 \le j \le 2m. \tag{3.4}$$

The next lemma is a standard result in the superconvergence theory for ordinary differential equations and classical (regular) Fredholm and Volterra integral equations of the second kind.

LEMMA 3.4. *Let $1 \le k \le m$. Assume that the collocation parameters $\{c_i\}$ satisfy the orthogonality conditions*

$$\int_0^1 s^r \prod_{i=1}^m (s-c_i)ds = 0, \quad 0 \le r \le k-1.$$

If $\psi \in C^k(J)$ and $\varphi \in C^{m+k}(J)$, then the following estimate is valid for all e_n $(1 \le n \le N)$:

$$\left|\int_{e_n} \psi(t)(\pi - \mathcal{I})\varphi(t)dt\right| \le Ch_n^{m+k+1}\|\psi\|_{k,\infty} \cdot \|\varphi\|_{m+k,\infty}. \tag{3.5}$$

Proof of Theorem 3.1. Since $u \in S_{m-1}^{(-1)}(J_N)$, it follows from the definition of π_h that $\pi_h u = u$. Equations (2.1) and (2.2) may be written in operator form as

$$y = \mathcal{K}y + f \tag{3.6}$$

and

$$u = \pi_h\mathcal{K}u + \pi_h f, \tag{3.7}$$

respectively.

If we subtract (3.6) from (3.7) and set $e := u - y$, we obtain

$$e = \pi_h\mathcal{K}e + (\pi_h - \mathcal{I})(\mathcal{K}y + f).$$

Hence, by observing (3.6),

$$e = \pi_h\mathcal{K}e + (\pi_h - \mathcal{I})y. \tag{3.8}$$

Using an induction argument we first prove that for all $\varphi \in C^1(e_n)$,

$$\left|\int_{e_n} \varphi(s)e(s)ds\right| \le Ch^{m+2}\|\varphi\|_{1,e_n,\infty} \quad (1 \le n \le N). \tag{3.9}$$

It follows by (3.8) that for any $\psi \in C^1(e_1)$ we have

$$\left|\int_{e_1} \psi(s)e(s)ds\right| \le \int_{e_1} |\psi(s)| \cdot |\mathcal{K}e(s)|ds + \left|\int_{e_1} \psi(s)(\mathcal{I} - \pi_h)y(s)ds\right|,$$

which, together with (3.3) and (3.5), yields

$$\begin{aligned}\left|\int_{e_1} \psi(s)e(s)ds\right| \le{}& C(h_1^2\|\psi\|_{e_1,\infty} \cdot \|e\|_{e_1,\infty} \\ &+h_1^{m+2}\|\psi\|_{1,e_1,\infty} \cdot \|y\|_{m+1,e_1,\infty})\end{aligned}$$

$$\begin{aligned} &\leq Ch_1^{m+2}\|\psi\|_{1,e_1,\infty} \cdot \|y\|_{m+1,e_1,\infty} \\ &\leq Ch^{m+2}\|\psi\|_{1,e_1,\infty}. \end{aligned}$$

Here, we have used the fact that $\theta(t) = t - \tau(t) \leq t_1$ for $t \in e_1$.

Assuming that the inequality

$$\left| \int_{e_i} \varphi(s)e(s)ds \right| \leq Ch^{m+2}\|\varphi\|_{1,e_i,\infty} \tag{3.10}$$

is valid for every $\varphi \in C^1(e_i)$ $(1 \leq i \leq n)$, we need to show that, for any $\psi \in C^1(e_{n+1})$,

$$\left| \int_{e_{n+1}} \psi(s)e(s)ds \right| \leq Ch^{m+2}\|\psi\|_{1,e_{n+1},\infty}. \tag{3.11}$$

In fact, by (3.8) we have

$$\left| \int_{e_{n+1}} \psi(s)e(s)ds \right| \leq \int_{e_{n+1}} |\psi(s)| \cdot |\mathcal{K}e(s)|ds + \left| \int_{e_{n+1}} \psi(s)(\mathcal{I} - \pi_h)y(s)ds \right|. \tag{3.12}$$

It is clear that

$$\int_{e_{n+1}} |\psi(s)| \cdot |\mathcal{K}e(s)|ds \leq h_{n+1}\|\psi\|_{e_{n+1},\infty} \cdot \sup_{s\in e_{n+1}} |\mathcal{K}e(s)|. \tag{3.13}$$

For $s \in e_{n+1}$, there is an index n_s such that $\theta(s) \in (t_{n_s}, t_{n_s+1}]$. Using (3.3) and the inductive assumption (3.10), we readily derive the estimate

$$\begin{aligned} |Ke(s)| &\leq \left| \int_0^{t_n} K_1(s,\sigma)e(\sigma)d\sigma + \int_{t_n}^{s} K_1(s,\sigma)e(\sigma)d\sigma \right| \\ &\quad + \left| \int_0^{t_{n_s}} K_2(s,\sigma)e(\sigma)ds + \int_{t_{n_s}}^{\theta(s)} K_2(s,\sigma)e(\sigma)d\sigma \right| \\ &\leq C[n \cdot h^{m+2} + (s - t_n)h_{n+1}^m + n_s \cdot h^{m+2} + (\theta(s) - t_{n_s})h_{n_s+1}^m] \\ &\leq Ch^{m+1} \quad \text{for all } s \in e_{n+1}. \end{aligned} \tag{3.14}$$

Upon substitution of this estimate in (3.13) we obtain

$$\int_{e_{n+1}} |\psi(s)| \cdot |\mathcal{K}e(s)|ds \leq Ch^{m+2}\|\psi\|_{e_{n+1},\infty}.$$

This, together (3.12) and (3.5), gives (3.11). It then follows by the induction principle that the inequality (3.9) is indeed valid for all $n \leq N$.

Consider now the iterated collocation approximation u^{it}. Since (2.3) can be written as

$$u^{it} = f + \mathcal{K}u,$$

subtraction of (3.6) from this equality yields

$$e_{it} := y - u_{it} = \mathcal{K}e.$$

The estimate (3.2) then follows from (3.14) and (3.9).

4. Local superconvergence results. It is clear that the estimate (3.2) also represents the optimal local order of superconvergence result when $m = 1$, since $2m = m + 1$. Hence, we will now consider only collocation approximations $u \in S_{m-1}^{(-1)}(J_N)$ with $m \geq 2$. We will show that if collocation is based on the collocation parameters $\{c_i\}$ given by the m Gauss points in $(0, 1)$ (that is, the zeros of the shifted Legendre polynomial $P_m(2s-1)$), then the order of local superconvergence of u^{it} on Z_N cannot attain the "classical" value $p^* = 2m$ usually associated with the Gauss collocation points; we can achieve only $p^* \leq m + 2$. This implies that a key conjecture in [4] and [7] is false when $m > 2$.

Without loss of generality we may assume that $0 < \theta(t_n) \leq t_n$ for each t_n with $n \geq 1$. Thus, there is an index n' such that $\theta(t_n) \in (t_{n'}, t_{n'+1}]$. For such an index n', define $q_{n'} = (\theta(t_n) - t_{n'})/h_{n'} \in (0, 1]$. Let $L_{m+1}(s)$ be the polynomial of degree $m + 1$ defined by

$$L_{m+1}(s) = \frac{d^{m-1}}{ds^{m-1}}(s(s-1))^m, \quad s \in [0, 1].$$

It is well known that $L'_{m+1}(s) = P_m(2s-1)$.

THEOREM 4.1. *Let the functions f and K_i $(i = 1, 2)$ in* (2.1) *satisfy $f \in C^\nu(J)$, $K_i \in C^\nu(\Omega_i)$, and suppose that the delay τ is subject to the conditions* (D1) *and* (D2) *(section* 2.1*), with $\nu \geq m + 2$. Assume that the nodes $\{t_n\}$ are chosen so that*

$$|L_{m+1}(q_{n'})| \leq Ch_{n'}, \quad n = 1, \dots, N.$$

Then the iterated collocation approximation u^{it} corresponding to the collocation approximation $u \in S_{m-1}^{(-1)}(J_N)$ induces the estimate

$$\max_{t \in Z_N} |y(t) - u^{it}(t)| \leq Ch^{m+2} \quad (as\, h \to 0^+), \tag{4.1}$$

where the exponent $m + 2$ is best possible.

The proof of this key result will be given in the next section.

In the following we focus on Volterra integral equations (2.1) with *proportional delay* $\tau(t) = (1 - q)t$ $(0 < q < 1)$, corresponding to $\theta(t) = qt$.

THEOREM 4.2. *Let the functions f and K_i $(i = 1, 2)$ in* (2.1) *satisfy the assumptions stated in Theorem* 4.1. *Assume that the delay is given by $\tau(t) = (1 - q)t$, with $q = \frac{1}{2}$. If the $\{c_i\}$ are the Gauss points, then on any uniform mesh with sufficiently small diameter $h > 0$ the following is true:*

(1) *The estimate* (4.1) *is valid if and only if m is even.*

(2) *The exponent describing the convergence order in* (4.1) *cannot be replaced by $m + 3$.*

Remark 4.1. We remind the reader that if the original interval $J = [0, T]$ is replaced by $J_0 := [t_0, T]$, with $t_0 > 0$, then the delay τ is strictly positive on J. Thus, the order exponent $m + 2$ in (4.1) can be replaced, for any $m \geq 1$, by the "classical" optimal exponent $2m$ (compare [3], [5]; see also [2]).

In the proof of Theorem 4.2 we shall need the following auxiliary result.

LEMMA 4.3. *Let $k = 1$ or $k = 2$. Then for n' and $q_{n'}$ defined at the beginning of section* 4, *the estimate*

$$\left| \int_{t_{n'}}^{\theta(t_n)} \phi(s) \cdot (\mathcal{I} - \pi_h)\psi(s) ds \right| \leq Ch_{n'}^{m+k+1} \|\phi\|_{k,e_n,\infty} \cdot \|\psi\|_{m+k,e_n,\infty} \tag{4.2}$$

holds for any $\phi \in C^k(e_{n'})$ *and* $\psi \in C^{m+k}(e_{n'})$ *if and only if*

$$\left|\int_0^{q_{n'}} s^{l-1}\prod_{i=1}^{m}(s-c_i)ds\right| \le Ch_{n'}^{k+1-l}, \quad 1 \le l \le k. \tag{4.3}$$

Proof. Let $\tilde{\pi}: C[0,1] \to \mathcal{P}_m[0,1]$ denote the linear interpolation operator associated with the interpolation points $\{c_j\}$. Hence, for any $r \ge 1$ there is a polynomial $p_{r-1}(s)$ of degree $r-1$ such that

$$(I-\tilde{\pi})s^{m+r-1} = p_{r-1}(s)\cdot\prod_{j=1}^{m}(s-c_j), \quad s \in [0,1]. \tag{4.4}$$

Using (4.4) and an obvious transformation of variables, we see that the condition (4.3) is equivalent to the inequality

$$\left|\int_0^{q_{n'}} s^{r_1-1}(I-\tilde{\pi})s^{m+r_2-1}ds\right| \le Ch_{n'}^{k+2-(r_1+r_2)}, \quad r_1+r_2 \le k+1\ (r_1,\ r_2 \ge 1). \tag{4.5}$$

We will first prove that the inequality (4.4) implies the estimate (4.2). By Taylor's formula, we have

$$\phi(s) = \sum_{i=0}^{k-1}\frac{1}{i!}\phi^{(i)}(t_{n'})(s-t_{n'})^i + \frac{1}{k!}\phi^{(k)}(\xi_{n'})(s-t_{n'})^k, \quad \xi_{n'} \in (t_{n'}, s),$$

and

$$\psi(s) = \sum_{i=0}^{m+k-1}\frac{1}{i!}\psi^{(i)}(t_{n'})(s-t_{n'})^i + \frac{1}{(m+k)!}\psi^{(m+k)}(\eta_{n'})(s-t_{n'})^{m+k}, \quad \eta_{n'} \in (t_{n'}, s).$$

Noting that $(\mathcal{I}-\tilde{\pi})s^r = 0$ for $r \le m-1$, we obtain

$$\begin{aligned}
&\left|\int_{t_{n'}}^{\theta(t_n)} \phi(s)\cdot(\mathcal{I}-\pi_h)\psi(s)ds\right| \\
&\le \sum_{i=0}^{k-1}\sum_{j=m}^{m+k-1}\frac{1}{i!j!}\left|\phi^{(i)}(t_{n'})\psi^{(j)}(t_{n'})\int_{t_{n'}}^{\theta(t_n)}(s-t_{n'})^i(\mathcal{I}-\pi_h)(s-t_{n'})^j ds\right| \\
&\quad +\frac{1}{(m+k)!}\sum_{i=0}^{k-1}\frac{1}{i!}\left|\phi^{(i)}(t_{n'})\int_{t_{n'}}^{\theta(t_n)}\psi^{(m+k)}(\eta_{n'})(s-t_{n'})^{i+m+k}ds\right| \\
&\quad +\frac{1}{k!}\sum_{j=m}^{m+k-1}\frac{1}{j!}\left|\psi^{(j)}(t_{n'})\int_{t_{n'}}^{\theta(t_n)}\phi^{(k)}(\xi_{n'})(s-t_{n'})^k(\mathcal{I}-\pi_h)(s-t_{n'})^j ds\right| \\
&\quad +\frac{1}{k!(m+k)!}\left|\int_{t_{n'}}^{\theta(t_n)}\phi^{(k)}(\xi_{n'})(s-t_{n'})^k(\mathcal{I}-\pi_h)\psi^{(m+k)}(\eta_{n'})(s-t_{n'})^{m+k}ds\right| \\
&=: I_1+I_2+I_3+I_4.
\end{aligned} \tag{4.6}$$

The boundedness of the operator π_h implies that there exist constants C_j so that

$$|I_j| \le C_j h_{n'}^{m+k+1}\|\phi\|_{k,e_n,\infty}\cdot\|\psi\|_{m+k,e_n,\infty} \ (j=2,3,4). \tag{4.7}$$

The term I_1 can be estimated by resorting to (4.5): we find that

$$\begin{aligned}|I_1| &\le \sum_{i=0}^{k-1}\sum_{j=m}^{m+k-1}\frac{1}{i!j!}h_{n'}^{i+j+1}\left|\phi^{(i)}(t_{n'})\psi^{(j)}(t_{n'})\int_0^{q_{n'}}s^i(\mathcal{I}-\tilde{\pi})s^j ds\right|\\ &\le C_1\sum_{i+j<m+k}h_{n'}^{i+j+1}\left|\int_0^{q_{n'}}s^i(\mathcal{I}-\tilde{\pi})s^j ds\right|\cdot\|\phi\|_{k-1,e_n,\infty}\cdot\|\psi\|_{m+k-1,e_n,\infty}\\ &\quad +C_1\sum_{i+j\ge m+k}h_{n'}^{i+j+1}\left|\int_0^{q_{n'}}s^i(\mathcal{I}-\tilde{\pi})s^j ds\right|\cdot\|\phi\|_{k-1,e_n,\infty}\cdot\|\psi\|_{m+k-1,e_n,\infty}\\ &\le C_1 h_{n'}^{m+k+1}\|\phi\|_{k-1,e_n,\infty}\cdot\|\psi\|_{m+k-1,e_n,\infty}.\end{aligned}$$

The result (4.2) readily follows by substituting these four estimates into (4.6).

We now prove the reverse conclusion, namely that (4.2) implies (4.5). For any two positive integers r_1 and r_2 satisfying $r_1+r_2\le k+1$, we have

$$\begin{aligned}&\left|\int_{t_{n'}}^{\theta(t_n)}(s-t_{n'})^{r_1-1}(\mathcal{I}-\pi_h)(s-t_{n'})^{m+r_2-1}ds\right|\\ &\qquad = h_{n'}^{m+r_1+r_2-1}\left|\int_0^{q_{n'}}s^{r_1-1}(\mathcal{I}-\tilde{\pi})s^{m+r_2-1}ds\right|\end{aligned}$$

or

$$\begin{aligned}&\left|\int_0^{q_{n'}}s^{r_1-1}(\mathcal{I}-\tilde{\pi})s^{m+r_2-1}ds\right|\\ &\qquad = h_{n'}^{-(m+r_1+r_2-1)}\left|\int_{t_{n'}}^{\theta(t_n)}(s-t_{n'})^{r_1-1}(\mathcal{I}-\pi_h)(s-t_{n'})^{m+r_2-1}ds\right|.\end{aligned}$$

This, together with (4.2), setting $\phi(t)=(s-t_{n'})^{r_1-1}$ and $\psi(t)=(s-t_{n'})^{m+r_2-1}$, respectively, leads to (4.5).

Proof of Theorem 4.2. Since the mesh J_N is assumed to be uniform ($t_n=nh$ for $0\le n\le N$), and since $\theta(t)=qt$, we have $n'=\lfloor qn\rfloor$ (if qn is not an integer) or $n'=qn-1$ (if qn is an integer). Thus,

$$q_{n'}=qn-n'=\begin{cases} qn-\lfloor qn\rfloor & \text{if } qn \text{ is not an integer,}\\ 1 & \text{if } qn \text{ is an integer.}\end{cases}$$

In particular, for $q=\frac{1}{2}$ we obtain

$$q_{n'}=\begin{cases}\frac{1}{2} & \text{if } n \text{ is odd,}\\ 1 & \text{if } n \text{ is even.}\end{cases}$$

(1) It is well known that while the polynomial $L_{m+1}(s)$ has roots at $s=0$ and $s=1$ for every m, it has a root at $s=\frac{1}{2}$ if and only if m is even. This means that when $q=\frac{1}{2}$ and m is even, the number $q_{n'}$ is a root of the polynomial $L_{m+1}(s)$ for every n. Thus, it follows by Theorem 4.1 that the estimate (4.1) is valid when m is an even number.

To arrive at the reverse conclusion, we need only to show that the estimate

$$|e^{it}(t_1)|\le Ch^{m+2} \tag{4.8}$$

is not true when m is odd.

It follows from (3.8) that

$$e^{it}(t_1) = \mathcal{K}e(t_1) = (\mathcal{K}\pi_h\mathcal{K}e)(t_1) + (\mathcal{K}(\pi_h - \mathcal{I})y)(t_1).$$

Thus,

$$\int_0^{\theta(t_1)} (\pi_h - \mathcal{I})y(s)ds = e^{it}(t_1) - (\mathcal{K}\pi_h\mathcal{K}e)(t_1) - \int_0^{t_1} (\pi_h - \mathcal{I})y(s)ds. \tag{4.9}$$

It is easy to see (recalling the proof of Theorem 3.1) that

$$\left|(\mathcal{K}\pi_h\mathcal{K}e)(t_1) - \int_0^{t_1} (\pi_h - \mathcal{I})y(s)ds\right| \le Ch^{m+2}.$$

If the inequality (4.8) is true, it follows by (4.9) that

$$\left|\int_0^{\theta(t_1)} (\pi_h - \mathcal{I})y(s)ds\right| \le Ch^{m+2}.$$

This, together with Lemma 4.3, implies that

$$\left|\int_0^{\frac{1}{2}} \prod_{i=1}^{m} (s - c_i)ds\right| \le Ch$$

(note that $q_{1'} = \frac{1}{2}$). Thus,

$$\left|L_{m+1}\left(\frac{1}{2}\right)\right| = \left|\int_0^{\frac{1}{2}} P_m(2s-1)ds\right| = \left|\text{const} \cdot \int_0^{\frac{1}{2}} \prod_{i=1}^{m} (s - c_i)ds\right| \le Ch.$$

But $|L_{m+1}(\frac{1}{2})|$ is a positive constant when m is odd, independent of h, which contradicts the above inequality. Therefore, the inference

$$|e^{it}(t_1)| \le Ch^{m+2}$$

is false.

(2) Since the inequality (4.1) is already the optimal local superconvergence result when $m = 2$, it suffices to show that the estimate

$$|e^{it}(t_1)| \le Ch^{m+3} \tag{4.10}$$

is not valid when $m \ge 3$ is even.

Subtraction of (3.6) from (3.7) leads to

$$e = \mathcal{K}e + (\pi_h - \mathcal{I})(\mathcal{K}u + f) = \mathcal{K}e + (\pi_h - \mathcal{I})\tilde{u}. \tag{4.11}$$

Hereafter, we set $\tilde{u} := \mathcal{K}u + f$. Using the substitution technique introduced in [15], we are led to

$$e = \mathcal{K}^2 e + \mathcal{K}(\pi_h - \mathcal{I})\tilde{u} + (\pi_h - \mathcal{I})\tilde{u}.$$

Thus,

$$e^{it}(t_1) = (\mathcal{K}^3 e)(t_1) + (\mathcal{K}^2(\pi_h - \mathcal{I})\tilde{u})(t_1) + (\mathcal{K}(\pi_h - \mathcal{I})\tilde{u})(t_1). \tag{4.12}$$

It follows by (3.6)–(3.7) that

$$|(\mathcal{K}^3 e)(t_1)| \le Ch^{m+3}, \tag{4.13}$$

and hence

$$\left| \int_0^{t_1} K_1(t_1, s)(\pi_h - \mathcal{I})\tilde{u}(s)ds \right| \le Ch^{2m}. \tag{4.14}$$

Since the kernel $K_2(t,s)$ can be written as $K_2(t,s) = (t-s)\tilde{K}_2(t,s)$, where $\tilde{K}_2(t,s)$ is a (smooth) function, the use of (2.5) and (3.4)–(3.5) allows us readily to verify that

$$|(\mathcal{K}^2(\pi_h - \mathcal{I})\tilde{u})(t_1)| \le Ch^{m+3}. \tag{4.15}$$

We now deduce from (4.12) and (4.13)–(4.15) that estimate (4.10) is equivalent to

$$\left| \int_0^{\frac{1}{2}t_1} K_2(t_1, s)(\pi_h - \mathcal{I})\tilde{u}(s)ds \right| \le Ch^{m+3}. \tag{4.16}$$

It is easy to see that

$$\left| \int_0^{\frac{1}{2}} \left(s \cdot \prod_{i=1}^{m} (s - c_i) \right) ds \right| = \left| \text{const.} \int_0^{\frac{1}{2}} sP_m(2s-1)ds \right|$$

is a positive constant which cannot be bounded by Ch (as $h \to 0$). It follows by Lemma 4.3 with $k = 2$ that the inequality (4.16) does not hold. Thus, the estimate (4.10) is not valid.

Remark 4.2. The papers [4] and [7] deal with collocation methods for Volterra integral equations with proportional delays. They conjectured (based on extensive numerical experiments using $m = 2$) that when $q = \frac{1}{2}$, the optimal order of local superconvergence for the iterative collocation approximation is $2m$, provided collocation is at the Gauss points. Theorem 4.2 indicates that this conjecture is indeed true for $m = 2$ but becomes false for $m \ge 3$.

5. Proof of Theorem 4.1.

As we shall see, the proof of Theorem 4.1 has many similarities with that of Theorem 3.1 but is considerably more complex.

It suffices to prove inductively that, for all $\varphi \in C^2(J)$,

$$\left| \int_{t_{n'}}^{\theta(t_n)} \varphi(t)e(t)dt \right| \le Ch^{m+2} \|\varphi\|_{2,e_{n'},\infty} \quad (1 \le n \le N) \tag{5.1}$$

and

$$\left| \int_0^{t_n} \varphi(t)e(t)dt \right| \le Ch^{m+2} \|\varphi\|_{2,[0,t_n],\infty} \quad (1 \le n \le N). \tag{5.2}$$

To verify this we first rewrite (4.13) as

$$e(t) = \int_0^t K_1(t,s)e(s)ds + A(t), \qquad t \in J,$$

with

$$A(t) := (\pi_h - \mathcal{I})\tilde{u}(t) + \int_0^{\theta(t)} K_2(t,s)e(s)ds.$$

Let R_1 be the resolvent kernel of K_1. The classical Volterra theory (see, e.g., [6] or [5]) implies that R_1 inherits the smoothness of the kernel K_1. Thus, the solution of the above Volterra integral equation for $e(t)$ can be represented in the form

$$e(t) = A(t) + \int_0^t R_1(t,s)A(s)ds, \quad t \in J. \tag{5.3}$$

By our assumption and by Lemma 4.3 (with $k = 1$) we have

$$\left| \int_{t_{n'}}^{\theta(t_n)} \varphi(t)(\pi_h - \mathcal{I})\tilde{u}(t)dt \right| \leq Ch^{m+2}\|\varphi\|_{2,e_n,\infty}, \quad 1 \leq n \leq N \tag{5.4}$$

(recall (3.4)).

Consider first the case $n = 1$. From (5.3) we have, for any $\varphi {\in} C^2(e_1)$,

$$\begin{aligned}\int_0^{\theta(t_1)} \varphi(t)e(t)dt &= \int_0^{\theta(t_1)} \varphi(t)A(t)dt + \int_0^{\theta(t_1)} \left(\varphi(t)\int_0^t R_1(t,s)A(s)ds\right) dt \\ &= \int_0^{\theta(t_1)} \varphi(t)(\pi_h - \mathcal{I})\tilde{u}(t)dt + \int_0^{\theta(t_1)} \left(\varphi(t)\int_0^{\theta(t)} K_2(t,s)e(s)ds\right) dt \\ &\quad + \int_0^{\theta(t_1)} \left(\varphi(t)\int_0^t R_1(t,s)(\pi_h - \mathcal{I})\tilde{u}(s)ds\right) dt \\ &\quad + \int_0^{\theta(t_1)} \varphi(t)\int_0^t \left(R_1(t,s)\int_0^{\theta(s)} K_2(s,\sigma)e(\sigma)d\sigma\right) ds\, dt.\end{aligned}$$

This, together with (2.5), (3.3), and (5.4), leads to (5.1) with $n = 1$, since

$$\left| \int_0^{\theta(t_1)} \varphi(t)e(t)dt \right| \leq Ch^{m+2}\|\varphi\|_{2,e_1,\infty}. \tag{5.5}$$

In an analogous way we can now verify that (5.2) is valid for $n = 1$, by using (3.5) and (5.5).

Assume then that the inequalities

$$\left| \int_{t_{n'}}^{\theta(t_n)} \varphi(t)e(t)dt \right| \leq Ch^{m+2}\|\varphi\|_{2,e_{n'},\infty} \tag{5.6}$$

and

$$\left| \int_0^{t_n} \varphi(t)e(t)dt \right| \leq Ch^{m+2}\|\varphi\|_{2,[0,t_n],\infty} \tag{5.7}$$

hold for every $\varphi \in C^2[0,t_n]$ and for $1 \leq n < N$. We need to prove that, for any $\varphi \in C^2[0,t_{n+1}]$,

$$\left| \int_{t_{(n+1)'}}^{\theta(t_{n+1})} \varphi(t)e(t)dt \right| \leq Ch^{m+2}\|\varphi\|_{2,e_{(n+1)'},\infty} \tag{5.8}$$

and

$$\left|\int_0^{t_{n+1}} \varphi(t)e(t)dt\right| \le Ch^{m+2}\|\varphi\|_{2,[0,t_{n+1}],\infty}. \tag{5.9}$$

For any $\varphi \in C^2(e_{n+1})$, we have

$$\begin{aligned}\left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \varphi(t)e(t)dt\right| &= \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \varphi(t)A(t)dt + \int_{t_{(n+1)'}}^{\theta(t_{n+1})} \left(\varphi(t)\int_0^t R_1(t,s)A(s)ds\right)dt\right| \\ &\le \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \varphi(t)(\pi_h - \mathcal{I})\tilde{u}(t)dt\right| \\ &\quad + \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \left(\varphi(t)\int_0^{\theta(t)} K_2(t,s)e(s)ds\right)dt\right| \\ &\quad + \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \left(\varphi(t)\int_0^t R_1(t,s)(\pi_h - \mathcal{I})\tilde{u}(s)ds\right)dt\right| \\ &\quad + \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \varphi(t)\int_0^t \left(R_1(t,s)\int_0^{\theta(s)} K_2(s,\sigma)e(\sigma)d\sigma\right)ds\,dt\right| \\ &=: I_1 + I_2 + I_3 + I_4.\end{aligned} \tag{5.10}$$

By (5.4) and (3.4) we readily obtain

$$I_1 \le Ch^{m+2}\|\varphi\|_{2,e_{(n+1)'},\infty}. \tag{5.11}$$

It follows by (2.5) and (3.5) that

$$\begin{aligned}I_3 &\le \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \left(\varphi(t)\int_0^{t_{(n+1)'}} R_1(t,s)(\pi_h - \mathcal{I})\tilde{u}(s)ds\right)dt\right| \\ &\quad + \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \left(\varphi(t)\int_{t_{(n+1)'}}^t R_1(t,s)(\pi_h - \mathcal{I})\tilde{u}(s)ds\right)dt\right| \\ &\le Ch^{m+2}\|\varphi\|_{2,e_{(n+1)'},\infty}.\end{aligned} \tag{5.12}$$

Here, we have used the relation

$$t - t_{(n+1)'} \le \theta(t_{n+1}) - t_{(n+1)'} \le h_{(n+1)'}, \quad t_{(n+1)'} \le t \le \theta(t_{n+1}).$$

The estimate of I_2 is similar to the one for I_3. It is clear that

$$\begin{aligned}I_2 &\le \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \left(\varphi(t)\int_{\theta(t_{(n+1)'})}^{\theta(t)} K_2(t,s)e(s)ds\right)dt\right| \\ &\quad + \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})} \left(\varphi(t)\int_0^{\theta(t_{(n+1)'})} K_2(t,s)e(s)ds\right)dt\right|.\end{aligned} \tag{5.13}$$

Since $\theta \in C^1(J)$, there is a constant γ such that

$$|\theta(t) - \theta(t_{(n+1)'})| \le \gamma|t - t_{(n+1)'}| \le \gamma h_{(n+1)'}, \quad t_{(n+1)'} \le t \le \theta(t_{n+1}).$$

It follows by (3.3) that

$$\left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})}\left(\varphi(t)\int_{\theta(t_{(n+1)'})}^{\theta(t)}K_2(t,s)e(s)ds\right)dt\right| \le Ch^{m+2}\|\varphi\|_{2,e_{(n+1)'},\infty}. \tag{5.14}$$

There is an index n'' such that $\theta(t_{(n+1)'}) \in (t_{n''}, t_{n''}+1]$. Thus,

$$\begin{aligned}
&\left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})}\left(\varphi(t)\int_{0}^{\theta(t_{(n+1)'})}K_2(t,s)e(s)ds\right)dt\right| \\
&\qquad\le \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})}\left(\varphi(t)\int_{t_{n''}}^{\theta(t_{(n+1)'})}K_2(t,s)e(s)ds\right)dt\right| \\
&\qquad\quad+ \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})}\left(\varphi(t)\int_{0}^{t_{n''}}K_2(t,s)e(s)ds\right)dt\right|.
\end{aligned}$$

If we now use (5.4) and (5.7) and observe that $(n+1)' \le n$, we find the estimate

$$\left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})}\left(\varphi(t)\int_{0}^{\theta(t_{(n+1)'})}K_2(t,s)e(s)ds\right)dt\right| \le Ch^{m+3}\|\varphi\|_{2,e_{(n+1)'},\infty}.$$

This result and (5.14) and (5.13) lead to

$$I_2 \le Ch^{m+3}\|\varphi\|_{2,e_{(n+1)'},\infty}. \tag{5.15}$$

Analogous to (5.15) we can also verify that

$$I_4 \le Ch^{m+3}\|\varphi\|_{2,e_{(n+1)'},\infty}.$$

Substitution of (5.11)–(5.12) and (5.15) and using the above estimate in (5.10) gives (5.8). Now consider (5.9): For any $\varphi \in C^2[0,t_{n+1}]$, we have

$$\begin{aligned}
\left|\int_0^{t_{n+1}}\varphi(t)e(t)dt\right| &= \left|\int_0^{t_{n+1}}\varphi(t)A(t)dt + \int_0^{t_{n+1}}\left(\varphi(t)\int_0^t R_1(t,s)A(s)ds\right)dt\right| \\
&\le \left|\int_0^{t_{n+1}}\varphi(t)(\pi_h-\mathcal{I})\tilde{u}(t)dt\right| \\
&\quad+\left|\int_0^{t_{n+1}}\left(\varphi(t)\int_0^{\theta(t)}K_2(t,s)e(s)ds\right)dt\right| \\
&\quad+\left|\int_0^{t_{n+1}}\left(\varphi(t)\int_0^{t}R_1(t,s)(\pi_h-\mathcal{I})\tilde{u}(s)ds\right)dt\right| \qquad (5.16)\\
&\quad+\left|\int_0^{t_{n+1}}\varphi(t)\int_0^{t}\left(R_1(t,s)\int_0^{\theta(s)}K_2(s,\sigma)e(\sigma)d\sigma\right)ds\,dt\right| \\
&=: I_1' + I_2' + I_3' + I_4'.
\end{aligned}$$

It follows by (3.5) that

$$I_1' \le Ch_{n+1}h^{m+2}\|\varphi\|_{2,[0,t_{n+1}],\infty}. \tag{5.17}$$

If we change the order of integration in I_3' (compare also [7]) and use (3.5), we obtain

$$I_3' \le Ch^{m+2}\|\varphi\|_{2,e_{(n+1)'},\infty}. \tag{5.18}$$

The estimate of I_2' is similar to that of I_3', but we now need to use the assumption on θ. By changing the order of integration, and noting that $\theta(t)$ is an increasing function, we are led to

$$\begin{aligned} I_2' &= \left|\int_0^{\theta(t_{n+1})}\left(\int_{\theta^{-1}(s)}^{t_{n+1}}\varphi(t)K_2(t,s)dt\cdot e(s)\right)ds\right| \\ &\le \left|\int_0^{t_{(n+1)'}}\left(\int_{\theta^{-1}(s)}^{t_{n+1}}\varphi(t)K_2(t,s)dt\cdot e(s)\right)ds\right| \\ &\quad + \left|\int_{t_{(n+1)'}}^{\theta(t_{n+1})}\left(\int_{\theta^{-1}(s)}^{t_{n+1}}\varphi(t)K_2(t,s)dt\cdot e(s)\right)ds\right|. \end{aligned}$$

This, together with (5.7) and (5.8), leads to

$$I_2' \le Ch^{m+2}\|\varphi\|_{2,[0,t_{n+1}],\infty} \tag{5.19}$$

(note that $\theta^{-1}\in C^2(J)$). Similarly, it is easy to verify that

$$I_4' \le Ch^{m+2}\|\varphi\|_{2,[0,t_{n+1}],\infty}.$$

Substitution of (5.17)–(5.19) and the above inequality into (5.16) yields the estimate (5.9).

Finally, it follows by the induction principle that the estimates (5.1)–(5.2) hold; hence, employing (5.1)–(5.2) in the relation

$$e_{it} = u_{it} - y = \mathcal{K}e$$

produces the desired result.

6. Extensions and open problems. The techniques underlying the proofs of the superconvergence results in the previous sections can be adapted to obtain analogous results for continuous piecewise polynomial collocation approximations to the solution of the Volterra integrodifferential equation

$$y'(t) = a(t)y(t)+b(t)y(\theta(t))+\int_0^t K_1(t,s)y(s)ds+\int_0^{\theta(t)} K_2(t,s)y(s)ds, \quad t\in J, \tag{6.1}$$

with $y(0)=y_0$ and with vanishing delay θ satisfying conditions (D1) and (D2). The corresponding superconvergence results will then also apply to an important special case of (6.1), namely the pantograph equation (1.3) (for which no superconvergence results have yet been proved). The detailed analysis will be given elsewhere.

There is, however, a class of more general Volterra functional integral equations, with a representative example given by

$$y(t) = a(t)y(pt) + f(t) + \int_{qt}^t K(t,s)y(s)ds, \quad t\in[0,T], \tag{6.2}$$

with $0<p,q<1$ (see Volterra [24], Denisov and Lorenzi [10]), for which the superconvergence analysis of (iterated) collocation solutions is not yet understood. One of the inherent difficulties lies in the fact that even the existence and uniqueness theory for the exact solution to (6.2) and its collocation approximation is no longer an elementary problem.

Acknowledgment. The authors are grateful to the three referees for their valuable comments and suggestions, which led to a considerably improved version of the paper.

REFERENCES

[1] G. Andreoli, *Sulle equazioni integrali*, Rend. Circ. Mat. Palermo, 37 (1914), pp. 76–112.

[2] A. Bellen, *One-step collocation for delay differential equations*, J. Comput. Appl. Math., 10 (1984), pp. 275–283.

[3] H. Brunner, *Iterated collocation methods for Volterra integral equations with delay arguments*, Math. Comp., 62 (1994), pp. 581–599.

[4] H. Brunner, *On the discretization of differential and Volterra integral equations with variable delay*, BIT, 37 (1997), pp. 1–12.

[5] H. Brunner, *Collocation Methods for Volterra Integral and Related Functional Differential Equations*, Cambridge University Press, Cambridge, 2004.

[6] H. Brunner and P. J. van der Houwen, *The Numerical Solution of Volterra Equations*, CWI Monographs 3, North-Holland, Amsterdam, 1986.

[7] H. Brunner, Q. Hu, and Q. Lin, *Geometric meshes in collocation methods for Volterra integral equations with proportional delays*, IMA J. Numer. Anal., 21 (2001), pp. 783–798.

[8] Ll. G. Chambers, *Some properties of the functional equation* $\phi(x) = f(x)+\int_0^{\lambda x} g(x, y, \phi(y))dy$, Internat. J. Math. Math. Sci., 14 (1990), pp. 27–44.

[9] F. Chatelin and R. Lebbar, *Superconvergence results for the iterated projection method applied to a Fredholm integral equation of the second kind and the corresponding eigenvalue problem*, J. Integral Equations, 6 (1984), pp. 71–91.

[10] A. M. Denisov and A. Lorenzi, *Existence results and regularization techniques for severely ill-posed integrofunctional equations*, Boll. Un. Mat. Ital. B (7), 11 (1997), pp. 713–732.

[11] R. Esser, *Numerische Behandlung einer Volterraschen Integralgleichung*, Computing, 19 (1978), pp. 269–284.

[12] L. Fox, D. F. Mayers, J. R. Ockendon, and A. B. Tayler, *On a functional differential equation*, J. Inst. Math. Appl., 8 (1971), pp. 271–307.

[13] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I: *Nonstiff Problems*, 2nd ed., Springer-Verlag, Berlin, 1993.

[14] Q. Hu, *Interpolation correction for collocation solutions of Fredholm integro-differential equations*, Math. Comp., 67 (1998), pp. 987–999.

[15] Q. Hu, *Multilevel correction for discrete collocation solutions of Volterra integral equations with delay arguments*, Appl. Numer. Math., 31 (1999), pp. 159–171.

[16] A. Iserles, *Numerical analysis of delay differential equations with variable delays*, Ann. Numer. Math., 1 (1994), pp. 133–152.

[17] A. Iserles and Y. Liu, *On pantograph integro-differential equations*, J. Integral Equations Appl., 6 (1994), pp. 213–237.

[18] G. R. Morris, A. Feldstein, and E. W. Bowen, *The Phragmén-Lindelöf principle and a class of functional differential equations*, in Ordinary Differential Equations (Proc. Conf., Math. Res. Center, Naval Res. Lab., Washington, DC, 1971), L. Weiss, ed., Academic Press, New York, 1972, pp. 513–540.

[19] V. Mureşan, *On a class of Volterra integral equations with deviating argument*, Studia Univ. Babeş-Bolyai Math., 44 (1999), pp. 47–54.

[20] Y. Muroya, E. Ishiwata, and H. Brunner, *On the attainable order of collocation methods for pantograph integro-differential equations*, J. Comput. Appl. Math., 152 (2003), pp. 347–366.

[21] J. Piila, *Characterization of the membrane theory of a clamped shell. The hyperbolic case*, Math. Methods Appl. Sci., 6 (1996), pp. 169–194.

[22] J. Piila and J. Pitkäranta, *On the integral equation* $f(x) - (c/L(x)) \int_{L(x)}^{x} f(y)dy = g(x)$, *where* $L(x) = \min\{ax, 1\}$, $a > 1$, J. Integral Equations Appl., 8 (1996), pp. 363–378.

[23] N. Takama, Y. Muroya, and E. Ishiwata, *On the attainable order of collocation methods for delay differential equation with proportional delay*, BIT, 40 (2000), pp. 374–394.

[24] V. Volterra, *Sopra alcune questioni di inversione di integrali definite*, Ann. Mat. Pura Appl. (2), 25 (1897), pp. 139–178.

[25] V. Volterra, *Leçons sur les équations intégrales et les équations intégro-différentielles*, Gauthier-Villars, Paris, 1913.

SIAM J. Numer. Anal.
Vol. 43, No. 5, pp. 1950–1968

THE EFFECTIVE STABILITY OF ADAPTIVE TIMESTEPPING ODE SOLVERS[*]

HARBIR LAMBA[†]

Abstract. We consider the behavior of certain adaptive timestepping methods, based upon embedded explicit Runge–Kutta pairs, when applied to dissipative ODEs. It has been observed numerically that the standard local error controls can impart desirable stability properties, but this has been rigorously verified only for very special, low-order, Runge–Kutta pairs.

The rooted-tree expansion of a certain quadratic form, central to the stability theory of Runge–Kutta methods, is derived. This, together with key assumptions on the sequence of accepted timesteps and the local error estimate, provides a general explanation for the observed stability of such algorithms on dissipative problems. Under these assumptions, which are expected to hold for "typical" numerical trajectories, two different results are proved. First, for a large class of embedded Runge–Kutta pairs of order $(1, 2)$, controlled on an error-per-unit-step basis, all such numerical trajectories will eventually enter a particular bounded set. This occurs for sufficiently small tolerances independent of the initial conditions. Second, for pairs of arbitrary orders $(p-1, p)$, operating under either error-per-step or error-per-unit-step control, similar results are obtained when an additional structural assumption (that should be valid for many cases of interest) is imposed on the dissipative vector field. Numerical results support both the analysis and the assumptions made.

Key words. error control, stability, numerical integration, ordinary differential equations

AMS subject classifications. 65L06, 65L20

DOI. 10.1137/S0036142903435648

1. Introduction.

We consider adaptive timestepping ODE solvers applied to initial value problems for an autonomous system of ODEs,

$$\frac{du}{dt} = f(u), \quad u(0) = U, \tag{1.1}$$

where $u(t) \in \mathbb{R}^m$. Furthermore, the Lipschitz continuous vector field f satisfies the following structural assumption:

$$\text{(D)} \qquad \exists \alpha \geq 0,\ \beta > 0 : \ \forall u \in \mathbb{R}^m, \quad \langle f(u), u\rangle \ \leq \alpha - \beta\|u\|^2,$$

where the norm $\|\cdot\|$ is induced by the inner product $\langle\cdot,\cdot\rangle$.

A bounded closed set $\mathcal{B}$ is a *bounded absorbing set* for (1.1) if $\forall U \in \mathbb{R}^m$, $\exists t^* = t^*(U)$ such that $u(t) \in \mathcal{B}\ \forall t \geq t^*$. If a bounded absorbing set exists, then (1.1) is termed *dissipative*. Under the structural assumption (D), (1.1) is dissipative as stated in the following theorem [16].

Theorem 1.1. *Let* $\overline{B}(v, r)$ *be the closed ball with center* v*, radius* r *using the norm* $\|\cdot\|$. *Then assumption* (D) *implies the existence of bounded absorbing sets* $\mathcal{B} = \overline{B}(0, \sqrt{(\alpha+\epsilon)/\beta})\ \forall\epsilon > 0$.

The structural assumption (D) has played an important role in nonlinear stability theory, where the aim is to find conditions under which numerical schemes, when regarded as discrete dynamical systems, preserve various qualitative asymptotic features of the original ODE (such as the existence of bounded absorbing sets). However, the

[*] Received by the editors September 27, 2003; accepted for publication (in revised form) April 25, 2005; published electronically December 8, 2005.
http://www.siam.org/journals/sinum/43-5/43564.html

[†] Department of Mathematical Sciences, George Mason University, MS 3F2, 4400 University Drive, Fairfax, VA 22030 (hlamba@gmu.edu).

vast majority of this body of work only applies to methods employing a fixed timestep, whereas most algorithms used in practice allow the timesteps to change from one step to the next. In the algorithms considered here, the timesteps are chosen so as to control an estimate of the local (one-step) error and this adaptive timestepping approach can result in extremely impressive efficiency gains.

Even though the standard error controls were not designed with stability in mind, it has been observed that such adaptive timestepping algorithms often have much better stability properties than their fixed-timestepping counterparts. This paper addresses the questions of when, and how, the stepsizes induced by the local error control will confer desirable stability properties upon the adaptive numerical method.

As mentioned above, there have been many investigations into the stability properties of Runge–Kutta methods with a fixed timestep, under various structural assumptions (e.g., [2, 1, 5, 9, 18]). We now provide a brief outline of the relevant results for dissipativity. Consider a general (implicit or explicit) s-stage Runge–Kutta scheme for (1.1) with timestep h,

$$\eta_i = U_n + h\sum_{j=1}^{s} a_{ij} f(\eta_j), \quad i = 1, \ldots, s, \tag{1.2}$$

$$U_{n+1} = U_n + h\sum_{i=1}^{s} b_i f(\eta_i), \tag{1.3}$$

and define the vector $b = (b_1, \ldots, b_s)^T$ and matrices A and B by $A(i,j) = a_{ij}$ and $B = \text{diag}(b)$.

Equations (1.2) and (1.3), after standard manipulations (see, for example, [18]), imply that

$$\|U_{n+1}\|^2 = \|U_n\|^2 + 2h\sum_{i=1}^{s} b_i\langle \eta_i, f(\eta_i)\rangle - h^2 \sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle,$$

where $m_{ij} = M(i,j)$ with $M = BA + A^T B - b^T b$. Under the structural assumption (D), with B positive semidefinite, and using the same norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$, we obtain

$$\|U_{n+1}\|^2 \leq \|U_n\|^2 + 2h\sum_{i=1}^{s} b_i(\alpha - \beta\|\eta_i\|^2) - h^2 \sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle. \tag{1.4}$$

The Runge–Kutta method is termed *algebraically stable* if the matrices M and B are both positive semidefinite. The condition on M ensures that the quadratic form

$$h^2 \sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle \tag{1.5}$$

is nonnegative, while the condition on B is used to show that the quantities $b_i(\alpha - \beta\|\eta_i\|^2)$ are negative outside a ball of sufficiently large radius in the norm $\|\cdot\|$. Together these imply the existence of bounded absorbing sets for the discrete dynamical system defined by the numerical scheme $\forall h > 0$. Thus the algebraic stability of the numerical scheme ensures that the property of dissipativity is transferred to the numerical approximation. However, M cannot be positive semidefinite for explicit Runge–Kutta methods and so all algebraically stable methods are necessarily implicit. Indeed, explicit Runge–Kutta methods using a fixed timestep often have very poor stability properties.

We now return to our discussion of adaptive schemes. While the quadratic form (1.5) cannot be forced to be nonnegative for a nonalgebraically stable method, we shall show that, under certain conditions, the constraints imposed upon the timestep sizes by the local error control will also effectively bound the *magnitude* of (1.5) (as opposed to its sign). Then, for a Runge–Kutta method with B positive definite, outside a ball of sufficiently large radius the single-summation term in (1.4) will be shown to dominate and the norm of the numerical solution to decrease. This idea underlies the approach introduced in this paper.

The class of adaptive schemes that will be analyzed is now defined. We set $U_0 = U$ and iteratively generate U_{n+1} from U_n using a timestep h_n. The equations defining a general embedded explicit Runge–Kutta pair with s stages are

$$\eta_i = U_n + h_n \sum_{j=1}^{s} a_{ij} f(\eta_j), \quad i = 1, \dots, s, \tag{1.6}$$

$$V_{n+1} = U_n + h_n \sum_{i=1}^{s} b_i f(\eta_i), \tag{1.7}$$

$$W_{n+1} = U_n + h_n \sum_{i=1}^{s} \overline{b}_i f(\eta_i). \tag{1.8}$$

Such a Runge–Kutta pair, with orders $p-1$ and p, will be referred to as a $(p-1,p)$ pair. We shall assume that the higher-order method is represented by the weights $b_1, \dots, b_s$ and the lower-order method by $\overline{b}_1, \dots, \overline{b}_s$. Thus $U_{n+1} = V_{n+1}$ when the higher-order method is used to advance the solution (extrapolation mode) and $U_{n+1} = W_{n+1}$ otherwise (nonextrapolation mode). To complement the definitions of A, B, and b, let $\overline{b} = (\overline{b}_1, \dots, \overline{b}_s)^T$ and $\overline{B} = \operatorname{diag}(\overline{b})$.

The *local error estimate* $E(U_n, h_n)$ is defined as the difference between the two approximations,

$$E(U_n, h_n) := W_{n+1} - V_{n+1}.$$

The user defines a tolerance τ, and the timesteps must satisfy the following standard local error control:

$$\|E(U_n, h_n)\| \le \sigma(\tau, U_n) h_n^{\rho}, \tag{1.9}$$

where $\rho = 0$ for error-per-step control and $\rho = 1$ for error-per-unit-step control. The quantity $\sigma(\tau, U_n)$ is a quantity closely related to the tolerance τ, and indeed may simply be equal to τ. However, we wish to allow for the possibility of absolute, relative, and mixed error controls. There are various ways in which this can be done but for simplicity we shall require only that there exists some constant $C_1 > 0$ such that

$$\sigma(\tau, u) \le C_1 \tau \|u\| \quad \forall u \in \mathbb{R}^m. \tag{1.10}$$

It should be noted that absolute or mixed error controls will need to be modified on some neighborhood of the origin in order to satisfy (1.10). However, we will be concerned exclusively with trajectories that lie entirely outside of (large) balls centered upon the origin and the choice of (1.10) will help to streamline the analysis. Also, the norms used in (1.9) and (1.10) and throughout the rest of the paper are the same as in (D).

We thus have four possible modes of operation depending upon the choice of solution-advancing method and type of error control. EPS and EPUS will denote error-per-step and error-per-unit step modes, respectively, in nonextrapolation mode, while XEPS and XEPUS are their extrapolation counterparts. We also assume the existence of a maximum timestep $h_{\max}$, independent of τ, which is a very common feature of adaptive algorithms. Throughout, we assume that the vector field f is sufficiently smooth on $\mathbb{R}^m$. These smoothness requirements are determined only by the order of the Runge–Kutta methods used to form the local error estimate. Note that no further details of the algorithm need to be specified, in particular, the way in which candidate timesteps are generated. All that is required is that the error control (1.9) is satisfied at every timestep.

While no explicit Runge–Kutta method can be algebraically stable, it has been observed [6, 18] that adaptive timestepping methods based upon explicit schemes do, for certain combinations of dissipative test problems and mode of operation, seem to have some desirable stability properties (see also [13] for a discussion of stability with regard to the existence of spurious fixed points). In particular, the numerical schemes appear to be dissipative. Of course, no amount of numerical testing can prove the existence of a bounded absorbing set for all initial data but the results do suggest that, with an extremely high degree of certainty, numerical trajectories enter and then remain within an "absorbing set" close to $\overline{B}(0, \sqrt{\alpha/\beta})$.

There have been previous analyses of the behavior of adaptive methods on dissipative ODEs that have attempted to explain this phenomenon. In [17], it was proved that very special, embedded explicit Runge–Kutta pairs generate a solution that, at each step, is a small perturbation of the solution generated by using a corresponding (implicit) algebraically stable method. In this way, the stability characteristics of this related scheme are transferred to the explicit pair. Such pairs were termed *essentially algebraically stable* and an order barrier for $(p-1, p)$ pairs, namely, that $p \leq 5$, was proved. For these pairs, applied to ODEs satisfying (D), under no additional assumptions and with an absolute error control, two different results were proved. The first, which is a discrete analogue of Theorem 1.1, stated that when such a pair is used in EPUS or XEPUS modes the numerical scheme has a bounded absorbing set for all sufficiently small tolerances τ, *independent of the initial data*. The second result, which is significantly weaker, states that for the same pairs operating in EPS or XEPS modes each numerical trajectory will again eventually enter a particular bounded absorbing set, but now the required tolerance does depend upon the initial data.

The independence of τ with respect to initial conditions is desirable, not just from a computational point of view, but also from a theoretical one, since it allows us to consider the numerical method, for a fixed sufficiently small tolerance, as a dynamical system with similar asymptotic behavior to the underlying ODE for all initial conditions. However, the set of essentially algebraically stable pairs forms a very small subset of pairs currently employed and are necessarily of low-order.

A second analysis [8] took a different approach. There it was assumed, for a general adaptive method under EPUS control, that the *actual* one-step truncation errors $T(U_n, h_n)$ (rather than the one-step error estimates) were correctly controlled at every step, in particular, that

$$T(U_n, h_n) \leq K(U)\tau h_n$$

occurred at every timestep for some constant $K(U)$. Using this assumption that the error control works correctly, positive stability results were proved for general adaptive schemes but only in the much weaker sense that the required tolerance depended upon the initial data.

Any stability properties introduced to an explicit Runge–Kutta method via a local error control are due to the size of the accepted steps. However, neither of the analyses described above explicitly considers the actual timestep sequences generated by the method (and they also only considered the case of absolute error control). In order to obtain tighter and/or more general results it is therefore natural to consider closely the timestep sequence itself, and this forms another motivation for our analysis.

The paper is organized as follows. In section 2, for a Runge–Kutta method of order r, an expansion of the quadratic form (1.5) is derived and the leading order term is proved to be at least $\mathcal{O}(h^{r+1})$. In section 3, we then use this expansion, together with the corresponding expansion of the local error estimate $E(U_n, h_n)$ at each timestep, to state and justify our two key assumptions on the numerical trajectory. The first assumption takes the form of an upper bound on the timesteps used at each point in the phase space. The second assumption is that controlling the local error estimate also bounds the magnitude of the quadratic form (1.5) at each timestep. It must be emphasized that the justification for these assumptions is that they are expected to hold for every timestep along "typical" numerical trajectories, but it seems likely that for most vector fields satisfying (D) there will be "atypical" numerical trajectories where, at one or more timesteps, they do not hold. Even when these extreme events occur, the fact that the assumptions hold for most of the timesteps should help to preserve the qualitative asymptotic features of the numerical trajectory.

We do not attempt to quantify the ways in which our assumptions can be violated and this is unsatisfactory from a rigorous mathematical viewpoint. However, using these assumptions, we shall gain valuable insights into how these algorithms behave on most simulations. The studies [15, 11, 12] have shown that even when considering the convergence to the exact solution, as $\tau \to 0$, of adaptive timestepping algorithms over finite time intervals and compact sets of initial data—arguably a more fundamental property—there are mechanisms that can give rise to the breakdown of convergence. These arise because of the possibility that the leading term of the error estimate may vanish at some point along the exact trajectory, resulting in a local increase in the size of the accepted timesteps and potential loss of convergence (or, more likely, a reduction in the rate of convergence). However, at least for generic vector fields, the probability of convergence failure is extremely small. These previous studies have therefore already demonstrated that a "worst-case analysis" is not necessarily appropriate in the context of ODE solvers, since the very small probability of failure to converge is outweighed by the superior efficiency of adaptive algorithms. In fact the situation here, where we are concerned with stability properties, is much better than that for convergence properties. This is because convergence can be destroyed by a single "bad" timestep whereas asymptotic qualitative properties are very likely to be robust in the presence of such extreme events. Nevertheless, it is hoped that the analysis presented here will stimulate further work into justifying or weakening the assumptions made.

In section 4, we present the main results. First, for embedded explicit Runge–Kutta pairs of order (1, 2), operating in EPUS or XEPUS modes with B positive-definite, any numerical trajectory satisfying the assumptions of section 3 will eventually enter a particular bounded set for all sufficiently small τ independent of U. Second, motivated by the analysis, we introduce an additional structural assumption on the vector field f:

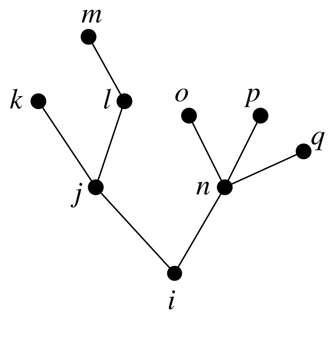
$$(\mathrm{D}') \qquad \exists \gamma > 0, R > 0 : \ \langle f(u), u \rangle \ \leq \ -\gamma \|f(u)\| \|u\| \quad \forall \|u\| \geq R.$$

Intuitively this structural assumption states that, for sufficiently large $\|u\|$, the vector field points inwards everywhere at some definite minimum nonzero angle and holds for many ODEs of interest. In particular, vector fields satisfying (D) or (D$'$) (or both) are not necessarily globally Lipschitz. Assuming that (D), (D$'$), and the assumptions on the numerical trajectory hold, then for sufficiently small τ independent of initial data, for arbitrary embedded $(p-1, p)$ pairs with B positive-definite in any mode of operation, a similar result is proved. Finally, in section 5, we present numerical results that support both the assumptions made in section 3 and the results of section 4.

2. Order conditions and the matrix M.

The Taylor series expansions in powers of h of both the exact solution to (1.1) and the one-step Runge–Kutta approximation, over some time interval $[s, s+h]$, consist of multiples of expressions involving f and its higher derivatives which rapidly become very complicated. We therefore first recall some necessary definitions and terminology from the rooted tree description of Taylor series expansions. This theory was developed by Butcher, and the reader is referred to [3, 4] for full details of all the notation, definitions, and results up to and including (2.3).

A rooted tree is an unlabeled connected graph containing no cycles and with one node identified as the "root." Each rooted tree with precisely n nodes corresponds uniquely to one term (of many) appearing at order h^n in the Taylor series. Each term is a multiple of an elementary differential of order n and this correspondence is achieved as follows. Let $f^i_{j_1, j_2, \ldots, j_r}$ denote the rth partial derivative of the ith component of f with respect to the components $j_1, j_2, \ldots, j_r$. Now attach the label i to the root of the tree and labels $j, k, l, \ldots$ to the other nodes. Then for each node, write down f with a superscript equal to the label of that node and subscripts given by the other nodes that are directly connected to it on the side away from the root node. For example, the rooted tree



corresponds to the product $f^i_{jn} f^j_{kl} f^k f^l_m f^m f^n_{opq} f^o f^p f^q$ (using the summation convention over repeated indices), which is the ith component of one particular elementary differential of order 9. Repeating the above process for each value of the index i provides each component of the elementary differential corresponding to the above (unlabeled) rooted tree. The elementary differential corresponding to a particular tree t will be denoted by the function $F(t) : \mathbb{R}^m \to \mathbb{R}^m$.

The set of all rooted trees, denoted by $\mathcal{T}$, is defined recursively as follows. The rooted tree consisting of a single node is defined as τ and any rooted tree t can be built up by joining trees $t_1, \dots, t_k$ to a new root. The rooted tree t is then written as $t = [t_1, \dots, t_k]$ (note that the order is unimportant) and m repetitions of a tree t_i are denoted by t_i^m.

We now recall some important functions that can be defined on the set $\mathcal{T}$. The function $\rho(t)$ is simply the number of nodes in t. The next three functions $\gamma(t), \sigma(t)$, and $\alpha(t)$ have important combinatorial interpretations (see [3, section 144]) and also allow for an elegant statement of Taylor series expansions. However, the following recursive definitions, also due to Butcher, are more relevant for our purposes:

$$\gamma(\tau) = 1, \quad \gamma([t_1, \dots, t_k]) = \rho([t_1, \dots, t_k]) \prod_{j=1}^{k} \gamma(t_j) \tag{2.1}$$

and

$$\sigma(\tau) = 1, \quad \sigma([t_1^{n_1}, \dots, t_k^{n_k}]) = n_1! n_2! \dots n_k! \prod_{j=1}^{k} \sigma(t_j)^{n_j},$$

where the trees $t_1, \dots, t_k$ are all distinct. Finally, the function $\alpha(t)$ is defined by

$$\alpha(t) = \frac{\rho(t)!}{\gamma(t)\sigma(t)}.$$

In [3] it is then proved that the Taylor series for the exact solution of (1.1) at time $s + h$ is

$$u(s+h) = u(s) + \sum_{t \in \mathcal{T}} \frac{\alpha(t)}{\rho(t)!} h^{\rho(t)} F(t)(u(s)). \tag{2.2}$$

The one-step numerical approximation, $\tilde{u}(s+h)$ can also be expressed in terms of elementary differentials. For a given rooted tree t and Runge–Kutta method (determined by (1.2) and (1.3)) we define the elementary weight $\Phi(t)$ as follows. Label the root of the tree i and attach labels to the other vertices. For every edge connecting vertices u and v, write down a factor a_{uv}, where u is the vertex closer to the root. Insert a final factor b_i, corresponding to the root, form the product of the above factors, and then sum every index over all of the stages. Thus the elementary weight corresponding to the tree drawn above is

$$\Phi(t) = \sum_{i,j,k,l,m,n,o,p,q=1}^{s} b_i a_{ij} a_{jk} a_{jl} a_{lm} a_{in} a_{no} a_{np} a_{nq}.$$

Now the numerical approximation can be expanded as

$$\tilde{u}(s+h) = u(s) + \sum_{t \in \mathcal{T}} \frac{\gamma(t)\alpha(t)\Phi(t)}{\rho(t)!} h^{\rho(t)} F(t)(u(s)). \tag{2.3}$$

By comparing (2.2) and (2.3), Butcher proved that a necessary and sufficient condition for a Runge–Kutta method to be of order precisely p is that $\Phi(t) = 1/\gamma(t)$ for all rooted trees t with $\rho(t) \leq p$, but not for at least one tree t with $\rho(t) = p + 1$.

The above definitions and results now enable us to prove a new expansion for the quadratic form (1.5).

LEMMA 2.1. *Let the stages $\eta_1, \ldots, \eta_s$ be generated by a Runge–Kutta method of order r using a timestep h from a solution value u. Then there exists an integer $q \geq r+1$ and scalar-valued functions $G_1(u)$ and $G_2(u,h)$ such that $G_2(u,0) = 0$ and*

$$h^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j) \rangle \; = \; h^q (G_1(u) + hG_2(u,h)). \tag{2.4}$$

Proof. Consider an arbitrary rooted tree t. Comparison of the Taylor series expansion of the numerical solution (2.3) with (1.3) shows that each of the terms $hf(\eta_i)$ can be expanded as

$$hf(\eta_i) = \sum_{t \in \mathcal{T}} \frac{\gamma(t)\alpha(t)\Phi_i(t)}{\rho(t)!} h^{\rho(t)} F(t)(u),$$

where each term $\Phi_i(t)$ is derived from $\Phi(t)$ by deleting both the factor b_i and the summation over the index i. Note that each $\Phi_i(t)$ has precisely $\rho(t) - 1$ factors. Let us now fix trees T_1 and T_2 (not necessarily distinct) and consider the coefficient of $\langle F(T_1)(u), F(T_2)(u) \rangle$ in the expansion of (1.5). Using the definition of the matrix M, this is

$$\begin{aligned} &(2 - \mathcal{I}_{T_1=T_2})\, h^{\rho(T_1)+\rho(T_2)} \frac{\alpha(T_1)\alpha(T_2)\gamma(T_1)\gamma(T_2)}{\rho(T_1)!\rho(T_2)!} \sum_{i,j=1}^{s} [\Phi_i(T_1)\Phi_j(T_2) b_i a_{ij} \\ &\qquad + \Phi_i(T_1)\Phi_j(T_2) b_j a_{ji} - \Phi_i(T_1)\Phi_j(T_2) b_i b_j], \end{aligned} \tag{2.5}$$

where $\mathcal{I}_{T_1=T_2} = 1$ if $T_1 = T_2$ and 0 otherwise.

We now introduce some new notation. Given two trees $T_1 = [s_1, \ldots, s_m]$ and $T_2 = [t_1, \ldots, t_n]$ (where $m = 0$ or $n = 0$ correspond to $T_1 = \tau$ or $T_2 = \tau$, respectively) we define the tree $T_1 \nearrow T_2 := [s_1, \ldots, s_m, T_2]$, which is the tree with $\rho(T_1) + \rho(T_2)$ nodes obtained by adding a single edge between the roots of T_1 and T_2 and keeping the root of T_1 as the root of the new tree. Similarly, $T_2 \nearrow T_1 := [t_1, \ldots, t_n, T_1]$. Thus the first term in the summand of (2.5), after summation, corresponds to $\Phi(T_1 \nearrow T_2)$, the second term corresponds to $\Phi(T_2 \nearrow T_1)$, and the third term to $\Phi(T_1)\Phi(T_2)$.

Let us now assume that $\rho(T_1) + \rho(T_2) \leq r$. Then this coefficient vanishes if

$$\Phi(T_1 \nearrow T_2) + \Phi(T_2 \nearrow T_1) = \Phi(T_1)\Phi(T_2). \tag{2.6}$$

But since the Runge–Kutta method is of order r this is equivalent to the condition that

$$\frac{1}{\gamma(T_1 \nearrow T_2)} + \frac{1}{\gamma(T_2 \nearrow T_1)} = \frac{1}{\gamma(T_1)\gamma(T_2)}. \tag{2.7}$$

This is easily proved via (2.1), the recursive definition of γ. For let us suppose first that $T_1 \neq \tau \neq T_2$. Then

$$\begin{aligned} \gamma(T_1) &= \rho(T_1)\gamma(s_1)\ldots\gamma(s_m), \\ \gamma(T_2) &= \rho(T_2)\gamma(t_1)\ldots\gamma(t_n), \\ \gamma(T_1 \nearrow T_2) &= [\rho(T_1) + \rho(T_2)]\rho(T_2)\gamma(s_1)\ldots\gamma(s_m)\gamma(t_1)\ldots\gamma(t_n), \\ \gamma(T_2 \nearrow T_1) &= [\rho(T_1) + \rho(T_2)]\rho(T_1)\gamma(t_1)\ldots\gamma(t_n)\gamma(s_1)\ldots\gamma(s_m), \end{aligned}$$

and (2.7) easily follows. The remaining cases when either $T_1 = \tau$ or $T_2 = \tau$ are also easily verified.

Thus

$$h^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j) \rangle \;=\; h^q (G_1(u) + hG_2(u,h))$$

for some $q \geq r+1$, with the function $G_1(u)$ being the sum of inner products of elementary differentials where the nodes in the corresponding rooted trees sum to precisely q, and the function $G_2(u,h)$ comprising the higher-order terms. □

The possibility of $q > r+1$ in the statement of Lemma 2.1 arises because, for a pair of trees T_1, T_2 with $\rho(T_1) + \rho(T_2) = n > r$, the Runge–Kutta method being of order n is a sufficient, but not necessary, condition for (2.6) to be satisfied. This is in fact the case for the improved Euler (Heun) method where (2.6) is also satisfied for the unique pair of rooted trees whose nodes sum to 3. Thus $q = 4$ even though $r = 2$.

3. Assumptions.

We now turn to the assumptions necessary for the analysis and results of section 4. Once again, the purpose of these results is to provide an explanation of the observed behavior of explicit Runge–Kutta pairs for "typical" numerical trajectories of "typical" vector fields. For the vast majority of adaptive schemes (i.e., apart from ones utilizing essentially algebraically stable pairs) it would appear that no results are possible without such assumptions. As mentioned in the introduction, similar problems arise when proving convergence results for adaptive algorithms, even for finite-time initial value problems on compact domains. This is because any method based upon a local error estimate can behave badly, even if only for a single timestep, by a sufficiently unfortunate (or devious) combination of vector field, solution value, and candidate timestep. However, both of the assumptions stated and justified below are numerically verified for every single timestep used to advance the solutions in the numerical experiments of section 5.

Assumption 1. If the local error estimate is derived from a $(p-1, p)$ explicit Runge–Kutta pair, then, for all sufficiently small $\tau > 0$, there exists a constant $K_1 > 0$, independent of U, such that for each accepted timestep h_n,

$$h_n^{p-\rho} \leq K_1 \frac{\sigma(\tau, U_n)}{\|f(U_n)\|}. \tag{3.1}$$

The intuitive reason for this assumption can be seen by following [15, 11] and expanding the local error estimate as

$$E(U_n, h_n) = h_n^p \left(B_1(U_n) + h_n B_2(U_n, h_n)\right) \tag{3.2}$$

$$= h_n^p \|f(U_n)\| (\tilde{B}_1(U_n) + h_n \tilde{B}_2(U_n, h_n)). \tag{3.3}$$

In (3.3) the expansion has simply been rescaled by a factor of $\|f(U_n)\|$. Now let us suppose that the function $\|B_1(u)\|$ is bounded away from zero along the numerical trajectory. Then if the error control is working correctly (for sufficiently small τ), and the accepted timesteps are controlled by the (nonvanishing) leading-order term of the expansion (3.2), we see that (3.1) immediately follows.

In [15, 11, 10], rigorous proofs of the upper bound (3.1) on the sequence of accepted timesteps are obtained via induction arguments for sufficiently small τ, but only for numerical trajectories lying inside a predefined compact set on which $B_1(u)$ is bounded away from zero. By restricting ourselves to ODEs satisfying (D), we now argue that (3.1) will only fail to hold in exceptional cases for any initial data.

Note first that under assumption (D), $f(u) \neq 0$ outside the ball $\overline{B}(0, \sqrt{\alpha/\beta})$. Thus, outside this ball, the leading order term of the error estimate can vanish only if $\tilde{B}_1$ does. But $\tilde{B}_1 : \mathbb{R}^m \to \mathbb{R}^m$ and so for typical vector fields will vanish only at isolated points in the phase space. In order to obtain (3.1) from (3.3), we assume the existence of constants $K, K' > 0$ (independent of τ and U_n) for τ sufficiently small such that at each step of the numerical trajectory

$$\begin{aligned}\sigma(U_n, \tau) \geq \|E(U_n, h_n)\|/h_n^\rho &\geq K h_n^{p-\rho} \max\left(B_1(u), h_n B_2(U_n, h_n)\right)\\ &\geq K h_n^{p-\rho} \|B_1(u)\|\\ &\geq KK' \|f(U_n)\| h_n^{p-\rho},\end{aligned}$$

leading immediately to (3.1) with $K_1 = 1/KK'$. The existence of the constant $K > 0$ is equivalent to assuming that at each step no catastrophic cancellation occurs between B_1 and $h_n B_2$. In order to justify the existence of the constant $K' > 0$ we need to demonstrate that, under assumption (D), $\|\tilde{B}_1(u)\|$ does not tend to 0 as $\|u\| \to \infty$ in any direction. We achieve this by showing that at least one of the rescaled elementary differentials comprising $\tilde{B}_1(u)$ cannot vanish as $\|u\| \to \infty$.

Let us suppose that the lower-order method of the pair does not increase its order on linear constant-coefficient problems.[1] Then $B_1(u)$ must contain an elementary differential of the form $cf'(u)^{p-1}f(u)$ with coefficient $c \neq 0$ (the rooted trees corresponding to such elementary differentials are often referred to as "tall trees" and contain no branches). Under the structural assumption (D), $\|f(u)\|$ must increase at least as fast as $\mathcal{O}(\|u\|)$ for sufficiently large $\|u\|$ in any given direction. Thus $\|f'(u)\|$ and $\|cf'(u)^{p-1}f(u)\|/\|f(u)\|$ cannot tend to 0 as $\|u\| \to \infty$ (although for pathological vector fields, $f'(u)$ may equal zero on arbitrarily large compact sets in the phase space). We now approach once again to the principle that catastrophic cancellation (this time between the weighted and rescaled elementary differentials comprising $\tilde{B}_1(u)$) occurs negligibly often, giving $\tilde{B}_1(u) \not\to 0$ as $\|u\| \to \infty$ in any direction. This completes our justification of (3.1).

It should be noted that for a linear constant-coefficient ODE satisfying (D), $\|\tilde{B}_1(u)\|$ is a nonzero constant, but for certain nonlinear problems we can expect $\|\tilde{B}_1(u)\|$ to grow as $\|u\|$ grows. Thus for particular classes of nonlinear problem it may be possible to strengthen the upper bound on the timestep sequence in Assumption 1 considerably (this is confirmed by numerical computations but we shall not explore this point further).

The second assumption states that the error control, which is of course designed to bound the local error estimate, also provides a bound on the magnitude of the quadratic form (1.5) for typical timesteps.

Assumption 2. For all sufficiently small τ there exists a constant $K_2 > 0$, independent of U, such that at each timestep along the numerical trajectory

$$\left| h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j)\rangle \right| \leq K_2 \sigma(\tau, U_n) h_n^{1+\rho} \|f(U_n)\|, \tag{3.4}$$

[1] If this mild condition is violated, then the method behaves substantially differently for such problems. Indeed if the $(p-1, p)$ pair has precisely p stages, then the local error estimate $E(u, h) \equiv 0$ and the error control fails completely. The reader is referred to [11] for further discussion of this point. We simply note that most embedded pairs used in practice satisfy this criterion.

where the matrix M is the stability matrix for the higher-order method of the $(p-1,\ p)$ Runge–Kutta pair.

We start our justification of Assumption 2 by defining

$$\hat{E}(U_n,h_n) := \left| h_n \sum_{j=1}^{s} \langle E(U_n,h_n), f(\eta_j)\rangle \right| \tag{3.5}$$

$$= \left| h_n^2 \sum_{i,j=1}^{s} (b_i - \bar{b}_i)\langle f(\eta_i), f(\eta_j)\rangle \right|. \tag{3.6}$$

The enforcement of the local error control (1.9) now allows us to bound $\hat{E}(U_n,h_n)$ from above, since

$$\begin{aligned} \hat{E}(U_n,h_n) &\le h_n \sum_{j=1}^{s} \|E\|\ \|f(\eta_j)\| \\ &\le s\sigma(\tau,U_n) h_n^{1+\rho} \max_{j=1,\dots,s} \|f(\eta_j)\|. \end{aligned} \tag{3.7}$$

We now compare the expansion (3.6) for $\hat{E}$ with that of the (absolute value of the) quadratic form for the higher-order method (1.5). From the proof of Lemma 2.1, (1.5) is a linear combination of inner products of elementary differentials. Reverting to the rooted-tree description of elementary differentials, the only inner products $\langle F(T_1)(u), F(T_2)(u)\rangle$ appearing in the expansion are those for which $\rho(T_1)+\rho(T_2) \ge p+1$, and their coefficients are of order $h_n^{\rho(T_1)+\rho(T_2)}$. The corresponding expansion for $\hat{E}$ contains those inner products $\langle F(T_1)(u), F(T_2)(u)\rangle$ for which $\max(\rho(T_1),\rho(T_2)) \ge p$, once again with coefficients of order $h_n^{\rho(T_1)+\rho(T_2)}$.

Note that the expansion of (1.5) therefore contains a (finite) number of additional inner products not appearing in that of $\hat{E}$. However, these inner products are closely related to others that are common to both expansions and so our assumption reduces to the observation that the control of the quantity $\hat{E}$ should effectively control (1.5) to within some constant. Assumption 2 now follows immediately from (3.7) by assuming that $\max_{j=1,\dots,s} \|f(\eta_j)\|$ is always close to $\|f(U_n)\|$, which of course should be the case, barring any catastrophic cancellations in the formation of the error estimate.

In principle, Assumptions 1 and 2 could be weakened considerably by, for example, only requiring that (3.1) and (3.4) hold, for a given K_1 and K_2, on a sufficiently large proportion of the numerical timesteps. However, an analysis resting on such assumptions would become far more difficult without generating any new insights into the mechanisms leading to effective numerical stability.

4. Results.

Using Assumptions 1 and 2 we are now ready to prove the main results. We start by considering the case of embedded explicit Runge–Kutta pairs with order (1, 2) in either EPUS or XEPUS mode.

THEOREM 4.1. *Consider an embedded explicit Runge–Kutta pair of order* $(1,2)$, *under either EPUS or XEPUS control, where the higher-order method has positive weights. If Assumptions* 1 *and* 2 *are satisfied and the ODE* (1.1) *satisfies* (D), *then* $\exists \tau^* > 0$ *such that* $\forall \tau \le \tau^*$, *the numerical trajectory eventually enters a compact set independent of* τ, U.

Proof. The tolerance τ is chosen sufficiently small such that Assumptions 1 and 2 are satisfied. We first consider advancing the numerical solution using the higher-order method (extrapolation mode). From (D) we have

$$\|V_{n+1}\|^2 = \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i \langle \eta_i, f(\eta_i)\rangle - h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j)\rangle \tag{4.1}$$

$$\leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i (\alpha - \beta\|\eta_i\|^2) - h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j)\rangle. \tag{4.2}$$

We now proceed by bounding the absolute value of the last term and, for sufficiently small τ, absorbing it into the previous one. From Assumptions 1 and 2 and (1.10),

$$\left| h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j)\rangle \right| \leq K_2 \sigma(\tau, U_n) h_n^2 \|f(U_n)\| \tag{4.3}$$

$$\leq K_1 K_2 h_n \sigma(\tau, U_n)^2 \tag{4.4}$$

$$\leq C_1^2 K_1 K_2 h_n \tau^2 \|U_n\|^2. \tag{4.5}$$

Now fix $0 < \tilde{\beta} < \beta$ and substitute (4.5) into the last term of (4.2) with

$$\tau < \sqrt{\frac{2b_1(\beta - \tilde{\beta})}{C_1^2 K_1 K_2}}.$$

Noting that $\eta_1 = U_n$ for explicit Runge–Kutta methods, we obtain

$$\|V_{n+1}\|^2 \leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i (\alpha - \tilde{\beta}\|\eta_i\|^2). \tag{4.6}$$

The proof now proceeds exactly as in [17, Lemma 4.2 and Theorem DC1] by showing that the norm of the numerical solution strictly decreases until it enters, for any $\epsilon > 0$, the compact set $S = \overline{B}(0, \sqrt{((\alpha + \epsilon)/\tilde{\beta}) + h_{\max} K})$, where

$$K = \max_{\|\eta_i\| \leq \gamma_i} \left(2 \sum_{i,j=1}^{s} b_i e_{ij} \langle \eta_i, f(\eta_i)\rangle + h_{\max} \sum_{i=1}^{s} b_i \left\| \sum_{j=1}^{s} e_{ij} f(\eta_j) \right\|^2 \right) \tag{4.7}$$

and

$$e_{ij} := b_j - a_{ij}, \qquad \gamma_i^2 := \frac{\alpha}{\tilde{\beta} b_i}.$$

We now consider the nonextrapolation case. From the local error control (1.9),

$$\begin{aligned} \|W_{n+1}\|^2 - \|V_{n+1}\|^2 &= \langle W_{n+1} + V_{n+1}, W_{n+1} - V_{n+1}\rangle \\ &\leq \|W_{n+1} + V_{n+1}\| \, \|W_{n+1} - V_{n+1}\| \\ &\leq \|W_{n+1} + V_{n+1}\| \sigma(\tau, U_n) h_n \\ &\leq 2\|V_{n+1}\| \sigma(\tau, U_n) h_n + \sigma^2(\tau, U_n) h_n^2. \end{aligned}$$

While the numerical trajectory is outside the compact set $\overline{B}(0, \sqrt{((\alpha + \epsilon)/\tilde{\beta}) + h_{\max} K})$, we have already proved that, for sufficiently small τ, $\|V_{n+1}\| \leq \|U_n\|$ implying

$$\begin{aligned} \|W_{n+1}\|^2 - \|V_{n+1}\|^2 &\leq 2\|U_n\| \sigma(\tau, U_n) h_n + \sigma^2(\tau, U_n) h_n^2 \\ &\leq 2C_1 \tau \|U_n\|^2 h_n + C_1^2 \tau^2 \|U_n\|^2 h_n^2. \end{aligned} \tag{4.8}$$

Thus the bound on $\|V_{n+1}\|^2$ from (4.6) may be invoked in (4.8) to give

(4.9)
$$\|W_{n+1}\|^2 \leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i(\alpha - \tilde{\beta}\|\eta_i\|^2) + 2C_1\tau\|U_n\|^2 h_n + C_1^2\tau^2\|U_n\|^2 h_n h_{\max}.$$

The argument now concludes in a very similar fashion to the extrapolation case. After reducing the tolerance τ further if necessary, the last two terms of (4.9) can be absorbed into the preceding term by reducing $\tilde{\beta}$ once again, and then redefining (increasing) K to $\tilde{K}$ via (4.7). This then proves that the numerical solution enters a set $\overline{B}(0, \sqrt{((\alpha+\epsilon)/\tilde{\beta}) + h_{\max}\tilde{K}})$ as required. □

Theorem 4.1 only states that numerical trajectories will enter a particular compact set, which is not necessarily close to the set $\overline{B}(0, \sqrt{\alpha/\beta})$. However, once a numerical trajectory has entered this set, finite-time convergence results, such as those contained in [15, 12], can be applied to prove that typical numerical trajectories (possibly after a further reduction in τ) will enter and remain within $\mathcal{O}(\tau)$ of the absorbing set $B(0, \sqrt{\alpha/\beta})$ of the ODE (1.1). Furthermore in [10], and under additional assumptions, the existence of a (local) numerical attractor that is upper-semicontinuous to the global attractor of (1.1) can be proved.

We now prove a more general result, applicable to embedded Runge–Kutta pairs of any order and under any mode of operation. Note also that Assumption 1 is no longer required.

THEOREM 4.2. *Consider an adaptive embedded Runge–Kutta pair of any order* $(p-1,p)$*, operating in EPS, XEPS, EPUS, or XEPUS mode, where the higher-order method has positive weights. If Assumption* 2 *holds and the ODE* (1.1) *satisfies both* (D) *and* (D′)*, then* $\exists\tau^* > 0$ *such that* $\forall\tau \leq \tau^*$*, the numerical trajectory eventually enters a compact set independent of* τ, U*.*

Proof. Again we consider the extrapolation case first. From Assumption 2 and (1.10),

$$\left| h_n^2 \sum_{i,j=1}^{s} m_{ij}\langle f(\eta_i), f(\eta_j)\rangle \right| \leq K_2 C_1 \|U_n\|\,\|f(U_n)\|\tau h_n^{1+\rho},$$

which upon substituting into (4.1) gives

$$\begin{aligned}\|V_{n+1}\|^2 &\leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i\langle\eta_i, f(\eta_i)\rangle + K_2C_1\|U_n\|\,\|f(U_n)\|\tau h_n^{1+\rho}\\ &= \|U_n\|^2 + h_n b_1\langle\eta_1, f(\eta_1)\rangle + 2h_n\sum_{i=1}^{s}\hat{b}_i\langle\eta_i, f(\eta_i)\rangle\\ &\quad + K_2C_1\|U_n\|\,\|f(U_n)\|\tau h_n^{1+\rho},\end{aligned}$$

where $\hat{b}_1 = \frac{1}{2}b_1$ and $\hat{b}_i = b_i,\ i = 2, \ldots, s$. We now assume that $\|U_n\| > R$ and using (D′) obtain

$$\begin{aligned}\|V_{n+1}\|^2 \leq \|U_n\|^2 - h_n b_1\gamma\|U_n\|\,\|f(U_n)\| + 2h_n\sum_{i=1}^{s}\hat{b}_i\langle\eta_i, f(\eta_i)\rangle\\ + K_2C_1\|U_n\|\,\|f(U_n)\|\tau h_n^{1+\rho}.\end{aligned}$$

Choosing

$$\tau \leq \frac{b_1 \gamma}{K_2 C_1 \max(1, h_{\max})^\rho}$$

and applying (D) we have

$$\|V_{n+1}\|^2 \leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} \hat{b}_i(\alpha - \beta\|\eta_i\|^2).$$

Next we define $\tilde{\beta} = \beta/2$ to give

$$\begin{aligned} \|V_{n+1}\|^2 &\leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i(\alpha - \beta\|\eta_i\|^2) - h_n b_1(\alpha - \beta\|\eta_1\|^2) \\ &\leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i(\alpha - \tilde{\beta}\|\eta_i\|^2) - h_n b_1 \alpha \\ &\leq \|U_n\|^2 + 2h_n \sum_{i=1}^{s} b_i(\alpha - \tilde{\beta}\|\eta_i\|^2), \end{aligned} \tag{4.10}$$

which is identical to (4.6). Thus the proof continues in a very similar manner to that of Theorem 4.1, via the construction of a compact set S outside of which

$$2h_n \sum_{i=1}^{s} b_i(\alpha - \tilde{\beta}\|\eta_i\|^2) \leq 0.$$

While the numerical trajectory is outside the set $S \cup \overline{B}(0, R)$, its norm strictly decreases until the set is eventually entered.

For the nonextrapolation case, from (4.10) and (4.8) we once again obtain (4.9). Again, reducing $\tau, \tilde{\beta}$ and increasing K, if necessary, the numerical trajectory eventually enters some compact set $S' \cup \overline{B}(0, R)$. □

5. Numerical results. Some numerical examples are now presented to support Assumptions 1 and 2 and Theorems 4.1 and 4.2. We shall consider various embedded Runge–Kutta pairs in different operational modes. The algorithms used are all modifications of the ode23 routine supplied with MATLAB Version 4.2. This code was used (rather than, for example, the more sophisticated ODE routines in later MATLAB versions) because the timestep mechanism is particularly straightforward, containing only elements common to all such adaptive algorithms. Note that none of the previous analysis relies upon a detailed description of the timestep selection mechanism, merely that the local error control is satisfied.

Two examples of vector fields that satisfy both (D) and (D′) are the scalar ODE

$$u_t = -u|u| \tag{5.1}$$

and the linear constant-coefficient problem

$$\begin{aligned} x_t &= -y - \epsilon x, \\ y_t &= x - \epsilon y \end{aligned} \tag{5.2}$$

for $\epsilon > 0$. Note that for scalar ODEs (D) implies (D′).

A vector field that satisfies (D) but not (D$'$) is, for $\epsilon > 0$,

$$x_t = -y\sqrt{x^2+y^2} - \epsilon x, \qquad (5.3)$$
$$y_t = x\sqrt{x^2+y^2} - \epsilon y,$$

while a vector field that satisfies neither, yet has an absorbing set, is

$$x_t = y - \epsilon\frac{x}{\sqrt{x^2+y^2}}, \qquad (5.4)$$
$$y_t = -x - \epsilon\frac{y}{\sqrt{x^2+y^2}}.$$

Up to this point, we have not considered how the numerical algorithm generates candidate timesteps since we require only that the error control be satisfied. However, for the sake of completeness, we shall explicitly describe the timestep selection mechanism used in the numerical simulations. This algorithm is based upon asymptotic considerations (see, for example, [14, 7, 12]) as the tolerance τ, and thus the timesteps, tend to zero. If h_{last} was the last attempted timestep (successful or otherwise), then the next attempted timestep is defined by

$$h_{\text{next}} = \min\left(h_{\max}, \theta\left(\frac{\sigma(\tau,U)}{E(U,h_{\text{last}})}\right)^{\frac{1}{p-\rho}} h_{\text{last}}\right),$$

where U is the most recent solution value. The constant $\theta < 1$ is a "safety factor" ensuring that, provided the exact solution lies in a compact set, the proportion of rejected timesteps along numerical approximations will tend to 0 as $\tau \to 0$.

We first consider the behavior of order (1, 2) pairs with error-per-unit-step control. Figure 5.1 plots the Euclidean norm of the numerical solution against integration time for the ODEs (5.1)–(5.4) using the embedded Runge–Kutta pair consisting of the forward Euler and Heun methods, defined by

$$A = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix}, \qquad (5.5)$$

in extrapolation mode with $\tau = 0.1$, $\theta = 0.9$, and the norm of the initial data set to $\|U\| = 10^5$. Here, as in all subsequent results, a relative error criterion defined by

$$\sigma(\tau, u) = \tau\|u\|_2$$

was used as this results in larger timesteps and thus provides a more severe (and, arguably, more relevant) test than a pure absolute error control. Even for this relatively large value of τ, the results are in agreement with Theorem 4.1. The reduction in norm of the numerical solution for sufficiently small τ is guaranteed for (5.1)–(5.3) since this pair is essentially algebraically stable. For (5.4) the norm of the solution increases with this value of τ. If τ is reduced sufficiently, then stability of the numerical solution is recovered for this initial data but the instability reappears as $\|U\|$ is increased further, i.e., τ depends upon the initial data. In Figure 5.2, we test Assumptions 1 and 2 by plotting the calculated values of

$$k_1(U_n, h_n) = \frac{h_n^{p-\rho}\|f(U_n)\|}{\sigma(\tau, U_n)}$$

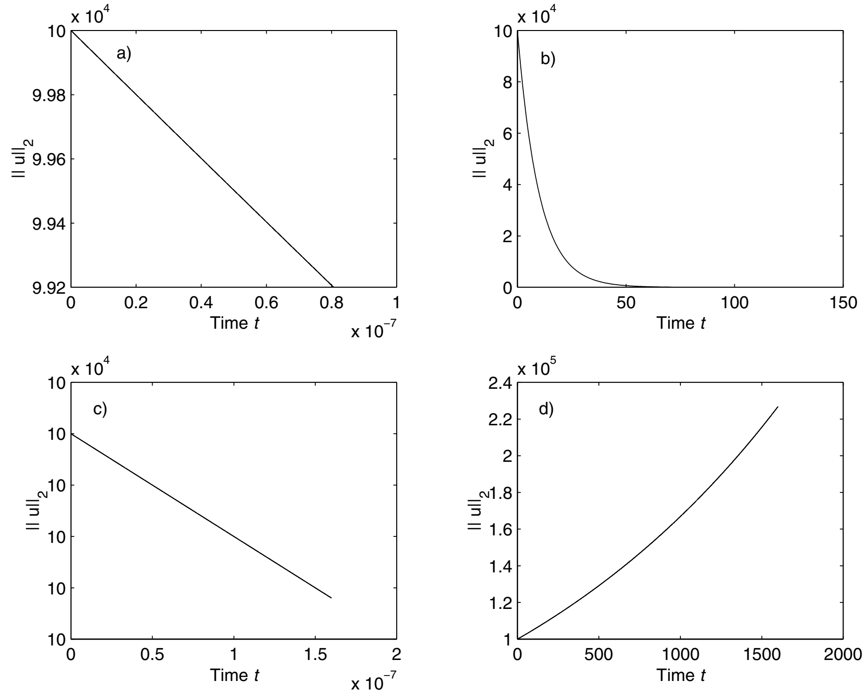


FIG. 5.1. (a)–(d) *plot the norms of the numerical solution using the embedded pair* (5.5) *in XEPUS mode for* (5.1)–(5.4)*, respectively. The values of* ϵ *used in* (b), (c), *and* (d) *are* 0.1, 1, *and* 1, *and in each case* $\tau = 0.1$. *(Note that the norm of the solution in* (c) *does decrease, as expected, but extremely slowly.)*

and

$$k_2(U_n, h_n) = \frac{\left| h_n^2 \sum_{i,j=1}^{s} m_{ij} \langle f(\eta_i), f(\eta_j) \rangle \right|}{\sigma(\tau, U_n) h_n^{1+\rho} \|f(U_n)\|}.$$

The maxima of these quantities along the numerical trajectory are the effective values of K_1 and K_2, respectively, and, if Assumptions 1 and 2 are justified, these quantities should remain bounded as $\|U_n\| \to \infty$. This is indeed the case for all four trajectories in Figure 5.1, and in Figure 5.2, k_1 and k_2 are plotted for just two of the test problems, namely, (5.1) and (5.3) (for the linear ODE (5.2), these quantities are constant along the entire numerical trajectory).

Figure 5.3 is generated exactly as Figure 5.1 but using the nonessentially algebraically stable embedded pair

$$A = \begin{pmatrix} 0 & 0 \\ 2 & 0 \end{pmatrix}, \quad \bar{b} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad b = \begin{pmatrix} \frac{3}{4} \\ \frac{1}{4} \end{pmatrix}. \tag{5.6}$$

As can be seen, the results are very similar to those using the essentially algebraically stable (EAS) pair (5.5) and suggest that, although EAS pairs have guaranteed stability properties, there is little difference between EAS and non-EAS pairs in practice.

We now consider Theorem 4.2. Figure 5.4 is generated identically to Figure 5.1 except that now the method (5.5) is being used in XEPS mode rather than XEPUS.

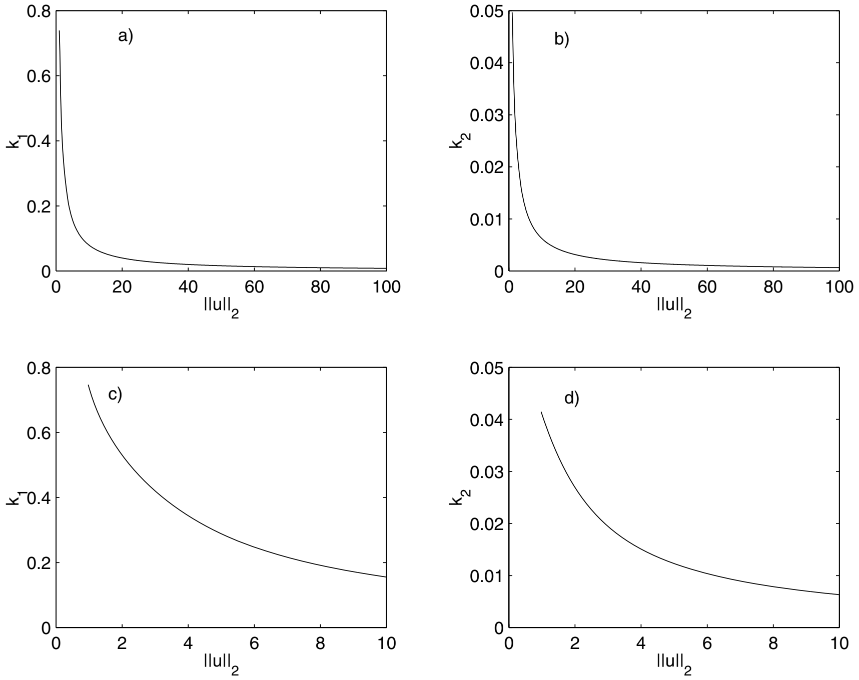


FIG. 5.2. (a) *and* (b) *show the computed values of* k_1 *and* k_2 *for a numerical trajectory of* (5.1) *while* (c) *and* (d) *are for* (5.3). *Apart from the initial data all the parameters are the same as used in Figure* 5.1(a) *and* (c).

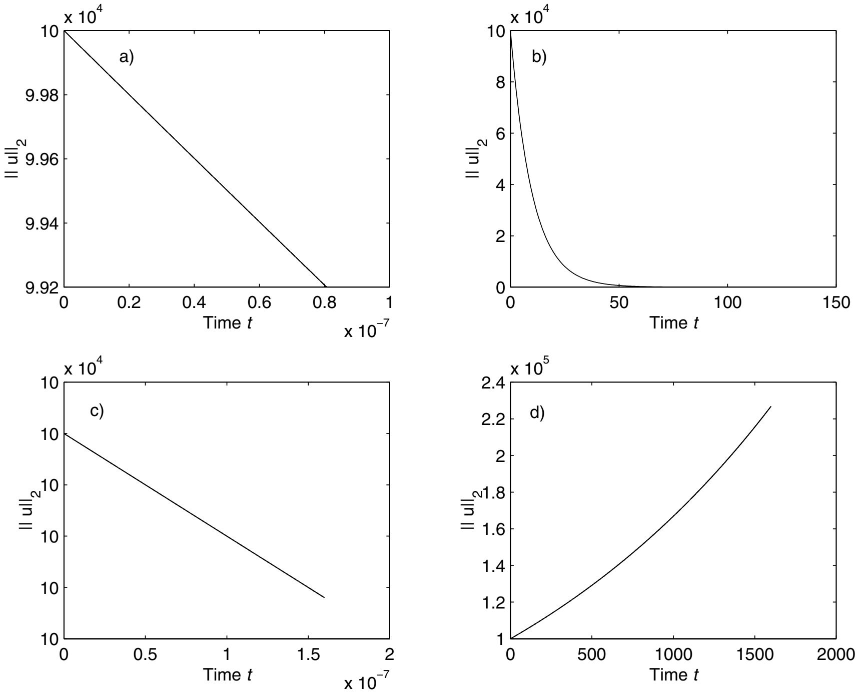


FIG. 5.3. (a)–(d) *are generated exactly as in Figure* 5.1 *but using the non-EAS pair* (5.6).

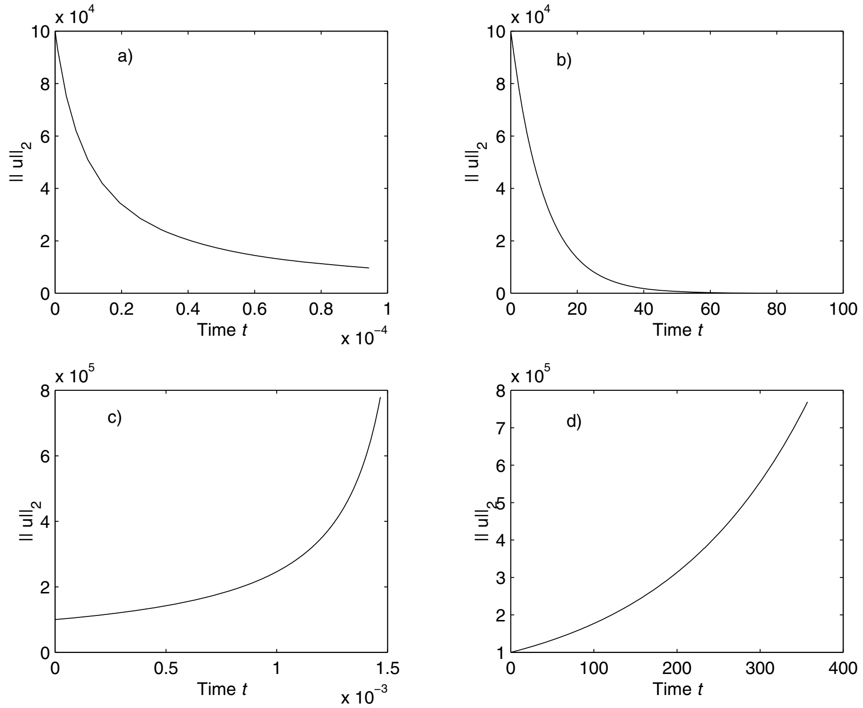


Fig. 5.4. (a)–(d) *are generated exactly as in Figure* 5.1 *but using the method* (5.5) *in XEPS rather than XEPUS mode.*

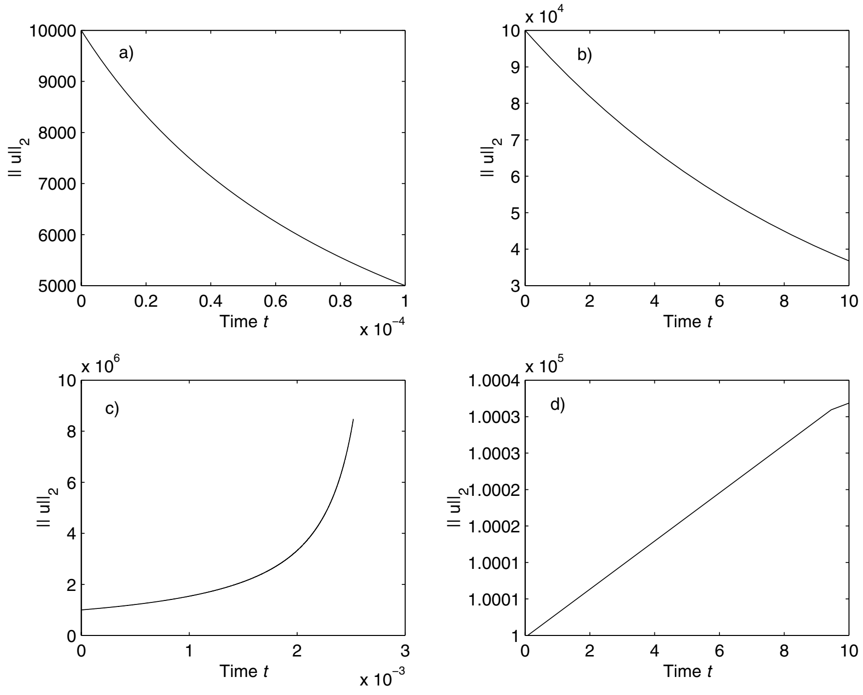


Fig. 5.5. (a)–(d) *are generated exactly as in Figure* 5.1 *but using the Fehlberg* (4, 5) *pair in XEPS mode.*

The interesting case is Figure 5.4(c), corresponding to the vector field (5.3), which satisfies (D) but not (D′). Now the norm of the numerical solution increases rather than decreases and, for any given tolerance, this phenomenon appears to occur for sufficiently large initial data.

Finally, we present results for a higher-order pair. Figure 5.5 shows the numerical results obtained using the Fehlberg (4, 5) pair in XEPS mode. Note that this embedded Runge–Kutta pair, whose coefficients are listed in [3, p. 306], does not satisfy the condition that the weights of the higher-order method are positive, but the results are similar to those obtained for other pairs that do satisfy this condition, suggesting that this condition could be weakened somewhat. Again, the importance of the additional structural assumption (D′) is revealed in Figure 5.5(c).

Acknowledgment. The author thanks one of the anonymous referees for their extremely careful reading of the original manuscript and many valuable suggestions.

REFERENCES

[1] K. Burrage and J.C. Butcher, *Stability criteria for implicit Runge–Kutta processes*, SIAM J. Numer. Anal., 19 (1979), pp. 46–57.

[2] J.C. Butcher, *A stability property of implicit Runge–Kutta methods,* BIT, 15 (1975), pp. 358–361.

[3] J.C. Butcher, *The Numerical Analysis of Ordinary Differential Equations,* Wiley, New York, 1992.

[4] J.C. Butcher, *Numerical Methods for Ordinary Differential Equations,* Wiley, New York, 2003.

[5] K. Dekker and J.G. Verwer, *Stability of Runge Kutta Methods for Stiff Nonlinear Differential Equations,* North-Holland, Amsterdam, 1984.

[6] D.F. Griffiths, *The dynamics of some linear multistep methods with step-size control,* in Numerical Analysis 1987, D.F. Griffiths and G.A. Watson, eds., Longman Scientific and Technical, Harlow, UK, 1988, pp. 115–134.

[7] E. Hairer, S.P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I. *Nonstiff Problems,* 2nd ed., Springer-Verlag, Berlin, 1993.

[8] D.J. Higham and A.M. Stuart, *Analysis of the dynamics of local error control via a piecewise continuous residual,* BIT, 38 (1998), pp. 44–57.

[9] A.R. Humphries and A.M. Stuart, *Runge–Kutta methods for dissipative and gradient dynamical systems,* SIAM J. Numer. Anal., 31 (1994), pp. 1452–1485.

[10] H. Lamba, *Dynamical systems and adaptive time-stepping in ODE solvers,* BIT, 40 (2000), pp. 314–335.

[11] H. Lamba and A.M. Stuart, *Convergence results for the MATLAB ode*23 *routine,* BIT, 38 (1998), pp. 751–780.

[12] H. Lamba and A.M. Stuart, *Convergence proofs for numerical IVP software,* in Dynamics of Algorithms, IMA Vol. Math. Appl. 118, Springer-Verlag, New York, 1999, pp. 107–127.

[13] J.M. Sanz-Serna, *Numerical ordinary differential equations vs. dynamical systems,* in The Dynamics of Numerics and the Numerics of Dynamics, D.S. Broomhead and A. Iserles, eds., Clarendon Press, Oxford, 1992, pp. 81–106.

[14] L.F. Shampine, *Numerical Solution of Ordinary Differential Equations,* Chapman and Hall, New York, 1994.

[15] A.M. Stuart, *Probabilistic and deterministic convergence proofs for software for initial value problems,* Numer. Algorithms, 14 (1997), pp. 227–260.

[16] A.M. Stuart and A.R. Humphries, *Model problems in numerical stability theory for initial value problems,* SIAM Rev., 36 (1994), pp. 226–257.

[17] A.M. Stuart and A.R. Humphries, *The essential stability of local error control for dynamical systems,* SIAM J. Numer. Anal., 32 (1995), pp. 1940–1971.

[18] A.M. Stuart and A.R. Humphries, *Dynamical Systems and Numerical Analysis,* Cambridge University Press, Cambridge, UK, 1996.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 1969–1987

ROOTS OF POLYNOMIALS EXPRESSED IN TERMS OF ORTHOGONAL POLYNOMIALS*

DAVID DAY† AND LOUIS ROMERO†

Abstract. A technique is presented for determining the roots of a polynomial $p(x)$ that is expressed in terms of an expansion in orthogonal polynomials. The roots are expressed as the eigenvalues of a nonstandard companion matrix $\mathbf{B}_n$ whose coefficients depend on the recurrence formula for the orthogonal polynomials, and on the coefficients of the orthogonal expansion. Some questions on the numerical stability of the eigenvalue problem to which they give rise are discussed. The problem of finding the roots of a transcendental function $f(x)$ can be reduced to the problem considered by approximating $f(x)$ by a Chebyshev polynomial. We illustrate the effectiveness of this convert-to-Chebyshev strategy by solving several transcendental equations using this plus our new algorithm. We analyze the numerical stability through both linear algebra theory and numerical experiments and find that this method is very well conditioned.

Key words. rootfinding, Chebyshev polynomial, Legendre polynomial, single transcendental equation, global methods, companion matrix, eigenvalue problem

AMS subject classifications. 65H05, 42C10, 65H20, 65F15

DOI. 10.1137/040609847

1. Introduction.

Suppose we want to find the real roots (especially those in $[-1,1]$) of a polynomial expressed by its Chebyshev coefficients

$$p(x) = \sum_{i=0}^{n} \gamma_i T_i(x).$$

Or more generally, $p(x)$ may be expressed in terms of polynomials $\{\phi_m(x)\}_{m\geq 0}$, each $\phi_m(x)$ of exact degree m, that are orthogonal with respect to an inner product, e.g.,

$$\langle f, g\rangle_\rho = \int_a^b f(x)\overline{g(x)}\rho(x)dx \tag{1.1}$$

for some real and positive weight function $\rho(x)$.

One way to find the roots of $p(x)$ is to express $p(x)$ as a sum of monomials, and then to calculate the roots as the eigenvalues of the standard companion matrix. However, expressing a polynomial by its monomial coefficients is not as well conditioned as the expression in terms of Chebyshev polynomials. The transformation between a polynomial of degree n in $[-1,1]$ and its expansion coefficients with respect to the monomials [12] has $\mathcal{O}((1+\sqrt{2})^{n+1})$ condition number with respect to maximum norms (over $[-1,1]$) and with respect to Chebyshev polynomials [10] has $\mathcal{O}(n)$ condition number.

*Received by the editors June 10, 2004; accepted for publication (in revised form) March 18, 2005; published electronically December 8, 2005. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sinum/43-5/60984.html

†Computational Mathematics and Algorithms Department, Sandia National Laboratories, P.O. Box 5800, Albuquerque, NM 87185-1110 (dmday@sandia.gov, lromero@sandia.gov).

For the case of Chebyshev polynomials, Boyd [6] and also Battles and Trefethen [2] have proposed solving this problem by projecting to the unit circle in the complex z-plane with $x = (z + z^{-1})/2$, and using the fact that $T_k(x) = \cos(k \cos^{-1}(x))$. Their technique allows them to find the roots of $p(x)$ in terms of a standard companion matrix that depends on the coefficients $\{\gamma_k\}_{k=0}^{n}$ of the orthogonal expansion. These authors have found that this is a very successful algorithm, but the trouble is that it makes use of an eigenvalue problem of size $2n$ for a rootfinding problem of size n.

The present manuscript proposes an alternative formulation based on a nonstandard companion matrix $\mathbf{B}_n$ of dimension n. The algorithm is an extension of the technique [13] for finding the roots of the nth orthogonal polynomial $\phi_n(x)$. The technique uses the fact that any set of orthogonal polynomials satisfies a recurrence formula of the form

$$x\phi_{n-1}(x) = \sum_{i=0}^{n} \phi_i(x) h_{i,n-1}. \tag{1.2}$$

The coefficients determine an n by n matrix $\mathbf{H}_n = [h_{i,j}]_{0 \le i,j < n}$ whose eigenvalues are the roots of the nth orthogonal polynomial $\phi_n(x)$. For ortho*normal* polynomials based on certain inner products such as (1.1), $\mathbf{H}_n$ is symmetric and tridiagonal. For a general inner products, $\mathbf{H}_n$ is upper Hessenberg, that is, $h_{i,j} = 0$ for $i > j + 1 > 0$.

As a specific example, the Chebyshev polynomials satisfy the three-term recurrence $T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$ for $n \ge 1$, or recast in the form of (1.2), $xT_n(x) = T_{n-1}(x)\frac{1}{2} + T_{n+1}(x)\frac{1}{2}$. In this case, there holds $h_{0,1} = \frac{1}{2}$, $h_{0,1} = 1$ for $i > 0$, $h_{i,i+1} = h_{i+1,i} = \frac{1}{2}$, and otherwise $h_{i,j} = 0$. The asymmetry of $h_{0,1}$ and $h_{1,0}$ reflects the nonconstant normalization of $\{T_k\}_{k \ge 0}$: for $\rho(x) = \sqrt{1 - x^2}$ there holds $\langle T_k, T_k \rangle_\rho = \frac{\pi}{4}(1 + \delta_{k,0})$.

Our technique for finding the roots of $p(x)$ is a modification of the technique for finding the roots of $\phi_n(x)$. To express our result we use the notation

$$\mathbf{f}_n(x) = [\phi_0(x), \dots, \phi_{n-1}(x)]^T \tag{1.3}$$

for the column vector-valued function containing the first n orthogonal polynomials, and the notation

$$\mathbf{c}^T = [\gamma_0, \gamma_1, \dots, \gamma_{n-1}] \tag{1.4}$$

for the column vector containing the first n coefficients of the polynomial $p(x)$. Using this notation we have

$$p(x) = \mathbf{f}_n(x)^T \mathbf{c} + \gamma_n \phi_n(x). \tag{1.5}$$

In section 2, Theorem 2.3 shows that the roots of $p(x)$ are the eigenvalues of the nonstandard companion matrix

$$\mathbf{B}_n = \mathbf{H}_n - h_{n,n-1} \frac{\mathbf{c}}{\gamma_n} \mathbf{e}_{n-1}^T, \tag{1.6}$$

where $\mathbf{e}_{n-1} = [0, \dots, 0, 1]^T$ is a column vector of dimension n. When applied to finding roots of polynomials expressed in terms of Chebyshev polynomials for large values of n the new method promises to be something like eight times faster than the method proposed by Boyd and Battles and Trefethen. It is somewhat faster than the direct conversion to a monomials (without doubling the degree), which is unstable for large values of n.

Orthogonal polynomials have many applications. Transcendental equations may be solved with Chebyshev polynomials, as was proposed in [4] and developed further in the follow-up papers [5] and [6]. Battles and Trefethen automate, through MATLAB calls, a suite of operators on functions. The implementation is accomplished using Chebyshev polynomials of very high degree. And the operator that finds the real roots of a function is (now) implemented along the lines described here. Battles and Trefethen have pointed out that certain applications of polynomials based on the monomial form may be significantly improved by using another form based on a specific family of orthogonal polynomials.

Although the technique we present in this paper finds all of the roots of the polynomial $p(x)$, we will see that it only has desirable stability properties for finding roots in an appropriate region of the complex plane. For example, for Chebyshev polynomials we only have desirable stability properties for finding real roots in or near the interval $[-1, 1]$. Similarly, transcendental equation solvers based on the rootfinding by Chebyshev expansions have desirable stability properties only for roots in or near interval $[-1, 1]$ (see Theorem 4.2).

When we approximate a transcendental function in terms of an orthogonal polynomial expansion, the highest order coefficient γ_n converges to zero (see Theorem 4.1). For this reason, many cases of interest are near the division by zero singularity in (1.6) for $\mathbf{B}_n$. The singularity is avoided by solving a generalized eigenvalue problem as described in section 2 or [19]. However, in Theorem 4.2, we will show that if a transcendental function is approximated as a finite sum of Jacobi polynomials, the roots found by using the corresponding matrix $\mathbf{B}_n$ accurately approximate the transcendental equation roots in or near $[-1, 1]$.

If the cost of solving the eigenvalue problem becomes a computational bottleneck, then one may use a subdivision algorithm (see [7]) that decomposes the rootfinding problem into several subproblems and applies Chebyshev polynomials of lower order in each subinterval.

1.1. Summary. We begin in section 2 by reviewing a process for finding the roots of the nth orthogonal polynomial $\phi_n(x)$ as the eigenvalues of the matrix $\mathbf{H}_n$. We then show how to modify this process to construct the nonstandard companion matrix $\mathbf{B}_n$ whose eigenvalues are given by the roots of the polynomial $p(x)$ (cf. Theorem 2.3). Although classical orthogonal polynomials are emphasized over all, we abstractly define “orthogonal polynomial” (see Definition 2.1) so that our results include the monomials, and hence our results include the standard companion matrix. Lemma 2.4 presents analytical expressions for both the left and right eigenvectors in terms of the eigenvalues. In section 3 the sensitivities of polynomial roots and matrix eigenvalues are compared. Theorem 3.3 demonstrates how eigenvalue and polynomial root sensitivities coincide in certain cases.

The algorithms presented herein are not so much new as they are not widely known. The companion matrices for orthogonal polynomials were independently discovered by Hans Stetter. For a derivation of the nonstandard companion matrix based on quotient rings in algebraic geometry (see [21]). Exercise 1c on p. 148 of [21] asks the reader to derive the companion matrix for Chebyshev polynomials. On the other hand, Stetter emphasizes application to polynomials of modest degree, say 10 (cf. p. 146). The observation that the roots of the nth member of a family of orthogonal polynomials must be the eigenvalues of a companion matrix whose elements come from the coefficients of the recurrence relation for the orthogonal polynomials was well known to Jacobi [13]. Like Stetter, we show how to define “orthogonal” polynomial broadly enough to apply the observation to any polynomial. Our contribution is some analysis of the numerical stability of such methods. Example 3 of section 5 uses a degree 256 polynomal to solve a transcendental equation.

In order to concentrate on issues of interest in applications using orthogonal polynomials, we discuss the representative application of finding the roots of a scalar transcendental equation in a real interval. Representing a function by the partial sum of an exponentially convergent orthogonal expansion raises issues that must be addressed. In particular, ill-conditioning is manifested in the roots that we do not want. However, roots in a specific domain of the complex plane are well conditioned in a certain sense.

In section 3.3 representation with respect to Chebyshev polynomials, or any Jacobi polynomial, are shown to be ideal for finding roots in or quite near $[-1, 1]$. Away from $[-1, 1]$, the Jacobi polynomials are not recommended. The prerequisite results for classical orthogonal polynomials are reviewed. It is shown that for rootfinding in an interval, representing polynomials with respect to Jacobi orthogonal polynomials is ideal. But monomials are better for rootfinding in the unit disk. In particular the algorithms described herein are not designed to find all of the roots of a polynomial.

In section 4 we discuss how matrix balancing is desirable in computing the eigenvalues of $\mathbf{B}_n$. The upper Hessenberg structure of $\mathbf{B}_n$ is crucial in the explanation of the success of matrix balancing. Theorem 4.2 shows how partial sums of orthogonal expansions lead to companion matrices that are amenable to matrix balancing. We use our analytical expressions for the left and the right eigenvalues to show that the polynomial and eigenvalue sensitivities differ by a computable (and benign) factor, related to the associated Lagrange interpolation polynomials. The companion matrix formulation is numerically stable in this case.

Analysis is also included intended for a posteriori use in solving transcendental equations. An algorithm for finding the roots of a transcendental equation in $[-1, 1]$ using expansions in terms of Chebyshev polynomials is presented in section 4.2. Numerical experiments are presented in section 5 that demonstrate the reliability of the algorithm. Our results are summarized in section 6.

For expansions of transcendental equations, we explain why the companion matrix is amenable to balancing. The exponential convergence rate is related to the distance to the nearest singularity of the locally analytic function, and also applies to the (right) eigenfunctions. Balancing "factors out" the dependence of $\mathbf{B}_n$ on $1/\gamma_n$, and the balanced companion matrix eigenvalue problem is numerically stable. An algorithm for finding the roots of a transcendental equation on $[-1, 1]$ using expansions in terms of Chebyshev polynomials is presented in section 4.2. Numerical experiments are presented in section 5 that demonstrate the reliability of the algorithm.

2. Companion matrices.

Starting from a general definition of orthogonal polynomials, we review the procedure for finding the roots of orthogonal polynomials as the eigenvalues of the matrix $\mathbf{H}_n$ containing the coefficients in the recurrence formula. The discussion closely follows [13]. Next we construct a nonstandard companion matrix corresponding to a sequence of orthogonal polynomials and a given polynomial. In Theorem 2.3 we establish the equivalence between the roots of the polynomial equation and the companion matrix spectrum. In Lemma 2.4 we give an analytical expression for the right eigenvectors of $\mathbf{B}_n$. The proof exploits the connection between Vandermonde matrices and Lagrange interpolation polynomials.

Orthogonal polynomials are broadly defined here to emphasize the connection between the numerical stability of a companion matrix eigenvalue problem and the associated inner product. There is a one-to-one correspondence between inner products on polynomials and the set of sequences of univariate polynomials $\{\phi_i(x)\}_{i\geq 0}$ such that each $p_k(x)$ has degree k. The polynomials are orthonormal with respect to the polynomial inner product that is the ordinary vector inner product of expansion coefficients.

We will work over the space of complex valued continuous functions on a bounded subdomain of the complex plane.

DEFINITION 2.1. *With respect to the inner product* $\langle , \rangle$ *the sequence* $\{\phi_n(x)\}_{n\geq 0}$ *are orthogonal polynomials if each* $\phi_n(x)$ *is a polynomial of exact degree* n *and* $\langle \phi_n, \phi_m\rangle = \delta_{n,m}\sigma_n^2$. *Here* $\delta_{i,j}$ *is Kronecker's delta and* $\{\sigma_n\}_{n\geq 0}$ *is a sequence of positive real numbers. The polynomials* $\{\phi_n(x)\}_{n\geq 0}$ *are orthonormal if each* σ_n *is one. The norm induced by the inner product is denoted by* $\||\ \psi\ \|| = \langle \psi, \psi\rangle^{1/2}$.

Orthogonality implies that for $i \leq n-1$, $\||\ \phi_i\ \||^2 \ \ h_{i,n-1} = \langle \phi_i(x), x\phi_{n-1}(x)\rangle$.

Usually when discussing orthogonal polynomials we will be concerned with inner products of the form in (1.1). Orthonormal polynomials with respect to this type of inner product must satisfy a symmetric three-term recurrence formula. This is a consequence of the fact that such an inner product is symmetric with respect to multiplication; that is, $\langle xf(x), g(x)\rangle = \langle f(x), xg(x)\rangle$.

For rootfinding problems over bounded complex domains, we recommend the inner product that arises in Bergman's theory of (reproducing) kernel functions (see [18, p. 36], [22, section 11.2], or [17, Lemma 17.2.3]). For example, the monomials are orthogonal polynomials with respect to the inner product

$$\langle f, g\rangle = \frac{1}{2\pi i}\int_\Gamma f(z)\overline{g}(z)dz,$$

where the integral is taken over the circle Γ centered around the origin in the complex plane. Note that this inner product is not symmetric with respect to multiplication.

2.1. Roots of orthogonal polynomials: A review.

A way to find the roots of the nth orthogonal polynomial $\phi_n(x)$ uses the recurrence formula in (1.2). The first n instances of (1.2) combine using the matrix $\mathbf{H}_n$, the column vector $\mathbf{f}_n$, and the coefficient $h_{n,n-1}$ into the matrix equation

$$x\mathbf{f}_n^T(x) = \mathbf{f}_n^T(x)\mathbf{H}_n + \phi_n(x)h_{n,n-1}\mathbf{e}_{n-1}^T. \tag{2.1}$$

Equation (2.1) exposes the equivalence between the roots ξ root of $\phi_n(x) = 0$ and the eigenvalues of $\mathbf{H}_n$,

$$\xi\mathbf{f}_n^T(\xi) = \mathbf{f}_n^T(\xi)\mathbf{H}_n.$$

We conclude with the following result that Gautschi [13] attributes to Jacobi.

THEOREM 2.2. *The algebraic eigenvalues of* H_n *defined in* (2.1) *coincide with the algebraic roots of the degree* n *orthogonal polynomial* $\phi_n(x)$.

Proof. The result is a corollary of Theorem 2.3. □

2.2. Nonstandard companion matrices.

Assuming that $\gamma_n \neq 0$, we can use (1.5) to express $\phi_n(x)$ as

$$\phi_n(x) = \frac{p(x) - \mathbf{f}_n(x)^T\mathbf{c}}{\gamma_n}. \tag{2.2}$$

If we substitute this expression for $\phi_n(x)$ into (2.1), we arrive at the equation

$$x\mathbf{f}_n^T(x) = \mathbf{f}_n^T(x)\mathbf{H}_n + \frac{p(x) - \mathbf{f}_n(x)^T\mathbf{c}}{\gamma_n}h_{n,n-1}\mathbf{e}_{n-1}^T. \tag{2.3}$$

We now see that if ξ is a root of $p(x) = 0$, then

$$\xi \mathbf{f}_n^T(\xi) = \mathbf{f}_n(\xi)^T \mathbf{B}_n, \tag{2.4}$$

where as in (1.6)

$$\mathbf{B}_n = \mathbf{H}_n - h_{n,n-1} \frac{\mathbf{c}}{\gamma_n} \mathbf{e}_{n-1}^T.$$

This shows that if ξ is a root of $p(x)$, then it must be an eigenvalue of $\mathbf{B}_n$ with left eigenvector $\mathbf{f}_n^T(\xi)$. The converse is established in Theorem 2.3.

Equivalently, we could express the first n terms of our recurrence formula as

$$x\mathbf{f}_{n+1}(x)^T = \mathbf{f}_{n+1}^T(x) \begin{bmatrix} \mathbf{H}_n \\ h_{n,n-1}\mathbf{e}_{n-1}^T \end{bmatrix}.$$

When we combine this with the requirement that $p(x) = 0$ using (1.5), we get the system of equations

$$\mathbf{f}_{n+1}^T(\xi) \begin{bmatrix} \mathbf{H}_n & \mathbf{c} \\ h_{n,n-1}\mathbf{e}_{n-1}^T & \gamma_n \end{bmatrix} = \xi \mathbf{f}_{n+1}^T(\xi) \begin{bmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix}. \tag{2.5}$$

Any root ξ of $p(x) = 0$ must be an eigenvalue of this generalized eigenvalue problem with left eigenvector $\mathbf{f}_{n+1}(\xi)$.

As is the case for companion matrices, one may either solve the generalized eigenvalue problem in (2.5) and discard an infinite eigenvalue or find the eigenvalues as defined in (2.4).

The standard backward stable algorithms for generalized and ordinary eigenvalue problems in (2.5) and (2.4) are the QZ and QR algorithms, respectively. Because of the upper Hessenberg form of these matrices, no initial transformation to Hessenberg form is required for either QZ or QR. For computing eigenvalues only in the average case, QR is three times faster than QZ [14]. The question of which formulation to use is not an entirely solved problem. The fact that $\mathbf{B}_n$ may have a very large norm suggests that the formulation of (2.5) has superior stability properties. Numerical experiments do not confirm this hypothesis. Polynomial equations for which the formulation of (2.5) is advantageous do exist [19], but do not arise in the solution of transcendental equations. We performed numerical experiments comparing the residual norms of the polynomials evaluated at the eigenvalues computed by either QZ or QR. We observed that if QR is used without balancing, then the eigenvalues computed from the ordinary eigenvalue problem suffer roundoff errors proportional to $\|\mathbf{c}\|/\gamma_n$. QR with balancing and QZ always computed eigenvalues of the same quality even if $\|\mathbf{c}\|/\gamma_n$ is very large. Explanations are provided in Theorem 3.3 and in section 4.

Next the equivalence of the roots of the polynomial $p(x)$ and the eigenvalues of the matrix $\mathbf{B}_n$ defined in (1.6) and (2.4) is demonstrated.

THEOREM 2.3. *The roots of a polynomial p of exact degree n coincide with the eigenvalues of the generalized companion matrix $\mathbf{B}_n$ counting algebraic multiplicity.*

Proof. We have already shown that a root ξ of $p(x)$ is an eigenvalue of $\mathbf{B}_n$ with left eigenvector $\mathbf{f}_n(\xi)$. The converse follows from two properties of unreduced upper Hessenberg matrices, including $\mathbf{B}_n - \xi I_n$ for any ξ: first a nontrivial (right) null vector has nonzero last component, and second the nullity is at most one. The properties of Hessenberg matrices are discussed in [14, section 7.4.5] and, in particular, [14, Theorem 7.4.4]. If ξ is an eigenvalue of $\mathbf{H}_n$ with nontrivial (right) eigenvector $\mathbf{v} = [v_0, \ldots, v_{n-1}]^T$, then $v_{n-1} \neq 0$. Substitution of (2.2) into (2.1) yields

$$x\mathbf{f}_n^T(x) = \mathbf{f}_n^T(x)\left(\mathbf{H}_n - \frac{\mathbf{c}}{\gamma_n} h_{n,n-1}\mathbf{e}_{n-1}^T\right) + \frac{p(x)}{\gamma_n} h_{n,n-1}\mathbf{e}_{n-1}^T. \tag{2.6}$$

Inspection of the product of (2.6) and $\mathbf{v}$ implies that $p(\xi) = 0$. As a consequence of the second property, a left eigenvector of ξ must be proportional to $\mathbf{f}_n(\xi)$. □

2.3. The right eigenvectors. We have already shown that if ξ_j is the jth root of the polynomial $p(x)$, then $\mathbf{v}_j^T = \mathbf{f}_n^T(\xi_j)$ is the left eigenvector associated with the eigenvalue ξ_j of $\mathbf{B}_n$. We will now give a simple expression for the right eigenvectors $\mathbf{w}_j$ associated with this eigenvalue.

The matrix of left eigenvectors, $\mathbf{V}$, has jth row $\mathbf{v}_j$. The ith row of the inverse contains the right eigenvector $\mathbf{w}_i$. Note that the n by n matrix $\mathbf{V} = [\nu_{i,j}]$ is called a generalized Vandermonde matrix due to $\nu_{i,j} = \phi_j(\xi_i)$.

The right eigenvectors can be expressed using the interpolating polynomials. Assuming that ξ_j is a simple root of $p(x)$, the jth Lagrange interpolating polynomial associated with the roots of $p(x)$ is

$$l_j(x) = \frac{p(x)}{p'(\xi_j)(x - \xi_j)}.$$

Each interpolating polynomial $l_j(x)$ has degree $n - 1$ and satisfies $l_j(\xi_k) = \delta_{jk}$ for $0 \leq j, k < n$.

Each polynomial $l_j(x)$ has degree $n-1$ and is a linear combination of $\{\phi_k(x)\}_{k=0}^n$. Define the column vector $\mathbf{w}_j$ to contain the expansion coefficients of $l_j(x)$,

$$l_j(x) = \mathbf{f}_n^T(x)\mathbf{w}_j.$$

It follows that $\delta_{ij} = l_j(\xi_i) = \mathbf{f}_n^T(\xi_i)\mathbf{w}_j = \mathbf{v}_i^T\mathbf{w}_j$. This shows that the vector $\mathbf{w}_i$ is in fact the ith column of the inverse matrix of $\mathbf{V}$, and hence $\mathbf{w}_i$ is the right eigenvector associated with the eigenvalue ξ_i. This proves the following lemma.

LEMMA 2.4. *The companion matrix $\mathbf{B}_n$ defined in (2.4) corresponding to a polynomial $p(x)$ of exact degree n and with distinct roots, $\{\xi_j\}_{0\leq j<n}$, has as eigenvalues the roots of $p(x)$. The left eigenvector corresponding to ξ_k is $\mathbf{f}_n(\xi_k)$. Moreover, if $\{\phi_j\}$ are orthonormal polynomials, then for $l_j(x) = \frac{p(x)}{p'(\xi_j)(x-\xi_j)}$ the jth right eigenvector $\mathbf{w}_j$ has components $\mathbf{e}_k^T\mathbf{w}_j = \langle l_j, \phi_k\rangle$ and $\|\mathbf{w}_j\|_2 = |||l_j|||$.*

Proof. The comments directly before the lemma establish the following: the n by n matrix $\mathbf{W} = [\mathbf{w}_0, \ldots, \mathbf{w}_{n-1}]$ whose jth column, $\mathbf{w}_j$, contains the expansion coefficients of $l_j(x)$ with respect to $\{\phi_j(x)\}_{0\leq j<n}$, i.e., $l_j(x) = \mathbf{f}_n^T(x)w_j$, is $\mathbf{W} = \mathbf{V}^{-1}$, thus is the matrix of right eigenvectors. Next take the inner product of $l_j(x)$ with $\phi_k(x)$. By Definition 2.1, $\langle l_j, \phi_k\rangle = |||\phi_k|||^2 e_k^T w_j$. The desired representation follows in the orthonormal case. Parseval's formula readily furnishes the equivalence between the norms. □

The relationship between $|||l_j|||$ and $p(x)$ is further developed in section 4.

3. Sensitivity analysis.

Rootfinding by eigenvalue problems is popular due to its favorable stability properties compared to other methods. A stability analysis for the standard companion matrix formulation has been performed in [23]. The companion matrix form is viewed as a rank one perturbation of a bidiagonal matrix, and the inverses of the shifted companion matrices are analyzed. Here we perform a local analysis of the nonstandard companion matrix, assessing stability by comparing the polynomial root and eigenvalue the polynomial root sensitivities. The limitation of a local analysis is that it is only valid for eigenvalues that are well separated.

3.1. Polynomial root sensitivity. The perturbation theory for polynomial roots is considered following [11]. We consider some particular zero ξ of $p(x)$ as a complex valued function of the expansion coefficients with respect to the orthogonal polynomials. Denote the dependence on the coefficients $\mathbf{c}$ by $\xi(\mathbf{c})$.

LEMMA 3.1. *Suppose that a polynomial of exact degree n, $p(x) = \mathbf{f}_{n+1}^T(x)\mathbf{c}_o$, where $\mathbf{c}_o$ is an $(n+1)$-dimensional column vector, has a simple root ξ_o such that $p(\xi_o) = 0 \neq p^{'}(\xi_o)$. There exists a smooth function $\xi(\mathbf{c})$ such that $\xi(\mathbf{c}_o) = \xi_o$ and $\mathbf{f}_{n+1}^T(\xi(\mathbf{c}))\mathbf{c} = 0$ with gradient $\nabla_{\mathbf{c}}\xi(\mathbf{c}_o) = -\mathbf{f}_{n+1}^T(\xi_o)/p^{'}(\xi_o)$.*

Proof. See Example 3.10 in [21] for the proof. □

We will discuss several aspects of Lemma 3.1, starting with the excluded case of multiple roots. In the case of a root of algebraic multiplicity $m > 1$, there exist infinitesimal coefficient perturbations that change the multiplicity, and the roots are Hölder continuous with exponent $1/m$ (see Proposition 5.1 in [21] or Theorem 4.1 in [6]). Approximations to roots with nontrivial multiplicity correspond to small values of $p(\xi_o)$. Monitoring the value of the polynomial derivative at all approximate roots of interest for small values is required.

Next, the implication of Lemma 3.1 is that the norm of the gradient of a root with respect to the coefficients is proportional to $\|\mathbf{f}_{n+1}(\xi_o)\|_2$. In other words, a root ξ_o is not very sensitive to the polynomial coefficients if both $\|\mathbf{f}_{n+1}(\xi_o)\|_2$ is not "large" and $|p'(\xi_o)|$ is not "small." For all the orthogonal polynomials familiar to the authors, for sufficiently large x, $\|\mathbf{f}_{n+1}(x)\|_2$ grows exponentially as a function of n. This observation reflects the intrinsic difficulty of finding all the roots of an arbitrary polynomial. On the other hand, using a Jacobi polynomial, for a root $-1 \leq \xi_o \leq 1$, $\|\mathbf{f}_{n+1}(\xi_o)\|_2$ is not "large" (clarified in section 3.3) and no well-separated root is very sensitive to the coefficients.

3.2. Eigenvalue sensitivity. In section 3.1 we showed that the polynomial root sensitivity with respect to the coefficients is $\|\mathbf{f}_{n+1}\|$. Here the condition number of a simple eigenvalue is related to the corresponding left and right eigenvectors $\mathbf{f}_n = \mathbf{v}$ and $\mathbf{w}$. A standard result from the perturbation theory of simple eigenvalues is that under an infinitesimal perturbation $\delta\mathbf{A}$ of a matrix $\mathbf{A}$, a simple eigenvalue λ changes to $\lambda + \delta\lambda$, where $\delta\lambda = \mathbf{v}^T\delta\mathbf{A}\mathbf{w}\ /\mathbf{v}^T\mathbf{w}$. Lemma 3.2 restates the result without using infinitesimals.

LEMMA 3.2. *If a square matrix $\mathbf{A}$ has a simple eigenvalue λ with corresponding left $\mathbf{v}^T$ and right $\mathbf{w}$ eigenvectors, $\mathbf{v}^T\mathbf{A} = \lambda\mathbf{v}^T$ and $\mathbf{A}\mathbf{w} = \mathbf{w}\lambda$, then $\nabla_{\mathbf{A}}\lambda = \mathbf{v}\mathbf{w}^T/\mathbf{v}^T\mathbf{w}$.*

Proof. See [14, p. 344] for the proof. □

Next Theorem 3.3 gives a sufficient condition for the numerical stability of root-finding based on a nonstandard companion matrix eigenvalue problem. It suffices for the computed eigenvalues to be the eigenvalues of $\mathbf{B}_n + \mathbf{E}_n$ nearby to $\mathbf{B}_n$ in a componentwise or relative sense. That is, there exists a tiny $\tau > 0$ such that $|\mathbf{E}_n| \leq \tau|\mathbf{B}_n|$. The result applies for any nonzero γ_n. The idea of the proof is that the factor of $1/\gamma_n$ in column n of $\mathbf{E}_n$ cancels with a factor of γ_n in row n of $\mathbf{w}$.

THEOREM 3.3. *Suppose that the degree n polynomial $p(x)$ has a simple root λ. Recall the notation of* (1.5), *in particular the definition of the column* $\mathbf{c}$ *in* (1.4), *and the definition of* $\mathbf{H}_n$ *in* (2.1). *The corresponding companion matrix* $\mathbf{B}_n$ *has left* *and right eigenvectors* $\mathbf{v}^T = \mathbf{f}_n(\lambda)^T$ *and* $\mathbf{w}$ *as in Lemma* 2.4 *so that* $\mathbf{v}^T\mathbf{w} = 1$. *A perturbation* $\mathbf{E}_n$ *of* $\mathbf{B}_n$ *such that* $|\mathbf{E}_n| \le \tau|\mathbf{B}_n|$ *perturbs* λ *by* $\delta\lambda$ *such that*

$$|\delta\lambda| \le \tau|\mathbf{v}|^T|\mathbf{H}_n|\ |\mathbf{w}| + \tau|\mathbf{v}|^T|\mathbf{c}|/|p'(\lambda)| + \mathcal{O}(\tau^2).$$

Proof. We start by establishing the following claim. For $h_{n,n-1}$ defined in (1.2) there holds

$$h_{n,n-1}\mathbf{e}_{n-1}^T\mathbf{w} = \gamma_n/p'(\lambda). \tag{3.1}$$

Substitute the expansion $l(x) = \mathbf{f}_n(x)\mathbf{w}$ below, and simplify to find that $\gamma_n = \langle p, \phi_n\rangle = \langle l(x)(x-\lambda)p'(\lambda), \phi_n\rangle = \langle l(x)xp'(\lambda), \phi_n\rangle = \langle \phi_{n-1}(x)xp'(\lambda), \phi_n\rangle\mathbf{e}_{n-1}^T\mathbf{w}$. Rewrite (1.2) in matrix form,

$$x\phi_{n-1}(x) = \mathbf{f}_n^T(x)\mathbf{H}_n e_{n-1} + \phi_n(x)h_{n,n-1}. \tag{3.2}$$

Equation (3.2) implies that $\gamma_n = h_{n,n-1}p'(\lambda)\mathbf{e}_{n-1}^T\mathbf{w}$. Divide by $p'(\lambda)$ to establish the claim.

By Lemma 3.2, to first order in τ there holds $|\delta\lambda| = |\mathbf{v}^T\mathbf{E}_n\mathbf{w}| \le |\mathbf{v}|^T|\mathbf{E}_n||\mathbf{w}| \le \tau|\mathbf{v}|^T|\mathbf{B}_n||\mathbf{w}|$. The proof is completed by substituting (1.6), applying the triangle inequality, and then the claim. □

Theorem 4.2 will show how for transcendental equations, QR iteration with balancing solves a nearby eigenvalue problem, $\mathbf{B}_n + \mathbf{E}_n$, such that only the last column of $\mathbf{E}_n$ is proportional to $1/\gamma_n$.

3.3. Classical orthogonal polynomials. In general, the term $\|\mathbf{f}_{n+1}(\xi)\|_2$ arising in polynomial roots sensitivity is associated with the "kernel polynomials." The kernel polynomials are defined by $\mathbf{K}_n(x_o, x) = \overline{\mathbf{f}}_{n+1}^T(x_o)\ \mathbf{f}_{n+1}(x)$. If x_o is a constant, then $\mathbf{K}_n(x_o, x)$ is a polynomial. The kernel polynomials maximize the ratio $|p(x_o)|/\|\|p(x)\|\|$ over all polynomials of exact degree n (see [22, Theorem 3.1.3]), and the maximum ratio is $\|\mathbf{f}_{n+1}(x_o)\|_2$. As we shall see, the asymptotic properties of the kernel polynomials indicates that the classical orthogonal polynomials over $[-1, 1]$, the Jacobi polynomials, are suitable for rootfinding. And conversely, for rootfinding over domains that are topologically different from intervals, none of the classical polynomials orthogonal over an interval is desirable, and a different inner product is needed [1, 9].

The Jacobi polynomials are orthogonal with respect to the inner product

$$\langle f, g\rangle_{\alpha,\beta} = \int_{-1}^{1} f(t)g(t)(1-t)^\alpha(1+t)^\beta dt$$

for $\alpha > -1, \beta > -1$ (see [22, section 2.4]). The case $\alpha = \beta = -1/2$ corresponds to the Chebyshev polynomials (of the first kind). The comparison of spectral methods for partial differential equations based on either Chebyshev polynomials or Legendre polynomials ($\alpha = \beta = 0$) in [3] demonstrates the advantages of Legendre polynomials. The Legendre orthonormal polynomials satisfy the three-term recurrence $\gamma_{n+1}\phi_{n+1}(x) = x\phi_n(x) - \gamma_n\phi_{n-1}(x)$ for $\gamma_n = n/\sqrt{4n^2-1}$. By rearranging the three-term recurrence into the form of (1.2), one may show that for $i \ge 0$, $h_{i+1,i} = h_{i,i+1} = \gamma_{i+1}$, and otherwise $h_{i,j} = 0$.

Theorem 7.71.2 in [22] states that $\max_{-1\le x_o\le 1}\|\mathbf{f}_{n+1}(x_o)\|_2 = \mathcal{O}(n^\kappa)$ for $\kappa = \max(\alpha+1, 1/2)$. The result is a consequence of the connection to Sturm–Liouville problems. The exponent is minimal for the Chebyshev polynomials, and for any admissible (α, β) the sensitivity bound grows like a polynomial. On the other hand, away from the interval $[-1, 1]$, the Jacobi polynomials grow exponentially with degree and are undesirable for rootfinding problems.

4. Solving transcendental equations.

We will discuss in depth a representative application of nonstandard companion matrices to rootfinding problems involving polynomial equations expressed with respect to Chebyshev polynomials. We discuss the solution of a transcendental equation $\psi(\xi) = 0$. Section 4.2 presents a complete description of the corresponding algorithm for a scalar transcendental equation. In numerical experiments, we find that the algorithm is numerically stable if matrix balancing is used (the default in MATLAB). However, it is crucial to use balancing with the QR algorithm in solving the eigenvalue problems that arise in the solution of transcendental equations.

4.1. Analysis of balancing.

Next some of the linear algebra issues associated with computing the eigenvalues of the B matrices are discussed in detail. Readers interested only in the solution of transcendental equations may choose to skip the section.

As a sequence of polynomials converge uniformly to $\psi(x)$ on some bounded domain, certain roots of the polynomials converge to $\{\xi : \psi(\xi) = 0\}$ [5]. The order of the approximation, n, is chosen to be sufficiently large such that the trailing γ_n is negligible [5]. For solving transcendental equations, balancing the generalized companion matrix (cf. [23]) usually employs an alarmingly ill-conditioned diagonal similarity transform, and extraordinarily reduces the condition number of the matrix of eigenvectors. Theorem 4.2 presents an explanation of the success of balancing for companion matrices arising in the solution of transcendental equations.

A transcendental equation $\psi(\zeta) = 0$ arises from $\psi(x)$ that is analytic in an open simply connected domain containing $[-1, 1]$. The orthogonal polynomials used are eigenfunctions of singular Sturm–Liouville problems in $[-1, 1]$, namely, the Jacobi polynomials (corresponding to one value of (α, β)) and here denoted $\{\phi_n(x)\}_{n\geq 0}$. The simplest case, $\{\phi_n(x)\}_{n\geq 0}$ orthonormal, is discussed. The convergence properties of Jacobi series expansion $\sum_{n\geq 0} \phi_n(x)\gamma_n$ with $\gamma_n = \langle \psi, \phi_n \rangle$ of $\psi(x)$ is described by Theorem 4.1.

Theorem 4.1. *Let $\psi(x)$ be an analytic function with an open domain containing $[-1, 1]$. The expansion of $\psi(x)$ in a Jacobi series is convergent in the interior of the greatest ellipse with foci at ± 1, in which $\psi(x)$ is regular. The expansion is divergent in the exterior of the ellipse. If $\psi(x) = \sum_{n\geq 0} \gamma_n \phi_n$, then we have the following representation of the sum R of the semiaxes of ellipse of convergence $R = \liminf_{n\to+\infty} |\gamma_n|^{-1/n}$.*

Proof. See [22, Theorem 9.1.1] for the proof. □

In Theorem 4.1, $R = A + B$ for an ellipse $(x/A)^2 + (y/B)^2 = 1$ in the (x, y)-plane with $A^2 = B^2 + 1$ and $A > B > 0$ (see [15, p. 37]). Roughly speaking, there holds $\sum_{j\geq n} |\gamma_j|^2 = \mathcal{O}(R^{-2n})$. The hypothesis that $\psi(x)$ is an analytic function on an open domain containing $[-1, 1]$ ensures that for the greatest ellipse $B > 0$ and $R > 1$.

An analytic function $\psi(x)$ with root ξ, $\psi(\xi) = 0$, has a Jacobi series. The Jacobi series has partial sums of the form $p_n(x) = \sum_{j=0}^{n} \phi_j(x)\gamma_j$. Each $p_n(x)$ has at least one root ξ_n nearest to ξ. In the case $R > 1$, Theorem 4.1 implies that in $[-1, 1]$, $\{p_n(x)\}_{n\geq 0}$ converges uniformly to $\psi(x)$. Furthermore, each derivative $p_n^{(m)}(x)$ converges uniformly to $\psi^{(m)}(x)$ in $[-1, 1]$.

If ξ is a simple root of $\psi(x)$ (i.e., $\psi'(\xi) \neq 0$), then $\xi_n \to \xi$. Moreover, for

n sufficiently large that $\psi'(\xi_n)\psi'(\xi) > (\psi'(\xi))^2/2$ the relationship between residual error and approximate solution error implies that $\xi_n - \xi = \mathcal{O}(R^{-n})$.

The algebraic eigenvalue problem $\mathbf{B}_n\mathbf{W}_n = \mathbf{W}_n\Lambda_n$ is solved by applying the QR algorithm to the balanced generalized companion matrix. In MATLAB, the default configuration of the QR algorithm applies balancing. Balancing refers to determining a diagonal matrix $\mathbf{\Sigma}_n$ such that the similar eigenvalue problem $\mathbf{\Sigma}_n^{-1}\mathbf{B}_n\mathbf{\Sigma}_n$ is (hopefully) much better conditioned. A nearly optimal diagonal similarity transformations $\mathbf{\Sigma}_n$ produces $\mathbf{\Sigma}_n^{-1}\mathbf{W}_n$ with equal row norms (see [16, section 12]), but $\mathbf{W}_n$ is not known a priori. Instead a diagonal similarity transformation $\mathbf{\Sigma}_n$ that nearly minimizes a norm of $\mathbf{\Sigma}_n^{-1}\mathbf{B}_n\mathbf{\Sigma}_n$ is determined.

To illustrate matrix balancing, consider $\mathbf{B}_4$ whose coefficients are chosen to reflect the asymptotic equation $\gamma_k = \mathcal{O}(R^{-k})$ for $R > 1$. A rootfinding algorithm based on Chebyshev polynomials is used. We use slightly more complicated coefficients, $\mathbf{c}^T = [2, 2R^{-1}, 2R^{-2} + \frac{1}{2}R^{-4}, 2R^{-3}]$ and $\gamma_4 = -R^4$, so that $\mathbf{B}_4$ takes the simple form in (4.1). We have included an extra nonzero element in the southwest term to illustrate the essential contribution of the upper Hessenberg structure of $\mathbf{B}_n$ to the success of the matrix balancing algorithm. We approximately balance this matrix using $\mathbf{\Sigma}_4 = \mathrm{diag}(R^3, R^2, R^1, 1)$,

$$\mathbf{\Sigma}_4^{-1}\begin{bmatrix} 0 & 1/2 & 0 & R^4 \\ 1 & 0 & 1/2 & R^3 \\ 0 & 1/2 & 0 & R^2 \\ S & 0 & 0 & R \end{bmatrix}, \quad \mathbf{\Sigma}_4 = \begin{bmatrix} 0 & \frac{1}{2R} & 0 & R \\ R & 0 & \frac{1}{2R} & R \\ 0 & R/2 & & R \\ SR^3 & 0 & 0 & R \end{bmatrix}. \tag{4.1}$$

Note that because $\mathbf{B}_4$ is upper Hessenberg, $S = 0$, so that balancing reduces $\mathbf{B}_4$ in norm from $\mathcal{O}(R^4)$ to $\mathcal{O}(R)$. In this example, as R increases, the diagonal matrix determined by the balancing algorithm converges to $\mathbf{\Sigma}_4$. For polynomials of degree n, the norm of $\mathbf{B}_n$ is proportional to R^k for k possibly as large as n.

Different normalizations of the orthogonal polynomials correspond to the different diagonal similarity transformations applied to $\mathbf{B}_n$. The product of (2.3) and $\mathbf{\Sigma}_n = \mathrm{diag}(\sigma_0, \ldots, \sigma_{n-1})$ has the form

$$x \;\; \mathbf{f}_n^T(x)\mathbf{\Sigma}_n = \mathbf{f}_n^T(x)\mathbf{\Sigma}_n \;\; \mathbf{\Sigma}_n^{-1}\mathbf{B}_n\mathbf{\Sigma}_n + \frac{p(x)}{\gamma_n} h_{n,n-1}\sigma_{n-1}\mathbf{e}_{n-1}^T. \tag{4.2}$$

In this sense, the balancing algorithm determines a suitable normalization of the orthogonal polynomials (cf. Definition 2.1). Bear in mind that $e_j^T\mathbf{B}_n e_{n-1}$ is proportional to γ_j/γ_n.

The next theorem will show that asymptotically the right eigenvectors all are graded in exactly the same way, decreasing from term to term by a ratio of approximately $1/R$. For such $\mathbf{B}_n$ for an optimal $\mathbf{\Sigma}_n$, which approximately equalizes the row norms of $\mathbf{\Sigma}_n^{-1}\mathbf{W}_n$, σ_i/σ_i is asymptotically R. In general, it is not necessarily the case that the diagonal $\mathbf{\Sigma}_n$ that approximately minimizes a norm of $\mathbf{\Sigma}_n^{-1}\mathbf{B}_n\mathbf{\Sigma}_n$ is nearly optimal for the eigenvalue problem. For transcendental equation solving, asymptotically the polynomial coefficients also decrease by a factor of $1/R$ from coefficient to coefficient. Equation (4.1) illustrates how in this case the balancing algorithm determines a nearly optimal scaling for eigenvalue problems.

By Theorem 4.1, $(\gamma_n)_{n\geq 0}$ decays exponentially. Not surprisingly, in practice the diagonal elements of $\mathbf{\Sigma}_n$ exhibits similar exponential decay. The resulting $\mathbf{\Sigma}_n$ has an alarmingly large condition number. Next Theorem 4.2 will show that the rows of $(\mathbf{e}_j^T\mathbf{W}_n\mathbf{e}_k)_{0\leq j<n}$ decay at the same exponential rate as $(\gamma_n)_{n\geq 0}$. The transformation

from $\mathbf{W}_n$ to $\mathbf{\Sigma}_n^{-1}\mathbf{W}_n$ reduces the variation in the norms of the rows of $\mathbf{W}_n$ and improves the condition number of the eigenvalue problem.

THEOREM 4.2. *Suppose that $\psi(x)$ is analytic in an ellipse with foci at ± 1 and that ξ is a simple root of $\psi(x)$ within the ellipse. Suppose in addition that for each partial sum of the Jacobi series expansion of $\psi(x)$, $p_n(x)$, all of the roots of $p_n(x)$ are within the ellipse. Choose a root ξ_n of $p_n(x)$ nearest to ξ. The generalized companion matrix corresponding to $p_n(x)$, $\mathbf{B}_n$, has an eigenvector $\mathbf{w}_n$ such that $\mathbf{B}_n\mathbf{w}_n = \mathbf{w}_n\xi_n$. Then $\mathbf{e}_j^T\mathbf{w}_n = \mathcal{O}(R^{-n})$.*

Proof The maximal ellipses for $l(t) = (\psi(t) - \psi(\xi))/(t - \xi)$ and $\psi(t)$ coincide. For $l(t) = \sum_{n\geq 0}\phi_n(t)\mu_n$, by Theorem 4.1, there holds $R = \liminf|\mu_n|^{-1/n}$. The partial sums are $l_n(x) = \sum_{0\leq j\leq n}\phi_j(x)\mu_j$. By careful accounting, one may show that for each fixed $\epsilon > 0$, and for $|x - \xi| > \epsilon$, there holds $l_n(x) - l(x) = \mathcal{O}(R^{-n})$. Furthermore, a similar argument shows that $l_n(\xi_n) - l(\xi_n) = \mathcal{O}(R^{-n})$, from which $|||l_n(x) - l(x)||| = \mathcal{O}(R^{-n})$ follows. By theorem

$$\begin{aligned}|\mathbf{e}_j^T\mathbf{w}_n| &= |\langle\phi_j, l_n\rangle| = |\langle\phi_j, l_n - l\rangle + \langle\phi_j, l\rangle| \\ &= |\langle\phi_j, l_n - l\rangle + \mu_j| \leq |||l_n - l||| + |\mu_j| = \mathcal{O}(R^{-n}). \qquad \square\end{aligned}$$

Note that if the Jacobi series converges superexponentially, or even if R is very large, our justification of the balancing algorithm breaks down. We performed many numerical experiments, in floating point arithmetic with machine precision 2^{-54}, attempting to cause the balancing algorithm to fail. The expansion coefficients in the Jacobi series of a transcendental function coefficients converge to zero. We assume that each expansion coefficient, γ_m, with the maximal absolute value, $\sup_k|\gamma_k| = |\gamma_m|$, arises for $m \ll n$. For an entire function, $\lim_{n\to+\infty}\gamma_n/\gamma_{n+1} = +\infty$. The values of $\{\gamma_n\}_{n\geq 0}$ computed in finite precision arithmetic do not share this asymptotic property. The absolute error in each nonzero γ_n is, very roughly, the product of the machine precision and $\sup_k|\gamma_k|$. In our numerical experiments, we never observed a huge value of γ_n/γ_{n+1} for the nonzero approximate values of $\{\gamma_n\}_{n\geq 0}$. In other words, although matrix balancing has always worked for us, one must check that balancing determines a B_n not much larger in norm than H_n.

4.2. An algorithm for transcendental equations. An algorithm is implemented as a MATLAB script that approximates the roots in an interval of a transcendental equation. Modified companion matrices are used to find the roots in or very near $[-1, 1]$. The case of a polynomial expressed as a sum of Chebyshev polynomials is considered. The algorithm to approximate a function by a polynomial is reviewed briefly. Many subtle numerical analysis issues are discussed that are crucial for readers who actually want to solve a transcendental equation, such as rules for when to discard some of the eigenvalues. Readers more interested in concrete information on how to find the roots of a given polynomial expressed as a sum of Chebyshev coefficients are directed to the paragraph directly following the algorithm.

A collocation method is used to determine the expansion coefficients with respect to the Chebyshev polynomials of a polynomial approximation of a scalar function $\psi(x)$ whose domain includes $[-1, 1]$ (see Appendix A in [6]). For completeness, we briefly review the popular method here. The Chebyshev–Gauss–Lobatto (CGL) points, $\cos(k\pi/n)_{k=0}^n$, are unisolvent. A unique nth degree polynomial interpolates $\psi(x)$ at the CGL points. The column vector of expansion coefficients $[\gamma_0, \ldots, \gamma_n]^T$ is the product of discrete Chebyshev transformation matrix, Π_n, and the column vector, $[\psi(\cos(0\pi/n)), \ldots, \psi(\cos(n\pi/n))]^T$, where $\Pi_n = [\cos(ij\pi/n)2/(q_iq_jn)]_{0\leq i,j\leq n}$, and $q_0 = q_n = 2$ and $q_i = 1$ otherwise. Other issues including spectral convergence, the adaptive selection of n, and the subdivision of the interval are discussed elsewhere [6].

Techniques for discarding some of the eigenvalues are discussed. There are two reasons to discard certain computed eigenvalues. First, equations with n_p roots in or near $[-1, 1]$ may be approximated by a polynomial of higher degree $n > n_p$. In finite precision arithmetic, the $n - n_p$ additional eigenvalues do not necessarily solve the polynomial equation. We would like to be able to reliably determine n_p. Second, in many applications the cost of evaluating the function is significant.

An important application of Chebyshev polynomials is solving transcendental equations in a way that minimizes the number of function evaluations [4]. For polynomials of high degree, some definitions of nearness to $[-1, 1]$ will classify a large percentage of the eigenvalues as potential polynomial roots, significantly increasing the number of function evaluations needed for equation residuals. The spurious eigenvalues are in a region in the complex plane in which Chebyshev polynomials of a given degree are wildly unstable. We select only eigenvalues within a domain of interest; here we discard eigenvalues outside of $(-2, 2) \times (-.2, .2)$. On the other hand, discarding roots may also be discarding part of the answer. Real roots may be approximated by complex eigenvalues near to $[-1, 1]$. For example, if the complex QR algorithm is applied to the real matrix $\mathbf{B}_n$ (for robustness), the set of computed roots is not closed under conjugation. The problem is addressed by using a partial condition number of the eigenvalues. The condition number of an eigenvalue, ξ, is the product of two terms, $\|\mathbf{f}_n(\xi)\|_2$ (defined in (1.3)) and a term that involves the Lagrange interpolation polynomial whose support contains the eigenvalue. Our solution is to add a test that discards ξ such that $\|\mathbf{f}_n(\xi)\|_2$ is enormous (compare the definition of `cond_max`). Chebyshev polynomials are wildly unstable in such regions in the complex plane. At multiple roots near $[-1, 1]$, the norm of the vector values of the orthogonal polynomial evaluated at the roots is of order one and is not discarded.

The parameters are chosen here to avoid large numbers of spurious roots. No attempt is made to find all of the roots of the polynomial.

```
n              = 2^4;                                     % polynomial degree
[CGLpoints, ChebTransMat] = setupChebyshev(n);
FunctValues    = problemRod(CGLpoints);                   % evaluate @ CGL pts
ExpansionCoeff = FunctValues * ChebTransMat;
if ExpansionCoeff(n+1) == 0,
   error(’leading expansion coefficient vanishes; try --n’);
end
ExpansionCoeff = ExpansionCoeff/(-2*ExpansionCoeff(n+1)); % normalize
H = diag(ones(n-1, 1)/2, 1) + diag(ones(n-1, 1)/2, -1); H(1, 2) = 1;
C = H; C(n, :) = C(n, :) + ExpansionCoeff(1:n);           % nonstandard
Eigenvalues    = eig( C );                                % ...companion matrix
Vandermonde    = evalCheb(n,Eigenvalues);                 % generalized...
Vcolsums       = sum( abs(Vandermonde’) );                % Vandermonde matrix
tube_index     = find((abs(imag(Eigenvalues))<.2) & ...
                     (abs(real(Eigenvalues))< 2));
Solutions      = Eigenvalues( tube_index );               % Spectrum in
Vcolsums       = Vcolsums( tube_index );                  % ...(-2,2)x(-.2,.2)
cond_max       = min( 2^(n/2), 10^6 );                    % Cluster threshold
condEigs_index = find ( Vcolsums < cond_max );            % Select
Solutions      = Solutions( condEigs_index );             % ...roots
```

Suppose we want to find the roots of the polynomial

$$p(x) = eT_0(x) + 2\pi T_1(x) + 2\gamma T_2(x) - 2T_3(x),$$

where $\gamma = 0.57721566\ldots$ is Euler's constant. The example corresponds to $n = 3$ at line 1. Lines 2 and 3 are replaced by ExpansionCoeff $= [e, 2\pi, 2\gamma, -2]$;. The roots are approximately 1.44, -1.02, and -0.13. The two roots outside of $[-1, 1]$ are due to their large condition numbers (estimated by Vcolsums).

In the algorithm, three user supplied external functions are called. The function setupChebyshev() determines the CGL points and the matrix that transforms function values to expansion coefficients.

```
function [cgl,CT] = setupChebyshev(n)
y = [0:n]*pi/n;
cgl = cos(y);
for i=0:n,
   CT(i+1,:) = cos(y*i);
end
pp = ones(n+1,1); pp(1) = 1/2; pp(n+1) = 1/2;
CT = diag(pp) * CT * diag(pp);
CT = CT * (2/n); %                     End of function setupChebyshev
```

The function problemRod evaluates user functions at specified points in the domain. Here a problem associated with the vibration of an elastic rod is solved.

```
function functValues = problemRod(cgl)
[one, ncol] = size(cgl);
n = ncol-1;
first = cgl*3 + ones(1,n+1)*(3+1);
functValues = cos(first*pi) - sech(first*pi);
%                                           End of function problemRod
```

The vector ExpansionCoeff is the vector of coefficients in the collocation approximation by Chebyshev polynomials of degree up to n. The function evalCheb evaluates the Chebyshev polynomials at a specified set of points.

```
function V = evalCheb(degree_max,z)
% Input: vector of points, z, and the polynomial degree, degree_max.
% Output: Vandermonde matrix, m by degree_max + 1, V(j+1,k)=T_j(z_k)
[m,one] = size(z);
if m*degree_max >= 0,
   V(:,1) = ones(m,1);
   if degree_max >= 1,
      V(:,2) = z;
      if degree_max >= 2,
          index = find( log(abs(z))  >= 100/degree_max ); %  avoid
          si = size(index,1);                              % overflow
          if si > 0
             z(index) = ones(si,1)*exp(100/degree_max);
          end
          for i=2:degree_max,
             V(:,i+1) = V(:,i).*(2*z)  - V(:,i-1);
          end
      end
   end
else
   V = [];
end %                                       End of function evalCheb
```

In practice, five figures are recommended for verification purposes. Graphs of the expansion coefficients of the function and its derivative versus the indices are useful for assessing convergence. The exponential decay of the expansion coefficients indicates the convergence of the series to the transcendental equation. A plot of the cubic spline interpolants at the CGL points to the function and its derivative over the interval $[-1, 1]$ is useful for detecting clusters of roots. Later the (real parts of the) roots and the (real parts of the) derivative values at the roots may be overlaid onto the respective graphs. In selecting the eigenvalues that approximate roots of the polynomial or transcendental equation, it is helpful to compare the distribution of the eigenvalues and the row sums of the generalized Vandermonde matrix. The roots in $[-1, 1]$ and the other roots appear as two sets that are easier to see than to quantify. Finally one must check the transcendental equation residuals at the selected eigenvalues.

5. Examples. We apply our rootfinding technique to Chebyshev expansions arising in the solution of some transcendental equations. Two representative applications of the algorithm are discussed, followed by two more challenging applications. For a transcendental equation $\psi(\xi) = 0$ with clustered or multiple roots, an added issue is that roots of $p_n(x)$ that are not near roots of $\psi(x)$ appear among the approximate solutions of $\psi(\xi) = 0$. Such problems demonstrate the importance of monitoring $\psi'(x)$. Examples 1 and 2 concern problems with well-separated simple roots. For polynomials with multiple roots, $p_n(x)$ may very accurately approximate $\psi(x)$ and still have spurious roots near roots of $\psi(x)$, as will be shown in Example 3. Last, numerical experiments on computing the eigenvalues of $\mathbf{B}_n$ are discussed.

The methodology of the experiments is as follows. The number of function evaluations is carefully minimized. The degree of the polynomial n is doubled. In fact n is a power of 2 in the examples, except Example 2. Although doubling the polynomial order with a spectral method is overkill, it is done here for two reasons. First, if the function is evaluated at n points, it is possible to reuse the previous $n/2$ function values [6], carefully minimizing the number of function evaluations. Second, we wish to illustrate the properties of the nonstandard companion matrices for all values of n, not just special values. We find that using excessively large values of n, up to 2^{10}, does not effect the accuracy of the approximate roots in $[-1, 1]$. For transcendental equations, we discuss at length the minimal values of n for which the series is (almost) converged.

Example 1 concerns the transverse vibrations (u) of a homogeneous rod of length π with both ends free (u'' and u''' vanish at endpoints). The first six flexible modes are found by solving the secular equation $\cos(\pi x) - \operatorname{sech}(\pi x)$ in the interval $1 \leq x \leq 7$ [8, p. 296]. For $n = 16$ or $n = 32$ all six roots are approximated to within 5 or 14 significant digits, respectively.

Example 2 reproduces results from [6]. The roots of Bessel's function of the first kind, $J_\nu(\xi) = 0$, are computed without doubling the polynomial degree. In the first numerical experiments $J_0(x) = 0$ is solved over three intervals, $[0, w]$, for $w = 20$, 60, and 180; $J_0(x)$ has 6, 19, and 57 roots in the respective intervals. The computed roots J_0 are compared to the roots computed by a stable algorithm. Here we can exactly reproduce the results of [6].

Example 3 originated in [20] and is the nonlinear eigenvalue problem for $T(\lambda) = \lambda^2 B^{(2)} + (e^\lambda - 1)B^{(1)} - B^{(0)}$. The transcendental equation here is $\det T(\lambda) = 0$. The determinant is evaluated by factoring $T(\lambda)$ (not by Cramer's rule). A large interval is chosen to test the numerical stability of the rootfinding algorithm.

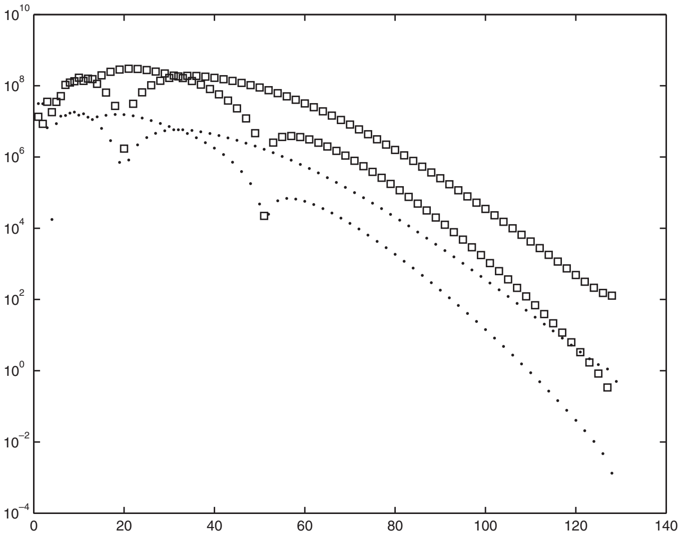


FIG. 5.1. *The function $\psi(x)$ of Example 3 is approximated by a Chebyshev series $p(x)$ of degree 128. The absolute values of the expansion coefficients of $p(x)$ (dots) and $p'(x)$ (squares) are plotted on a logarithmic scale.*

The problem is an example of an overdamped system. Each $B^{(i)}$ is symmetric positive definite. At $\lambda = -\infty$, $\lambda = 0$, and $\lambda = +\infty$ the normalized matrix $T(\lambda)/\|T(\lambda)\|$ is positive definite, negative definite, and positive definite. The matrices are 8 by 8 with $B^{(0)} = 100I_8$, and for $0 \le i, j < 8$,

$$B^{(1)}_{i,j} = (i+1)(j+1)(9 - \max(i+1, j+1)) \quad \text{and} \quad B^{(2)}_{i,j} = 8\delta_{i,j} - 1/(i+j+2).$$

Six roots are clustered near -3.7 with average absolute gap 0.1. As is carefully documented in [4], the the exponential growth of $\det T(\lambda)$ as $\lambda \to +\infty$ impedes resolution on certain intervals. The problem illustrates the rewards for choosing a suitable interval. On the interval $[-8, 8]$ in double precision arithmetic the roots of the polynomial approximation of $\det T(\lambda)$ poorly approximate the solutions, but in the interval $[-8, 4]$, the roots of the polynomial approximation converge rapidly to the solutions.

Another approach, pursued here, concerns the alternative scaled problem $\psi(\lambda) = \det(T(\Lambda)/\sigma(\lambda))$ for

$$\sigma(\lambda) = (\det B^{(0)})^{1/8} + (\det B^{(1)})^{1/8}(e^{\lambda} - 1) + (\det B^{(2)})^{1/8}\lambda^2.$$

The interval $[-10, 10]$ containing all 16 of the roots is used. We will discuss in detail the results obtained using a Chebyshev series expansions of degree 128 (see Figure 5.1). Of the 128 eigenvalues, 108 are discarded and 20 are potential solutions (see Figure 5.2).

Inspection of the graphs of the $\psi(x)$ and $\psi'(x)$, shown in Figures 5.3 and 5.4, respectively, is helpful. In a large neighborhood of the root cluster around -3.7, there holds $|\psi| < 10^{-9}$ and $|\psi'| < 10^{-5}$. A degree 128 polynomial is insufficient to resolve each root in the cluster. On the other hand, with a degree 256 polynomial, 240 eigenvalues are discarded. The remaining 16 eigenvalues approximate the 16 roots, each with residual norms below 10^{-13}.

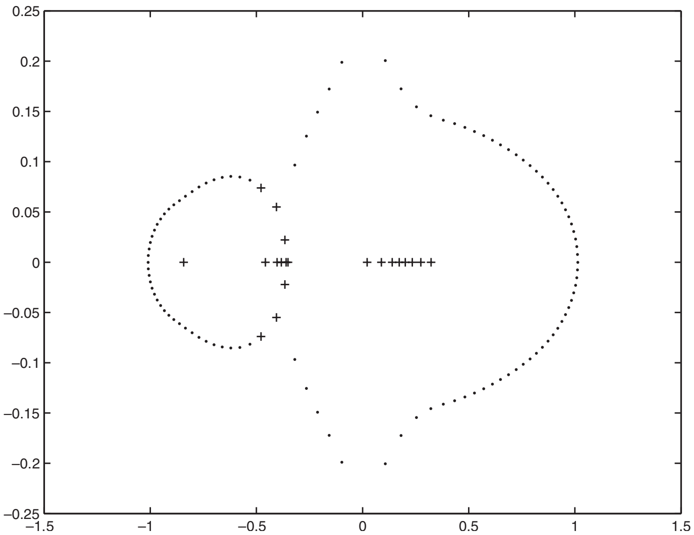


FIG. 5.2. *In Example* 3 *using a Chebyshev series of degree* 128 *results in a companion matrix* $\mathbf{B}_{128}$. *The figure displays the eigenvalues of* $\mathbf{B}_{128}$ *in the complex plane. A "$\cdot$" indicates each of the* 108 *discarded eigenvalues, and a "+" indicates each of the* 20 *potential roots.*

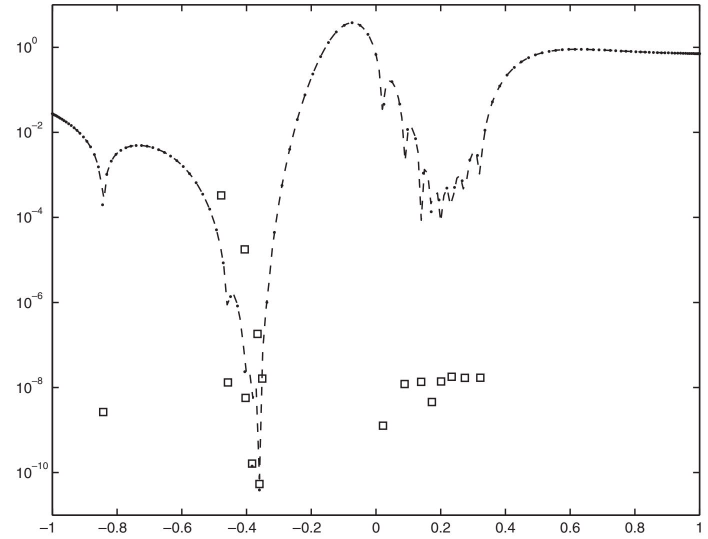


FIG. 5.3. *The value of* $|p(x)|$ *for* $-1 \le x \le 1$ *is shown on a logarithmic scale for the degree* 128 *Chebyshev series approximation of the function* $\psi(x)$ *of Example* 3. *The function values at CGL points (dots), a spline interpolant to the CGL points (dashed line), and the residuals at the* 20 *potential roots (squares) are each presented. A complex root* ξ *is displayed at* $x = \Re(\xi)$. *Because of this discrepancy, such function values appear above the spline interpolant in* $[-1, 1]$.

Before closing, we make an observation about accelerating the convergence of the QR algorithm for computing the eigenvalues of $\mathbf{B}_n$ arising from the solution of transcendental equations. Loosely speaking, a matrix is graded (by diagonal) if the norms of the diagonals increase geometrically. Note that $\mathbf{B}_n$ is graded by diagonal. We have

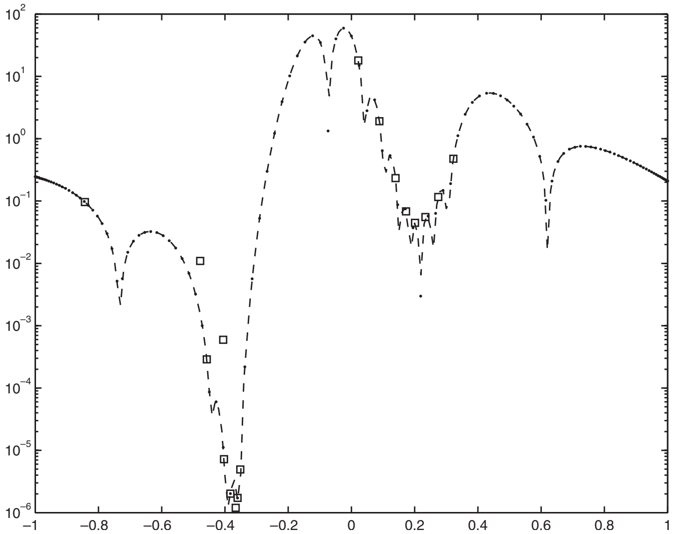


FIG. 5.4. *Example* 3*, the absolute value the derivative* $|p'(x)|$ *for* $-1 \le x \le 1$ *is shown on a logarithmic scale for the degree* 128 *Chebyshev series approximation of the function* $\psi(x)$ *of Example* 3*. The function values at CGL points (dots), a spline interpolant to the CGL points (dashed line), and the derivative values at the* 20 *potential roots (squares) are all shown. Complex roots are displayed at the real part of the root, and because of this discrepancy, such function values differ from the spline interpolant.*

observed that the computed Schur form of $\mathbf{B}_n$ is also graded. Careful examination of the QR iterates (say with zero shifts) from $\mathbf{B}_n$ (without balancing) indicates that along with the accuracy of the computed eigenvalues, the graded structure is also lost. Better results are obtained using the matrix $\mathbf{M}_n\mathbf{B}_n^T\mathbf{M}_n^T$ determined using the antidiagonal matrix $\mathbf{M}_n = [\mu_{i,j}]_{0\le i,j<n}$ with $\mu_{i,j} = \delta_{i,n-i-1}$. The matrix $\mathbf{M}_n\mathbf{B}_n^T\mathbf{M}_n^T$ is similar to $\mathbf{B}_n$ and inherits its unreduced upper Hessenberg and graded structure. We observe that the QR iteration applied to $\mathbf{M}_n\mathbf{B}_n^T\mathbf{M}_n^T$ (without balancing) preserves the graded structure of $\mathbf{B}_n$ and converges in many fewer iterations.

6. Conclusion.

We have shown how to find the roots of a degree n polynomial $p(x)$ expressed in terms of orthogonal polynomials. In particular, we have shown that these roots are the eigenvalues of a nonstandard companion matrix $\mathbf{B}_n$. This companion matrix gets infinitely large as the highest order coefficient γ_n in our orthogonal expansion goes to zero. However, we have analyzed the numerical stability of this algorithm for Jacobi polynomials and found that it has good numerical stability properties as long as we are interested only in roots in the interval $[-1, 1]$. This makes the algorithm particularly suited for finding the roots of transcendental equations.

We have presented an algorithm for finding the roots of a scalar transcendental equation by expressing it in terms of orthogonal polynomials, and using the companion matrix $\mathbf{B}_n$. We have given several numerical examples that illustrate the stability of this algorithm. For a more detailed summary, see section 1.1.

Acknowledgments. It is a pleasure to acknowledge the assistance of the referee Prof. L.N. Trefethen and the anonymous referee in the preparation of the document. We wish to thank Prof. Holly Dison for assistance in performing the numerical experiments and Prof. Ming Gu, who we heard had used the Chebyshev companion matrix, though he has not published anything on any nonstandard companion matrix.

REFERENCES

[1] G.S. Ammar, D. Calvetti, W.B. Gragg, and L. Reichel, *Polynomial zerofinders based on Szegö polynomials*, J. Comput. Appl. Math., 127 (2001), pp. 1–16.

[2] Z. Battles and L. N. Trefethen, *An extension of MATLAB to continuous functions and operators*, SIAM J. Sci. Comput., 25 (2004), pp. 1743–1770.

[3] C. Bernardi and Y. Maday, *Collection Mathématiques et Applications, Vol.* 10: *Approximations Spectrals de Problèms aux Limites Elliptiques*, Springer, Paris, 1992.

[4] J.P. Boyd, *A Chebyshev polynomial interval-searching method ("Lanczos economization") for solving a nonlinear equation with application to the nonlinear eigenvalue problem*, J. Comput. Phys., 118 (1995), pp. 1–8.

[5] J.P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, New York, 2001.

[6] J.P. Boyd, *Computing zeros on a real interval through Chebyshev expansions and polynomial root finding*, SIAM J. Numer. Anal., 40 (2002), pp. 1666–1682.

[7] J.P. Boyd, *Computing real roots of a polynomial in Chebyshev series form through subdivision*, Appl. Numer. Math., submitted.

[8] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, 1st English ed., Vol. 1, Interscience, New York, 1937.

[9] S. Fortune, *An iterated eigenvalue algorithm for approximating roots of univariate polynomials*, J. Symbolic Comput., 33 (2002), pp. 627–646.

[10] W. Gautschi, *The condition of orthogonal polynomials*, Math. Comp., 26 (1972), pp. 923–924.

[11] W. Gautschi, *The condition of algebraic equations*, Numer. Math., 21 (1973), pp. 405–424.

[12] W. Gautschi, *The condition of polynomials in power form*, Math. Comp., 33 (1979), pp. 343–352.

[13] W. Gautschi, *Orthogonal polynomials: Applications and computation*, Acta Numer., 5 (1996), pp. 45–120.

[14] G.H. Golub and C. Van Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[15] D. Gottlieb and S.A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, 2nd ed., SIAM, Philadelphia, 1989.

[16] N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

[17] E. Hille, *Analytic Function Theory*, Vol. II, Ginn and Company, Boston, 1962.

[18] H. Hochstadt, *The Functions of Mathematical Physics*, Dover, New York, 1986.

[19] G.F. Jónsson and S. Vavasis, *Solving polynomials with small leading coefficients*, SIAM J. Matrix Anal. Appl., 26 (2004), pp. 400–414.

[20] K.P. Hadeler, *Mehrparametrige und nichtlineare eigenwertaufgaben*, Arch. Ration. Mech. Anal., 27 (1967), pp. 306–328.

[21] H.J. Stetter, *Numerical Polynomial Algebra*, SIAM, Philadelphia, 2004.

[22] G. Szegö, *Orthogonal Polynomials*, 4th ed., AMS, Providence, RI, 1975.

[23] K.-C. Toh and L.N. Trefethen, *Pseudozeros of polynomials and pseudospectra of companion matrices*, Numer. Math., 68 (1994), pp. 403–425.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 1988–2011

A FINITE ELEMENT, MULTIRESOLUTION VISCOSITY METHOD FOR HYPERBOLIC CONSERVATION LAWS[*]

MARCUS CALHOUN-LOPEZ[†] AND MAX D. GUNZBURGER[‡]

Abstract. It is well known that the classic Galerkin finite element method is unstable when applied to hyperbolic conservation laws such as the Euler equations for compressible flow. It is also well known that naively adding artificial diffusion to the equations stabilizes the method but sacrifices too much accuracy to be of any practical use. An elegant approach, referred to as spectral viscosity methods, has been developed for spectral methods in which one adds diffusion only to the high-frequency modes of the solution, the result being that stabilization is effected without sacrificing accuracy. We extend this idea into the finite element framework by using hierarchical finite element functions as a multifrequency basis. The result is a new finite element method for solving hyperbolic conservation laws in which artificial diffusion can be applied selectively only to the high-frequency modes of the approximation. As for spectral viscosity methods, this results in stability without compromising accuracy. In the context of a one-dimensional scalar hyperbolic conservation law, we prove the convergence of approximate solutions, obtained using the new method, to the entropy solution of the conservation law. To illustrate the method, the results of a computational experiment for a one-dimensional hyperbolic conservation law are provided.

Key words. hyperbolic conservation laws, finite element methods, multiresolution viscosity, hierarchical basis functions

AMS subject classifications. 65N60, 35L65

DOI. 10.1137/S0036142904439380

1. Introduction.

We consider a new finite element method, based on hierarchical basis functions and a scale-dependent artificial viscosity, for hyperbolic conservation laws. With respect to a given triangulation of a domain, standard nodal basis functions are all of the same scale; i.e., their support is roughly equal. In contrast, hierarchical basis functions can be clustered into groups such that basis functions within a particular group are of a different scale from those of the other groups. The multiscale nature of the hierarchical basis functions allows for the selective addition of viscosity only at the smallest scales, very much in the spirit of spectral viscosity methods. It is hoped that such flexibility will be sufficient for stabilizing discrete Galerkin finite element approximations while, at the same time, results in more accurate approximations with respect to both convergence rates in regions where the solution is smooth and the sharpness of the resolution of discontinuities in the solution.

This paper is a first step at verifying that finite element methods based on hierarchical basis functions and a scale-dependent artificial viscosity do indeed fulfill the promise mentioned in the previous paragraph. We provide some background

[*] Received by the editors January 5, 2004; accepted for publication (in revised form) October 18, 2004; published electronically December 15, 2005.
http://www.siam.org/journals/sinum/43-5/43938.html

[†] Mathematics Department, University of Maryland, College Park, MD 20742-4015 (mcalhoun@math.umd.edu). The research of this author was supported in part by National Science Foundation grant DMS-0240049 and by the U.S. Department of Energy, Office of Energy Research and performed at Sandia National Laboratories, a multiprogram laboratory operated by Sandia Corporation, a Lockheed-Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC-94AL85000.

[‡] School of Computational Science, Florida State University, Tallahassee, FL 32306-4120 (gunzburg@csit.fsu.edu). The research of this author was supported in part by National Science Foundation grant DMS-0308845 and by CSRI, Sandia National Laboratories, contract 18407.

material and then describe the new method. We then analyze it for the case of a one-dimensional, periodic, scalar hyperbolic conservation law, showing that, under appropriate hypotheses, the approximate solution converges to the entropy solution of the conservation law. We then provide a simple example of the use of the method. Several issues concerning the efficient implementation of the new method as well as the results of more extensive computational testing are provided in [3,4].

1.1. Hyperbolic conservation laws. Let $\Omega \subseteq \mathbb{R}^d$ be a bounded domain. A general system of conservation laws has the form

$$\frac{\partial \mathbf{q}}{\partial t} + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \mathbf{f}_j (\mathbf{q}) = \mathbf{0} \quad \text{in } \Omega \times (0,\infty) \qquad \text{and} \qquad \mathbf{q}(\cdot, 0) = \mathbf{g} \quad \text{in } \Omega \tag{1.1}$$

along with appropriate boundary conditions. Here, $\mathbf{q} : \Omega \times [0,\infty) \to \mathbb{R}^p$ denotes the vector-valued conserved variable, $\mathbf{f}_j : \mathbb{R}^p \to \mathbb{R}^p$, $j = 1, \ldots, d$, denote the d flux functions, and $\mathbf{g} : \Omega \to \mathbb{R}^p$ denotes the given initial data. For $\widetilde{\mathbf{q}} \in \mathbb{R}^p$, let $\mathbf{A}_j(\widetilde{\mathbf{q}}) : \mathbb{R}^p \to \mathbb{R}^{p\times p}$ denote the $p \times p$ Jacobian matrix of $\mathbf{f}_j$, i.e., $\mathbf{A}_j(\widetilde{\mathbf{q}}) = \frac{\partial \mathbf{f}_j}{\partial \mathbf{q}}(\widetilde{\mathbf{q}})$. The system (1.1) is hyperbolic if, for all solutions $\mathbf{q}$, any linear combination of $\{\mathbf{A}_j(\mathbf{q})\}_{j=1}^{d}$ has real eigenvalues with eigenvectors that span $\mathbb{R}^p$. The system (1.1) is strictly hyperbolic if the eigenvalues are distinct. See, e.g., [8,10] for details.

The system (1.1) does not, in general, have a classical solution because of the spontaneous formation of discontinuities. Instead, one must look for a solution $\mathbf{q} \in L^\infty(\Omega \times (0,\infty); \mathbb{R}^p)$ which satisfies (1.1) in the distributional sense:

$$\int_0^\infty \int_\Omega \left[\mathbf{q} \cdot \frac{\partial \boldsymbol{\phi}}{\partial t} + \sum_{j=1}^{d} \mathbf{f}_j(\mathbf{q}) \cdot \frac{\partial \boldsymbol{\phi}}{\partial x_j} \right] d\Omega\, dt + \int_\Omega \mathbf{g} \cdot \boldsymbol{\phi}(\cdot, 0)\, d\Omega = 0 \tag{1.2}$$

for all test functions $\boldsymbol{\phi} \in C_0^\infty(\Omega \times [0,\infty); \mathbb{R}^p)$. It is clear that for a smooth enough solution, (1.1) and (1.2) are equivalent.

In the presence of discontinuities, solutions of the system (1.1) or of the weak formulation (1.2) are not uniquely determined. Additional conditions must be imposed to determine the unique, physically relevant solution. The second law of thermodynamics tells us that the entropy of the system should not decrease; satisfying this requirement suffices to allow one to obtain the unique, physically relevant entropy solution.

Let $\Phi, \{\Psi_j\}_{j=1}^{d} : \mathbb{R}^p \to \mathbb{R}$ be smooth functions; for (1.1), Φ is an entropy with entropy fluxes $\{\Psi_j\}_{j=1}^{d}$ if Φ is convex and $\boldsymbol{\nabla}_{\mathbf{q}} \Phi^T \frac{\partial \mathbf{f}_j}{\partial \mathbf{q}} = \boldsymbol{\nabla}_{\mathbf{q}} \Psi_j$ in $\mathbb{R}^p$ for $1 \leq j \leq d$. A simple calculation states that if $\mathbf{q}$ is a smooth solution to (1.1), then $\Phi(\mathbf{q})$ satisfies a scalar conservation law with flux $\Psi(\mathbf{q})$:

$$\frac{\partial}{\partial t} \Phi(\mathbf{q}) + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \Psi_j(\mathbf{q}) = 0 \text{ in } \mathbb{R}^d \times (0,\infty). \tag{1.3}$$

In some instances, Φ can be interpreted as the negative of the physical entropy, so (1.3) says that if $\mathbf{u}$ is a smooth solution, then $\Phi \circ \mathbf{q}$ satisfies a conservation law with flux functions $\{\Psi_j \circ \mathbf{q}\}_{j=1}^{d}$.

For solutions with discontinuities, we impose the *entropy condition* on $\mathbf{q}$ that requires the physical entropy to be nondecreasing:

$$\frac{\partial}{\partial t} \Phi(\mathbf{q}) + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \Psi_j(\mathbf{q}) \leq 0 \tag{1.4}$$

for every entropy function Φ with entropy fluxes $\{\Psi\}_{j=1}^{d}$; (1.4) is an inequality in the distributional sense:

$$\int_0^\infty \int_\Omega \left[\Phi\left(\mathbf{q}\right)\frac{\partial\phi}{\partial t}+\sum_{j=1}^{d}\Psi_j\left(\mathbf{q}\right)\frac{\partial\phi}{\partial x_j}\right]d\Omega\,dt \geq 0\ \forall\,\phi\in C_0^\infty\left(\Omega\times(0,\infty)\right),\ \phi\geq 0. \tag{1.5}$$

In (1.1), viscous effects are ignored. For the class of phenomena that are modeled with hyperbolic conservation laws, viscous effects are generally small, but they are present and play a role when sharp gradients (such as shocks) of the solution are present. An alternate and equivalent means of characterizing the unique, physically relevant solution of (1.1) is to let $\mathbf{q}=\lim_{\varepsilon\to 0}\mathbf{q}^\varepsilon$ a.e., where $\mathbf{q}^\varepsilon:\Omega\times[0,\infty)\to\mathbb{R}^p$ is the solution of the perturbed equation

$$\frac{\partial\mathbf{q}^\varepsilon}{\partial t}+\sum_{j=1}^{d}\frac{\partial}{\partial x_j}\mathbf{f}_j\left(\mathbf{q}^\varepsilon\right)-\varepsilon\,\Delta\mathbf{q}^\varepsilon=\mathbf{0}\ \text{ in } \Omega\times(0,\infty) \quad \text{and} \quad \mathbf{q}^\varepsilon\left(\cdot,0\right)=\mathbf{g}\ \text{ in } \Omega \tag{1.6}$$

along with boundary conditions. In other words, the entropy solution is the limit of the *viscous solution* as the viscosity goes to zero.

1.2. Numerical methods for hyperbolic conservation laws. Direct discretizations of (1.1) lead to unstable approximations. The most obvious stabilization approach is to instead discretize the perturbed system (1.6) but, as is well known, this leads to severe smearing of discontinuities and to low accuracy even in regions in which the solution is smooth. Of course, there have been many methods proposed for determining improved stabilized approximation solutions of hyperbolic conservation laws; see, e.g., [6,10,13,17,18,24,27].

Finite difference (FD) methods are the oldest of the numerical methods, so many variations have been developed. Many successful strategies for solving hyperbolic conservation laws were originally developed in the FD framework then adapted to other methods. However, FD schemes tend to have difficulties with complex geometries, satisfying prescribed boundary conditions, and rigorous analyses. In fluid dynamics, complex geometries are common, and, as shown in [6], poor approximation of boundary conditions can severely affect a numerical method.

Finite volume (FV) methods inherently capture many of the important aspects of conservation laws; FV methods are locally conservative. Information is propagated along the characteristic curves, at least approximately. FV methods use unstructured grids, so they can handle complex geometries. High-order schemes, however, are difficult to attain.

Finite element (FE) methods are well suited to handle complex geometries and prescribed boundary conditions. Formally high-order schemes can be defined by simply increasing the degree of the approximating polynomials used. The price paid is a large increase in the number of unknowns. In discontinuous Galerkin (DG) methods (see, e.g., [6]), no continuity restrictions are placed on the approximating solution, which results in several advantages. DG methods are easy to parallelize; adaptive strategies are relatively easy to implement; and the mass matrix is block diagonal, so explicit time-stepping schemes are possible. The lack of continuity of solutions, however, is also the cause of the biggest drawback of DG methods: The number of unknowns is drastically increased compared to, say, nodal finite element methods and other types of methods. The shock capturing streamline diffusion method adds a diffusion term to the conservation law, but unlike (1.6), diffusion is added in different amounts in the direction of the characteristic curves (streamline diffusion) and its normal direction (crosswind diffusion). Streamline diffusion is added everywhere, while crosswind diffusion is added only near discontinuities. To determine characteristic curves, space-time elements must be used, which increases the number of unknowns and results in an implicit time scheme. See [13, 14, 23].

Spectral methods, including the spectral viscosity (SV) method, provide another class of methods. Since incorporating the ideas from the SV method into the FE framework is the subject of this paper, we discuss SV methods in a little more detail in section 1.2.1.

The distinctions between the various methods are not always sharp. Some FV and FE schemes can be written into an equivalent FD formulation. As noted in [6], some FV methods can be considered to be special types of DG methods. Furthermore, there are other methods, e.g., kinetic methods [22], for hyperbolic conservation laws that do not fall within the classes just mentioned.

1.2.1. Spectral viscosity methods.

In [24], SV methods were introduced as a scheme to obtain approximate solutions of the periodic Burgers equation using Fourier spectral basis functions. The theory was further refined and extended in a series of papers [5, 9, 12, 19, 20, 25, 26]. Of particular importance to us are [12, 19], in which Legendre polynomials are used. The variational formulation of the Legendre SV method is closest to our FE formulation.

We present the most basic SV method, which uses Fourier spectral basis functions. Using standard notation from Fourier spectral methods, we define

$$u_N = P_N u\left(x,t\right), \quad P_N u = \sum_{|k|\le N} \widehat{u}_k\left(t\right) e^{ik\pi x}, \quad \widehat{u}_k\left(t\right) = \frac{1}{2}\int_{-1}^{+1} u\left(x,t\right) e^{-ik\pi x}\, d\Omega.$$

We seek u_N such that

$$\frac{\partial u_N}{\partial t} + \frac{\partial}{\partial x}\left(P_N \frac{{u_N}^2}{2}\right) = \varepsilon \frac{\partial}{\partial x}\left(Q_N \frac{\partial u_N}{\partial x}\right).$$

Q_N denotes the spectral viscosity operator defined as a convolution with the viscosity kernel, $Q_N\left(x\right)$, so that

$$Q_N \frac{\partial u_N}{\partial x} = Q_N\left(x\right) * \frac{\partial u_N\left(x,t\right)}{\partial x} \qquad \text{and} \qquad Q_N\left(x\right) = \sum_{|k|\le N} \widehat{Q}_k e^{ik\pi x}.$$

We choose $0 \le \widehat{Q}_k \le 1$ and $\widehat{Q}_k = 0$ for small $|k|$. It is easy to see the effect of Q_N if we write the diffusion term in Fourier space:

$$\varepsilon \frac{\partial}{\partial x}\left(Q_N \frac{\partial u_N}{\partial x}\right) = -\varepsilon \sum_{|k|\le N}\left(k^2\,\pi^2\,\widehat{Q}_k\,\widehat{u}_k\,e^{ik\pi x}\right).$$

Since $\widehat{Q}_k = 0$ for all but large $|k|$, Q_N dampens or eliminates the low frequency modes of u_N in the diffusion term. So, we see that the SV diffusion term is a compromise between not adding diffusion, which leads to instability, and adding full diffusion, which limits the convergence rate and smears out discontinuities in the solution. Ideally, one would like to add diffusion only in the vicinity of a discontinuity. However, the global nature of the basis functions does not allow for an adaptive viscosity kernel.

The SV solution u_N does not converge to the exact solution u at the optimal rate because of the poor convergence of $P_N u$. $P_N u$ is limited to first-order convergence in smooth regions and has $O(1)$ Gibbs oscillations near a discontinuity. Post-processing u_N recovers spectral convergence. The post-processing scheme can be enhanced by knowing the locations of discontinuities, as in [9]. Because of the global nature of spectral basis functions, this edge detection task is not trivial.

1.3. Hierarchical finite element basis functions. The usual (nodal) basis functions used in FE methods all have the same frequency. In order to have multi-frequency basis functions, we use hierarchical basis functions. In the elliptic partial differential equation setting, an early analysis of hierarchical basis functions, especially in one dimension, is given in [31]. For two dimensions, see [29]. A good overview of hierarchical basis functions can be found in [30].

First consider a polygonal domain Ω. Let $\mathcal{T}_0$ be a coarse grid approximation of Ω. The nth-level triangulation $\mathcal{T}_n$ is obtained by subdividing the elements of $\mathcal{T}_{n-1}$. Let S^N be the space of continuous functions which are polynomials of degree p on the elements of $\mathcal{T}_N$. Let $\mathcal{N}_N \subseteq \overline{\Omega}$ be the nodes of the elements of $\mathcal{T}_N$. The nodal basis functions of S^N are defined by $\phi_i \in S^N$ such that $\phi_i(x_j) = \delta_{ij}$ for all $x_j \in \mathcal{N}_N$. It is well known that $S^N = \operatorname{span}\{\phi_i\}_i$. The use of nodal bases leads to many nice numerical properties, such as sparse matrices and the local assembly of matrices. However, we cannot use the nodal bases for our purposes because, as we noted earlier, the elements of $\{\phi_i\}_i$ all have the same frequency.

Let $\mathcal{N}_n$ denote the nodes of the nth-level triangulation, $\mathcal{T}_n$, $\mathcal{S}^n$ denote the corresponding finite element space, and B^n denote the nodal basis of $\mathcal{S}^n$. The hierarchical basis functions are defined by

$$\psi_{n,i} \in B^n \quad \text{such that} \quad \psi_{n,i}(x_j) = 0 \quad \forall\, x_j \in \mathcal{N}_{n-1}.$$

For $0 \le n \le N$, $\{\psi_{n,i}\}_{n,i} \subseteq S^N$ is a linearly independent set with the same dimension as S^N, so $S^N = \operatorname{span}\{\psi_{n,i}\}_{n,i}$. See Figure 1 for a comparison of the nodal and hierarchical bases for linear elements in one dimension. As can be seen from the figure, $\psi_{n,i}$ is a low frequency function for small n and a high frequency function for large n.

The strategy just outlined works for polynomials of any degree. For example, the first column of Figure 2 consists of quadratic hierarchical basis functions. An alternate strategy is to use linear hierarchical basis functions for $n < N$, as in the second column of Figure 2.

In order to determine $\mathcal{T}_{n+1}$ from $\mathcal{T}_n$, we must decide, for a given $T \in \mathcal{T}_n$, how many subelements to divide T into. For linear and quadratic rectangular-type elements in $\mathbb{R}^d$, the natural choice is 2^d subelements. For cubic elements, the natural choice is 3^d since the vertices of an element will then be a subset of the vertices of its parent. Here, we limit our attention to linear and quadratic basis functions.

For domains with curved boundaries, the situation is more complicated. Let Ω be our potentially complicated domain. One strategy is to use the hierarchical decomposition of some polygonal domain Ω' such that $\Omega \subseteq \Omega'$, as in [15]. Another strategy is to try and impose a hierarchical structure on an unstructured mesh, as in [1]. The hierarchical structure could also be imposed on the mapping of Ω to a polygonal domain.

Several important properties of hierarchical finite element basis functions and several issues that arise in efficient implementations of finite element methods based on these kinds of bases are discussed in [3, 4].

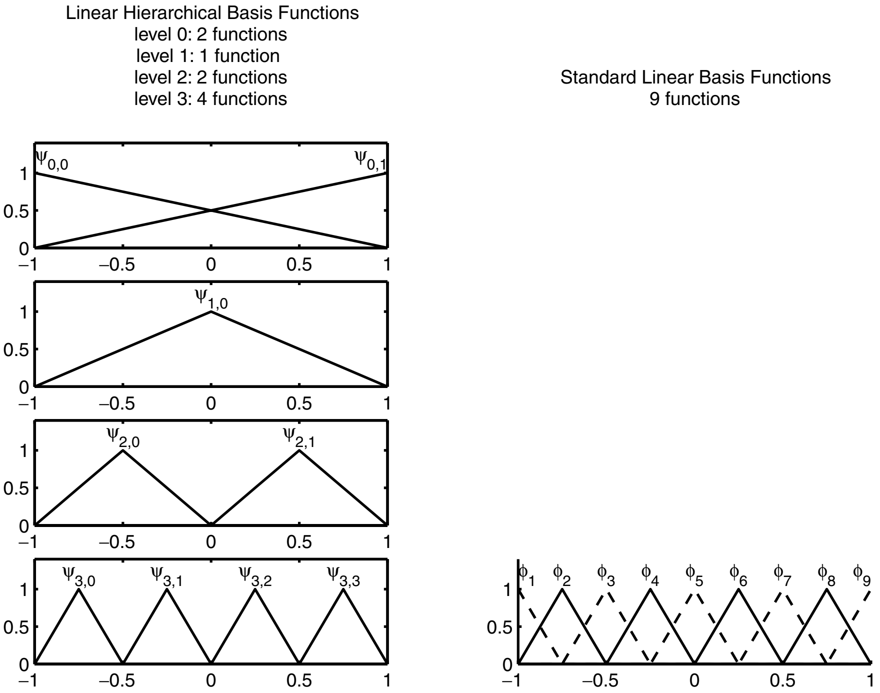


Fig. 1. *Hierarchical (left) and nodal (right) linear basis functions spanning the same nine-dimensional finite element space.*

2. Finite element multiresolution viscosity method. Assume we have a hierarchical sequence of partitions $\{\mathcal{T}_n\}_{n=0}^N$ of the open bounded set $\Omega \subseteq \mathbb{R}^d$. Let $\mathbf{S}^N$ be the space of continuous vector-valued functions whose components are in S^N. We seek an approximate solution to the hyperbolic conservation law (1.1). The finite element multiresolution viscosity method is defined as follows: seek $\mathbf{q}^N \in \mathbf{S}^N$ such that

$$\begin{aligned}(2.1)\qquad &\frac{d}{dt}\int_\Omega \mathbf{q}^N \cdot \mathbf{v}\, d\Omega + \sum_{j=1}^d \int_\Omega \frac{\partial}{\partial x_j}\mathbf{f}_j\left(\mathbf{q}^N\right)\cdot \mathbf{v}\, d\Omega \\ &\quad + \varepsilon_N \sum_{i=1}^p \sum_{j,k=1}^d \int_\Omega \frac{\partial}{\partial x_j}\left(Q_N^{j,k} q_i^N\right) \frac{\partial v_i}{\partial x_k}\, d\Omega \\ &\quad - \varepsilon_N \sum_{i=1}^p \sum_{j,k=1}^d \int_{\partial\Omega} \frac{\partial}{\partial x_j}\left(Q_N^{j,k} q_i^N\right) v_i\, \widehat{n}_k\, ds = 0 \qquad \forall\, \mathbf{v} \in \mathbf{S}^N,\end{aligned}$$

where $\widehat{\mathbf{n}}$ is the unit normal to the boundary $\partial\Omega$ of Ω. As in the SV method, $Q_N^{j,k}$ is chosen to dampen or eliminate the low frequency modes of a function:

$$(2.2)\qquad \begin{cases} Q_N^{j,k} : S^N \to S^N, \qquad \displaystyle\sum_{n=0}^N \sum_i \beta_{n,i}\, \psi_{n,i} \mapsto \sum_{n=0}^N \sum_i Q_{N;n,i}^{j,k}\, \beta_{n,i}\, \psi_{n,i}, \\ 0 \le Q_{N;n,i}^{j,k} \le 1, \qquad \text{and} \qquad Q_{N;n,i}^{j,k} = \begin{cases} 0 & \text{for small } n \ (n \text{ near } 0), \\ 1 & \text{for large } n \ (n \text{ near } N). \end{cases} \end{cases}$$

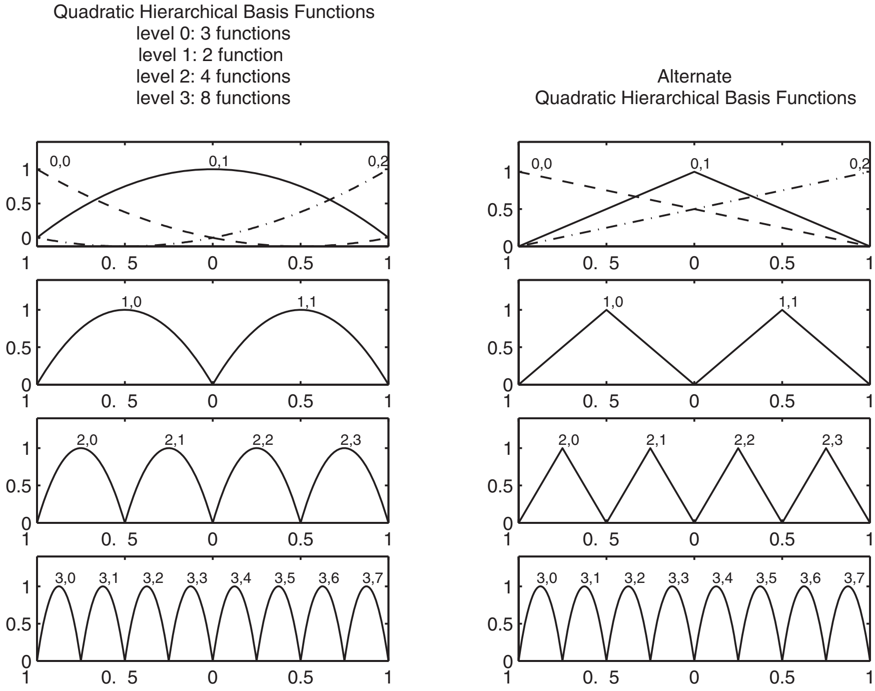


FIG. 2. *Two sets (left and right) of quadratic hierarchical basis functions spanning the same 17-dimensional, quadratic finite element space.*

To account for boundary conditions imposed along with (1.1), a subspace of S_N might need to be used (for essential boundary conditions) or the boundary integral in (2.1) might be reduced to one over part of $\partial\Omega$ (for natural boundary conditions).

Equation (2.1) is a weak formulation of the modified system

$$\frac{\partial \mathbf{q}}{\partial t} + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{q}) - \varepsilon_N \sum_{j,k=1}^{d} \frac{\partial^2}{\partial x_j \partial x_k} \left(Q_N^{j,k} \mathbf{q} \right) = \mathbf{0}, \tag{2.3}$$

where $[Q_N^{j,k}\mathbf{q}]_i = Q_N^{j,k} q_i$ for $1 \le i \le p$. The dependence of $Q_N^{j,k}$ on both j and k allows for the flexibility of introducing directional bias in the diffusion which can result in reduced crosswind diffusion. As in the streamline diffusion method, this would probably require the use of entropy variables. Here, we simplify our formulation by using an isotropic diffusion term, Q_N, such that

$$Q_{N;n,i}^{j,k} = Q_{N;n,i}\, \delta_{j,k}.$$

Then, (2.1) and (2.3), respectively, reduce to

$$\begin{aligned} &\frac{d}{dt} \int_\Omega \mathbf{q} \cdot \mathbf{v}\, d\Omega + \sum_{j=1}^{d} \int_\Omega \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{q}) \cdot \mathbf{v}\, d\Omega + \varepsilon_N \int_\Omega \boldsymbol{\nabla}(Q_N \mathbf{q}) : \boldsymbol{\nabla} \mathbf{v}\, d\Omega \\ &\qquad - \varepsilon_N \int_{\partial\Omega} \frac{\partial}{\partial n} (Q_N \mathbf{q}) \cdot \mathbf{v}\, ds = 0 \qquad \forall\, \mathbf{v} \in \mathbf{S}^N \end{aligned} \tag{2.4}$$

and

$$\frac{\partial \mathbf{q}}{\partial t} + \sum_{j=1}^{d} \frac{\partial}{\partial x_j} \mathbf{f}_j(\mathbf{q}) - \varepsilon_N \Delta (Q_N \mathbf{q}) = \mathbf{0}.$$

After choosing a time discretization technique, (2.4) is equivalent to a nonlinear system of equations that may be solved, e.g., by Newton's method. The relevant Jacobian matrix has the form $J^H = \widetilde{J}^H + K^H Q$, where $\widetilde{J}^H$ is the Jacobian of the time dependent and flux terms, K^H is the Laplacian stiffness matrix, and Q is a diagonal matrix whose nonzero elements are $\{Q_{N;n,i}\}$. For ease of presentation, we have ignored the boundary term. Note that for an explicit time integration method, the Jacobian matrix reduces to the mass matrix.

The solution of the discrete equations resulting from our hierarchical finite element discretization may be implemented using matrices arising from the corresponding nodal basis. See [3, 4] for details. Here, we merely observe that the Jacobian matrix J^H and residual vector $\vec{R}^H$ may be expressed, respectively, in terms of their nodal basis counterparts J^D and $\vec{R}^D$ through the relations $J^H = S^T(\widetilde{J}^D S + K^D S Q)$ and $\vec{R}^H = S^T \vec{R}^D$, where S is the change of basis matrix such that $\vec{X}^D = S\vec{X}^H$. The determination of $(J^H)^{-1}\vec{R}^H = (\widetilde{J}^D S + K^D S Q)^{-1}\vec{R}^D$ by an iterative solver then requires the calculation of the matrix-vector multiplication $(\widetilde{J}^D S + K^D S Q)\vec{X}$ and possibly $(\widetilde{J}^D S + K^D S Q)^T \vec{X}$ that only involve the nodal matrices $\widetilde{J}^D$ and K^D and the transfer matrix S. Compared to K^D and $\widetilde{J}^D$, S is not sparse, so one does not want to explicitly construct S. So, making the iterative linear solvers efficient requires being able to calculate $S\vec{X}$ and $S^T\vec{X}$ quickly. Algorithms for this purpose can be found in [3, 4].

2.1. Advantages of hierarchical bases. We use hierarchical bases (instead of nodal bases) because of their multifrequency property. Nodal bases, however, have important computational advantages such as producing matrices that are much more sparse and that can be locally assembled. However, as just discussed, one can retain most of the advantages of the nodal bases. One assembles and stores all of the system matrices as nodal basis matrices and uses the transfer matrix S in an iterative linear solver; S does not even have to be stored.

In the SV method, there is only one function, with global support, at a given frequency. In the hierarchical FE formulation, there are many functions, with local support, at a given frequency. Compared to SV methods, the hierarchical FE formulation offers two advantages: Diffusion can be added locally, and edge detection is trivial. For large values of n, the hierarchical basis function $\psi_{n,i}$ has local support, so $Q^{j,k}_{N;n,i}$ only has a local effect. One can therefore add more diffusion near a discontinuity and less or no diffusion in smooth regions. This should improve the accuracy of the method. As we are about to see, the size of $|\beta_{n,i}|$ (where $\beta_{n,i}$ is the coefficient of $\psi_{n,i}$ in the hierarchical basis expansion of a function; see (2.2)) can be used to determine if the support of $\psi_{n,i}$ resides in a smooth region or is near a discontinuity.

2.1.1. Edge detection. Using hierarchical bases, edge detection is a trivial task. Near a discontinuity, the high frequency hierarchical coefficients are of order one. In smooth regions, they shrink exponentially. Figure 3 illustrates a hierarchical decomposition of a piecewise smooth function containing a discontinuity. One can easily determine the location of discontinuities by looking at the magnitude of the coefficients of the high frequency basis functions.

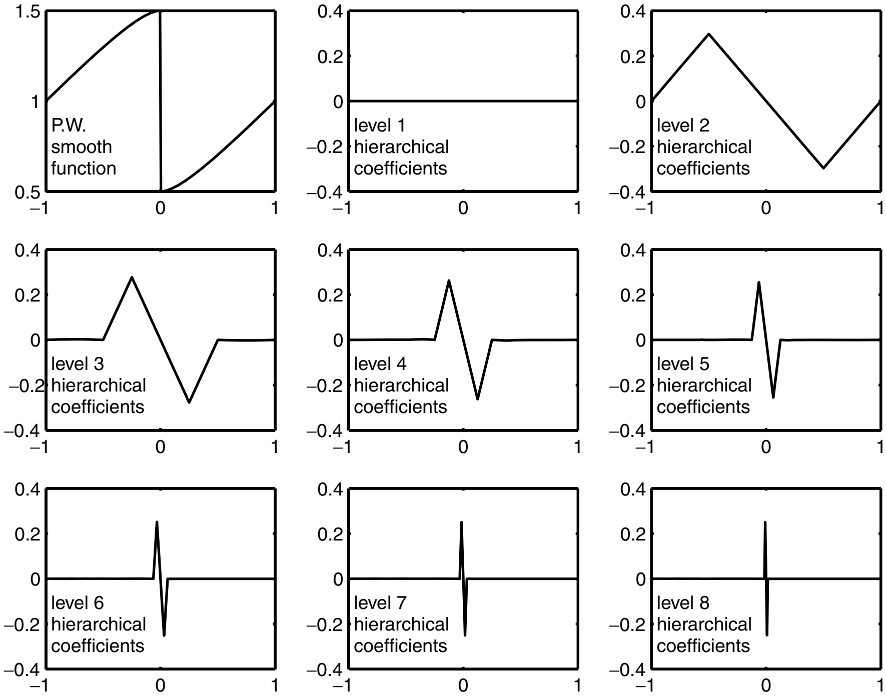


FIG. 3. *Hierarchical decomposition of a piecewise smooth function with a discontinuity.*

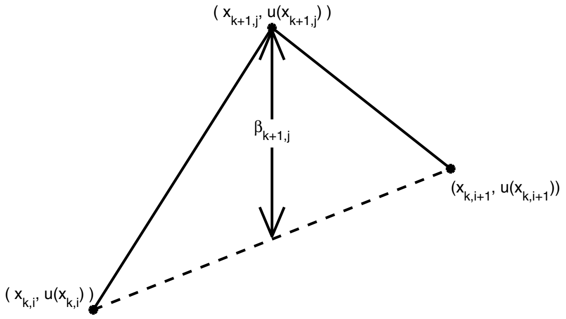


FIG. 4. *Determination of a linear hierarchical coefficient from function values.*

Edge detection for piecewise linear polynomials. Let us examine the behavior of the hierarchical coefficients for linear hierarchical basis functions. As one can see in Figure 4, $\beta_{k+1,j}$ can be calculated from the value of the function u at $x_{k+1,j}$ and at the node points of the parent cell. For a uniform partition, the cell size at level k is $\Delta x_k = |\Omega|\ 2^{-k}$. Thus, $x_{k+1,j} - x_{k,i} = x_{k,i+1} - x_{k+1,j} = \Delta x_{k+1}$ and

$$\beta_{k+1,j} = \frac{u\left(x_{k+1,j}\right) - u\left(x_{k,i}\right)}{2} - \frac{u\left(x_{k,i+1}\right) - u\left(x_{k+1,j}\right)}{2}. \tag{2.5}$$

Let $T_{k,i} = (x_{k,i}, x_{k,i+1})$.

Assume that u *is discontinuous* and has a discontinuity in $T_{k,i}$. Then, at least one of the two terms in (2.5) will have a relatively large value so that $|\beta_{k+1,j}|$ will be of the same order as the jump of u, independent of k.

Now, assume that u is *continuously differentiable*, i.e., $u \in C^1(\overline{T_{k,i}})$. Then, since (2.5) can be written as

$$\beta_{k+1,j} = \frac{|\Omega|}{4}\, 2^{-k} \left(\frac{u\,(x_{k+1,j}) - u\,(x_{k,i})}{x_{k+1,j} - x_{k,i}} - \frac{u\,(x_{k,i+1}) - u\,(x_{k+1,j})}{x_{k,i+1} - x_{k+1,j}} \right),$$

the mean value theorem yields that

$$\beta_{k+1,j} = \frac{|\Omega|}{4}\, 2^{-k}\, \left[u'\,(\widetilde{x}_1) - u'\,(\widetilde{x}_2)\right] \tag{2.6}$$

for some $\widetilde{x}_1, \widetilde{x}_2 \in T_{k,i}$. Therefore, $\beta_{k+1,j}$ is of order 2^{-k}, i.e.,

$$|\beta_{k+1,j}| \leq |\Omega|\ \|u'\|_{L^\infty(T_{k,i})}\ 2^{-k-1}$$

so that it decays exponentially with k.

If u is *twice continuously differentiable*, i.e., $u \in C^2(\overline{T_{k,i}})$, the mean value theorem applied to (2.6) yields that

$$\beta_{k+1,j} = -\frac{|\Omega|}{4}\, 2^{-k}\, (\widetilde{x}_2 - \widetilde{x}_1)\ u''\,(\widetilde{x})$$

for some $\widetilde{x} \in T_{k,i}$. Therefore, $\beta_{k+1,j}$ is of order 4^{-k}, i.e.,

$$|\beta_{k+1,j}| \leq \frac{|\Omega|}{4}\ |\widetilde{x}_2 - \widetilde{x}_1|\ |u''\,(\widetilde{x})|\ 2^{-k} \leq |\Omega|^2\ \|u''\|_{L^\infty(T_{k,i})}\ 4^{-k-1}$$

so that it again decays exponentially with k.

Similar results can be obtained for quadratic hierarchical basis function; see [3,4].

3. Convergence to entropy solutions for one-dimensional scalar conservation laws.

In this section, we prove that the solution of the hierarchical finite element discretization introduced in section 2 converges to the entropy solution of the one-dimensional, periodic Burgers equation. We will make use of the method of compensated compactness; in particular, we will use the div-curl lemma [7,21,28] and Murat's lemma [7,13,19,21,28]. The broad outlines of the proof follow that of [24].

3.1. The periodic, one-dimensional Burgers equation.

Let $\Omega = (a,b)$ be an open bounded interval and let $\Omega_T = (a,b) \times (0,T)$ for some finite time interval $(0,T)$. We seek a finite element (FE) approximation to $u\,(x,t)$, the entropy solution of the periodic hyperbolic conservation law:

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0 \quad \text{in } \Omega_T, \tag{3.1}$$

$$u\,(a,t) = u\,(b,t) \quad \text{in } (0,T), \qquad \text{and} \qquad u\,(x,0) = g \quad \text{in } (a,b), \tag{3.2}$$

along with the entropy condition

$$\frac{\partial}{\partial t}\left(\frac{u^2}{2}\right) + \frac{\partial}{\partial x}\left(\frac{u^3}{3}\right) \leq 0 \quad \text{in } \Omega_T, \tag{3.3}$$

where (3.1) and (3.3) hold in the distributional sense. We will assume the $g \in H^1\,(a,b)$ and that g is space-periodic. The entropy solution of Burgers' equation can be found using (3.3) instead of the more general entropy condition (1.4). This greatly simplifies our analysis since we now require an entropy-type inequality for one entropy/entropy flux pair, $(\frac{u^2}{2}, \frac{u^3}{3})$, instead of all of them. A weak form of the problem (3.1) is given by

$$\int_{\Omega_T} \left(\varphi \frac{\partial u}{\partial t} + \varphi \frac{\partial}{\partial x} \Big(\frac{u^2}{2} \Big) \right) dx\, dt = 0 \quad \forall \varphi \in C_0^\infty(\Omega_T). \tag{3.4}$$

Correspondingly, (3.3) can be expressed in the weak form

$$\int_{\Omega_T} \left(\varphi \frac{\partial}{\partial t} \Big(\frac{u^2}{2} \Big) + \varphi \frac{\partial}{\partial x} \Big(\frac{u^3}{3} \Big) \right) dx\, dt \le 0 \quad \forall \varphi \in C_0^\infty(\Omega_T),\ \varphi \ge 0. \tag{3.5}$$

3.1.1. The hierarchical finite element discretization. To formulate the FE approximation, we need some notation. Let $\mathcal{T}_0 = (a,b)$; $\mathcal{T}_N$ is obtained by subdividing the elements of $\mathcal{T}_{N-1}$ into M distinct elements. Let $|b-a|\, h_N$ be the maximal diameter of the elements of $\mathcal{T}_N$. Since we assume that the partition is quasi-uniform, there exists a positive constant ν such that $M^{-N} \le h_N \le \nu M^{-N}$ for all N.

Let $\{\psi_{k,i}\}_{k,i}$ be a hierarchical basis of S_p^N. Let us define $Q_N : S_p^N \to S_p^N$ as a damping operator $Q_N u_N = \sum_{k,i} (Q_{k,i}\, \beta_{k,i}\, \psi_{k,i})$ for $u = \sum_{k,i} (\beta_{k,i}\, \psi_{k,i}) \in S_p^N$, where $0 \le Q_{k,i} \le 1$ and $Q_{k,i} = 1$ for $k > m_H$. Thus, Q_N dampens (or eliminates) the low frequencies of a function while keeping the high frequencies above the level m_H. Occasionally, when the level of a basis function is unimportant, we will switch to the less cumbersome notation $\{\psi_i\}_i$ and $\{Q_i\}_i$ for the basis functions and damping coefficients, respectively.

We will also use the following convention: C will denote any positive constant which depends on known quantities and is independent of any indexing variables.

Let $g_N \in S_p^N$ be the interpolant of g. The hierarchical finite element approximation of (3.1)–(3.2) is given by the following: seek $u_N(x,t)$ with $u_N(\cdot,0) = g_N$ such that, for all $v \in S_p^N$,

$$\int_a^b \left[\frac{\partial u_N}{\partial t} + \frac{\partial}{\partial x} \left(\frac{{u_N}^2}{2} \right) \right] v\, dx + \varepsilon_N \int_a^b \left[\frac{\partial}{\partial x} (Q_N u_N) \frac{\partial v}{\partial x} \right] dx = 0. \tag{3.6}$$

3.2. Convergence theorem. We prove the following convergence results for hierarchical finite element approximations of the entropy solution of (3.1) and (3.2).

THEOREM 3.1. *Let $\{u_N\}_{N=0}^\infty$ denote a sequence of hierarchical finite element approximations determined by* (3.6). *Assume that $\|u_N\|_{L^\infty(\Omega_T)} \le C$ and assume that*

$$\varepsilon_N, h_N \to 0 \text{ as } N \to \infty, \tag{3.7}$$

$$\frac{\varepsilon_N}{h_N} \ge C, \tag{3.8}$$

$$\sqrt{\varepsilon_N} \left\| \frac{\partial}{\partial x} [(I - Q_N)\, v_N] \right\|_{L^2(a,b)} \le C\, \|v_N\|_{L^2(a,b)} \text{ for } v_N \in \left\{ u_N, \frac{\partial u_N}{\partial t} \right\}, \tag{3.9}$$

$$\left\| \frac{d}{dx} (Q_N g_N) \right\|_{L^2(U)} \le C \left\| \frac{d g_N}{dx} \right\|_{L^2(U)}. \tag{3.10}$$

Then, there exists a subsequence of $\{u_N\}_{N=0}^\infty$ that converges strongly in $L^2(\Omega_T)$ to a solution $u \in L^2(\Omega_T)$ of (3.1) *and* (3.2). *Further, assume that*

$$\frac{\varepsilon_N}{h_N} \to \infty \text{ as } N \to \infty \text{ and} \tag{3.11}$$

$$\sqrt{\varepsilon_N}\left\|\frac{\partial}{\partial x}\left[\left(I-Q_N\right)u_N\right]\right\|_{L^2(a,b)} \to 0 \ \textit{as } N\to\infty. \tag{3.12}$$

Then the subsequence of $\{u_N\}_N^\infty$ *converges strongly in* $L^2(\Omega_T)$ *to the entropy solution of* (3.1) *and* (3.2), *i.e., to the solution of* (3.1)–(3.3).

For the moment, we assume that (3.7)–(3.12) hold, and we prove, in sections 3.3 to 3.7, the theorem. Subsequently, in section 3.8, we will show that these conditions are satisfied in our context.

3.3. Existence of the finite element approximation.

The discrete FE equations (3.6) are equivalent to the following: seek $\vec{\alpha}:(0,T)\to\mathbb{R}^s$, where $s=\dim S_p^N$, such that

$$\dot{\vec{\alpha}}+M^{-1}\vec{F}\left(\vec{\alpha}\right)+M^{-1}\,K\,Q\,\vec{\alpha}=\vec{0}, \tag{3.13}$$

where M is the mass matrix, K is the stiffness matrix, Q is a diagonal matrix whose elements are $\{Q_i\}$, and $\vec{F}$ is the flux term: $[\vec{F}(\vec{\alpha})]_i=\int_a^b\psi_i\frac{1}{2}\frac{\partial}{\partial x}\left(\sum_j\alpha_j\,\psi_j\right)^2dx=\vec{\alpha}^TA_i\vec{\alpha}$, where A_i is the symmetric matrix $(A_i)_{j,k}=\frac{1}{2}\int_a^b\psi_i\frac{\partial}{\partial x}\left(\psi_j\,\psi_k\right)dx$. Evidently, the diffusion term is globally Lipschitz continuous. We now show that the flux term is locally Lipschitz continuous. For all $\vec{\alpha},\vec{\beta}\in\mathbb{R}^s$ and all i,

$$\begin{aligned}\left|\left[\vec{F}(\vec{\beta})-\vec{F}(\vec{\alpha})\right]_i\right| &= \left|\vec{\beta}^T A_i\,\vec{\beta}-\vec{\alpha}^T A_i\,\vec{\alpha}\right| = \left|(\vec{\beta}+\vec{\alpha})^T A_i\,(\vec{\beta}-\vec{\alpha})\right|\\ &\le \|\vec{\beta}+\vec{\alpha}\|_2\|A_i\|_2\|\vec{\beta}-\vec{\alpha}\|_2 \le \sqrt{s}\|\vec{\beta}+\vec{\alpha}\|_2\|A_i\|_2\|\vec{\beta}-\vec{\alpha}\|_\infty\end{aligned}$$

so that $\|\vec{F}(\vec{\beta})-\vec{F}(\vec{\alpha})\|_\infty\le\sqrt{s}\|\vec{\beta}+\vec{\alpha}\|_2\max_{1\le i\le s}\|A_i\|_2\|\vec{\beta}-\vec{\alpha}\|_\infty$. Since $\|A_i\|_2$ and s are independent of $\vec{\alpha}$ and $\vec{\beta}$, the flux term is locally Lipschitz continuous for any T.

Lipschitz continuity together with $|\vec{\alpha}\,(t)|<\infty$, by (3.15) and (3.17), yields that there exists a unique $\vec{C}^1\,[0,T]$ solution of (3.13) or, equivalently, of (3.6).

3.4. Estimates for u_N.

In (3.6), choose $v=u_N$; then

$$\int_a^b\left[\frac{\partial}{\partial t}\left(\frac{{u_N}^2}{2}\right)+\frac{\partial}{\partial x}\left(\frac{{u_N}^3}{3}\right)\right]dx+\varepsilon_N\int_a^b\frac{\partial}{\partial x}\left(Q_Nu_N\right)\frac{\partial u_N}{\partial x}\,dx=0.$$

Since u_N is periodic, $\int_a^b\frac{\partial}{\partial x}\left(\frac{{u_N}^3}{3}\right)dx=0$ so that

$$\frac{1}{2}\frac{d}{dt}\|u_N\|^2_{L^2(a,b)}+\varepsilon_N\int_a^b\frac{\partial}{\partial x}\left(Q_Nu_N\right)\frac{\partial u_N}{\partial x}\,dx=0. \tag{3.14}$$

3.4.1. $H^1\left(\Omega_T\right)$ estimates for u_N for linear polynomials.

The elements of the piecewise linear hierarchical basis are orthogonal with respect to the $H^1\left(a,b\right)$ seminorm. As a result,

$$\begin{aligned}\int_a^b\frac{\partial}{\partial x}\left(Q_Nu_N\right)\frac{\partial u_N}{\partial x}\,dx &= \sum_i\sum_j Q_i\beta_i\beta_j\int_a^b\psi_i'\psi_j'\,dx\\ &= \sum_i Q_i\beta_i^2\int_a^b\left(\psi_i'\right)^2\,dx \ge \sum_i Q_i^2\beta_i^2\int_a^b\left(\psi_i'\right)^2\,dx\\ &= \sum_i\sum_j Q_iQ_j\beta_i\beta_j\int_a^b\psi_i'\psi_j'\,dx = \left\|\frac{\partial}{\partial x}\left(Q_Nu_N\right)\right\|^2_{L^2(a,b)}.\end{aligned}$$

Integrating (3.14) over time, we obtain

$$\begin{aligned} C\left\|g\right\|^2_{L^2(a,b)} \geq \left\|g_N\right\|^2_{L^2(a,b)} &= \left\|u_N\left(\cdot,t\right)\right\|^2_{L^2(a,b)} + 2\varepsilon_N \int_0^t \int_a^b \frac{\partial}{\partial x}\left(Q_N u_N\right) \frac{\partial u_N}{\partial x}\, dx\, dt \\ &\geq \left\|u_N\left(\cdot,t\right)\right\|^2_{L^2(a,b)} + 2\varepsilon_N \int_0^t \left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|^2_{L^2(a,b)} dt \end{aligned}$$

so that

$$\left\|u_N\right\|_{L^2(\Omega_T)} \leq C\sqrt{T}\, \left\|g\right\|_{L^2(a,b)}, \quad \sqrt{\varepsilon_N}\left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|_{L^2(\Omega_T)} \leq C\left\|g\right\|_{L^2(a,b)}. \tag{3.15}$$

3.4.2. $H^1\left(\Omega_T\right)$ estimates for u_N for quadratic polynomials. The quadratic hierarchical basis functions are not orthogonal, but we can still obtain an estimate similar to (3.15). We now have that

$$\begin{aligned} \int_a^b \frac{\partial}{\partial x}\left(Q_N u_N\right) \frac{\partial u_N}{\partial x}\, dx &= \int_a^b \frac{\partial}{\partial x}\left(Q_N u_N\right) \frac{\partial}{\partial x}\left(Q_N u_N\right)\, dx \\ &\quad + \int_a^b \frac{\partial}{\partial x}\left(Q_N u_N\right) \frac{\partial}{\partial x}\left[\left(I-Q_N\right)u_N\right]\, dx \\ &\geq \int_a^b \left|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right|^2 dx \\ &\quad -\frac{1}{2}\int_a^b \left|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right|^2 dx - \frac{1}{2}\int_a^b \left|\frac{\partial}{\partial x}\left[\left(I-Q_N\right)u_N\right]\right|^2 dx \\ &= \frac{1}{2}\left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|^2_{L^2(a,b)} - \frac{1}{2}\left\|\frac{\partial}{\partial x}\left[\left(I-Q_N\right)u_N\right]\right\|^2_{L^2(a,b)} \\ &\geq \frac{1}{2}\left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|^2_{L^2(a,b)} - \frac{C}{2\,\varepsilon_N}\left\|u_N\right\|^2_{L^2(a,b)}. \end{aligned}$$

Substituting this result into (3.14), we obtain

$$\frac{d}{dt}\left\|u_N\right\|^2_{L^2(a,b)} \leq C\left\|u_N\right\|^2_{L^2(a,b)} - \varepsilon_N\left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|^2_{L^2(a,b)}. \tag{3.16}$$

We require a nonstandard formulation of the differential form of Gronwall's inequality. A proof is given in [8].

LEMMA 3.2. *Let $\eta\left(t\right)$ be an absolutely continuous function on $[0,T]$ such that for a.e. $t \in [0,T]$, $\eta'\left(t\right) \leq \phi\left(t\right)\eta\left(t\right) + \psi\left(t\right)$, where $\phi\left(t\right)$ and $\psi\left(t\right)$ are summable functions on $[0,T]$. Then, $\eta\left(t\right) \leq e^{\int_0^t \phi(r)\, dr}[\eta\left(0\right) + \int_0^t e^{-\int_0^s \phi(r)\, dr}\psi\left(s\right)\, ds]\ \forall\ t \in [0,T]$.*

Let us now assume that $\phi = C$ is a positive constant, and $\psi \leq 0$ is never positive. We then have

$$\begin{aligned} \eta\left(t\right) &\leq e^{Ct}\left[\eta\left(0\right) + \int_0^t e^{-Cs}\psi\left(s\right)\, ds\right] \\ &\leq e^{Ct}\left[\eta\left(0\right) + \int_0^t e^{-Ct}\psi\left(s\right)\, ds\right] = e^{Ct}\,\eta\left(0\right) + \int_0^t \psi\left(s\right)\, ds. \end{aligned}$$

Using this result with (3.16), we obtain

$$\left\|u_N\right\|^2_{L^2(a,b)} \leq e^{Ct}\left\|g_N\right\|^2_{L^2(a,b)} - \varepsilon_N \int_0^t \left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|^2_{L^2(a,b)} ds.$$

Since $\|g_N\|_{L^2(a,b)} \leq C\,\|g\|_{L^2(a,b)}$,

$$\|u_N\|^2_{L^2(a,b)} + \varepsilon_N \int_0^t \left\| \frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|^2_{L^2(a,b)} ds \leq Ce^{Ct}\,\|g\|^2_{L^2(a,b)} .$$

Therefore, we have that

$$\text{(3.17)} \qquad \left\{ \begin{array}{l} \|u_N\|_{L^2(\Omega_T)} \leq C\sqrt{e^{CT}-1}\ \|g\|_{L^2(a,b)}, \\ \sqrt{\varepsilon_N}\Big\| \dfrac{\partial}{\partial x}\left(Q_N u_N\right)\Big\|_{L^2(\Omega_T)} \leq C\sqrt{e^{CT}}\ \|g\|_{L^2(a,b)} . \end{array} \right.$$

3.5. Strong convergence of $\{u_N\}$. Let $\mathbf{v}_N = (\frac{{u_N}^2}{2}, u_N)$ and $\mathbf{w}_N = (\frac{{u_N}^2}{2}, -\frac{{u_N}^3}{3})$ so that

$$\operatorname{div}\mathbf{v}_N = \frac{\partial u_N}{\partial t} + \frac{\partial}{\partial x}\left(\frac{{u_N}^2}{2}\right) \qquad \text{and} \qquad \operatorname{curl}\mathbf{w}_N = \frac{\partial}{\partial t}\left(\frac{{u_N}^2}{2}\right) + \frac{\partial}{\partial x}\left(\frac{{u_N}^3}{3}\right).$$

3.5.1. $L^2(\Omega_T)$ bound on $\{\operatorname{div}\mathbf{v}_N\}$. In (3.6), choose $v = \frac{\partial u_N}{\partial t}$; then

$$\begin{aligned}
&\int_a^b \left|\frac{\partial u_N}{\partial t}\right|^2 dx + \int_a^b \frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\frac{\partial u_N}{\partial t}\,dx = -\,\varepsilon_N \int_a^b \frac{\partial}{\partial x}\left(Q_N u_N\right)\frac{\partial^2 u_N}{\partial x \partial t}\,dx \\
&=-\,\varepsilon_N \int_a^b \frac{\partial}{\partial x}\left(Q_N u_N\right)\frac{\partial^2}{\partial x\partial t}\left(Q_N u_N\right)dx \\
&\quad -\varepsilon_N \int_a^b \frac{\partial}{\partial x}\left(Q_N u_N\right)\frac{\partial^2}{\partial x\partial t}\left[\left(I-Q_N\right)u_N\right]dx \\
&=-\,\frac{\varepsilon_N}{2}\int_a^b \frac{\partial}{\partial t}\Big\{\Big[\frac{\partial}{\partial x}\left(Q_N u_N\right)\Big]^2\Big\}dx - \varepsilon_N \int_a^b \frac{\partial}{\partial x}\left(Q_N u_N\right)\frac{\partial^2}{\partial x\partial t}\left[\left(I-Q_N\right)u_N\right]dx \\
&\leq-\,\frac{\varepsilon_N}{2}\int_a^b \frac{\partial}{\partial t}\Big\{\Big[\frac{\partial}{\partial x}\left(Q_N u_N\right)\Big]^2\Big\}dx \\
&\quad +\varepsilon_N\Big\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\Big\|_{L^2(a,b)}\Big\|\frac{\partial^2}{\partial x\partial t}\left[\left(I-Q_N\right)u_N\right]\Big\|_{L^2(a,b)} \\
&\leq-\,\frac{\varepsilon_N}{2}\int_a^b \frac{\partial}{\partial t}\Big\{\Big[\frac{\partial}{\partial x}\left(Q_N u_N\right)\Big]^2\Big\}dx + C\sqrt{\varepsilon_N}\left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|_{L^2(a,b)}\left\|\frac{\partial u_N}{\partial t}\right\|_{L^2(a,b)} \\
&\leq-\,\frac{\varepsilon_N}{2}\int_a^b \frac{\partial}{\partial t}\left\{\left[\frac{\partial}{\partial x}\left(Q_N u_N\right)\right]^2\right\}dx + C\Big\|\frac{\partial u_N}{\partial t}\Big\|_{L^2(a,b)} \\
&\leq-\,\frac{\varepsilon_N}{2}\int_a^b \frac{\partial}{\partial t}\Big\{\Big[\frac{\partial}{\partial x}\left(Q_N u_N\right)\Big]^2\Big\}dx + C^2 + \frac{1}{4}\Big\|\frac{\partial u_N}{\partial t}\Big\|^2_{L^2(a,b)}.
\end{aligned}$$

Rearranging terms in the last expression, we obtain

$$\begin{aligned}
\frac{3}{4}\left\|\frac{\partial u_N}{\partial t}\right\|^2_{L^2(a,b)} + \frac{\varepsilon_N}{2}\frac{d}{dt}\left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|^2_{L^2(a,b)} - C &\leq -\int_a^b \frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\frac{\partial u_N}{\partial t}\,dx \\
&\leq \Big\|\frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\Big\|_{L^2(a,b)}\Big\|\frac{\partial u_N}{\partial t}\Big\|_{L^2(a,b)} \leq \Big\|\frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\Big\|^2_{L^2(a,b)} + \frac{1}{4}\Big\|\frac{\partial u_N}{\partial t}\Big\|^2_{L^2(a,b)}.
\end{aligned}$$

Rearranging terms again, we obtain

$$\left\|\frac{\partial u_N}{\partial t}\right\|^2_{L^2(a,b)} \leq C - \varepsilon_N\frac{d}{dt}\left\|\frac{\partial}{\partial x}\left(Q_N u_N\right)\right\|^2_{L^2(a,b)} + 2\left\|\frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\right\|^2_{L^2(a,b)} .$$

Integrating over time, we obtain

$$\begin{aligned}\left\|\frac{\partial u_N}{\partial t}\right\|^2_{L^2(\Omega_T)} &\le C + 2\int_{\Omega_T}\left|\frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\right|^2 dx\,dt \\ &\quad - \varepsilon_N\left\|\frac{\partial}{\partial x}\left(Q_N u_N(\cdot,T)\right)\right\|^2_{L^2(a,b)} + \varepsilon_N\left\|\frac{d}{dx}\left(Q_N g_N\right)\right\|^2_{L^2(a,b)} \qquad (3.18)\\ &\le C + 2\int_{U_T}\left|\frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\right|^2 dx\,dt + \varepsilon_N\left\|\frac{d}{dx}\left(Q_N g_N\right)\right\|^2_{L^2(a,b)} \qquad (3.19)\\ &\le C + 2\int_{U_T}\left|\frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\right|^2 dx\,dt + C\,\varepsilon_N\left\|\frac{dg_N}{dx}\right\|^2_{L^2(a,b)} \qquad (3.20)\\ &\le C + 2\int_{U_T}\left|\frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\right|^2 dx\,dt + C\,\varepsilon_N\left\|\frac{dg}{dx}\right\|^2_{L^2(a,b)}. \qquad (3.21)\end{aligned}$$

Now,

$$\begin{aligned}\int_{\Omega_T}\left|\frac{\partial}{\partial x}\left(\frac{u_N^2}{2}\right)\right|^2 dx\,dt &= \int_{\Omega_T}|u_N|^2\left|\frac{\partial u_N}{\partial x}\right|^2 dx\,dt \\ &\le \|u_N\|^2_{L^\infty(\Omega_T)}\left\|\frac{\partial u_N}{\partial x}\right\|^2_{L^2(\Omega_T)} \le \frac{C}{\varepsilon_N}.\end{aligned} \tag{3.22}$$

Combining the last two results, we obtain $\varepsilon_N\|\frac{\partial u_N}{\partial t}\|^2_{L^2(a,b)} \le C\left(1+\varepsilon_N+\varepsilon_N{}^2\right)$ so that

$$\sqrt{\varepsilon_N}\left\|\frac{\partial u_N}{\partial t}\right\|_{L^2(\Omega_T)} \le C. \tag{3.23}$$

Combining (3.22) and (3.23), we obtain

$$\begin{aligned}\sqrt{\varepsilon_N}\left\|\text{div }\mathbf{v}_N\right\|_{L^2(\Omega_T)} &= \sqrt{\varepsilon_N}\left\|\frac{\partial u_N}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u_N{}^2}{2}\right)\right\|_{L^2(\Omega_T)} \\ &\le \sqrt{\varepsilon_N}\left\|\frac{\partial u_N}{\partial t}\right\|_{L^2(\Omega_T)} + \sqrt{\varepsilon_N}\left\|\frac{\partial}{\partial x}\left(\frac{u_N{}^2}{2}\right)\right\|_{L^2(\Omega_T)} \le C.\end{aligned}$$

3.5.2. $\{\mathbf{div\,v}_N\}$ lies in a compact subset of $H^{-1}(\Omega_T)$. Let $\widetilde{\varphi} \in H_0^1(\Omega_T)$. For all $t \in (0,T)$, let $\widetilde{\varphi}_N(\cdot,t) \in S_p^N \cap H_0^1(a,b)$ be the $H^1(a,b)$ projection of $\widetilde{\varphi}$ so that

$$\int_a^b \frac{\partial\widetilde{\varphi}_N(\cdot,t)}{\partial x}\frac{\partial v}{\partial x}\,dx = \int_a^b \frac{\partial\widetilde{\varphi}(\cdot,t)}{\partial x}\frac{\partial v}{\partial x}\,dx \qquad \forall\, v \in S_p^N \cap H_0^1(a,b).$$

We need the $H^1(a,b)$ projection into S_p^N of an arbitrary $\widetilde{\varphi} \in H_0^1(\Omega_T)$ in order to use our FE formulation:

$$\begin{aligned}&\int_{\Omega_T}(\text{div }\mathbf{v}_N)\,\widetilde{\varphi}\,dx\,dt = \int_{\Omega_T}(\text{div }\mathbf{v}_N)\,\widetilde{\varphi}_N\,dx\,dt + \int_{\Omega_T}(\text{div }\mathbf{v}_N)\left(\widetilde{\varphi}-\widetilde{\varphi}_N\right)dx\,dt \\ &= -\varepsilon_N\int_{\Omega_T}\frac{\partial}{\partial x}\left(Q_N u_N\right)\frac{\partial\widetilde{\varphi}_N}{\partial x}\,dx\,dt + \int_{\Omega_T}(\text{div }\mathbf{v}_N)\left(\widetilde{\varphi}-\widetilde{\varphi}_N\right)dx\,dt \\ &= -\varepsilon_N\int_{\Omega_T}\frac{\partial}{\partial x}\left(Q_N u_N\right)\frac{\partial\widetilde{\varphi}}{\partial x}\,dx\,dt + \int_{\Omega_T}(\text{div }\mathbf{v}_N)\left(\widetilde{\varphi}-\widetilde{\varphi}_N\right)dx\,dt\end{aligned}$$

$$\begin{aligned}
&\le \varepsilon_N \Big\| \frac{\partial}{\partial x}\left(Q_N u_N\right)\Big\|_{L^2(\Omega_T)} \Big\|\frac{\partial \widetilde{\varphi}}{\partial x}\Big\|_{L^2(\Omega_T)} + \|\text{div } v_N\|_{L^2(\Omega_T)} \|\widetilde{\varphi} - \widetilde{\varphi}_N\|_{L^2(\Omega_T)} \\
&\le \varepsilon_N \Big\| \frac{\partial}{\partial x}\left(Q_N u_N\right)\Big\|_{L^2(\Omega_T)} \Big\|\frac{\partial \widetilde{\varphi}}{\partial x}\Big\|_{L^2(\Omega_T)} + C\, h_N \,\|\text{div } v_N\|_{L^2(\Omega_T)} \Big\|\frac{\partial \widetilde{\varphi}}{\partial x}\Big\|_{L^2(\Omega_T)} \\
&\le C \left(\sqrt{\varepsilon_N} + \frac{h_N}{\sqrt{\varepsilon_N}}\right) \Big\|\frac{\partial \widetilde{\varphi}}{\partial x}\Big\|_{L^2(\Omega_T)} = C\sqrt{\varepsilon_N}\left(1 + \frac{h_N}{\varepsilon_N}\right)\Big\|\frac{\partial \widetilde{\varphi}}{\partial x}\Big\|_{L^2(\Omega_T)}.
\end{aligned}$$

From (3.8), we have that $\frac{h_N}{\varepsilon_N} \le C$ so that

$$\int_{\Omega_T} (\text{div}\, \mathbf{v}_N)\, \widetilde{\varphi}\, dx\, dt \le C\sqrt{\varepsilon_N}\,\Big\|\frac{\partial \widetilde{\varphi}}{\partial x}\Big\|_{L^2(\Omega_T)}. \tag{3.24}$$

Let $\widetilde{\varphi} \in H_0^1(\Omega_T)$ with $\|\widetilde{\varphi}\|_{H^1(\Omega_T)} \le 1$. Then, from (3.24), we have that

$$\|\text{div } \mathbf{v}_N\|_{H^{-1}(\Omega_T)} \le C\sqrt{\varepsilon_N} \to 0 \text{ as } N \to \infty$$

so that $\{\text{div}\, \mathbf{v}_N\}$ lies in a compact subset of $H^{-1}(\Omega_T)$.

3.5.3. $\{\text{curl}\, \mathbf{w}_N\}$ lies in a compact subset of $H^{-1}(\Omega_T)$. Let $\varphi \in C_0^\infty(\Omega_T)$ be a test function. Since $u_N\,\varphi \in H_0^1(\Omega_T)$, we can choose $\widetilde{\varphi} = u_N\varphi$ in (3.24). Then,

$$\begin{aligned}
&\int_{\Omega_T} (\text{curl } \mathbf{w}_N)\, \varphi\, dx\, dt = \int_{\Omega_T} (\text{div } \mathbf{v}_N)\, u_N\, \varphi\, dx\, dt \\
&\le C\sqrt{\varepsilon_N}\,\Big\|\frac{\partial}{\partial x}(u_N\varphi)\Big\|_{L^2(\Omega_T)} = C\sqrt{\varepsilon_N}\Big\|u_N\frac{\partial \varphi}{\partial x} + \varphi\frac{\partial u_N}{\partial x}\Big\|_{L^2(\Omega_T)} \\
&\le C\sqrt{\varepsilon_N}\Big(\Big\|u_N\frac{\partial \varphi}{\partial x}\Big\|_{L^2(\Omega_T)} + \Big\|\varphi\frac{\partial u_N}{\partial x}\Big\|_{L^2(\Omega_T)}\Big) \\
&\le C\sqrt{\varepsilon_N}\Big(\|u_N\|_{L^\infty(\Omega_T)}\,\Big\|\frac{\partial \varphi}{\partial x}\Big\|_{L^2(\Omega_T)} + \|\varphi\|_{L^\infty(\Omega_T)}\,\Big\|\frac{\partial u_N}{\partial x}\Big\|_{L^2(\Omega_T)}\Big) \\
&\le C\Big(\sqrt{\varepsilon_N}\Big\|\frac{\partial \varphi}{\partial x}\Big\|_{L^2(\Omega_T)} + \|\varphi\|_{L^\infty(\Omega_T)}\Big).
\end{aligned}$$

Using a variational form of Murat's lemma [13, 19] gives, from the last result, that $\{\text{curl}\, \mathbf{w}_N\}$ lies in a compact subset of $H^{-1}(\Omega_T)$.

3.5.4. Strong convergence in $L^2(\Omega_T)$ of a subsequence of $\{u_N\}$. Since $\|u_N\|_{L^\infty(\Omega_T)} \le C$, there exists a subsequence $\{u_{N_k}\}$ of $\{u_N\}$ such that for $1 \le p \le 4$, $\{u_{N_k}^p\}$ converges weakly in $L^2(\Omega_T)$. Let $\overline{u^{(p)}} \in L^2(\Omega_T)$ be the weak limit of $u_{N_k}^p$. Then, $\mathbf{v}_{N_k}$ and $\mathbf{w}_{N_k}$ converge weakly:

$$\mathbf{v}_{N_k} \rightharpoonup \left(\frac{\overline{u^{(2)}}}{2}, \overline{u^{(1)}}\right) =: \overline{\mathbf{v}} \qquad \text{and} \qquad \mathbf{w}_{N_k} \rightharpoonup \left(\frac{\overline{u^{(2)}}}{2}, -\frac{\overline{u^{(3)}}}{3}\right) =: \overline{\mathbf{w}}.$$

By the div-curl lemma [7, 21, 28], we have

$$\lim_{k\to\infty}\int_{\Omega_T} (\mathbf{v}_{N_k}\cdot \mathbf{w}_{N_k})\, \varphi\, dx\, dt = \int_{\Omega_T} (\overline{\mathbf{v}}\cdot\overline{\mathbf{w}})\, \varphi\, dx\, dt \quad \forall\, \varphi \in C_0^\infty(\Omega_T)\,. \tag{3.25}$$

For all $\varphi \in C_0^\infty(\Omega_T)$,

$$\begin{aligned}
\lim_{k\to\infty}\int_{\Omega_T} (\mathbf{v}_{N_k}\cdot \mathbf{w}_{N_k})\, \varphi\, dx\, dt &= \lim_{k\to\infty}\int_{\Omega_T}\left(\frac{{u_{N_k}}^4}{4} - \frac{{u_{N_k}}^4}{3}\right)\varphi\, dx\, dt \\
&= \lim_{k\to\infty}\int_{\Omega_T} -\frac{{u_{N_k}}^4}{12}\varphi\, dx\, dt = \int_{\Omega_T} -\frac{\overline{u^{(4)}}}{12}\varphi\, dx\, dt.
\end{aligned} \tag{3.26}$$

For the right-hand side of (3.25) we have

$$\int_{\Omega_T} (\overline{\mathbf{v}}\cdot\overline{\mathbf{w}})\,\varphi\,dx\,dt = \int_{\Omega_T} \left(\frac{1}{4}\left(\overline{u^{(2)}}\right)^2 - \frac{1}{3}\overline{u^{(1)}}\,\overline{u^{(3)}}\right)\varphi\,dx\,dt. \tag{3.27}$$

Combining (3.25)–(3.27) yields that

$$\overline{u^{(4)}} = 4\,\overline{u^{(1)}}\,\overline{u^{(3)}} - 3\,\left(\overline{u^{(2)}}\right)^2 \quad \text{a.e.,}$$

which can be used to show that u_{N_k} converges strongly to $\overline{u^{(1)}}$ in $L^2(\Omega_T)$. First, we have

$$\left(u_{N_k} - \overline{u^{(1)}}\right)^4 = {u_{N_k}}^4 - 4\,{u_{N_k}}^3\,\overline{u^{(1)}} + 6\,{u_{N_k}}^2\,\left(\overline{u^{(1)}}\right)^2 - 4\,u_{N_k}\,\left(\overline{u^{(1)}}\right)^3 + \left(\overline{u^{(1)}}\right)^4.$$

Taking the weak limit of both sides of the last equation, we have that

$$\begin{aligned} & w - \lim_{k\to\infty}\left(u_{N_k} - \overline{u^{(1)}}\right)^4 \\ &\quad = \overline{u^{(4)}} - 4\,\overline{u^{(3)}}\,\overline{u^{(1)}} + 6\,\overline{u^{(2)}}\,\left(\overline{u^{(1)}}\right)^2 - 4\,\left(\overline{u^{(1)}}\right)^4 + \left(\overline{u^{(1)}}\right)^4 \\ &\quad = 4\,\overline{u^{(1)}}\,\overline{u^{(3)}} - 3\,\left(\overline{u^{(2)}}\right)^2 - 4\,\overline{u^{(3)}}\,\overline{u^{(1)}} + 6\,\overline{u^{(2)}}\,\left(\overline{u^{(1)}}\right)^2 - 4\,\left(\overline{u^{(1)}}\right)^4 + \left(\overline{u^{(1)}}\right)^4 \\ &\quad = -3\,\left(\overline{u^{(2)}}\right)^2 + 6\,\overline{u^{(2)}}\,\left(\overline{u^{(1)}}\right)^2 - 3\,\left(\overline{u^{(1)}}\right)^4 = -3\,\left[\overline{u^{(2)}} - \left(\overline{u^{(1)}}\right)^2\right]^2 \le 0. \end{aligned}$$

Then

$$0 \le \lim_{k\to\infty}\int_{\Omega_T}\left(u_{N_k} - \overline{u^{(1)}}\right)^4 dx\,dt = \int_{\Omega_T} -3\,\left[\overline{u^{(2)}} - \left(\overline{u^{(1)}}\right)^2\right]^2 dx\,dt \le 0.$$

We now have $\overline{u^{(2)}} = \left(\overline{u^{(1)}}\right)^2$ a.e., which gives us

$$\left\|\overline{u^{(1)}}\right\|^2_{L^2(\Omega_T)} = \int_{\Omega_T}\overline{u^{(2)}}\,dx\,dt = \lim_{k\to\infty}\int_{\Omega_T}{u_{N_k}}^2\,dx\,dt = \lim_{k\to\infty}\|u_{N_k}\|^2_{L^2(\Omega_T)}.$$

Therefore, $\overline{u} := \overline{u^{(1)}}$ is the strong limit of u_{N_k} in $L^2(\Omega_T)$.

3.6. Convergence to a solution of the hyperbolic conservation law. We now show that $\overline{u}$ is a solution of the conservation law (3.4). For all test functions $\varphi \in C_0^\infty(\Omega_T)$,

$$\begin{aligned} &\int_{\Omega_T}\left[\frac{\partial\overline{u}}{\partial t}\varphi + \frac{\partial}{\partial x}\left(\frac{\overline{u}^2}{2}\right)\varphi\right]dx\,dt = -\int_{\Omega_T}\left[\overline{u}\frac{\partial\varphi}{\partial t} + \frac{\overline{u}^2}{2}\frac{\partial\varphi}{\partial x}\right]dx\,dt \\ &\quad = -\int_{\Omega_T}\left[\overline{u^{(1)}}\frac{\partial\varphi}{\partial t} + \frac{\overline{u^{(2)}}}{2}\frac{\partial\varphi}{\partial x}\right]dx\,dt = -\lim_{k\to\infty}\int_{\Omega_T}\left[u_{N_k}\frac{\partial\varphi}{\partial t} + \frac{u_{N_k}^2}{2}\frac{\partial\varphi}{\partial x}\right]dx\,dt \\ &\quad = \lim_{k\to\infty}\int_{\Omega_T}\left[\frac{\partial u_{N_k}}{\partial t}\varphi + \frac{\partial}{\partial x}\left(\frac{u_{N_k}^2}{2}\right)\varphi\right]dx\,dt = \lim_{k\to\infty}\int_{\Omega_T}(\operatorname{div}\mathbf{v}_{N_k})\,\varphi\,dx\,dt. \end{aligned}$$

The right-hand side of last expression vanishes since

$$\begin{aligned} 0 \le \left|\int_{\Omega_T}(\operatorname{div}\mathbf{v}_{N_k})\,\varphi\,dx\,dt\right| &\le \|\operatorname{div}\mathbf{v}_{N_k}\|_{H^{-1}(\Omega_T)}\,\|\varphi\|_{H^1(\Omega_T)} \\ &\le C\,\sqrt{\varepsilon_N}\,\|\varphi\|_{H^1(\Omega_T)} \to 0 \text{ as } k\to\infty \end{aligned}$$

so that $\overline{u}$ is a solution of (3.4).

3.7. Convergence to the entropy solution. We showed in sections 3.5 and 3.6 that $\{u_{N_k}\}$ converges strongly to a solution $\overline{u}$ of the conservation law. We now show that, if the strengthened requirements (3.11) and (3.12) are satisfied, then $\overline{u}$ is the physically relevant entropy solution.

Let $\varphi \in C_0^\infty(\Omega_T)$; then

$$\begin{aligned}\left|\int_{\Omega_T} \left(\overline{u}^3 - {u_{N_k}}^3\right)\varphi\,dx\,dt\right| &= \left|\int_{\Omega_T} \left(\overline{u} - u_{N_k}\right)\left(\overline{u}^2 + \overline{u}\,u_{N_k} + {u_{N_k}}^2\right)\varphi\,dx\,dt\right| \\ &\le \|\overline{u} - u_{N_k}\|_{L^2(\Omega_T)} \left\|\left(\overline{u}^2 + \overline{u}\,u_{N_k} + {u_{N_k}}^2\right)\varphi\right\|_{L^2(\Omega_T)} \\ &\le \|\varphi\|_{L^\infty(\Omega_T)} \|\overline{u} - u_{N_k}\|_{L^2(\Omega_T)} \left\|\overline{u}^2 + \overline{u}\,u_{N_k} + {u_{N_k}}^2\right\|_{L^2(\Omega_T)} \\ &\le \|\varphi\|_{L^\infty(\Omega_T)} \|\overline{u} - u_{N_k}\|_{L^2(\Omega_T)} \left(\left\|\overline{u}^2\right\|_{L^2(\Omega_T)} + \|\overline{u}\,u_{N_k}\|_{L^2(\Omega_T)} + \left\|{u_{N_k}}^2\right\|_{L^2(\Omega_T)}\right) \\ &\le \|\varphi\|_{L^\infty(\Omega_T)} \|\overline{u} - u_{N_k}\|_{L^2(\Omega_T)} \\ &\qquad \left(\|\overline{u}\|^2_{L^4(\Omega_T)} + \|u_{N_k}\|_{L^\infty(\Omega_T)} \|\overline{u}\|_{L^2(\Omega_T)} + \|u_{N_k}\|^2_{L^\infty(\Omega_T)} \sqrt{|\Omega_T|}\right).\end{aligned}$$

Since $\{u_N\}$ is uniformly bounded, we have that $\|u_{N_k}\|_{L^\infty(\Omega_T)} \le C$. Also, $\overline{u}$ is in $L^4(\Omega_T)$ since $\overline{u}^2 = \overline{u^{(2)}} \in L^2(\Omega_T)$. Then, since $\lim_{k\to\infty} \|\overline{u} - u_{N_k}\|_{L^2(\Omega_T)} = 0$, we have that

$$\lim_{k\to\infty} \int_{\Omega_T} (u_{N_k})^3\,\varphi\,dx\,dt = \int_{\Omega_T} \overline{u}^3 \varphi\,dx\,dt.$$

Now, let $\varphi \in C_0^\infty(\Omega_T)$ with $\varphi \ge 0$; then,

$$\begin{aligned}\int_{\Omega_T} \left[\frac{\partial}{\partial t}\left(\frac{\overline{u}^2}{2}\right) + \frac{\partial}{\partial x}\left(\frac{\overline{u}^3}{3}\right)\right]\varphi\,dx\,dt &= -\int_{\Omega_T} \left(\frac{\overline{u}^2}{2}\frac{\partial\varphi}{\partial t} + \frac{\overline{u}^3}{3}\frac{\partial\varphi}{\partial x}\right)dx\,dt \\ &= -\lim_{k\to\infty} \int_{\Omega_T} \left(\frac{{u_{N_k}}^2}{2}\frac{\partial\varphi}{\partial t} + \frac{{u_{N_k}}^3}{3}\frac{\partial\varphi}{\partial x}\right)dx\,dt \\ &= \lim_{k\to\infty} \int_{\Omega_T} \left[\frac{\partial}{\partial t}\left(\frac{{u_{N_k}}^2}{2}\right) + \frac{\partial}{\partial x}\left(\frac{{u_{N_k}}^3}{3}\right)\right]\varphi\,dx\,dt \\ &= \lim_{k\to\infty} \int_{\Omega_T} \left(\operatorname{curl} \mathbf{w}_{N_k}\right)\varphi\,dx\,dt = \lim_{k\to\infty} \int_{\Omega_T} \left(\operatorname{div} \mathbf{v}_{N_k}\right) u_{N_k}\,\varphi\,dx\,dt.\end{aligned} \tag{3.28}$$

Let $z_N = u_N\,\varphi$. For all $t \in (0,T)$, let $z_N^h(\cdot,t) \in S_p^N \cap H_0^1(a,b)$ be the $H^1(a,b)$ projection of z_N so that, for all $v \in S_p^N \cap H_0^1(a,b)$,

$$\int_a^b \frac{\partial z_N^h(\cdot,t)}{\partial x}\,\frac{\partial v}{\partial x}\,dx = \int_a^b \frac{\partial z_N(\cdot,t)}{\partial x}\,\frac{\partial v}{\partial x}\,dx.$$

We then show that, as $k \to \infty$, the right-hand side of (3.28) is nonpositive:

$$
\begin{aligned}
&\int_{\Omega_T} (\text{div } u_N)\; u_N\, \varphi\, dx\, dt = \int_{\Omega_T} (\text{div } u_N)\; z_N\, dx\, dt \\
&= \int_{\Omega_T} (\text{div } u_N)\; z_N^h\, dx\, dt + \int_{\Omega_T} (\text{div } u_{N_k})\; \left(z_N - z_N^h\right)\, dx\, dt \\
&= -\varepsilon_N \int_{\Omega_T} \frac{\partial z_N^h}{\partial x}\, \frac{\partial}{\partial x}\left(Q_N u_N\right)\, dx\, dt + \int_{\Omega_T} (\text{div } u_N)\; \left(z_N - z_N^h\right)\, dx\, dt \\
&= -\varepsilon_N \int_{\Omega_T} \frac{\partial z_N}{\partial x}\, \frac{\partial}{\partial x}\left(Q_N u_N\right)\, dx\, dt + \int_{\Omega_T} (\text{div } u_N)\; \left(z_N - z_N^h\right)\, dx\, dt \\
&= -\varepsilon_N \int_{\Omega_T} \frac{\partial z_N}{\partial x}\, \frac{\partial u_N}{\partial x}\, dx\, dt + \varepsilon_N \int_{\Omega_T} \frac{\partial z_N}{\partial x}\, \frac{\partial}{\partial x}\left[\left(I - Q_N\right) u_N\right]\, dx\, dt \\
&\quad + \int_{\Omega_T} (\text{div } u_N)\; \left(z_N - z_N^h\right)\, dx\, dt \\
&= -\varepsilon_N \int_{\Omega_T} \frac{\partial}{\partial x}\left(u_N\, \varphi\right)\, \frac{\partial u_N}{\partial x}\, dx\, dt \\
&\quad + \varepsilon_N \int_{\Omega_T} \frac{\partial z_N}{\partial x}\, \frac{\partial}{\partial x}\left[\left(I - Q_N\right) u_N\right]\, dx\, dt + \int_{\Omega_T} (\text{div } u_N)\; \left(z_N - z_N^h\right)\, dx\, dt \\
&= -\varepsilon_N \int_{\Omega_T} \varphi\, \left|\frac{\partial u_N}{\partial x}\right|^2\, dx\, dt - \varepsilon_N \int_{\Omega_T} u_N \frac{\partial \varphi}{\partial x}\, \frac{\partial u_N}{\partial x}\, dx\, dt \\
&\quad + \varepsilon_N \int_{\Omega_T} \frac{\partial z_N}{\partial x}\, \frac{\partial}{\partial x}\left[\left(I - Q_N\right) u_N\right]\, dx\, dt + \int_{\Omega_T} (\text{div } u_N)\; \left(z_N - z_N^h\right)\, dx\, dt.
\end{aligned}
\tag{3.29}
$$

For the second term on the right-hand side of (3.29), we have that

$$
\begin{aligned}
\left| -\varepsilon_N \int_{\Omega_T} u_N \frac{\partial \varphi}{\partial x}\, \frac{\partial u_N}{\partial x}\, dx\, dt \right| &\le \varepsilon_N\, \|u_N\|_{L^\infty(\Omega_T)} \left\| \frac{\partial u_N}{\partial x} \right\|_{L^2(\Omega_T)} \left\| \frac{\partial \varphi}{\partial x} \right\|_{L^2(\Omega_T)} \\
&\le C\, \sqrt{\varepsilon_N} \left\| \frac{\partial \varphi}{\partial x} \right\|_{L^2(\Omega_T)} \to 0 \text{ as } N \to \infty.
\end{aligned}
$$

For the third term on the right-hand side of (3.29), we have that

$$
\begin{aligned}
&\left| \varepsilon_N \int_{\Omega_T} \frac{\partial z_N}{\partial x}\, \frac{\partial}{\partial x}\left[\left(I - Q_N\right) u_N\right]\, dx\, dt \right| \le \varepsilon_N \left\| \frac{\partial z_N}{\partial x} \right\|_{L^2(\Omega_T)} \left\| \frac{\partial}{\partial x}\left[\left(I - Q_N\right) u_N\right] \right\|_{L^2(\Omega_T)} \\
&= \varepsilon_N \left\| \frac{\partial}{\partial x}\left(u_N\, \varphi\right) \right\|_{L^2(\Omega_T)} \left\| \frac{\partial}{\partial x}\left[\left(I - Q_N\right) u_N\right] \right\|_{L^2(\Omega_T)} \le \varepsilon_N \left\| \frac{\partial}{\partial x}\left[\left(I - Q_N\right) u_N\right] \right\|_{L^2(\Omega_T)} \\
&\qquad\qquad \left(\|\varphi\|_{L^\infty(\Omega_T)} \left\| \frac{\partial u_N}{\partial x} \right\|_{L^2(\Omega_T)} + \|u_N\|_{L^\infty(\Omega_T)} \left\| \frac{\partial \varphi}{\partial x} \right\|_{L^2(\Omega_T)} \right) \\
&\le C\, \sqrt{\varepsilon_N} \left\| \frac{\partial}{\partial x}\left[\left(I - Q_N\right) u_N\right] \right\|_{L^2(\Omega_T)} \left(\|\varphi\|_{L^\infty(\Omega_T)} + \sqrt{\varepsilon_N} \left\| \frac{\partial \varphi}{\partial x} \right\|_{L^2(\Omega_T)} \right) \to 0 \text{ as } N \to \infty.
\end{aligned}
$$

For the fourth term on the right-hand side of (3.29), we have that

$$
\begin{aligned}
&\left| \int_{\Omega_T} (\text{div } u_N)\; \left(z_N - z_N^h\right)\, dx\, dt \right| \le \|\text{div } u_N\|_{L^2(\Omega_T)}\, \left\|z_N - z_N^h\right\|_{L^2(\Omega_T)} \\
&\le C\, h_N\, \|\text{div } u_N\|_{L^2(\Omega_T)} \left\| \frac{\partial z_N}{\partial x} \right\|_{L^2(\Omega_T)} \le C\, \frac{h_N}{\sqrt{\varepsilon_N}} \left\| \frac{\partial z_N}{\partial x} \right\|_{L^2(\Omega_T)}
\end{aligned}
$$

$$= C\,\frac{h_N}{\sqrt{\varepsilon_N}}\left\|\frac{\partial}{\partial x}\left(u_N\,\varphi\right)\right\|_{L^2(\Omega_T)}$$
$$\leq C\,\frac{h_N}{\sqrt{\varepsilon_N}}\left(\|u_N\|_{L^\infty(\Omega_T)}\left\|\frac{\partial\varphi}{\partial x}\right\|_{L^2(\Omega_T)}+\|\varphi\|_{L^\infty(\Omega_T)}\left\|\frac{\partial u_N}{\partial x}\right\|_{L^2(\Omega_T)}\right)$$
$$\leq C\Big(\frac{h_N}{\sqrt{\varepsilon_N}}\left\|\frac{\partial\varphi}{\partial x}\right\|_{L^2(\Omega_T)}+\frac{h_N}{\varepsilon_N}\,\|\varphi\|_{L^\infty(\Omega_T)}\Big)\to 0 \text{ as } N\to\infty.$$

Thus, we have shown that the second, third, and fourth terms on the right-hand side of (3.29) vanish as $N\to\infty$. Since the first term is clearly nonpositive, we have that

$$\int_{\Omega_T}\left[\frac{\partial}{\partial t}\left(\frac{\overline{u}^2}{2}\right)+\frac{\partial}{\partial x}\left(\frac{\overline{u}^3}{3}\right)\right]\varphi\,dx\,dt=\liminf_{k\to\infty}\int_{\Omega_T}\left(\operatorname{div}\,u_{N_k}\right)\,u_{N_k}\,\varphi\,dx\,dt\leq 0$$

so that the $\overline{u}$ is the entropy solution of (3.4). This completes the proof of Theorem 3.1.

3.8. Verifying the hypotheses of Theorem 3.1. In the hierarchical finite element formulation (3.6), we have to choose ε_N, m, and the form of Q_N for $k\leq m$. Let $0<\delta\leq\theta\leq 1$. We then choose

$$\varepsilon_N=C\,{h_N}^{\theta},\qquad m_H\leq\frac{\delta\,N}{2},\qquad\text{and}\qquad Q_{k,i}=\begin{cases}0, & k\leq m_H,\\ 1, & k>m_H.\end{cases}$$

It is then evident that $\varepsilon_N, h_N\to 0$ as $N\to\infty$ so (3.7) holds. Since $0<\theta\leq 1$, $\frac{\varepsilon_N}{h_N}=C\,{h_N}^{\theta-1}\geq C$ so that (3.8) also holds.

Let $v\in S_p^N$; $(I-Q_N)$ is simply an interpolation operator on a coarse grid so that

$$\|(I-Q_N)\,v\|_{L^2(a,b)}\leq C\,\|v\|_{L^2(a,b)}\,.$$

Q_N retains the high frequencies of a function, so $(I-Q_N)$ eliminates them: $Q_{k,i}=1\Rightarrow 1-Q_{k,i}=0$ for $k>m_H$ so that $(I-Q_N)\,v\in S_p^{m_H}$. Using a standard inverse estimate,

$$\left\|\frac{\partial}{\partial x}\left[(I-Q_N)\,u_N\right]\right\|_{L^2(a,b)}\leq C\,\left(h_{m_H}\right)^{-1}\|(I-Q_N)\,v\|_{L^2(a,b)}$$
$$\leq C\,M^{m_H}\,\|(I-Q_N)\,v\|_{L^2(a,b)}\leq C\,M^{\frac{N\,\delta}{2}}\,\|(I-Q_N)\,v\|_{L^2(a,b)}$$
$$\leq C\,\left(\frac{\nu}{h_N}\right)^{\frac{\delta}{2}}\|(I-Q_N)\,v\|_{L^2(a,b)}\leq C\,\left(\frac{\nu}{h_N}\right)^{\frac{\delta}{2}}\|v\|_{L^2(a,b)}$$
$$=C\,\sqrt{\nu^\delta}\,{h_N}^{-\frac{\delta}{2}}\,\|v\|_{L^2(a,b)}$$

and

$$\begin{aligned}\sqrt{\varepsilon_N}\left\|\frac{\partial}{\partial x}\left[(I-Q_N)\,u_N\right]\right\|_{L^2(a,b)}&\leq C\,\sqrt{\nu^\delta}\,\sqrt{\frac{\varepsilon_N}{h_N^\delta}}\,\|v\|_{L^2(a,b)}\\&=C\,\sqrt{{h_N}^{\theta-\delta}}\,\|v\|_{L^2(a,b)}\,.\end{aligned}\tag{3.30}$$

Since $\delta\leq\theta$, we have that ${h_N}^{\theta-\delta}\leq C$; therefore, (3.9) is satisfied.

Since $(I-Q_N)$ is an interpolation operator on a coarse grid, for all $v\in S_p^N$,

$$\left\|\frac{d}{dx}\left(Q_N v\right)\right\|_{L^2(a,b)}=\left\|\frac{dv}{dx}-\frac{d}{dx}\left[(I-Q_N)\,v\right]\right\|_{L^2(a,b)}\leq C\left\|\frac{dv}{dx}\right\|_{L^2(a,b)}.$$

Therefore, (3.10) is satisfied. This completes the verification of the hypotheses (3.7)–(3.10) of Theorem 3.1 that are used to prove the convergence of the hierarchical finite element approximations.

To verify the hypotheses (3.11) and (3.12) of Theorem 3.1 that are used to prove the convergence of the hierarchical finite element approximations to the entropy solution, we must choose $0 < \theta < 1$. In this case, $\frac{\varepsilon_N}{h_N} = C\,{h_N}^{\theta-1} \to \infty$ as $N \to \infty$ so that (3.11) holds. Since now $\delta < \theta$ so that ${h_N}^{\theta-\delta} \to 0$ as $N \to \infty$, (3.30) implies that (3.12) holds.

4. A simple computational illustration. We consider the simple periodic problem for the Burgers equation in one dimension:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x}\left(\dfrac{u^2}{2}\right) = 0 \text{ on } (-1,+1) \times (0,T) \\ u\,(-1,t) = u\,(+1,t) \text{ for all } t \in (0,T) \\ u\,(x,0) = 1 + \dfrac{1}{2}\sin\left(\pi x\right). \end{cases} \tag{4.1}$$

A means for establishing the exact solution of this problem is given in [8, 16]. All of our numerical results were generated using the finite element library `deal.II` [2]. More extensive computational experimentations are provided in [3, 4].

In (2.1), the values of several parameters must be chosen. We set $\varepsilon_N = h_N$ and add diffusion only to the finest level: $Q_{N;n,\imath} = 0$ for $n < N$. Neither of these choices satisfies the requirements of Theorem 3.1 for convergence to the entropy solution. Even with the smaller diffusion term, however, our numerical experiments indicate that the approximations still converge to the correct solution.

After spatial discretization is effected using linear hierarchical finite element functions, the resulting system of ordinary differential equations is integrated using a third-order, strong, stability-preserving Runge–Kutta method found in [11], with time step Δt that satisfies the CFL condition $\Delta t/h_N \sup u \le 0.2$.

The exact and discrete solutions of (4.1) are given in Figure 5. Because we are approximating a discontinuous solution with continuous piecewise polynomials, we see Gibbs oscillations near the discontinuity; see Figure 5. A simple post-processing strategy to remove the oscillations is to set the coefficients of the hierarchical expansion to zero around the discontinuity. The question then becomes how to determine the location of the discontinuity. Let $\beta_{n+1,\imath}$ be a high frequency hierarchical coefficient in the discrete solution. Let $\beta_{n,\jmath}$ be the parent hierarchical coefficient, so the support of $\psi_{n+1,\imath}$ is a subset of $\psi_{n,\jmath}$. If the solution is continuously differentiable in the region of the support of $\psi_{n,\jmath}$, then $\beta_{n,\jmath}/\beta_{n+1,\imath} \approx 2$. Thus, our simple post-processing strategy is as follows: For the highest four frequencies, if a hierarchical coefficient is larger than half the value of its parent, then it is set to zero. Our simple post-processing strategy only affects the region around a discontinuity, but it has the disadvantage of smoothing across the discontinuity. See Figure 6. We note that post-processing strategies must also be applied in the spectral viscosity method in order to reduce the size of the Gibbs oscillations.

In Table 1, we use the L^1 norm to measure errors in the approximate solution. Near a discontinuity, we are limited to how well a piecewise polynomial can approximate a solution. We are more interested in the convergence rates in the smooth regions. We therefore exclude a region of length 0.2 around the discontinuity in our error calculations. We see that away from the discontinuity, we achieve the optimal

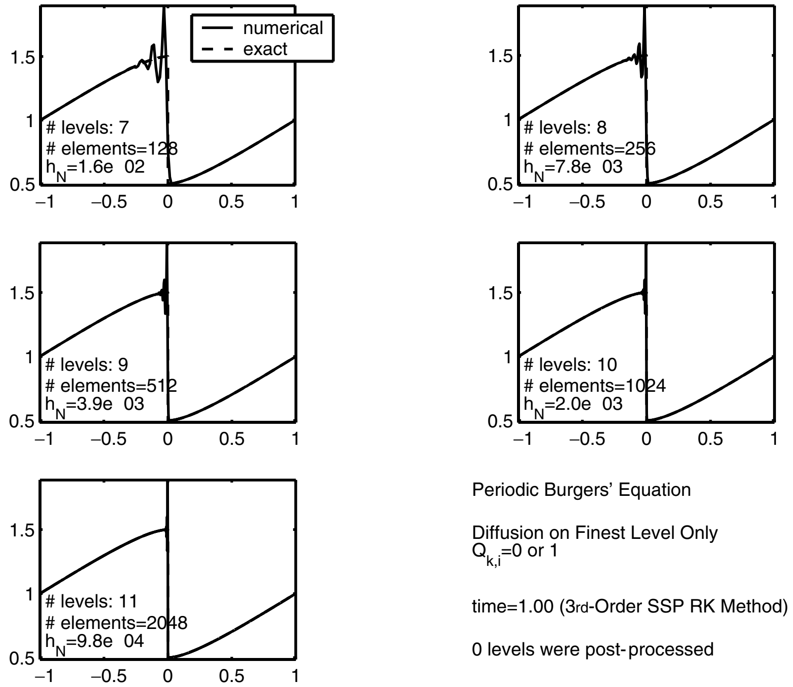


FIG. 5. *Solution of periodic Burgers equation with linear polynomials (without post-processing).*

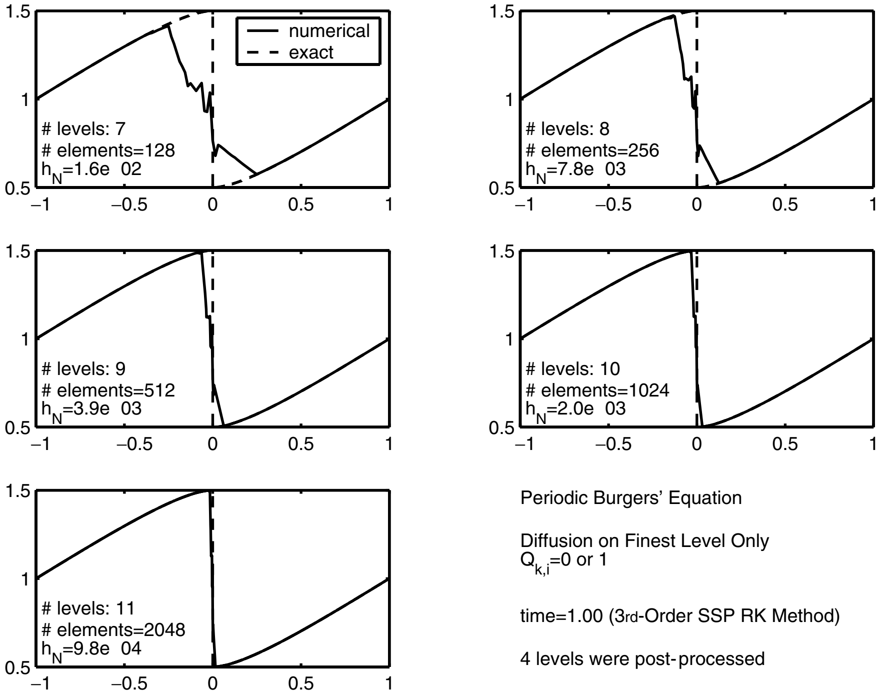


FIG. 6. *Solution of periodic Burgers equation with linear polynomials (with post-processing).*

Table 1
Convergence rate for the periodic Burgers equation using linear hierarchical basis functions.

	without post-processing		with post-processing	
levels	L^1 error	rate	L^1 error	rate
8	5.775e-03	-	5.238e-02	-
9	5.321e-04	3.44	3.124e-03	4.07
10	2.221e-05	4.58	2.517e-05	6.96
11	3.691e-06	2.59	3.691e-06	2.77
12	9.230e-07	2.00	9.230e-07	2.00
13	2.307e-07	2.00	2.307e-07	2.00
14	5.768e-08	2.00	5.768e-08	2.00

error rate with or without post-processing. We have no theoretical justification for these convergence rates, but this is a common failing for conservation laws.

5. Concluding remarks. Initial results for the new method seem promising. We have a stable finite element method which, in some cases, attains quasi-optimal convergence rates in smooth regions. We also have developed a theoretical foundation for understanding why the method works. These results, however, are preliminary. There are potential pitfalls awaiting in more complicated problems, but there is also untapped potential within the framework. Hierarchical bases, for example, should provide a suitable environment for implementing adaptive strategies, both for the grid and the diffusion term.

Acknowledgments. The authors wish to thank Professor Eitan Tadmor and Dr. Curtis Ober for many helpful discussions.

REFERENCES

[1] R. Bank and J. Xu, *An algorithm for coarsening unstructured meshes*, Numer. Math., 73 (1996), pp. 1–36.

[2] W. Bangerth, R. Hartmann, and G. Kanschat, deal.II *Differential Equations Analysis Library, Technical Reference*, IWR, Heidelberg, http://www.dealii.org.

[3] M. Calhoun-Lopez, *Numerical Solutions of Hyperbolic Conservation Laws: Incorporating Multi-Resolution Viscosity Methods into the Finite Element Framework*, Ph.D. thesis, Iowa State University, Ames, IA, 2003.

[4] M. Calhoun-Lopez and M. Gunzburger, *Implementation and testing of finite element, multi-resolution viscosity method for hyperbolic conservation laws*, Comp. Methods Appl. Mech. Engrg., submitted.

[5] G. Chen, Q. Du, and E. Tadmor, *Spectral viscosity approximations to multidimensional scalar conservation laws*, Math. Comp., 61 (1993), pp. 629–643.

[6] B. Cockburn, G. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods Theory, Computation, and Applications, B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., Springer, Berlin, 2000, pp. 3–50.

[7] L. Evans, *Weak Convergence Methods for Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1990.

[8] L. Evans, *Partial Differential Equations*, AMS, Providence, RI, 1998.

[9] A. Gelb and E. Tadmor, *Enhanced spectral viscosity approximations for conservation laws*, Appl. Numer. Math., 33 (2000), pp. 3–21.

[10] E. Godlewski and P.-A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer, New York, 1996.

[11] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[12] B.-Y. Guo, H.-P. Ma, and E. Tadmor, *Spectral vanishing viscosity method for nonlinear conservation laws*, SIAM J. Numer. Anal., 39 (2001), pp. 1254–1268.

[13] C. Johnson and A. Szepessy, *On the convergence of a finite element method for a nonlinear hyperbolic conservation law*, Math. Comp., 49 (1987), pp. 427–444.

[14] C. Johnson, A. Szepessy, and P. Hansbo, *On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws*, Math. Comp., 54 (1990), pp. 107–129.

[15] R. Kornhuber and H. Yserentant, *Multilevel methods for elliptic problems on domains not resolved by the coarse grid*, in Domain Decomposition Methods in Scientific and Engineering Computing, AMS, Providence, RI, 1994, pp. 49–60.

[16] P. Lax, *Hyperbolic Systems of Conservation Laws and the Mathematical Theory of Shock Waves*, SIAM, Philadelphia, 1973.

[17] R. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser, Basel, 1992.

[18] R. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.

[19] Y. Maday, S. M. Ould Kaber, and E. Tadmor, *Legendre pseudospectral viscosity method for nonlinear conservation laws*, SIAM J. Numer. Anal., 30 (1993), pp. 321–342.

[20] Y. Maday and E. Tadmor, *Analysis of the spectral vanishing viscosity method for periodic conservation laws*, SIAM J. Numer. Anal., 26 (1989), pp. 854–870.

[21] F. Murat, *L'injection du cône positif de H^{-1} dans $W^{-1,q}$ est compacte pour tout $q < 2$*, J. Math. Pures Appl. (9), 60 (1981), pp. 309–322.

[22] B. Perthame and E. Tadmor, *A kinetic equation with kinetic entropy functions for scalar conservation laws*, Comm. Math. Phys., 136 (1991), pp. 501–517.

[23] A. Szepessy, *Convergence of a shock-capturing streamline diffusion finite element method for a scalar conservation law in two space dimensions*, Math. Comp., 53 (1989), pp. 527–545.

[24] E. Tadmor, *Convergence of spectral methods for nonlinear conservation laws*, SIAM J. Numer. Anal., 26 (1989), pp. 30–44.

[25] E. Tadmor, *Total variation and error estimates for spectral viscosity approximations*, Math. Comp., 60 (1993), pp. 245–256.

[26] E. Tadmor, *Super-viscosity and spectral approximations of nonlinear conservation laws*, in Numerical Methods for Fluid Dynamics, Vol. IV, M. J. Baines and K. W. Morton, eds., Oxford University Press, New York, 1993, pp. 69–81.

[27] E. Tadmor, *Approximate solutions of nonlinear conservation laws and related equations*, in Recent Advances in Partial Differential Equations, AMS, Providence, RI, 1998, pp. 325–368.

[28] L. Tartar, *Compensated compactness and applications to partial differential equations*, in Nonlinear Analysis and Mechanics: Heriot-Watt Symposium, Vol. IV, Pitman, Boston, 1979, pp. 136–212.

[29] H. Yserentant, *On the multilevel splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

[30] H. Yserentant, *Hierarchical bases*, in Proceedings of ICIAM 91, R. O'Malley, ed., SIAM, Philadelphia, 1992, pp. 256–276.

[31] O. Zienkiewicz, D. Kelly, J. Gago, and I. Babuška, *Hierarchical finite element approaches, error estimates and adaptive refinement*, in The Mathematics of Finite Elements and Applications, Vol. IV, J. Whiteman, ed., Academic Press, New York, 1982, pp. 313–346.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2012–2033

A UNIFIED ANALYSIS FOR CONFORMING AND NONCONFORMING STABILIZED FINITE ELEMENT METHODS USING INTERIOR PENALTY[*]

ERIK BURMAN[†]

Abstract. We discuss stabilized Galerkin approximations in a new framework, widening the scope from the usual dichotomy of the discontinuous Galerkin method on the one hand and Petrov–Galerkin methods such as the SUPG method on the other. The idea is to use interior penalty terms as a means of stabilizing the finite element method using conforming or nonconforming approximation, thus circumventing the need of a Petrov–Galerkin-type choice of spaces. This is made possible by adding a higher-order penalty term giving L^2-control of the jumps in the gradients between adjacent elements. We consider convection-diffusion-reaction problems using piecewise linear approximations and prove optimal order a priori error estimates for two different finite element spaces, the standard H^1-conforming space of piecewise linears and the nonconforming space of piecewise linear elements where the nodes are situated at the midpoint of the element sides (the Crouzeix–Raviart element). Moreover, we show how the formulation extends to discontinuous Galerkin interior penalty methods in a natural way by domain decomposition using Nitsche's method.

Key words. convection-diffusion problem, interior penalty, finite element approximation, Crouzeix–Raviart element

AMS subject classifications. 65N30, 65N12, 35L50

DOI. 10.1137/S0036142903437374

1. Introduction.

The solution of convection-diffusion problems with dominating convection using finite element methods has been the object of much research during the last 30 years. Essentially, the field has been separated into two main branches, Petrov–Galerkin methods in cases where conforming approximation is used [4, 16] and discontinuous Galerkin with interior penalty when nonconforming approximation is used [19, 17, 13]. Of course the discontinuous Galerkin method may also be supplied with an SUPG-type stabilization as in [25], and there is the SUPG method using the Crouzeix–Raviart element [15, 14, 20], which needs both interior penalty and Petrov–Galerkin-type approximation spaces to be stable in the limit of vanishing diffusion. So the current state of affairs seems to be that Petrov–Galerkin-type approximations are necessary for all approximations except the discontinuous Galerkin method. This is not satisfactory since the SUPG-method in practice suffers from several shortcomings:

- The mass matrix may not be lumped. This may severely reduce performance when solving large reactive systems using low-order elements.
- The consistency requirements practically impose the use of a space-time finite element approach for time-stepping, using discontinuous approximation in time. The practical implementation of such techniques is rather involved and requires additional unknowns.
- The stabilization parameter depends on the diffusion. This may lead to complications when computing the solution of large coupled systems with a

[*] Received by the editors November 4, 2003; accepted for publication (in revised form) November 12, 2004; published electronically December 15, 2005.
http://www.siam.org/journals/sinum/43-5/43737.html

[†] Institut d'analyse et calcul scientifique, Ecole Polytechnique Federale de Lausanne, CH 1015 Lausanne, Switzerland (Erik.Burman@epfl.ch).

complex diffusion matrix or in cases when the diffusion/viscosity depends in a strongly nonlinear way on the solution.

The discontinuous Galerkin (DG) method, on the other hand, behaves well with respect to these above mentioned points but suffers from the fact that it involves a larger number of degrees of freedom due to the discontinuous approximation space. In fact, memory requirements of the DG method are typically a factor 7–10 larger than those of the SUPG method. Hence there is a strong motivation to find methods that use more economic spaces that do not suffer from the same disadvantages as the SUPG method.

In this paper we will go beyond this dichotomy between conforming finite element methods using Petrov–Galerkin-type stabilizations and discontinuous Galerkin methods using interior penalty-type stabilization and adopt a different point of view, where the interior penalty is the main stabilization. We also show that interior penalty stabilization is sufficient not only for the discontinuous Galerkin method but also for conforming piecewise linear finite element approximations, even in the case when the same trial and test spaces are used. The outline of the paper is as follows: In the next section we introduce the model problem and discuss in more detail this new framework; in section 3 we then consider the limiting case of conforming piecewise linear approximation stabilized by using only an interior penalty term; we prove stability and a priori error estimates. Then we use these results in the general framework and extend the method to the nonconforming case of Crouzeix–Raviart-type finite element approximation (continuity at the midpoints of the element sides) in section 4. In section 5 we discuss domain decomposition using Nitsche's method and how this naturally leads to discontinuous Galerkin-type interior penalty methods. The performance of the method is shown numerically in section 6. Finally, we draw some conclusions in section 7.

2. A new framework.

As a model problem we propose the convection-diffusion-reaction equation

$$\left\{ \begin{aligned} \beta\cdot\nabla u + \sigma u - \varepsilon\Delta u &= f \qquad \text{in } \Omega,\\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned}\right. \tag{2.1}$$

where Ω is a bounded open connected subset of $\mathbb{R}^d$ with a Lipschitz boundary $\partial\Omega$, $d = 2$ or 3 is the space dimension, $\beta \in [W^{1,\infty}(\Omega)]^d$ is a velocity field, $\varepsilon > 0$ is a diffusion coefficient (that may be zero if the boundary conditions are modified), and $\sigma > 0$ is the reaction coefficient, $f \in L^2(\Omega)$. We will use the notation $\partial\Omega_{in}$ ($\partial\Omega_{out}$) for the subset of $\partial\Omega$ such that $\beta\cdot n < 0$ ($\beta\cdot n > 0$). We assume that the following standard coercivity condition holds:

$$\sigma - \frac{1}{2}\nabla\cdot\beta \geq \sigma_0 > 0, \tag{2.2}$$

and we define the associated parameter σ_1 by

$$\sigma_1 = \text{ess sup}_{x\in\Omega}\frac{|\sigma - \nabla\cdot\beta|^2}{\sigma_0}.$$

Problem (2.1) is well-posed thanks to the Lax–Milgram lemma, and we will always assume that the solution is sufficiently smooth, i.e., $u \in H^2(\Omega)$.

Remark 2.1. An analysis including the case $\sigma_0 = 0$ could be undertaken using exponentially weighted test functions following [21] but is beyond the scope of the present paper.

For the finite element formulation of this problem, we introduce some additional notation. Let $\mathcal{T}_h$ be a triangulation of Ω, without hanging nodes, and let V_h be a space of (conforming or nonconforming) piecewise linear functions defined on $\mathcal{T}_h$. Let S_i be a vertex of $\mathcal{T}_h$, $\varphi_i \in V_h$ the associated nodal basis function, and denote by Ω_i the macro-element formed by the elements K in $\mathcal{T}_h$ sharing vertex S_i. Let $\mathcal{E}_i$ denote the set of faces connected to S_i. Let h_K denote the diameter of an element K and set $h = \max_{K\in\mathcal{T}_h} h_K$. Moreover, we shall assume that there exists a constant $\rho > 0$ such that for all vertices S_i in $\mathcal{T}_h$, we have

$$\max_{e\in\mathcal{E}_i} h_e \leq \rho \min_{e\in\mathcal{E}_i} h_e, \tag{2.3}$$

where $h_e = |e|$ is the length of edge e. Property (2.3) was introduced in [5] and is a local quasi-uniformity property of the mesh. It implies that for each node S_i there is a finite number, n_ρ, of elements in Ω_i. The jump $[x]_f$ of a quantity x over an interior face f will be defined by $[x(\xi)]_f = \lim_{\epsilon\to 0^+}(x(\xi - n_f\epsilon) - x(\xi + n_f\epsilon))$, where n_f denotes a normal vector to the face f having an arbitrary but fixed orientation and $\xi \in f$. The subscript is omitted when there is no ambiguity. For faces such that $f \subset \partial\Omega$, we define n_f as the outward pointing normal and set $[x]_f \equiv 0$. By $\{x\}_f$ we denote the average value of x over face f, $\{x(\xi)\}_f = \lim_{\epsilon\to 0^+}\frac{1}{2}(x(\xi - n_f\epsilon) + x(\xi + n_f\epsilon))$. Tangential vectors of a face f will be denoted τ_f ($\tau_f \cdot n_f = 0$). Furthermore, we will use the notation $(x,y)_X = \int_X x\cdot y \,\mathrm{dx}$, $\quad\langle x,y\rangle_{\partial X} = \int_{\partial X} x\cdot y \,\mathrm{ds}$ with the elementwise counterparts $(x,y)_{X,h} = \sum_{K\in X}\int_K x\cdot y\,\mathrm{dx}$ and $\quad\langle x,y\rangle_{\partial X,h} = \sum_{f\in\partial X}\int_f x\cdot y\,\mathrm{ds}$. Let $\|x\|_{\partial X} = (x,x)_X^{1/2}$ denote the L^2-norm over X and $|x|_X = \langle x,x\rangle_{\partial X}^{1/2}$ the L^2-norm over ∂X with the elementwise counterparts $\|x\|_{X,h} = (x,x)_{X,h}^{1/2}$ and $|x|_{\partial X,h} = \langle x,x\rangle_{\partial X,h}^{1/2}$, respectively. When the subscript X or ∂X is omitted, the norm is taken over the domain Ω or its boundary $\partial\Omega$. The norm of the space $H^i(X)$ will be denoted $\|x\|_{i,X}$ with $i = 1,2$. We will use c and C to denote generic positive constants independent of h_K but not necessarily of the local mesh geometry.

The general discretization for (2.1) typically takes the following form: Find $u_h \in V_h$ such that

$$A(u_h,v_h) + \sum_{i=0}^{1} J_i(u_h,v_h) = (f,v_h) \quad \forall v_h \in W_h, \tag{2.4}$$

where

$$\begin{aligned}(2.5)\qquad & \\ A(u_h,v_h) &= (\sigma u_h, v_h) + (\varepsilon\nabla u_h, \nabla v_h)_h + (\beta\cdot\nabla u_h, v_h)_h \\ -\tfrac{1}{2}\textstyle\sum_K\Big(\langle\beta\cdot n[u_h],\{v_h\}\rangle_{\partial K\setminus\partial\Omega} &+ \langle\{\varepsilon\nabla u_h\cdot n\},[v_h]\rangle_{\partial K\setminus\partial\Omega} + \langle\{\varepsilon\nabla v_h\cdot n\},[u_h]\rangle_{\partial K\setminus\partial\Omega}\Big)^* \\ &- \langle\varepsilon\nabla u_h\cdot n, v_h\rangle_h - \langle\varepsilon\nabla v_h\cdot n, u_h\rangle_h \\ &+ \left\langle\gamma_{bc}\tfrac{\varepsilon}{h}u_h, v_h\right\rangle + \langle|\beta\cdot n|u_h, v_h\rangle_{\partial\Omega_{in}},\end{aligned}$$

$$J_0(u_h,v_h) = \sum_K \langle\gamma_0(h)[u_h],[v_h]\ \rangle_{\partial K\setminus\partial\Omega},$$

and

$$J_1(u_h,v_h) = \sum_K \langle\gamma_1(h)[\nabla u_h],[\nabla v_h]\ \rangle_{\partial K\setminus\partial\Omega}, \tag{2.6}$$

with $\gamma_i(h) = \tilde{\gamma}_i h^{s_i}$ and s_i chosen so as to obtain optimal stability and approximation properties. W_h denote some test space, the choice of which will be discussed later. γ_{bc} denotes the penalization parameter for the weakly imposed boundary condition. Moreover, we have marked with an asterisk the terms that are present only when nonconforming approximation spaces are used. It should be noted that the term $J_1(u_h, v_h)$ can be decomposed in the streamline and the crosswind part. To this end, we assume that $|\beta| > 0$ and define the unit vector parallel to β as $e_\beta = \frac{\beta}{|\beta|}$ and the unit vector orthogonal to β such that $e_{\beta^\perp}$. Clearly we may decompose the gradient in the orthogonal basis formed by $\{e_\beta, e_{\beta^\perp}\}$ (in two space dimensions).

$$\nabla u_h = (e_\beta \cdot \nabla u_h) e_\beta + (e_{\beta^\perp} \cdot \nabla u_h) e_{\beta^\perp}.$$

Plugging this into (2.6) yields for the jumps

$$\begin{aligned}[\nabla u_h] \cdot [\nabla v_h] &= [(e_\beta \cdot \nabla u_h) e_\beta + (e_{\beta^\perp} \cdot \nabla u_h) e_{\beta^\perp}] \cdot [(e_\beta \cdot \nabla v_h) e_\beta + (e_{\beta^\perp} \cdot \nabla v_h) e_{\beta^\perp}] \\ &= [e_\beta \cdot \nabla u_h][e_\beta \cdot \nabla v_h] + [e_{\beta^\perp} \cdot \nabla u_h][e_{\beta^\perp} \cdot \nabla v_h].\end{aligned}$$

This implies that one may use the following form of the stabilization term:

$$\begin{aligned}J_1(u_h, v_h) &= \textstyle\sum_K \left\langle \gamma_{1,\beta}(h)[e_\beta \cdot \nabla u_h], [e_\beta \cdot \nabla v_h] \right\rangle_{\partial K \setminus \partial\Omega} \\ &\quad + \textstyle\sum_K \left\langle \gamma_{1,\beta^\perp}(h)[e_{\beta^\perp} \cdot \nabla u_h], [e_{\beta^\perp} \cdot \nabla v_h] \right\rangle_{\partial K \setminus \partial\Omega},\end{aligned}$$

which coincides with (2.6) when $\gamma_{1,\beta}(h) = \gamma_{1,\beta^\perp}(h)$. For stability, however, it is only essential that the parameter $\gamma_{1,\beta}(h)$ is large enough. The parameter $\gamma_{1,\beta^\perp}(h)$ may be set to zero. We note that in the case of piecewise linear continuous functions u_h, there holds $[\tau_f \cdot \nabla u_h]_f = 0$. Using this observation, we may introduce some further simplifications of the stabilization term. This time, consider the decomposition of the gradient in the directions normal and tangential to the element edge; for the jump in the streamline derivative we then obtain

$$\begin{aligned}[\beta \cdot \nabla u_h]_f &= [\beta \cdot ((n_f \cdot \nabla u_h) n_f + (\tau_f \cdot \nabla u_h) \tau_f)]_f \\ &= [\beta \cdot n_f (n_f \cdot \nabla u_h)]_f + [\beta \cdot \tau_f (\tau_f \cdot \nabla u_h)]_f.\end{aligned}$$

However, since the tangential jump is zero, the second term in the right-hand side vanishes and we may readily deduce that on each face we have $[\beta \cdot \nabla u_h]_f [\beta \cdot \nabla v_h]_f = |\beta \cdot n_f|^2 [n_f \cdot \nabla u_h]_f [n_f \cdot \nabla v_h]_f$. Using once again the fact that the tangential jump is zero, it follows that the product of the jumps in the normal component equals the scalar product of the jump in the full gradient; $[\beta \cdot \nabla u_h]_f [\beta \cdot \nabla v_h]_f = |\beta \cdot n_f|^2 [\nabla u_h]_f \cdot [\nabla v_h]_f$. Hence in this case the streamline diffusion character of the stabilization may be included in the parameter γ_1 in (2.6). We also recall that the addition of stabilization in the crosswind direction increases the accuracy of the approximation close to interior layers; see [18]. The first term on the second line of the expression for $A(u_h, v_h)$ is related to the consistency error of the convective term. This term can be chosen in a variety of different ways, related to what numerical flux one wishes to use in the nonconforming method. We will not pursue this further here, but only point out that the whole parenthesis marked * vanishes for conforming approximation. Note that in the above formulation we impose the boundary conditions weakly, see [22]; this is natural when considering the general framework, since the finite element solution may be nonconforming. It also has some advantages from the point of view of the analysis. For results using conforming piecewise linear approximation and strongly imposed boundary conditions, we refer to [6]. The SUPG method is typically obtained by choosing $\gamma_1 = 0$ and taking $W_h = \{w_h : w_h = v_h + \delta\beta \cdot \nabla v_h,\, v_h \in V_h\}$. Then γ_0 has to be chosen correctly in order to ensure the coercivity of $A(u_h, v_h)$ in the nonconforming case. However, our main concern in this paper is the case $W_h = V_h$. We will show that the use of approximation spaces that previously needed SUPG-type stabilization may, in fact, be stabilized using interior penalty only. The key observation is that the following inequality holds:

$$\inf_{\zeta_h \in V_h} \|h^{1/2}(\beta \cdot \nabla u_h - \zeta_h)\|^2 \leq J_1(u_h, u_h). \tag{2.7}$$

This means that we only stabilize the scales that are not already resolved by the finite element space; in this sense this is a minimal stabilizing procedure [3]. Other methods following similar ideas but using hierarchic meshes or projections have been proposed in [11, 9]. The inequality (2.7) was originally proved in [6] but only for constant velocities and uniform meshes. In this paper the essential restrictions that we impose are that β should belong to the space of piecewise linear continuous functions and that the computational mesh is locally quasi-uniform. Moreover, the general framework allows us to circumvent the *inf-sup* condition proved in [6]. The technique of proof introduced in this paper is flexible and may be used in the analysis of more complex problems. The low-order interior penalty term $J_0(u_h, v_h)$ should ensure coercivity and continuity of the bilinear form whereas the term $J_1(u_h, v_h)$ is what makes the method stable in the hyperbolic limit. Clearly for continuous approximations $J_0(u_h, v_h) = 0$, but in this case $A(u_h, v_h)$ is coercive without stabilization (if we discard the boundary conditions for the moment). For the discontinuous Galerkin method, on the other hand, $\beta \cdot \nabla u_h \in V_h$ so that (2.7) holds with $\gamma_1 \equiv 0$.

So it seems that the right dichotomy is between methods using Petrov–Galerkin-type stabilization and methods using interior penalty-type stabilization and not between conforming and nonconforming approximations. In this new framework the guideline is to add only the amount of stabilization needed to control the part of the streamline derivative that cannot be represented by the approximation space. This can be seen in the analysis leading to (2.7): a big space yields a small value of γ_1 and a small space yields a big value of γ_1. The Petrov–Galerkin approach, on the other hand, enforces stability in a much stronger sense when modifying the test space, and the stabilization will be the same regardless of the properties of the approximating space. We will first prove inequality (2.7) in the case where the space of piecewise linear H^1-conforming functions is a subspace of V_h. Let

$$P_c^1 = \{v_h : v_h \in H^1(\Omega); v_h|_K \in P_1(K)\}.$$

The crucial part is to prove that the jumps in the gradient can control some interpolation error of the streamline derivative, $\|h^{1/2}(\beta \cdot \nabla u_h - \pi_h^*(\beta \cdot \nabla u_h))\|$. For simplicity we will consider the case of two-space dimensions; the extension to three-space dimensions is straightforward.

THEOREM 2.2 (stability). *Assume that $P_c^1 \subset V_h$. Let $\beta \in [P_c^1]^2$ and let $u_h \in V_h$. Then there exists an interpolation operator $\pi_h^* : \beta \cdot \nabla V_h \to P_c^1$ and a constant $\tilde{\gamma}_1 \geq c_0 > 0$, depending only on the local mesh geometry, such that*

$$\|h^{1/2}(\beta \cdot \nabla u_h - \pi_h^*(\beta \cdot \nabla u_h))\|^2 \leq J_1(u_h, u_h)$$

with

$$J_1(u_h,u_h) = \sum_K \int_{\partial K\setminus\partial\Omega} \tilde{\gamma}_1 h_{\partial K}^2 [\beta\cdot\nabla u_h]^2 \; ds. \tag{2.8}$$

Proof. First we will define the operator π_h^*. To this end we recall a quasi-interpolant due to Oswald (see [23, 12]). Consider a node S_i and let $\nabla u_h(S_i)|_K$ denote the value of ∇u_h in the element K and in node S_i. Then let

$$\pi_h^*(\beta\cdot\nabla u_h)(S_i) = \frac{1}{n_i}\sum_{K\subset\Omega_i} \beta(S_i)\cdot\nabla u_h(S_i)|_K, \tag{2.9}$$

where n_i denotes the number of triangles in Ω_i. Let φ_j, $j=1,2,3$, be the basis functions on some arbitrary element K' in $\mathcal{T}_h$. Denoting the locally numbered nodes of K' by s_i, with associated macro-elements Ω_i, $i=1,2,3$, there holds $\varphi_j(s_i)=\delta_{ij}$, where δ_{ij} denotes the Kronecker delta. We now consider the projection error on the element K'.

$$\begin{aligned} &\|h_{K'}^{1/2}(\beta\cdot\nabla u_h - \pi_h^*(\beta\cdot\nabla u_h))\|_{K'}^2 \\ &\quad= \int_{K'} h_{K'}\left(\sum_{j=1}^3\left(\beta(s_j)\cdot\nabla u_h(s_j)|_{K'} - \frac{1}{n_j}\sum_{K\subset\Omega_j}\beta(s_j)\cdot\nabla u_h(s_j)|_K\right)\varphi_j\right)^2 \mathrm{d}x \\ &\quad= \int_{K'} h_{K'}\left(\sum_{j=1}^3\frac{1}{n_j}\sum_{K\subset\Omega_j}\Big(\beta(s_j)\cdot\nabla u_h(s_j)|_{K'} - \beta(s_j)\cdot\nabla u_h(s_j)|_K\Big)\varphi_j\right)^2 \mathrm{d}x. \end{aligned} \tag{2.10}$$

Clearly for any K and K', the difference of the streamline derivatives may be rewritten

$$\beta(s_j)\cdot(\nabla u_h(s_j)|_{K'} - \nabla u_h(s_j)|_K) = \sum_{e\in P(K,K')} [\beta(s_j)\cdot\nabla u_h(s_j)]_e,$$

where $P(K,K')$ is the set of edges between the elements connecting K and K' (the shortest path; see Figure 1) and we may write

$$\begin{aligned} &\|h_{K'}^{1/2}(\beta\cdot\nabla u_h - \pi_h^*(\beta\cdot\nabla u_h))\|_{K'}^2 \\ &\qquad= \int_{K'} h_{K'}\left(\sum_{j=1}^3\frac{1}{n_j}\left(\sum_{K\subset\Omega_j}\sum_{e\in P(K,K')}[\beta(s_j)\cdot\nabla u_h(s_j)]_e\right)\varphi_j\right)^2 \mathrm{d}x. \end{aligned} \tag{2.11}$$

Since V_h is a space of piecewise linears and $\beta\in[P_c^1]^2$, the integrand is a quadratic polynomial on K' and we may use the midpoints on the element sides to evaluate the integral. We let x_k denote the midpoints of the edges and write

$$\begin{aligned} &\|h_{K'}^{1/2}(\beta\cdot\nabla u_h - \pi_h^*(\beta\cdot\nabla u_h))\|_{K'}^2 \\ &\quad= \sum_{k=1}^3\frac{\mathrm{meas}(K')}{3}h_{K'}\left(\sum_{j=1}^3\frac{1}{n_j}\left(\sum_{K\subset\Omega_j}\sum_{e\in P(K,K')}[\beta(s_j)\cdot\nabla u_h(s_j)]_e\right)\varphi_j(x_k)\right)^2. \end{aligned}$$

We now consider $k=3$ and assume that this is the midpoint between s_1 and s_2 (see Figure 1). Using the inequality $(\sum_{i=1}^N a_i)^2 \le N\sum_{i=1}^N a_i^2$ and the inequality

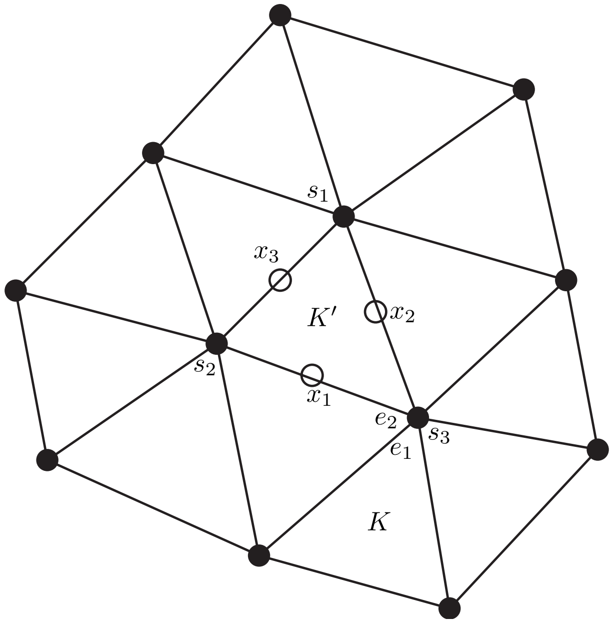


Fig. 1. *Example of element K' with nodes s_1, s_2, and s_3 and the three associated macro-elements Ω_1, Ω_2, and Ω_3. The edges e_1, e_2 separating K' and another triangle K are illustrated, $P(K,K') = \{e_1, e_2\}$, as well as the edge midpoint quadrature points x_1, x_2, and x_3.*

$|\sum_{e\in P(K,K')}[\beta(s_j)\cdot\nabla u_h(s_j)]_e| \le \frac{1}{2}\sum_{e\in\mathcal{E}_j}|[\beta(s_j)\cdot\nabla u_h(s_j)]_e|$, we obtain (recalling that we have $\varphi_1(x_3) = \varphi_2(x_3) = 1/2$ and $\varphi_3(x_3) = 0$ and that in two space dimensions card $\mathcal{E}_j = n_j$)

$$\begin{aligned}
&\frac{\mathrm{meas}(K')}{3} h_{K'} \left(\sum_{j=1}^{3} \frac{1}{n_j}\left(\sum_{K\subset\Omega_j}\sum_{e\in P(K,K')}[\beta(s_j)\cdot\nabla u_h(s_j)]_e\right)\varphi_j(x_3)\right)^2 \\
&\qquad\le \frac{\mathrm{meas}(K')}{3n_j^2} h_{K'} 2\sum_{j=1}^{2} n_j \sum_{K\subset\Omega_j}\frac{n_j}{4}\sum_{e\in\mathcal{E}_j}[\beta(s_j)\cdot\nabla u_h(s_j)]_e^2 \frac{1}{4} \\
&\qquad\qquad\le \frac{\mathrm{meas}(K')}{24} h_{K'} \sum_{j=1}^{2} n_j \sum_{e\in\mathcal{E}_j}[\beta(s_j)\cdot\nabla u_h(s_j)]_e^2.
\end{aligned}$$

It follows from the local quasi-uniformity of the mesh, using three-point quadrature for the edge integral (weights 1/6, 4/6, 1/6), that

$$\frac{\mathrm{meas}(K')}{24} h_{K'} \sum_{j=1}^{2} n_j \sum_{e\in\mathcal{E}_j}[\beta(s_j)\cdot\nabla u_h(s_j)]_e^2 \le \sum_{j=1}^{2}\sum_{e\in\mathcal{E}_j}\int_e \tilde{\gamma}_{1,j} h_e^2 [\beta\cdot\nabla u_h]_e^2 \,\mathrm{d}s,$$

where $\tilde{\gamma}_{1,j} \le \frac{\rho^3 n_j}{4}$. We complete the proof by summing over all Gauss points and all elements leading to a final upper bound on the parameter of $\gamma_1(h) = \tilde{\gamma}_1 h_{\partial K}^2$, with $\tilde{\gamma}_1 \le \frac{\rho^3 n_\rho^2}{4}$. ▯

Remark 2.3. Theorem 2.2 may be extended to finite element spaces using higher-order polynomial approximations. The dependence of the stabilization parameter on the polynomial order is, however, nontrivial and will be a subject for future work.

3. A crucial limit case: Piecewise linear H^1-conforming approximation.

The case of H^1-conforming piecewise linear approximation is important since it is the space for which Petrov–Galerkin-type approximations generally have been used. We will show that this approximation is stable for (2.1) and has (quasi-) optimal convergence properties. We consider H^1-conforming, piecewise-affine finite elements, $V_h = P_c^1$. In (2.4) we take $W_h = V_h$ and $\gamma_0 = 0$, $\gamma_1 = \tilde{\gamma}_1 h_{\partial K}^2$, where $\tilde{\gamma}_1$ scales as $\|\beta\|_\infty^{-1}$ and depends on the local mesh geometry (but not on the mesh size). This results in an interior penalty method originally proposed in [10] and analyzed in [6].

The finite element formulation now takes the following form: Find $u_h \in V_h$ such that

$$A(u_h, v_h) + J_1(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h, \tag{3.1}$$

where $A(u_h, v_h)$ is given by (2.5) with the terms marked (*) left out (being zero) and $J_1(u_h, v_h)$ is given by (2.8).

3.1. Analysis.

We will prove the three preliminary lemmas (Lemmas 3.1, 3.3, and 3.4) giving an approximation result, coercivity of the bilinear form, and Galerkin orthogonality. Using these preliminary results and the stability Theorem 2.2, we then prove the convergence in Theorem 3.5, which is the main result of this section. We first recall a trace inequality that we will use repeatedly:

$$\|v\|_{0,\partial K}^2 \le C\Big(h_K^{-1}\,\|v\|_{0,K}^2 + h_K\,\|v\|_{1,K}^2\Big) \quad \forall v \in H^1(K). \tag{3.2}$$

For a proof of this result, we refer to [26]. The triple norm takes the form

$$\begin{aligned} |||w_h|||^2 = \|\sigma_0^{1/2} w_h\|^2 + \|\varepsilon^{1/2}\nabla w_h\|_h^2 + |(h\varepsilon)^{1/2}\nabla w_h \cdot n|_h^2 \\ + J_1(w_h, w_h) + |\delta(\varepsilon,\beta) w_h|^2, \end{aligned} \tag{3.3}$$

where

$$\delta(\varepsilon,\beta)^2 = \Big(\gamma_{bc}\frac{\varepsilon}{h}\Big) + \frac{1}{2}|\beta \cdot n|.$$

For the continuity of the bilinear form, we will also use the modified norm

$$\begin{aligned} |]w_h[|^2 = \|\sigma_1^{1/2} w_h\|^2 + \|\beta\|_\infty \|h^{-1/2} w_h\|^2 + \|\varepsilon^{1/2}\nabla w_h\|_h^2 \\ + |(h\varepsilon)^{1/2}\nabla w_h \cdot n|_h^2 + J_1(w_h, w_h) + |\delta(\varepsilon,\beta) w_h|^2. \end{aligned}$$

Note that we have used the broken norm for the definition of the triple norms. This is not necessary in the conforming case, but it allows us to use the same triple norm also for the nonconforming approximation.

LEMMA 3.1 (approximation). *Assume that the mesh $\mathcal{T}_h$ is locally quasi-uniform. Let $u \in H^2(\Omega)$ and let $\pi_h u$ denote the standard L_2-projection of u onto V_h; then, if $\tilde{\gamma}_1 \le C\|\beta\|_\infty^{-1}$, we have that*

$$|||\pi_h u - u||| \le Ch(\sigma_0^{1/2} h + \varepsilon^{1/2} + \|\beta\|_\infty^{1/2} h^{1/2})\|u\|_{2,\Omega},$$

where C is independent of σ, ε, β, and h but depends on the mesh geometry.

Proof. It follows from standard interpolation results that $\|\sigma_0^{1/2}(\pi_h u - u)\| \leq \sigma_0^{1/2} h^2 \|u\|_{2,\Omega}$. We then write $\xi_h = \pi_h u - \pi_h^n u$, where π_h^n denotes the nodal interpolant, and note that $\xi_h = \pi_h(u - \pi_h^n u)$. By the H^1-stability of the L^2-projection on locally quasi-uniform meshes [2], we may write

$$\|\nabla \xi_h\| \leq \|\nabla(u - \pi_h^n u)\| \leq Ch\|u\|_{2,\Omega}. \tag{3.4}$$

It immediately follows that

$$\|\varepsilon^{1/2}\nabla(u - \pi_h u)\| \leq C\varepsilon^{1/2} h \|u\|_{2,\Omega},$$

and, using the trace inequality (3.2) and (3.4),

$$|(\varepsilon h)^{1/2}\nabla(\pi_h u - u)|_h^2 \leq \sum_{K \in \mathcal{T}_h} \left(\varepsilon\|\nabla(\pi_h u - u)\|_K^2 + \varepsilon h_K^2 |u|_{2,K}^2\right) \leq C\varepsilon h^2 \|u\|_{2,\Omega}^2.$$

Using once again (3.2) and (3.4) we get in a similar fashion

$$J_1(u - \pi_h u, u - \pi_h u) \leq c\tilde{\gamma}_1\Big(h^{-1}h^2\|\beta\|_\infty^2\|\nabla(u - \pi_h u)\|^2 + h^3\|\beta\|_\infty^2|u|_{2,\Omega}^2\Big)$$
$$\leq \|\beta\|_\infty h^3 \|u\|_{2,\Omega}^2.$$

Finally we note that for the boundary term we have, using (3.2),

$$\langle \pi_h u - u, \pi_h u - u\rangle_{\partial\Omega} \leq h^{-1}\|\pi_h u - u\|^2 + h\|\nabla(\pi_h u - u)\|^2 \leq Ch^3\|u\|_{2,\Omega}^2,$$

which concludes the proof. □

As an immediate consequence of the above result we have the following corollary.

COROLLARY 3.2. *Under the same assumptions as in Lemma* 3.1 *we have that*

$$|]\pi_h u - u[| \leq Ch(\sigma_1^{1/2} h + \varepsilon^{1/2} + \|\beta\|_\infty^{1/2} h^{1/2})\|u\|_{2,\Omega},$$

where C is independent of σ, ε, β, and h but depends on the mesh geometry.

LEMMA 3.3 (coercivity). *The bilinear form $A(u_h, v_h) + J(u_h, v_h)$ is coercive: There exists c, independent of $\varepsilon, \sigma, \beta$, and of h, such that*

$$c|||w_h|||^2 \leq A(w_h, w_h) + J_1(w_h, w_h) \quad \forall w_h \in V_h.$$

Proof. We essentially only need to show that the weakly imposed boundary conditions do not destroy coercivity. We have

$$\begin{aligned} A(w_h, w_h) &= \|\sigma^{1/2} w_h\|^2 + \|\varepsilon^{1/2}\nabla w_h\|^2 + (\beta \cdot \nabla w_h, w_h) \\ &\quad - 2\left\langle \varepsilon \nabla w_h \cdot n, w_h\right\rangle + \left\langle \gamma_{bc}\frac{\varepsilon}{h} w_h, w_h\right\rangle + \left\langle |\beta \cdot n| w_h, w_h\right\rangle_{\partial\Omega_{in}}. \end{aligned} \tag{3.5}$$

Consider the third term and the last term on the right-hand side. Integration by parts yields

$$\begin{aligned} &(\beta \cdot \nabla w_h, w_h) + \left\langle |\beta \cdot n| w_h, w_h\right\rangle_{\partial\Omega_{in}} \\ &\quad = -\frac{1}{2}(\nabla \cdot \beta\; w_h, w_h) + \frac{1}{2}\left\langle \beta \cdot n\, w_h, w_h\right\rangle + \left\langle |\beta \cdot n| w_h, w_h\right\rangle_{\partial\Omega_{in}} \\ &\quad = -\frac{1}{2}(\nabla \cdot \beta\; w_h, w_h) + \frac{1}{2}\left\langle |\beta \cdot n| w_h, w_h\right\rangle. \end{aligned} \tag{3.6}$$

We now consider the second, fourth, and fifth terms of (3.5). The nonsymmetric boundary integral is split using a Cauchy–Schwarz inequality followed by Young's inequality and controlled by the symmetric terms in the following fashion:

$$\begin{aligned}(3.7)\quad &\|\varepsilon^{1/2}\nabla w_h\|^2 - 2\left\langle \varepsilon\nabla w_h\cdot n, w_h\right\rangle + \left\langle \gamma_{bc}\frac{\varepsilon}{h}w_h, w_h\right\rangle\\ &\geq \|\varepsilon^{1/2}\nabla w_h\|^2 - \alpha|(h\varepsilon)^{1/2}\nabla w_h\cdot n|^2 + \left\langle \left(\gamma_{bc}-\frac{1}{\alpha}\right)\frac{\varepsilon}{h}w_h, w_h\right\rangle.\end{aligned}$$

As a consequence of the trace inequality (3.2) we have

$$(3.8)\qquad |(h\varepsilon)^{1/2}\nabla w_h\cdot n|^2 \leq C_t\|\varepsilon^{1/2}\nabla w_h\|^2,$$

and by choosing $\alpha = (2C_t)^{-1}$ and $\gamma_{bc} = 2C_t(\frac{2+2C_t}{1+2C_t})$ we conclude that

$$\begin{aligned}(3.9)\quad &\|\varepsilon^{1/2}\nabla w_h\|^2 - 2\left\langle \varepsilon\nabla w_h\cdot n, w_h\right\rangle + \left\langle \gamma_{bc}\frac{\varepsilon}{h}w_h, w_h\right\rangle\\ &\geq \frac{1}{2(1+C_t)}\left(\|\varepsilon^{1/2}\nabla w_h\|^2 + |(h\varepsilon)^{1/2}\nabla w_h\cdot n|^2 + \left\langle \gamma_{bc}\frac{\varepsilon}{h}w_h, w_h\right\rangle\right).\end{aligned}$$

Combining the results of (3.5), (3.6), (3.9), and the condition (2.2), the lemma follows with the coercivity constant $c = \frac{1}{2(1+C_t)}$. □

Lemma 3.4 (Galerkin orthogonality). *Let u be the solution of* (2.1) *and $u_h \in V_h$ the solution of* (3.1); *then we have that*

$$(3.10)\qquad A(u-u_h, w_h) + J_1(u-u_h, w_h) = 0 \quad \forall w_h \in V_h.$$

Proof. First note that since $u \in H^2(\Omega)$, the trace of ∇u is well-defined, and hence $J_1(u, w_h) = 0$. Since $u = 0$ on $\partial\Omega$, we have that

$$A(u, w_h) = (\sigma u + \beta\cdot\nabla u, w_h) + (\varepsilon\nabla u, \nabla w_h) - \left\langle \varepsilon\nabla u\cdot n, w_h\right\rangle.$$

By an integration by parts in the second term on the right-hand side, we conclude that

$$A(u, w_h) = (\sigma u + \beta\cdot\nabla u - \varepsilon\Delta u, w_h) = (f, w_h),$$

and the lemma is an immediate consequence of (3.1). □

Theorem 3.5. *Let $u \in H^2(\Omega)$ be the solution of* (2.1) *and let $u_h \in V_h$ be the solution of* (3.1); *then, the following a priori error estimate holds:*

$$|||u-u_h||| \leq Ch(\tilde{\sigma}^{1/2}h + \varepsilon^{1/2} + \|\beta\|_\infty^{1/2}h^{1/2})\|u\|_{2,\Omega},$$

where $\tilde{\sigma} = \max(\sigma_0, \sigma_1)$.

Proof. Let $\pi_h u$ be the L^2-projection of u onto V_h. Consider $\xi_h = u_h - \pi_h u$ and $\eta = u - \pi_h u$. By the triangle inequality we have

$$|||u-u_h||| \leq |||\eta||| + |||\xi_h|||$$

and hence by Lemma 3.1 we only need to control $|||\xi_h|||$. We now use the coercivity lemma, Lemma 3.3, followed by Galerkin orthogonality, Lemma 3.4, to obtain

$$(3.11)\qquad c|||\xi_h|||^2 \leq A(\xi_h, \xi_h) + J_1(\xi_h, \xi_h) = A(\eta, \xi_h) + J_1(\eta, \xi_h).$$

Note that after integration by parts in the convective term followed by the application of the Cauchy–Schwarz inequality in $A(\eta,\xi_h)+J_1(\eta,\xi_h)$ we have

$$\begin{aligned}A(\eta,\xi_h)+J_1(\eta,\xi_h) &\leq \|\sigma_1^{1/2}\eta\|\|\sigma_0^{1/2}\xi_h\| + \|\varepsilon^{1/2}\nabla\eta\|\|\varepsilon^{1/2}\nabla\xi_h\| \\ &\quad + J_1(\eta,\eta)^{1/2}J_1(\xi_h,\xi_h)^{1/2} + |(\eta,\beta\cdot\nabla\xi_h)| \\ &\quad + |\delta(\varepsilon,\beta)\xi_h||\delta(\varepsilon,\beta)\eta| \\ &\quad + C|(\varepsilon h)^{1/2}\nabla\eta\cdot n||\delta(\varepsilon,\beta)\xi_h| + C|(\varepsilon h)^{1/2}\nabla\xi_h\cdot n||\delta(\varepsilon,\beta)\eta| \\ &\leq C|]\eta[|\;|||\xi_h||| + |(\eta,\beta\cdot\nabla\xi_h)|.\end{aligned}$$

In the second term in the right-hand side of the last inequality, we now use the orthogonality of the L_2-projection to subtract the Oswald quasi-interpolant from the streamline derivative of ξ_h

$$\begin{aligned}|(\eta,\beta\cdot\nabla\xi_h)| &= |(\eta,\beta\cdot\nabla\xi_h-\pi_h^*(\beta\cdot\nabla\xi_h))| \\ &\leq \|\beta\|_\infty^{1/2}\|h^{-1/2}\eta\|\,\|\beta\|_\infty^{-1/2}\|h^{1/2}(\beta\cdot\nabla\xi_h-\pi_h^*(\beta\cdot\nabla\xi_h))\|.\end{aligned}$$

By Theorem 2.2 we then conclude that

$$\begin{aligned}c|||\xi_h|||^2 &\leq C|]\eta[|\;|||\xi_h||| + \|\beta\|_\infty^{1/2}\|h^{-1/2}\eta\|\,J_1(\xi_h,\xi_h)^{1/2} \\ &\leq C|]\eta[|\;|||\xi_h|||,\end{aligned}$$

and the claim follows by the approximation Corollary 3.2. □

4. An intermediate space: The nonconforming P_1-Crouzeix–Raviart element. Stabilized finite element methods using the Crouzeix–Raviart element have been considered in a number of articles [15, 14, 20], all from the Petrov–Galerkin standpoint. Here we will show how this discretization enters the interior penalty framework using only a penalization on the jump in the gradients, together with a (numerical flux) term involving the jump in the solution assuring coercivity of the convective term. The space of Crouzeix–Raviart finite elements is defined by

$$V_h^{CR} = \left\{ v : v|_K \subset P_1(K), \int_{\partial K\setminus\partial\Omega} [v]\,\mathrm{d}s = 0 \right\}.$$

It is well known that on triangular meshes $P_c^1 \subset V_h^{CR}$, and we will use this fact to simplify our analysis. We propose the following scheme, obtained by taking V_h^{CR} as test and trial space in (2.4): Find $u_h \in V_h^{CR}$ such that

$$A(u_h,v_h)+J_1(u_h,v_h)=(f,v_h) \quad \forall v_h \in V_h^{CR}, \tag{4.1}$$

where the bilinear form is given by

$$\begin{aligned}A(u_h,v_h) &= (\sigma u_h,v_h)+(\varepsilon\nabla u_h,\nabla v_h)_h+(\beta\cdot\nabla u_h,v_h)_h \\ &\quad -\frac{1}{2}\sum_K \langle \beta\cdot n[u_h],\{v_h\}\rangle_{\partial K\setminus\partial\Omega} - \langle\varepsilon\nabla u_h\cdot n,v_h\rangle_h - \langle\varepsilon\nabla v_h\cdot n,u_h\rangle_h \\ &\quad + \left\langle \gamma_{bc}\frac{\varepsilon}{h}u_h,v_h\right\rangle + \langle|\beta\cdot n|u_h,v_h\rangle_{\partial\Omega_{in}}.\end{aligned} \tag{4.2}$$

This time we choose the following form of the interior penalty term:

$$J_1(u_h, v_h) \;=\; \sum_K \Big(\langle \gamma_\tau(h)[\nabla u_h \cdot \tau], [\nabla v_h \cdot \tau]\rangle_{\partial K} \;+\; \langle \gamma_1(h)[\beta \cdot \nabla u_h], [\beta \cdot \nabla v_h]\rangle_{\partial K} \Big). \tag{4.3}$$

Note that we do not add any terms penalizing the jump in the solution; this is because for the Crouzeix–Raviart discretization the jump in the solution is bounded by the jump in the tangential derivative as shown in the following lemma.

Lemma 4.1. *The jump in the solution over element edges satisfies*

$$\int_e \alpha [u_h]^2 \ \mathrm{d}s = \frac{1}{12} \int_e \alpha h_e^2 [\nabla u_h \cdot \tau_e]^2 \ \mathrm{d}s.$$

Proof. Let x_e denote the midpoint on edge e. Clearly $[u_h(x)]^2 = [\nabla u_h \cdot \tau_e]^2 (x - x_e)^2$ for all $x \in e$ and the lemma follows by integration. □

The parameter $\gamma_1(h)$ may be chosen as in the previous section and $\gamma_\tau(h)|_{\partial K} = \tilde{\gamma}_\tau h^2 \|\beta \cdot n\|_{\infty, \partial K}$, $\tilde{\gamma}_\tau = 1/12$. For the analysis we also need the operator $\pi_h^0 : L^2(K) \to P_0(K)$ that denotes the L^2-projection onto the space $P_0(K)$ of piecewise constant functions on the element K. As an immediate consequence of (3.2) we have the estimate

$$\|v - \pi_h^0 v\|_{0,\partial K} \le C h_K^{1/2} \|\nabla v\|_K, \quad v \in H^1(K), \tag{4.4}$$

which we will use to prove that the consistency error is of optimal order. For the convergence proof we will use the same triple norm (3.3) (but with the slightly modified $J_1(u_h, v_h)$ given by (4.3) that has the same approximation properties) and the L_2-projection π_h onto the space P_c^1 so that Lemma 3.1 and Theorem 2.2 hold. We will now proceed to prove equivalents of Lemmas 3.3 and 3.4 for the formulation (4.1) using the Crouzeix–Raviart space. The convergence and, in particular, that the inconsistencies are of the correct order is then shown in Theorem 4.4.

Lemma 4.2. *The bilinear form of formulation* (4.1) *is coercive: There exists a constant c independent of ε, β, σ, and h such that*

$$c |||w_h|||^2 \le A(w_h, w_h) + J_1(w_h, w_h).$$

Proof. The boundary part is handled in the same way as in Lemma 3.3. The part that we need to show does not interfere with coercivity is, in this case, the convective term, but by partial integration we obtain, on using $[w_h^2] = 2[w_h]\{w_h\}$, that

$$\begin{aligned}(\beta \cdot \nabla w_h, w_h)_h = -(\nabla \cdot \beta \ w_h, w_h)_h - (w_h, \beta \cdot \nabla w_h)_h \\ + \sum_K \langle \beta \cdot n \ [w_h], \{w_h\} \rangle_{\partial K \setminus \partial\Omega} + \langle \beta \cdot n \ w_h, w_h \rangle .\end{aligned} \tag{4.5}$$

Using this in $A(w_h, w_h)$ gives

$$\begin{aligned}(\beta \cdot \nabla w_h, w_h)_h - \frac{1}{2} \sum_K \langle \beta \cdot n \ [w_h], \{w_h\} \rangle_{\partial K \setminus \partial\Omega} + \langle |\beta \cdot n| w_h, w_h \rangle_{\partial\Omega_{in}} \\ = -\frac{1}{2} (\nabla \cdot \beta \ w_h, w_h)_h + \frac{1}{2} \langle |\beta \cdot n| w_h, w_h \rangle\end{aligned} \tag{4.6}$$

and coercivity follows by the coercivity condition (2.2). □

Lemma 4.3 (Galerkin orthogonality). *Let u be the solution of* (2.1) *and let u_h be the solution of* (4.1)*; then, we have that*

$$A(u-u_h,w_h)+J_1(u-u_h,w_h)=\frac{1}{2}\sum_K\left\langle\varepsilon(\nabla u\cdot n-\pi_h^0(\nabla u\cdot n)),[w_h]\right\rangle_{\partial K\setminus\partial\Omega},$$

where π_h^0 denotes the projection onto piecewise constants on the element K.

Proof. We note that

$$\begin{aligned}A(u,w_h)&=(\sigma u+\beta\cdot\nabla u,w_h)+(\varepsilon\nabla u,\nabla w_h)_h-\langle\varepsilon\nabla u\cdot n,w_h\rangle\\&=(\sigma u+\beta\cdot\nabla u-\varepsilon\Delta u,w_h)+\frac{1}{2}\sum_{K\in\mathcal{T}_h}\langle\varepsilon\nabla u\cdot n,[w_h]\rangle_{\partial K\setminus\partial\Omega}.\end{aligned}$$

Using now (2.1), the zero mean value property of the jump $[w_h]$ and the fact that $J_1(u,w_h)=0$ for $u\in H^2(\Omega)$ we may write

$$A(u,w_h)=(f,w_h)+\frac{1}{2}\sum_{K\in\mathcal{T}_h}\left\langle\varepsilon(\nabla u\cdot n-\pi_h^0(\nabla u\cdot n)),[w_h]\right\rangle_{\partial K\setminus\partial\Omega},$$

which completes the proof. □

Theorem 4.4. *Let $u\in H^2(\Omega)$ be the solution of* (2.1) *and let $u_h\in V_h^{CR}$ be the solution of* (4.1)*; then, the following a priori error estimate holds:*

$$|||u-u_h|||\le Ch(\tilde\sigma^{1/2}h+\varepsilon^{1/2}+\|\beta\|_\infty^{1/2}h^{1/2})\|u\|_{2,\Omega},$$

where $\tilde\sigma=\max(\sigma_0,\sigma_1)$.

Proof. The proof is similar to the proof of Theorem 3.5. We only need to prove that the residual terms due to the inconsistency have the correct order of convergence. Consider $u-\pi_h u$ with π_h the L_2-projection onto P_c^1. Let $\xi_h=u_h-\pi_h u$ and $\eta=u-\pi_h u$. Following the previous convergence proof we obtain by Lemmas 4.2 and 4.3 and the continuity of the symmetric part that

$$\begin{aligned}|||\xi_h|||^2&\le A(\xi_h,\xi_h)+J_1(\xi_h,\xi_h)\\&=A(\eta,\xi_h)+J_1(\eta,\xi_h)+\frac{1}{2}\sum_K\left\langle\varepsilon(\nabla u\cdot n-\pi_h^0(\nabla u\cdot n)),[\xi_h]\right\rangle_{\partial K\setminus\partial\Omega}\\&\le C]|\eta[|\;|||\xi_h|||+\frac{1}{2}\sum_K\left\langle\varepsilon(\nabla u\cdot n-\pi_h^0(\nabla u\cdot n)),[\xi_h]\right\rangle_{\partial K\setminus\partial\Omega}\\&\qquad+\frac{1}{2}\sum_K\langle\beta\cdot n\eta,[\xi_h]\rangle_{\partial K\setminus\Omega}+|(\eta,\beta\cdot\nabla\xi_h)_h|.\end{aligned}$$

For the element boundary terms we readily obtain

$$\begin{aligned}&\sum_K\left\langle\varepsilon(\nabla u\cdot n-\pi_h^0(\nabla u\cdot n)),[\xi_h]\right\rangle_{\partial K\setminus\partial\Omega}\\&\le\left(\sum_K\|\varepsilon^{1/2}h^{1/2}(\nabla u\cdot n-\pi_h^0(\nabla u\cdot n))\|^2_{\partial K\setminus\partial\Omega}\right)^{1/2}\left(\sum_K\left\langle\varepsilon h^{-1}[\xi_h],[\xi_h]\right\rangle_{\partial K\setminus\partial\Omega}\right)^{1/2}\end{aligned}$$

and

$$\frac{1}{2}\sum_K \langle \beta\cdot n\eta, [\xi_h]\rangle_{\partial K\setminus\partial\Omega} \leq \frac{1}{2}\left(\sum_K \|\beta\|_\infty |\eta|^2_{\partial K}\right)^{1/2}\left(\sum_K \langle |\beta\cdot n|[\xi_h],[\xi_h]\rangle_{\partial K\setminus\partial\Omega}\right)^{1/2}.$$

Using now the projection estimate (4.4), approximation, and Lemma 4.1 we have

$$\left(\sum_K \|\varepsilon^{1/2}h^{1/2}(\nabla u\cdot n - \pi_h^0(\nabla u\cdot n))\|^2_{\partial K\setminus\partial\Omega}\right)^{1/2} \leq \varepsilon^{1/2}h\|u\|_{2,\Omega},$$

$$\frac{1}{2}\left(\sum_K \|\beta\|_\infty |\eta|^2_{\partial K}\right)^{1/2} \leq C\|\beta\|_\infty^{1/2}h^{3/2}\|u\|_{2,\Omega},$$

and

$$\left(\sum_K \big\langle(\varepsilon h^{-1}+|\beta\cdot n|)[\xi_h],[\xi_h]\big\rangle_{\partial K\setminus\partial\Omega}\right)^{1/2} \leq \Big(C\|\varepsilon^{1/2}\nabla\xi_h\|^2_h + J_1(\xi_h,\xi_h)\Big)^{1/2}.$$

Finally the convective term is handled exactly as in the proof of Theorem 3.5 using the orthogonality of the L^2-projection and Theorem 2.2, and we conclude the proof by an application of the approximation Corollary 3.2. □

Remark 4.5. The above analysis of the Crouzeix–Raviart discretization only shows that the method will converge with the same order as the conforming piecewise linear method. However, we expect a richer space to provide a better approximation of the streamline derivative and hence the upper bound on the parameter $\tilde{\gamma}_1$ to be smaller. A more precise analysis following the proof of Theorem 2.2 shows that this is indeed the case. For completeness below we add such a result, which is proven in [7]. What should be observed is that the richer space gives a sharper estimate: In this case the stabilization parameter is independent of the mesh geometry.

LEMMA 4.6. *Let $\beta\in[P_c^1]^d$ and $w_h\in V_h^{CR}$; then*

$$\|h^{1/2}(\beta\cdot\nabla w_h - \pi_h^{CR}(\beta\cdot\nabla w_h))\|^2_h \leq j_\beta(w_h,w_h),$$

where π_h^{CR} denotes the averaging interpolation operator of (2.9) *defined on the Crouzeix–Raviart space and $j_\beta(w_h,w_h)$ is given by*

$$j_\beta(w_h,w_h) = \sum_K \gamma_\beta \int_{\partial K\setminus\partial\Omega} h_K h_{\partial K^\perp}[\beta\cdot\nabla w_h]^2\ ds$$

with $h_{\partial K^\perp}$ denoting the triangle size perpendicular to the side on ∂K and γ_β depends only on the space dimension.

5. Domain decomposition and the relation to discontinuous Galerkin methods. In this section we will show how domain decomposition using Nitsche's method leads to discontinuous Galerkin-type penalty methods in a natural way. For the Poisson problem this method was analyzed in [1]. Below we will briefly sketch how the results of [1] may be extended to the case of convection-diffusion problems using the interior penalty framework. Consider a decomposition of the domain Ω into the disjoint subdomains ω_i, $i=1,\dots,N$, with corresponding triangulations $\mathcal{T}_{h,i}$ such that $\cup_{i=1}^N \mathcal{T}_{h,i} = \cup_{i=1}^N \bar\omega_i = \bar\Omega$. Note that we do not suppose that neighboring meshes are conforming over the intersubdomain boundary. On each triangulation we define a finite element space $V_{h,i}$ associated with the subdomain ω_i.

$$V_{h,i} = \{v_h : v_h \in H^1(\omega_i); v_h|_K \in P_1(K)\}$$

and we let $V_h = \sum_{i=1}^N V_{h,i}$. We now consider problem (2.1) on Ω and, by taking V_h as trial and test space in the formulation (2.4), we propose the finite element method: Find $u_h \in V_h$ such that

$$A(u_h, v_h) + J(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h, \tag{5.1}$$

where

$$\begin{aligned} A(u_h, v_h) = \textstyle\sum_{i=1}^N \left((\sigma u_h, v_h)_{\omega_i} + (\varepsilon \nabla u_h, \nabla v_h)_{\omega_i} + (\beta \cdot \nabla u_h, v_h)_{\omega_i}\right) \\ -\tfrac{1}{2} \textstyle\sum_{i=1}^N \left(\langle \beta \cdot n[u_h], \{v_h\}\rangle_{\partial\omega_i} + \langle \{\varepsilon \nabla u_h \cdot n\}, [v_h]\rangle_{\partial\omega_i} + \langle \{\varepsilon \nabla v_h \cdot n\}, [u_h]\rangle_{\partial\omega_i}\right) \\ - \langle \varepsilon \nabla u_h \cdot n, v_h\rangle - \langle \varepsilon \nabla v_h \cdot n, u_h\rangle \\ + \left\langle \gamma_{bc} \tfrac{\varepsilon}{h} u_h, v_h \right\rangle + \langle |\beta \cdot n| u_h, v_h\rangle_{\partial\Omega_{in}} \end{aligned}$$

and

$$\begin{aligned} J(u_h, v_h) = \sum_{i=1}^N \Bigg(\sum_{K \in \mathcal{T}_{h,i}} \langle \gamma_{1,i}(h) [\beta \cdot \nabla u_h], [\beta \cdot \nabla v_h]\ \rangle_{\partial K \setminus \partial\omega_j} \\ + \left\langle \delta(\varepsilon, \beta)^2 [u_h], [v_h] \right\rangle_{\partial\omega_i \setminus \partial\Omega} \Bigg). \end{aligned} \tag{5.2}$$

Note that the bilinear form A corresponds to a standard Galerkin formulation in each subdomain, supplemented with boundary terms on the inner and outer boundaries that appear naturally in the formulation to assure coercivity or consistency. The interior penalty term $J(u_h, v_h)$ has been decomposed into a term controlling the jumps in the gradient over *interior* edges of each subdomain ω_i and another term controlling the jump of the solution over interior boundaries of neighboring subdomains. The stabilization parameter $\gamma_{1,i}(h) = \tilde\gamma_{1,i} h_K^2$ is now dependent on the mesh geometry of the subdomain triangulation $\mathcal{T}_{h,i}$. We define the triple norm

$$\begin{aligned} |||w_h|||^2 = \sum_{i=1}^N \left(\|\sigma_0^{1/2} w_h\|_{\omega_i}^2 + \|\varepsilon^{1/2} \nabla w_h\|_{\omega_i}^2 + |(h\varepsilon)^{1/2} \nabla w_h \cdot n|_{\partial\omega_i}^2 \right) \\ + J(w_h, w_h) + |\delta(\varepsilon, \beta) w_h|_{\partial\Omega} \end{aligned} \tag{5.3}$$

and obtain the following a priori error estimate.

THEOREM 5.1. *Let $u \in H^2(\Omega)$ be the solution of* (2.1) *and let $u_h \in V_h$ be the solution of* (5.1); *then, the following a priori error estimate holds:*

$$|||u - u_h||| \leq Ch \left(\tilde\sigma^{1/2} h + \varepsilon^{1/2} + \|\beta\|_\infty^{1/2} h^{1/2} \right) \|u\|_{2,\Omega},$$

where $\tilde\sigma = \max(\sigma_0, \sigma_1)$.

Proof. We will not give the details of the proof here, but note that it follows by applying the techniques of Theorem 3.5 in each subdomain ω_i. The internal boundary terms are treated in the same fashion as the outer boundary terms. The added penalty terms on the jump of the solution over internal boundaries ensures the coercivity and continuity of the bilinear form. For a detailed analysis of the method in the case of the Poisson problem, we refer to [1]. □

COROLLARY 5.2. *If the triangulation of each subdomain consists of a single triangle, then the formulation* (5.1) *is equivalent to an interior penalty discontinuous Galerkin method for* (2.1).

Proof. This result is immediate by noting that the interior penalty term on the gradient jumps vanishes since there are no interior edges in the subdomains. ∎

Remark 5.3. The substructuring iterative method for parallel solution naturally associated to (5.1) will be analyzed in a forthcoming work [8].

6. Numerical examples.

In this section we illustrate the numerical performance of the interior penalty method on some academic test cases. We will only consider the case of conforming piecewise linear approximation. In these test cases we have used weakly imposed boundary conditions. For results using strongly imposed boundary conditions, or comparisons between stabilization using the jump in the streamline derivative versus the jump in the whole gradient, see [6]. For results on shock-capturing and discrete maximum principles, see [5], and for results using the Crouzeix–Raviart element, see [7]. First we consider three problems with known exact solution, the first two on structured meshes and the third on the so-called Peterson meshes. The reason we consider Peterson meshes is because we wish to verify that our a priori error estimate is sharp. Finally we will show qualitatively the effect of the weakly imposed boundary conditions. We have applied the finite element method (3.1) to (2.1) using the stabilizing term

$$J_{tot}(u_h, v_h) = \sum_K \int_{\partial K \setminus \partial \Omega} \gamma_1(h)[\nabla u_h] \cdot [\nabla v_h] \mathrm{d}s$$

with $\gamma_1(h) = 0.025\ h_K^2$. The a priori error estimate of Theorem 3.5 holds also for this choice, but some consistent crosswind diffusion is added, giving better control of the gradient. The parameter γ_{bc} is set to unity.

6.1. Convergence tests, smooth solutions.

Consider problem (2.1) with $\beta = (1, 0)$, $\sigma = 1$, and $\varepsilon = 1.\mathrm{E} - 5$ in a square with unit sidelength. To examine the convergence behavior of our method we propose two smooth test cases with known solution. The exact solutions are as follows (see Figure 2):

- test case 1: $u = \exp(-\frac{(x-0.5)^2}{a_w} - \frac{3(y-0.5)^2}{a_w})$, $a_w = 0.2$;
- test case 2: $u = \frac{1}{2}(1 - \tanh(\frac{x-0.5}{a_w}))$, $a_w = 0.05$.

These functions have then been inserted into the equations and the corresponding source terms have been computed. The solution has been computed on a series of structured meshes having 20, 40, 80, 160, and 320 elements, respectively, on each side. A typical mesh is presented in Figure 4. In Tables 6.1 and 6.2 we report the errors in the L^2-norm and the H^1-seminorm as well as the convergence of the jumps in the gradients over element edges given by $J_{tot}(u_h, u_h)$ (with $\gamma_1(h) = h_K^2$ for simplicity). Note that $J_{tot}(u_h, u_h)$ and $J_{tot}(u_h - \pi_h u, u_h - \pi_h u)$ have the same convergence order. The observed order of convergence is denoted by α indicating that the rate is of order $O(h^\alpha)$.

We observe second-order convergence of the error in the L_2-norm and first-order convergence in the H^1-norm. For the stabilization term we obtain the convergence order $h^{3/2}$. In Figure 3 we present a comparison between the numerical results obtained using the continuous interior penalty (CIP) method (for the corresponding theoretical result see Theorem 3.5) and those obtained by solving the problem using a standard SUPG approach. For these simple test cases, the numerical performance of the two methods is nearly identical.

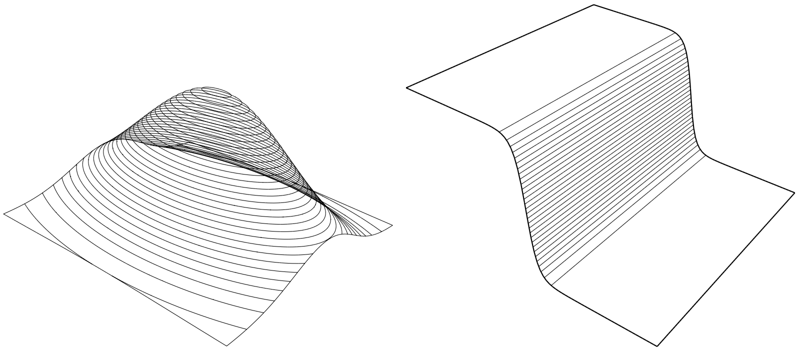


Fig. 2. *The two exact solutions: the Gaussian (left), the hyperbolic tangent (right).*

Table 6.1
Convergence results for test case 1.

N	L_2	α	H^1	α	$J_{tot}(u_h,u_h)^{1/2}$	α
20	0.1618E−02	–	0.1482E+00	–	0.6300E−02	–
40	0.3458E−03	2.22	0.7333E−01	1.02	0.2241E−02	1.49
80	0.8236E−04	2.07	0.3647E−01	1.01	0.7933E−03	1.50
160	0.2045E−04	2.01	0.1817E−01	1.01	0.2806E−03	1.50
320	0.5117E−05	2.00	0.9058E−02	1.00	0.9920E−04	1.50

Table 6.2
Convergence results for test case 2.

N	L_2	α	H^1	α	$J_{tot}(u_h,u_h)^{1/2}$	α
20	0.7382E−02	–	0.6678E+00	–	0.2447E−01	–
40	0.1267E−02	2.54	0.2913E+00	1.20	0.8485E−02	1.53
80	0.2985E−03	2.09	0.1442E+00	1.01	0.3000E−02	1.50
160	0.7370E−04	2.02	0.7198E−01	1.00	0.1061E−02	1.50
320	0.1838E−04	2.00	0.3596E−01	1.00	0.3752E−03	1.50

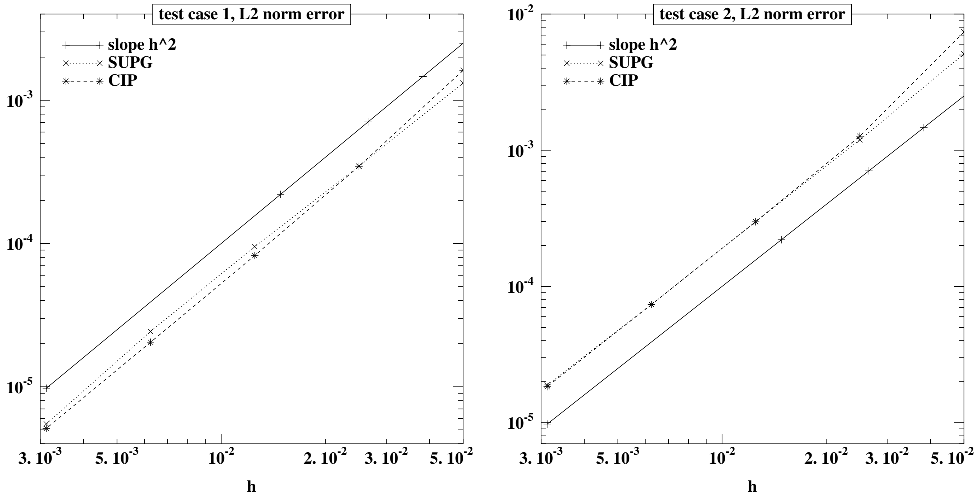


Fig. 3. *Comparisons with the SUPG method: test case* 1 *(left); test case* 2 *(right).*

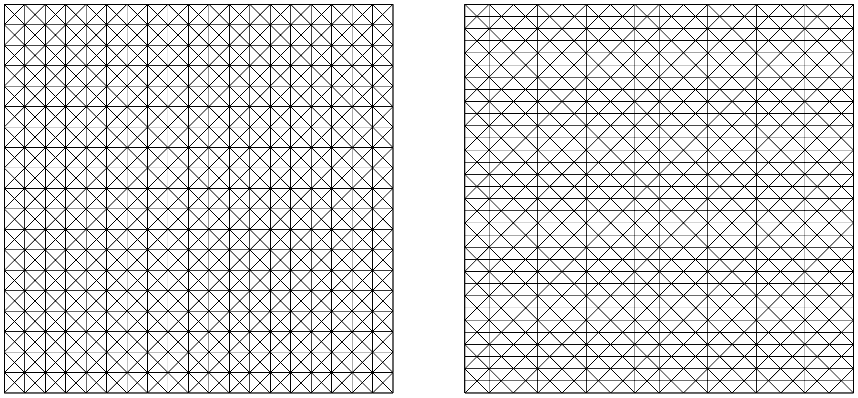


Fig. 4. *Example of the meshes used: structured crisscross mesh (left); Peterson mesh (right).*

6.2. Peterson meshes. Naively, one might hope that the consistently added crosswind diffusion in the jump term $J_1(u_h, v_h)$ obtained by taking the jump of the whole gradient and not only the streamline derivative would result in a numerical scheme for which the error in the L^2-norm never degenerates to $O(h^{3/2})$. However, as the following numerical example shows, this is not the case. We recall the test cases of [27, 24] on the so-called Peterson meshes. In Figure 4 we show an example of a Peterson mesh. In [27] it was shown that the convergence order of the streamline diffusion method on Peterson meshes depends on the number of vertical lines in the mesh. In fact, the streamline diffusion method can be made to converge with any rate $O(h^{3/2}) - O(h^2)$ depending on the distribution of the vertical edges. Here we only consider the worst case where the number of inserted lines is given by $m \approx h^{-3/4}$. Following [27] we chose $\beta = (0, 1)$, $\sigma = 1$, and $\varepsilon = 0$ in (2.1). Moreover we choose $f = x^2$ and the inflow boundary condition $u_{in} = x^2$. The exact solution is given by $u(x, y) = x^2$. In Table 6.3 we report the errors obtained in different norms and the corresponding convergence orders. We note that the convergence rate of the method degenerates to almost $O(h^{3/2})$ in the L^2-norm and to $O(h^{0.88})$ in the H^1-norm. The jump term has a slightly suboptimal convergence rate of $\alpha = 1.4$ but seems to be increasing toward the asymptotic value $\alpha = 1.5$ as the mesh is refined.

Table 6.3
Convergence results on Peterson meshes.

N	m	L_2	α	H^1	α	$J_{tot}(u_h)$	α
8	5	0.7958E−02	–	0.1601E+00	–	0.5065E−02	–
16	8	0.2602E−02	1.61	0.8728E−01	0.88	0.2013E−02	1.33
32	13	0.8178E−03	1.67	0.4726E−01	0.89	0.7613E−03	1.40
64	23	0.2654E−03	1.62	0.2543E−01	0.89	0.2839E−03	1.42
128	38	0.8365E−04	1.67	0.1375E−01	0.89	0.1054E−03	1.43
256	64	0.2712E−04	1.63	0.7465E−02	0.88	0.3878E−04	1.44

6.3. Nonsmooth solutions, weak boundary conditions. In this last numerical example we show the effect of the weakly imposed boundary condition and compare with the case when the boundary condition is imposed strongly. We consider a classical problem with an interior layer and an outflow layer. In this case we choose $\varepsilon = 2.\text{E} - 3$, $\sigma = 0$, $\beta = (-\cos 55°, -\sin 55°)$. The boundary conditions and the computational domain are specified in Figure 5. In Figure 6 we present solutions on three different meshes, having 20, 80, and 320 elements per side, respectively. On the coarsest mesh we show the carpet plot of the mesh and on the finer meshes we only show elevations of the contour plots. Note how the strongly imposed boundary conditions induce significant overshoots in the outflow layer. When the boundary conditions are imposed weakly there are hardly any overshoots, but the approximate solution will satisfy the boundary condition only when the layer is fully resolved. The parameter γ_{bc} can be tuned to impose the satisfaction of the boundary condition on a given scale. However, if the penalty parameter is chosen too large, the oscillations will reappear. The spurious oscillations on the interior layer are suppressed thanks to the added crosswind diffusion but disappear completely only when the mesh is sufficiently fine.

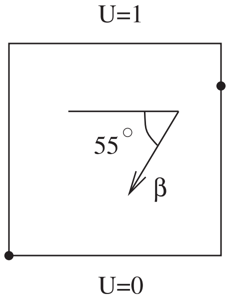


Fig. 5. *Problem data specification, outflow layer test case. At the points the boundary data changes linearly from $U = 1$ to $U = 0$ over an interval of size ε.*

7. Conclusion. We have proposed a new framework for stabilized methods based on interior penalty and conforming or nonconforming approximation. In order to avoid Petrov–Galerkin-type discretizations we added a term giving L^2-control of the jumps in the solution gradient over element boundaries when using spaces V_h that do not satisfy $\beta \cdot \nabla v_h \in V_h \; \forall v_h \in V_h$. We proved that this results in a method that is stable in the hyperbolic limit with optimal order convergence for continuous piecewise linear approximation. The stabilization is symmetric, uniform in the diffusion parameter ε and lumped mass may be used for efficient time stepping. The framework also allows for nonconforming approximations and we proved optimal order a priori error estimates for the first-order Crouzeix–Raviart element using the theory developed for the conforming case. Moreover we discussed domain decomposition using Nitsche's method and the relation to discontinuous Galerkin methods. Finally we considered some numerical examples for the continuous piecewise linear case. We showed that the method has optimal convergence order of $O(h^2)$ in the L^2 norm for smooth test problems on structured meshes, but degenerates to $O(h^{3/2})$ on the so-called Peterson meshes, indicating that our a priori error estimates are sharp. We believe that this form of stabilization offers an attractive compromise between the SUPG method and the discontinuous Galerkin method. Compared to SUPG we are more flexible with respect to time-stepping schemes and mass lumping; however, we pay a price in the size of the system matrix which increases in size by a factor of two in two space dimensions and a factor three in three space dimensions. The implementation also differs since one needs a data structure containing the elements neighboring to a given element in order to compute the gradient jumps. The method

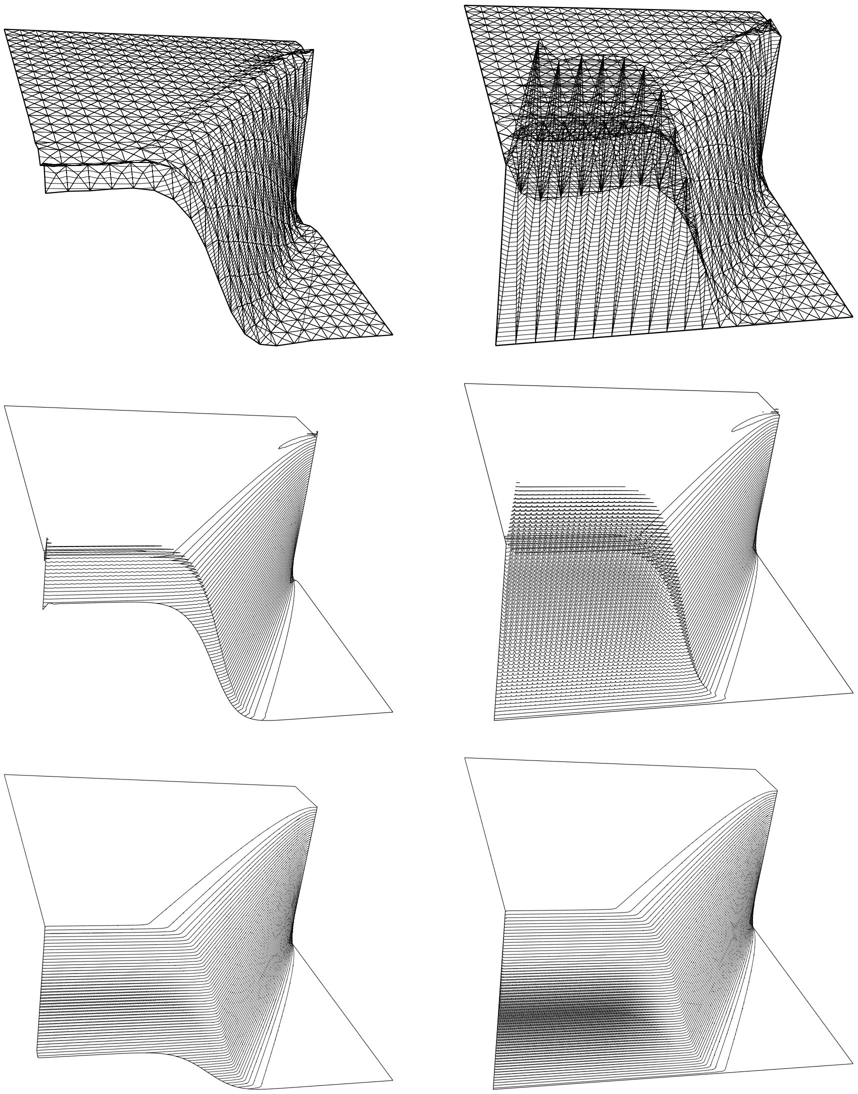


FIG. 6. *Outflow layer test case: weakly imposed boundary conditions (left); strongly imposed boundary conditions; resolutions from top down:* 20×20, 80×80, *and* 320×320 *(right).*

enjoys many of the advantages of the discontinuous Galerkin method. Two important exceptions, however, are the local conservation properties of the discontinuous Galerkin method and the ease by which one may couple finite elements with different polynomial degree. On the other hand, in the continuous interior penalty method we can control the number of degrees of freedom we use by choosing our approximation spaces judiciously. A particularly interesting feature of the method is the way in which it can be combined with discontinuous Galerkin approximations using a Nitsche-type coupling.

Acknowledgment. The author thanks an anonymous referee for the careful reading of the manuscript and the constructive criticism that helped to improve the paper.

REFERENCES

[1] R. Becker, P. Hansbo, and R. Stenberg, *A finite element method for domain decomposition with non-matching grids*, M2AN Math. Model. Numer. Anal., 37 (2003), pp. 209–225.

[2] J. H. Bramble, J. E. Pasciak, and O. Steinbach, *On the stability of the L^2 projection in $H^1(\Omega)$*, Math. Comp., 71 (2002), pp. 147–156.

[3] F. Brezzi and M. Fortin, *A minimal stabilisation procedure for mixed finite element methods*, Numer. Math., 89 (2001), pp. 457–491.

[4] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[5] E. Burman and A. Ern, *Stabilized Galerkin approximation of convection–diffusion–reaction equations: Discrete maximum principle and convergence*, Math. Comp., 74 (2005), pp. 1637–1652.

[6] E. Burman and P. Hansbo, *Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1437–1453.

[7] E. Burman and P. Hansbo, *A stabilized non-conforming finite element method for incompressible flow*, Comput. Methods Appl. Mech. Engrg., (2004), to appear.

[8] E. Burman and P. Zunino, *Iterative Substructuring Methods Based on Nitsche's Matching Conditions for the Solution of Advection–Diffusion Problems Using Interior Penalty Stabilization*, IACS tech. report, 2005.

[9] R. Codina, *Stabilization of incompressibility and convection through orthogonal sub-scales in finite element methods*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 1579–1599.

[10] J. Douglas, Jr. and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences, Second Internat. Symposium, Versailles, 1975, Lecture Notes in Phys. 58, Springer-Verlag, Berlin, 1976, pp. 207–216.

[11] J.-L. Guermond, *Stabilization of Galerkin approximations of transport equations by subgrid modeling*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1293–1316.

[12] R. H. W. Hoppe and B. Wohlmuth, *Element-oriented and edge-oriented local error estimators for nonconforming finite element methods*, RAIRO Modél. Math. Anal. Numér., 30 (1996), pp. 237–263.

[13] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[14] V. John, G. Matthies, F. Schieweck, and L. Tobiska, *A streamline-diffusion method for nonconforming finite element approximations applied to convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 85–97.

[15] V. John, J. M. Maubach, and L. Tobiska, *Nonconforming streamline-diffusion-finite-element-methods for convection-diffusion problems*, Numer. Math., 78 (1997), pp. 165–188.

[16] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[17] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[18] C. Johnson, A. H. Schatz, and L. B. Wahlbin, *Crosswind smear and pointwise errors in streamline diffusion finite element methods*, Math. Comp., 49 (1987), pp. 25–38.

[19] P. Lesaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1974), Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974, pp. 89–123.

[20] G. Matthies and L. Tobiska, *The streamline-diffusion method for conforming and nonconforming finite elements of lowest order applied to convection-diffusion problems*, Computing, 66 (2001), pp. 343–364.

[21] U. Nävert, *A Finite Element Method for Convection-Diffusion Problems*, Ph.D. thesis, Chalmers University of Technology, Göteborg, Sweden, 1982.

[22] J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.

[23] P. Oswald, *On a bpx-Preconditioner for p1 Elements*, Tech. report, FSU Jena, Jena, Germany, 1991.

[24] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[25] E. Süli, P. Houston, and C. Schwab, *hp-finite element methods for hyperbolic problems*, in The Mathematics of Finite Elements and Applications X, MAFELAP 1999 (Uxbridge), Elsevier, Oxford, 2000, pp. 143–162.

[26] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Ser. Comput. Math. 25, Springer-Verlag, Berlin, 1997.

[27] G. Zhou, *How accurate is the streamline diffusion finite element method?* Math. Comp., 66 (1997), pp. 31–44.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2034–2059

A LINEAR ALGEBRAIC ANALYSIS OF DIFFUSION SYNTHETIC ACCELERATION FOR THREE-DIMENSIONAL TRANSPORT EQUATIONS*

S. VAN CRIEKINGEN†

Abstract. The effectiveness of the three-dimensional (3-D) diffusion synthetic acceleration preconditioning procedure is proved in various asymptotic regimes for the discretized, mono-energetic, steady-state, linear Boltzmann transport equation with isotropic scattering. The discretizations consist of a discrete ordinate collocation in angle and a Petrov–Galerkin finite element method in space. Following the path initiated by Faber and Manteuffel, we pursue the 3-D development of Brown by providing a 3-D extension of the slab geometry convergence results of Ashby et al. Our theoretical results confirm the good numerical results of Brown in thin and thick limits and hold for problems with nonconstant coefficients and nonuniform spatial zoning posed on finite domains with an incident flux prescribed at the boundaries.

Key words. three-dimensional Boltzmann transport, diffusion limit, diffusion synthetic acceleration, linear algebra, asymptotic analysis

AMS subject classifications. 65F10, 65R20

DOI. 10.1137/S0036142901398441

1. Introduction.

The integro-differential linear Boltzmann transport equation (BTE) models neutral and charged particle transport. In deterministic approaches, a spatio-angular discretization of the BTE yields a system of linear algebraic equations that is solved iteratively. Diffusion synthetic acceleration (DSA) is known to be effective in speeding up the iterative solution of the discretized BTE. A recent review paper by Adams and Larsen [2] gives an extensive survey of the DSA history together with other acceleration methods using discrete ordinate angular discretizations.

We follow here the path initiated by Faber and Manteuffel [7], who popularized the equivalence between DSA and preconditioning techniques. Their linear algebraic setting then allows benefiting from well-known linear system solution techniques such as conjugate gradients. Ashby et al. [4] extended the viewpoint of [7] by incorporating diamond-difference (or Petrov–Galerkin) spatial discretization and discrete ordinate angular discretization. Their matrix formulation of the resulting discretization enabled a linear algebraic derivation of the preconditioner and initial guess. They also showed the relationship between their work and the four-step method of Larsen [12].

While this was restricted to slab geometry (one-dimensional Cartesian), Brown [5] extended this viewpoint to three-dimensional (3-D) geometry. He derived the preconditioning matrix yielding the DSA procedure, and presented numerical experiments demonstrating the effectiveness of this 3-D DSA preconditioner on some example problems. Specifically, this preconditioner was shown to exhibit very good behavior

*Received by the editors November 16, 2001; accepted for publication (in revised form) October 4, 2004; published electronically December 15, 2005. This work was performed under the auspices of the U.S. Department of Energy at the Lawrence Livermore National Laboratory under contract W-7405-Eng-48. The U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes. Copyright is owned by SIAM to the extent not limited by these rights.

http://www.siam.org/journals/sinum/43-5/39844.html

†Commissariat à l'énergie atomique (CEA-Saclay), DEN/DM2S/SERMA/LENR (bat 470), 91191 Gif-sur-Yvette Cedex, France (serge.van-criekingen@cea.fr).

in the thin and thick limits, that is, for total cross-sections tending to zero or infinity, respectively.

Nevertheless, the 3-D DSA preconditioner derivation of Brown [5] lacks a detailed mathematical analysis of the thin and thick limits similar to that given by Ashby et al. [4] for the slab geometry. In this paper, we intend to remedy this lack so as to confirm the encouraging numerical results in [5]. Opposite to typical Fourier analyses, our results hold for problems posed on finite domains, with nonconstant coefficients and nonuniform spatial zoning.

We consider discretizations consisting of a standard discrete ordinate collocation of the angular variable and a Petrov–Galerkin finite element approximation of the spatial variable. This discretization scheme [5] is equivalent to the well-known discrete ordinate diamond-difference scheme in [13], here in a matrix formulation. As is well known [3, 12], a "consistent" discretization of a limiting diffusion approximation to the BTE leads to effective DSA algorithms. While for slab geometry the consistently-differenced diffusion problem is nonsingular, Brown [5] showed that the "consistently" differenced 3-D diffusion approximation is actually singular, although the DSA preconditioner itself remains nonsingular. This fact makes the 3-D convergence analysis challenging, forcing us to use pseudo-inverses while standard inverses could be used in slab geometry.

Similar to what was done to obtain slab geometry theoretical results [4], we will consider particular values of the cross-sections, namely the thick and thin limits. Within the thick limit, we more precisely investigate the asymptotic diffusion limit, where sources tend to zero and scattering ratios (fraction of "losses" due to scattering processes) tend to unity while total cross-sections tend to infinity. Besides, the effectiveness of the 3-D DSA preconditioner is proved in another thick limit, with scattering ratios bounded away from unity. Finally, we discuss the behavior of the system matrix and DSA preconditioner in the thin regime.

The paper is organized as follows. After some preliminaries in section 2, the 3-D BTE is introduced in section 3. Isotropic scattering is assumed, and Dirichlet boundary conditions are applied. Section 4 gives an integrated presentation of the BTE discretization, and section 5 displays the 3-D DSA preconditioner and related initial guess derived in [5]. Section 6 contains our 3-D convergence analysis results for the different problem regimes described above. Despite their importance, the proofs have been put in an appendix for ease of readability.

2. Preliminaries.

We review some notations and results used throughout the paper. For a linear operator A, $\mathcal{N}(A)$ denotes its null space and $\mathcal{R}(A)$ its range. For matrices $A = (a_{ij}) \in \mathbf{R}^{m\times n}$ and $B \in \mathbf{R}^{k\times l}$, the Kronecker (or tensor, or direct) product of A and B is the $mk \times nl$ matrix defined by

$$A \otimes B \equiv \begin{pmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{pmatrix}.$$

We recall the following Kronecker product properties [9]:

- If A and B are nonsingular, then $A \otimes B$ is nonsingular with $(A\otimes B)^{-1} = A^{-1} \otimes B^{-1}$,
- $(A \otimes B)^T = A^T \otimes B^T$,
- $(A \otimes B) \cdot (C \otimes D) = AC \otimes BD$ as long as both sides of the equation make sense,

- $(A+B)\otimes C = A\otimes C + B\otimes C$,
- $A\otimes(B+C) = A\otimes B + A\otimes C$, and
- for any two vectors u and v, $\|u\otimes v\|_2 = \|u\|_2 \cdot \|v\|_2$, where $\|\cdot\|_2$ denotes the usual Euclidean norm of a vector.

Finally, for diagonal matrices A and B, $A < (\leq)\ B$ means $a_{ii} < (\leq)\ b_{ii}$ for all i.

3. Problem definition. We begin with the mono-energetic, steady-state, linear BTE in a 3-D box geometry with isotropic scattering [13]. The spatial domain is the box $\mathcal{D} \equiv \{r = (x,y,z) | a_x \leq x \leq b_x, a_y \leq y \leq b_y, \text{ and } a_z \leq z \leq b_z\}$, the direction variable is $\Omega \in \mathcal{S}^2$ (the unit sphere in $\mathbf{R}^3$), and the equation in the flux ψ is given by

$$\Omega \cdot \nabla\psi(r,\Omega) + \sigma(r)\psi = \sigma_s(r) \int_{\mathcal{S}^2} \psi(r,\Omega')d\Omega' + q(r,\Omega), \tag{3.1}$$

where $\nabla\psi \equiv (\partial\psi/\partial x, \partial\psi/\partial y, \partial\psi/\partial z)$. The functions $\sigma(r)$, $\sigma_s(r)$ and $q(r,\Omega)$ are assumed known. The flux $\psi(r,\Omega)$ is expanded in surface harmonics $Y_n^m(\Omega)$ according to

$$\psi(r,\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} \phi_n^m(r) Y_n^m(\Omega),$$

where

$$\phi_n^m(r) \equiv \int_{\mathcal{S}^2} \psi(r,\Omega) Y_n^m(\Omega) d\Omega$$

is the (n,m)th moment of ψ. Similarly, the source q is expanded as

$$q(r,\Omega) = \sum_{n=0}^{\infty} \sum_{m=-n}^{n} q_n^m(r) Y_n^m(\Omega),$$

where

$$q_n^m(r) \equiv \int_{\mathcal{S}^2} q(r,\Omega) Y_n^m(\Omega) d\Omega.$$

For ease of exposition in what follows, we have elected to use real-valued surface harmonics, all scaled to have unit norm in $L^2(\mathcal{S}^2)$. See [5] for a detailed definition of these harmonics. To relate to notations in [5], we also define $\sigma_{s,0} = 4\pi\sigma_s$.

Given ψ in the above form, one is able to rewrite the scattering term in the form

$$\sigma_s(r) \int_{\mathcal{S}^2} \psi(r,\Omega')d\Omega' = \sigma_{s,0}(r)\phi_0^0(r) Y_0^0$$

with $Y_0^0 = \frac{1}{\sqrt{4\pi}}$. The total cross-section σ is given by

$$\sigma(r) = \sigma_a(r) + \sigma_{s,0}(r),$$

where σ_a is the absorption cross-section.

Boundary conditions must also be specified so as to make (3.1) well-posed. We consider Dirichlet boundary conditions in which the incident flux $g(r,\Omega)$ is specified on a face. That is,

$$\psi(r,\Omega) = g(r,\Omega) \quad \text{for all} \quad r \in \partial\mathcal{D} \quad \text{and} \quad \Omega \in \mathcal{S}^2 \quad \text{with } \vec{n}(r)\cdot\Omega < 0, \tag{3.2}$$

where $\vec{n}(r)$ is the outward pointing unit normal at $r \in \partial\mathcal{D}$. Another common choice of boundary conditions is a reflecting condition on a face. Note that a problem with a reflecting condition can be transformed into one with Dirichlet conditions by reflecting the problem data about the reflecting boundary.

4. Discretization of the 3-D problem.

We give here an integrated presentation of the discretization scheme developed in [5] for (3.1)–(3.2). This scheme is equivalent to the discrete ordinate diamond-difference scheme in [13]. On the other hand, the Petrov–Galerkin spatial discretization used here can also be related to the finite volume element method analyzed by Cai, Mandel, and McCormick [6].

4.1. Spatio-angular discretization of the BTE.

The angular variable is discretized using a discrete ordinate collocation method (see [13] for a complete overview of such methods). It consists simply of evaluating the BTE (3.1) in discrete angular directions $\Omega_\ell \in \mathcal{S}^2$ $(\ell = 1, \dots, L)$, each of them characterized by direction cosines $(\mu_\ell, \eta_\ell, \xi_\ell)$. The same set of (nonzero) direction cosines is used with respect to each of the three coordinate axes, i.e., $\{\mu_\ell\} = \{\eta_\ell\} = \{\xi_\ell\}$. Also, the direction cosines are assumed to be symmetrically placed (with respect to the origin) along each axis. Integrals over the unit sphere are approximated by a quadrature rule

$$\int_{\mathcal{S}^2} \psi(\Omega) d\Omega \approx \sum_{\ell=1}^{L} w_\ell \psi(\Omega_\ell). \tag{4.1}$$

As in [5], we consider either weights invariant under 90°-rotations about any coordinate axis, or weights that are all equal. In either case, we require all the weights to be positive and such that

$$\sum_{\ell=1}^{L} w_\ell \xi_\ell^{2n} = \frac{4\pi}{2n+1} \tag{4.2}$$

for $n = 0$ and 1. It follows from the symmetrical placement of the direction cosines that

$$\sum_{\ell=1}^{L} w_\ell \mu_\ell = 0, \quad \sum_{\ell=1}^{L} w_\ell \eta_\ell = 0, \quad \text{and} \quad \sum_{\ell=1}^{L} w_\ell \xi_\ell = 0. \tag{4.3}$$

For the spatial variable, we first discretize the box-shaped domain $\mathcal{D}$ into zones delimited by the coordinate lines $x = x_i$ $(i = 0, \dots, M)$, $y = y_j$ $(j = 0, \dots, J)$, and $z = z_k$ $(k = 0, \dots, K)$, and define $r_{ijk} = (x_i, y_j, z_k)$. Next, we define $\Delta x_i = x_i - x_{i-1}$ for $i = 1, \dots, M$, and similarly Δy_j and Δz_k. We also define $\Delta r_{ijk} \equiv \Delta x_i \Delta y_j \Delta z_k$. The $\{r_{ijk}\}$ are referred to as *nodes*, and function values at these points are called *nodal values*. Assume that σ, σ_s and q have constant values on each *zone*

$$\mathcal{Z}_{ijk} \equiv \{r | x_{i-1} < x < x_i, y_{j-1} < y < y_j, z_{k-1} < z < z_k\},$$

denoted by σ_{ijk}, $\sigma_{s,ijk}$, and q_{ijk}, respectively. In what follows, we assume $\sigma_{ijk} \neq 0$ for all i, j, k, even though we examine the thin limit (where total cross-sections tend to zero as described below) in section 6.2. Function values that are constant on zones will be referred to as *zone-centered* values. We use $\psi_{ijk,\ell}$ to denote the approximation to $\psi(r_{ijk}, \Omega_\ell)$, the true solution at r_{ijk} in the direction Ω_ℓ.

For the spatial expansions, we use the continuous piecewise-trilinear elements

$$P_{ijk}(r) \equiv p_i(x) p_j(y) p_k(z),$$

where $p_i(x)$, $p_j(y)$, and $p_k(z)$ are the standard continuous piecewise-linear one-dimensional basis functions. We then approximate $\psi(r, \Omega_\ell)$ by the piecewise-trilinear function ψ_a given by

$$\psi(r, \Omega_\ell) \approx \psi_a(r, \Omega_\ell) \equiv \sum_{i,j,k} \psi_{ijk,\ell} P_{ijk}(r). \tag{4.4}$$

Note that $\psi_a(r_{ijk}, \Omega_\ell) = \psi_{ijk,\ell}$.

The Petrov–Galerkin method consists of substituting ψ_a for ψ in the BTE (3.1) and then averaging over zone $\mathcal{Z}_{ijk}$. For each direction Ω_ℓ, this procedure yields a set of MJK zonal equations in the $(M+1)(J+1)(K+1)$ unknowns $\psi_{ijk,\ell}$. The boundary conditions provide the extra equations necessary to make the problem well-posed. In what follows, *nodes* and *zones* used as sub- or superscripts will respectively denote the number of nodes $(M+1)(J+1)(K+1)$ and zones MJK.

To obtain a compact notation, we define the vector of nodal values

$$\Psi \equiv \begin{pmatrix} \Psi_1 \\ \vdots \\ \Psi_L \end{pmatrix} \quad \text{with} \quad \Psi_\ell \equiv \begin{pmatrix} \Psi_{0,\ell} \\ \vdots \\ \Psi_{K,\ell} \end{pmatrix} \in \mathbf{R}^{nodes},$$

$$\Psi_{k,\ell} \equiv \begin{pmatrix} \Psi_{0k,\ell} \\ \vdots \\ \Psi_{Jk,\ell} \end{pmatrix} \in \mathbf{R}^{(M+1)(J+1)}, \quad \text{and} \quad \Psi_{jk,\ell} \equiv \begin{pmatrix} \psi_{0jk,\ell} \\ \vdots \\ \psi_{Mjk,\ell} \end{pmatrix} \in \mathbf{R}^{M+1}.$$

Next, we define

$$\begin{aligned} \Delta x &\equiv \operatorname{diag}(\Delta x_1, \dots, \Delta x_M), \\ \Delta y &\equiv \operatorname{diag}(\Delta y_1, \dots, \Delta y_J), \\ \Delta z &\equiv \operatorname{diag}(\Delta z_1, \dots, \Delta z_K), \\ \Delta r &= \Delta z \otimes \Delta y \otimes \Delta x, \end{aligned}$$

as well as the matrices

$$D_M \equiv \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbf{R}^{M\times(M+1)}, \tag{4.5}$$

$$S_M \equiv \frac{1}{2}\begin{pmatrix} 1 & 1 & & \\ & \ddots & \ddots & \\ & & 1 & 1 \end{pmatrix} \in \mathbf{R}^{M\times(M+1)}, \tag{4.6}$$

and similarly the matrices D_J, S_J, D_K, and S_K for the other two directions. Furthermore, we introduce

$$\Sigma \equiv \operatorname{diag}(\sigma_{111}, \dots, \sigma_{zones}) \in \mathbf{R}^{zones},$$

and

$$\begin{aligned} S &\equiv S_K \otimes S_J \otimes S_M, \\ C_x &\equiv S_K \otimes S_J \otimes \Delta x^{-1} D_M, \\ C_y &\equiv S_K \otimes \Delta y^{-1} D_J \otimes S_M, \\ C_z &\equiv \Delta z^{-1} D_K \otimes S_J \otimes S_M. \end{aligned}$$

Note that S represents an averaging matrix taking nodal vectors into zone-centered vectors, while C_x, C_y, and C_z represent the discretized versions of the differentiation operators $\partial/\partial x$, $\partial/\partial y$, and $\partial/\partial z$, respectively. Then, for each ℓ,

$$C_\ell \equiv \mu_\ell C_x + \eta_\ell C_y + \xi_\ell C_z$$

is the discretized version of the $\Omega_\ell \cdot \nabla$ operator on the left-hand side of (3.1).

To represent the source term, we define the zone-centered vector

$$Q \equiv \begin{pmatrix} Q_1 \\ \vdots \\ Q_L \end{pmatrix} \quad \text{with} \quad Q_\ell \equiv \begin{pmatrix} Q_{1,\ell} \\ \vdots \\ Q_{K,\ell} \end{pmatrix} \in \mathbf{R}^{zones},$$

$$Q_{k,\ell} \equiv \begin{pmatrix} Q_{1k,\ell} \\ \vdots \\ Q_{Jk,\ell} \end{pmatrix} \in \mathbf{R}^{MJ}, \quad \text{and} \quad Q_{jk,\ell} \equiv \begin{pmatrix} q_{1jk,\ell} \\ \vdots \\ q_{Mjk,\ell} \end{pmatrix} \in \mathbf{R}^{M},$$

where $q_{ijk,\ell} \equiv q(r_{ijk}, \Omega_\ell)$ is given.

With I_{zones} the $MJK \times MJK$ identity matrix, we then define the $MJK \times LMJK$ matrices

$$L_{n,m} \equiv (w_1 Y_n^m(\Omega_1) I_{zones} \,|\, w_2 Y_n^m(\Omega_2) I_{zones} \,|\, \cdots \,|\, w_L Y_n^m(\Omega_L) I_{zones}) \tag{4.7}$$

and the $LMJK \times MJK$ matrices

$$L_{n,m}^{+} \equiv \begin{pmatrix} Y_n^m(\Omega_1) I_{zones} \\ \vdots \\ Y_n^m(\Omega_L) I_{zones} \end{pmatrix}. \tag{4.8}$$

Note that

$$L_{n,m} L_{n',m'}^{+} = \sum_{\ell=1}^{L} w_\ell Y_n^m(\Omega_\ell) Y_{n'}^{m'}(\Omega_\ell) I_{zones},$$

and it was proven in [5] that, for $n, n' = 0, 1$ and $|m| \le n, |m'| \le n'$, we have

$$L_{n,m} L_{n',m'}^{+} = \delta_{n,n'} \delta_{m,m'} I_{zones}. \tag{4.9}$$

We also define the grouped matrices L_n and L_n^+ as

$$L_n = \begin{pmatrix} L_{n,-n} \\ \vdots \\ L_{n,n} \end{pmatrix}$$

and

$$L_n^{+} = (L_{n,-n}^{+}, \dots, L_{n,n}^{+}).$$

Furthermore we need

$$\begin{aligned} &\Gamma_0 = \operatorname{diag}(\gamma_{0,111}, \dots, \gamma_{0,MJK}) \text{ with } \gamma_0 = \sigma_{s,0}/\sigma \le 1, \\ &\bar{\Sigma} \equiv I_L \otimes \Sigma, \text{ and} \\ &\bar{S} \equiv I_L \otimes S \end{aligned} \tag{4.10}$$

with I_L the $L \times L$ identity matrix. Then, with $D_\ell \equiv (\Sigma^{-1} C_\ell + S) \in \mathbf{R}^{zones \times nodes}$ and

$$D \equiv \operatorname{diag}(D_1, \dots, D_L),$$

the discretized version of the BTE equation (3.1) reads

$$D\Psi = L_0^{+} \Gamma_0 L_0 \bar{S} \Psi + \bar{\Sigma}^{-1} Q. \tag{4.11}$$

4.2. Discretized Dirichlet boundary conditions. For the boundary conditions (3.2), when $x = x_0$, the normal $\vec{n}(r_{0jk}) = (-1, 0, 0)$ for all j, k. Hence, $\vec{n}(r_{0jk}) \cdot \Omega_\ell = -\mu_\ell$, and for $\mu_\ell > 0$ we have

$$\psi_{0jk,\ell} = g_{0jk,\ell} (\equiv g(r_{0jk}, \Omega_\ell)). \tag{4.12}$$

The other cases are treated similarly. We let G be the nodal vector made out of the $g_{ijk,\ell}$ and built as the Ψ vector.

To isolate the boundary values, first note that for a direction vector Ω_ℓ with all its components positive, ψ satisfies a Dirichlet condition for all $r = r_{0jk}$, r_{i0k}, or r_{ij0}, i.e., for an r on any one of the three faces $x = x_0$, $y = y_0$, or $z = z_0$. In this case, after discretization, the boundary conditions (4.12) and their relevant $y-$ and $z-$ counterparts can be written in tensor notation as $E_{000}^T(\Psi - G) = 0$, where

$$E_{000}^T \equiv \begin{pmatrix} e_{0K}^T \otimes I_{J+1} \otimes I_{M+1} \\ (0, I_K) \otimes e_{0J}^T \otimes I_{M+1} \\ (0, I_K) \otimes (0, I_J) \otimes e_{0M}^T \end{pmatrix}$$

with $e_{0M} = (1, 0, \ldots, 0)^T \in \mathbf{R}^{M+1}$, and similarly for the vectors e_{0J} and e_{0K}. More generally, e_{IJ} designates a column vector of size $J + 1$, filled with zeros everywhere except for a 1 on the $(I + 1)$th row. There are different E matrices for the other possible cases. For example, for a direction Ω_ℓ with $\mu_\ell > 0$, $\eta_\ell > 0$ and $\xi_\ell < 0$, ψ satisfies a Dirichlet condition for r on any one of the three faces $x = x_0$, $y = y_0$, or $z = z_K$. We then have $E_{00K}(\Psi - G) = 0$, where

$$E_{00K}^T \equiv \begin{pmatrix} e_{KK}^T \otimes I_{J+1} \otimes I_{M+1} \\ (I_K, 0) \otimes e_{0J}^T \otimes I_{M+1} \\ (I_K, 0) \otimes (0, I_J) \otimes e_{0M}^T \end{pmatrix}.$$

In all, there are eight different E_{ijk} matrices, with $i = 0$ or M, $j = 0$ or J, and $k = 0$ or K. Defining

$$B_\ell = \begin{cases} E_{000}^T\,, \text{ if } \mu_l > 0, \eta_l > 0, \text{ and } \xi_l > 0, \\ E_{M00}^T\,, \text{ if } \mu_l < 0, \eta_l > 0, \text{ and } \xi_l > 0, \\ E_{0J0}^T\,, \text{ if } \mu_l > 0, \eta_l < 0, \text{ and } \xi_l > 0, \\ E_{00K}^T\,, \text{ if } \mu_l > 0, \eta_l > 0, \text{ and } \xi_l < 0, \\ E_{MJ0}^T\,, \text{ if } \mu_l < 0, \eta_l < 0, \text{ and } \xi_l > 0, \\ E_{M0K}^T\,, \text{ if } \mu_l < 0, \eta_l > 0, \text{ and } \xi_l < 0, \\ E_{0JK}^T\,, \text{ if } \mu_l > 0, \eta_l < 0, \text{ and } \xi_l < 0, \\ E_{MJK}^T\,, \text{ if } \mu_l < 0, \eta_l < 0, \text{ and } \xi_l < 0, \end{cases} \tag{4.13}$$

the discretized Dirichlet boundary conditions for each Ω_ℓ thus read

$$B_\ell \Psi_\ell = B_\ell G_\ell. \tag{4.14}$$

We have $B_l \in \mathbf{R}^{(nodes-zones)\times nodes}$.

4.3. The complete discretized system. We now combine (4.11) and the boundary conditions (4.14) into a single matrix notation. In this view we define

$$H \equiv \operatorname{diag}(H_1, \dots, H_L),$$

where

$$H_\ell \equiv \begin{pmatrix} D_\ell \\ B_\ell \end{pmatrix}.$$

Note that H_ℓ operates on nodal vectors. Let also

$$Z \equiv I_L \otimes Z_0, \quad \text{where } Z_0 \equiv \begin{pmatrix} I_{zones} \\ 0 \end{pmatrix} \in \mathbf{R}^{nodes \times zones}, \tag{4.15}$$

such that the matrix Z injects zone-centered vectors into the nodal vector space. For the boundary terms, define the block diagonal matrix $\bar{B}$ by

$$\bar{B} \equiv \operatorname{diag}(B_1^0, \dots, B_L^0), \quad \text{where } B_\ell^0 \equiv \begin{pmatrix} 0 \\ B_\ell \end{pmatrix}$$

for all ℓ, with each $B_\ell^0 \in \mathbf{R}^{nodes \times nodes}$, as well as $\beta \equiv \bar{B}G$.

The complete discretization of (3.1)–(3.2) can now be written in the compact form

$$H\Psi = ZL_0^+ \Gamma_0 L_0 \bar{S}\Psi + Z\bar{\Sigma}^{-1}Q + \beta. \tag{4.16}$$

Equivalently with $T = H - ZL_0^+\Gamma_0 L_0 \bar{S}$, we have

$$T\Psi = Z\bar{\Sigma}^{-1}Q + \beta. \tag{4.17}$$

It is shown in [5] that H is invertible. So, we can multiply (4.16) by $L_0\bar{S}H^{-1}$, to yield

$$L_0\bar{S}\Psi = L_0\bar{S}H^{-1}ZL_0^+\Gamma_0 L_0\bar{S}\Psi + L_0\bar{S}H^{-1}(Z\bar{\Sigma}^{-1}Q + \beta).$$

Defining $\Phi_0 \equiv L_0\bar{S}\Psi$, $R_0 \equiv L_0\bar{S}H^{-1}(Z\bar{\Sigma}^{-1}Q + \beta)$, and $K_{0,0} \equiv L_0\bar{S}H^{-1}ZL_0^+$, we obtain

$$\Phi_0 = K_{0,0}\Gamma_0\Phi_0 + R_0$$

or

$$A_0\Phi_0 = R_0 \tag{4.18}$$

with

$$A_0 \equiv I_{zones} - K_{0,0}\Gamma_0. \tag{4.19}$$

It is also shown in [5] that A_0 is nonsingular. Once Φ_0 is obtained by solving (4.18), Ψ is recovered by solving the equation

$$H\Psi = ZL_0^+\Gamma_0\Phi_0 + Z\bar{\Sigma}^{-1}Q + \beta.$$

5. Preconditioners and the initial guess. Due to the form of the matrix A_0 in equation (4.18), an iterative solution method is normally used, and a simple Richardson or source iteration is the most common choice. The selection of a preconditioner and an initial guess to speed up the iterative solving of (4.18) has been discussed in [5]. The preconditioner was derived through a "consistent" discretization of a limiting diffusion approximation to the BTE, and its use in Richardson's method is called DSA. We state the results of this discussion here, referring to [5] for the proofs, and restricting our attention to isotropic scattering.

5.1. Discrete DSA results. We first note that the definitions of the discrete moment matrices $L_{n,m}$ and $L^+_{n,m}$ can be easily modified to operate directly on the nodal vector Ψ. This is accomplished by replacing the I_{zones} identity matrix by the I_{nodes} identity in equations (4.7) and (4.8). We denote these nodal moment operators by $\tilde{L}_{n,m}$ and $\tilde{L}^+_{n,m}$, respectively, and analogously the operators $\tilde{L}_n$ and $\tilde{L}^+_n$. We can then define the nodal moments $\tilde{\Phi}_n$ of Ψ by

$$\tilde{\Phi}_n \equiv \tilde{L}_n \Psi.$$

To take discrete moments of the boundary conditions, we need to introduce the matrix operator $L_{\vec{n}}$ defined by

$$L_{\vec{n}} \equiv (v_1 w_1 Y_0^0 S_{\vec{n}} \,|\, v_2 w_2 Y_0^0 S_{\vec{n}} \,|\, \cdots \,|\, v_L w_L Y_0^0 S_{\vec{n}}),$$

with $v_\ell \equiv \vec{n} \cdot \Omega_\ell$ if $\vec{n} \cdot \Omega_\ell < 0$ and 0 otherwise and with $S_{\vec{n}}$ defined so that $S_{\vec{n}}\Psi$ gives a vector containing only two-dimensional zone-averaged entries of Ψ on the face of the box corresponding to $\vec{n}$. We have for

$$\vec{n} = \vec{n}_{x_0} \equiv (-1,0,0) : S_{\vec{n}_{x_0}} \equiv S_K \otimes S_J \otimes e^T_{0M}, \tag{5.1}$$

$$\vec{n} = \vec{n}_{x_M} \equiv (1,0,0) : S_{\vec{n}_{x_M}} \equiv S_K \otimes S_J \otimes e^T_{MM}, \tag{5.2}$$

and similarly for the other two directions.

The DSA preconditioner is obtained using a P_1 approximation to (4.17) that consists in approximating Ψ by $\tilde{L}^+_0\tilde{\Phi}_0 + \tilde{L}^+_1\tilde{\Phi}_1$. The following discrete P_1 system was derived in [5]:

$$T_1 \begin{pmatrix} \tilde{\Phi}_0 \\ \tilde{\Phi}_1 \end{pmatrix} \equiv \begin{pmatrix} L_{x_0,0} & L_{x_0,1} \\ L_{y_0,0} & L_{y_0,1} \\ L_{z_0,0} & L_{z_0,1} \\ T_{00} & T_{01} \\ T_{10} & T_{11} \\ L_{x_M,0} & L_{x_M,1} \\ L_{y_J,0} & L_{y_J,1} \\ L_{z_K,0} & L_{z_K,1} \end{pmatrix} \begin{pmatrix} \tilde{\Phi}_0 \\ \tilde{\Phi}_1 \end{pmatrix} = \begin{pmatrix} G_{x_0} \\ G_{y_0} \\ G_{z_0} \\ (\Sigma^0)^{-1}Q_0 \\ (\Sigma^1)^{-1}Q_1 \\ G_{x_M} \\ G_{y_J} \\ G_{z_K} \end{pmatrix}, \tag{5.3}$$

where

$$\begin{aligned} T_{00} &= (I_{zones} - \Gamma_0)S^0 \\ T_{01} &= \frac{1}{\sqrt{3}}[-\Sigma^{-1}C_y, \Sigma^{-1}C_z, -\Sigma^{-1}C_x] \\ T_{10} &= \frac{1}{\sqrt{3}}[-C_y^T\Sigma^{-1}, C_z^T\Sigma^{-1}, -C_x^T\Sigma^{-1}]^T \\ T_{11} &= S^1, \end{aligned}$$

with $S^n \equiv I_{2n+1} \otimes S$, $\Sigma^n \equiv I_{2n+1} \otimes \Sigma$, and $Q_n = L_n Q$ $(n = 0, 1)$. Also, $L_{x_0,0} \equiv L_{\vec{n}} \tilde{L}_0^+$, $L_{x_0,1} \equiv L_{\vec{n}} \tilde{L}_1^+$, and $G_{x_0} \equiv L_{\vec{n}} G$ for $\vec{n} = \vec{n}_{x_0} \equiv (-1, 0, 0)$, and similarly for the other directions. The T_1 matrix defined in (5.3) is the P_1 approximation to T. Its first and last lines correspond to the boundary conditions.

The system (5.3) can be reduced [5] to the solution of a diffusion problem involving only $\tilde{\Phi}_0$, namely

$$\begin{aligned} D_{co}\tilde{\Phi}_0 = {} & S^T \Delta r^0 Q_0 - T_{10}^T \Delta r^1 Q_1 \\ & + 2\frac{1}{\sqrt{4\pi}} \sum_{\ell=1}^{L} w_\ell S_K^T \Delta z S_K \otimes S_J^T \Delta y S_J \otimes (v_{x_0,\ell} e_{0M} e_{0M}^T + v_{x_M,\ell} e_{MM} e_{MM}^T) \cdot G_\ell \\ & + 2\frac{1}{\sqrt{4\pi}} \sum_{\ell=1}^{L} w_\ell S_K^T \Delta z S_K \otimes (v_{y_0,\ell} e_{0J} e_{0J}^T + v_{y_J,\ell} e_{JJ} e_{JJ}^T) \otimes S_M^T \Delta x S_M \cdot G_\ell \\ & + 2\frac{1}{\sqrt{4\pi}} \sum_{\ell=1}^{L} w_\ell (v_{z_0,\ell} e_{0K} e_{0K}^T + v_{z_K,\ell} e_{KK} e_{KK}^T) \otimes S_J^T \Delta y S_J \otimes S_M^T \Delta x S_M \cdot G_\ell, \end{aligned} \tag{5.4}$$

where $\Delta r^n \equiv I_{2n+1} \otimes \Delta r$ for $n = 0, 1$, $v_{x_0,\ell} = \vec{n}_{x_0} \cdot \Omega_\ell$ if $\vec{n}_{x_0} \cdot \Omega_\ell < 0$ and 0 otherwise, etc., $G_\ell \in \mathbf{R}^{nodes}$ are the elements of the boundary condition nodal vector G, and

$$\begin{aligned} D_{co} \equiv {} & S^T \Delta r \Sigma_{a,0} S + 2\alpha A \\ & + \frac{1}{3}\left(C_x^T \Delta r \Sigma^{-1} C_x + C_y^T \Delta r \Sigma^{-1} C_y + C_z^T \Delta r \Sigma^{-1} C_z\right), \end{aligned} \tag{5.5}$$

with $\alpha = \frac{1}{4\pi} \sum_{\mu_l > 0} w_l \mu_l$, $\Sigma_{a,0} \equiv \Sigma(I - \Gamma_0)$, and

$$A = (\hat{E}_{MM}\hat{E}_{MM}^T + \hat{E}_{0M}\hat{E}_{0M}^T) + (\hat{E}_{JJ}\hat{E}_{JJ}^T + \hat{E}_{0J}\hat{E}_{0J}^T) + (\hat{E}_{KK}\hat{E}_{KK}^T + \hat{E}_{0K}\hat{E}_{0K}^T), \tag{5.6}$$

where

$$\hat{E}_{0M} \equiv S_{\vec{n}_{x_0}}^T (\Delta z \otimes \Delta y \otimes I_M)^{1/2}, \quad \hat{E}_{MM} \equiv S_{\vec{n}_{x_M}}^T (\Delta z \otimes \Delta y \otimes I_M)^{1/2}, \tag{5.7}$$

and similarly for other directions.

It is shown in [5] that D_{co} is singular. Nevertheless, the system (5.4) has solutions and can be solved in a least-squares sense using the pseudo-inverse D_{co}^+ of D_{co} [9, 8]. Then, the DSA preconditioner for A_0 in (4.18) can be obtained from the P_1 approximation to the transport problem. This preconditioner reads [5]

$$C_0 = I_{zones} + S D_{co}^+ S^T \Sigma \Delta r \Gamma_0. \tag{5.8}$$

It is also shown in [5] that a good initial guess is given by

$$\Phi_0 = S D_{co}^+ \left\{ \begin{pmatrix} \Sigma^0 \Delta r S \\ \Delta r \Sigma^1 T_{10} \end{pmatrix}^T \begin{pmatrix} (\Sigma^0)^{-1} Q_0 \\ (\Sigma^1)^{-1} Q_1 \end{pmatrix} + 2\frac{1}{\sqrt{4\pi}} \sum_{\ell=1}^{L} W_\ell G_\ell \right\}, \tag{5.9}$$

where all the boundary terms are lumped into the sum over ℓ.

6. Limiting behavior of A_0 and C_0.

In this section, we demonstrate that the DSA preconditioner C_0 works well in different problem regimes. Our results are 3-D generalizations of the slab geometry results of Ashby et al. [4]. For ease of readability, the proofs were put in an appendix. Throughout the analysis, the number of quadrature points (L) and the number of spatial zones in each direction (M, J, and K) are assumed constant, as well as the mesh size.

Recall that the matrix A_0 is given by

$$A_0 = I_{zones} - K_{0,0}\Gamma_0,$$

where $K_{0,0} = L_0\bar{S}H^{-1}ZL_0^+$. For a fixed mesh size, H is a function only of Σ, hence A_0 is a function of the diagonal matrices Σ and Γ_0. We write $A_0 = A_0(\Sigma, \Gamma_0)$ to denote this dependence. Also, in what follows, I stands for I_{zones}.

6.1. The thick regime.

The thick regime refers to problems with large total cross sections, i.e., where $\|\Sigma^{-1}\|_2$ is small. We prove in section A.2 the following general result for the thick limit.

THEOREM 6.1. *Let A_0, Σ, and Γ_0 be defined as above. Then*

$$\|A_0(\Sigma,\Gamma_0) - (I - \Gamma_0)\|_2 \to 0 \quad \textit{as} \quad \|\Sigma^{-1}\|_2 \to 0, \tag{6.1}$$

uniformly for $0 \le \Gamma_0 \le I$.

6.1.1. The asymptotic diffusion limit.

A particular case of the thick regime is the asymptotic diffusion limit, which is defined by setting

$$\begin{aligned} \Sigma &= \epsilon^{-1}\hat{\Sigma}, \\ \Gamma_0 &= I - \epsilon^2\hat{\Gamma}_0, \\ Q &= \epsilon\hat{Q}, \end{aligned} \tag{6.2}$$

where $\hat{\Sigma}$ and $\hat{\Gamma}_0$ are fixed diagonal matrices, $\hat{Q}$ is a fixed zone-centered vector, and letting $\epsilon \to 0$. In view of Theorem 6.1, $\Gamma_0 \to I$ in a thick regime implies $A_0 \to 0$ so that C_0 has to converge to infinity to be a good preconditioner. The following theorem, proved in section A.2.1, shows the effectiveness of the DSA preconditioner in the asymptotic diffusion limit (even when $\hat{\Gamma}_0 = 0$).

THEOREM 6.2. *Assume that $\Sigma = \epsilon^{-1}\hat{\Sigma}$ and that $\Gamma_0 = I - \epsilon^2\hat{\Gamma}_0$, where $\hat{\Sigma} > 0$ and $\hat{\Gamma}_0 \ge 0$ are fixed diagonal matrices. Then*

$$\|C_0(\epsilon)A_0(\epsilon) - I\|_2 \to 0 \quad \textit{as} \quad \epsilon \to 0. \tag{6.3}$$

We also have (see section A.2.1) that in the asymptotic diffusion limit, $\bar{S}\Psi$ is well approximated (in L^2 norm) by $L_0^+\Phi_0$, where Ψ is the solution of (4.16) and Φ_0 is the solution of (4.18). This is formalized as follows.

THEOREM 6.3. *Assume that $\Sigma = \epsilon^{-1}\hat{\Sigma}$, $\Gamma_0 = I - \epsilon^2\hat{\Gamma}_0$, and $Q = \epsilon\hat{Q}$ for any fixed diagonal matrices $\hat{\Sigma} > 0$ and $\hat{\Gamma}_0 \ge 0$, and fixed vector $\hat{Q}$. Let Δr be fixed. Then*

$$\|\bar{S}\Psi_\epsilon - L_0^+\Phi_{0,\epsilon}\|_2 \to 0 \tag{6.4}$$

as $\epsilon \to 0$, where Ψ_ϵ and $\Phi_{0,\epsilon}$ are the respective solutions of (4.16) *and* (4.18). *If we additionally assume that $g \equiv 0$ in* (3.2) (*i.e., zero Dirichlet boundary conditions*), *then it also follows that*

$$\|\Phi_{0,\epsilon}^{init} - \Phi_{0,\epsilon}\|_2 \to 0, \tag{6.5}$$

where $\Phi_{0,\epsilon}^{init}$ is given by (5.9).

6.1.2. Another thick limit. We here consider the case of a thick regime where Γ_0 is bounded away from the identity. Then A_0 is bounded away from zero, and the analysis is somewhat simpler. Theorem 6.4 is proved in section A.2.2.

THEOREM 6.4. *Let C_0 be defined as above. Then for any $0 < \epsilon < 1$,*

$$\|C_0(\Sigma, \Gamma_0) - (I - \Gamma_0)^{-1}\|_2 \to 0 \quad \text{as} \quad \|\Sigma^{-1}\|_2 \to 0, \tag{6.6}$$

uniformly for $0 \leq \Gamma_0 \leq (1 - \epsilon)I$.

In view of Theorem 6.1, we have as an immediate corollary.

COROLLARY 6.5. *Under the assumptions of Theorem 6.1, given any $0 < \epsilon < 1$, we have*

$$\|C_0(\Sigma, \Gamma_0)A_0(\Sigma, \Gamma_0) - I\|_2 \to 0 \quad \text{as} \quad \|\Sigma^{-1}\|_2 \to 0,$$

uniformly for $0 \leq \Gamma_0 \leq (1 - \epsilon)I$.

6.2. The thin regime. The thin regime refers to problems with small total cross-sections, i.e., where $\|\Sigma\|_2$ is small. In this limit, the source iteration procedure is known to converge in one iteration [13]. It is therefore not surprising to have the following theorem.

THEOREM 6.6. *Let A_0 be defined as above. Then*

$$\|A_0(\Sigma, \Gamma_0) - I\|_2 \to 0$$

as $\|\Sigma\|_2 \to 0$, uniformly for $0 \leq \Gamma_0 \leq I$.

As in slab geometry [4], the proof involves the observation that H_ℓ^{-1} has a limiting value that is annihilated by Z_0 on the right so that $H^{-1}Z \to 0$, and so $\|A_0 - I\|_2 \to 0$ as $\|\Sigma\|_2 \to 0$.

Note that this theorem remains valid in pure absorbing media, i.e., for $\|\Gamma_0\|_2 \to 0$ and $K_1 < \|\Sigma\|_2 < K_2$ (with K_1, K_2 positive constants). This is also not surprising since the source iteration procedure is known to converge in one iteration in this limit as well [13].

Theorem 6.6 implies that the system (4.18) needs progressively less preconditioning as $\|\Sigma\|_2$ gets small. A preconditioner for (4.18) should thus converge to the identity as $\|\Sigma\|_2 \to 0$. The next result says that this indeed holds for C_0.

THEOREM 6.7. *Let C_0 be defined as above, and assume $\Sigma = \epsilon\hat{\Sigma}$. Then*

$$\|C_0(\epsilon, \Gamma_0) - I\|_2 \to 0$$

as $\epsilon \to 0$, uniformly for $0 \leq \Gamma_0 \leq I$.

The proof of this theorem is provided in section A.3. Again, this remains true in the pure absorbing case.

7. Discussion. Despite the singularity of the "consistently" differenced 3-D diffusion approximation required by the DSA, we could extend the slab geometry results of Ashby et al. [4] to 3-D geometry. We proved the efficiency of the 3-D DSA preconditioner in the asymptotic diffusion limit even though A_0 tends to zero. We also proved that in this limit, the discrete BTE solution is well-approximated (in the L^2 sense of Theorem 6.3) by its first discrete angular moment. Furthermore, the 3-D DSA preconditioner was proved efficient in another thick limit where A_0 is bounded away from zero. Our theoretical results thus confirm the good numerical results obtained by Brown [5]. Even if it does not pertain to the preconditioner performances, one should be aware that the discrete ordinate approximation does not necessarily provide accurate solutions in the thick limit. This was demonstrated by Larsen and Morel [10, 11] for the diamond-difference scheme, in case of unresolved boundary layers with anisotropic incoming fluxes. One should thus be careful in interpreting the output in this limit. In the thin limit, acceleration becomes unnecessary and the 3-D DSA preconditioner accordingly tends to the identity.

Possible theoretical extensions include the introduction of linear anisotropy [5] at the price of more cumbersome notations. Also, reflected boundary conditions could be introduced. This was done in slab geometry [4] using a mirror domain extension obtained by reflection about the boundary. Ashby et al. then obtained relationships between the A_0 and C_0 matrices and their reflected counterparts on the mirror domain. Kronecker product notations would certainly appear extremely useful in trying to generalize such results to 3-D geometry. Finally, the behavior of our DSA preconditioner for any value of the cross-sections still has to be theoretically investigated in slab geometry before possible 3-D generalizations can be addressed.

Another context where DSA can be found useful is the corner balance spatial discretization scheme [1, 2]. We started addressing the extension of the slab geometry DSA analysis to this scheme.

Appendix.

A.1. Additional notation. Before proving the theorems, we need to introduce some additional notation. Since

$$H_\ell \equiv \begin{pmatrix} S + \Sigma^{-1} C_\ell \\ B_\ell \end{pmatrix} \in \mathbf{R}^{nodes \times nodes}$$

for all $l = 1, \dots, L$, it can also be written as

$$H_\ell = H_{\ell,0} + Z_0(\Sigma^{-1} C_\ell),$$

with Z_0 defined in (4.15) and

$$H_{\ell,0} \equiv \begin{pmatrix} S \\ B_\ell \end{pmatrix}$$

independent of Σ. Note that H_ℓ and $H_{\ell,0}$ were shown to be nonsingular in [5]. Note also that what was defined as E_{000} in [5] is now defined as E_{000}^T, in order to resemble the one-dimensional case of [4] (where $e_{..}$ plays the role of $E_{...}$ here). Next, we define the matrices V_ℓ and W_ℓ so that

$$(V_\ell, W_\ell) = H_{\ell,0}^{-1},$$

where $V_\ell \in \mathbf{R}^{nodes \times zones}$ and $W_\ell \in \mathbf{R}^{nodes \times (nodes - zones)}$. Hence the following identities are true:

$$\begin{aligned} & V_\ell S + W_\ell B_\ell = I_{nodes}, \\ \text{(A.1)} \qquad & S V_\ell = I \qquad (i.e., I_{zones}), \\ & S W_\ell = 0\,(\in \mathbf{R}^{zones \times zones}), \\ \text{(A.2)} \qquad & B_\ell V_\ell = 0\,(\in \mathbf{R}^{[nodes-zones] \times [nodes-zones]}), \text{ and} \\ & B_\ell W_\ell = I_{[nodes-zones]}. \end{aligned}$$

Since B_ℓ defined in (4.13) can take eight different forms depending on the sign of μ_ℓ, η_ℓ and ξ_ℓ, V_ℓ and W_ℓ can each also take eight different forms and are constant in one octant. Further developments require a closed form for V_ℓ, which we derive in the following lemma.

LEMMA A.1. *We have*

$$V_\ell = \begin{cases} L_K \otimes L_J \otimes L_M, & \textit{if } \mu_\ell \geq 0, \eta_\ell \geq 0, \textit{ and } \xi_\ell \geq 0, \\ L_K \otimes L_J \otimes U_M, & \textit{if } \mu_\ell \leq 0, \eta_\ell \geq 0, \textit{ and } \xi_\ell \geq 0, \\ L_K \otimes U_J \otimes L_M, & \textit{if } \mu_\ell \geq 0, \eta_\ell \leq 0, \textit{ and } \xi_\ell \geq 0, \\ U_K \otimes L_J \otimes L_M, & \textit{if } \mu_\ell \geq 0, \eta_\ell \geq 0, \textit{ and } \xi_\ell \leq 0, \\ L_K \otimes U_J \otimes U_M, & \textit{if } \mu_\ell \leq 0, \eta_\ell \leq 0, \textit{ and } \xi_\ell \geq 0, \\ U_K \otimes L_J \otimes U_M, & \textit{if } \mu_\ell \leq 0, \eta_\ell \geq 0, \textit{ and } \xi_\ell \leq 0, \\ U_K \otimes U_J \otimes L_M, & \textit{if } \mu_\ell \geq 0, \eta_\ell \leq 0, \textit{ and } \xi_\ell \leq 0, \\ U_K \otimes U_J \otimes U_M, & \textit{if } \mu_\ell \leq 0, \eta_\ell \leq 0, \textit{ and } \xi_\ell \leq 0, \end{cases} \tag{A.3}$$

where

$$L_M \equiv 2 \begin{pmatrix} 0 & 0 & \cdots & 0 \\ (-1)^2 & 0 & \cdots & 0 \\ (-1)^3 & (-1)^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ (-1)^M & (-1)^{M-1} & \cdots & 0 \\ (-1)^{M+1} & (-1)^M & \cdots & (-1)^2 \end{pmatrix} \in \mathbf{R}^{(M+1)\times M},$$

$$U_M \equiv 2 \begin{pmatrix} (-1)^2 & (-1)^3 & \cdots & (-1)^{M+1} \\ 0 & (-1)^2 & \cdots & (-1)^M \\ 0 & 0 & \cdots & (-1)^{M-1} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & (-1)^2 \\ 0 & 0 & \cdots & 0 \end{pmatrix} \in \mathbf{R}^{(M+1)\times M}.$$

Proof. First the V_ℓ in (A.3) verify $B_\ell V_\ell = 0$ for all ℓ. Indeed, for instance,

$$E_{00K}^T \cdot (U_K \otimes L_J \otimes L_M) = \begin{pmatrix} e_{KK}^T \otimes I_{J+1} \otimes I_{M+1} \\ (I_K, 0) \otimes e_{0J}^T \otimes I_{M+1} \\ (I_K, 0) \otimes (0, I_J) \otimes e_{0M}^T \end{pmatrix} (U_K \otimes L_J \otimes L_M) = 0$$

since $e_{KK}^T U_K = 0$, $e_{0J}^T L_J = 0$, and $e_{0M}^T L_M = 0$. Then the V_ℓ in (A.3) also verify $SV_\ell = I$ for all ℓ, because $S_K L_K = I_K = S_K U_K$, and similarly for the other two directions. The fact that $H_{\ell,0}$ is nonsingular assures that these V_ℓ are unique. □

Furthermore, we define

$$\begin{aligned} F_x &\equiv S^T \Delta r C_x + C_x^T \Delta r S, \\ F_y &\equiv S^T \Delta r C_y + C_y^T \Delta r S, \\ F_z &\equiv S^T \Delta r C_z + C_z^T \Delta r S, \end{aligned} \tag{A.4}$$

which can also be written in view of Lemma A.1 in [5] as

$$\begin{aligned} F_x &= S_K^T \Delta z S_K \otimes S_J^T \Delta y S_J \otimes (-e_{0M} e_{0M}^T + e_{MM} e_{MM}^T), \\ F_y &= S_K^T \Delta z S_K \otimes (-e_{0J} e_{0J}^T + e_{JJ} e_{JJ}^T) \otimes S_M^T \Delta x S_M, \\ F_z &= (-e_{0K} e_{0K}^T + e_{KK} e_{KK}^T) \otimes S_J^T \Delta y S_J \otimes S_M^T \Delta x S_M. \end{aligned}$$

Given (5.1) and (5.7), we have

$$S_K^T \Delta z S_K \otimes S_J^T \Delta y S_J \otimes e_{0M} e_{0M}^T = S_{\vec{n}_{x_0}}^T (\Delta z \otimes \Delta y \otimes I_M) S_{\vec{n}_{x_0}} = \hat{E}_{0M} \hat{E}_{0M}^T.$$

Thus, we end up with

$$\begin{aligned} F_x &= \hat{E}_{MM} \hat{E}_{MM}^T - \hat{E}_{0M} \hat{E}_{0M}^T, \\ F_y &= \hat{E}_{JJ} \hat{E}_{JJ}^T - \hat{E}_{0J} \hat{E}_{0J}^T, \\ F_z &= \hat{E}_{KK} \hat{E}_{KK}^T - \hat{E}_{0K} \hat{E}_{0K}^T. \end{aligned}$$

A.2. Proofs for the thick regime. We first prove Theorem 6.1. We need two technical lemmas.

LEMMA A.2. *Let $0 < \delta < 1$ be given. Then for any $\Sigma > 0$ such that*

$$\|\Sigma^{-1}\|_2 \leq \epsilon_1 \equiv \frac{\delta}{\max_\ell \{\|C_\ell \cdot (V_\ell, W_\ell)\|_2\}}, \tag{A.5}$$

we have

$$SH_\ell^{-1} Z_0 = \sum_{k=0}^{\infty} (-1)^k (\Sigma^{-1} C_\ell V_\ell)^k$$

for all $\ell = 1, \dots, L$. Furthermore, the convergence is uniform in Σ with $\|\Sigma^{-1}\|_2 \leq \epsilon_1$.

Proof. For $\|\Sigma^{-1}\|_2 \leq \epsilon_1$, we have

$$\begin{aligned} H_\ell^{-1} &= H_{\ell,0}^{-1} (I + Z_0 \Sigma^{-1} C_\ell H_{\ell,0}^{-1})^{-1} \\ &= (V_\ell, W_\ell) \sum_{k=0}^{\infty} (-1)^k \left(Z_0 \Sigma^{-1} C_\ell \cdot (V_\ell, W_\ell) \right)^k, \end{aligned}$$

where the Neumann series converges uniformly since

$$\|Z_0 \Sigma^{-1} C_\ell \cdot (V_\ell, W_\ell)\|_2 \leq \|Z_0\|_2 \cdot \|\Sigma^{-1}\|_2 \|C_\ell \cdot (V_\ell, W_\ell)\|_2 \leq \delta < 1,$$

using $\|Z_0\|_2 = 1$. Thus,

$$\begin{aligned} SH_\ell^{-1} Z_0 &= S \cdot (V_\ell, W_\ell) \cdot \sum_{k=0}^{\infty} (-1)^k \left(Z_0 \Sigma^{-1} C_\ell \cdot (V_\ell, W_\ell) \right)^k Z_0 \\ &= S \cdot (V_\ell, W_\ell) \cdot Z_0 \sum_{k=0}^{\infty} (-1)^k \left(\Sigma^{-1} C_\ell \cdot (V_\ell, W_\ell) \cdot Z_0 \right)^k \\ &= \sum_{k=0}^{\infty} (-1)^k \left(\Sigma^{-1} C_\ell V_\ell \right)^k \end{aligned}$$

from the fact that $S \cdot (V_\ell, W_\ell) \cdot Z_0 = I$ (A.1) and $C_\ell \cdot (V_\ell, W_\ell) \cdot Z_0 = C_\ell V_\ell$. □

LEMMA A.3. *Let $0 \leq \Gamma_0 \leq I$, and let ϵ_1 be defined by* (A.5). *Then*

$$A_0 = (I - \Gamma_0) - \frac{1}{4\pi} F(\Sigma^{-1}) \cdot \Gamma_0, \tag{A.6}$$

where

$$F(\Sigma^{-1}) \equiv \sum_{\ell=1}^{L} w_\ell \sum_{k=1}^{\infty} (-1)^k \left(\Sigma^{-1} C_\ell V_\ell\right)^k, \tag{A.7}$$

whenever $\|\Sigma^{-1}\|_2 \leq \epsilon_1$, *with the series converging uniformly.*

Proof. We first consider $K_{0,0}$ defined in section 4.3. We have

$$\begin{aligned} K_{0,0} &\equiv L_0 \bar{S} H^{-1} Z L_0^+ \\ &= \frac{1}{4\pi} \sum_{\ell=1}^{L} w_\ell S H_\ell^{-1} Z_0 \\ &= \frac{1}{4\pi} \sum_{\ell=1}^{L} w_\ell \sum_{k=0}^{\infty} (-1)^k \left(\Sigma^{-1} C_\ell V_\ell\right)^k \\ &= I + \frac{1}{4\pi} \sum_{\ell=1}^{L} w_\ell \sum_{k=1}^{\infty} (-1)^k \left(\Sigma^{-1} C_\ell V_\ell\right)^k \end{aligned}$$

where we used successively Lemma A.2 and the fact that $\sum_{\ell=1}^{L} w_\ell = 4\pi$. The lemma now follows easily using the definition (4.19) of A_0. □

Proof of Theorem 6.1. Using (A.6), $\|\Gamma_0\|_2 \leq 1$ and $\frac{w_\ell}{4\pi} \leq 1$, it now follows that

$$\begin{aligned} \|A_0 - (I - \Gamma_0)\|_2 &\leq \frac{1}{4\pi} \left\| \sum_{\ell=1}^{L} w_\ell \sum_{k=1}^{\infty} (-1)^k \left(\Sigma^{-1} C_\ell V_\ell\right)^k \right\|_2 \|\Gamma_0\|_2 \\ &\leq \sum_{\ell=1}^{L} \sum_{k=1}^{\infty} \| \left(\Sigma^{-1} C_\ell V_\ell\right)^k \|_2 \to 0 \end{aligned}$$

as $\|\Sigma^{-1}\|_2 \to 0$, and the convergence is uniform in $0 \leq \Gamma_0 \leq I$. □

A.2.1. Proofs for the asymptotic diffusion limit. Here, we prove Theorems 6.2 and 6.3 for the asymptotic diffusion limit. Recall from (5.8) that

$$C_0 = I + S D_{co}^{+} S^T \Sigma \Delta r \Gamma_0. \tag{A.8}$$

Using (A.6), we can write

$$C_0 A_0 - I = -\Gamma_0 - \frac{1}{4\pi} F(\Sigma^{-1}) \Gamma_0 + S D_{co}^{+} S^T \Sigma \Delta r \Gamma_0 A_0. \tag{A.9}$$

We will prove Proposition A.6, which establishes an extended form for $S^T \Sigma \Delta r \Gamma_0 A_0$. This proposition requires two preliminary lemmas, which we now establish.

Lemma A.4. *With the above notations, we have*

$$\sum_{\ell=1}^{L} w_\ell \mu_\ell F_x V_\ell = \pi \alpha (\hat{E}_{MM} \hat{E}_{MM}^T + \hat{E}_{0M} \hat{E}_{0M}^T) \sum_{i=1}^{8} V_{\mathcal{L}_i},$$

where $\alpha \equiv \frac{1}{4\pi} \sum_{\mu_\ell > 0} w_\ell \mu_\ell$, *and* $V_{\mathcal{L}_i}$ *is the (unique) value of* V_ℓ *for all* ℓ *corresponding to a direction* Ω_ℓ *in octant* i $(i = 1, 2, \ldots, 8)$. *We have similar results for the* y *and* z *directions.*

Proof. We have

$$\sum_{l=1}^{L} w_l \mu_l F_x V_l = \sum_{\mu_l > 0} w_l \mu_l F_x V_l + \sum_{\mu_l < 0} w_l \mu_l F_x V_l.$$

We showed in section A.1 that F_x can be expressed as

$$F_x = \hat{E}_{MM} \hat{E}_{MM}^T - \hat{E}_{0M} \hat{E}_{0M}^T. \tag{A.10}$$

Recall that $S_{\vec{n}_{x_0}} \equiv S_K \otimes S_J \otimes e_{0M}^T$ selects the $x = x_0$ face and then forms two-dimensional zone averages on this face. Meanwhile, E_{000}^T also selects (among others) the $x = x_0$ face. Thus, there exists a matrix $M_{000,\vec{n}_{x_0}}$ restricting E_{000}^T such that

$$S_{\vec{n}_{x_0}} = M_{000,\vec{n}_{x_0}} E_{000}^T. \tag{A.11}$$

Now, in view of (4.13) and (A.2), $E_{000}^T V_\ell = 0$ for any ℓ such that $\mu_\ell > 0$, $\eta_\ell > 0$, and $\xi_\ell > 0$. Thus, $S_{\vec{n}_{x_0}} V_\ell = 0$ and given (5.7), $\hat{E}_{0M} \hat{E}_{0M}^T V_\ell = 0$ for such ℓ. Moreover, $S_{\vec{n}_{x_0}}$ can be obtained as in (A.11) from any B_ℓ defined in (4.13) that selects the $x = x_0$ face, that is, from any B_ℓ with ℓ such that $\mu_l > 0$. Therefore, $\hat{E}_{0M} \hat{E}_{0M}^T V_\ell = 0$ for any ℓ such that $\mu_\ell > 0$. Similarly, $\hat{E}_{MM} \hat{E}_{MM} V_\ell = 0$ for any ℓ such that $\mu_\ell < 0$. Thus, flipping the sign of the vanishing terms, we can write

$$\begin{aligned} \sum_{\ell=1}^{L} w_\ell \mu_\ell F_x V_\ell = & \sum_{\mu_\ell > 0} w_\ell \mu_\ell (\hat{E}_{MM} \hat{E}_{MM}^T + \hat{E}_{0M} \hat{E}_{0M}^T) V_\ell \\ & - \sum_{\mu_l < 0} w_l \mu_\ell (\hat{E}_{MM} \hat{E}_{MM}^T + \hat{E}_{0M} \hat{E}_{0M}^T) V_\ell. \end{aligned} \tag{A.12}$$

We have

$$\sum_{\ell=1}^{L} = \sum_{i=1}^{8} \sum_{\ell \in \mathcal{L}_i}, \text{ where } \mathcal{L}_i \text{ includes all the } \ell \text{ corresponding to a } \Omega_\ell \text{ in one octant } i.$$

Let us take the convention that $i = 1, 2, 3$, and 4 correspond to the four octants where $\mu_\ell > 0$. We introduce α

$$\begin{aligned} \alpha &\equiv \frac{1}{4\pi} \sum_{\mu_\ell > 0} w_\ell \mu_\ell \\ &= \frac{1}{4\pi} \sum_{i=1}^{4} \sum_{\ell \in \mathcal{L}_i} w_\ell \mu_\ell \\ &= \frac{1}{\pi} \sum_{\ell \in \mathcal{L}_i} w_\ell \mu_\ell \quad (i = 1, 2, 3 \text{ or } 4) \end{aligned}$$

by symmetry over the four octants $i = 1, 2, 3$, and 4. Due to the symmetrical placement of the μ_ℓ along the x-axis, we have for the other four octants

$$\alpha = -\frac{1}{\pi} \sum_{\ell \in \mathcal{L}_i} w_\ell \mu_\ell \quad (i = 5, 6, 7 \text{ or } 8).$$

Since V_ℓ is constant in each octant, we introduce the notation $V_{\mathcal{L}_i}$ to denote its value for all $\ell \in \mathcal{L}_i$, i.e., for all ℓ corresponding to a direction Ω_ℓ in octant i. Then

$$\begin{aligned}\sum_{\mu_\ell>0} w_\ell \mu_\ell (\hat{E}_{MM}\hat{E}_{MM}^T + \hat{E}_{0M}\hat{E}_{0M}^T)V_\ell &= \sum_{i=1}^{4}\sum_{\ell\in\mathcal{L}_i} w_\ell \mu_\ell (\hat{E}_{MM}\hat{E}_{MM}^T + \hat{E}_{0M}\hat{E}_{0M}^T)V_{\mathcal{L}_i} \\ &= \pi\alpha(\hat{E}_{MM}\hat{E}_{MM}^T + \hat{E}_{0M}\hat{E}_{0M}^T)\sum_{i=1}^{4} V_{\mathcal{L}_i}.\end{aligned}$$

Similarly, for $\mu_\ell < 0$, we have

$$\sum_{\mu_\ell<0} w_\ell \mu_\ell (\hat{E}_{MM}\hat{E}_{MM}^T + \hat{E}_{0M}\hat{E}_{0M}^T)V_\ell = -\pi\alpha \sum_{i=5}^{8}(\hat{E}_{MM}\hat{E}_{MM}^T + \hat{E}_{0M}\hat{E}_{0M}^T)V_{\mathcal{L}_i}.$$

The conclusion now follows directly from (A.12). □

Lemma A.5.

$$\sum_{\ell=1}^{L} w_\ell C_\ell^T \Sigma^{-1}\Delta r C_\ell V_\ell = \frac{4\pi}{3}\frac{1}{8}\left[C_x^T\Sigma^{-1}\Delta r C_x + C_y^T\Sigma^{-1}\Delta r C_y + C_z^T\Sigma^{-1}\Delta r C_z\right]\sum_{i=1}^{8} V_{\mathcal{L}_i},$$

where $V_{\mathcal{L}_i}$ is defined as in Lemma A.4.

Proof. We have

$$\begin{aligned}\sum_{\ell=1}^{L} w_\ell C_\ell^T \Sigma^{-1}\Delta r C_\ell V_\ell = \sum_{\ell=1}^{L} w_\ell[&(\mu_\ell)^2 C_x^T\Sigma^{-1}\Delta r C_x + \mu_\ell\eta_\ell C_x^T\Sigma^{-1}\Delta r C_y \\ &+ \mu_\ell\xi_\ell C_x^T\Sigma^{-1}\Delta r C_z + \eta_\ell\mu_\ell C_y^T\Sigma^{-1}\Delta r C_x \\ &+ (\eta_\ell)^2 C_y^T\Sigma^{-1}\Delta r C_y + \eta_\ell\xi_\ell C_y^T\Sigma^{-1}\Delta r C_z \\ &+ \xi_\ell\mu_\ell C_z^T\Sigma^{-1}\Delta r C_x + \xi_\ell\eta_\ell C_z^T\Sigma^{-1}\Delta r C_y \\ &+ (\xi_l)^2 C_z^T\Sigma^{-1}\Delta r C_z]V_\ell.\end{aligned} \tag{A.13}$$

First we look at the "diagonal" terms in (A.13). We have for the x-direction

$$\begin{aligned}\sum_{\ell=1}^{L} w_\ell(\mu_\ell)^2 C_x^T\Sigma^{-1}\Delta r C_x V_\ell &= \sum_{i=1}^{8}\sum_{\ell\in\mathcal{L}_i} w_\ell(\mu_\ell)^2 C_x^T\Sigma^{-1}\Delta r C_x V_\ell \\ &= \frac{4\pi}{3}\frac{1}{8}\sum_{i=1}^{8} C_x^T\Sigma^{-1}\Delta r C_x V_{\mathcal{L}_i}\end{aligned}$$

using (4.2) and the symmetry over the eight octants. Combining the three directions, we obtain Lemma A.5 provided the "nondiagonal" terms in (A.13) vanish. Indeed, take for instance

$$\begin{aligned}&\sum_{\ell=1}^{L} w_\ell\eta_\ell\mu_\ell C_y^T\Sigma^{-1}\Delta r C_x V_\ell \\ &\quad = C_y^T\Sigma^{-1}\Delta r\left\{\sum_{i=1}^{4}\sum_{\ell\in\mathcal{L}_i} w_\ell\eta_\ell\mu_\ell C_x V_\ell + \sum_{i=5}^{8}\sum_{\ell\in\mathcal{L}_i} w_\ell\eta_\ell\mu_\ell C_x V_\ell\right\}.\end{aligned}$$

If the four octants $i = 1, 2, 3$ and 4 correspond to positive values of μ_ℓ, we have

$$\sum_{i=1}^{4}\sum_{\ell\in\mathcal{L}_i} w_\ell\eta_\ell\mu_\ell C_x V_\ell = (I_K \otimes I_J \otimes (\Delta x)^{-1} D_M L_M)\sum_{i=1}^{4}\sum_{\ell\in\mathcal{L}_i} w_\ell\eta_\ell\mu_\ell,$$

using (A.3) as well as $S_K L_K = I_K = S_K U_K$ and $S_J L_J = I_J = S_J U_J$. Since by symmetry $\sum_{i=1}^{4}\sum_{\ell\in\mathcal{L}_i} w_\ell\eta_\ell\mu_\ell = 0$, the "nondiagonal" terms vanish, and Lemma A.5 is proved. $\square$

We are now prepared to establish the following proposition.

PROPOSITION A.6. *For* $\|\Sigma^{-1}\|_2 \le \epsilon_1$*, with* ϵ_1 *defined in* (A.5)*, we have*

$$\begin{aligned}
&S^T\Sigma\Delta r\Gamma_0 A_0\\
&\quad = \left\{\frac{1}{8}D_{co}\sum_{i=1}^{8} V_{\mathcal{L}_i} + \frac{1}{4\pi}\sum_{k=2}^{\infty}\sum_{\ell=1}^{L} w_\ell\left\{C_\ell^T\Delta r\Sigma^{-1}C_\ell + |\mu_\ell| A\right\}V_\ell((-1)\Sigma^{-1}C_\ell V_\ell)^{k-1}\right.\\
&\qquad\left. + \frac{1}{4\pi}S^T\Sigma\Delta r(I-\Gamma_0)F(\Sigma^{-1})\right\}\cdot\Gamma_0,
\end{aligned}$$

where D_{co} *and* A *were defined in* (5.5) *and* (5.6)*, respectively.*

Proof. From (A.6), we have

$$\begin{aligned}
S^T\Sigma\Delta r\Gamma_0 A_0 &= \left\{S^T\Sigma\Delta r(I-\Gamma_0) - \frac{1}{4\pi}S^T\Sigma\Delta r\Gamma_0 F(\Sigma^{-1})\right\}\Gamma_0\\
&= \left\{S^T\Delta r\Sigma_{a,0} - \frac{1}{4\pi}S^T\Sigma\Delta r F(\Sigma^{-1}) + \frac{1}{4\pi}S^T\Sigma\Delta r(I-\Gamma_0)F(\Sigma^{-1})\right\}\Gamma_0, \qquad \text{(A.14)}
\end{aligned}$$

where $\Sigma_{a,0} = \Sigma(I-\Gamma_0)$.

Let us look now at the $S^T\Sigma\Delta r F(\Sigma^{-1})$ term

$$S^T\Sigma\Delta r F(\Sigma^{-1}) = \sum_{k=1}^{\infty}(-1)^k\sum_{\ell=1}^{L} w_\ell S^T\Sigma\Delta r\left(\Sigma^{-1}C_\ell V_\ell\right)^k.$$

We have

$$\begin{aligned}
S^T\Sigma\Delta r\left(\Sigma^{-1}C_\ell V_\ell\right)^k &= S^T\Delta r C_\ell V_\ell\left(\Sigma^{-1}C_\ell V_\ell\right)^{k-1}\\
&= S^T\Delta r(\mu_\ell C_x + \eta_\ell C_y + \xi_\ell C_z)V_\ell\left(\Sigma^{-1}C_\ell V_\ell\right)^{k-1}\\
&= [-C_\ell^T\Delta r S + \mu_\ell F_x + \eta_\ell F_y + \xi_\ell F_z]V_\ell\left(\Sigma^{-1}C_\ell V_\ell\right)^{k-1},
\end{aligned}$$

where F_x, F_y, and F_z were defined in (A.4). Recalling that $SV_\ell = I$ (A.1), we can write after some manipulations

$$\begin{aligned}
S^T\Sigma\Delta r F(\Sigma^{-1}) = &-\sum_{\ell=1}^{L} w_\ell[-C_\ell^T\Delta r + (\mu_\ell F_x + \eta_\ell F_y + \xi_\ell F_z)V_\ell]\\
&+\sum_{\ell=1}^{L} w_\ell[-C_\ell^T\Delta r\Sigma^{-1}C_\ell V_\ell] \qquad \text{(A.15)}\\
&+\sum_{k=2}^{\infty}\sum_{\ell=1}^{L}[-w_\ell C_\ell^T\Delta r(-1)^{k+1}(\Sigma^{-1}C_\ell V_\ell)^k\\
&+(-1)^k w_l(\mu_\ell F_x + \eta_l F_y + \xi_l F_z)V_l(\Sigma^{-1}C_l V_l)^{k-1}].
\end{aligned}$$

In the first sum of (A.15), the first term vanishes since

$$\sum_{l=1}^{L} w_l C_l^T = \sum_{l=1}^{L} w_l \mu_l C_x^T + \sum_{l=1}^{L} w_l \eta_l C_y^T + \sum_{l=1}^{L} w_l \xi_l C_z^T = 0,$$

where we have used (4.3). Lemma A.4 takes care of the second term in the first sum, while we can apply Lemma A.5 to the second sum. With A defined in (5.6), we obtain

$$\begin{aligned} S^T \Sigma \Delta r F(\Sigma^{-1}) = &-\pi\alpha A \sum_{i=1}^{8} V_{\mathcal{L}_i} \\ &- \frac{4\pi}{3}\frac{1}{8}\left[C_x^T \Sigma^{-1} \Delta r C_x + C_y^T \Sigma^{-1} \Delta r C_y + C_z^T \Sigma^{-1} \Delta r C_z\right] \sum_{i=1}^{8} V_{\mathcal{L}_i} \\ &+ \sum_{k=2}^{\infty} \sum_{\ell=1}^{L} \left[-w_\ell C_\ell^T \Delta r (-1)^{k+1} (\Sigma^{-1} C_\ell V_\ell)^k \right. \\ &\left. + (-1)^k w_\ell (\mu_\ell F_x + \eta_\ell F_y + \xi_\ell F_z) V_\ell (\Sigma^{-1} C_\ell V_\ell)^{k-1}\right]. \end{aligned}$$

Since $SV_\ell = I$ for all ℓ we have $I = \frac{1}{8} S \sum_{i=1}^{8} V_{\mathcal{L}_i}$. Then we have, using (A.14) and recalling definition (5.5) of D_{co},

$$\begin{aligned} S^T \Sigma \Delta r \Gamma_0 A_0 = &\left\{ \frac{1}{8} D_{co} \sum_{i=1}^{8} V_{\mathcal{L}_i} + \frac{1}{4\pi} \sum_{k=2}^{\infty} \sum_{\ell=1}^{L} w_\ell \{ C_\ell^T \Delta r (-1)^{k+1} (\Sigma^{-1} C_\ell V_\ell)^k \right. \\ &+ (\mu_\ell F_x + \eta_\ell F_y + \xi_\ell F_z) V_\ell ((-1) \Sigma^{-1} C_\ell V_\ell)^{k-1} \} \\ &\left. + \frac{1}{4\pi} S^T \Sigma \Delta r (I - \Gamma_0) F(\Sigma^{-1}) \right\} \cdot \Gamma_0. \end{aligned} \tag{A.16}$$

Now, proceeding as in (A.12) and using symmetry properties, the proposition follows. □

Before using Proposition A.6, we need to notice that

$$\frac{1}{8} S D_{co}^{+} D_{co} \sum_{i=1}^{8} V_{\mathcal{L}_i} = I. \tag{A.17}$$

Indeed, writing $V_{\mathcal{L}_i}$ as the direct sum $V_{\mathcal{L}_i} = R_1 \oplus R_2$, with $R_1 \in \mathcal{N}(D_{co})$ and $R_2 \in \mathcal{N}(D_{co})^\perp = \mathcal{R}(D_{co}^T)$, we have for $i = 1, \ldots, 8$ that $SD_{co}^{+} D_{co} V_{\mathcal{L}_i} = SR_2$ since $D_{co}^{+} D_{co}$ is the orthogonal projection on the range of $D_{co}^T = D_{co}$ [8]. Also, one can show [5] that $\mathcal{N}(D_{co}) \subset \mathcal{N}(S)$ which implies $SR_2 = SV_{\mathcal{L}_i} = I$ using (A.1). Thus (A.17) is verified, and we can write from (A.9) and Proposition A.6

$$\begin{aligned} C_0 A_0 - I = &-\frac{1}{4\pi} F(\Sigma^{-1}) \Gamma_0 \\ &+ S D_{co}^{+} \frac{1}{4\pi} \sum_{k=2}^{\infty} \sum_{\ell=1}^{L} w_\ell \left\{ C_\ell^T \Delta r \Sigma^{-1} C_\ell + |\mu_\ell| A \right\} V_\ell ((-1) \Sigma^{-1} C_\ell V_\ell)^{k-1} \Gamma_0 \\ &+ \frac{1}{4\pi} S D_{co}^{+} S^T \Sigma \Delta r (I - \Gamma_0)(SS^T)(SS^T)^{-1} F(\Sigma^{-1}) \Gamma_0. \end{aligned} \tag{A.18}$$

Given its definition in (A.7), $F(\Sigma^{-1}) \to 0$ when $\Sigma^{-1} \to 0$. Therefore, to prove our theorem, we will show that

$$\|D_{co}^{+} S^T \Sigma \Delta r (I - \Gamma_0) S\| \tag{A.19}$$

and

$$\|D_{co}^{+} w_\ell \left\{ C_\ell^T \Delta r \Sigma^{-1} C_\ell + |\mu_\ell| A \right\} \| \tag{A.20}$$

remain bounded as $\Sigma^{-1} \to 0$.

We need a few more intermediate results. Since D_{co} is symmetric and singular, its singular value decomposition [8] can be written

$$D_{co} = [U, V] \begin{pmatrix} \Theta & 0 \\ 0 & 0 \end{pmatrix} [U, V]^T, \tag{A.21}$$

where the vectors of V span the null space of D_{co} and those of U its range. The diagonal matrix Θ contains the nonzero singular values of D_{co}. Then, the pseudoinverse [8] reads

$$D_{co}^{+} = [U, V] \begin{pmatrix} \Theta^{-1} & 0 \\ 0 & 0 \end{pmatrix} [U, V]^T = U \Theta^{-1} U^T.$$

We have the following lemma.

LEMMA A.7. *With V defined as above, if R is such that $V^T R = 0$, then $D_{co}^{+} R = (D_{co} + \delta V V^T)^{-1} R$ for any $\delta > 0$.*

Proof.

$$(D_{co} + \delta V V^T)^{-1} = [U, V] \begin{pmatrix} \Theta^{-1} & 0 \\ 0 & \delta^{-1} I \end{pmatrix} [U, V]^T = U \Theta^{-1} U^T + \delta^{-1} V V^T,$$

thus

$$(D_{co} + \delta V V^T)^{-1} R = U \Theta^{-1} U^T R + \delta^{-1} V V^T R = D_{co}^{+} R,$$

which completes the proof. □

Plugging the asymptotic diffusion limit assumptions $\Sigma = \epsilon^{-1} \hat{\Sigma}$ and $\Gamma_0 = I - \epsilon^2 \hat{\Gamma}_0$ in the definition (5.5) of D_{co} yields

$$D_{co} + \epsilon V V^T = 2\alpha (A + \epsilon E) \tag{A.22}$$

with A defined in (5.6), and

$$E = \frac{1}{2\alpha} \left(\frac{1}{3} (C_x^T \hat{\Sigma}^{-1} \Delta r C_x + C_y^T \hat{\Sigma}^{-1} \Delta r C_y + C_z^T \hat{\Sigma}^{-1} \Delta r C_z) + S^T \hat{\Sigma} \hat{\Gamma}_0 \Delta r S + V V^T \right).$$

Arguments in [5] show that $\mathcal{N}(D_{co}) \subset \mathcal{N}(A)$. We define U_1, U_2 through the singular value decomposition of A:

$$A = [U_1, U_2] \begin{pmatrix} 0 & 0 \\ 0 & T_2 \end{pmatrix} [U_1, U_2]^T = U_2 T_2 U_2^T,$$

where the vectors of U_1 span the null space of A, and those of U_2 its range. The diagonal matrix T_2 contains the singular values of A. The next lemma then derives an expression for $(A + \epsilon E)^{-1}$ in terms of a power series in ϵ, for all ϵ sufficiently small.

Lemma A.8. *Let U_1, U_2, A, and E be defined as above. Define*

$$E_1 \equiv U_1^T E U_1,$$

and define the projections

$$P \equiv I - EU_1E_1^{-1}U_1^T \text{ and } Q \equiv I - U_1E_1^{-1}U_1^T E.$$

Also, introducing $\tilde{A} = U_2(T_2)^{-1}U_2^T$, define

$$X_{-1} \equiv U_1E_1^{-1}U_1^T \text{ and } X_k \equiv (-1)^k Q\tilde{A}P(E\tilde{A}P)^k.$$

Then

$$(A+\epsilon E)^{-1} = X \equiv \epsilon^{-1}X_{-1} + \sum_{k=0}^{\infty} \epsilon^k X_k, \tag{A.23}$$

and the convergence is uniform in ϵ for all ϵ sufficiently small.

Proof. First we show that the matrix E_1 is nonsingular. Suppose $E_1p = 0$ for some vector p. Then from the form of E_1, we must have that $U_1^T(C_x^T\hat{\Sigma}^{-1}\Delta r C_x + C_y^T\hat{\Sigma}^{-1}\Delta r C_y + C_z^T\hat{\Sigma}^{-1}\Delta r C_z)U_1p = 0$, and in turn that $C_xU_1p = 0$, $C_yU_1p = 0$, and $C_zU_1p = 0$. However, the description of $\mathcal{N}(A)$ in [5] show that the range of U_1 and the null space of the C_x, C_y, and C_z matrices intersect only trivially. Hence, $p = 0$ and E_1 is nonsingular.

Next, it is clear that the series in (A.23) converges uniformly in ϵ for any $\epsilon \leq \epsilon_{max} < 1/\|EAP\|_2$. To check that X in (A.23) is the inverse of $A + \epsilon E$, note that since the series converges in norm (i.e., absolutely), we can write

$$\begin{aligned}(A+\epsilon E)X - I &= \epsilon^{-1}AX_{-1} + (AX_0 + EX_{-1} - I)\\ &\quad + \sum_{k=0}^{\infty}\epsilon^{k+1}(AX_{k+1} + EX_k).\end{aligned} \tag{A.24}$$

From the definition of X_{-1}, it is immediate that $AX_{-1} = 0$. Moreover,

$$AX_0 + EX_{-1} - I = AQ\tilde{A}P + EU_1E_1^{-1}U_1^T - I = AQ\tilde{A}P - P.$$

Since $AQ = A$, we have $AQ\tilde{A}P - P = (A\tilde{A} - I)P$. But $U_1U_1^T + U_2U_2^T = I$ and $A\tilde{A} = U_2U_2^T$. Thus $(A\tilde{A} - I)P = -U_1U_1^TP = 0$ by definition of P, and the second term in (A.24) vanishes. For the remaining terms in (A.24),

$$\begin{aligned}AX_{k+1} + EX_k &= (-1)^{k+1}AQ\tilde{A}P(E\tilde{A}P)^{k+1} + (-1)^k EQ\tilde{A}P(E\tilde{A}P)^k\\ &= (-1)^k(-AQ\tilde{A}P + P)(E\tilde{A}P)^{k+1}\\ &= 0.\end{aligned}$$

using the fact that $EQ = PE$. Thus, all the terms in (A.24) vanish, and the lemma is proved. □

Using Lemma A.7, (A.22), and Lemma A.8, we obtain in the asymptotic diffusion limit, provided R verifies $V^TR = 0$ with V defined in (A.21),

$$D_{co}^{+}R = \left(\epsilon^{-1}\hat{X}_{-1} + \sum_{k=0}^{\infty}\epsilon^k\hat{X}_k\right)R, \tag{A.25}$$

where $\hat{X}_{-1} = (2\alpha)^{-1}X_{-1}$ and $\hat{X}_k = (2\alpha)^{-1}X_k$.

We now go back to (A.19) and (A.20).

LEMMA A.9. *Assume that* $\Sigma = \epsilon^{-1}\hat{\Sigma}$. *There exist positive constants* c_1 *and* c_2 *such that*

$$\|D_{co}^{+} S^T \Sigma \Delta r (I - \Gamma_0) S\| \leq c_1$$

and

$$\|D_{co}^{+} w_\ell \left\{ C_\ell^T \Delta r \Sigma^{-1} C_\ell + |\mu_\ell| A \right\}\| \leq c_2$$

for all $\hat{\Sigma} > 0$ *and* $\epsilon \to 0$.

Proof. It was shown in [5] that $\mathcal{N}(D_{co}) \subset \mathcal{N}(S)$, $\mathcal{N}(D_{co}) \subset \mathcal{N}(C_x)$, $\mathcal{N}(D_{co}) \subset \mathcal{N}(C_y)$, $\mathcal{N}(D_{co}) \subset \mathcal{N}(C_z)$, and $\mathcal{N}(D_{co}) \subset \mathcal{N}(A)$. Consequently, $V^T S^T = 0$, $V^T C_l^T = 0$, and $V^T A = 0$ so that we can apply (A.25). With the asymptotic diffusion limit assumptions, we obtain

$$\begin{aligned} D_{co}^{+} S^T \Sigma \Delta r (I - \Gamma_0) S &= \left(\epsilon^{-1} \hat{X}_{-1} + \sum_{k=0}^{\infty} \epsilon^k \hat{X}_k \right) \epsilon S^T \hat{\Sigma} \Delta r \hat{\Gamma}_0 S \\ &= \hat{X}_{-1} S^T \hat{\Sigma} \Delta r \hat{\Gamma}_0 S + O(\epsilon), \end{aligned}$$

and, using also (A.22),

$$\begin{aligned} & D_{co}^{+}(w_\ell \{ C_\ell^T \Delta r \Sigma^{-1} C_\ell + |\mu_\ell| A \}) \\ &\quad = \left(\epsilon^{-1} \hat{X}_{-1} + \sum_{k=0}^{\infty} \epsilon^k \hat{X}_k \right) (w_\ell \{ C_\ell^T \Delta r \epsilon \hat{\Sigma}^{-1} C_\ell + |\mu_\ell| A \}) \\ &\quad = \hat{X}_{-1} \left(w_\ell \left\{ C_\ell^T \Delta r \hat{\Sigma}^{-1} C_\ell + \frac{|\mu_\ell|}{2\alpha} V V^T - |\mu_\ell| E \right\} \right) + w_\ell \frac{|\mu_\ell|}{2\alpha} D_{co}^{+} D_{co} + O(\epsilon). \end{aligned}$$

Since $D_{co}^{+} D_{co}$ is a projection, its norm is bounded. Thus both terms remain bounded as $\epsilon \to 0$. ◻

We are now prepared to conclude the proof of Theorem 6.2.

Proof of Theorem 6.2. From (A.18) and Lemma A.9 we can write

$$\begin{aligned} \|C_0 A_0 - I\|_2 \leq & \frac{1}{4\pi} \|F(\Sigma^{-1}) \Gamma_0\|_2 + \frac{1}{4\pi} \|S\|_2 \sum_{k=2}^{\infty} \sum_{\ell=1}^{L} c_2 \|V_\ell\|_2 \, \|\Sigma^{-1} C_\ell V_\ell\|_2^{k-1} \, \|\Gamma_0\|_2 \\ & + \frac{1}{4\pi} \|S\|_2 \, c_1 \, \|S^T (S S^T)^{-1}\|_2 \, \|F(\Sigma^{-1})\|_2 \|\Gamma_0\|_2. \end{aligned}$$

Since $F(\Sigma^{-1}) \to 0$ when $\Sigma^{-1} \to 0$ (A.7), the conclusion follows. ◻

For Theorem 6.3, it follows from (A.25) that the preconditioner C_0 given by (5.8) can be written, in the asymptotic diffusion limit,

$$\begin{aligned} C_0 &= I + S D_{co}^{+} S^T \epsilon^{-1} \hat{\Sigma} \Delta r (I - \epsilon^2 \hat{\Gamma}_0) \\ &= \epsilon^{-2} C_{0,-2} + \epsilon^{-1} C_{0,-1} + C_{0,0} + O(\epsilon), \end{aligned}$$

where

$$\begin{aligned} C_{0,-2} &= S \hat{X}_{-1} S^T \hat{\Sigma} \Delta r, \\ C_{0,-1} &= S \hat{X}_0 S^T \hat{\Sigma} \Delta r, \text{ and} \\ C_{0,0} &= I - S \hat{X}_{-1} S^T \hat{\Sigma} \Delta r \hat{\Gamma}_0 + S \hat{X}_1 S^T \hat{\Sigma} \Delta r. \end{aligned}$$

Then, the proof of Theorem 6.3 parallels almost literally the one of Theorem 6.15 in [4]. We therefore do not repeat it here.

A.2.2. Proof for the other thick limit. To prove Theorem 6.4, we need to introduce the following lemma, proved in [14].

LEMMA A.10. *Given two matrices, A and B, a necessary and sufficient condition that*

$$\lim_{B\to A} B^+ = A^+ \tag{A.26}$$

is that $\operatorname{rank}(B) = \operatorname{rank}(A)$ *as B approaches A.*

Also, with T_1 defined in (5.3), we introduce $T_{1,0}$ as the limit of T_1 for $\|\Sigma^{-1}\|_2 \to 0$, i.e.,

$$T_{1,0} = \begin{pmatrix} L_{x_0,0} & L_{x_0,1} \\ L_{y_0,0} & L_{y_0,1} \\ L_{z_0,0} & L_{z_0,1} \\ \tilde{\Gamma}_0 S^0 & 0 \\ 0 & S^1 \\ L_{x_M,0} & L_{x_M,1} \\ L_{y_J,0} & L_{y_J,1} \\ L_{z_K,0} & L_{z_K,1} \end{pmatrix}, \tag{A.27}$$

where $\tilde{\Gamma}_0 = I - \Gamma_0 \geq \epsilon I$. From the developments in appendix A.2 of [5], one can show that T_1 and $T_{1,0}$ have full rank. Also, their pseudo-inverse T_1^+ and $T_{1,0}^+$ are such that $T_1 T_1^+ = I$ and $T_{1,0}T_{1,0}^+ = I$.

Proof of Theorem 6.4. We have that $T_1 - T_{1,0} \to 0$ uniformly as $\|\Sigma^{-1}\|_2 \to 0$. Developments in [5] show that C_0 defined in (5.8) can equivalently be written as

$$C_0 = I + \begin{pmatrix} S & 0_{LA} \end{pmatrix} T_1^+ \begin{pmatrix} 0_{BC} \\ I \\ 0_{LA} \\ 0_{BC} \end{pmatrix} \Gamma_0, \tag{A.28}$$

where the zero matrix $0_{LA} \in \mathbf{R}^{3MJK \times MJK}$ corresponds to the linear anisotropy terms (vanishing here since we assume isotropic scattering), and the zero matrix $0_{BC} \in \mathbf{R}^{KJ+MJ+MK\times MJK}$ corresponds to the boundary conditions. Then we look at the difference

$$E_0 \equiv C_0 - (I - \Gamma_0)^{-1} = C_0 - \tilde{\Gamma}_0^{-1}.$$

We have

$$\tilde{\Gamma}_0 E_0 = \tilde{\Gamma}_0 C_0 - I = \tilde{\Gamma}_0 \begin{pmatrix} S & 0_{LA} \end{pmatrix} T_1^+ \begin{pmatrix} 0_{BC} \\ I \\ 0_{LA} \\ 0_{BC} \end{pmatrix} \Gamma_0 + \tilde{\Gamma}_0 - I.$$

Thus

$$E_0 = \tilde{\Gamma}_0^{-1} \left[\tilde{\Gamma}_0 \begin{pmatrix} S & 0_{LA} \end{pmatrix} T_1^+ \begin{pmatrix} 0_{BC} \\ I \\ 0_{LA} \\ 0_{BC} \end{pmatrix} - I \right] \Gamma_0.$$

Using Lemma A.10 and the fact that T_1 and $T_{1,0}$ have the same (full) rank, the bracketed factor in the last expression tends to

$$\left[\tilde{\Gamma}_0 \begin{pmatrix} S & 0_{LA} \end{pmatrix} T_{1,0}^{+} \begin{pmatrix} 0_{BC} \\ I \\ 0_{LA} \\ 0_{BC} \end{pmatrix} - I \right] \equiv E_*,$$

as $\|\Sigma^{-1}\|_2 \to 0$, uniformly for $\tilde{\Gamma}_0 \geq \epsilon I$. But $E_* = 0$. Indeed, for any $w \in \mathbf{R}^{zones}$ define

$$v \equiv \begin{pmatrix} v_0 \\ v_1 \end{pmatrix} \equiv T_{1,0}^{+} \begin{pmatrix} 0_{BC} \\ I \\ 0_{LA} \\ 0_{BC} \end{pmatrix} w \qquad (v_0 \in \mathbf{R}^{zones}, v_1 \in \mathbf{R}^{3*zones}).$$

Then multiplying both sides by $T_{1,0}$ yields $\tilde{\Gamma}_0 S^0 v_0 = w$ (from (A.27) and $T_{1,0}T_{1,0}^{+} = I$) so that $E_* w = \tilde{\Gamma}_0 S^0 v_0 - w = 0$. Thus $E_0 \to 0$ as $\|\Sigma^{-1}\|_2 \to 0$, uniformly for $\tilde{\Gamma}_0 \geq \epsilon I$. □

A.3. Proof for the thin regime. We can here go straight to the proof.

Proof of Theorem 6.7. From (5.8) and Lemma A.7, we get

$$C_0 = I + S(D_{co} + \epsilon^{-1} V V^T)^{-1} S^T \Sigma \Delta r \Gamma_0.$$

From the definition (5.5) of D_{co}, we see that, for $\|\Sigma\|_2 \to 0$,

$$D_{co} \to \frac{1}{3}\left(C_x^T \Delta r \Sigma^{-1} C_x + C_y^T \Delta r \Sigma^{-1} C_y + C_z^T \Delta r \Sigma^{-1} C_z \right).$$

Thus, for $\Sigma = \epsilon \hat{\Sigma}$ and $\epsilon \to 0$,

$$C_0 - I \to \epsilon^2 S \left(\frac{1}{3} (C_x^T \Delta r \hat{\Sigma}^{-1} C_x + C_y^T \Delta r \hat{\Sigma}^{-1} C_y + C_z^T \Delta r \hat{\Sigma}^{-1} C_z) + V V^T \right)^{-1} S^T \hat{\Sigma} \Delta r \Gamma_0,$$

which leads to the conclusion for $0 \leq \Gamma_0 \leq I$. □

Acknowledgments. The author acknowledges the participation of Peter N. Brown in this paper.

REFERENCES

[1] M. Adams, *Subcell balance methods for radiative transfer on arbitrary grids*, Transport Theory Statist. Phys., 26 (1997), pp. 385–431.

[2] M. Adams and E. Larsen, *Fast iterative methods for discrete-ordinates particle transport calculations*, Progress in Nuclear Energy, 40 (2002), pp. 3–159.

[3] R. E. Alcouffe, *Diffusion synthetic acceleration methods for the diamond-differenced discrete ordinates equations*, Nucl. Sci. Eng., 64 (1977), pp. 344–355.

[4] S. F. Ashby, P. N. Brown, M. R. Dorr, and A. C. Hindmarsh, *A linear algebraic analysis of diffusion synthetic acceleration for the Boltzmann transport equation*, SIAM J. Numer. Anal., 32 (1995), pp. 128–178.

[5] P. N. Brown, *A linear algebraic development of diffusion synthetic acceleration for three-dimensional transport equations*, SIAM J. Numer. Anal., 32 (1995), pp. 179–214.

[6] Z. Cai, J. Mandel, and S. McCormick, *The finite volume element method for diffusion equations on general triangulations*, SIAM J. Numer. Anal., 28 (1991), pp. 392–402.

[7] V. Faber and T. A. Manteuffel, *A look at transport theory from the point of view of linear algebra*, in Transport Theory, Invariant Imbedding, and Integral Equations, P. Nelson et al., eds., Marcel Dekker, New York, 1989, pp. 37–61.

[8] G. H. Golub and C. F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press, Baltimore, MD, 1983.

[9] R. Horn and C. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1991.

[10] E. Larsen and J. Morel, *Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes* II, J. Comput. Phys., 83 (1989), pp. 212–236.

[11] E. Larsen and J. Morel, *Corrigendum on asymptotic solutions of numerical transport problems in optically thick, diffusive regimes* II, J. Comput. Phys., 91 (1990), p. 246.

[12] E. W. Larsen, *Unconditionally stable diffusion-synthetic acceleration methods for the slab geometry discrete ordinates equations. Part* I*: Theory*, Nucl. Sci. Eng., 82 (1982), pp. 47–63.

[13] E. Lewis and W. Miller, Jr., *Computational Methods of Neutron Transport*, John Wiley & Sons, New York, 1984.

[14] G. W. Stewart, *On the perturbation of pseudo-inverses, projections and linear least squares problems*, SIAM Rev., 19 (1977), pp. 634–662.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2060–2076

ON THE ADAPTIVE SELECTION OF THE PARAMETER IN REGULARIZATION OF ILL-POSED PROBLEMS*

SERGEI PEREVERZEV[†] AND EBERHARD SCHOCK[‡]

Abstract. We study the possibility of using the structure of the regularization error for a posteriori choice of the regularization parameter. As a result, a rather general form of a selection criterion is proposed, and its relation to the heuristical quasi-optimality principle of Tikhonov and Glasko [*Z. Vychisl. Mat. Mat. Fiz.*, 4 (1964), pp. 564–571] and to an adaptation scheme proposed in a statistical context by Lepskii [*Theory Probab. Appl.*, 36 (1990), pp. 454–466] is discussed. The advantages of the proposed criterion are illustrated by using such examples as self-regularization of the trapezoidal rule for noisy Abel-type integral equations, Lavrentiev regularization for nonlinear ill-posed problems, and an inverse problem of the two-dimensional profile reconstruction.

Key words. inverse problems in Banach spaces, parameter choice, Abel integral equations, Lavrentiev regularization for equations with monotone operators, scattering, profile reconstruction

AMS subject classification. 65J22

DOI. 10.1137/S0036142903433819

1. Introduction.

How do we choose a posteriori a suitable value for the regularization parameter in ill-posed problems without knowledge about the solution's smoothness that may not be accessible? This question is discussed extensively in the regularization theory. A first a posteriori rule of choice is described in the paper by Phillips [21], which predates even Tikhonov's paper [26] recognized as a reference point of regularization theory.

We define an operator equation

$$Ax = y \tag{1.1}$$

with a linear operator $A \in \mathcal{L}(X, Y)$ between Banach spaces X and Y as essentially ill-posed if the range $R(A)$ of A is not closed in Y. If A is invertible, this nonclosed range is associated with the discontinuity of the inverse operator A^{-1}. In general, the best approximate solution $A^{+}y$, where A^{+} is the Moore–Penrose inverse of A, does not depend continuously on the right-hand side y. Since in practice data will almost never be available exactly, because of measurement error, one has to be aware of numerical instabilities when a noisy observation $y_\delta \in Y$ instead of y with

$$\|y - y_\delta\|_Y \le \delta \tag{1.2}$$

is known. Hence, in order to approximate $A^{+}y$ in a stable way, regularization methods should be applied. In general, regularization methods for the solution of (1.1) replace the generalized inverse A^{+} by a family of continuous operators R_α, which converge pointwise to A^{+}. The standard regularization methods have in common that the

[*] Received by the editors August 28, 2003; accepted for publication (in revised form) October 27, 2004; published electronically December 15, 2005.

http://www.siam.org/journals/sinum/43-5/43381.html

[†] Johann-Radon Institute for Computational and Applied Mathematics, Altenbergerstraße 69, A-4040 Linz, Austria (sergei.pereverzyev@oeaw.ac.at).

[‡] Department of Mathematics, University of Kaiserslautern, P.O. Box 3049, D-67653 Kaiserslautern, Germany (schock@mathematik.uni-kl.de).

approximation error $\|A^+y - R_\alpha y\|_X$ is monotonically decreasing for decreasing α-values. In general, it is natural to assume that there exists an increasing continuous function $\varphi(\alpha) = \varphi(\alpha; A, y)$ such that $0 = \varphi(0) \leq \varphi(\alpha) \leq 1$, and

$$\|A^+y - R_\alpha y\|_X \leq \varphi(\alpha). \tag{1.3}$$

This property is no longer true for the regularization error $\|A^+y - R_\alpha y_\delta\|_X$. The regularized solutions $R_\alpha y_\delta$ converge to A^+y as $\delta \to 0$ only if the regularization parameter α is properly chosen dependent upon the noise level and possibly upon the data, i.e., $\alpha = \alpha(\delta, y_\delta)$. There are several methods that have been proposed and used for the a posteriori choice of the regularization parameter α as a function of the noise level and the data. These include the discrepancy principle (DP) originally proposed by Phillips [21] and later reinvented by Morozov [20] and Marti [18], a method developed by Gferer [9], Engl and Gfrerer [8], and Raus [23], which is sometimes called the minimum-bound (MB) method [17], and the monotone error rule (ME) proposed recently by Tautenhahn and Hämarik [25]. The MB and ME methods have been designed for ill-posed problems in Hilbert spaces. The DP method is more universal, because Plato (see, e.g., [22]) demonstrated that the DP method can be successfully applied to problems in Banach spaces. However, the DP method does not provide the best order of approximation for all problems, which could be, in principle, treated by a fixed regularization method with optimal order of accuracy; see, e.g., [11]. The MB and ME methods are free from this drawback of the discrepancy principle, but a disadvantage of both methods is that they require the knowledge of an additional approximate solution obtained within the framework of the regularization method of higher qualification. For example, Tautenhahn and Hämarik [25] select the regularization parameter for the ordinary Tikhonov regularization by constructing an additional approximate solution using iterated Tikhonov regularization; i.e., another regularization method should be involved in the choice procedure, and that is not always reasonable.

At the same time the structure of regularization error is very similar to the loss function of statistical estimation, where some parameter always controls the trade-off between the bias and the variance of the risk. It gives a hint that the statistical art of bias-variance balancing can be used for choosing the regularization parameter.

Indeed, the regularization error can be estimated by

$$\|A^+y - R_\alpha y_\delta\|_X \leq \|A^+y - R_\alpha y\|_X + \|R_\alpha y - R_\alpha y_\delta\|_X, \tag{1.4}$$

where the first term on the right-hand side is an approximation error, whereas the second term is a stability bound on the regularizing operator R_α. If R_α possesses a locally uniformly bounded Fréchet derivative R'_α in a ball of radius δ around the exact free term y then

$$\|R_\alpha y - R_\alpha y_\delta\|_X \leq \delta \|R'_\alpha\|_{Y \to X} + o(\delta).$$

For linear problems (1.1) R_α is usually linear, and $R'_\alpha = R_\alpha$. Keeping in mind that $\{R_\alpha\}$ approximates the unbounded Moore–Penrose inverse A^+, it is easy to realize that $\|R_\alpha\|$ (or $\|R'_\alpha\|$) should increase for $\alpha \to 0$. Thus, there exists an increasing continuous function $\lambda(\alpha)$ such that $\lambda(0) = 0$, and

$$\|R_\alpha y - R_\alpha y_\delta\| \leq \frac{\delta}{\lambda(\alpha)}. \tag{1.5}$$

For each regularization method $\lambda(\alpha)$ is known, or at least it can be estimated effectively. For the standard regularization methods $\lambda(\alpha) = \gamma\sqrt{\alpha}$, where γ is a known constant. Another forms of $\lambda(\alpha)$ will be discussed later.

Thus, from (1.3)–(1.5) it follows that

$$\|A^+y - R_\alpha y_\delta\|_X \leq \varphi(\alpha) + \frac{\delta}{\lambda(\alpha)}. \tag{1.6}$$

Almost all existing results about the accuracy of regularization methods are asymptotic results in δ. These results indicate that a choice of

$$\alpha = \alpha_{\text{opt}} = (\varphi\lambda)^{-1}(\delta), \tag{1.7}$$

that balances $\varphi(\alpha)$ with $\frac{\delta}{\lambda(\alpha)}$, leads to the error estimate

$$\|A^+y - R_{\alpha_{\text{opt}}} y_\delta\|_X \leq 2\varphi((\varphi\lambda)^{-1}(\delta)), \tag{1.8}$$

which has at least optimal order with respect to δ. Unfortunately, an a priori parameter choice (1.7) can seldom be used in practice because the smoothness properties of the unknown solution A^+y reflected in function φ from (1.3) are generally unknown. On the other hand, an error estimate (1.8) can be considered as a benchmark for a posteriori parameter choice strategies, because it indicates the order of accuracy that cannot be beaten by any of them within the framework of assumptions (1.3), (1.5). The outline of the paper is as follows. In the next section our focus will be on the question of how to adapt the regularization parameter to the unknown smoothness in such a way that the optimal order of accuracy (1.8) would be reached automatically. We shall present two adaptive procedures solving this question. Then in section 3 the advantages of the proposed procedures will be illustrated by using several examples of linear and nonlinear ill-posed problems. We close this paper with a short conclusion.

2. General theorems.

In practical applications, different regularization parameters α_i are often selected from some finite set

$$\Delta_N = \{\alpha_i : 0 < \alpha_0 < \alpha_1 < \cdots < \alpha_N\}$$

and the corresponding regularization solutions

$$x^\delta_{\alpha_i} = R_{\alpha_i} y_\delta, \quad i = 1, 2, \ldots, N,$$

are studied online. In view of the representation

$$\alpha_{\text{opt}} = \max\left\{\alpha : \varphi(\alpha) \leq \frac{\delta}{\lambda(\alpha)}\right\}$$

the optimal choice of α_i from Δ_N is

$$\alpha_* = \alpha_\ell = \max\{\alpha_i : \alpha_i \in M(\Delta_N)\},$$

where

$$M(\Delta_N) := \left\{\alpha_i : \alpha_i \in \Delta_N,\ \varphi(\alpha_i) \leq \frac{\delta}{\lambda(\alpha_i)}\right\}.$$

But if φ is unknown, such a choice is also not feasible. At the same time, for any $\alpha_i, \alpha_j,\ \alpha_i \geq \alpha_j$, from the set $M(\Delta_N)$, containing α_* as an upper bound, the estimation of the norm $\|x^\delta_{\alpha_i} - x^\delta_{\alpha_j}\|$ does not require knowledge of φ. Indeed, due to the monotonicity of $\varphi(\alpha)$, $\lambda(\alpha)$ from (1.6), it follows that

$$\begin{aligned}\|x^\delta_{\alpha_i} - x^\delta_{\alpha_j}\| &\leq \|A^+y - R_{\alpha_i}y_\delta\| + \|A^+y - R_{\alpha_j}y_\delta\|\\ &\leq \varphi(\alpha_i) + \varphi(\alpha_j) + \frac{\delta}{\lambda(\alpha_i)} + \frac{\delta}{\lambda(\alpha_j)}\\ &\leq 2\varphi(\alpha_i) + \frac{\delta}{\lambda(\alpha_i)} + \frac{\delta}{\lambda(\alpha_j)}\\ &\leq \frac{4\delta}{\lambda(\alpha_j)}.\end{aligned}$$

This gives a hint that the upper bound of the subset

$$M^+(\Delta_N) := \left\{\alpha_i \in \Delta_N : \|x^\delta_{\alpha_i} - x^\delta_{\alpha_j}\| \leq \frac{4\delta}{\lambda(\alpha_j)},\ j = 0, 1, 2, \dots, i\right\} \tag{2.1}$$

should be sufficiently close to a desirable value α_*. The following proposition justifies this conjecture on α_i.

Theorem 2.1. *Let $\Delta_N = \Delta_N^{\lambda,q}$ be such that $M(\Delta_N) \neq \varnothing$, $\Delta_N \backslash M(\Delta_N) \neq \varnothing$, and for any $\alpha_i \in \Delta_N$, $i = 1, 2, \dots, N$,*

$$\lambda(\alpha_i) \leq q\lambda(\alpha_{i-1}), \tag{2.2}$$

where q is some fixed constant. Then under the assumptions (1.2), (1.3), (1.5) *for $\alpha_+ = \alpha_k \in \Delta_N$ chosen as*

$$\alpha_+ = \max\{\alpha_i : \alpha_i \in M^+(\Delta_N)\} \tag{2.3}$$

the following estimate holds:

$$\|A^+y - x^\delta_{\alpha_+}\| \leq 6q\varphi((\varphi\lambda)^{-1}(\delta)). \tag{2.4}$$

Proof. From the definition of $\alpha_* = \alpha_\ell$ we have that for $\alpha_{\ell+1} > \alpha_\ell$

$$\varphi(\alpha_{\ell+1})\lambda(\alpha_{\ell+1}) > \delta = \varphi(\alpha_{\text{opt}})\lambda(\alpha_{\text{opt}}),$$

and using the monotonicity of $\varphi(\alpha)$, $\lambda(\alpha)$ we deduce $\alpha_{\ell+1} > \alpha_{\text{opt}}$. Then under our hypothesis

$$\lambda(\alpha_{\text{opt}}) < \lambda(\alpha_{\ell+1}) \leq q\lambda(\alpha_\ell) = q\lambda(\alpha_*).$$

Hence

$$\frac{\delta}{\lambda(\alpha_*)} \leq q\frac{\delta}{\lambda(\alpha_{\text{opt}})}. \tag{2.5}$$

As already shown above, $M(\Delta_N) \subset M^+(\Delta_N)$, and therefore

$$\alpha_* = \alpha_\ell = \max\{\alpha_i \in M(\Delta_N)\} \leq \alpha_+ = \alpha_k = \max\{\alpha_i \in M^+(\Delta_N)\}.$$

From the definition of $M^+(\Delta_N)$ and (2.5) we conclude

$$\begin{aligned}\|A^+y - x^\delta_{\alpha_+}\| &= \|A^+y - x^\delta_{\alpha_k}\| \leq \|A^+y - x^\delta_{\alpha_\ell}\| + \|x^\delta_{\alpha_\ell} - x^\delta_{\alpha_k}\| \\ &\leq \varphi(\alpha_\ell) + \frac{\delta}{\lambda(\alpha_\ell)} + \frac{4\delta}{\lambda(\alpha_\ell)} \leq \frac{6\delta}{\lambda(\alpha_*)} \leq 6q\frac{\delta}{\lambda(\alpha_{\text{opt}})} \\ &= 6q\varphi((\varphi\lambda)^{-1}(\delta)),\end{aligned}$$

and the theorem is proved. □

If we would know in advance that the function $\varphi(\alpha)$ reflecting the smoothness properties of the unknown solution A^+y, then we may achieve the accuracy of the optimal order given in (1.8). Comparing (1.8) with (2.4) we can conclude that the choice of the regularization parameter $\alpha = \alpha_+$ is also order optimal in the sense of accuracy. We would like to stress, however, that the selection criterion (2.1), (2.3) producing α_+ is adaptive to the unknown smoothness, because φ is not involved in its construction. Observe, that α_+ depends only on the noisy data y_δ, on the noise level δ, and on the discrete set $\Delta_N = \Delta_N^{\lambda,q}$ which should meet the conditions of Theorem 2.1. The conditions $M(\Delta_N) \neq \varnothing$, $\Delta_N \backslash M(\Delta_N) \neq \varnothing$ are rather natural. It is satisfied if, for example, $\alpha_0 = \lambda^{-1}(\delta) \in \Delta_N$, $\alpha_N = \lambda^{-1}(1) \in \Delta_N$. The condition (2.2) is also not so restrictive. Recall that for the standard regularization methods $\lambda(\alpha) = \gamma\sqrt{\alpha}$. Then to meet (2.2) one can take Δ_N in the form of a geometric sequence

$$\Delta_N = \{\alpha_i : \alpha_i = \mu^i \alpha_0,\ i = 0, 1, \dots, N\} \tag{2.6}$$

with $\mu = q^2 > 1$.

We remark that the first time a geometric sequence was used as a set of regularization parameters in the papers by Tikhonov and Glasko [27, 28], where a method of choosing a paramter $\alpha_T = \alpha_m = \mu^m \alpha_0$ from such a sequence, termed quasi-optimality criterion, was suggested for which

$$\sigma(\alpha_i) := \|x^\delta_{\alpha_i} - x^\delta_{\alpha_{i-1}}\| \tag{2.7}$$

has the minimum value $\sigma(\alpha_m)$ in the chosen set (2.6). It is worth to mention that this quasi-optimality criterion is chronologically the first in the class of the heuristically motivated regularization parameter choice rules that seek to avoid any a priori knowledge of the noise level δ. There is, however, a negative result of Bakushinskii [1], which tells us that no convergence theory and error estimates as above can exist for noise level-free rules, and for the quasi-optimality criterion in particular.

At the same time the quasi-optimality criterion gives a hint that the quantities (2.7) can be used as indicators for the order optimal regularization parameter choice. Indeed, if α_{i-1}, $\alpha_i = \mu\alpha_{i-1}$ belong to the set $M(\Delta_N)$ containing the optimal parameter value $\alpha = \alpha_*$ then the quantity (2.7) can be estimated as

$$\|x^\delta_{\alpha_i} - x^\delta_{\alpha_{i-1}}\| \leq \frac{4\delta}{\lambda(\alpha_{i-1})}. \tag{2.8}$$

The right-hand side of (2.8) is a decreasing function of α. Therefore, the largest $\alpha_i \in \Delta_N$ satisfying (2.8) cannot be far from α_T minimizing (2.7). This observation leads to the following noise level-dependent analog of the quasi-optimality criterion:

$$\overline{\alpha} = \max\left\{\alpha_j \in \Delta_N : \|x^\delta_{\alpha_i} - x^\delta_{\alpha_{i-1}}\| \leq \frac{4\delta}{\lambda(\alpha_{i-1})},\ i = 0, 1, 2, \dots, j\right\}. \tag{2.9}$$

THEOREM 2.2. *Assume* (1.2), (1.3), (1.5) *to hold. Assume, furthermore, that* $\lambda(\alpha)$ *from* (1.5) *obeys a strong* Δ_2*-condition, i.e., there are* $\kappa_1 > \kappa > 1$ *such that for any* $\alpha > 0$, $\lambda(2\alpha)/\kappa_1 \leq \lambda(\alpha) \leq \lambda(2\alpha)/\kappa$. *If the geometric sequence* (2.6) *meets the condition of Theorem* 2.1, *then*

$$\|A^+y - x^\delta_{\overline{\alpha}}\| \leq c\varphi((\varphi\lambda)^{-1}(\delta)), \tag{2.10}$$

where the constant c depends only on q, κ, κ_1, μ.

Proof. Let $\overline{\alpha} = \alpha_m \in \Delta_N$. From (2.1), (2.3), and (2.9) it follows that $\overline{\alpha} \geq \alpha_+$. Then, as in the proof of Theorem 2.1, one can deduce $\overline{\alpha} = \alpha_m \geq \alpha_+ = \alpha_k \geq \alpha_* = \alpha_\ell$, and using the triangle inequality successively, we arrive at

$$\begin{aligned}\|A^+y - x^\delta_{\overline{\alpha}}\| &\leq \|A^+y - x^\delta_{\alpha_*}\| + \sum_{i=\ell+1}^{m} \|x^\delta_{\alpha_{i-1}} - x^\delta_{\alpha_i}\| \\ &\leq \|A^+y - x^\delta_{\alpha_*}\| + \sum_{i=\ell+1}^{m} \frac{4\delta}{\lambda(\alpha_{i-1})} \\ &\leq \|A^+y - x^\delta_{\alpha_*}\| + \sum_{\nu=0}^{m-\ell-1} \frac{4\delta}{\lambda(\alpha_*\mu^\nu)}.\end{aligned}$$

On the other hand, for any $\mu > 1$, $b > 1$ and integers n, j such that $2^n \leq \mu \leq 2^{n+1}$, $2^j \leq b \leq 2^{j+1}$ iterating the strong Δ_2-condition, if necessary, one obtains

$$\begin{aligned}&\frac{1}{\lambda(b\alpha_*)} \leq \frac{1}{\lambda(2^j\alpha_*)} \leq \frac{1}{\kappa^j\lambda(\alpha_*)} \leq \frac{\kappa}{\kappa^{\log_2 b}\lambda(\alpha_*)}; \\ &\lambda(\alpha_i) = \lambda(\alpha_{i-1}\mu) \leq \kappa_1^{n+1}\lambda(\alpha_{i-1}) \leq \kappa_1^{\log_2 2\mu}\lambda(\alpha_{i-1}).\end{aligned}$$

The last inequality means that (2.2) is satisfied with $q = \kappa_1^{\log_2 2\mu}$. Using these observations and (2.5) we conclude

$$\begin{aligned}\|A^+y - x^\delta_{\overline{\alpha}}\| &\leq \varphi(\alpha_*) + \frac{\delta}{\lambda(\alpha_*)} + \frac{4\kappa\delta}{\lambda(\alpha_*)} \sum_{\nu=0}^{m-\ell-1} \left(\frac{1}{\kappa^{\log_2 \mu}}\right)^\nu \\ &\leq \frac{\delta}{\lambda(\alpha_*)}\left[2 + \frac{4\kappa^{\log_2 2\mu}}{\kappa^{\log_2^{\mu}-1}}\right] = \frac{c_1\delta}{\lambda(\alpha_*)} \\ &\leq \frac{c_1\kappa_1^{\log_2 2\mu}}{\lambda(\alpha_{\mathrm{opt}})}\delta = c\varphi((\varphi\lambda)^{-1}(\delta)).\end{aligned}$$

The theorem is proved. □

At first glance the rule (2.9) looks like a simplified version of (2.1), (2.3), because it requires us to compare the regularized solutions $x^\delta_{\alpha_i}$ corresponding to parameters with adjacent numbers only. But as has been mentioned above, there are two different ideas behind these rules. The rule (2.9) is related to the heuristical quasi-optimality criterion. Up to a certain extent it supports heuristic theoretically. Moreover, numerical tests from [12] show that in some important particular cases both these criteria give the same value of regularization parameter. At the same time the rule (2.1), (2.3) has a statistical root. This rule was first studied in the paper [15] by Lepskii, devoted to statistical estimation from direct white noise observations that corresponds to (1.1) with identity operator A, but with random noisy data y_δ. Since then many authors have adopted this approach toward various statistical applications, we mention only [10, 5, 29], where the same idea has been realized in the context of ill-posed problems of the form (1.1) with compact operator A, but still with random noise. Deterministic noise model (1.2) allows to improve the order of accuracy of the regularized solution, as has been shown in [12, 4, 19] for the Hilbert space setting. Theorems 2.1 and 2.2 contain all these results as particular cases. Moreover, these theorems provide an uniform approach to order-optimal regularization parameter choice for linear ill-posed problems in Banach spaces.

3. Examples. This section applies Theorems 2.1 and 2.2 to several new examples such as self-regularization of the trapezoidal rule in the Banach space of continuous functions, Lavrentiev regularization for nonlinear problems, and an inverse problem of profile reconstruction.

3.1. Example 1: Self-regularization of the trapezoidal rule for noisy Abel-type integral equations. Consider an equation of the form (1.1) with the Abel-type integral operator

$$Ax(t) = A_\beta x(t) := \int_0^t \frac{a(t,\tau)}{(t-\tau)^\beta} x(\tau) d\tau, \quad t \in [0,1], \tag{3.1}$$

in Banach spaces $X = Y = C = C_{[0,1]}$, where $a(t,\tau)$ is at least Lipschitz-continuous on $0 \le \tau \le t \le 1$, and

$$|a(t,t)| \ge a_0 > 0. \tag{3.2}$$

The parameter β satisfies $0 < \beta < 1$.

The trapezoidal-discretization method for (1.1), (3.1) has been intensively studied in [2, 30, 7]. It consists of replacing (1.1), (3.1) by a set of linear equations

$$\int_0^{\frac{i}{n}} \frac{a\left(\frac{i}{n},\tau\right)}{\left(\frac{i}{n}-\tau\right)^\beta} x(\tau) d\tau = y\left(\frac{i}{n}\right), \quad i = 1, 2, \dots, n.$$

Then one replaces each of them by means of discretizing the integral on the left as follows:

$$n \sum_{j=1}^{i} \int_{\frac{j-1}{n}}^{\frac{j}{n}} \frac{\left(\tau - \frac{j-1}{n}\right) a_{ij}^n x_{n,j} + \left(\frac{j}{n} - \tau\right) a_{ij-1}^n x_{n,j-1}}{\left(\frac{i}{n}-\tau\right)^\beta} d\tau = y\left(\frac{i}{n}\right), \quad i = 1, 2, \dots, n, \tag{3.3}$$

where $a_{ij}^n = a(\frac{i}{n}, \frac{j}{n})$, and $x_{n,j}$ denotes the numerical approximation to $x(\frac{j}{n})$, $j = 0, 1, 2, \dots, n$. Thus, (3.3) is a system of n equations in $n+1$ unknown. For starting value $x_{n,0}$ one can take

$$x(0) = \lim_{t \to 0} \frac{(1-\beta)}{a(0,0)} \frac{y(t)}{t^{1-\beta}},$$

which exists, whenever (1.1), (3.1) has a continuous solution, or, as in [7],

$$x_{n,0} = \frac{(1-\beta)}{a(0,0)} \left\{ 3g\left(\frac{1}{n}\right) - 3g\left(\frac{2}{n}\right) + g\left(\frac{3}{n}\right) \right\}, \tag{3.4}$$

with $g(t) = t^{\beta-1}y(t)$. This yields the following triangular system for the approximations $\bar{x}_n = (x_{n,1}, x_{n,2}, \dots, x_{n,n})^T$

$$\frac{n^{\beta-1}}{(1-\beta)(2-\beta)}\,\bar{A}_n\bar{x}_n = \bar{y}_n - \frac{n^{\beta-1}}{(1-\beta)(2-\beta)}\,\bar{b}_n, \tag{3.5}$$

where $\bar{y}_n = (y(\frac{1}{n}), y(\frac{2}{n}), \dots, y(1))^T$, $\bar{b}_n = (b_{n,1}, b_{n,2}, \dots, b_{n,n})^T$,

$$\begin{aligned}
b_{n,i} &= a\left(\frac{i}{n}, 0\right) x_{n,0}, \quad i = 1, 2, \dots, n,\\
(\bar{A}_n)_{i,j} &= \begin{cases} a^n_{i,j}\kappa_{i-j}, & 1 \le j \le i \le n,\\ 0, & \text{otherwise},\end{cases}\\
\kappa_\ell &= (\ell+1)^{2-\beta} - 2\ell^{2-\beta} + (\ell-1)^{2-\beta}, \quad \ell \ge 1,\\
\kappa_0 &= 1.
\end{aligned}$$

The question of the existence and uniqueness of a solution (3.5) is summarized in the following proposition.

PROPOSITION 3.1 (see [7]). *If $a(t,\tau)$ is Lipschitz-continuous on $0 \le \tau \le t \le 1$, then there is a constant $\tilde{c}_{\beta,a}$ depending on β and such that $\|(\bar{A}_n)^{-1}\|_\infty \le \tilde{c}_{\beta,a}$, i.e., for any $\bar{f}_n = (f_1, f_2, \dots, f_n)^T$, $\|\bar{f}_n\|_\infty := \max_i |f_i|$,*

$$\|(\bar{A}_n)^{-1}\bar{f}_n\|_\infty \le \tilde{c}_{\beta,a}\|\bar{f}_n\|_\infty.$$

Moreover, in [2] (see also [31]) the convergence of the trapezoidal-discretization method has been shown to hold when the solution $x(t)$ of (1.1), (3.1) has only Lipschitz continuity and the same conditions on $a(t,\tau)$ apply. It means that there exists an increasing continuous function $\psi_{a,\beta}(x;t)$ such that $\psi_{a,\beta}(x;0) = 0$ and

$$\max_{0\le i\le n}\left|x\left(\frac{i}{n}\right) - x_{n,i}\right| \le \psi_{a,\beta}\left(x, \frac{1}{n}\right). \tag{3.6}$$

Let us turn to the case of the noisy equation

$$Ax(t) = y_\delta(t), \tag{3.7}$$

where A has the form (3.1), and y_δ can be only element from $Y = C_{[0,1]}$ such that (1.2) holds.

The trapezoidal-discretization method can be applied directly to (3.7) if in (3.4), (3.5) $y(\frac{i}{n})$ will be replaced by $y_\delta(\frac{i}{n})$, $i = 1, 2, \dots, n$. Then from Proposition 3.1 it follows that there is always a unique solution $\bar{x}^\delta_n$ of the system

$$\frac{n^{\beta-1}}{(1-\beta)(2-\beta)}\,\bar{A}_n\bar{x}^\delta_n = \bar{y}^\delta_n - \frac{n^{\beta-1}}{(1-\beta)(2-\beta)}\,\bar{b}^\delta_n.$$

It is easy to see that $\|\bar{y}_n - \bar{y}^\delta_n\|_\infty \le \delta$,

$$|x_{n,0} - x^\delta_{n,0}| \le \frac{(1-\beta)}{|a(0,0)|}n^{1-\beta}\delta(3 + 3\cdot 2^{\beta-1} + 3^{\beta-1}) = c_{\beta,a,1}n^{1-\beta}\delta, \tag{3.8}$$

and

$$\|\bar{b}_n - \bar{b}^\delta_n\|_\infty \le \|a(\cdot,0)\|_C|x_{n,0} - x^\delta_{n,0}| \le c_{\beta,a,2}n^{1-\beta}\delta.$$

Thus, under the condition of Proposition 3.1, the following bound holds:

$$\begin{aligned}\max_{0\le i\le n}|x_{n,i}-x^{\delta}_{n,i}| &\le n^{1-\beta}\|(\bar{A}_n)^{-1}\|_{\infty}((1-\beta)(2-\beta)+c_{\beta,a,2})\delta \\ &\le c_{\beta,a,3}n^{1-\beta}\delta.\end{aligned}\tag{3.9}$$

Within the framework of trapezoidal-discretization method the approximate solution of (1.1), (3.1) can be taken as piecewise linear interpolation spline $x_n(t)$ with uniform interpolation knots such that $x_n(\frac{i}{n})=x_{n,i}$, $i=0,1,2,\dots,n$. If only noisy right-hand side $y_\delta(t)$ is available then such a spline will interpolate $x^{\delta}_{n,i}$ and have the form

$$x^{\delta}_n(t)=\sum_{i=0}^{n}x^{\delta}_{n,i}\ell_{n,i}(t),$$

where $\ell_{n,i}(t)$ are so-called fundamental linear splines with knots $\{\frac{i}{n}\}_{i=0}^{n}$ such that $\ell_{n,i}(t)\ge 0$ for $t\in[0,1]$, and $\ell_{n,i}(\frac{j}{n})=\delta_{ij}$. From (3.8), (3.9) it follows that

$$\begin{aligned}|x_n(t)-x^{\delta}_n(t)| &\le \sum_{i=0}^{n}|x_{n,i}-x^{\delta}_{n,i}|\ell_{n,i}(t)\le \max_{0\le i\le n}|x_{n,i}-x^{\delta}_{n,i}| \\ &\le n^{1-\beta}\delta\max\{c_{\beta,a,1},c_{\beta,a,3}\}=c_{\beta,a}n^{1-\beta}\delta.\end{aligned}\tag{3.10}$$

Let now $s_n(x;t)$ be a piecewise linear spline with knots $\{\frac{i}{n}\}_{i=0}^{n}$ interpolating the values $x(\frac{i}{n})$, $i=0,1,\dots,n$, of the solution (1.1), (3.1). It is well-known that

$$|x(t)-s_n(x;t)|\le c\omega_2\left(x;\frac{1}{n}\right),$$

where $\omega_2(x;h)$ is the second-order modulus of smoothness, $\omega_2(x;h)\to 0$, and c is some absolute constant. Using (3.6) this yields

$$\begin{aligned}|x(t)-x_n(t)| &\le |x(t)-s_n(x;t)|+|s_n(x;t)-x_n(t)| \\ &\le c\omega_2\left(x;\frac{1}{n}\right)+\sum_{i=0}^{n}\left|x\left(\frac{i}{n}\right)-x_{n,i}\right|\ell_{n,i}(t) \\ &\le c\omega_2\left(x;\frac{1}{n}\right)+\psi_{a,\beta}\left(x;\frac{1}{n}\right)=\varphi\left(\frac{1}{n}\right).\end{aligned}$$

Combining it with (3.10), we obtain

$$|x(t)-x^{\delta}_n(t)|\le\varphi\left(\frac{1}{n}\right)+c_{\beta,a}n^{1-\beta}\delta.\tag{3.11}$$

Here the function φ depends on the smoothness of the solution (1.1), (3.1) and usually is unknown. But (3.11) has the same form as (1.6), where $\alpha=\frac{1}{n}$, $\lambda(\alpha)=c^{-1}_{\beta,a}\alpha^{1-\beta}$. For such α and $\lambda(\alpha)$ we have $\Delta_N=\{\alpha_i=\frac{1}{N-i+1}\}_{i=0}^{N}$ and

$$\begin{aligned}M^{+}(\Delta_N)=\Bigg\{\alpha_i=\frac{1}{N-i+1} &: \|x^{\delta}_{N-i+1}-x^{\delta}_{N-j+1}\|_C\le 4\delta c_{\beta,a}(N-j+1)^{1-\beta}, \\ &j=0,1,2,\dots,i\Bigg\} \\ =\{n:\|x^{\delta}_n-x^{\delta}_m\|_C &\le 4\delta c_{\beta,a}m^{1-\beta},\ m=N+1,N,\dots,n\}.\end{aligned}$$

Hence the selection criterion (2.1), (2.3) can be written as

$$n_+ := \min\{n : \|x_n^\delta - x_m^\delta\|_C \leq 4\delta c_{\beta,a} m^{1-\beta},\ m = N+1, N, \dots, n\} \tag{3.12}$$

and the conditions of Theorem 2.1 are satisfied with $q = 2^{1-\beta}$, $N > (c_{\beta,a}\delta)^{\frac{1}{\beta-1}}$. Thus, we have Theorem 3.2.

THEOREM 3.2. *If $a(t,\tau)$ and the solution $x(t)$ of* (1.1), (3.1) *are Lipschitz-continuous on $0 \leq \tau \leq t \leq 1$, then for $N > (c_{\beta,a}\delta)^{\frac{1}{\beta-1}}$ and n_+ chosen as* (3.12)

$$|x(t) - x_{n+}^\delta(t)| \leq 6 \cdot 2^{1-\beta}\varphi((\varphi\lambda)^{-1}(\delta)),$$

where $\lambda(\alpha) = c_{\beta,a}^{-1}\alpha^{1-\beta}$, $\varphi(\alpha) = c\omega_2(x;\alpha) + \psi_{a,\beta}(x;\alpha)$, and $\psi_{a,\beta}$ is the function from (3.6).

Remark 3.3. We can indicate only one case when the order of the function φ is known. Namely, in [7] it has been shown that $\psi_{a,\beta}(x, \frac{1}{n}) = c_1 n^{-2}$ for all $\beta \in (0,1)$, and for $x(t)$, $a(t,\tau)$ having Lipschitz-continuous second derivatives. For such $x(t)$ $\omega_2(x; \frac{1}{n})$ has the best possible order $\omega_2(x; \frac{1}{n}) = c_2 n^{-2}$. Thus, in the considered case $\varphi(\frac{1}{n}) = c_3 n^{-2}$ and to balance both terms in (3.11), one should take $n = n_{\text{opt}} \asymp \delta^{\frac{1}{\beta-3}}$ that gives an accuracy of order $O(\delta^{\frac{2}{3-\beta}})$. Note that Theorem 3.2 gives the same order of accuracy automatically without knowledge of φ.

Theorem 3.2 shows that the regularization of ill-posed problem (1.1), (3.1) with noisy right-hand side y_δ can be achieved by just choosing the number of knots in the trapezoidal rule properly. This is called self-regularization. Self-regularization adapted to unknown smoothness in a Hilbert space has been discussed recently in [13, 12, 4]. To the best of our knowledge, Theorem 3.2 is the first example of adaptive self-regularization in Banach space.

3.2. Example 2: Lavrentiev regularization for nonlinear ill-posed problems with monotone operators.

Throughout this section we assume that $A : D(A) \to X$ is a nonlinear monotone operator with domain $D(A)$ in a real Hilbert space X. Monotonicity means that for all $x_1, x_2 \in D(A)$

$$\langle A(x_1) - A(x_2),\ x_1 - x_2\rangle \geq 0,$$

where $\langle\cdot,\cdot\rangle$ is the inner product associated with norm $\|\cdot\| = \|\cdot\|_X$.

We further assume throughout that the nonlinear equation

$$A(x) = y$$

has a solution x^+, but only a noisy data y_δ with a known noise level δ is available, i.e., $\|y - y_\delta\| \leq \delta$. We do not assume that x^+ depends continuously on the data. It means that the stable reconstruction of x^+ from the noisy equation

$$A(x) = y_\delta \tag{3.13}$$

requires the application of special regularization methods. In the well-known Tikhonov regularization method a regularized approximation x_α^δ is obtained by minimizing the functional

$$J_\alpha(x) = \|A(x) - y_\delta\|^2 + \alpha\|x - \bar{x}\|^2,$$

with some initial guess $\bar{x} \in X$ and some properly chosen regularization parameter $\alpha > 0$. If A is Fréchet-differentiable in some ball $B_\rho(x^+)$ of radius ρ around x^+, and

x_α^δ is an interior point of $D(A)$ then it can be found from the (nonlinear) normal equation of Tikhonov's functional $J_\alpha(x)$

$$[A'(x)]^*[A(x) - y^\delta] + \alpha(x - \bar{x}) = 0,$$

where $[A'(x)]^*$ is the adjoint of the Fréchet derivative $A'(x)$. As has been indicated in [16, 24], for the problems with monotone operators the least squares minimization (and hence the use of the Fréchet derivatives) can be avoided and one can use the simpler regularized equation

$$A(x) + \alpha(x - \bar{x}) = y_\delta \tag{3.14}$$

known as Lavrentiev regularization.

If $D(A) = X$ and $A(x)$ is a continuous operator, then as has been shown in [6, pp. 97, 100], the monotonicity implies that for $\alpha > 0$ the operator $F(x) = \alpha x + A(x)$, $x \in X$, is strongly monotone, and $F^{-1}(x) = (\alpha I + A)^{-1}(x)$ is Lipschitz with constant $\frac{1}{\alpha}$; i.e., for any $u, v \in X$

$$\|(\alpha I + A)^{-1}(u - v)\| \le \frac{1}{\alpha}\|u - v\|. \tag{3.15}$$

Applying Lavrentiev regularization one usually assumes that for pure data $y = A(x^+)$ it produces an approximate solution $x_\alpha = (\alpha I + A)^{-1}(y + \alpha\bar{x})$ converging to x^+ as $\alpha \to 0$. It means that there exists an increasing continuous function $\varphi(\alpha) = \varphi(x^+; \alpha)$ such that $\varphi(0) = 0$ and

$$\|x^+ - x_\alpha\| \le \varphi(\alpha). \tag{3.16}$$

THEOREM 3.4. *Let $A(x)$ be a continuous monotone operator in a real Hilbert space X. Consider $\Delta_N = \{\alpha_i = q^i\delta,\ i = 0, 1, \dots, N\}$, $q > 1$, $\alpha_N \simeq 1$, and $\bar{\alpha} = \max\{\alpha_j \in \Delta_N : \|x_{\alpha_i}^\delta - x_{\alpha_{i-1}}^\delta\| \le 4q^{1-i},\ i = 1, 2, \dots, j\}$, where $x_{\alpha_i}^\delta$ is the unique solution of* (3.14) *for $\alpha = \alpha_i$. Then under the assumption* (3.16)

$$\|x^+ - x_{\bar{\alpha}}^\delta\| \le \frac{(6q - 2)q}{q - 1}\varphi(x^+; \theta_\varphi^{-1}(\delta)), \tag{3.17}$$

where $\theta_\varphi(t) = \varphi(t)t$.

Proof. From (3.15) and (3.16), it follows that for any $\alpha > 0$

$$\begin{aligned}\|x^+ - x_\alpha^\delta\| &\le \|x^+ - x_\alpha\| + \|x_\alpha - x_\alpha^\delta\| \\ &\le \varphi(\alpha) + \|(\alpha I + A)^{-1}(y - y_\delta)\| \\ &\le \varphi(\alpha) + \frac{\delta}{\alpha}.\end{aligned}$$

Hence, error bound has the form (1.6) with $\lambda(\alpha) = \alpha$. It is easy to see that for such $\lambda(\alpha)$, all conditions of Theorem 2.2 are satisfied. Moreover, for $\lambda(\alpha) = \alpha$ the arguments from the proof can be simplified, and it gives an explicit form of the constant c near the optimal order. □

Remark 3.5. Lavrentiev regularization is usually studied under the assumption (3.16) with $\varphi(\alpha) = c\alpha^p$, $p \in (0, 1]$. For example, the case of unknown p has been discussed recently in [24], where it has been shown that for α chosen as the solution of the nonlinear equation

$$\|\alpha(\alpha I + A'(x_\alpha^\delta))^{-1}(A(x_\alpha^\delta) - y_\delta)\| = c_1\delta, \quad c_1 > 1,$$

one has

$$\|x^+ - x_\alpha^\delta\| \le c_p \delta^{\frac{p}{p+1}}.$$

The disadvantage of this a posteriori rule is that its combination with the Lavrentiev regularization (3.14) does not allow to avoid the use of the Fréchet derivatives. At the same time, an a posteriori rule presented in Theorem 3.4 is free from the above mentioned drawback and gives the same order of accuracy $0(\delta^{\frac{p}{p+1}})$ for $\varphi(\alpha) = c\alpha^p$.

Moreover, to our knowledge, the rule from Theorem 3.4 is the only one that allows to reach the best possible order of accuracy of Lavrentiev regularization automatically, and does not involve another regularization methods.

3.3. Example 3: Inverse problem of profile reconstruction in diffractive optics. The statement of the problem discussed in this section is borrowed from [3]. Let the profile of two-dimensional diffraction grating be described by the curve

$$\Lambda_f := \{(x_1, f(x_1)) : x_1 \in \mathbb{R}\}$$

with 2π-periodic function f. Let

$$\Omega_f := \{x = (x_1, x_2) : x_2 > f(x_1),\ x_1 \in \mathbb{R}\}$$

be filled with a material whose index of refraction k is some positive constant. Suppose that a plane wave given by

$$u^{in}(x) = \exp(i\alpha x_1 - i\beta x_2)$$

is incident on Λ_f from the top, where $\alpha = k \sin\theta$, $\beta = k\cos\theta$, and $\theta \in (-\frac{\pi}{2}, \frac{\pi}{2})$ is the incident angle. Then the scattering of this wave by Λ_f is modelled by the Dirichlet problem for the Helmholz equation

$$\Delta u + k^2 u = 0 \quad \text{in } \Omega_f, \qquad u = -u^{in} \quad \text{on } \Lambda_f \tag{3.18}$$

where the scattered field u is assumed to satisfy a radiation condition, i.e., u is composed of bounded outgoing plane waves:

$$u(x_1, x_2) = \sum_{n\in\mathbb{Z}} A_n \exp[i(n+\alpha)x_1 + i\beta_n x_2], \tag{3.19}$$

with $\beta_n = \sqrt{k^2 - (n+\alpha)^2} \in \mathbb{C}$, and the Rayleigh coefficients $A_n \in \mathbb{C}$. To exclude resonances one assumes that $\beta_n \ne 0$, $n \in \mathbb{Z}$.

The inverse problem of profile reconstruction is to recover the profile function f from the trace $u_b(x) = u(x, b)$ of the scattered field $u(x_1, x_2)$ on the line $x_2 = b$ for a given incident wave u^{in}. Without loss of generality we can assume that the unknown profile Λ_f lies above the line $x_2 = b_0$ and below $x_2 = b$, i.e.,

$$b_0 < f(x_1) < b, \quad x_1 \in \mathbb{R}. \tag{3.20}$$

Representing the scattered field as a single layer potential

$$u(x_1, x_2) = \int_0^{2\pi} z(t) G(x_1, x_2, t, 0) dt$$

with an unknown density function $z \in L_2(0, 2\pi)$ and the free space quasi-periodic Green function

$$G(x_1, x_2, y_1, y_2) = \frac{i}{2\pi} \sum_{n \in \mathbb{Z}} \frac{1}{\beta_n} \exp[i(n + \alpha)(x_1 - y_1) + i\beta_n(x_2 - y_2)],$$

one can reduce the inverse problem of profile reconstruction to the following system of integral equations

$$\begin{aligned} Tz(x_1) &:= \int_0^{2\pi} z(t)G(x_1, b, t, 0)dt = u_b(x_1), \\ S_f z(x_1) &:= \int_0^{2\pi} z(t)G(x_1, f(x_1), t, 0)dt = -u^{in} \circ f(x_1), \end{aligned} \tag{3.21}$$

which is nonlinear with respect to f. Here $u^{in} \circ f(x_1) = \exp(i\alpha x_1 - i\beta f(x_1))$. Applying the arguments from the proof of [3, Lemma 4.1 and Theorem 4.2] one can obtain the following proposition.

PROPOSITION 3.6. *Let $u_b(x_1)$ be the exact pattern of the scattered field $u(x_1, x_2)$ on $x_2 = b$ that corresponds to some 2π-periodic profile functions $f \in C^2(\mathbb{R})$ meeting* (3.20). *Then there exists a solution (z_0, f_0) of the system* (3.21). *If in addition the inverse problem of profile reconstruction is uniquely solvable then $f = f_0$.*

Note that in problem (3.21) the knowledge of all Rayleigh coefficients A_n of the scattered waves is required. At the same time, the Fourier coefficients of $u_b(x_1) = u(x_1, b)$ with respect to orthonormal basis $\{\exp[i(n+\alpha)x_1]\}_{n\in\mathbb{Z}}$ of the complex Hilbert space $L_2(0, 2\pi)$ have the form $A_n e^{i\beta_n b}$, $n \in \mathbb{Z}$, and decay exponentially. Therefore, in practice one is able to measure only a finite number of $A_n, n \in U$, corresponding to outgoing plane waves (modes) of the scattered field (3.19) that can be observed on the line $x_2 = b$. Here U is some finite index set. Moreover, even these coefficients will not be given exactly but will be perturbed by measurement errors. To be more precise, we have only a vector $(A_n^\delta)_{n\in U}$ determining the "noisy trace"

$$u_b^\delta(x_1) = \sum_{n \in U} A_n^\delta \exp[i(n + \alpha)x_1 + i\beta_n b]$$

such that

$$\|u_b - u_b^\delta\| \le \delta, \tag{3.22}$$

where $\|\cdot\|$ denotes the norm in the complex Hilbert space $L_2(0, 2\pi)$.

Thus, replacing the scattered field u_b by u_b^δ one obtains the system (3.21) containing the noisy equation $Tz = u_b^\delta$, and for a stable profile reconstruction its regularized version should be considered. Such an approach was first proposed by Kirsch and Kress [14] for acoustic obstacle scattering. For the profile reconstruction problem it has been developed recently in [3]. These authors have observed that the structure of the system (3.21) allows the decomposition of the inverse problem of profile reconstruction into the severely ill-posed linear problem of estimating the scattered field potential density $z(t)$, and into the well-posed nonlinear problem of determining the unknown profile function as the location of the zeros of the total field; the later problem can then be replaced by the finite dimensional nonlinear least squares problem.

If $z(t)$ is given as a Fourier series

$$z(t) = \sum_{n\in\mathbb{Z}} z_n \exp[i(n+\alpha)t], \quad z_n \in \mathbb{C},$$

then the operators from the system (3.21) can be represented in the following form:

$$Tz(x_1) = i\sum_{n\in\mathbb{Z}} z_n\beta_n^{-1}\exp[i(n+\alpha)x_1 + i\beta_n b], \tag{3.23}$$

$$S_f z(x_1) = i\sum_{n\in\mathbb{Z}} z_n\beta_n^{-1}\exp[i(n+\alpha)x_1 + i\beta_n f(x_1)]. \tag{3.24}$$

Now it can be easily checked that T is an injective linear operator whose inverse T^{-1} acts continuously from $L_2 = L_2(0,2\pi)$ to the Hilbert space of generalized functions

$$L_{2,\exp}^{-b} := \left\{ z : \|z\|^2_{L^{-b}_{2,\exp}} := \sum_{n\in\mathbb{Z}} |z_n|^2 |e^{2i\beta_n b}||\beta_n|^{-2} < \infty \right\},$$

where z_n is the value of the functional $\langle z, e^{i(n+\alpha)}\rangle_{L_2(0,2\pi)}$, $n\in\mathbb{Z}$. Thus, if the problem was to find the solution of the equation $Tz = u_b^\delta$ in the space $L^{-b}_{2,\exp}$, it would be well-posed. But the second equation of (3.21) presumes $S_f z \in L_2(0,2\pi)$ for all admissible function meeting (3.20). One can guarantee it if $z \in L^{-b_0+h}_{2,\exp}$ for some $0 < h < b_0$. Indeed, $|\beta_m| \sim m$ and

$$\begin{aligned}\|S_f z\|^2 &= \int_0^{2\pi} \left| \sum_{n\in\mathbb{Z}} z_n \beta_n^{-1} e^{i(\alpha+n)x_1} e^{i\beta_n f(x_1)} \right|^2 dx_1 \\ &\le c\left(\sum_{n\in\mathbb{Z}} \frac{|z_n|^2}{|\beta_n|^2} |e^{2i\beta_n(b_0-h)}| \right) \int_0^{2\pi} \sum_{n\in\mathbb{Z}} e^{-2|\beta_n|(f(x_1)-b_0+h)} dx_1 \\ &\le \frac{c}{1-e^{-2h}} \|z\|^2_{L^{-b_0+h}_{2,\exp}} = c_h \|z\|^2_{L^{-b_0+h}_{2,\exp}},\end{aligned} \tag{3.25}$$

where the constant c_h depends only on h. Thus, it is reasonable to seek for solution of $Tz = u_b^\delta$ in the space $L^{-b_0+h}_{2,\exp}$.

Remark 3.7. In [3] it has been proposed to regularize the first equation of (3.21) in the space L_2. Keeping in mind that $L_2 \hookrightarrow L^{-b_0+h}_{2,\exp}$, it is easy to realize that for the pair (L_2, L_2) the problem $Tz = u_b^\delta$ is more ill-posed than for $(L^{-b_0+h}_{2,\exp}, L_2)$. Moreover, for any regularized solution z_δ of equation $Tz = u_b$ one has

$$\|z_0 - z_\delta\|_{L^{-b_0+h}_{2,\exp}} \le \|z_0 - z_\delta\|_{L_2},$$

where $z_0 = T^{-1}u_b$. At the same time, from (3.24) it follows that the perturbation of the left-hand side of the second equation of (3.21) caused by the replacement z_0 for z_δ can be estimated as

$$\|S_f z_0 - S_f z_\delta\| \le c_h \|z_0 - z_\delta\|_{L^{-b_0+h}_{2,\exp}}.$$

The equation supports the use of $L^{-b_0+h}_{2,\exp}$ as a more suitable space for the problem under consideration.

Singular value expansion (3.23) of the operator T allows us to apply the spectral cut-off scheme for the regularization of the equation $Tz = u_b^\delta$. It gives the following sequence of regularized solutions:

$$z_{m,\delta}(x_1) = -i \sum_{|n|<m} A_n^\delta \beta_n \exp[i(\alpha+n)x_1], \quad m = 1, 2, \ldots, M+1, \tag{3.26}$$

where $M = \max\{m : (-m, -m+1, \ldots, m-1, m) \subset U\}$. Replacing A_n^δ with A_n in (3.26), one obtains the partial sum $z_{m,0}$ of the Fourier series

$$z_0(x_1) = T^{-1}u_b(x_1) = -i \sum_{n \in \mathbb{Z}} A_n \beta_n \exp[i(\alpha+n)x_1].$$

Keeping in mind that $\|z_{m,0} - z_0\| \to 0$ as $m \to \infty$, and $\|z_{m,0} - z_0\|_{L_{2,\exp}^{-b_0+h}} \leq \|z_{m,0} - z_0\|$, we deduce that there exists an increasing continuous function $\varphi(\lambda)$ such that $\varphi(0) = 0$ and

$$\|z_0 - z_{m,0}\|_{L_{2,\exp}^{-b_0+h}} \leq \varphi\left(\frac{1}{m}\right). \tag{3.27}$$

Moreover, from (3.22) it follows that

$$\begin{aligned}
\|z_{m,0} - z_{m,\delta}\|^2_{L_{2,\exp}^{-b_0+h}} &= \sum_{|n|<m} |A_n - A_n^\delta|^2 |e^{2i\beta_n(b_0-h)}| \\
&= \sum_{|n|<m} |A_n - A_n^\delta|^2 |e^{2i\beta_n b}| |e^{2i\beta_n(b_0-h-b)}| \\
&\leq e^{2|\beta_m|(b+h-b_0)} \|u_b - u_b^\delta\|^2 \\
&\leq \delta^2 e^{2|\beta_m|(b+h-b_0)}.
\end{aligned}$$

Then

$$\|z_0 - z_{m,\delta}\|_{L_{2,\exp}^{-b_0+h}} \leq \varphi\left(\frac{1}{m}\right) + \delta e^{|\beta_m|(b+h-b_0)}.$$

This estimate has the form (1.6) with $\alpha = \frac{1}{m}$ and

$$\lambda(\alpha) = \exp\left[-\sqrt{|k^2 - (\alpha^{-1} + k\sin\theta)^2|}(b+h-b_0)\right]. \tag{3.28}$$

As in section 3.1 we consider $\Delta_M = \{\alpha_i = \frac{1}{M-i+1}\}_{i=0}^M$. Keeping in mind that

$$c_1 e^{-\frac{a}{\alpha}} \leq \lambda(\alpha) \leq c_1 e^{-\frac{a}{\alpha}}$$

with $a = (b+h-b_0)$ and some constants c_1, c_2 depending only on k and θ, it is easy to check that in considered case the condition (2.2) is satisfied with $q = \frac{c_2 e^a}{c_1}$. Then, as in section 3.1, the straightforward application of Theorem 2.1 gives the following theorem.

Theorem 3.8. *Assume that the inverse problem of profile reconstruction is uniquely solvable. If M is sufficiently large such that $M \sim (b+h-b_0)^{-1} \ln\frac{1}{\delta}$ then for m_+ chosen as*

$$m_+ = \min\left\{m : \|z_{m,\delta} - z_{n,\delta}\|^2_{L_{2,\exp}^{-b_0+h}} \leq 4\delta e^{|\beta_n|(b+h-b_0)},\ n = M+1, M, \ldots, m\right\}$$

one has

$$\|z_0 - z_{m_+,\delta}\|_{L_{2,\exp}^{-b_0+h}} \leq c\varphi((\varphi\lambda)^{-1}(\delta)),$$

where φ, λ are the functions from (3.27), (3.28), *and c depends only on b, b_0, h, k, θ.*

Remark 3.9. In the case under consideration, the spectral cut of scheme (3.26) can be combined with the discrepancy principle. Then the regularization parameter m would be chosen as

$$m_d = \min\{m : \|Tz_{m,\delta} - u_b^\delta\|_{L_2} \leq d\delta;\ m = 1, 2, \ldots, M+1\}. \tag{3.29}$$

It is easy to observe that the combination of (3.26) with (3.29) does not take into account our wish to regularize a problem in such an "exotic" space as $L_{2,\exp}^{-b_0+h}$. In this respect the parameter choice rule discussed above is much more flexible, and it is one more advantage of it.

4. Conclusion. As mentioned in the introduction, the a posteriori choice of the regularization parameter by several of the known principles may not yield the optimal order of accuracy for a given solution's smoothness. The principle proposed in the present paper is free from the above-mentioned drawback. Namely, for the first time one has a parameter choice rule that allows us to reach the best order of accuracy for all ill-posed problems that in principle can be treated in an optimal way by considered regularization methods.

We would like to emphasize that adaptive parameter choice strategy described in section 2 can be applied in a wide variety of situations where the goal is to find a balance between stability and approximation rate, and the latter one is unknown. Such a strategy may be of interest in other areas of numerical analysis. It seems that the problem of the weight choice in penalty finite element methods, for example, can be treated using the same idea.

Acknowledgments. The results of the paper were obtained while the first author was a DAAD guest professor at the University of Kaiserslautern. The support of the German Academic Exchange Service (DAAD) during the academic year 2002/2003 is gratefully acknowledged. The authors also thank Richard Lehoucq and two anonymous referees for careful reading and suggestions that have improved the presentation of the manuscript.

REFERENCES

[1] A. B. Bakushinskii, *Remarks on choosing a regularization parameter using the quasi-optimality and ratio criterion*, Comput. Math. Math. Phys., 24 (1984), pp. 181–182.

[2] M. B. Benson, *Errors in Numerical Quadrature for Certain Singular Integrands and the Numerical Solution of Abel Integral Equations*, Ph.D. thesis, Dept. of Math., University of Wisconsin, Madison, 1973.

[3] G. Bruckner, J. Elscher, and M. Yamamoto, *An Optimization Method for Grating Profile Reconstruction*, Preprint 682, WIAS Berlin, 2001.

[4] G. Bruckner and S. Pereverzev, *Self-regularization of projection methods with a posteriori discretization level choice for severally ill-posed problems*, Inverse Problems, 19 (2003), pp. 147–156.

[5] L. Cavalier and A. Tsybakov, *Sharp adaptation for inverse problems with random noise*, Probab. Theory Related Fields, 123 (2002), pp. 323–354.

[6] K. Deimling, *Nonlinear Functional Analysis*, Springer, New York, 1985.

[7] P. P. B. Eggermont, *A new analysis of the trapezoidal discretization method for the numerical solution of Abel-type integral equations*, J. Integral Equations, 3 (1981), pp. 317–332.

[8] H. W. Engl and H. Gfrerer, *A posteriori parameter choice for general regularization methods for solving linear ill-posed problems*, Appl. Numer. Math., 4 (1988), pp. 395–417.

[9] H. Gfrerer, *An a posteriori paramter choice for ordinary and iterated Tikhonov regularization of ill-posed problems leading to optimal convergence rates*, Math. Comput., 49 (1987), pp. 507–522.

[10] A. Goldenshluger and S. Pereverzev, *Adaptive estimation of linear functionals in Hilbert scales from indirect white noise observations*, Probab. Theory Related Fields, 118 (2000), pp. 169–186.

[11] C. W. Groetsch, *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*, Pitman, Boston, 1984.

[12] H. Harbrecht, S. Pereverzev, and R. Schneider, *Self-regularization by projection for noisy pseudodifferential equations of negative order*, Numer. Math., 95 (2003), pp. 123–143.

[13] B. Kaltenbacher, *Regularization by projection with a posteriori discretization level choice for linear and nonlinear ill-posed problems*, Inverse Problems, 16 (2000), pp. 137–155.

[14] A. Kirsch and R. Kress, *An optimization method in inverse acoustic scattering*, in Boundary Elements IX, Vol. 3, Springer, Berlin, 1987, pp. 3–18.

[15] O. Lepskii, *A problem of adaptive estimation in Gaussian white noise*, Theory Probab. Appl., 36 (1990), pp. 454–466.

[16] F. Liu and M. Z. Nashed, *Convergence of regularized solutions of nonlinear ill-posed problems with monotone operators*, in Partial Differential Equations and Applications, Dekker, New York, 1996, pp. 353–361.

[17] M. A. Lukas, *Comparison of parameter choice methods for regularization with discrete noisy data*, Inverse Problems, 14 (1998), pp. 161–184.

[18] J. T. Marti, *An algorithm for computing minimum norm solutions of Fredholm integral equations of the first kind*, SIAM J. Numer. Anal., 15 (1978), pp. 1071–1076.

[19] P. Mathé and S. Pereverzev, *Geometry of linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789–803.

[20] V. A. Morozov, *On the solution of functional equations by the method of regularization*, Soviet Math. Dokl., 7 (1966), pp. 414–417.

[21] D. L. Phillips, *A technique for the numerical solution of certain integral equations of the first kind*, J. Assoc. Comput. Mach., 9 (1962), pp. 84–97.

[22] R. Plato and U. Hämarik, *On pseudo-optimal parameter choices and stopping rules for regularization methods in Banach spaces*, Numer. Funct. Anal. Optim., 17 (1996), pp. 181–195.

[23] T. Raus, *Residue principle for ill-posed problems*, Acta Comment. Univ. Tartu. Math., 672 (1984), pp. 16–26.

[24] U. Tautenhahn, *On the method of Lavrentiev regularization for nonlinear ill-posed problems*, Inverse Problems, 18 (2002), pp. 191–207.

[25] U. Tautenhahn and U. Hämarik, *The use of monotonicity for choosing the regularization parameter in ill-posed problems*, Inverse Problems, 15 (1999), pp. 1487–1505.

[26] A. N. Tikhonov, *Regularization of incorrectly posed problems*, Soviet Math. Dokl., 4 (1963), pp. 1624–1627.

[27] A. N. Tikhonov and V. B. Glasko, *An approximate solution of Fredholm integral equations of the first kind*, Zh. Vychisl. Mat. Mat. Fiz., 4 (1964), pp. 564–571.

[28] A. N. Tikhonov and V. B. Glasko, *Use of the regularization method in non-linear problems*, Zh. Vychisl. Mat. Mat. Fiz., 5 (1965), pp. 463–473.

[29] A. Tsybakov, *On the best rate of adaptive estimation in some inverse problems*, C. R. Acad. Sci. Paris Sér. I Math., 330 (2000), pp. 835–840.

[30] R. Weiss, *Product integration for the generalized Abel equation*, Math. Comp., 26 (1972), pp. 177–190.

[31] R. Weiss and R. S. Anderssen, *A product integration for a class of first kind Volterra equations*, Numer. Math., 18 (1972), pp. 442–456.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2077–2097

A PROLONGATION/RESTRICTION OPERATOR FOR WHITNEY ELEMENTS ON SIMPLICIAL MESHES[*]

ALAIN BOSSAVIT[†] AND FRANCESCA RAPETTI[‡]

Abstract. The paper is mainly focused on the construction of two transfer operators between nested grids in the case of Whitney finite elements (node-, edge-, face-, or volume-based). These transfer operators, instances of what is called "chain map" in homology, have duals acting on cochains, that is to say, arrays of degrees of freedom in the context of the finite-element discretization of variational problems. We show how these duals can act as restriction/prolongation operators in a multigrid approach to such problems, especially those involving vector fields (e.g., electromagnetism). The duality between the operation of mesh refinement of a simplicial complex and that of restriction/prolongation of degrees of freedom from one mesh to a nested one is thus analyzed and explained. We use the language of p-forms, with frequent explanatory references to the more traditional vector-fields formalism.

Key words. mesh refinement, p-chains, p-forms, Whitney elements, multigrid

AMS subject classification. 65N30

DOI. 10.1137/040604923

1. Introduction.

In the approximation of a given differential problem by a finite element method, solving the final algebraic linear system is a delicate step. It is well-known that the associated matrix is sparse and can be of large size so that iterative solvers are preferable to direct ones. However, the convergence of iterative solvers strongly depends on the matrix condition number and slows down when the latter is large. Moreover, classical iterative methods fail to be effective whenever the spectral radius of the iteration matrix is close to one. A Fourier analysis shows that the reduction in the error depends on the spatial frequency. Errors with high frequency are rapidly damped whereas low frequency errors are slowly reduced and hold back convergence.

The multigrid algorithm [16] is an iterative technique well-adapted to solving linear systems arising from a finite element discretization of differential equations over a given grid. The basic idea of the method is to change the grid in such a way that low frequency (smooth) errors on a grid with elements of maximal diameter h can be singled out and cut down on a coarser grid, while high frequency errors that are not visible on the coarse grid with elements of maximal diameter $H > h$, for example, can be resolved on the fine grid. The exchange of information between the two meshes is done by means of two linear operators, one behaving as a prolongation and the other as a restriction. These operators are well known for nodal finite elements on nested or nonnested grids [19] but have still to be fully understood for edge or face finite elements.

It must be remarked that recovering the coarse grid from the fine one can be a very demanding operation. Therefore, in this paper we will address this problem the other

[*] Received by the editors March 9, 2004; accepted for publication (in revised form) November 1, 2004; published electronically December 15, 2005.
http://www.siam.org/journals/sinum/43-5/60492.html

[†] Laboratoire de Génie Electrique de Paris, UMR 8507 CNRS, Plateau de Moulon, Gif-sur-Yvette, Paris 91192, France (bossavit@lgep.supelec.fr).

[‡] Laboratoire de Mathématiques, UMR 6621 CNRS and Université de Nice et Sophia-Antipolis, Parc Valrose, 06108 Nice cedex 02, France (frapetti@math.unice.fr).

way around; i.e., we suppose that we have a coarse grid and we refine it repeatedly by means of a fixed procedure. This way, by using a suitable system of labels for the mesh nodes, we can know at which refinement level we are. Any other situation is not considered here since we wish to focus on the transfer of degrees of freedom from one mesh to the other, rather than on the coarsening process itself. However, the proposed analysis does not depend on the refinement or coarsening process. In short, we focus on a specific criterion to present the theory, but the theory is independent from the chosen criterion.

The paper is organized as follows. In section 2, suitable algebraic tools are introduced to lead the reader into the "world" of Whitney elements on simplicial meshes, including an appropriate formulation of the Stokes theorem. In section 3, we consider the problem of subdividing a simplicial mesh. The core of the paper is section 4, where we construct the information exchange operators between two "nested" meshes (by which we mean, two meshes m and $\tilde{m}$, the latter a conforming refinement of the former). Notions thus developed are applied in section 5, where we define the two transfer operators for Whitney elements on two nested simplicial meshes. The multigrid algorithm then comes as a straightforward application of these notions. Analyzing its performances is a difficult and technical issue, which we do not address. (Relevant references are given in section 5.)

2. Algebraic tools.

In this section, we recall some basic notions in algebraic topology (see, e.g., [1, 17]) and explain our notation. We restrict ourselves to a three-dimensional domain Ω (but the same notions can be defined in any dimension). For all integrals, we omit specifying the integration variable when this can be done without ambiguity. We shall denote by $\int_\gamma \mathbf{u}\cdot\mathbf{t}_\gamma$ and $\int_\sigma \mathbf{u}\cdot\mathbf{n}_\sigma$, respectively, the circulation and the flux of a vector field $\mathbf{u}$, where $\mathbf{t}_\gamma$ is the unit tangent to the smooth curve γ and $\mathbf{n}_\sigma$ the outward unit normal to the surface σ. Moreover, we shall emphasize the maps $\gamma \to \int_\gamma \mathbf{u}\cdot\mathbf{t}_\gamma$ and $\sigma \to \int_\sigma \mathbf{u}\cdot\mathbf{n}_\sigma$, that is to say, the differential forms of degree 1 and 2, respectively, which one can associate with a given vector field $\mathbf{u}$, and we occasionally use notations specific to exterior calculus, such as the exterior derivative d, as used in the Stokes theorem.

2.1. Triangulations and Whitney finite elements.

Given a domain $\Omega \subset \mathbb{R}^3$ with boundary Γ, a simplicial mesh m in Ω is a tessellation of $\overline{\Omega}$ by tetrahedra, under the condition that any two of them may intersect along a common face, edge, or node, but in no other way. We denote by $\mathcal{N}_m$, $\mathcal{E}_m$, $\mathcal{F}_m$, $\mathcal{T}_m$ (nodes, edges, faces, and tetrahedra, respectively) the sets of simplices of dimension 0 to 3 thus obtained (see Figure 1 for an example), each with its own orientation (more on this will follow), and by N_m, E_m, F_m, T_m their cardinalities. Alternatively, we may use $\mathcal{S}^p_m$ to denote the set of p-dimensional simplices in m and $\#\mathcal{S}^p_m$ for its cardinality. The importance of simplicial meshes lies in the fact that any triangulated domain is homeomorphic to one in which the triangles are flat and the edges straight. Note that the triangulation for Ω is not unique, but topological properties of a triangulated domain do not depend on the triangulation used to investigate them. (For such "homological" computations, using a definite triangulation but yielding mesh-independent results, which we believe are relevant to engineering practice, see [12, 15].)

For what follows, we need to underline some combinatorial properties of the simplicial mesh. Besides the list of nodes and their positions, the mesh data structure also contains incidence matrices, saying which node lies at an end of which edge, which edge bounds which face, etc. [4]. This encodes the orientation of each simplex, as we now explain. In short, an oriented edge is not only a two-node subset of $\mathcal{N}_m$,

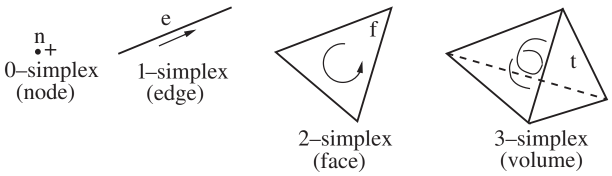


FIG. 1. *Examples of oriented p-simplex, $p = 0, \dots, 3$.*

but an ordered such set, where the order implies an orientation. Let $e = \{\ell, n\}$ be an edge of the mesh oriented from the node ℓ to n. We can define the incidence numbers $G_{e,n} = 1$, $G_{e,\ell} = -1$, and $G_{e,k} = 0$ for all nodes k other than ℓ and n. These numbers form a $(E_m \times N_m)$-matrix G, which describes how edges connect to nodes. A face $f = \{\ell, n, k\}$ has three vertices which are the nodes ℓ, n, k. Note that $\{n, k, l\}$ and $\{k, l, n\}$ denote the same face f, whereas $\{n, l, k\}$ denotes an oppositely oriented face, which is not supposed to belong to $\mathcal{F}_m$ if f does. An orientation of f induces an orientation of its boundary. So, with respect to the orientation of the face f, the one of the edge $\{l, n\}$ is positive and that of $\{k, n\}$ is negative. So we can define the incidence number $R_{f,e} = 1$ (resp., -1) if the orientation of e matches (resp., does not match) the one on the boundary of f and $R_{f,e} = 0$ if e is not an edge of f. These numbers form a $(F_m \times E_m)$-matrix R. Finally, let us consider the tetrahedron $t = \{k, l, m, n\}$, positively oriented if the three vectors $\{k, l\}$, $\{k, m\}$, and $\{k, n\}$ define a positive frame ($t' = \{l, m, n, k\}$ has a negative orientation and does not belong to $\mathcal{T}_m$ if t does). A $(V_m \times F_m)$-matrix D can be defined by setting $D_{t,f} = \pm 1$ if face f bounds the tetrahedron t, with the sign depending on whether the orientation of f and of the boundary of t match or not, and $D_{t,f} = 0$ in case f does not bound t. (For consistency, we may attribute an orientation to nodes as well—a sign, ± 1. Implicitly, we have been orienting all nodes the same way ($+1$) up to now. Note that a sign (-1) to node n changes the sign of all entries of column n in the above G.) It can easily be proved that $RG = 0$ and $DR = 0$ [4].

We now define the Whitney finite elements we use [4, 9, 10, 13]: They are scalar functions or vector fields associated to all the simplices of the mesh m. Given the node n, the edge $e = \{\ell, m\}$, the face $f = \{\ell, m, k\}$, and the tetrahedron $t = \{i, j, k, \ell\}$, we define the following scalar or vector functions (λ_n is the barycentric coordinate associated to node n):

$$\begin{aligned}
w^n &= \lambda_n,\\
w^e &= \lambda_\ell \,\mathbf{grad}\,\lambda_m - \lambda_m \mathbf{grad}\,\lambda_\ell,\\
w^f &= 2\,(\lambda_\ell \,\mathbf{grad}\,\lambda_m \times \mathbf{grad}\,\lambda_k + \lambda_m \,\mathbf{grad}\,\lambda_k \times \mathbf{grad}\,\lambda_\ell + \lambda_k \,\mathbf{grad}\,\lambda_\ell \times \mathbf{grad}\,\lambda_m),\\
w^t &= 6\,(\lambda_i \,\mathbf{grad}\,\lambda_j \times \mathbf{grad}\,\lambda_k \cdot \mathbf{grad}\,\lambda_\ell + \lambda_j \,\mathbf{grad}\,\lambda_k \times \mathbf{grad}\,\lambda_\ell \cdot \mathbf{grad}\,\lambda_i\\
&\qquad + \lambda_k \,\mathbf{grad}\,\lambda_\ell \times \mathbf{grad}\,\lambda_i \cdot \mathbf{grad}\,\lambda_j + \lambda_\ell \,\mathbf{grad}\,\lambda_i \times \mathbf{grad}\,\lambda_j \cdot \mathbf{grad}\,\lambda_k)
\end{aligned}$$

(w^t is just the constant $1/\mathrm{vol}(t)$). We define $W^p_m = \mathrm{span}\,\{w^s : s \in \mathcal{S}^p_m\}$, $p = 0, 1, 2, 3$ (the simplicial dimension, e.g., $p = 0$ for nodes). It can be verified that the value (resp., circulation, flux, integral) of w^n (resp., w^e, w^f, w^t) on its supporting simplex is 1, and 0 on other simplices of matching dimension, a fact we shall be able to state more compactly in a moment.

Given two adjacent tetrahedra t and t' sharing a face f, the function w^n and both the tangential component of w^e and the normal component of w^f are continuous across f, whereas the function w^t is not. Thanks to these continuity properties, $W^0_m \subset H^1(\Omega)$, $W^1_m \subset H(\mathbf{curl}, \Omega)$, $W^2_m \subset H(\mathrm{div}, \Omega)$, $W^3_m \subset L^2(\Omega)$. The spaces W^p_m, $p = 0, 1, 2, 3$, have finite dimension given by N_m, E_m, F_m, T_m, respectively, and they play the role of Galerkin approximation spaces for the functional spaces just mentioned. Therefore, a scalar field $k \in H^1(\Omega)$ can be represented in W^0_m by the approximation $\sum_{n\in\mathcal{N}_m} k_n\, w^n$ where $\{k_n : n \in \mathcal{N}_m\}$ are the values of k at the mesh nodes (i.e., the degrees of freedom of k on the mesh m). Similarly, a vector field $\mathbf{v} \in H(\mathbf{curl}, \Omega)$ can be represented in W^1_m by $\sum_{e\in\mathcal{E}_m} v_e\, w^e$, where $\{v_e : e \in \mathcal{E}_m\}$ are the circulations of $\mathbf{v}$ along the mesh edges. A vector field $\mathbf{v} \in H(\mathrm{div}, \Omega)$ can be represented in W^2_m by $\sum_{f\in\mathcal{F}_m} v_f\, w^f$, where $\{v_f : f \in \mathcal{F}_m\}$ are the fluxes of $\mathbf{v}$ across the mesh faces. Finally, a scalar field $k \in L^2(\Omega)$ can be represented in W^3_m by $\sum_{t\in\mathcal{T}_m} k_t\, w^t$, where $\{k_t : t \in \mathcal{T}_m\}$ are the integrals of k on the mesh tetrahedra.

Properties discussed so far concern the spaces W^p_m taken one by one. Properties of the structure made of the spaces W^p_m when taken together should also be mentioned. We know that the following inclusions hold:

$$\mathbf{grad}\, W^0_m \subset W^1_m, \qquad \mathbf{curl}\, W^1_m \subset W^2_m, \qquad \mathrm{div}\, W^2_m \subset W^3_m.$$

It is natural to ask whether the sequence

$$\{0\} \longrightarrow W^0_m \xrightarrow{\mathbf{grad}} W^1_m \xrightarrow{\mathbf{curl}} W^2_m \xrightarrow{\mathrm{div}} W^3_m \longrightarrow \{0\}$$

is *exact* at levels 1 and 2, i.e., when it happens that

$$\ker(\mathbf{curl}; W^1_m) = \mathbf{grad}\, W^0_m, \qquad \ker(\mathrm{div}; W^2_m) = \mathbf{curl}\, W^1_m,$$

where $\ker(\mathbf{curl}; W^1_m) := W^1_m \cap \ker(\mathbf{curl})$ and $\ker(\mathrm{div}; W^2_m) := W^2_m \cap \ker(\mathrm{div})$. (At level 0, the gradient operator is not injective. At level 3, the divergence operator is surjective.) The Poincaré lemma states that, when the domain Ω is contractible, the image fills the kernel in both cases. This may fail to happen: With Ω a solid torus, for example, $\mathbf{grad}\,(W^0_m)$ is a proper subset of $\ker(\mathbf{curl}; W^1_m)$. If so, it tells us something on the topology of Ω, namely the presence of b_1 "loops," where $b_1 = \dim[\ker(\mathbf{curl}; W^1_m)/\mathbf{grad}\,(W^0_m)]$ is the Betti number of dimension 1 of the domain. Solenoidal fields that are not **curl**s indicate the presence of b_2 "holes," where $b_2 = \dim[\ker(\mathrm{div}; W^2_m)/\mathbf{curl}\,(W^1_m)]$ is the Betti number of dimension 2 of the domain. (One may add that $b_0 = \dim[\ker(\mathbf{grad}; W^0_m)]$ is the number of connected components, here assumed to be 1, of Ω.) These are global topological properties of the meshed domain: They depend on Ω, but not on which mesh is used to compute them. The sequences are thus an algebraic tool by which the topology of Ω can be explored (which was the point of inventing Whitney forms [18]).

2.2. Chains and homology groups. We now introduce chains over the mesh m. A p-chain c is an assignment to each p-simplex s of a rational integer α_s. This can be denoted by $c = \sum_{s\in\mathcal{S}^p_m} \alpha_s\, s$. Let $C_p(m)$ be the set of all p-chains. This set has a structure of Abelian group with respect to the addition of p-chains: Two p-chains are added by adding the corresponding coefficients.

If s is an oriented simplex, the *elementary chain* corresponding to s is the assignment $\alpha_s = 1$ and $\alpha'_s = 0$ for all $s' \neq s$. In what follows, we will use the same symbol s (or n, e, etc., depending) to denote the oriented simplex and the associated elementary chain. Note how this is consistent with the above expansion of c as a formal weighted sum of simplices. Which meaning is implied will hopefully be clear from the context.

The *boundary* of an oriented p-simplex of m is a $(p-1)$-chain determined by the sum of its $(p-1)$-dimensional faces, each taken with the orientation induced from that of the whole simplex. So, the boundary ∂s of a single simplex s is

$$\partial e = \sum_{n \in \mathcal{N}_m} G_{e,n}\, n, \qquad \partial f = \sum_{e \in \mathcal{E}_m} R_{f,e}\, e, \qquad \partial t = \sum_{f \in \mathcal{F}_m} D_{t,f}\, f.$$

By linearity, the boundary operator ∂ defines a group homomorphism $C_p(m) \to C_{p-1}(m)$ as follows:

$$\partial c = \partial \left(\sum_{s \in \mathcal{S}_m^p} \alpha_s\, s \right) = \sum_{s \in \mathcal{S}_m^p} \alpha_s\, \partial s.$$

Note that ∂ is represented by a matrix, which is G^t, R^t, or D^t, depending on the dimension $p > 0$. We remark that $\partial \circ \partial = 0$, i.e., the boundary of a boundary is the null chain. When $p = 0$, we define the boundary of a single vertex to be zero and $C_{-1}(m) = \{0\}$.

The kernel of $\partial : C_p(m) \to C_{p-1}(m)$ is denoted by $Z_p(m)$ and is called the group of *p-cycles* of m. The image of $\partial : C_{p+1}(m) \to C_q(m)$ is denoted by $B_q(m)$ and is called the group of *p-boundaries* of m. The property $\partial \circ \partial = 0$ implies that $B_p(m)$ is a subgroup of $Z_p(m)$. The quotient $H_p(m) = Z_p(m)/B_p(m)$ is the *homology group* of order p of the mesh m and the Betti number b_p is equal to the rank of $H_p(m)$. Not all cycles bound, as a rule (think again of the solid torus, for $p = 1$), so b_p need not be zero.

By linearity, integration over simplices extends to chains as follows (let's deal with 2-chains for definiteness). If $c = \Sigma_{f \in \mathcal{F}_m} c_f f$, the integral of a vector field w over c is, by definition (and with some notational abuse for which we shall be rewarded later),

$$\int_c w = \sum_{f \in \mathcal{F}_m} c_f \int_f w \cdot \mathbf{n}_f. \tag{1}$$

Substituting the Whitney form w^f for w there, one sees that $\int_c w^f$ is just c_f. A similar definition can be stated for node-based, edge-based, or volume-based chains. So we now have $\int_{s'} w^s = 1$ if $s' = s$ and 0 if $s' \neq s$ for all p-simplices s' and Whitney elements w^s—the promised compact expression of their main property.

Remark 2.1. We note that (1) amounts to considering the vector field w as a differential form, as defined at the beginning of section 1. The functions and vector fields w^n, w^e, w^f, w^t of section 2.1 are thus differential forms, known as Whitney forms in the mathematical literature [18].

2.3. Cochains and cohomology groups. In this section, we introduce the dual concept of *p-cochain*. A *p-cochain* is a linear functional on the vector space of p-chains. For instance, given an array $\mathbf{b} = \{b^s : s \in \mathcal{S}_m^p\}$ of real numbers, we can define the p-cochain $c \to \sum_{s \in \mathcal{S}_m^p} b^s\, c_s$ acting on p-chains c with coefficients c_s. Also, as in (1), given a differential form w, the mapping $c \to \int_c w$ defines a p-cochain. More generally, the p-cochain coefficients are obtained by integrating the differential form w on the elements of the p-chain c; i.e., the map $c \to \sum_{s \in \mathcal{S}_m^p} c_s \int_s w$ is a cochain. We shall denote the latter value as $\langle w\ ;\ c \rangle$.

Once a metric is introduced on the ambient affine space, differential forms are in correspondence with scalar and vector fields (called "proxy fields"—metric dependent, of course). The coefficients of p-cochains become the degrees of freedom of scalar and vector fields (and this is exactly what occurs with Whitney finite elements in section 2.1). Let $W^p(m)$ denote the set of p-cochains (or p-forms) defined on Ω when triangulated by m. Then, $C_p(m)$ and $W^p(m)$ are *in duality* via the bilinear continuous map $\langle \cdot\ ;\ \cdot \rangle : W^p(m) \times C_p(m) \rightarrow \mathbb{R}$ defined as $\langle w\ ;\ c\rangle = \int_c w$ where the integral must be interpreted as in (1) in the example case $p = 2$. A duality product should be *nondegenerate*, i.e., $\langle w\ ;\ c\rangle = 0$ for all c implies $w = 0$, and $\langle w\ ;\ c\rangle = 0$ for all w implies $c = 0$. The former property holds true by definition, and the latter is satisfied because, if $c \neq 0$, one can construct an ad-hoc smooth vector field or function with nonzero integral and hence a nonzero form w such that $\langle w\ ;\ c\rangle \neq 0$.

For $p > 0$, the *exterior derivative* of the $(p-1)$-form w is the p-form $\mathrm{d}\, w$. The integral $\int_c w$ is treated in two ways: If $c = \partial\tau$ and w is smooth, one may go forward and integrate dw over τ. Alternatively, if the form $w = dv$, one may go backward and integrate v over ∂c. In particular, we have $\int_{\partial c} w = \int_c \mathrm{d}\, w$, which is the common form of Stokes' theorem [7], or equivalently,

$$\langle w\ ;\ \partial c\rangle = \langle \mathrm{d}\, w\ ;\ c\rangle \qquad \forall c \in C_p \quad \text{and} \quad \forall w \in W^{p-1}. \tag{2}$$

Equation (2) reveals that d is the *dual* of ∂ (in the sense of Yosida [20, p. 194]). As a corollary of the boundary operator property $\partial \circ \partial = 0$, we have that $\mathrm{d} \circ \mathrm{d} = 0$. A form w is *closed* if $\mathrm{d}\, w = 0$, *exact* if $w = \mathrm{d}\, v$ for some v (in the first case we have a *cocycle* and in the second case a *coboundary*). Denoting by $\mathcal{Z}^p(m)$ the vector space of all closed p-forms and by $\mathcal{B}^p(m)$ the subspace made of all exact p-forms, the property $\mathrm{d} \circ \mathrm{d} = 0$ implies that $\mathcal{B}^p(m) \subset \mathcal{Z}^p(m)$; i.e., the integral of a cocycle over a boundary vanishes. In domains Ω that are topologically trivial, all closed p-forms are exact (this is the Poincaré lemma). But closed forms need not be exact in general manifolds: This is the dual aspect of the above "not all cycles bound" (section 2.2). The quotient space $\mathcal{H}^p(m) = \mathcal{Z}^p(m)/\mathcal{B}^p(m)$ is (considered as an additive group) the *De Rham's* p*th cohomology group* of Ω or equivalently of m.

3. Refinement of a triangulation and simplicial maps.

A mesh refinement is a procedure to subdivide each simplex of a given mesh (referred to as the "coarse" one) m into a finite number of smaller ones, whose assembly is still a proper mesh (the "fine" one). We consider here conforming refinements, i.e., such that the set $\tilde{m}$ of all simplices of the fine mesh, is itself a cellular complex (no hanging nodes). Moreover, we are interested in subdividing a simplicial mesh in a way that will not deteriorate the aspect ratio of the new smaller and smaller tetrahedra that appear during the division process. In this framework, we speak of *uniform* refinement procedure if there is a *finite* catalog of "model cells" such that any cell in any $\tilde{m}$ is similar to one of them, for all meshes $\tilde{m}$ in the family $\mathcal{M}$ of meshes potentially created in the process of iterated refinement.

The *barycentric* (or *regular*) *refinement* is an example of conforming refinement procedure where the small cells are more and more stretched (see Figure 2 for a face) and hence not uniform in that sense. In three dimensions, each tetrahedron T is divided into 24 tetrahedra, and we can understand that after the first step of refinement, the new tetrahedra are more stretched toward the barycenter o: When their aspect ratio becomes too small, the classical a priori error estimates for finite elements do not apply and convergence is not warranted.

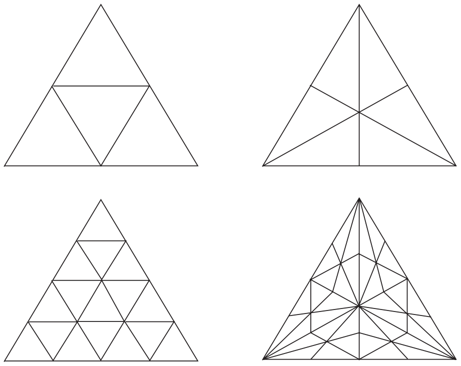


FIG. 2. *Two-level refinement of normal (left) and barycentric (right) type. A face is divided, respectively, into four faces (left) and six faces (right) at each refinement level.*

The *normal refinement*[1] presented in Figure 2 for a face, and in Figure 5 for a tetrahedron, is an example of conforming refinement procedure that enjoys uniformity. In three dimensions, let us consider a tetrahedron T built on four nodes k, l, m, n. Call o the center, lm, ln, etc., the midpoints. The big tetrahedron $T = \{k, l, m, n\}$ subdivides into four midsize ones, such as $\{kn, ln, mn, n\}$, and a midsize core octahedron (Figure 3), itself a half-size reduction of a big one O circumscribed to T. In turn, the core divides into six small octahedra and eight small tetrahedra, all similar to O and T, respectively, with a reduction factor of 4 (Figure 4). Hence there are two basic shapes, that of T and that of O, which are found again and again.

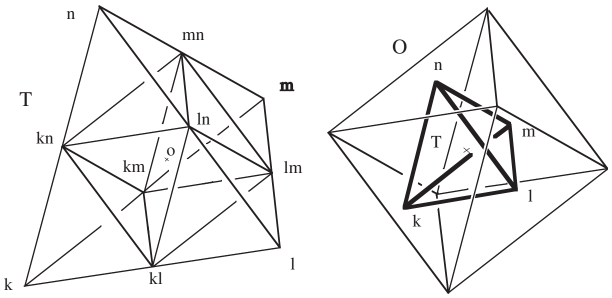


FIG. 3. *Cutting T into four midsize tetrahedra plus a core octahedron, similar to the circumscribed one, O. Note that the common center o of T and O is four times closer to face $\{k, l, m\}$ than node n was. Faces of O are similar to those of T, twice as big.*

All that is left to do, in order to get a series of nested simplicial meshes, is to cut the octahedra into tetrahedra, either eight (Figure 5) or just four. The latter solution simply consists of adding an edge joining two opposite nodes of the octahedron. As there are three nonequivalent ways to do that, one must be careful to draw all these

[1] Whitney defines in [18, pp. 358–360] a *standard subdivision* that guarantees uniformity but does not treat nodes symmetrically, the way ours does, hence the introduction of the adjective *normal* for definiteness. Normal subdivision can be done in dimensions $d > 3$, where it also involves, as can be inferred from Figure 3, convex hulls of barycenters of p-faces of the reference d-simplex. Denoting such convex polytopes by T_p (the reference simplex thus being T_0, and the O of Figure 5 a T_1), it can be shown that each T_p can be dissected into a finite number (bounded by a function of d only) of polytopes similar to one of the T_q, $1 \leq q \leq d$; hence there is uniformity.

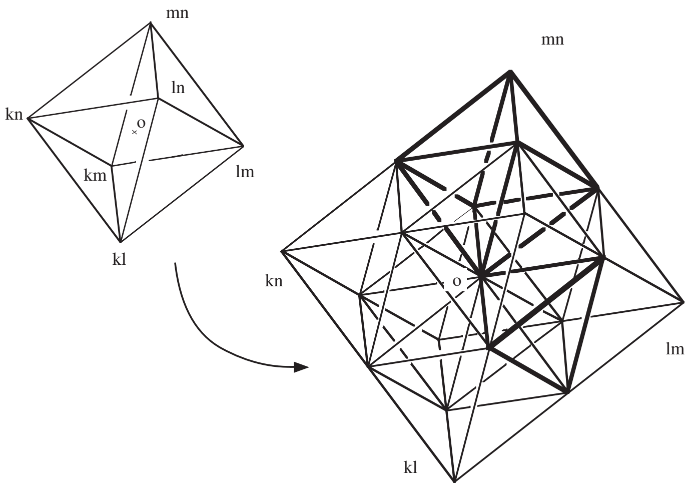


FIG. 4. *Cutting the octahedral core O into six small octahedra (one per node of the core, or edge of T) plus eight small tetrahedra (one per face of the core), all similar to T and O, respectively, and four times smaller.*

diagonals parallel to a same direction if one wants to minimize the number of different shapes of tetrahedra. Cutting in eight is a more symmetrical procedure. With both methods, the number of shapes is kept down to five. (That is the generic number, of course lower if T had some symmetry to start with.)

In any event, it is only at the latest stage of the subdivision that the final cut of octahedra should be involved. Conceptually, we have two cell shapes, T and O. Each T-cell breaks into four smaller T-cells and one O-cell. Each O-cell splits into six O-cells and eight T-cells. At the generic step $\kappa \geq 1$, we get

$$T \rightarrow \alpha \, \frac{T}{2^\kappa} + \beta \, \frac{O}{2^{\kappa-1}}, \tag{3}$$

where α, β are two positive integers. As a last step, O-cells are chopped.

If a tetrahedron t born from this last subdivision is earmarked for refinement by the error-estimator, one must look upward to its ancestry before dividing it. If t is a T-cell, apply normal subdivision. Otherwise, backtrack to its mother O-cell and subdivide the latter. Apart from those that are T-cells, tetrahedra of the subdivision are mules, not able to reproduce by division. The same strategy applies to the transition layer of tetrahedra that touch divided ones, and need division for conformity. They should be cut in two or more tetrahedra, depending on how many of their edges belong to divided tetrahedra. Here, for simplicity, we consider only two cases: a T-cell with a divided face results in four tetrahedra and the one with a divided edge results in two. Products of this subdivision can be mules as well as tetrahedra with divided edges or faces and the procedure goes on. All other T-cells presenting two or more divided faces are cut according to the normal subdivision (at worst, the normal subdivision applies to the whole set of tetrahedra). If one of the two or four tetrahedra t that compose a T-cell in the transition layer is pointed at for subdivision, one backtracks to its (mother) T-cell and applies normal subdivision. (Refining tetrahedra t the same way, by the normal subdivision that served for the T-cell, would make a mess.) The number of different tetrahedral shapes is thus kept small, whatever the depth of the subdivision procedure.

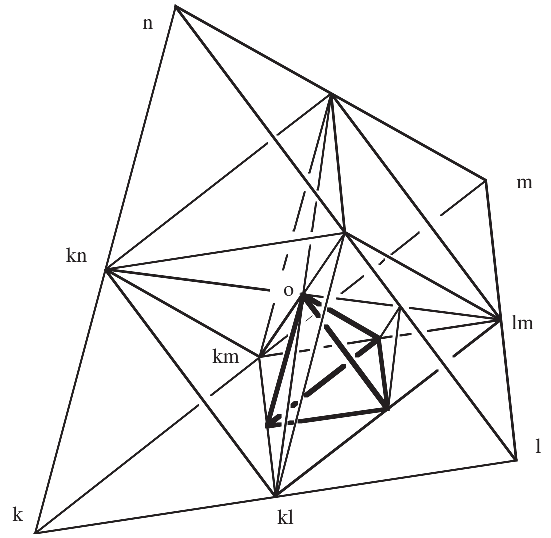


FIG. 5. *Normal refinement for the tetrahedron $T = \{k, l, m, n\}$. Mid-edges are denoted kl, lm, etc., and o is the barycenter. A first halving of edges generates four small tetrahedra and a core octahedron, which itself can be divided into eight "octants" such as $O = \{o, kl, lm, mk\}$, of at most four different shapes. At this point, we have twelve small tetrahedra, only eight in the octahedron. Now, octants like O should be subdivided as follows: Divide the face in front of o into four triangles and join to o; hence we have a tetrahedron similar to T, and three peripheral tetrahedra. These, in turn, are halved as shown for the one hanging from edge $\{o, lm\}$. Its two parts are similar to O and to the neighbor octant $\{o, kn, kl, mk\}$, respectively. At the end of the second step, we have* 56 *tetrahedra for the core octahedron.*

Note that, starting from a given mesh, the barycentric subdivision as well as the normal one do not change the homology group of a complex, since the triangulated domain is always the same. This is the very point of homology (see, for example, [1]).

4. Construction of a restriction/prolongation operator between two nested meshes. Recall that the collection of groups and homomorphisms

$$\{0\} \longrightarrow \ldots \xrightarrow{\partial} C_p(m) \xrightarrow{\partial} C_{p-1}(m) \xrightarrow{\partial} \ldots \longrightarrow \{0\}$$

is usually referred to as the *chain complex* of the mesh m and denoted by $C(m)$. Here, we shall consider *two* meshes, the coarse one m and the fine one $\tilde{m}$, as obtained from m by a given refinement technique; hence we have two complexes $C(m)$ and $C(\tilde{m})$. We use capital letters to denote nodes, edges, faces, and volumes in m and lowercase letters to denote analogous cells in $\tilde{m}$. Incidence matrices for $\tilde{m}$ are denoted g, r, d. Recall that the elementary chain associated with a simplex of m (or $\tilde{m}$) and the simplex itself are denoted by the same symbol. Last, we shall use the shorthand "$s \subset S$" when simplex s is, *as a subset of the three-dimensional space*, a part of S. (Thus, $N \subset E$ means N is an endpoint of E. In the case of nodes, $n \subset N$ just means that n and N sit at the same point.)

In what follows, we first introduce two maps, $\chi : C(m) \to C(\tilde{m})$ and $\pi : C(\tilde{m}) \to C(m)$; we next prove that χ and π are "chain maps," as defined below, and that $\pi\,\chi = 1$.

Definition 4.1. *Given a p-simplex S of the coarse complex m, set*

$$\chi(S) = \sum_{s\in\mathcal{S}^p_{\tilde{m}}} \chi^s_S\, s \quad \textit{with} \quad \chi^s_S = \begin{cases} 0 & \textit{if} \quad s \not\subset S, \\ +1 & \textit{if} \quad s \subset S \textit{ and same orientation}, \\ -1 & \textit{if} \quad s \subset S \textit{ and opposite orientation}. \end{cases} \tag{4}$$

Of course, the map χ is the natural way to embed m into $\tilde{m}$: Chop the large simplex into small ones, and build a chain from these, with weights ±1 according to respective orientations. For nodes, we assumed positive orientation for all of them, so χ^n_N is 1 if n coincides with N, 0 otherwise.

Next, let w^S denote the Whitney form associated with a p-simplex S of the coarse mesh so that $\langle w^S\ ;\ S'\rangle = \delta_{S,S'}$ for all p-simplices $S' \in m$. Then we have the following definition.

Definition 4.2. *Given a p-simplex s of the fine complex $\tilde{m}$, set*

$$\pi(s) = \sum_{S\in\mathcal{S}^p_m} \langle w^S\ ;\ s\rangle\, S \equiv \sum_{S\in\mathcal{S}^p_m} \pi^S_s\, S. \tag{5}$$

A small simplex is thus represented by a chain of big ones. (The use of Whitney forms for this is natural: As argued elsewhere [5], Whitney forms are best viewed as a device to represent manifolds by simplicial chains. Here, the manifold is the small simplex s.) We now prove three propositions.

Proposition 4.3. *One has $\pi\,\chi = 1$.*

Proof. We must show that $\pi(\chi(S)) = S$ for any coarse p-simplex S. Indeed,

$$\pi(\chi(S)) = \pi\left(\sum_{s\in\mathcal{S}^p_{\tilde{m}}} \chi^s_S\, s\right)$$

$$= \sum_{s\in\mathcal{S}^p_{\tilde{m}}} \chi^s_S \sum_{S'\in\mathcal{S}^p_m} \langle w^{S'}\ ;\ s\rangle\, S'$$

$$= \sum_{S'\in\mathcal{S}^p_m} \left\langle w^{S'}\ ;\ \sum_{s\in\mathcal{S}^p_{\tilde{m}}} \chi^s_S s \right\rangle S' \tag{6}$$

$$= \sum_{S'\in\mathcal{S}^p_m} \langle w^{S'}\ ;\ S\rangle\, S' \tag{7}$$

$$= S,$$

thanks to the fundamental property of Whitney forms, $\langle w^{S'}\ ;\ S\rangle = \delta_{S,S'}$. To pass from (6) to (7), use the equality $\langle w\ ;\ S\rangle = \langle w\ ;\ \chi(S)\rangle$, for any given p-form w, which stems from additivity of the integral. □

It is important to remark that Proposition 4.3 and its proof do not depend on the refinement *technique*, but just on the fact that coarse cells are tessellations of small ones.

Proposition 4.4. *The map χ defined in* (4) *is a chain map, i.e., $\partial\,\chi = \chi\,\partial$.*

Proof. Although both statement and proof are independent of the refinement procedure, we suppose here that the normal subdivision is considered at the first step (where each tetrahedron gives 12 small tetrahedra) to build up the fine complex $\tilde{m}$

from the coarse one m. We also treat separately the cases $p = 1, 2, 3$, where p is the dimension of the chain on which $\partial\chi$ and $\chi\partial$ act, though as will be apparent a generic proof (much shorter, but perhaps less informative) could be given. Our purpose is to help understand, on concrete examples, what is going on.

For $p = 1$, we have

$$\begin{aligned}\chi\,\partial\,E &= \chi\,(\textstyle\sum_{N\in\mathcal{N}_m} G_{E,N}\,N)\\ &= \textstyle\sum_{N\in\mathcal{N}_m} G_{E,N}\sum_{n\in\mathcal{N}_{\tilde m}} \chi_N^n\,n\\ &= \textstyle\sum_{n\in\mathcal{N}_{\tilde m}} [\sum_{N\in\mathcal{N}_m} G_{E,N}\,\chi_N^n\,]\,n = \sum_{n\in\mathcal{N}_{\tilde m}} [\sum_{N\subset E} G_{E,N}\,\chi_N^n\,]\,n,\end{aligned}$$

since only those nodes N that are, as sets, part of E, make $G_{E,N} \neq 0$, and thus contribute to the sum. On the other hand, we obtain:

$$\begin{aligned}\partial\,\chi\,E &= \partial\,(\textstyle\sum_{e\in\mathcal{E}_{\tilde m}}\ \chi_E^e\ e)\\ &= \textstyle\sum_{e\in\mathcal{E}_{\tilde m}}\ \chi_E^e\ \sum_{n\in\mathcal{N}_{\tilde m}}\ g_{e,n}\ n\\ &= \textstyle\sum_{n\in\mathcal{N}_{\tilde m}} [\sum_{e\in\mathcal{E}_{\tilde m}} \chi_E^e\ g_{e,n}]\,n = \sum_{n\in\mathcal{N}_{\tilde m}} [\sum_{e\subset E} \chi_E^e\ g_{e,n}]\,n,\end{aligned}$$

since only those e that are, as sets, part of E, make $\chi_E^e \neq 0$, and thus contribute to the sum. The conclusion comes from the equality between bracketed terms above, which stems from the interplay between incidence numbers on both meshes, as we now show in detail.

If $n \not\subset E$, there is no N such that $n \subset N \subset E$, so the first bracket vanishes. There is no e either such that $n \subset e \subset E$, so the second bracket vanishes too. Assuming therefore $n \subset E$, we have two cases to examine, illustrated by the center part and the right-hand part of Figure 6, where E is supposed to be E_2: either $n \subset N$ for N one of the endpoints of E_2 (say N_1 or N_2), or not (see Figure 6, center and right-hand part, respectively).

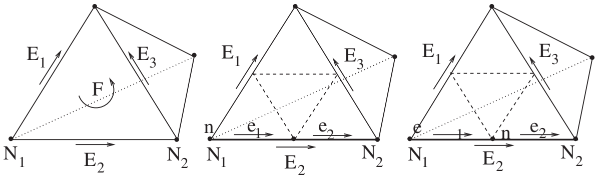


FIG. 6. *For node $n \in \mathcal{N}_{\tilde m}$ belonging to edge E_2, either there exists $N \subset E_2$ such that $n \subset N$ (center) or not (right).*

According to the situation at the center of Figure 6 ($n \subset N_1$), we have

$$\sum_{N\subset E_2} G_{E_2,N}\,\chi_N^n = G_{E_2,N_1}\,\chi_{N_1}^n + G_{E_2,N_2}\,\chi_{N_2}^n = (-1)(1) + (1)(0) = -1.$$

For the situation at the right-hand side of Figure 6, since $n \not\subset N_i$, for $i = 1$ or 2, we have

$$\chi_{N_i}^n = 0 \quad \forall\, i \quad \text{so that} \quad \sum_{N\subset E} G_{E,N}\,\chi_N^n = 0.$$

Let us do the same reasoning for the other quantity, looking at Figure 6. For the situation at the center of Figure 6, we have (recalling that g is the “fine” incidence matrix)

$$\sum_{e\subset E_2} g_{e,n}\, \chi^{e}_{E_2} = g_{e_1,n}\, \chi^{e_1}_{E_2} + g_{e_2,n}\, \chi^{e_2}_{E_2} = (-1)(1) + (0)(1) = -1.$$

For the situation at the right-hand side of Figure 6, we have

$$\sum_{e\subset E_2} g_{e,n}\, \chi^{e}_{E_2} = g_{e_1,n}\, \chi^{e_1}_{E_2} + g_{e_2,n}\, \chi^{e_2}_{E_2} = (1)(1) + (-1)(1) = 0.$$

Summing up, for a given $n \in \mathcal{N}_{\tilde{m}}$, the two quantities in brackets take the same value (-1, 1 or 0), due to the definition of the incidence matrices G and g and coefficients χ^n_N and χ^e_E.

For $p = 2$, we can write that

$$\begin{aligned}\chi\,\partial\, F &= \chi\,(\textstyle\sum_{E\in\mathcal{E}_m} R_{F,E}\, E)\\ &= \textstyle\sum_{E\in\mathcal{E}_m} R_{F,E} \sum_{e\in\mathcal{E}_{\tilde{m}}} \chi^e_E\, e\\ &= \textstyle\sum_{e\in\mathcal{E}_{\tilde{m}}} [\sum_{E\in\mathcal{E}_m} R_{F,E}\, \chi^e_E\,]\, e = \sum_{e\in\mathcal{E}_{\tilde{m}}} [\sum_{E\subset F} R_{F,E}\, \chi^e_E\,]\, e.\end{aligned}$$

On the other hand, we have

$$\begin{aligned}\partial\,\chi\, F &= \partial\,(\textstyle\sum_{f\in\mathcal{F}_{\tilde{m}}} \chi^f_F\, f)\\ &= \textstyle\sum_{f\in\mathcal{F}_{\tilde{m}}} \chi^f_F \sum_{e\in\mathcal{E}_{\tilde{m}}} r_{f,e}\, e\\ &= \textstyle\sum_{e\in\mathcal{E}_{\tilde{m}}} [\sum_{f\in\mathcal{F}_{\tilde{m}}} \chi^f_F\, r_{f,e}]\, e = \sum_{e\in\mathcal{E}_{\tilde{m}}} [\sum_{f\subset F} \chi^f_F\, r_{f,e}]\, e.\end{aligned}$$

We compare again the two quantities in brackets, noting again that both brackets vanish for each $e \in \mathcal{E}_{\tilde{m}}$ such that $e \not\subset F$. Assuming therefore $e \subset F$ (Figure 7), we have two cases: Either there exists $E \in \mathcal{E}_m$ such that $e \subset E$ and $E \subset F$, or not (see Figures 7 and 8, center and right-hand part, respectively).

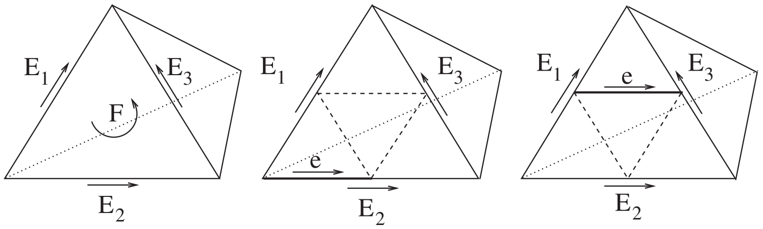


FIG. 7. *For edge $e \in \mathcal{E}_{\tilde{m}}$ belonging to face F, either there exists $E \subset F$ such that $e \subset E$ (center) or not (right).*

According to the situation at the center of Figure 7, we have

$$\begin{aligned}\textstyle\sum_{E\subset F} R_{F,E}\, \chi^e_E &= R_{F,E_1}\, \chi^e_{E_1} + R_{F,E_2}\, \chi^e_{E_2} + R_{F,E_3}\, \chi^e_{E_3}\\ &= (-1)(0) + (1)(1) + (1)(0) = 1.\end{aligned}$$

For the situation at the right-hand side of Figure 7, since $e \not\subset E_i$, $i = 1, 2, 3$, we have

$$\chi^e_{E_i} = 0 \quad \forall\, i \quad \text{so that} \quad \sum_{E\subset F} R_{F,E}\, \chi^e_E = 0.$$

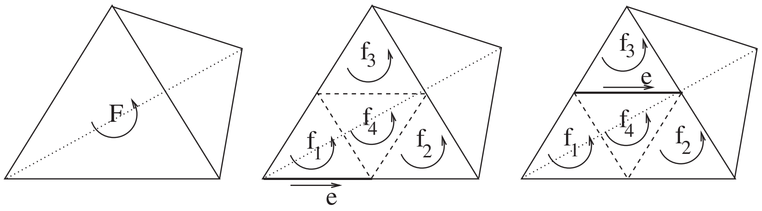


FIG. 8. *For edge $e \in \mathcal{E}_{\tilde{m}}$ belonging to face F, either there exists only one face $f \in \mathcal{E}_{\tilde{m}}$ such that $e \subset f$ (center) or two (right).*

Let us do the same reasoning for the other quantity, looking at Figure 8. For the situation at the center of Figure 8, we have

$$\begin{aligned}\textstyle\sum_{f\subset F} \chi_F^f\, r_{f,e} &= \chi_F^{f_1}\, r_{f_1,e} + \chi_F^{f_2}\, r_{f_2,e} + \chi_F^{f_3}\, r_{f_3,e} + \chi_F^{f_4}\, r_{f_4,e}\\ &= (1)(1) + (0)(1) + (0)(1) + (0)(1) = 1.\end{aligned}$$

For the situation at the right-hand side of Figure 8, we have

$$\begin{aligned}\textstyle\sum_{f\subset F} r_{f,e}\, \chi_F^f &= \chi_F^{f_1}\, r_{f_1,e} + \chi_F^{f_2}\, r_{f_2,e} + \chi_F^{f_3}\, r_{f_3,e} + \chi_F^{f_4}\, r_{f_4,e}\\ &= (0)(1) + (0)(1) + (1)(1) + (-1)(1) = 0.\end{aligned}$$

Summing up, for a given $e \in \mathcal{E}_{\tilde{m}}$, the two quantities in brackets take the same value (1 or 0), due to the definition of the incidence matrices R and r and coefficients χ_E^e and χ_F^f.

Finally, for $p = 3$, we have

$$\begin{aligned}\chi\,\partial T &= \chi\,(\textstyle\sum_{F\in\mathcal{F}_m} D_{T,F}\, F)\\ &= \textstyle\sum_{F\in\mathcal{F}_m} D_{T,F} \sum_{f\in\mathcal{F}_{\tilde{m}}} \chi_F^f\, f\\ &= \textstyle\sum_{f\in\mathcal{F}_{\tilde{m}}} (\sum_{F\in\mathcal{F}_m} D_{T,F}\, \chi_F^f)\, f = \sum_{f\in\mathcal{F}_{\tilde{m}}} [\sum_{F\subset T} D_{T,F}\, \chi_F^f]\, f.\end{aligned}$$

On the other hand,

$$\begin{aligned}\partial\,\chi\, T &= \partial\,(\textstyle\sum_{t\in\mathcal{T}_{\tilde{m}}} \chi_T^t\, t)\\ &= \textstyle\sum_{t\in\mathcal{T}_{\tilde{m}}} \chi_T^t \sum_{f\in\mathcal{F}_{\tilde{m}}} d_{t,f}\, f\\ &= \textstyle\sum_{f\in\mathcal{F}_{\tilde{m}}} (\sum_{t\in\mathcal{T}_{\tilde{m}}} \chi_T^t\, d_{t,f})\, f = \sum_{f\in\mathcal{F}_{\tilde{m}}} [\sum_{t\subset T} \chi_T^t\, d_{t,f}]\, f.\end{aligned}$$

We compare again the two quantities in brackets, assuming $f \subset T$. We have two cases: either there exists $F \in \mathcal{F}_m$ such that $f \subset F$ or not (see Figure 9's left-hand and right-hand part, respectively). If f is part of, say, F_1, then

$$\begin{aligned}\textstyle\sum_{F\subset T} D_{T,F}\, \chi_F^f &= D_{T,F_1}\, \chi_{F_1}^f + D_{T,F_2}\, \chi_{F_2}^f + D_{T,F_3}\, \chi_{F_3}^f + D_{T,F_4}\, \chi_{F_4}^f\\ &= (1)(1) + (1)(0) + (1)(0) + (1)(0) = 1.\end{aligned}$$

If $f \not\subset F_i$, whatever $F_i \subset T$, then

$$\chi_{F_i}^f = 0 \quad \forall\, i; \qquad \text{hence} \quad \sum_{F\subset T} D_{T,F}\, \chi_F^f = 0.$$

For the other bracketed term, either f is part of some F_1, and there exists only one $t \in \mathcal{T}_{\tilde{m}}$ containing f, namely t^*, so that

$$\sum_{t\subset T} d_{t,f}\, \chi_T^t = d_{t^*,f}\, \chi_T^{t^*} = (1)(1) = 1,$$

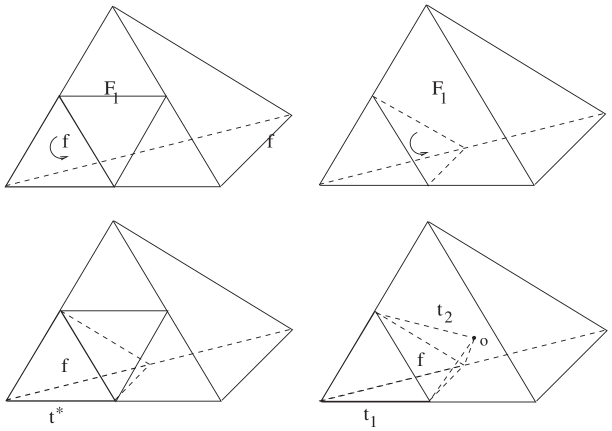


FIG. 9. *For face $f \in \mathcal{F}_{\tilde{m}}$ belonging to tetrahedron $T \in \mathcal{T}_m$, either there exists $F \subset T$ such that $f \subset F$ (left) or not (right). In the first case, there exists a unique tetrahedron $t^* \in \mathcal{T}_{\tilde{m}}$ containing f, and in the second case, two tetrahedra $t_1, t_2 \in \mathcal{T}_{\tilde{m}}$. The normal subdivision of T is not completely shown to make the figure clearer (o is the barycenter of T).*

or $f \not\subset F_i$ whatever $F_i \subset T$. Then, f is inside T and is thus shared by two tetrahedra of $\tilde{m}$, say t_1 and t_2. So,

$$\sum_{t \subset T} d_{t,f}\, \chi_{T}^{t} = d_{t_1,f}\, \chi_{T}^{t_1} + d_{t_2,f}\, \chi_{T}^{t_2} = (1)(1) + (-1)(1) = 0.$$

Summing up, for a given $f \in \mathcal{F}_{\tilde{m}}$, the two quantities in brackets take the same value (1 or 0), owing to the definition of the incidence matrices D and d and coefficients χ_{F}^{f} and χ_{T}^{t}.

This completes the proof, which has been detailed for all cases, much beyond logical necessity, to show how the incidence matrices and the two maps interact. Later on, we will consider only one case, the others being on the same pattern. ∎

PROPOSITION 4.5. *The map π defined in* (5) *is a chain map, i.e., $\partial \pi = \pi \partial$.*

Proof. We consider the case $p = 2$ to detail the proof. Then

$$\begin{aligned} \pi\, \partial\, f &= \pi\, \big(\textstyle\sum_{e \in \mathcal{E}_{\tilde{m}}} r_{f,e}\, e\big) \\ &= \textstyle\sum_{e \in \mathcal{E}_{\tilde{m}}} r_{f,e} \sum_{E \in \mathcal{E}_m} \langle w^{E}\ ;\ e \rangle\, E \\ &= \textstyle\sum_{E \in \mathcal{E}_m} \big[\sum_{e \in \mathcal{E}_{\tilde{m}}} r_{f,e}\, \langle w^{E}\ ;\ e \rangle\big]\, E. \end{aligned}$$

On the other hand, we can write

$$\begin{aligned} \partial\, \pi\, f &= \partial\, \big(\textstyle\sum_{F \in \mathcal{F}_m} \langle w^{F}\ ;\ f \rangle\, F\big) \\ &= \textstyle\sum_{F \in \mathcal{F}_m} \langle w^{F}\ ;\ f \rangle \sum_{E \in \mathcal{E}_m} R_{F,E}\, E \\ &= \textstyle\sum_{E \in \mathcal{E}_m} \big[\sum_{F \in \mathcal{F}_m} R_{F,E}\, \langle w^{F}\ ;\ f \rangle\big]\, E. \end{aligned}$$

We now recall that, when $p = 2$,

$$\mathrm{d}\, w^{E} = \sum_{F \in \mathcal{F}_m} R_{F,E}\, w^{F} \tag{8}$$

so that

$$\begin{aligned}\partial\,\pi\, f &= \textstyle\sum_{E\in\mathcal{E}_m} [\,\sum_{F\in\mathcal{F}_m} R_{F,E}\,\langle w^{F}\ ;\ f\rangle\,]\,E\\ &= \textstyle\sum_{E\in\mathcal{E}_m} [\,\langle \mathrm{d}\, w^{E}\ ;\ f\rangle\,]\,E\\ &= \textstyle\sum_{E\in\mathcal{E}_m} [\,\langle w^{E}\ ;\ \partial f\rangle\,]\,E\\ &= \textstyle\sum_{E\in\mathcal{E}_m} [\,\sum_{e\in\mathcal{E}_{\tilde m}} r_{f,e}\,\langle w^{E}\ ;\ e\rangle\,]\,E\\ &= \pi\,\partial\, f.\end{aligned}$$

Note the two ingredients of the proof: the Stokes theorem and the structural property, (8), of the Whitney complex. For other dimensions, the proof is similar: Just change R and r into G and g if $p = 1$, into D and d if $p = 3$. ∎

Remark 4.6. The chain map $\chi : C(m) \to C(\tilde m)$ can be defined, similarly to π, as follows: Given a p-simplex S of the coarse complex m, set

$$\chi(S) = \sum_{s\in\mathcal{S}^p_{\tilde m}} \langle w^s\ ;\ S\rangle\, s = \sum_{s\in\mathcal{S}^p_{\tilde m}} \chi^s_{S}\, s. \tag{9}$$

In the nested case, definitions (9) and (4) coincide. In the nonnested case, (9) is a generalization of (4); the property $\pi\,\chi = 1$ is lost, and the coefficients π^S_s and χ^s_S cannot be computed "by hands" as we shall see in the next section for nested grids.

5. Application.

We explain how the map π can be used to design a multigrid algorithm for the solution of linear systems arising from the use of Whitney elements on tetrahedra to discretize a given differential (e.g., electromagnetic) problem. The detailed analysis of the mesh-independent convergence and performances of the multigrid algorithm based on π is not considered here. We refer to [3, 8, 14] for rigorous theoretical and numerical results in the edge element framework, and to [6] for a formulation and application of the multigrid algorithm on hexahedral meshes.

As already pointed out in the introduction, the motivation for this approach comes from the analysis of the error on the numerical solution in the frequency domain. We recall the basic multigrid algorithm, assuming a two-grid method for simplicity. Let h and H denote, respectively, the maximal diameter of tetrahedra in the fine $\tilde m$ and coarse m grids. Let $\mathcal{V}_h$ and $\mathcal{V}_{H}$ be the underlying finite dimensional spaces of cochains, with $\dim(\mathcal{V}_h) > \dim(\mathcal{V}_{H})$, consistent with $h < H$. One wishes to solve the linear system $A_h\mathbf{u}_h = \mathbf{b}_h$ in $\mathcal{V}_h$, assuming that the matrix A_h is symmetric and positive definite (as is usually the case for matrices resulting from finite element discretizations of a variational problem). Denoted by $(\mathbf{u},\mathbf{v})$ the scalar product of $\mathbf{u},\mathbf{v} \in \mathcal{V}_h$, solving $A_h\mathbf{u}_h = \mathbf{b}_h$ in $\mathcal{V}_h$ is then equivalent to finding the minimizer $\mathbf{u}_h \in \mathcal{V}_h$ of the quadratic functional $\Phi(\mathbf{u}) = \frac{1}{2}(A_h\mathbf{u},\mathbf{u}) - (\mathbf{b}_h,\mathbf{u})$. In what follows, M_h represents a suitable preconditioner for A_h. The maps $R^{H}_h : \mathcal{V}_h \to \mathcal{V}_{H}$, usually called *restriction operator*, and $P^h_{H} : \mathcal{V}_{H} \to \mathcal{V}_h$, called *prolongation operator*, are full-rank linear cochain-to-cochain operators. The so-called *two-level V-cycle* of the multigrid procedure reads as follows:

1. *Fine grid presmoothing*: from $\mathbf{u}_h^0 \in \mathcal{V}_h$ and for $k = 1, \ldots, n_1$, do

$$\mathbf{u}_h^k = \mathbf{u}_h^{k-1} + M_h\,(\mathbf{b}_h - A_h \mathbf{u}_h^{k-1}). \tag{10}$$

2. *Coarse grid correction*: given $\mathbf{r}_h^{n_1} = \mathbf{b}_h - A_h \mathbf{u}_h^{n_1}$ in $\mathcal{V}_h$, do

   restict the residual on the coarse grid: $\mathbf{r}_H \leftarrow R_h^H \mathbf{r}_h^{n_1}$.
   solve the residual problem: $A_H \mathbf{z}_H = \mathbf{r}_H$.
   correct the solution in $\mathcal{V}_h$: $\mathbf{u}_h^{n_1} \leftarrow \mathbf{u}_h^{n_1} + P_H^h\, \mathbf{z}_H$.

3. *Fine grid postsmoothing*: from $\mathbf{u}_h^{n_1} \in \mathcal{V}_h$ and for $k = 1, \ldots n_2$, do (10).

In the *fine grid presmoothing* step, one iteratively solves $A_h \mathbf{u}_h = \mathbf{b}_h$ in $\mathcal{V}_h$ by a basic iterative method. High frequency errors are thus well eliminated, and once this is achieved in, e.g., n_1 iterations, further fine grid iterations would not improve significantly the convergence rate. In the *coarse grid correction*, one tries to correct $\mathbf{u}_h^{n_1}$ on the coarse space $\mathcal{V}_H$. The coarse correction $\mathbf{z}_H$ minimizes $\Phi(\mathbf{u}_h^{n_1} + P_H^h \mathbf{z}_H)$ over $\mathcal{V}_H$. This is equivalent to solving $(P_H^h)^t A_h P_H^h \mathbf{z}_H = (P_H^h)^t \mathbf{r}_h^{n_1}$ on $\mathcal{V}_H$. Thus, $A_H = (P_H^h)^t A_h P_H^h$ and the R_h^H of step 2 is the transpose $(P_H^h)^t$ [16]. On m, the low frequency errors of $\tilde{m}$ manifest themselves as relatively high frequency errors and are thus eliminated efficiently using again simple iterative smoothing methods. If the coarsest grid has been reached, the coarse system has to be solved exactly, by a direct solver, which can be done with little computational effort due to the small number of unknowns. Otherwise, the three-step multigrid procedure can be repeated recursively to solve the residual problem, as many times as the number of coarsening levels m one considers, starting from the fine one $\tilde{m}$. Each grid level is responsible for eliminating a particular frequency bandwidth of the error. Finally, in the *fine grid postsmoothing* step, one solves iteratively n_2 times in $\mathcal{V}_h$ the system $A_h \mathbf{u}_h = \mathbf{b}_h$, starting from $\mathbf{u}_h^{n_1} + P_H^h \mathbf{z}_H$, to eliminate high frequency errors on the term $P_H^h \mathbf{z}_H$.

Our proposal is now to define the operator P_H^h as the dual of the chain map π. Indeed, recall that $\mathcal{V}_h$ and $\mathcal{V}_H$ are spaces of p-cochains, while π is defined on p-chains (see Definition 4.2). There is therefore a natural prolongation operator P_H^h, defined as the dual of π, i.e., by $\langle \mathbf{u}\,;\, \pi(s) \rangle = \langle P_H^h \mathbf{u}\,;\, s \rangle$ for all p-chains s and p-cochains $\mathbf{u} \in \mathcal{V}_H$, as suggested by the diagram below.

$$\begin{tikzcd}
\mathcal{V}_h \arrow[r, phantom, ""] & C(\tilde{m}) \arrow[d, "\pi"] \\
\mathcal{V}_H \arrow[u, "P_H^h"] \arrow[r, phantom, "\langle \cdot\,;\,\cdot \rangle"] & C(m)
\end{tikzcd}$$

Taking dual bases on both $\mathcal{V}_h$ and $\mathcal{V}_H$ as explained in section 2.3, the matrix representation of P_H^h has entries $(P_H^h)_S^s = \pi_s^S$ (cf. (5)). Recall that S and s here are two simplices of same dimension p in m and $\tilde{m}$, respectively, so that there are distinct prolongation operators for each p, i.e., for degrees of freedom based on nodes, edges, faces, and volumes.

We now detail the calculation of π in the case where a number of coarse tetrahedra have undergone one normal subdivision, i.e., by using (3) with $\kappa = 1$ (see Figure 5), thus being divided into twelve small ones, while tetrahedra in the transition layer are split into four or two small ones; hence we have the three cases considered below. It is important to remark that the chain-map coefficients π_s^S we search, defined as $\langle w^S\,;\, s \rangle$ in (5), do not depend on the shape of S and s but on their relative position and orientation. Their computation relies on the following two obvious lemmas.

LEMMA 5.1. *Let s be a p-simplex and w a linear p-differential form, linear with respect to position x. Then $\int_s w = \int_s w(\mathbf{x}_s)$, where $\mathbf{x}_s$ denotes the barycenter of s.*

This replaces a linear differential form by a constant one. For these, one has:

LEMMA 5.2. *Let s be a p-simplex and w a constant p-differential form, L a linear map which sends simplex s to simplex s'. Then $\int_{s'} w = \det(L) \int_s w$.*

Case I. Tetrahedron T has been divided into twelve small tetrahedra t.

For $p = 3$, let w^T be the scalar function associated to T (section 2.1), that is the constant such that $\langle w^T \; ; \; T\rangle = 1$. Computing $\langle w^T \; ; \; t\rangle$ thus amounts to finding the relative volume of t (an affine notion, not a metric one) with respect to T. This is $\frac{1}{8}$ for the four tetrahedra t sharing a vertex with T (scaling factor $\frac{1}{2}$, to the cube), which leaves $\frac{1}{2}$ to be shared equally between the 8 tetrahedra with a vertex at the barycenter o of T. So one has

$\langle w^T \; ; \; t\rangle = \pm\frac{1}{8}$ for all t not contained in the core octahedron;

$\langle w^T \; ; \; t\rangle = \pm\frac{1}{16}$ for all t contained in the core octahedron.

The sign ± 1 depends on the relative orientation between t and T.

Note that the Lebesgue measure of t or of T played no role here: Considerations of *scaling* and *symmetry* suffice to do the job, as will also be the case for other values of p. We give only the results without further comments. Only the nonzero coefficients are displayed.

For $p = 2$, there are four different situations, depending on where the small face f is located with respect to the big one F and we refer again to Figure 5, $\kappa = 1$. The coefficients $\langle w^F \; ; \; f\rangle$ are the fluxes of the vector function w^F across the small faces f. Let F be $\{k, l, n\}$ for definiteness. Using $\langle w^F \; ; \; F\rangle = 1$, scaling, and symmetry, then

$\langle w^F \; ; \; f\rangle = \pm\frac{1}{4}$ for all $f \subset F$ such as $f = \{k, kl, kn\}$;

$\langle w^F \; ; \; f\rangle = \pm\frac{1}{8}$ for all $f \not\subset F$ and with three vertices at mid-points not in F, such as $f = \{mk, lm, mn\}$;

$\langle w^F \; ; \; f\rangle = \pm\frac{1}{8}$ for all $f \not\subset F$ and with three vertices at mid-points and one edge on F, such as $f = \{kl, kn, km\}$;

$\langle w^F \; ; \; f\rangle = \pm\frac{1}{16}$ for all $f \not\subset F$ and with two vertices at mid-points and the third one at o, such as $f = \{kl, kn, o\}$.

For $p = 1$, the coefficients $\langle w^E \; ; \; e\rangle$ are the circulations of the vector function w^E along the small edges e. Consider $E = \{k, n\}$. Again, $\langle w^E \; ; \; E\rangle = 1$, scaling and symmetry yield

$\langle w^E \; ; \; e\rangle = \pm\frac{1}{2}$ for all $e \subset E$, such as $e = \{k, kn\}$;

$\langle w^E \; ; \; e\rangle = \pm\frac{1}{4}$ for all $e \not\subset E$ with vertices at mid-points, one of which belongs to E, such as $e = \{kn, kl\}$;

$\langle w^E \; ; \; e\rangle = \pm\frac{1}{4}$ for all $e \not\subset E$ with vertices at mid-points and parallel to E, such as $e = \{kl, ln\}$;

$\langle w^E \; ; \; e\rangle = \pm\frac{1}{8}$ for all $e \not\subset E$ with one vertex at o and one at a mid-point not in E, such as $e = \{ln, o\}$ or $e = \{mn, o\}$.

For $p = 0$, the coefficient $\langle w^N \; ; \; n\rangle$ is the value of the scalar function w^N at node

n. Take for example $N = k$. Using $\langle w^{N} \; ; \; N\rangle = 1$ and linearity, one gets

$\langle w^{N} \; ; \; n\rangle = 1$ for $n \subset N$;

$\langle w^{N} \; ; \; n\rangle = \frac{1}{2}$ for n at the middle of an edge
incident on N, such as kn or km;

$\langle w^{N} \; ; \; n\rangle = \frac{1}{4}$ for n at o.

Case II. Here, T is a tetrahedron of the transition layer, divided into four small tetrahedra t, as shown in Figure 10 (left and right).

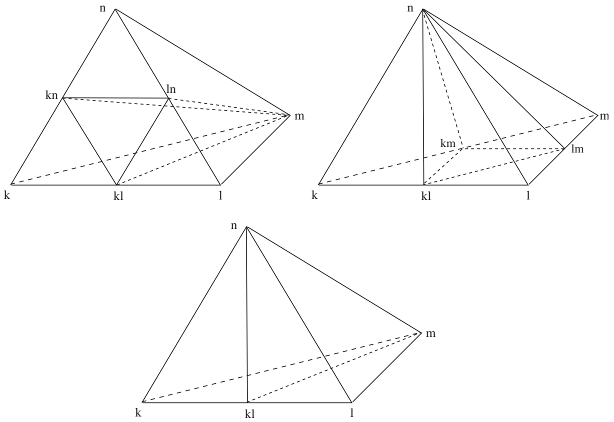


FIG. 10. *Tetrahedron T in the transition layer and division in four and two tetrahedra t.*

For $p = 3$, by symmetry, $\langle w^{T} \; ; \; t\rangle = \pm\frac{1}{4}$ for all t contained in T.

For $p = 2$, two cases. If F (taken here as $\{k, l, n\}$ for the sake of the example) is divided in four as in Figure 10 (left), then

$\langle w^{F} \; ; \; f\rangle = \pm\frac{1}{4}$ for all $f \subset F$, such as $f = \{k, kl, kn\}$,

while F is divided in two, as in Figure 10 (right), then

$\langle w^{F} \; ; \; f\rangle = \pm\frac{1}{2}$ for all $f \subset F$, such as $f = \{k, kl, n\}$;

$\langle w^{F} \; ; \; f\rangle = \pm\frac{1}{4}$ for all f neither in F nor in $F' \neq F$, such as $f = \{kl, km, n\}$.

For $p = 1$, two cases again. If $E \subset F$ and F is divided in four, then (with $E = \{k, n\}$ for illustration)

$\langle w^{E} \; ; \; e\rangle = \pm\frac{1}{2}$ for all $e \subset E$, such as $e = \{k, kn\}$;

$\langle w^{E} \; ; \; e\rangle = \pm\frac{1}{4}$ for all $e \not\subset E$ with vertices at mid-points,
one of which belongs to E, such as $e = \{kn, kl\}$;

$\langle w^{E} \; ; \; e\rangle = \pm\frac{1}{4}$ for all $e \not\subset E$ parallel to E with
vertices at mid-points, such as $e = \{kl, ln\}$,

while if F is divided in two, then

$\langle w^{E} \; ; \; e\rangle = 1$ for $e = E$;
$\langle w^{E} \; ; \; e\rangle = \pm\frac{1}{2}$ for $e = \{kl, n\}$ and $\{km, n\}$.

For $p = 0$, by linearity,

$\langle w^{N} \; ; \; n\rangle = 1$ for n at N ;
$\langle w^{N} \; ; \; n\rangle = \frac{1}{2}$ for all n mid-points of edges with one extremity at N, such as kn or kl.

Case III. Now T, is halved, as shown in Figure 10 (bottom).

For $p = 3$, $\langle w^{T} \; ; \; t\rangle = \pm\frac{1}{2}$ for all t contained in T.

For $p=2$ and F halved (as in Figure 10, bottom, take $F = \{k, l, n\}$ for illustration),

$\langle w^{F} \; ; \; f\rangle = \pm\frac{1}{2}$ for all $f \subset F$, such as $f = \{k, kl, n\}$.

If $F' \neq F$ is halved, then

$\langle w^{F} \; ; \; f\rangle = 1$ for $f \equiv F$;
$\langle w^{F} \; ; \; f\rangle = \pm\frac{1}{2}$ for $f = \{k, lm, n\}$.

For $p = 1$, if $E \subset F$ and F is divided in two but not E, then (consider $E = \{k, n\}$)

$\langle w^{E} \; ; \; e\rangle = 1$ for $e = E$;
$\langle w^{E} \; ; \; e\rangle = \pm\frac{1}{2}$ for $e = \{kl, n\}$.

If $E \subset F$ and F is not divided in two, then

$\langle w^{E} \; ; \; e\rangle = 1$ for $e = E$

and last, if $E \subset F$ and F is divided in two along E, then

$\langle w^{E} \; ; \; e\rangle = \pm\frac{1}{2}$ for $e \subset E$, such as $e = \{k, kn\}$.

For $p = 0$, finally

$\langle w^{N} \; ; \; n\rangle = 1$ for n at N;
$\langle w^{N} \; ; \; n\rangle = \frac{1}{2}$ for all n mid-points of edges with one extremity at N, such as kn, kl or km.

Remark 5.3. The strategy adopted to compute the coefficients π_s^S, $S \in \mathcal{S}_m^p$ and $s \in \mathcal{S}_{\tilde{m}}^p$, can be used when dealing with quadratic, cubic, etc., differential forms, provided that the integration rule is modified accordingly. Therefore, the main problem with Whitney elements of order $r > 1$ (see, e.g., [9]) is the definition on a p-simplex, $p = 2, 3$, of an integration rule which is exact for all polynomials of degree r (on 1-simplices we can use Gaussian quadratures). This problem is far from being trivial and is linked to another one, namely, the location in a p-simplex of the degrees of freedom associated with Whitney elements of order $r > 1$. Both problems will be addressed in future work.

As we have stressed, computing the coefficients of the chain-map π is a metric-independent process. Implementation, however, may have to be done in a code conceived in terms of proxy vector fields, with an underlying metric, instead of differential forms. Hence we have the following description of the procedure, where $|t|$ denotes the volume of tetrahedron t, $|f|$ the area of face f, and $|e|$ the length of edge e. We use $\mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_t$ to denote the barycenters of edge e, face f, and tetrahedron t, respectively. Points $\mathbf{x}_k, \mathbf{x}_\ell, \mathbf{x}_m, \mathbf{x}_n$ are the vertices of t or T. Moreover, $\mathbf{t}_e$ denotes the unit vector along the mesh side e, and $\mathbf{n}_f$ stands for the unit vector normal to the mesh face f. For completeness, we throw in the computation of the other chain map, χ. Thanks to Lemmas 5.1 and 5.2, the following algorithm, though relying on metric elements such as dot product, etc., does implement in the nested case (up to floating-point errors, and barring clerical mistakes of ours . . .) the metric-free computation of the prolongation/restriction operator we have detailed.

```
Loop over $S$, the $p$-simplices of $m$
  Loop over $s \subset S$, with $s$ the $p$-simplices of $\tilde{m}$

    Computation of $\pi_s^S$

        $p = 0,$        $\pi_n^N = w^N(\mathbf{x}_n),$                              $S = N,\ s = n;$
          $p = 1,$        $\pi_e^E = |e|(w^E(\mathbf{x}_e) \cdot \mathbf{t}_e),$          $S = E,\ s = e;$
            $p = 2,$        $\pi_f^F = |f|(w^F(\mathbf{x}_f) \cdot \mathbf{n}_f),$          $S = F,\ s = f;$
              $p = 3,$        $\pi_t^T = |t|,$                                          $S = T,\ s = t.$

    Computation of $\chi_S^s$

      $p = 0,$        $\chi_N^n = 1$              $S = N,\ s = n,\ n \equiv N;$
        $p = 1,$        $\chi_E^e = 1\ (-1)$        $S = E,\ s = e,\ \mathbf{t}_e \cdot \mathbf{t}_E > 0\ (< 0),$
          $p = 2,$        $\chi_F^f = 1\ (-1)$        $S = F,\ s = f,\ \mathbf{n}_f \cdot \mathbf{n}_F > 0\ (< 0),$
            $p = 3,$        $\chi_T^t = 1\ (-1)$        $S = T,\ s = t$  and

              $(\mathbf{x}_\ell - \mathbf{x}_k) \cdot [(\mathbf{x}_m - \mathbf{x}_k) \times (\mathbf{x}_n - \mathbf{x}_k)] > 0\ (< 0)$  for both tetrahedra.

  end loop over $s$
end loop over $S$.
```

REFERENCES

[1] M. A. Armstrong, *Basic Topology*, Springer-Verlag, New York, 1983.

[2] M. Berger and B. Gostiaux, *Géométrie différentielle: variétés, courbes et surfaces*, Presses Universitaires de France, Paris, 1987.

[3] P. B. Bochev, C. J. Garasi, J. J. Hu, A. C. Robinson, and R. S. Tuminaro, *An improved algebraic multigrid method for solving Maxwell's equations*, SIAM J. Sci. Comp., 25 (2003), pp. 623–642.

[4] A. Bossavit, *Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements*, Academic Press, New York, 1998.

[5] A. Bossavit, *Generating Whitney forms of polynomial degree one and higher*, IEEE Trans. Magn., 38 (2002), pp. 341–344.

[6] M. Clemens, S. Feigh, and T. Weiland, *Geometric multigrid algorithms using the conformal finite integration technique*, IEEE Trans. Magn., 40 (2004), pp. 1065–1068.

[7] J. Harrison, *Stokes' theorem for nonsmooth chains*, Bull. Amer. Math. Soc. (N.S.), 29 (1993), pp. 235–242.

[8] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.

[9] J. C. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.

[10] J. C. Nédélec, *A new family of mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1986), pp. 57–81.

[11] G. Nicolas, F. Arnoux-Guisse, and O. Bonnin, *Adaptive meshing for 3D finite element software*, in 9th International Conference in Finite Elements in Fluids, Venezia, 1995.

[12] F. Rapetti, F. Dubois, and A. Bossavit, *Integer matrix factorization for mesh defects detection*, C. R. Acad. Sci. Paris Sér. I math., 334 (2002), pp. 717–720.

[13] S. M. Rao, D. R. Wilton, and A. W. Glisson, *Electromagnetic scattering by surfaces of arbitrary shape*, IEEE Trans. Antennas and Propagation, 30 (1982), pp. 409–418.

[14] S. Reitzinger and J. Schöberl, *An algebraic multigrid method for finite element discretizations with edge elements*, Numer. Linear Algebra Appl., 9 (2002), pp. 223–238.

[15] S. Suuriniemi, *Homological Computations in Electromagnetic Modeling*, Ph.D. thesis, Tampere University of Technology, Tampere, Finland, 2004.

[16] B. Smith, P. Bjørstad, and W. Gropp, *Domain Decomposition*, Cambridge University Press, New York, 1996.

[17] J. Stillwell, *Classical Topology and Combinatorial Group Theory*, Grad. Texts in Math. 72, Springer-Verlag, 1993.

[18] H. Whitney, *Geometric Integration Theory*, Princeton University Press, 1957.

[19] B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Lect. Notes Comput. Sci. Eng. 17, Springer, Heidelberg, 2001.

[20] K. Yosida, *Functional Analysis*, 6th ed., Springer, 1980.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2098–2120

FIRST-ORDER SYSTEM $\mathcal{LL}^*$ (FOSLL*) FOR GENERAL SCALAR ELLIPTIC PROBLEMS IN THE PLANE*

T. A. MANTEUFFEL[†], S. F. MCCORMICK[†], J. RUGE[†], AND J. G. SCHMIDT[‡]

Abstract. This paper develops new first-order system $\mathcal{LL}^*$ (FOSLL*) formulations for scalar elliptic partial differential equations. It extends the work of [Z. Cal et al., *SIAM J. Numer. Anal.*, 39 (2001), pp. 1418–1445], where the FOSLL* methodology was first introduced. One focus of that paper was to develop FOSLL* formulations that allow the use of H^1-conforming finite element spaces and optimal multigrid solution techniques to construct L^2 approximations of the dependent variables in the presence of discontinuous coefficients. The problems for which this goal was achieved were limited to those with no reaction term and with Dirichlet and Neumann boundaries that were individually connected; that is, each had at most one component. Here, new FOSLL* formulations are developed to achieve the same goals on a wider class of problems, including problems with reaction terms, Dirichlet and Neumann boundaries with multiple components, reentrant corners, and points at which Dirichlet and Neumann boundaries meet with an inner angle greater than $\pi/2$. The efficiency of the improved FOSLL* formulations is illustrated by a series of numerical examples.

Key words. first-order systems, least-squares methods, finite element methods, discontinuous coefficients, reduced regularity, nonsmooth solutions

AMS subject classification. 65N30

DOI. 10.1137/S0036142903430402

1. Introduction.

First-order system $\mathcal{LL}^*$ (FOSLL*) was developed in an earlier paper [12] as a numerical method for solving partial differential equations (PDEs) that do not exhibit the regularity required by standard first-order system least squares (FOSLS [10, 11]). The purpose here is to extend the class of problems to which the FOSLL* approach can be efficiently applied. While we include a brief discussion of the context of this development below, the interested reader should consult [12] for more background and historical perspective.

Standard FOSLS recasts the original problem as an expanded first-order system, $Lu = f$, to which a least-squares principle is then applied. The usual goal is to reformulate the original problem as the minimization of a functional, $\|Lu - f\|^2$, whose bilinear part is equivalent to the product H^1 norm (i.e., the square root of the bilinear part is continuous and coercive in the norm formed by summing the H^1 norms applied to each dependent variable). This product H^1-equivalence means that the minimization process amounts to solving a weakly coupled system of scalar elliptic equations, which, in turn, implies that H^1-conforming finite element spaces and multigrid solvers can be used to full efficiency. Unfortunately, standard FOSLS is product H^1-equivalent only under sufficient smoothness assumptions on the original problem (e.g., the domain, coefficients, and data). Inverse-norm versions of FOSLS could be used when the problem lacks sufficient smoothness, but these methods tend to lose

[*] Received by the editors June 27, 2003; accepted for publication (in revised form) April 5, 2005; published electronically December 30, 2005. This work was sponsored by the National Science Foundation under grant DMS-0420873 and by the Department of Energy under grants DE-FC02-01ER25479 and DE-FG02-03ER25574.

http://www.siam.org/journals/sinum/43-5/43040.html

[†] Department of Applied Mathematics, University of Colorado at Boulder, Boulder, CO 80309-0526 (tmanteuf@colorado.edu, stevem@colorado.edu, john.ruge@colorado.edu).

[‡] C&C Research Lab, NEC Europe Ltd., Rathausallee 10, 53757 St. Augustin, Germany (schmidt@ccrl-nece.de).

efficiency, especially for problems with widely varying coefficients.

Our purpose here is to continue the development of a potentially more efficient alternative, FOSLL*. As with FOSLS, the FOSLL* approach begins by recasting the original problem as an expanded first-order system, $Lu = f$. Now, however, instead of applying a least-squares principle to this *primal* problem, we introduce the *dual normal equations*, $LL^*w = f$, defined in terms of dual variable w and adjoint L^*. Note that $f = Lu$, so that $LL^*w = Lu$, which are the normal equations for the dual problem, $L^*w = u$. The original problem can now be recast as one of minimizing the functional, $\|L^*w - u\|^2$, which has the same minimizer as the functional $\|L^*w\|^2 - 2\,\langle w\,, f\rangle$.

If H^1-conforming finite element spaces are used in a standard FOSLS formulation, then it must fail when u is not in H^1. For this choice of finite element spaces, the discrete approximation produced by FOSLS cannot converge to the solution, u. It will, instead, converge optimally to the minimizer of $\|Lv - f\|^2$, that is, to $v \in H^1$ that minimizes $\|L(u-v)\|$. FOSLL* attempts to overcome this limitation by recasting the primal problem in terms of a dual variable, w such that $L^*w = u$. The aim is to use L^* to lift the smoothness of u so that w is in H^1.

Consider the following scalar elliptic problem:

$$\begin{aligned} \nabla\cdot(A\nabla p) - \mathbf{b}\cdot\nabla p - cp &= f \text{ in } \Omega,\\ p &= 0 \text{ on } \Gamma_D,\\ \mathbf{n}\cdot A\nabla p &= 0 \text{ on } \Gamma_N, \end{aligned}$$

and define the flux variable $\mathbf{u} = \nabla p$ (for a complete list of assumptions, see (2.1)–(2.3)). One focus of the earlier paper [12] was to develop the FOSLL* methodology for problems of this type with the reduced regularity that arises by allowing discontinuous A. In that paper, the goal of using H^1-conforming finite element spaces to approximate the flux variable, $\mathbf{u}$, in the L^2 norm and the primal variable, p, in the H^1 norm was achieved through a two-stage procedure. The two-stage procedure described there is applicable only when $c = 0$ and when Γ_D and Γ_N each have at most one component.

The aim of this paper is to expand the class of problems for which H^1-conforming finite element spaces and optimal multigrid solvers can be efficiently used. In section 2, we show that when $c \neq 0$, the original FOSLL* formulation can be modified to achieve this goal, provided that the domain, Ω, is sufficiently smooth. By this we mean that the boundary of Ω contains no reentrant corners or corners at which $\overline{\Gamma}_D$ and $\overline{\Gamma}_N$ meet with an inner angle bigger than $\pi/2$. Such points are referred to as irregular boundary points. These are precisely the conditions under which $H(\nabla\cdot) \cap H(\nabla\times) \subset (H^1)^2$.

In section 3, we develop a new FOSLL* formulation that achieves the goal of allowing accurate approximation using H^1-conforming finite element spaces in the presence of irregular boundary points. The key idea behind this new approach is to first expand the domain of the primal problem in such a manner that the domain of the dual problem remains in a subspace of H^1. Generally, at this point the primal operator, L, is not bijective and the dual operator, L^*, is not surjective. The next step is to apply additional boundary conditions to the slack variables in the primal equations so that fewer boundary conditions are needed for the dual problem. The aim is that the primal operator, L, becomes bijective and the dual operator becomes surjective. This process generally means that the dual equations are not uniquely solvable. However, this is not an issue for the FOSLL* formulation, since any one solution of the dual problem, say, w, yields the primal solution, $L^*w = u$. This approach is limited to problems for which $\Gamma_D \neq \emptyset$. The pure Neumann case remains an open question. In section 3.2, we show that, in the case $\mathbf{b} = \mathbf{0}$ and either $c > 0$ or $c = 0$, the FOSLL*approximation is equivalent to a Galerkin formulation of the original boundary value problem (2.1)–(2.3).

The numerical results presented in section 4 confirm the optimality of H^1-conforming finite element spaces and multigrid solvers for the new FOSLL*formulation. The loss of unique solution for the dual problem is not an issue for the FOSLL* approximation in that we seek any dual solution for which $L^*w = u$. However, the loss of uniqueness does affect the multigrid solution algorithm. In section 4.2, we develop an additional modification that mitigates this effect. Section 5 contains conclusions.

Alternatives to the approach we develop here are described in detail in [1] and include adding H^1 singular basis functions in standard Galerkin methods to enhance the rate of convergence (cf. [23], [14], [8], and [9]) the use of $H(\text{div})$-conforming finite element spaces with mixed formulations (see [7]) or with FOSLS functionals that are based on $H(\text{div})$ (see [10], [20], and [21]) and including $H(\nabla\cdot) \cap H(\nabla\times)$ singular functions in a FOSLS formulation (see [1], [2]). Standard finite element spaces can be used with FOSLS functionals that are weighted to eliminate the overall impact on accuracy of the singular behavior of the flux [14], [19], [24]. Alternatives similar to the FOSLL* formulation use FOSLS based on inverse norms [3], [5], [6], [13].

We begin in the next section with a brief overview of the current theory underlying FOSLS and FOSLL* as a way of exposing the need for modifications of the original FOSLL* approach.

2. General FOSLS and FOSLL* theory.

This section summarizes the principles and theory underlying the FOSLS and FOSLL* methods. For more detail and historical perspective, see [10], [11], [12], [4]. The main goal of this section is to clarify the need for modifying the FOSLL* method introduced in [12].

2.1. Model problem.

Let Ω be a bounded, open, simply connected domain in $\mathbb{R}^2$ with Lipschitz boundary, $\partial\Omega$. Let $\bigcup_{i=1}^{M}(\overline{\Gamma}_{D,i} \cup \overline{\Gamma}_{N,i}) = \partial\Omega$ be a partition of the boundary, interlaced so that every pair $(\Gamma_{D,i}, \Gamma_{D,i+1})$ is separated by a Neumann boundary segment $\Gamma_{N,i}$ and every Neumann pair is similarly separated by a Dirichlet segment. The Neumann and Dirichlet boundaries of the problem are defined by $\Gamma_N := \bigcup_{i=1}^{M} \Gamma_{N,i}$ and $\Gamma_D := \bigcup_{i=1}^{M} \Gamma_{D,i}$, respectively. Let $\mathbf{n}$ be the outward unit normal vector and $\mathbf{t}$ the counterclockwise-oriented tangent vector on $\partial\Omega$. We do not consider the pure Neumann case here, so Γ_D is assumed to have positive measure.

The FOSLL* methodology has application in many contexts, including elliptic systems of PDEs. However, in this paper, we restrict our considerations to the following reaction–convection–diffusion boundary value problem (BVP):

$$\nabla\cdot(A\nabla p) - \mathbf{b}\cdot\nabla p - cp = f \text{ in } \Omega, \tag{2.1}$$

$$p = 0 \text{ on } \Gamma_D, \tag{2.2}$$

$$\mathbf{n}\cdot A\nabla p = 0 \text{ on } \Gamma_N, \tag{2.3}$$

where $f \in L^2(\Omega)$, $0 \le c \in L^\infty(\Omega)$, $\mathbf{b} \in L^\infty(\Omega) \cap H(\nabla\cdot)$, and $A(\mathbf{x})$ is a 2×2 matrix of $L^\infty(\Omega)$-functions that is uniformly symmetric positive definite; i.e., there exists $\lambda_1 \ge \lambda_0 > 0$ such that

$$\lambda_0 \boldsymbol{\xi}\cdot\boldsymbol{\xi} \le \boldsymbol{\xi}\cdot A(\mathbf{x})\boldsymbol{\xi} \le \lambda_1 \boldsymbol{\xi}\cdot\boldsymbol{\xi}$$

for all $\boldsymbol{\xi} \in \mathbb{R}^2$ and $\mathbf{x} \in \Omega$. We also assume that both (2.1)–(2.3) and the adjoint problem,

$$\nabla\cdot(A\nabla p) + \nabla\cdot(\mathbf{b}p) - cp = f \text{ in } \Omega, \tag{2.4}$$

$$p = 0 \text{ on } \Gamma_D, \tag{2.5}$$

$$\mathbf{n}\cdot(A\nabla p + \mathbf{b}p) = 0 \text{ on } \Gamma_N, \tag{2.6}$$

have unique solutions in $H^1(\Omega)$.

We make use of the following standard differential operators:

$$\nabla s = \operatorname{grad} s = \begin{pmatrix}\partial_x s\\ \partial_y s\end{pmatrix}, \qquad \nabla\cdot\mathbf{v} = \operatorname{div}\begin{pmatrix}v_1\\ v_2\end{pmatrix} = \partial_x v_1 + \partial_y v_2,$$

$$\nabla^\perp s = \operatorname{rot} s = \begin{pmatrix}\partial_y s\\ -\partial_x s\end{pmatrix}, \qquad \nabla\times\mathbf{v} = \operatorname{curl}\begin{pmatrix}v_1\\ v_2\end{pmatrix} = -\partial_y v_1 + \partial_x v_2.$$

We use $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ to denote the respective L^2 inner product and norm and $\mathcal{D}$, $\mathcal{R}$, and $\mathcal{N}$ for the respective domain, range, and null space of an operator. We also use $\|\cdot\|_1$ to denote the $H^1(\Omega)$ norm: $\|s\|_1^2 = \|s\|^2 + \|\nabla s\|^2$. As usual, norms of vectors are meant to be taken componentwise, so that $\|\nabla s\| = (\|\frac{\partial s}{\partial x}\|^2 + \|\frac{\partial s}{\partial y}\|^2)^{1/2}$, for example.

2.2. FOSLS. We begin with a brief introduction to the main ideas of FOSLS as a way of providing a foundation for the FOSLL* methodology. We describe how it works for domains without irregular boundary points and show why it may fail for domains with irregular boundary points.

Standard FOSLS transforms BVP (2.1)–(2.3) into a first-order system to which an L^2 norm minimization principle is applied. This transformation can be done by introducing the gradient, $\mathbf{u} = \nabla p$, as a dependent variable and adding the curl constraint, $\nabla\times\mathbf{u} = 0$ on Ω, and tangential boundary condition, $\mathbf{t}\cdot\mathbf{u} = 0$ on Γ_D. The resulting first-order system, then has the form

$$L_0(\mathbf{u},p) = (\mathbf{0}, f, 0)^t \text{ in } \Omega, \tag{2.7}$$

$$\mathbf{t}\cdot\mathbf{u} = 0 \text{ on } \Gamma_D, \tag{2.8}$$

$$\mathbf{n}\cdot A\mathbf{u} = 0 \text{ on } \Gamma_N, \tag{2.9}$$

$$p = 0 \text{ on } \Gamma_D, \tag{2.10}$$

where

$$L_0(\mathbf{u},p) := \begin{bmatrix} I & -\nabla\\ \nabla\cdot A - \mathbf{b}\cdot & -c\\ -\nabla\times & 0\end{bmatrix}\begin{pmatrix}\mathbf{u}\\ p\end{pmatrix} = \begin{pmatrix}\mathbf{u} - \nabla p\\ \nabla\cdot A\mathbf{u} - \mathbf{b}\cdot\mathbf{u} - cp\\ -\nabla\times\mathbf{u}\end{pmatrix}. \tag{2.11}$$

The least-squares functional to be minimized is

$$\mathcal{F}_0(\mathbf{v},t) = \left\|L_0(\mathbf{v},t) - (\mathbf{0},f,0)^t\right\|^2.$$

Since we want this functional to exist for all $(\mathbf{v},t) \in \mathcal{D}(L_0)$, we need

$$\mathcal{R}(L_0) \subseteq \left(L^2(\Omega)\right)^4. \tag{2.12}$$

We are, thus, lead to choose

$$\mathcal{D}(L_0) = (H_N(\nabla\cdot A;\Omega) \cap H_D(\nabla\times;\Omega)) \times H_D^1(\Omega),$$

where, for a general 2×2 matrix B, we define

$$\begin{aligned} H^1_J(\Omega) &:= \left\{ s \in H^1(\Omega) : s = 0 \text{ on } \Gamma_J \right\}, \\ H_J(\nabla\cdot B; \Omega) &:= \left\{ \mathbf{w} \in (L^2(\Omega))^2 : \nabla\cdot(B\mathbf{w}) \in L^2(\Omega), \mathbf{n}\cdot B\mathbf{w} = 0 \text{ on } \Gamma_J \right\}, \\ H_J(\nabla\times B; \Omega) &:= \left\{ \mathbf{w} \in (L^2(\Omega))^2 : \nabla\times(B\mathbf{w}) \in L^2(\Omega), \mathbf{t}\cdot B\mathbf{n} = 0 \text{ on } \Gamma_J \right\} \end{aligned}$$

for $J \in \{N, D\}$. Moreover, $H_J(\nabla\cdot; \Omega) := H_J(\nabla\cdot I; \Omega)$ and $H_J(\nabla\times; \Omega) := H_J(\nabla\times I; \Omega)$, where I is the 2×2 identity matrix. Clearly, $\mathcal{D}(L_0)$ is a Hilbert space under the norm

$$\|(\mathbf{v}, t)\|^2_{L_0} := \|\mathbf{v}\|^2 + \|\nabla\cdot A\mathbf{v}\|^2 + \|\nabla\times\mathbf{v}\|^2 + \|t\|^2_1.$$

Since BVP (2.1)–(2.3) is well posed by assumption, we know that (2.7)–(2.10) has a unique solution in $\mathcal{D}_0$. Thus, $\mathcal{F}_0$ has a unique minimizer in $\mathcal{D}_0$ with minimum value zero. We minimize functional $\mathcal{F}_0$ in the weak sense; i.e., we look for solutions of the corresponding variational problem:

Find $(\mathbf{u}, p) \in \mathcal{D}_0$ *such that*

$$\left\langle L_0(\mathbf{u}, p) - (\mathbf{0}, f, 0)^t, L_0(\mathbf{v}, t) \right\rangle = 0 \tag{2.13}$$

for all $(\mathbf{v}, t) \in \mathcal{D}_0$.

A convenient choice for this FOSLS formulation is to discretize variational problem (2.13), using H^1-conforming finite element spaces, such as bilinears on quadrilaterals or linears on triangles. As the mesh size of the discretization tends to zero, the use of H^1-conforming finite element spaces yields converging approximations of the solution provided that solution is in H^1. This approach requires

$$(\mathbf{u}, p) \in \left(H^1(\Omega)\right)^3 \tag{2.14}$$

for FOSLL* approximations using H^1-conforming finite elements to converge to the solution of primal problem (2.7)–(2.10). The requirement (2.14) is more restrictive than (2.12) and is, in general, not fulfilled for problems with irregular boundary points or discontinuous coefficient matrix A. In such cases, the FOSLS approximations, $(\mathbf{u}^h, p^h)$, do not converge to the solution of (2.7)–(2.10), as the following example illustrates.

Example 2.1. *Define the following L-shaped domain:*

$$\Omega = \left\{ \mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_\infty < 1 \text{ and } \theta(\mathbf{x}) \in (0, 3\pi/2) \right\}, \tag{2.15}$$

where $\theta = \arcsin(y/x)$. *Let* $A = I$, $c = 1$, *and* $\mathbf{b} = (-y/10, 10x)^t$. *The Neumann boundary consists of three parts,*

$$\begin{aligned} &\Gamma_{N,1} = \{(x, y) \in \partial\Omega : x \in (0, 1), y = 1\}, \quad \Gamma_{N,2} = \{(x, y) \in \partial\Omega : y = -1\}, \\ &\Gamma_{N,3} = \{(x, y) \in \partial\Omega : x = -1, y \in (0, 1)\}. \end{aligned}$$

The three remaining parts of $\partial\Omega$ *form* Γ_D. *This domain contains irregular boundary points at* $(0, 0)$, $(-1, 0)$, *and* $(0, 1)$.

Let (r, θ) *denote standard polar coordinates on* $\mathbb{R}^2$ *and let*

$$p = \delta(r) r^{2/3} \sin(2\theta/3). \tag{2.16}$$

Then, p *is a solution of BVP* (2.1)–(2.3) *when* $f = \sin(2\theta/3) r^{2/3}(\delta''(r) + \frac{7}{3r}\delta'(r)) - \mathbf{b}\cdot\nabla p - cp$. *Here,* $\delta(r) \in C^2(\Omega)$ *is a cut-off function that satisfies* $\delta(r) = 1$ *for* $r < 1/4$

TABLE 2.1
Error norms for Example 2.1 *on a sequence of uniform meshes with mesh sizes* h.

h	1/4	1/8	1/16	1/32	1/64	1/128	1/256
$\|\|p - p^h\|\|_0$	0.1461	0.1521	0.1543	0.1550	0.1551	0.1550	0.1549
$\|\|\nabla p - \mathbf{u}^h\|\|_0$	0.8433	0.8267	0.8199	0.8173	0.8152	0.8137	0.8127

and $\delta(r) = 0$ *for* $r > 3/4$. *Clearly,* $f \in L^2(\Omega)$, *but* $\mathbf{u} = \nabla p$ *is not in* $(H^\alpha(\Omega))^2$ *for any* $\alpha \geq 2/3$. *Table* 2.1 *shows the results of numerical experiments for this problem with error norms* $||p - p^h||_0$ *and* $||\nabla p - \mathbf{u}^h||_0$ *for a sequence of uniform meshes with decreasing mesh sizes* h *and standard bilinear* H^1*-conforming finite element spaces. Standard FOSLS clearly fails for this example. A closer look at* $\mathbf{u}^h$ *shows that the FOSLS approximation is completely unaware of the singularities in the gradient at the reentrant corner of* Ω.

2.3. FOSLL*. The FOSLL* method was developed to overcome this difficulty with standard FOSLS, while continuing to use standard H^1-conforming finite element spaces in the discretization process. Clearly, H^1-conforming spaces cannot be used to approximate the nonsmooth primal solution, $(\mathbf{u}, p)$, so FOSLL* instead attempts to introduce a dual first-order system whose solution is in H^1.

The main idea can be motivated by looking at the simplest discrete analog, that is, a linear system of equations, $Ax = b$. Solving the corresponding least-squares problem of minimizing $\|Ax - b\|^2_{\ell^2}$ leads to the normal equations, $A^tAx = A^tb$, and the weak form, $\langle Ax, Az\rangle = \langle b, Az\rangle$ *for all* z. This is analogous to what FOSLS does at the PDE level. But another way to recast $Ax = b$ as a minimization problem is to recognize that if $Ax = b$ has a solution, then so does $AA^ty = b$. Note that this system for dual variable y is the normal equations for dual problem $A^ty = x$, and that it can be recast as the minimization of $\|A^ty - x\|^2_{\ell^2}$, which has the same minimizer as the functional $\langle A^ty, A^ty\rangle - 2\langle y, b\rangle$. This leads to the weak form, $\langle A^ty, A^tz\rangle = \langle b, z\rangle$ *for all* z. Note that $x = A^ty$ yields the minimal norm solution of the original problem, $Ax = b$. This idea is formally applicable at the PDE level since our primal problem surely has a solution.

While simpler approaches are possible in some cases, a fairly general methodology for applying FOSLL* is to attempt to reformulate the original BVP as a first-order primal problem whose associated operator, L_1, and adjoint, L_1^*, are bijective. This guarantees the existence of a unique solution for the dual normal equations, $L_1L_1^*w = f$.

This bijectivity is achieved for BVP (2.1)–(2.3) by incorporating a scalar slack variable, q, into the system and using the scaled gradient, $\tilde{\mathbf{u}} = A^{1/2}\nabla p$. (Here, we incorporate a slightly different scaling than in (2.11) because it has computational advantages.) This is done in such a way that $(\tilde{\mathbf{u}}, p, 0)$ solves the primal problem, which for BVP (2.1)–(2.3) is given by

$$L_1(\tilde{\mathbf{u}}, p, q) = (\mathbf{0}, f, 0)^t \text{ in } \Omega, \tag{2.17}$$

$$\mathbf{t} \cdot A^{-1/2}\tilde{\mathbf{u}} = 0 \text{ on } \Gamma_D, \tag{2.18}$$

$$\mathbf{n} \cdot A^{1/2}\tilde{\mathbf{u}} = 0 \text{ on } \Gamma_N, \tag{2.19}$$

$$p = 0 \text{ on } \Gamma_D, \tag{2.20}$$

$$q = 0 \text{ on } \Gamma_N, \tag{2.21}$$

where

$$L_1(\tilde{\mathbf{u}}, p, q) := \begin{bmatrix} A^{-1/2} & -\nabla & -\nabla^\perp \\ \nabla\cdot A^{1/2} - \mathbf{b}\cdot A^{-1/2} & -c & 0 \\ -\nabla\times A^{-1/2} & 0 & 0 \end{bmatrix} \begin{pmatrix} \tilde{\mathbf{u}} \\ p \\ q \end{pmatrix} = \begin{pmatrix} A^{-1/2}\tilde{\mathbf{u}} - \nabla p - \nabla^\perp q \\ \nabla\cdot(A^{1/2}\tilde{\mathbf{u}}) - \mathbf{b}\cdot A^{-1/2}\tilde{\mathbf{u}} - cp \\ -\nabla\times(A^{-1/2}\tilde{\mathbf{u}}) \end{pmatrix}. \tag{2.22}$$

The domain of L_1 is

$$\mathcal{D}(L_1) = \Big(H_N(\nabla\cdot A^{1/2};\Omega) \cap H_D(\nabla\times A^{-1/2};\Omega)\Big) \times H^1_D(\Omega) \times H^1_N(\Omega),$$

which is a Hilbert space under the norm

$$\|(\mathbf{v}, t, z)\|^2_{L_1} := \|\mathbf{v}\|^2 + \left\|\nabla\cdot(A^{1/2}\mathbf{v})\right\|^2 + \left\|\nabla\times(A^{-1/2}\mathbf{v})\right\|^2 + \|t\|_1^2 + \|z\|_1^2. \tag{2.23}$$

The FOSLL* approach is to approximate the solution, $(\mathbf{w}, r, s)$, of the corresponding dual problem,

$$L_1^*(\mathbf{w}, r, s) = (\tilde{\mathbf{u}}, p, q)^t = (A^{1/2}\nabla p, p, 0)^t \text{ in } \Omega, \tag{2.24}$$

$$\mathbf{t}\cdot A^{-1/2}\mathbf{w} = 0 \text{ on } \Gamma_D, \tag{2.25}$$

$$\mathbf{n}\cdot A^{1/2}\mathbf{w} = 0 \text{ on } \Gamma_N, \tag{2.26}$$

$$r = 0 \text{ on } \Gamma_D, \tag{2.27}$$

$$s = 0 \text{ on } \Gamma_N, \tag{2.28}$$

where the adjoint operator is defined by

$$L_1^*(\mathbf{w}, r, s) = \begin{bmatrix} A^{-1/2} & -A^{1/2}\nabla - A^{-1/2}\mathbf{b} & -A^{-1/2}\nabla^\perp \\ \nabla\cdot & -c & 0 \\ -\nabla\times & 0 & 0 \end{bmatrix} \begin{pmatrix} \mathbf{w} \\ r \\ s \end{pmatrix}. \tag{2.29}$$

The domain of L_1^* is

$$\mathcal{D}(L_1^*) = (H_N(\nabla\cdot;\Omega) \cap H_D(\nabla\times;\Omega)) \times H^1_D(\Omega) \times H^1_N(\Omega).$$

which is a Hilbert space under the norm

$$\|(\mathbf{v}, t, z)\|^2_{L_1^*} := \|\mathbf{v}\|^2 + \|\nabla\cdot(\mathbf{v})\|^2 + \|\nabla\times(\mathbf{v})\|^2 + \|t\|_1^2 + \|z\|_1^2. \tag{2.30}$$

This formulation is similar to the FOSLL$^*_e$ formulation described in [12]. The difference is that in (2.29), the coefficient matrix, A, only appears outside of the differential operators. Note, also, that this scaling yields $A^{1/2}\nabla r$ orthogonal to $A^{-1/2}\nabla^\perp s$, which produces better performance for the multigrid solvers. A minor modification of the proof of Theorem 4.1 in [12] yields the following result.

THEOREM 2.2. *Operators L_1 and L_1^* are bijective from $\mathcal{D}(L_1)$ and $\mathcal{D}(L_1^*)$, respectively, onto $(L^2(\Omega))^4$. Further, L_1 and L_1^* are coercive and continuous in the norms defined in* (2.23) *and* (2.30), *respectively.*

Proof. The proof requires the assumption that the adjoint problem (2.4)–(2.6) is well posed and follows with minor modifications from the proof of Theorem 4.1 in [12], together with an application of Lemma 2.1 in [12]. □

Solving the dual problem is equivalent to minimizing the dual functional,

$$\mathcal{F}_1^*(\mathbf{v}, t, z) = \left\| L_1^*(\mathbf{v}, t, z) - (\tilde{\mathbf{u}}, p, q)^t \right\|^2, \tag{2.31}$$

on $\mathcal{D}(L_1^*)$. The associated weak form is as follows:

Find $(\mathbf{w}, r, s) \in \mathcal{D}(L_1^*)$ *such that*

$$\langle L_1^*(\mathbf{w}, r, s), L_1^*(\mathbf{v}, t, z)\rangle = \left\langle (\tilde{\mathbf{u}}, p, q)^t, L_1^*(\mathbf{v}, t, z)\right\rangle \tag{2.32}$$

for all $(\mathbf{v}, t, z) \in \mathcal{D}(L_1^*)$.

The unknown solution, $(\tilde{\mathbf{u}}, p, q)$, is eliminated from the right side of (2.32) by rewriting the right side as follows:

$$\left\langle (\tilde{\mathbf{u}}, p, q)^t, L_1^*(\mathbf{v}, t, z)\right\rangle = \left\langle L_1(\tilde{\mathbf{u}}, p, q), (\mathbf{v}, t, z)^t\right\rangle = \left\langle (\mathbf{0}, f, 0)^t, (\mathbf{v}, t, z)^t\right\rangle.$$

After discretizing this variational form and computing an approximation, $(\mathbf{w}^h, r^h, s^h)$, for the dual unknowns, an L^2 approximation, $(\tilde{\mathbf{u}}^h, p^h, q^h)$, for the primal unknowns is computed easily by applying the adjoint: $(\tilde{\mathbf{u}}^h, p^h, q^h)^t = L_1^*(\mathbf{w}^h, r^h, s^h)$.

This formulation of FOSLL* works well with H^1-conforming finite element spaces if the violation of the crucial regularity condition, (2.14), is due only to the discontinuities in A. This can be most easily seen by noting that in (2.29) the coefficients are never differentiated. However, in the presence of irregular boundary points, we may be left with the difficulty that

$$H_N(\nabla\cdot; \Omega) \cap H_D(\nabla\times; \Omega) \not\subset \left(H^1(\Omega)\right)^2. \tag{2.33}$$

For example, (2.33) holds if the boundary of Ω contains re-entrant corners or points in $\overline{\Gamma_D} \cap \overline{\Gamma_N}$ with an inner angle bigger than $\pi/2$ (cf. [17]). If H^1-conforming finite element spaces are used to approximate the solution to (2.32), then the approximation will not, in general, converge to the solution, but rather to the closest element in $(H^1)^4$ to the solution. In general, this error will not have local support. In the next section, we introduce a modification to FOSLL* that overcomes this difficulty.

We close this section by demonstrating numerically how the FOSLL* formulation described above fails in the presence of irregular boundary points.

EXAMPLE 2.3. *We apply the FOSLL* method* (2.32), *using H^1-conforming finite elements, to the BVP from Example* 2.1. *Table* 2.2 *shows that the L^2 norm of the errors of the approximations for p and $\tilde{\mathbf{u}}$ stagnate as h decreases.*

TABLE 2.2
Error norms for the standard FOSLL* approximations for Example 2.3 *on a sequence of uniform meshes with mesh sizes h.*

h	1/4	1/8	1/16	1/32	1/64	1/128	1/256
$\|\|p - p^h\|\|_0$	0.0482	0.0259	0.0231	0.0230	0.0229	0.0228	0.0228
$\|\|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}^h\|\|_0$	0.6767	0.3202	0.1825	0.1220	0.1001	0.0933	0.0914

3. Improved FOSLL*.

We begin here by introducing modifications to the standard FOSLL* formulation that overcome the shortcomings for problems with irregular boundary points. We then describe how the method can be made more efficient for the special cases $\mathbf{b} = \mathbf{0}$ and $c = 0$.

As a starting point of our improvements, we revert to the scaling used in (2.11). While the scaling in (2.22) is preferable in practice, we use this simpler scaling for ease of exposition. All results in this section can be easily generalized to the scaling in (2.22).

Thus, we define the unscaled gradient, $\mathbf{u} = \nabla p$, as a dependent variable. The primal problem has the form $\mathcal{L}_0(\mathbf{u}, p, q) = (\mathbf{0}, f, 0)^t$, where q is a slack variable as introduced in the previous subsection,

$$\mathcal{L}_0 = \begin{bmatrix} I & -\nabla & -\nabla^{\perp} \\ (\nabla \cdot A - \mathbf{b}\cdot) & -c & 0 \\ -\nabla\times & 0 & -d \end{bmatrix}, \tag{3.1}$$

and d is a nonnegative analytic function on Ω.

Following the development for standard FOSLL*, the domain of $\mathcal{L}_0$ is given by

$$\mathcal{D}(\mathcal{L}_0) = (H_N(\nabla\cdot A; \Omega) \cap H_D(\nabla\times; \Omega)) \times H^1_D(\Omega) \times H^1_N(\Omega).$$

Clearly, $(\nabla p, p, 0)$ solves this problem when p is the solution of the BVP (2.1)–(2.3). The corresponding dual problem is

$$\mathcal{L}_0^*(\mathbf{w}, r, s) := \begin{bmatrix} I & -(A\nabla + \mathbf{b}) & -\nabla^{\perp} \\ \nabla\cdot & -c & 0 \\ -\nabla\times & 0 & -d \end{bmatrix} \begin{pmatrix} \mathbf{w} \\ r \\ s \end{pmatrix} = \begin{pmatrix} \mathbf{u} \\ p \\ q \end{pmatrix} \tag{3.2}$$

on the adjoint domain

$$\mathcal{D}(\mathcal{L}_0^*) = (H_N(\nabla\cdot; \Omega) \cap H_D(\nabla\times; \Omega)) \times H^1_D(\Omega) \times H^1_N(\Omega).$$

Formulating the FOSLL* method using $\mathcal{L}_0^*$ reveals exactly the same difficulty as the formulation using L_1^* in the last subsection. While discontinuous coefficients do not cause difficulties, irregular boundary points do, because they imply $H_N(\nabla\cdot; \Omega) \cap H_D (\nabla\times; \Omega) \not\subset \left(H^1(\Omega)\right)^2$, which in turn implies $\mathcal{D}(\mathcal{L}_1^*) \not\subset \left(H^1(\Omega)\right)^4$.

The next step is to introduce a modified operator, $\mathcal{L}_1$, that is formally identical to $\mathcal{L}_0$ in (3.1) but has a different domain. The aim is to expand the domain of $\mathcal{L}_1$ so that the domain of its adjoint shrinks to a subspace of $\left(H^1(\Omega)\right)^4$. To see how this is done, note that the second and third entries in the first row of $\mathcal{L}_0$ in (3.1) can be rewritten as follows:

$$\begin{bmatrix} -\nabla & -\nabla^{\perp} \end{bmatrix} = \begin{bmatrix} -\partial_x & -\partial_y \\ -\partial_y & \partial_x \end{bmatrix} = \begin{bmatrix} -\nabla\cdot \\ \nabla\times \end{bmatrix}.$$

Thus, instead of asking the gradients of p and q to be in $L^2(\Omega)$ individually, we may impose the more general condition that the div and curl of the pair (p, q) be in $L^2(\Omega)$. We are thus lead to rewrite $\mathcal{L}_0$ as

$$\mathcal{L}_1 = \begin{bmatrix} I & \begin{matrix} -\nabla\cdot \\ \nabla\times \end{matrix} \\ \begin{matrix} (\nabla\cdot A - \mathbf{b}\cdot) \\ -\nabla\times \end{matrix} & -B \end{bmatrix}, \qquad \text{where } B = \begin{pmatrix} c & 0 \\ 0 & d \end{pmatrix},$$

so that

$$\mathcal{D}(\mathcal{L}_1) = (H_N(\nabla\cdot A; \Omega) \cap H_D(\nabla\times; \Omega)) \times \mathcal{H}_{DN}(\Omega),$$

where

$$\mathcal{H}_{DN}(\Omega) := \{(v_1, v_2) \in H(\nabla\cdot; \Omega) \cap H(\nabla\times; \Omega) : v_1 = 0 \text{ on } \Gamma_D, v_2 = 0 \text{ on } \Gamma_N\}.$$

Integration by parts then shows that the domain of $\mathcal{L}_1^*$ is in $\left(H^1(\Omega)\right)^4$:

$$\mathcal{D}(\mathcal{L}_1^*) = (H_N(\nabla\cdot;\Omega) \cap H_D(\nabla\times;\Omega) \cap \left(H^1(\Omega)\right)^2) \times H_D^1(\Omega) \times H_N^1(\Omega).$$

Unfortunately, this approach is not yet viable because the adjoint, $\mathcal{L}_1^*$, is in general no longer surjective and we can no longer guarantee that $(\mathbf{u}, p, q) \in \mathcal{R}(\mathcal{L}_1^*)$, as the following example shows.

EXAMPLE 3.1. *Let Ω be the L-shaped domain from* (2.15) *and set $A = I$ and $\mathbf{b} = \mathbf{0}$. Let d be any positive analytic function and $c \in L^\infty$ with $0 < c < 1$ a.e. We choose homogeneous Dirichlet boundary conditions: $\Gamma_D = \partial\Omega$. Let Γ_H be the union of the three horizontal edges and Γ_V be the union of the three vertical edges of Ω. Thus, imposing $v_1 = 0$ on Γ_D is equivalent to imposing $\mathbf{n} \cdot (v_1, v_2) = 0$ on Γ_H and $\mathbf{t} \cdot (v_1, v_2) = 0$ on Γ_V, so*

$$\mathcal{H}_{DN}(\Omega) = H_H(\nabla\cdot;\Omega) \cap H_V(\nabla\times;\Omega)$$

holds for this example. The analysis of the div-curl *operator in* [11] *shows that $\left[\begin{smallmatrix}-\nabla\cdot\\ \nabla\times\end{smallmatrix}\right]$ has a nontrivial null space on $\mathcal{H}_{DN}(\Omega)$. (For example, let $\mathbf{z} = \nabla\phi$, where $\Delta\phi = 0$, $\mathbf{n} \cdot \nabla\phi = 0$ on Γ_H, $\phi = 0$ on $\Gamma_{V_1} \cup \Gamma_{V_2}$, and $\phi = 1$ on Γ_{V_3}.) Let $\mathbf{z} \neq \mathbf{0}$ be such a null space element. Note that $\mathcal{L}_1(\mathbf{0}, \mathbf{z}) = (\mathbf{0}, -cz_1, -dz_2)^t$ is in $\left(L^2(\Omega)\right)^4$. Since $\mathcal{L}_1 = \mathcal{L}_1^*$ formally holds, Lemma* 3.6 *in* [11] *implies the existence of a more regular preimage, $(\mathbf{v}, \mathbf{w}) \in \mathcal{D}(\mathcal{L}_1) \cap \left(H^1(\Omega)\right)^4$ with $\mathcal{L}_1(\mathbf{v}, \mathbf{w}) = (\mathbf{0}, -cz_1, -dz_2)^t$. Therefore, $(\mathbf{v}, \mathbf{w} - \mathbf{z})$ is a nontrivial element of null space $\mathcal{N}(\mathcal{L}_1)$. Since $(\mathcal{L}_1^*)^* = \mathcal{L}_1$, the closed range theorem implies that $\mathcal{R}(\mathcal{L}_1^*) = \mathcal{N}(\mathcal{L}_1)^\perp$, so $\mathcal{R}(\mathcal{L}_1^*)$ is not all of $\left(L^2(\Omega)\right)^4$ and $\mathcal{L}_1^*$ is not surjective.*

To prove that, in general, $U = (\nabla p, p, 0) \notin \mathcal{R}(\mathcal{L}_1^*)$ so that the dual problem is not solvable, assume otherwise: $U \in \mathcal{R}(\mathcal{L}_1^*)$ or, equivalently, $U \perp \mathcal{N}(\mathcal{L}_1)$ for all admissible p. Let $(\mathbf{v}, \mathbf{w})$ be an element of $\mathcal{N}(\mathcal{L}_1)$, i.e.,

$$\mathbf{v} + \begin{bmatrix}-\nabla\cdot\\ \nabla\times\end{bmatrix}\mathbf{w} = \mathbf{0} \tag{3.3}$$

$$\begin{bmatrix}\nabla\cdot\\ -\nabla\times\end{bmatrix}\mathbf{v} - \begin{pmatrix}cw_1\\ dw_2\end{pmatrix} = \mathbf{0}. \tag{3.4}$$

Now, $U \perp (\mathbf{v}, \mathbf{w})$ means $\langle\nabla p, \mathbf{v}\rangle + \langle p, w_1\rangle = 0$. Using the divergence theorem and (3.4), *we thus have $(c-1)\,\langle p, w_1\rangle = 0$. Since $\langle p, w_1\rangle$ must vanish for all admissible p, we must have $w_1 = 0$. From* (3.3) *and* (3.4), *we conclude that $w_2 \in H^1(\Omega)$ and*

$$\Delta w_2 - dw_2 = 0.$$

The definition of $\mathcal{D}(\mathcal{L}_1)$ and (3.3) *supply the boundary condition*

$$\mathbf{n} \cdot \nabla w_2 = -\mathbf{t} \cdot \nabla^\perp w_2 = -\mathbf{t} \cdot \mathbf{v} = 0 \text{ on } \Gamma_D = \partial\Omega.$$

Therefore, w_2 is a constant on Ω and (3.3)–(3.4) *yield $\mathbf{v} = \mathbf{w} = \mathbf{0}$. Since $\mathcal{N}(\mathcal{L}_1)$ is nontrivial, our assumption is wrong, and the dual problem is not, in general, solvable in $\mathcal{D}(\mathcal{L}_1^*)$.*

Our numerical experience supports the difficulty expressed in this example: it seems that $(\nabla p, p, 0)$ is in $\mathcal{R}(\mathcal{L}_1^*)$ only for very special choice of $A, \mathbf{b}, c, d$, and Ω, whenever $\partial\Omega$ contains irregular boundary points.

Nevertheless, the modifications that lead from $\mathcal{L}_0$ to $\mathcal{L}_1$ take a step in the right direction because we now have $\mathcal{D}(\mathcal{L}_1^*) \subset \left(H^1(\Omega)\right)^4$. This H^1-inclusion property guarantees that a dual solution, when it exists, can be easily approximated by standard H^1 finite element spaces. As the final step, we now modify the domain of the operators again to ensure solvability. The aim is to increase the domain of the new dual operator, $\mathcal{L}^*$, in order to make it surjective. This is done indirectly by reducing the number of boundary conditions on the dual domain. We do this by enforcing more boundary conditions on the domain of the primal operator, $\mathcal{L}$. Of course, we are only allowed to enforce additional boundary conditions on the primal problem that are fulfilled by the primal solution, $(\mathbf{u}, p, q)$. The key is to identify these allowable conditions and choose those that induce the appropriate $\mathcal{D}(\mathcal{L}^*)$.

First, we introduce the new modified operator, $\mathcal{L}$, then prove some useful lemmas, and finally we present our main results, the surjectivity of the dual operator, $\mathcal{L}^*$.

The two additional boundary conditions we enforce on the primal problem are

$$\int_{\Gamma_{N,i}} \mathbf{t} \cdot \mathbf{u}\, ds = 0, \quad i = 1, \dots, M, \tag{3.5}$$

$$q = 0 \text{ on } \Gamma_Q \subset \Gamma_D. \tag{3.6}$$

These additional conditions are allowable because the primal solution, $(\mathbf{u}, p, q)$, satisfies them:

$$\int_{\Gamma_{N,i}} \mathbf{t} \cdot \mathbf{u}\, ds = \int_{\Gamma_{N,i}} \mathbf{t} \cdot \nabla p\, ds = \int_{\Gamma_{N,i}} \frac{dp}{ds}\, ds = 0, \qquad i = 1, \dots, M,$$

and $q = 0$ on $\partial\Omega$. For theoretical purposes, we impose condition (3.6) only on a subset, $\Gamma_Q \subset \Gamma_D$, that does not contain any irregular boundary points in its closure but has positive measure. See Remark 3.2 for motivation.

The new operator, $\mathcal{L}$, has the same form as $\mathcal{L}_1$, but differs again by its domain. We define the form of $\mathcal{L}$ blockwise:

$$T = \begin{bmatrix} \nabla\cdot A - \mathbf{b}\cdot \\ -\nabla\times \end{bmatrix}, \qquad B = \begin{bmatrix} c & 0 \\ 0 & d \end{bmatrix}, \tag{3.7}$$

$$S = \begin{bmatrix} -\partial_x & -\partial_y \\ -\partial_y & \partial_x \end{bmatrix} = \begin{bmatrix} -\nabla\cdot \\ \nabla\times \end{bmatrix}, \tag{3.8}$$

$$\mathcal{L} = \begin{bmatrix} I & S \\ T & -B \end{bmatrix}. \tag{3.9}$$

The corresponding domains include the following additional boundary conditions:

$$\mathcal{D}(T) = \left\{ \mathbf{v} \in H_N(\nabla\cdot A; \Omega) \cap H_D(\nabla\times; \Omega) \ : \ \int_{\Gamma_{N,i}} \mathbf{t} \cdot \mathbf{v}\, ds = 0,\ 1 \le i \le M \right\}, \tag{3.10}$$

$$\mathcal{D}(S) = \left\{ (t, z) \in \mathcal{H}_{DN}(\Omega) \ : \ z = 0 \text{ on } \Gamma_Q \right\}, \tag{3.11}$$

$$\mathcal{D}(\mathcal{L}) = \mathcal{D}(T) \times \mathcal{D}(S). \tag{3.12}$$

These domains are Hilbert spaces under the div-curl norms,

$$\|\mathbf{v}\|_S^2 := \|\mathbf{v}\|^2 + \|\nabla\cdot\mathbf{v}\|^2 + \|\nabla\times\mathbf{v}\|^2, \tag{3.13}$$

$$\|\mathbf{v}\|_T^2 := \|\mathbf{v}\|^2 + \|\nabla\cdot(A\mathbf{v})\|^2 + \|\nabla\times\mathbf{v}\|^2, \tag{3.14}$$

$$\|(\mathbf{v}, \mathbf{w})\|_{\mathcal{L}}^2 := \|\mathbf{v}\|_T^2 + \|\mathbf{w}\|_S^2. \tag{3.15}$$

Integration by parts leads to the adjoint operators,

$$T^* = \begin{bmatrix} -(A\nabla + \mathbf{b}) & -\nabla^\perp \end{bmatrix}, \tag{3.16}$$

$$S^* = \begin{bmatrix} \partial_x & \partial_y \\ \partial_y & -\partial_x \end{bmatrix} = \begin{bmatrix} \nabla & \nabla^\perp \end{bmatrix}, \tag{3.17}$$

$$\mathcal{L}^* = \begin{bmatrix} I & T^* \\ S^* & -B \end{bmatrix}, \tag{3.18}$$

and the domains,

$$\mathcal{D}(S^*) = \left\{ \mathbf{v} \in \left(H^1(\Omega)\right)^2 \ : \ \mathbf{n}\cdot\mathbf{v} = 0 \text{ on } \Gamma_N \text{ and } \mathbf{t}\cdot\mathbf{v} = 0 \text{ on } \Gamma_D \backslash \Gamma_Q \right\}, \tag{3.19}$$

$$\mathcal{D}(T^*) = \left\{ (t, z) \in \left(H^1(\Omega)\right)^2 \ : \ t = 0 \text{ on } \Gamma_D \text{ and } z \equiv c_i \text{ on } \Gamma_{N,i} \right\}, \tag{3.20}$$

where $1 \leq i \leq M$, and c_i are arbitrary constants, and

$$\mathcal{D}(\mathcal{L}^*) = \mathcal{D}(S^*) \times \mathcal{D}(T^*). \tag{3.21}$$

Remark 3.2. We do not allow the closure of $\Gamma_Q \subset \Gamma_D$ to contain irregular boundary points because we would expect singular functions in $H(\nabla\times;\Omega) \cap H(\nabla\cdot;\Omega)$ to arise at these points. These singular functions would no longer be in $\mathcal{D}(S)$, but they would be in $\mathcal{D}(S^*)$. In practice, there seems to be no difficulty with allowing Γ_Q to touch irregular boundary points.

For the remainder of this section, we adopt the assumptions on BVP (2.1)–(2.3) made in subsection 2.1 and let d be any nonnegative analytic function. We now prove coercivity of the primal operator, $\mathcal{L}$. To this end, we need two auxiliary results.

LEMMA 3.3. *$\mathcal{L}$ is injective.*

Proof. Assume that there exists a $(\mathbf{v}, \mathbf{w}) \in \mathcal{D}(\mathcal{L})$ such that

$$\mathcal{L}(\mathbf{v}, \mathbf{w}) = \begin{bmatrix} I & S \\ T & -B \end{bmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{w} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}.$$

Then, $S\mathbf{w} \in \mathcal{D}(T)$, which, together with $\mathbf{w} \in \mathcal{D}(S)$, implies that

$$\mathbf{t} \cdot S\mathbf{w} = -\mathbf{n} \cdot \nabla w_2 = 0 \text{ on } \Gamma_Q, \tag{3.22}$$

$$\mathbf{n} \cdot AS\mathbf{w} = -\mathbf{n} \cdot A(\nabla w_1 + \nabla^\perp w_2) = 0 \text{ on } \Gamma_N. \tag{3.23}$$

Now, choose any open set, $\mathcal{O} \subset \Omega$ such that $\overline{\mathcal{O}} \cap \Gamma_Q$ has positive measure and $\partial\mathcal{O}$ contains no irregular points of $\partial\Omega$. On $\mathcal{O}$ we have $\mathbf{w} \in (H^1)^2$. If we only look at the set $\mathcal{O}$, eliminating $\mathbf{v}$ yields the following equation for $\mathbf{w}$:

$$\begin{aligned} -TS\mathbf{w} - B\mathbf{w} &= -\begin{bmatrix} \nabla\cdot A - \mathbf{b}\cdot \\ -\nabla\times \end{bmatrix} \begin{bmatrix} -\nabla & -\nabla^\perp \end{bmatrix} \mathbf{w} - \begin{bmatrix} c & 0 \\ 0 & d \end{bmatrix} \mathbf{w} \\ &= \begin{bmatrix} (\nabla\cdot A\nabla - \mathbf{b}\cdot\nabla - c) & (\nabla\cdot A\nabla^\perp - \mathbf{b}\cdot\nabla^\perp) \\ 0 & \Delta - d \end{bmatrix} \mathbf{w} = \mathbf{0}. \end{aligned} \tag{3.24}$$

Consider the second equation together with the boundary conditions to get

$$\begin{aligned} \Delta w_2 - d w_2 &= 0 \text{ in } \mathcal{O}, \\ w_2 &= 0 \text{ on } \partial\mathcal{O} \cap \Gamma_Q, \\ \mathbf{n} \cdot \nabla w_2 &= 0 \text{ on } \partial\mathcal{O} \cap \Gamma_Q. \end{aligned}$$

According to the unique continuation theorem (cf. Hörmander [18]), we must have $w_2 = 0$ in $\mathcal{O}$. Since every point of Ω is in some $\mathcal{O}$ of this type, we conclude that $w_2 = 0$ in Ω. Since $S\mathbf{w} \in (L^2)^2$, then $w_2 = 0$ implies that $w_1 \in H^1(\Omega)$. Equations (3.23) and (3.24) now yield

$$\begin{aligned}(\nabla\cdot A\nabla - \mathbf{b}\cdot\nabla - c)w_1 &= 0 \text{ in } \Omega,\\ w_1 &= 0 \text{ on } \Gamma_D,\\ \mathbf{n}\cdot A\nabla w_1 &= 0 \text{ on } \Gamma_N.\end{aligned}$$

The well-posedness of BVP (2.1)–(2.3) implies that $w_1 = 0$ and, therefore, $\mathbf{v} = \mathbf{0}$. Hence, $\mathcal{N}(\mathcal{L}) = \{\mathbf{0}\}$ and the lemma follows. □

Lemma 3.4. *S is injective.*

Proof. Assume that there is a $\mathbf{w} \in \mathcal{D}(S)$ such that

$$S\mathbf{w} = \begin{bmatrix} -\nabla\cdot \\ \nabla\times \end{bmatrix} \mathbf{w} = \mathbf{0}.$$

Since Ω is simply connected, the curl-free condition here implies that $\mathbf{w} = \nabla\phi$ for some $\phi \in H^1(\Omega)$, with ϕ determined uniquely up to a constant (cf. [16]). The div-free condition implies that ϕ is harmonic. The boundary conditions on $\mathcal{D}(S)$ imply that $\nabla\phi = \mathbf{0}$ on Γ_Q, so

$$\mathbf{n}\cdot\nabla\phi = 0 \text{ on } \Gamma_Q.$$

But $\mathbf{t}\cdot\nabla\phi = 0$ is also true on Γ_Q. Thus, ϕ is constant on Γ_Q and, without loss of generality, we may assume

$$\phi = 0 \text{ on } \Gamma_Q.$$

Applying the unique continuation theorem (cf. Hörmander [18]) yields $\phi = 0$, which completes the proof. □

We are now able to establish coercivity of $\mathcal{L}$.

Theorem 3.5. *Operators S, T, and $\mathcal{L}$ are coercive in the norms* (3.13), (3.14), *and* (3.15)*, respectively.*

Proof. We begin by proving coercivity of S and T. For S, it suffices to prove a Poincaré inequality of the following form:

There exists constant $C > 0$ such that

$$\|\mathbf{w}\|^2 \le C\left(\|\nabla\cdot\mathbf{w}\|^2 + \|\nabla\times\mathbf{w}\|^2\right) \tag{3.25}$$

for all $\mathbf{w} \in \mathcal{D}(S)$.

To establish (3.25), we assume that no such inequality exists, that is, that there exists $\{\mathbf{w}^{(i)}\}_{i=1,\infty} \in \mathcal{D}(S)$ such that, for all $i > 0$,

$$\|\nabla\cdot\mathbf{w}^{(i)}\|^2 + \|\nabla\times\mathbf{w}^{(i)}\|^2 = 1, \tag{3.26}$$

$$\|\mathbf{w}^{(i)}\|^2 \ge i. \tag{3.27}$$

Now, every $\mathbf{w}^{(i)} \in \mathcal{D}(S)$ can be written as

$$\mathbf{w}^{(i)} = \mathbf{z}^{(i)} + \sum_{j=1}^{K} \beta_{ij}\boldsymbol{\phi}^{(j)},$$

where $\mathbf{z}^{(i)} \in \mathcal{D}(S) \cap \left(H^1(\Omega)\right)^2$ and $\{\boldsymbol{\phi}^{(j)}\}_{j=1,K}$ is a basis of the finite-dimensional orthogonal complement of $\mathcal{D}(S) \cap \left(H^1(\Omega)\right)^2$ in $\mathcal{D}(S)$. Here, we take orthogonality in the $\mathcal{H}_{DN}(\Omega)$ sense, which is an inner product because S is injective. That is, we require

$$\left\langle \nabla\cdot\boldsymbol{\phi}^{(j)}, \nabla\cdot\mathbf{z} \right\rangle + \left\langle \nabla\times\boldsymbol{\phi}^{(j)}, \nabla\times\mathbf{z} \right\rangle = 0$$

for every $\mathbf{z} \in \mathcal{D}(S) \cap \left(H^1(\Omega)\right)^2$. Then, (3.26) becomes

$$\|\nabla\cdot\mathbf{z}^{(i)}\|^2 + \|\nabla\times\mathbf{z}^{(i)}\|^2 + \left\|\nabla\cdot\sum_{j=1}^{K}\beta_{ij}\boldsymbol{\phi}^{(j)}\right\|^2 + \left\|\nabla\times\sum_{j=1}^{K}\beta_{ij}\boldsymbol{\phi}^{(j)}\right\|^2 = 1. \tag{3.28}$$

Since $\mathbf{z}^{(i)} \in \left(H^1(\Omega)\right)^2$, we know that there exist constants $C_0, C_1 > 0$ such that, for all $i > 0$,

$$\|\mathbf{z}^{(i)}\|^2 \le C_0 \left(\|\nabla\mathbf{z}^{(i)}\|^2\right) \le C_1 \left(\|\nabla\cdot\mathbf{z}^{(i)}\|^2 + \|\nabla\times\mathbf{z}^{(i)}\|^2\right) \le C_1, \tag{3.29}$$

where the second inequality can be found in [17] and the last inequality follows from (3.28). In several places in this proof, we make use of the general inequality

$$\|\alpha + \beta\|^2 \le 2(\|\alpha\|^2 + \|\beta\|^2). \tag{3.30}$$

Now, to satisfy (3.27), we combine it with (3.28) and (3.29), using inequality (3.30), to see that we must have

$$\left\|\sum_{j=1}^{K}\beta_{ij}\boldsymbol{\phi}^{(j)}\right\|^2 \ge \frac{i}{2} - C_1 \tag{3.31}$$

for all $i > 0$. We now define $P, N \in \mathbb{R}^{(K\times K)}$ as

$$\begin{aligned} P &:= (p_{kl}) = \left\langle \boldsymbol{\phi}^{(k)},\, \boldsymbol{\phi}^{(l)} \right\rangle, \\ N &:= (n_{kl}) = \left\langle \nabla\cdot\boldsymbol{\phi}^{(k)}, \nabla\cdot\boldsymbol{\phi}^{(l)} \right\rangle + \left\langle \nabla\times\boldsymbol{\phi}^{(k)}, \nabla\times\boldsymbol{\phi}^{(l)} \right\rangle. \end{aligned}$$

Because the $\{\boldsymbol{\phi}^{(j)}\}_{j=1,K}$ are linearly independent and S has no null space (see Lemma 3.4), P and N must be symmetric positive definite matrices. Now, define the vectors

$$\mathbf{b}^{(i)} := (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iK})^t.$$

Equations (3.28) and (3.31) imply

$$\frac{\mathbf{b}^{(i)} \cdot P\mathbf{b}^{(i)}}{\mathbf{b}^{(i)} \cdot N\mathbf{b}^{(i)}} \ge \frac{i}{2} - C_1,$$

which contradicts positive definiteness of N. Therefore, (3.25) holds and S is coercive in the norm defined by (3.13).

To prove coercivity of T, note that, by inequality (3.30), there exists a constant $C_2 > 0$ (dependent only on $\|\mathbf{b}\|$) such that

$$\begin{aligned} \|\mathbf{v}\|_T^2 &\le \|\mathbf{v}\|^2 + 2\|\nabla\cdot(A\mathbf{v}) - \mathbf{b}\cdot\mathbf{v}\|^2 + \|\nabla\times\mathbf{v}\|^2 + 2\|\mathbf{b}\cdot\mathbf{v}\|^2 \\ &\le C_2 \left(\|T\mathbf{v}\|^2 + \|\mathbf{v}\|^2\right) \end{aligned}$$

for all $\mathbf{v} \in \mathcal{D}(T)$. Since T is injective (cf. [11]) and $\mathcal{D}(T)$ is compactly embedded in $\left(L^2(\Omega)\right)^2$, a standard compactness argument establishes coercivity of T.

By coercivity of T and S and the inequality (3.30), there exist constants $C_3, C_4 > 0$ (depending only on $\|\mathbf{b}\|$, $\|c\|$, and $\|d\|$) such that for all $(\mathbf{v}, \mathbf{w}) \in \mathcal{D}(\mathcal{L})$,

$$\begin{aligned} \|(\mathbf{v},\mathbf{w})\|_{\mathcal{L}}^2 = \|\mathbf{v}\|_T^2 + \|\mathbf{w}\|_S^2 &\leq C_3\left(\|T\mathbf{v}\|^2 + \|S\mathbf{w}\|^2\right) \\ &\leq C_4\left(\|T\mathbf{v} - B\mathbf{w}\|^2 + \|\mathbf{v} + S\mathbf{w}\|^2 + \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2\right). \end{aligned}$$

Since $\mathcal{L}$ is injective (see Lemma 3.3) and $\mathcal{D}(\mathcal{L})$ is compactly embedded in $\left(L^2(\Omega)\right)^4$, again we appeal to a standard compactness argument and establish the coercivity of $\mathcal{L}$. □

The main result of this section is the following theorem, which confirms the existence of a dual solution. It is a simple consequence of Theorem 3.5.

THEOREM 3.6. *$\mathcal{L}^* : \mathcal{D}(\mathcal{L}^*) \to \left(L^2(\Omega)\right)^4$ is surjective.*

Proof. It is clear that $(\mathcal{L}^*)^* = \mathcal{L}$. Thus, both $\mathcal{L}$ and $\mathcal{L}^*$ are closed and we may use the closed range theorem. Since $\mathcal{L}$ is coercive (see Theorem 3.5), then $\mathcal{R}(\mathcal{L})$ is closed. The closed range theorem implies that $\mathcal{R}(\mathcal{L}^*)$ is closed. Thus, we know that $\mathcal{R}(\mathcal{L}^*) = \mathcal{N}(\mathcal{L})^\perp$. Since $\mathcal{N}(\mathcal{L})$ is empty, then $\mathcal{L}^*$ is surjective. □

3.1. The case $c = 0$. This is the case that was examined in [12]. In this paper we remove the requirement that Γ_D and Γ_N have at most one component. When $c = 0$, it is useful to choose $d = 0$ because the second and third rows of $\mathcal{L}_0$ in (3.1) only involve $\mathbf{u}$. This allows us to write the primal problem, $\mathcal{L}U = F$, in two stages:

$$T\mathbf{u} = \begin{pmatrix} f \\ 0 \end{pmatrix}, \tag{3.32}$$

$$\nabla p = \mathbf{u}. \tag{3.33}$$

Since T is injective by itself (see the proof of Theorem 3.5), problem (3.32) alone is sufficient to determine $\mathbf{u}$. We can, thus, begin by solving this so-called first-stage equation. The second stage, (3.33), can be then solved for p if needed.

Discontinuous coefficients in A and irregular boundary points on $\partial\Omega$ imply $\mathbf{u} \notin \left(H^1(\Omega)\right)^2$, so we use a FOSLL* formulation to solve (3.32). To accommodate the possibility of multiply connected boundary components, (3.32) is posed on domain $\mathcal{D}(T)$ defined in (3.10). Note, then, that the dual problem for the first stage reads $T^*\mathbf{w} = \mathbf{u}$ and takes the variational form

$$\langle T^*\mathbf{w}, T^*\mathbf{v}\rangle = \left\langle (f,0)^t, \mathbf{v}\right\rangle \text{ for all } \mathbf{v} \in \mathcal{D}(T^*). \tag{3.34}$$

THEOREM 3.7. *Operator $T^* : \mathcal{D}(T^*) \to \left(L^2(\Omega)\right)^2$ is surjective.*

Proof. Coercivity of T was proved for Theorem 3.5 and, by arguments similar to those in Theorem 3.6, we can then prove surjectivity of T^*. □

This theorem establishes existence of a solution for the first stage (3.32). Note that recovery of the solution, p, of the original BVP, (2.1)–(2.3), can then be done by applying a standard FOSLS scheme for the second stage equation, (3.33), i.e., by minimizing $\|\nabla p - \mathbf{u}\|$, where $\mathbf{u} = T^*\mathbf{w}$ is the approximation obtained in the first stage. This minimization is done in an appropriate subspace of $H_D^1(\Omega)$ and leads to an H^1 approximation of p, which is clearly more desirable than the L^2 approximations for p obtained by the general FOSLL* approach for $c \neq 0$.

Remark 3.8. A closer look at the dual problem for the first stage, $T^*\mathbf{w} = \mathbf{u}$, shows that the second component of the dual variable, w_2, is only determined up to a constant. Therefore, without loss of generality, we can restrict the space in which we are looking for $\mathbf{w}$ to

$$\begin{aligned}\bigl\{\mathbf{w} \in \bigl(H^1(\Omega)\bigr)^2 \ :\ & w_1 = 0 \text{ on } \Gamma_D,\\ & w_2 = 0 \text{ on } \Gamma_{N,1}, w_2 \equiv c_i \text{ on } \Gamma_{N,i},\ 2 \le i \le M\bigr\}.\end{aligned}$$

Thus, for the case $c = 0$, standard FOSLL*, as proposed in [12], works well enough, unless Γ_D or Γ_N is not simply connected.

Remark 3.9. A scaled version of the first stage (3.32) that solves for the scaled flux, $\tilde{\mathbf{u}} = A^{1/2}\nabla p$ (see (3.36)), yields a dual problem with better computational performance when used in conjunction with multigrid solvers.

3.2. The case $\mathbf{b} = \mathbf{0}$. For $\mathbf{b} = \mathbf{0}$, we consider two cases: $c > 0$ and $c = 0$. (We exclude the case that c is neither 0 nor strictly positive.) We show in both cases that a scaled form of FOSLL* reduces to the standard Galerkin formulation of (2.1)–(2.3).

Consider, first the case $c > 0$. We can rescale the primal problem by using the scaled primal unknowns, $(\tilde{\mathbf{u}}, \tilde{p}, \tilde{q}) = (A^{1/2}\nabla p, c^{1/2}p, c^{1/2}q)$. The primal operator is then simply a scaled version of L_1 used in standard FOSLL*. The primal problem takes the form

$$\begin{aligned}\tilde{L}_1(\tilde{\mathbf{u}}, \tilde{p}, \tilde{q}) &= (\mathbf{0}, f, 0)^t \text{ in } \Omega,\\ \mathbf{t}\cdot A^{-1/2}\tilde{\mathbf{u}} &= 0 \text{ on } \Gamma_D,\\ \mathbf{n}\cdot A^{1/2}\tilde{\mathbf{u}} &= 0 \text{ on } \Gamma_N,\\ \tilde{p} &= 0 \text{ on } \Gamma_D,\\ \tilde{q} &= 0 \text{ on } \Gamma_N,\end{aligned}$$

where

$$\begin{aligned}\tilde{L}_1(\tilde{\mathbf{u}}, \tilde{p}, \tilde{q}) &:= \begin{bmatrix} A^{-1/2} & -\nabla c^{-1/2} & -\nabla^{\perp} c^{-1/2} \\ \nabla\cdot A^{1/2} & -c^{1/2} & 0 \\ -\nabla\times A^{-1/2} & 0 & 0 \end{bmatrix} \begin{pmatrix} \tilde{\mathbf{u}} \\ \tilde{p} \\ \tilde{q} \end{pmatrix}\\ &= \begin{pmatrix} A^{-1/2}\tilde{\mathbf{u}} - \nabla c^{-1/2}\tilde{p} - \nabla^{\perp} c^{-1/2}\tilde{q} \\ \nabla\cdot(A^{1/2}\tilde{\mathbf{u}}) - c^{1/2}\tilde{p} \\ -\nabla\times(A^{-1/2}\tilde{\mathbf{u}}) \end{pmatrix}.\end{aligned}$$

The domain of this operator is

$$\mathcal{D}(\tilde{L}_1) = \Bigl(H_N(\nabla\cdot A^{1/2};\Omega) \cap H_D(\nabla\times A^{-1/2};\Omega)\Bigr) \times H^1_D(c^{-1/2};\Omega) \times H^1_N(c^{-1/2};\Omega),$$

where $\phi \in H^1_J(c^{-1/2};\Omega)$ if and only if $c^{-1/2}\phi \in H^1_J(\Omega)$. Thus, the dual problem takes the form

$$\tilde{L}_1^*(\mathbf{w}, r, s) = (\tilde{\mathbf{u}}, \tilde{p}, \tilde{q})^t = (A^{1/2}\nabla p, c^{1/2}p, 0)^t \text{ in } \Omega \tag{3.35}$$

on

$$\mathcal{D}(\tilde{L}_1^*) = (H_N(\nabla\cdot;\Omega) \cap H_D(\nabla\times;\Omega)) \times H^1_D(\Omega) \times H^1_N(\Omega),$$

where the adjoint operator has the form

$$\tilde{L}_1^*(\mathbf{w}, r, s) = \begin{bmatrix} A^{-1/2} & -A^{1/2}\nabla & -A^{-1/2}\nabla^{\perp} \\ c^{-1/2}\nabla\cdot & -c^{1/2} & 0 \\ -c^{-1/2}\nabla\times & 0 & 0 \end{bmatrix} \begin{pmatrix} \mathbf{w} \\ r \\ s \end{pmatrix}.$$

Remarkably, this specially scaled problem has a dual solution in $\left(H^1(\Omega)\right)^4$, namely,

$$(\mathbf{w}, r, s) = (\mathbf{0}, -p, 0).$$

Knowing that $\mathbf{w}$ and s vanish in this special case, dual problem (3.35) takes the following variational form:

Find $p \in H^1_D(\Omega)$ such that

$$\langle A\nabla p, \nabla t\rangle + \langle cp, t\rangle = -\langle f, t\rangle$$

for all $t \in H^1_D(\Omega)$.

This is precisely the variational form of the Galerkin approach for BVP (2.1)–(2.3) with $\mathbf{b} = \mathbf{0}$. In other words, FOSLL* yields the same H^1 approximation, p^h, as the Galerkin approach.

Next, consider the case $\mathbf{b} = \mathbf{0}$ and $c = 0$. A scaled two-stage approach leads to the following first stage primal problem:

$$\begin{bmatrix} \nabla\cdot A^{1/2} \\ -\nabla\times A^{-1/2} \end{bmatrix} \tilde{\mathbf{u}} = \begin{pmatrix} f \\ 0 \end{pmatrix} \tag{3.36}$$

on $H_N(\nabla\cdot A^{1/2}; \Omega) \cap H_D(\nabla\times A^{-1/2}; \Omega)$. The corresponding dual problem is

$$-A^{1/2}\nabla w_1 - A^{-1/2}\nabla^{\perp} w_2 = \tilde{\mathbf{u}} = A^{1/2}\nabla p$$

on $H^1_D(\Omega) \times H^1_N(\Omega)$, which obviously has the solution $\mathbf{w} = (-p, 0)$. Knowing that w_2 vanishes leads to the following variational form:

Find $p \in H^1_D(\Omega)$ such that

$$\langle A\nabla p, \nabla t\rangle = -\langle f, t\rangle$$

for all $t \in H^1_D(\Omega)$.

This, again, is precisely the variational form of the Galerkin approach for BVP (2.1)–(2.3) when $\mathbf{b} = \mathbf{0}$ and $c = 0$. Thus, FOSLL* and Galerkin again yield the same H^1 approximation, p^h.

4. Numerical results. Here we report on various numerical results and discuss some implementation issues for the methods proposed in the previous section. All problems in this section were computed with FOSPACK [22]. The linear solver used for the discretized equations was a conjugate gradient iteration (PCG), preconditioned by algebraic multigrid (AMG) using one standard W(1,1)-cycle based on point Gauss–Seidel relaxation. In all cases, the PCG/AMG iterations were applied until the residual norm of the linear system was reduced by a factor of at least 10^{-10}. While this criterion is unnecessarily strong, and is not recommended in practice, it was used to eliminate algebraic error from the analysis of the convergence of the finite element approximations.

First, we show how the improved FOSLL* method performs on the problem proposed in Examples 2.1 and 2.3. One of our improvements was the introduction of

Table 4.1

Error norms, approximate order of discretization convergence, β, and asymptotic AMG convergence factors, ρ, for the improved FOSLL approximations for Example* 4.1 *on a sequence of uniform meshes with mesh sizes h.*

h	1/4	1/8	1/16	1/32	1/64	1/128	1/256
$\|p - p^h\|_0$	0.0475	0.0189	0.0113	0.0075	0.0047	0.0027	0.0016
β		1.328	0.738	0.606	0.674	0.767	0.809
$\|\mathbf{u} - \mathbf{u}^h\|_0$	0.6674	0.3051	0.1573	0.0810	0.0421	0.0222	0.0120
β		1.129	0.956	0.958	0.946	0.923	0.892
ρ	0.20	0.31	0.46	0.65	0.77	0.83	0.87

$\Gamma_Q \subset \Gamma_D$, an additional Dirichlet boundary for the slack variable. For our numerical tests, we chose the domain described in Example 2.1 and

$$\Gamma_Q = \{(x, y) \in \Gamma_D : x \in (0.5, 1) \text{ and } y = 0\}. \tag{4.1}$$

Example 4.1. *We apply the improved FOSLL* method to the BVP from Examples* 2.1 *and* 2.3. *We, thus, use the constructs defined in* (3.7)–(3.21), *with Γ_Q as in* (4.1) *and $d = 1$. The L^2 norms of the errors are shown in Table* 4.1. *Since the primal solution is in $(H^\alpha(\Omega))^4$ only for $\alpha < 2/3$, the optimal asymptotic bounds on these errors is in general proportional to $h^{2/3}$. We compute the approximate order of convergence by computing β such that $(1/2)^\beta$ is equal to the ratio of errors on consecutive grids. The table suggests that the improved FOSLL* approach does indeed achieve these optimal bounds, while the FOSLS and standard FOSLL* methods do not converge at all* (*cf. Tables* 2.1 *and* 2.2).

For the improved FOSLL* method, the four components of the dual solution, $(\mathbf{w}^h, r^h, s^h)$, on the $h = 1/32$ mesh are shown in Figure 4.1. By simply computing $(\mathbf{u}^h, p^h, q^h)^t = \mathcal{L}^*(\mathbf{w}^h, r^h, s^h)$, we obtained L^2 approximations for the primal variables, as shown in Figure 4.2. As these figures and tables show, the improved FOSLL* method yields converging L^2 approximations for the primal variables, $\mathbf{u}$ and p.

Unfortunately, convergence of $\|p-p^h\|_0$ tends to drop to a suboptimal rate if Γ_Q is chosen to be too small, especially when there are irregular points inside the Neumann boundary. Therefore, one has to take care in choosing Γ_Q sufficiently large. On the other hand, choosing such a large Γ_Q with a fixed length seems to inhibit optimal AMG performance: the average per-step residual error reduction factor, which we call ρ, seems to depend on the mesh size h. In fact, $1 - \rho$ seems to be proportional to h^α for some positive α. For Example 4.1, the multigrid reduction factor given in Table 4.1 suggests that $1 - \rho$ is proportional to $h^{3/4}$. This difficulty seems to come from the null space of $\mathcal{L}^*$ as defined in (3.7)–(3.21). This null space is nontrivial since there are no boundary conditions for the first two components of $\mathcal{D}(\mathcal{L}^*)$ on Γ_Q. One remedy could be to use linear solvers that can deal with nontrivial null spaces, such as the MINRES algorithm (cf. [15]). Another remedy is discussed in the next subsection.

Remark 4.2. For $c = 0$, the primal problem can be decomposed as shown in (3.32)–(3.33). Problem (3.32) can be solved by a FOSLL* method. Since this problem only involves the operator T, no Γ_Q is needed. Furthermore, the dual problem of this first stage has a full set of boundary conditions, namely, $w_1 = 0$ on Γ_D, $w_2 = 0$ on $\Gamma_{N,1}$, and $w_2 \equiv$ const. on $\Gamma_{N,i}$, $2 \leq i \leq M$. This full set of boundary conditions leads to optimal multigrid convergence for $c = 0$. To demonstrate this, we tested this approach on Example 4.1 with $c = 0$. Table 4.2 shows the approximation errors and

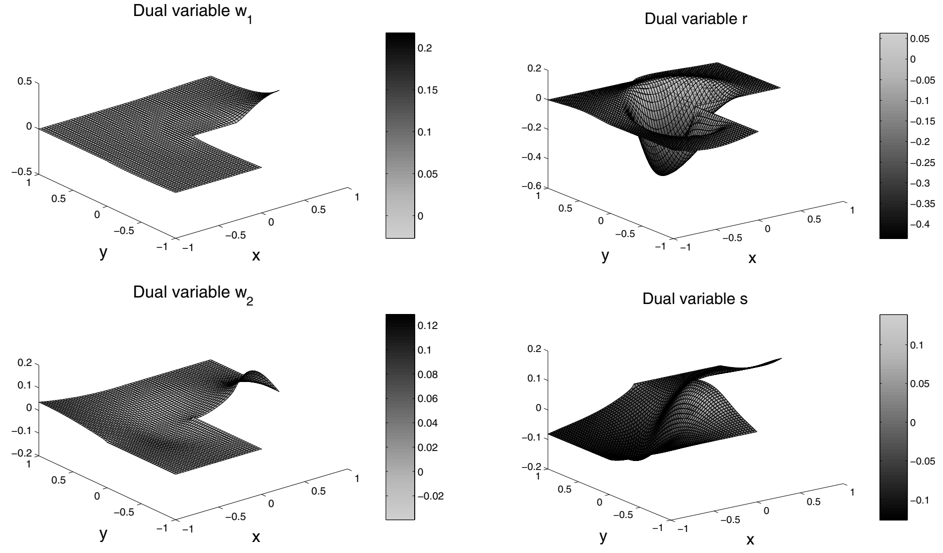


FIG. 4.1. *Approximations of the dual variables for Example* 4.1 *on a uniform mesh with* $h = 1/32$.

TABLE 4.2

Error norms and multigrid convergence for the approximations for Example 4.1 *for* $c = 0$ *on a sequence of uniform meshes with mesh sizes* h. *Upper half: First stage, using the improved FOSLL* method. Lower half: Second stage, using FOSLS method.*

h	1/4	1/8	1/16	1/32	1/64	1/128	1/256
$\|\mathbf{u}-\mathbf{u}^h\|_0$	0.4588	0.2036	0.1046	0.0539	0.0281	0.0149	0.0081
β		1.1718	0.961	0.956	0.941	0.916	0.882
ρ	0.12	0.10	0.054	0.041	0.032	0.040	0.040
$\|p-p^h\|_0$	4.41E-2	1.21E-2	3.34E-3	1.02E-3	3.34E-4	1.19E-4	4.63E-5
β		1.856	1.868	1.710	1.614	1.482	1.367
ρ	0.022	0.032	0.032	0.031	0.040	0.031	0.031

the multigrid convergence factors for both stages. Here, second-stage equation (3.33) is solved by FOSLS, since we know $p \in H^1(\Omega)$ and can, therefore, obtain H^1 approximations for p. Both finite element and multigrid convergence show optimal behavior.

4.1. Restoring optimal multigrid convergence. A heuristic approach for restoring optimal multigrid convergence (i.e., $\rho \ll 1$) is to choose different boundaries Γ_Q^h on different meshes so that $|\Gamma_Q^h| = O(h)$. The motivation for this is that such a choice for Γ_Q^h should control the dimension of the null space of the discrete operator since only a bounded number of elements could then intersect Γ_Q^h. These null space components that AMG cannot seem to eliminate by itself would then hopefully be attenuated by a fixed number of conjugate gradient steps.

The new difficulty that this choice introduces is that operators S^* and $\mathcal{L}^*$ lose surjectivity in the limit $h \to 0$. This, in turn, impairs finite element convergence as h decreases. Fortunately, this difficulty does not affect coercivity of T nor, as our observations show, convergence of $\|\mathbf{u}^h - \mathbf{u}\|$. Convergence of $\|p^h - p\|$ does degrade,

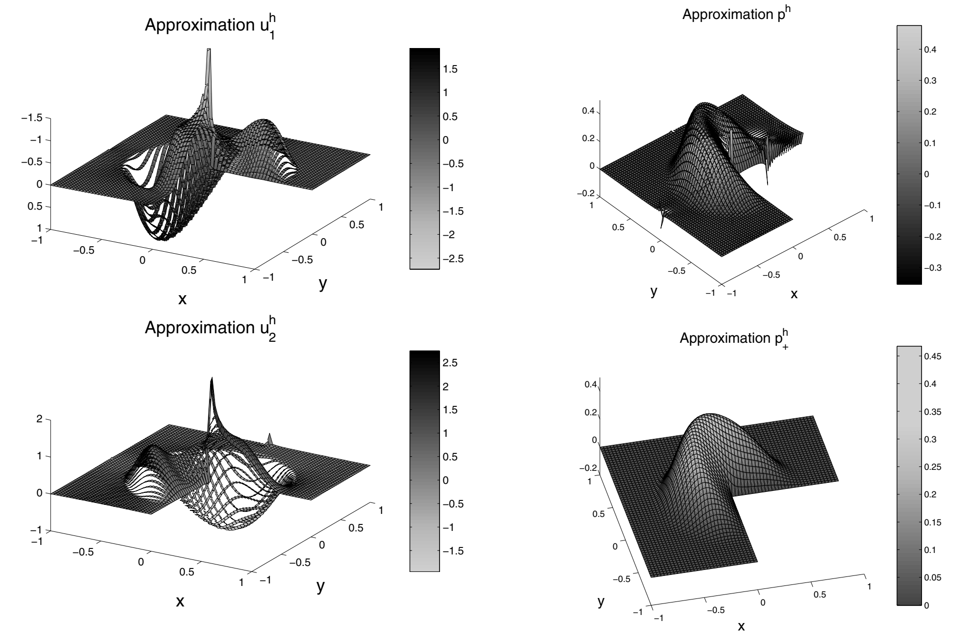


FIG. 4.2. *Approximations p^h and $\mathbf{u}^h$ of primal variables p and $\mathbf{u} = \nabla p$ as well as second-stage approximation p_+^h of p on a uniform mesh with $h = 1/32$ for Example* 4.1.

however, but this can be remedied by appealing to the relation $\nabla p = \mathbf{u}$. That is, we can simply replace p^h by a new approximation, denoted p_+^h, that approximately solves $\nabla p_+^h = \mathbf{u}^h$, where $\mathbf{u}^h$ is the approximation for $\mathbf{u}$ obtained by the improved FOSLL* method with variable Γ_Q^h. This postprocessing step is exactly the same as in (3.33), so we refer to it as the second stage. Since convergence of $\|\mathbf{u}^h - \mathbf{u}\|$ is still optimal for Γ_Q^h, then convergence of $\|p_+^h - p\|$ should be optimal as well. Our implementation solves this second stage by the FOSLS approach of finding $p_+^h = \arg\min \|\nabla z - \mathbf{u}^h\|_0$, where z is chosen from the same H^1-conforming finite element space that was used to approximate the dual solution.

EXAMPLE 4.3. *Using the same problem as in Example* 4.1*, we make a different choice for Γ_Q:*

$$\Gamma_Q^h = \{(x, y) \in \Gamma_D : x \in (1 - 4h, 1), y = 0\},$$

for which $|\Gamma_Q^h| = 4h$. In Table 4.3*, we list the L^2 errors associated with $\mathbf{u}^h$, p^h, and p_+^h. We also include the multigrid convergence factors, ρ, for the solution of the discretized dual problem and the computational cost of the second stage as a percentage of the computational cost of the solution of the discretized dual problem. The results show that our approach leads to optimal multigrid convergence and a very accurate approximation for p at very small additional cost. Approximations p_+^h and p^h for this problem on a mesh with $h = 1/32$ are shown in Figure* 4.2.

Remark 4.4. The second stage yields an H^1 approximation to p, while p^h is in general in $L^2 \backslash H^1$. This desirable feature of this new approach utilizes the higher regularity of p in an efficient way. The approximation is not only in a smoother space, but also more accurate. Thus, the second stage is generally an effective tool to improve convergence of FOSLL*, not only just for the case of variable Γ_Q^h.

TABLE 4.3

Upper half: Error norms, approximate order of discretization convergence, β, and multigrid convergence factors, ρ, for the improved FOSLL approximations for Example* 4.1 *on a sequence of uniform meshes with mesh sizes h and $\Gamma_Q^h \in O(h)$. Lower half: Error norms for the second-stage approximation, approximate order of discretization convergence, β, and work of the second stage as a percentage of the work of the FOSLL* method above.*

h	1/4	1/8	1/16	1/32	1/64	1/128	1/256
$\|p - p^h\|_0$	0.0475	0.0194	0.0125	0.0093	0.0071	0.0055	0.0043
β		1.29	0.639	0.430	0.385	0.372	0.365
$\|\mathbf{u} - \mathbf{u}^h\|_0$	0.6674	0.3051	0.1573	0.0810	0.0420	0.0221	0.0120
β		1.13	0.956	0.958	0.946	0.924	0.892
ρ	0.19	0.23	0.17	0.13	0.10	0.08	0.09
$\|p - p_+^h\|_0$	4.35E-2	1.20E-2	3.36E-3	1.09E-3	3.95E-4	1.64E-4	7.56E-5
β		1.857	1.837	1.625	1.464	1.271	1.113
$\|\mathbf{u} - \nabla p_+^h\|_0$	0.4640	0.2078	0.1071	0.0555	0.0290	0.0155	0.0084
β		1.159	0.955	0.950	0.934	0.908	0.874
ρ	0.022	0.032	0.032	0.031	0.040	0.031	0.031
*stage*2	3.1%	3.4%	3.7%	4.0%	5.0%	4.2%	4.5 %

TABLE 4.4

Error norms and AMG convergence factors for the approximations from Example 4.5 *for varying σ_0 on a sequence of uniform meshes with mesh sizes h.*

		h						
σ_0		1/4	1/8	1/16	1/32	1/64	1/128	1/256
10^0	$\|\mathbf{u} - \mathbf{u}^h\|_0$	0.6674	0.3051	0.1573	0.0810	0.0420	0.0221	0.0120
	β		1.13	0.956	0.958	0.946	0.924	0.892
	ρ	0.19	0.23	0.17	0.13	0.10	0.08	0.09
10^3	$\|\mathbf{u} - \mathbf{u}^h\|_0$	0.6732	0.3277	0.1672	0.0882	0.0483	0.0276	0.0166
	β		1.04	0.971	0.923	0.869	0.805	0.734
	ρ	0.38	0.47	0.53	0.49	0.40	0.31	0.23
10^6	$\|\mathbf{u} - \mathbf{u}^h\|_0$	0.6738	0.3280	0.1674	0.0883	0.0484	0.0277	0.0166
	β		1.04	0.971	0.923	0.869	0.805	0.734
	ρ	0.22	0.24	0.37	0.47	0.62	0.76	0.77

4.2. Dependence of A and b.

4.2. Dependence of A and b. Here we report on examples that demonstrate how FOSLL* depends on A and $\mathbf{b}$.

EXAMPLE 4.5. *In a first experiment, we slightly changed Example* 4.3 *by setting $A = \sigma I$ with $\sigma = 1$ for $x + y < 0$ and $\sigma = \sigma_0$ otherwise. The results are displayed in Table* 4.4 *and show that the AMG solver works well, even in the presence of huge jumps in the coefficients. It is remarkable that the AMG convergence factors are getting better for very fine meshes, where $1/h$ starts to dominate the convection and the jumping coefficients. For this example, we used the scaling mentioned in Remark* 3.9.

As a second experiment, we fixed $\sigma_0 = 1$ and varied the convection, $\mathbf{b}$. *The results of this experiment are shown in Table* 4.5. *Note again the relative insensitivity of the order of discretization error, now with respect to the size of* $\mathbf{b}$. *AMG performance does degrade with increasing size of* $\mathbf{b}$, *but this reflects the usual behavior of* standard *multigrid solvers for convection dominated problems. Again, as the mesh size tends to* 0, *the discretized differential operators dominate the convection and cause a steady*

TABLE 4.5

Error norms, approximate order of discretization convergence, β, and AMG convergence factors for the approximations from Example 4.5 for varying $\mathbf{b}$ on a sequence of uniform meshes with mesh sizes h.

		h				
$\mathbf{b}^t$		1/16	1/32	1/64	1/128	1/256
$(\frac{-y}{10}, 10x)$	$\|\mathbf{u} - \mathbf{u}^h\|_0$	0.1573	0.0810	0.0420	0.0221	0.0119
	β		0.958	0.946	0.924	0.892
	ρ	0.17	0.13	0.10	0.08	0.09
$10(\frac{-y}{10}, 10x)$	$\|\mathbf{u} - \mathbf{u}^h\|_0$	0.3356	0.1862	0.0968	0.0492	0.0250
	β		0.850	0.944	0.978	0.979
	ρ	0.57	0.60	0.58	0.45	0.31
$100(\frac{-y}{10}, 10x)$	$\|\mathbf{u} - \mathbf{u}^h\|_0$	0.4687	0.3371	0.2230	0.1342	0.0737
	β		0.476	0.596	0.732	0.864
	ρ	0.65	0.76	0.83	0.84	0.84

improvement of the AMG convergence rates.

5. Conclusions.

In this paper we have developed new FOSLL* formulations that allow the use of H^1-conforming finite element spaces and optimal multigrid solvers for constructing L^2 approximations of the primal variables on an extended class of scalar elliptic equations. This class includes problems with reaction terms, domains with Dirichlet and Neumann boundaries with multiple components, and irregular boundary points. The extension was accomplished by redefining the boundary conditions associated with the slack variables in the primal problem. Specifically, for domains with $\Gamma_D \neq \emptyset$, the slack variable, q, was given additional boundary conditions on $\Gamma_Q \subset \Gamma_D$. Our theory establishes the surjectivity of the adjoint operator, $\mathcal{L}^*$, as long as Γ_Q contains no irregular points. However, numerical results show that the multilevel solution techniques work better, and the finite element approximations are no worse, if Γ_Q is chosen to touch irregular boundary points and to shrink along with the mesh spacing, h. The case of pure Neumann boundary conditions remains an open problem.

The improved FOSLL* approach yields an L^2 approximation to the primal flux variable that achieves the optimal theoretical convergence rate. A postprocessing step was shown to yield optimal H^1 approximation to the original scalar variable, p, at a small additional cost.

We also showed that the FOSLL* formulation produces the same approximation as a Galerkin formulation of the original second-order boundary value problem, (2.1)–(2.3), in the absence of first order terms ($\mathbf{b} = \mathbf{0}$) and either no reaction term ($c = 0$) or strictly positive reaction term ($c > 0$).

The efficiency of the improved FOSLL* formulations was illustrated by a series of numerical examples.

REFERENCES

[1] M. Berndt, T. A. Manteuffel, S. F. McCormick, and G. Starke, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients: Part* I, SIAM J. Numer. Anal., 43 (2005), pp. 386–408.

[2] M. Berndt, T. A. Manteuffel, and S. F. McCormick, *Analysis of first-order system least squares (FOSLS) for elliptic problems with discontinuous coefficients: Part* II, SIAM J. Numer. Anal., 43 (2005), pp. 409–436.

[3] P. Bochev, Z. Cai, T. Manteuffel, and S. McCormick, *Analysis of velocity-flux least-squares principles for the Navier–Stokes equations: Part* II, SIAM J. Numer. Anal., 36 (1999), pp. 1125–1144.

[4] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[5] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *Least-squares methods for the Stokes equations based on a discrete minus one inner product*, J. Comput. Appl. Math., 74 (1996), pp. 155–173.

[6] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak, *A least-squares approach based on a discrete minus one inner product for first order systems*, Math. Comp., 66 (1997), pp. 935–955.

[7] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer, Berlin, 1994.

[8] S. C. Brenner and L. Y. Sung, *Multigrid methods for the computation of singular solutions and stress intensity factors* II, BIT, 37 (1997), pp. 623–643.

[9] Z. Cai and S. Kim, *A finite element method using singular functions for the poisson equation: Corner singularities*, SIAM J. Numer. Anal., 39 (2001), pp. 286–299.

[10] Z. Cai, R. Lazarov, T. Manteuffel, and S. McCormick, *First-order system least squares for second order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[11] Z. Cai, T. Manteuffel, and S. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[12] Z. Cai, T. Manteuffel, S. McCormick, and J. Ruge, *First-order system $\mathcal{LL}^*$ (FOSLL*): Scalar elliptic partial differential equations*, SIAM J. Numer. Anal., 39 (2001), pp. 1418–1445.

[13] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for the Stokes equations, with application to linear elasticity*, SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[14] C. L. Cox and G. J. Fix, *On the accuracy of least squares methods in the presence of corner singularities*, Comput. Math. Appl., 10 (1984), pp. 463–475.

[15] B. Fischer, *Polynomial Based Iteration Methods for Symmetric Linear Systems*, Wiley-Teubner, Chichester, Stuttgart, Germany, 1996.

[16] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations, Theory and Algorithms*, Springer, Berlin, 1986.

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

[18] L. Hörmander, *Linear Partial Differential Operators*, Springer, Berlin, 1963.

[19] T. A. Manteuffel, S. F. McCormick, and G. Starke, *First-order system least-squares for second-order elliptic problems with discontinuous coefficients*, in Proceedings of the Seventh Copper Mountain Conference on Multigrid Methods, N. D. Melson, T. A. Manteuffel, and S. F. McCormick, eds., NASA, Hampton, VA, 1995.

[20] A. I. Pehlivanov and G. F. Carey, *Error estimates for least-squares mixed finite elements*, RAIRO Modél. Math. Anal. Numér., 28 (1994), pp. 499–516.

[21] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[22] J. Ruge, *FOSPACK: A First-Order System Least-Squares (FOSLS) code*, in preparation.

[23] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice-Hall, Englewood Cliffs, NJ, 1973.

[24] C. R. Westphal, *First-Order System Least Squares (FOSLS) for Geometrically-Nonlinear Elasticity in Nonsmooth Domains*, Ph.D. thesis, University of Colorado at Boulder, 2004.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2121–2143

CONVERGENCE ANALYSIS OF THE PERFECTLY MATCHED LAYER PROBLEMS FOR TIME-HARMONIC MAXWELL'S EQUATIONS*

GANG BAO† AND HAIJUN WU‡

Abstract. This paper is concerned with convergence analysis of the perfectly matched layer (PML) problem in spherical coordinates for the three-dimensional electromagnetic scattering. Under some simple assumptions on the PML medium parameter, it is shown that the truncated PML problem attains a unique solution. The main result of the paper is to establish an explicit error estimate between the solution of the scattering problem and that of the truncated PML problem. The error estimate implies, in particular, that the PML solution converges exponentially to the scattering solution by increasing either the PML medium parameter or the PML layer thickness. The convergence result is expected to be useful for determining the PML medium parameter in the computational electromagnetic scattering problems.

Key words. Maxwell's equations, electromagnetic scattering, perfectly matched layer, convergence

AMS subject classifications. 65N15, 78A25, 35Q60

DOI. 10.1137/040604315

1. Introduction. Since the pioneering work of Bérenger [3, 4], the perfectly matched layer (PML) has become an increasingly important mesh termination technique in computational wave propagation due to its effectiveness, simplicity, and flexibility [5, 6, 7, 9, 10, 12, 13, 17, 18, 19]. The idea is to surround the computational domain by a nonphysical PML medium which has the remarkable property of being reflectionless for incident waves of any frequency or any incident direction, and the waves decay exponentially in magnitude into the PML medium. In practical computation, the PML medium must be truncated and the truncation boundary generates reflected waves which can pollute the solution in the computational domain. Therefore, it is imperative to study the error estimate in the computational domain between the solution of the wave propagation problem and that of the truncated PML problem.

For a simple two-dimensional model of the electromagnetic scattering by periodic structures, Chen and Wu have recently proved in [5] the exponential convergence with respect to a definite integral of the PML medium parameter. The exponential convergence can be achieved by increasing either the medium parameter or the PML layer thickness. Moreover, the error estimate given in [5] is explicit and can be used to determine the PML medium parameter according to the error tolerance in practical computation. Other convergence results of the PML problems for the Helmholtz scattering may be found in [9, 12, 13, 18]. To the best of our knowledge, there is no convergence result of the truncated PML problem for the three-dimensional

*Received by the editors February 18, 2004; accepted for publication (in revised form) January 28, 2005; published electronically December 30, 2005. This research was partially supported by NSF grant DMS 01-04001, ONR grant N000140210365, and NSF of China grants 10428105, 10401016.
http://www.siam.org/journals/sinum/43-5/60431.html

†Department of Mathematics, Michigan State University, East Lansing, MI 48824 (bao@math.msu.edu).

‡Department of Mathematics, Nanjing University, Jiangsu 210093, People's Republic of China (hjw@nju.edu.cn). This author's research was supported by the national basic research program under grant 2005CB32170X.

electromagnetic scattering, though the PML technique has been the subject of a substantial engineering literature.

Our goal in this paper is to analyze the convergence of PML solutions for the three-dimensional electromagnetic scattering. Attempts are made to generalize the results of [5] to the PML in spherical coordinates for Maxwell's equations. However, the techniques completely differ due to the more complicated model in the three-dimensional case. Under the assumption that there is a unique solution of the original scattering problem and a proper assumption on the PML medium parameter, we prove in this paper that the truncated PML problem attains a unique solution in $H(\mathrm{curl})$ and obtain an explicit error estimate between the solution of the scattering problem and the solution of the truncated PML problem in the computational domain. The error estimate implies particularly that the PML solution converges exponentially to the scattering solution when either the PML medium parameter or the layer thickness is increased. The significance of our main result is twofold:

- new error estimates for a fixed PML layer thickness;
- explicit constants in the estimates which may be used to determine the PML constants in the computational electromagnetic scattering.

Our proof is based on a variational approach. A crucial step is to conduct a careful analysis of special functions, for example, the Bessel functions and the spherical Hankel functions.

2. PML formulation in spherical coordinates.

We first introduce the scattering problem. Suppose that a bounded medium characterized by permittivity ε and permeability μ is illuminated by a time-harmonic electromagnetic wave (E^{in}, H^{in}), where E^{in} is the electric field and H^{in} is the magnetic field. The incoming wave (E^{in}, H^{in}) is assumed to be a classical solution of the Maxwell system

$$\operatorname{curl} E^{in} = \mathbf{i}\omega\mu_0 H^{in} \quad \text{and} \quad \operatorname{curl} H^{in} = -\mathbf{i}\omega\varepsilon_0 E^{in} \quad \text{in } \mathbb{R}^3. \tag{2.1}$$

Here ε_0 and μ_0 are two positive constants. The interaction of the incident field and the medium gives rise to the scattered field (E^{sc}, H^{sc}). Let ε, μ be two functions of the spatial variable $x = (x_1, x_2, x_3)^T$, and $r = |x|$. Then the scattered field satisfies the time-harmonic Maxwell system

$$\begin{cases} \operatorname{curl} E^{sc} = \mathbf{i}\omega\mu H^{sc} + \mathbf{i}\omega(\mu H^{in} - \mu_0 H^{in}) & \text{in } \mathbb{R}^3, \\ \operatorname{curl} H^{sc} = -\mathbf{i}\omega\varepsilon E^{sc} - \mathbf{i}\omega(\varepsilon E^{in} - \varepsilon_0 E^{in}) & \text{in } \mathbb{R}^3, \\ \lim_{r\to+\infty}(\sqrt{\mu_0} H^{sc} \wedge x - r\sqrt{\varepsilon_0} E^{sc}) = 0. \end{cases} \tag{2.2}$$

Assume that ε and μ, respectively, are in $(L^\infty(\mathbb{R}^3))^{3\times 3}$, and so are ε^{-1} and μ^{-1}. Assume also that the inhomogeneity is bounded so that there is a constant $R > 0$ such that $\varepsilon(x) = \varepsilon_0 I$ and $\mu(x) = \mu_0 I$ if $r = |x| \geq R$, where I is the 3×3 identity matrix. Denote the ball of radius by R and its surface by $\Omega = \{x \in \mathbb{R}^3, |x| < R\}$, respectively, and $S = \{x \in \mathbb{R}^3, |x| = R\}$.

Next, introduce the PML medium in spherical coordinates (r, θ, φ), where θ and φ are the Euler angles: $x_1 = r \sin\theta\cos\varphi,\ x_2 = r\sin\theta\sin\varphi,\ x_3 = r\cos\theta$.

For any point x, let $\vec{e_r}, \vec{e_\theta}$, and $\vec{e_\varphi}$ be the local orthonormal basis, i.e.,

$$\begin{cases} \vec{e_r} = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta)^T = x/r, \\ \vec{e_\theta} = (\cos\theta\cos\varphi, \cos\theta\sin\varphi, -\sin\theta)^T, \\ \vec{e_\varphi} = (-\sin\varphi, \cos\varphi, 0)^T. \end{cases} \tag{2.3}$$

For any vector field $v = (v_1, v_2, v_3)^T$, denote by v_r, v_θ, and v_φ the projections of v onto $\vec{e_r}$, $\vec{e_\theta}$, and $\vec{e_\varphi}$, respectively,

$$v_r = v \cdot \vec{e_r}, \quad v_\theta = v \cdot \vec{e_\theta}, \quad v_\varphi = v \cdot \vec{e_\varphi}. \tag{2.4}$$

Let $Q = \begin{pmatrix} \vec{e_r} & \vec{e_\theta} & \vec{e_\varphi} \end{pmatrix}$ be a 3×3 matrix composed of $\vec{e_r}, \vec{e_\theta}$, and $\vec{e_\varphi}$. It is clear that

$$\begin{pmatrix} v_r & v_\theta & v_\varphi \end{pmatrix}^T = Q^T \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}^T, \quad \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix}^T = Q \begin{pmatrix} v_r & v_\theta & v_\varphi \end{pmatrix}^T. \tag{2.5}$$

Furthermore,

$$\begin{aligned} \operatorname{curl} v = {} & \frac{1}{r \sin\theta} \left(\frac{\partial(\sin\theta v_\varphi)}{\partial\theta} - \frac{\partial v_\theta}{\partial\varphi} \right) \vec{e_r} + \left(\frac{1}{r\sin\theta} \frac{\partial v_r}{\partial\varphi} - \frac{1}{r} \frac{\partial(r v_\varphi)}{\partial r} \right) \vec{e_\theta} \\ & + \frac{1}{r} \left(\frac{\partial(r v_\theta)}{\partial r} - \frac{\partial v_r}{\partial\theta} \right) \vec{e_\varphi}. \end{aligned} \tag{2.6}$$

Obviously, the scattered field (E^{sc}, H^{sc}) outside the ball Ω satisfies the following Faraday equation and Ampere equation (Maxwell's equations):

$$\operatorname{curl} E^{sc} = \mathbf{i}\omega\mu_0 H^{sc} \quad \text{and} \quad \operatorname{curl} H^{sc} = -\mathbf{i}\omega\varepsilon_0 E^{sc} \quad \text{in } \mathbb{R}^3 \setminus \overline{\Omega}. \tag{2.7}$$

In the spherical coordinates, Faraday's equation becomes

$$\begin{cases} \dfrac{1}{r\sin\theta} \left(\dfrac{\partial(\sin\theta E^{sc}_\varphi)}{\partial\theta} - \dfrac{\partial E^{sc}_\theta}{\partial\varphi} \right) = \mathbf{i}\omega\mu_0 H^{sc}_r & \text{in } \mathbb{R}^3 \setminus \overline{\Omega}, \\ \dfrac{1}{r\sin\theta} \dfrac{\partial E^{sc}_r}{\partial\varphi} - \dfrac{1}{r} \dfrac{\partial(r E^{sc}_\varphi)}{\partial r} = \mathbf{i}\omega\mu_0 H^{sc}_\theta & \text{in } \mathbb{R}^3 \setminus \overline{\Omega}, \\ \dfrac{1}{r} \left(\dfrac{\partial(r E^{sc}_\theta)}{\partial r} - \dfrac{\partial E^{sc}_r}{\partial\theta} \right) = \mathbf{i}\omega\mu_0 H^{sc}_\varphi & \text{in } \mathbb{R}^3 \setminus \overline{\Omega}. \end{cases} \tag{2.8}$$

Following Teixeira and Chew [17], we introduce the PML problem by a change of variables,

$$r \to \int_0^{\hat r} s(\tau)\, d\tau, \tag{2.9}$$

where $s(\tau) = 1 + \mathbf{i}s_I(\tau)$ is continuous, $s_I(\tau) \geq 0$ and $s_I(\tau) = 0$ for $0 \leq \tau \leq R$. In the Cartesian coordinates, the change of variables is equivalent to

$$x \to \hat{x} = (\hat r \sin\theta\cos\varphi, \hat r \sin\theta\sin\varphi, \hat r\cos\theta). \tag{2.10}$$

It is clear that $s(\tau) = 1$ and $r = \hat r$, $x = \hat x$ for $0 \leq \hat r \leq R$.

Noting that $\partial/\partial r = (1/s(\hat r))\partial/\partial \hat r$ and $E^{sc}_r = E^{sc}_{\hat r}$, $H^{sc}_r = H^{sc}_{\hat r}$, we have from (2.8) that

$$\begin{cases} \dfrac{1}{r\sin\theta} \left(\dfrac{\partial(\sin\theta E^{sc}_\varphi)}{\partial\theta} - \dfrac{\partial E^{sc}_\theta}{\partial\varphi} \right) = \mathbf{i}\omega\mu_0 H^{sc}_{\hat r} & \text{in } \mathbb{R}^3 \setminus \overline{\Omega}, \\ \dfrac{1}{r\sin\theta} \dfrac{\partial E^{sc}_{\hat r}}{\partial\varphi} - \dfrac{1}{r s(\hat r)} \dfrac{\partial(r E^{sc}_\varphi)}{\partial \hat r} = \mathbf{i}\omega\mu_0 H^{sc}_\theta & \text{in } \mathbb{R}^3 \setminus \overline{\Omega}, \\ \dfrac{1}{r} \left(\dfrac{1}{s(\hat r)} \dfrac{\partial(r E^{sc}_\theta)}{\partial \hat r} - \dfrac{\partial E^{sc}_{\hat r}}{\partial\theta} \right) = \mathbf{i}\omega\mu_0 H^{sc}_\varphi & \text{in } \mathbb{R}^3 \setminus \overline{\Omega}. \end{cases}$$

Multiplying the first equation by $(r/\hat{r})^2$, the second and the third equations by $s(\hat{r})(r/\hat{r})$, and denoting by

$$\begin{aligned} E_{\hat{r}}^{sc,\mathrm{PML}} &= s(\hat{r})E_{\hat{r}}^{sc}, \quad E_{\theta}^{sc,\mathrm{PML}} = \frac{r}{\hat{r}}E_{\theta}^{sc}, \quad E_{\varphi}^{sc,\mathrm{PML}} = \frac{r}{\hat{r}}E_{\varphi}^{sc}, \\ H_{\hat{r}}^{sc,\mathrm{PML}} &= s(\hat{r})H_{\hat{r}}^{sc}, \quad H_{\theta}^{sc,\mathrm{PML}} = \frac{r}{\hat{r}}H_{\theta}^{sc}, \quad H_{\varphi}^{sc,\mathrm{PML}} = \frac{r}{\hat{r}}H_{\varphi}^{sc}, \end{aligned} \tag{2.11}$$

we obtain the Faraday equation for the PML medium in $\{\hat{x} \in \mathbb{R}^3 \setminus \overline{\Omega}\}$:

$$\begin{cases} \dfrac{1}{\hat{r}\sin\theta}\left(\dfrac{\partial(\sin\theta E_{\varphi}^{sc,\mathrm{PML}})}{\partial\theta} - \dfrac{\partial E_{\theta}^{sc,\mathrm{PML}}}{\partial\varphi}\right) = \mathbf{i}\omega\mu_0\left(\left(\dfrac{r}{\hat{r}}\right)^2\dfrac{1}{s(\hat{r})}\right)H_{\hat{r}}^{sc,\mathrm{PML}}, \\ \dfrac{1}{\hat{r}\sin\theta}\dfrac{\partial E_{\hat{r}}^{sc,\mathrm{PML}}}{\partial\varphi} - \dfrac{1}{\hat{r}}\dfrac{\partial(\hat{r}E_{\varphi}^{sc,\mathrm{PML}})}{\partial\hat{r}} = \mathbf{i}\omega\mu_0 s(\hat{r})H_{\theta}^{sc,\mathrm{PML}}, \\ \dfrac{1}{\hat{r}}\left(\dfrac{\partial(\hat{r}E_{\theta}^{sc,\mathrm{PML}})}{\partial\hat{r}} - \dfrac{\partial E_{\hat{r}}^{sc,\mathrm{PML}}}{\partial\theta}\right) = \mathbf{i}\omega\mu_0 s(\hat{r})H_{\varphi}^{sc,\mathrm{PML}}. \end{cases} \tag{2.12}$$

The Ampere equation for the PML medium may be derived similarly.

Furthermore, from (2.5) and (2.6), we rewrite Maxwell's equations for the PML medium in the Cartesian coordinates as

$$\mathrm{curl}_{\hat{x}}\, E^{sc,\mathrm{PML}} = \mathbf{i}\omega\hat{\mu}H^{sc,\mathrm{PML}} \quad \text{and} \quad \mathrm{curl}_{\hat{x}}\, H^{sc,\mathrm{PML}} = -\mathbf{i}\omega\hat{\varepsilon}E^{sc,\mathrm{PML}},$$

where

$$\hat{\varepsilon} = \widehat{Q}\varepsilon, \quad \hat{\mu} = \widehat{Q}\mu, \quad \widehat{Q} = Q\,\mathrm{diag}((r/\hat{r})^2/s(\hat{r}),\ s(\hat{r}),\ s(\hat{r}))Q^T, \tag{2.13}$$

and $Q = \begin{pmatrix}\vec{e_r} & \vec{e_\theta} & \vec{e_\varphi}\end{pmatrix}$. Note that $\hat{\varepsilon} = \varepsilon$ and $\hat{\mu} = \mu$ for $x \in \Omega$.

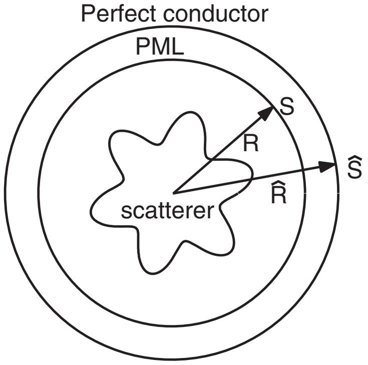


FIG. 2.1. *Geometry of the scattering with truncated PML.*

In practical computation, the PML medium is truncated by a perfect conductor boundary condition on $\widehat{S} = \{\hat{x} \in \mathbb{R}^3, |\hat{x}| = \widehat{R}\}$ for some $\widehat{R} > R$ (see Figure 2.1). Denote by $\widehat{\Omega} = \{\hat{x} \in \mathbb{R}^3, |\hat{x}| < \widehat{R}\}$. Then the scattering problem with a truncated PML takes the following form: find $(\widehat{E}^{sc,\mathrm{PML}}, \widehat{H}^{sc,\mathrm{PML}})$ such that

$$\begin{cases} \mathrm{curl}_{\hat{x}}\, \widehat{E}^{sc,\mathrm{PML}} = \mathbf{i}\omega\hat{\mu}\widehat{H}^{sc,\mathrm{PML}} + \mathbf{i}\omega(\mu H^{in} - \mu_0 H^{in}) & \text{in } \widehat{\Omega}, \\ \mathrm{curl}_{\hat{x}}\, \widehat{H}^{sc,\mathrm{PML}} = -\mathbf{i}\omega\hat{\varepsilon}\widehat{E}^{sc,\mathrm{PML}} - \mathbf{i}\omega(\varepsilon E^{in} - \varepsilon_0 E^{in}) & \text{in } \widehat{\Omega}, \\ \widehat{E}^{sc,\mathrm{PML}} \wedge n = 0 & \text{on } \widehat{S}. \end{cases} \tag{2.14}$$

The goal of the paper is to estimate the error between $(\widehat{E}^{sc,\mathrm{PML}}, \widehat{H}^{sc,\mathrm{PML}})$ and (E^{sc}, H^{sc}).

3. The main result. We begin with the variational form of the scattering problem (2.2). For any smooth vector field v, denote by v_S its tangential component on the surface S:

$$v_S = -(v \wedge n) \wedge n \quad \text{on } S, \tag{3.1}$$

where $n = \vec{e_r}$ is the unit outer normal vector to S. Introduce the following usual functional spaces:

$$\begin{aligned}
&H(\mathrm{curl}, \Omega) = \{u \in (L^2(\Omega))^3,\ \mathrm{curl}\, u \in (L^2(\Omega))^3\},\\
&TL^2(S) = \{u \in (L^2(S))^3,\ u \cdot n = 0\},\\
&TH^{-1/2}(\mathrm{curl}, S) = \{u \in (H^{-1/2}(S))^3,\ u \cdot n = 0,\ \mathrm{curl}_S\, u \in H^{-1/2}(S)\},\\
&TH^{-1/2}(\mathrm{div}, S) = \{u \in (H^{-1/2}(S))^3,\ u \cdot n = 0,\ \mathrm{div_S}\, u \in H^{-1/2}(S)\}.
\end{aligned}$$

For the definitions of the surface divergence $\mathrm{div_S}$ and the scalar rotational curl_S, we refer to [14]. Recall the following Paquet duality result [16]: $TH^{-1/2}(\mathrm{curl}, S) = TH^{-1/2}(\mathrm{div}, S)'$.

Following Ammari and Nédélec [1], define the capacity operator $\mathcal{T}_S$ from $TH^{-1/2}$ (curl, S) to $TH^{-1/2}(\mathrm{div}, S)$:

$$\mathcal{T}_S u = \mathcal{H} \wedge n, \tag{3.2}$$

where

$$\begin{cases}
\mathrm{curl}\, \mathcal{E} = \mathbf{i}\omega\mu_0 \mathcal{H} \quad \text{and} \quad \mathrm{curl}\, \mathcal{H} = -\mathbf{i}\omega\varepsilon_0 \mathcal{E} & \text{in } \mathbb{R}^3 \setminus \overline{\Omega},\\
\mathcal{E}_S = u & \text{on } S,\\
\lim_{r\to+\infty} \left(\sqrt{\mu_0}\mathcal{H} \wedge x - r\sqrt{\varepsilon_0}\mathcal{E}\right) = 0.
\end{cases} \tag{3.3}$$

From (2.2), it is easily seen that

$$H^{sc} \wedge n = \mathcal{T}_S E^{sc}_S \quad \text{on } S. \tag{3.4}$$

By eliminating the magnetic field H^{sc} from (2.2), we obtain

$$\mathrm{curl}(\mu^{-1}\, \mathrm{curl}\, E^{sc}) - \omega^2 \varepsilon E^{sc} = f^{in} \quad \text{in } \Omega, \tag{3.5}$$

where

$$f^{in} = -\mathbf{i}\omega\mu_0\, \mathrm{curl}(\mu^{-1} H^{in}) + \omega^2 \varepsilon E^{in}. \tag{3.6}$$

Multiplying (3.5) by a test function $\phi \in H(\mathrm{curl}, \Omega)$, integrating over Ω, and using integration by parts, we arrive at the variational form for the scattering problem (2.2): find $E^{sc} \in H(\mathrm{curl}, \Omega)$ such that

$$A(E^{sc}, \phi) = \langle f^{in}, \phi\rangle \quad \forall \phi \in H(\mathrm{curl}, \Omega), \tag{3.7}$$

where the bilinear form

$$A(v, \phi) = \int_\Omega \mu^{-1}\, \mathrm{curl}\, v \cdot \mathrm{curl}\, \phi - \omega^2 \int_\Omega \varepsilon v \cdot \phi - \mathbf{i}\omega \int_S \mathcal{T}_S v_S \cdot \phi_S \tag{3.8}$$

and

$$\langle f^{in}, \phi\rangle = -\mathbf{i}\omega\mu_0 \int_\Omega \mu^{-1} H^{in}\cdot \operatorname{curl}\phi + \mathbf{i}\omega \int_S H^{in}\wedge n\cdot \phi_S + \omega^2 \int_\Omega \varepsilon E^{in}\cdot\phi. \tag{3.9}$$

Assume in the following that the variational problem (3.7) attains a unique solution. Then the general theory in Babuška and Aziz [2, Chapter 5] implies that there exists a constant $\gamma_1 > 0$ such that the following inf-sup condition holds:

$$\sup_{0\neq\phi\in H(\operatorname{curl},\Omega)} \frac{|A(v,\phi)|}{\|\phi\|_{H(\operatorname{curl},\Omega)}} \geq \gamma_1 \|v\|_{H(\operatorname{curl},\Omega)} \quad \forall v\in H(\operatorname{curl},\Omega). \tag{3.10}$$

See also Kirsch and Monk [11] for additional discussions on the variational problem.

Similarly, we introduce a variational form for the truncated PML scattering problem (2.14) by defining a capacity operator $\mathcal{T}_S^{\mathrm{PML}}$ from $TH^{-1/2}(\operatorname{curl}, S)$ to $TH^{-1/2}(\operatorname{div}, S)$:

$$\mathcal{T}_S^{\mathrm{PML}} u = \mathcal{H}^{\mathrm{PML}}\wedge n, \tag{3.11}$$

where

$$\begin{cases} \operatorname{curl}_{\hat{x}} \mathcal{E}^{\mathrm{PML}} = \mathbf{i}\omega\hat{\mu}\mathcal{H}^{\mathrm{PML}} \quad \text{and} \quad \operatorname{curl}_{\hat{x}} \mathcal{H}^{\mathrm{PML}} = -\mathbf{i}\omega\hat{\varepsilon}\mathcal{E}^{\mathrm{PML}} & \text{in } \mathbb{R}^3\setminus\overline{\Omega}, \\ \mathcal{E}_S^{\mathrm{PML}} = u & \text{on } S, \\ \mathcal{E}^{\mathrm{PML}}\wedge n = 0 & \text{on } \widehat{S}. \end{cases} \tag{3.12}$$

It follows from (2.14) that

$$\widehat{H}^{sc,\mathrm{PML}}\wedge n = \mathcal{T}_S^{\mathrm{PML}} \widehat{E}_S^{sc,\mathrm{PML}} \quad \text{on } S. \tag{3.13}$$

For $x\in\Omega$, since $\hat{\varepsilon} = \varepsilon$ and $\hat{\mu} = \mu$, the fields $(\widehat{E}^{sc,\mathrm{PML}}, \widehat{H}^{sc,\mathrm{PML}})$ and (E^{sc}, H^{sc}) satisfy the same equation. Therefore, we have the variational form of (2.14): find $\widehat{E}^{sc,\mathrm{PML}}\in H(\operatorname{curl},\Omega)$ such that

$$A^{\mathrm{PML}}(\widehat{E}^{sc,\mathrm{PML}}, \phi) = \langle f^{in}, \phi\rangle \quad \forall\phi\in H(\operatorname{curl},\Omega), \tag{3.14}$$

where the bilinear form

$$A^{\mathrm{PML}}(v,\phi) = \int_\Omega \mu^{-1}\operatorname{curl} v\cdot\operatorname{curl}\phi - \omega^2\int_\Omega \varepsilon v\cdot\phi - \mathbf{i}\omega\int_S \mathcal{T}_S^{\mathrm{PML}} v_S\cdot\phi_S. \tag{3.15}$$

In order to estimate the error between $\widehat{E}^{sc,\mathrm{PML}}$ and E^{sc}, it is sufficient to estimate the error between the two capacity operators $\mathcal{T}_S^{\mathrm{PML}}$ and $\mathcal{T}_S$. We have the following important lemma that will be proved in section 6.

Lemma 3.1. *Let*

$$\widetilde{R}_I = \int_R^{\widehat{R}} s_I(\tau)\,d\tau \quad \textit{and} \quad a = \min\left\{\frac{1}{2}, \frac{kR}{5}\right\}. \tag{3.16}$$

Suppose

$$\widetilde{R}_I \geq \max\{7R/5,\ \widehat{R},\ 17/k\}. \tag{3.17}$$

Then for any $v_S, \phi_S \in TH^{-1/2}(\mathrm{curl}, S)$,

$$\left|\omega \int_S (\mathcal{T}_S^{\mathrm{PML}} - \mathcal{T}_S) v_S \cdot \phi_S\right| \leq \mathcal{M} \, \|v_S\|_{TH^{-1/2}(\mathrm{curl},S)} \, \|\phi_S\|_{TH^{-1/2}(\mathrm{curl},S)} \,,$$

where

$$\mathcal{M} = \frac{4k(a\mu_0)^{-1} \max\{(kR)^2(3kR+3/2)^2, 1\}}{e^{k\widetilde{R}_I[2-(a\widetilde{R}_I/R)^{-2}+(a\widetilde{R}_I/R)^{-4}/19]} - 10}. \tag{3.18}$$

Remark 3.1. The constant $\widetilde{R}_I$ is known as the PML parameter. Here, we examine the structure of the constant $\mathcal{M}$ which controls the modeling error between the PML equation and the original scattering problem (see Theorem 3.2). Obviously, the constant $\mathcal{M}$ approaches zero exponentially as the PML parameters $\widetilde{R}_I$ goes to infinity. From definition (3.16), $\widetilde{R}_I$ may be calculated by the medium property $s_I(\tau)$, which is usually taken as a power function

$$s_I(\tau) = \delta_m[(\tau - R)/(\widehat{R} - R)]^m \quad \text{for } \tau \geq R, \quad m \geq 1.$$

Thus,

$$\widetilde{R}_I = \delta_m(\widehat{R} - R)/(m+1). \tag{3.19}$$

It is obvious that the PML approximation error is reduced by either enlarging the medium parameter δ_m or increasing the layer thickness $\widehat{R} - R$.

Recall the trace regularity result for $H(\mathrm{curl}, \Omega)$ (cf. [14]):

$$\|v_S\|_{TH^{-1/2}(\mathrm{curl},S)} \leq \gamma_0 \, \|v\|_{H(\mathrm{curl},\Omega)} \quad \forall v \in H(\mathrm{curl}, \Omega), \tag{3.20}$$

where γ_0 is a positive constant. We are now ready to present the main result of this paper.

Theorem 3.2. *Assume that* (3.10) *and* (3.17) *hold.*

(i) *If* $\mathcal{M}\gamma_0^2 < \gamma_1$, *then the PML variational problem* (3.14) *attains a unique solution. Furthermore, the following a priori estimate (dependent on the original scattering solution) holds:*

$$\begin{aligned} |||\widehat{E}^{sc,\mathrm{PML}} - E^{sc}||| &:= \sup_{0 \neq \phi \in H(\mathrm{curl},\Omega)} \frac{|A(\widehat{E}^{sc,\mathrm{PML}} - E^{sc}, \phi)|}{\|\phi\|_{H(\mathrm{curl},\Omega)}} \\ &\leq \frac{\mathcal{M}\gamma_0^2}{1 - \mathcal{M}\gamma_0^2/\gamma_1} \|E^{sc}\|_{H(\mathrm{curl},\Omega)}. \end{aligned} \tag{3.21}$$

(ii) *If the PML variational problem* (3.14) *has a solution* $\widehat{E}^{sc,\mathrm{PML}} \in H(\mathrm{curl}, \Omega)$, *then the following a posteriori estimate (dependent on the PML solution) holds:*

$$|||\widehat{E}^{sc,\mathrm{PML}} - E^{sc}||| \leq \mathcal{M}\gamma_0^2 \|\widehat{E}^{sc,\mathrm{PML}}\|_{H(\mathrm{curl},\Omega)}, \tag{3.22}$$

where γ_0 *is defined in* (3.20), γ_1 *and* $\mathcal{M}$ *are defined in* (3.10) *and* (3.18), *respectively.*

Proof. We first prove (ii). By the definitions (3.8) and (3.15) of A and A^{PML}, Lemma 3.1, and the trace regularity (3.20), we have

$$\begin{aligned} |A^{\mathrm{PML}}(v, \phi) - A(v, \phi)| &= \left|\omega \int_S (\mathcal{T}_S^{\mathrm{PML}} - \mathcal{T}_S) v_S \cdot \phi_S\right| \\ &\leq \mathcal{M}\gamma_0^2 \, \|v\|_{H(\mathrm{curl},\Omega)} \, \|\phi\|_{H(\mathrm{curl},\Omega)}. \end{aligned} \tag{3.23}$$

Hence from (3.7) and (3.14), we conclude that

$$\begin{aligned}|A(\widehat{E}^{sc,\mathrm{PML}} - E^{sc}, \phi)| &= |A(\widehat{E}^{sc,\mathrm{PML}}, \phi) - A^{\mathrm{PML}}(\widehat{E}^{sc,\mathrm{PML}}, \phi)| \\ &\le \mathcal{M}\gamma_0^2 \|\widehat{E}^{sc,\mathrm{PML}}\|_{H(\mathrm{curl},\Omega)} \|\phi\|_{H(\mathrm{curl},\Omega)},\end{aligned}$$

which implies (3.22).

Next we prove (i). From (3.10), the assumption $\mathcal{M}\gamma_0^2 < \gamma_1$, and (3.23), we conclude that the bilinear form, $A^{\mathrm{PML}} : H(\mathrm{curl},\Omega) \times H(\mathrm{curl},\Omega) \to \mathbf{C}$, defined in (3.15) satisfies the inf-sup condition. Therefore, the PML variational problem (3.14) attains a unique solution. Finally, the error estimate (3.21) follows from (3.22) and (3.10). □

Remark 3.2. Since the norm $|||\cdot|||$ is equivalent to the norm $\|\cdot\|_{H(\mathrm{curl},\Omega)}$ with $|||\cdot||| \ge \gamma_1 \|\cdot\|_{H(\mathrm{curl},\Omega)}$ (cf. (3.10)), the error estimate between E^{sc} and $\widehat{E}^{sc,\mathrm{PML}}$ in $H(\mathrm{curl},\Omega)$ is easily obtained from Theorem 3.2. Having established the error estimate for the electric field, the error estimate between the magnetic fields H^{sc} and $\widehat{H}^{sc,\mathrm{PML}}$ in $H(\mathrm{curl},\Omega)$ may be obtained by examining the following system:

$$\begin{cases}\widehat{H}^{sc,\mathrm{PML}} - H^{sc} = (\mathbf{i}\omega\mu)^{-1}\,\mathrm{curl}(\widehat{E}^{sc,\mathrm{PML}} - E^{sc}) & \text{in } \Omega,\\ \mathrm{curl}(\widehat{H}^{sc,\mathrm{PML}} - H^{sc}) = -\mathbf{i}\omega\varepsilon(\widehat{E}^{sc,\mathrm{PML}} - E^{sc}) & \text{in } \Omega,\end{cases}$$

which is derived from (2.2), (2.14), and the fact that $\hat{x} = x$, $\hat{\varepsilon} = \varepsilon$, and $\hat{\mu} = \mu$ for $x \in \Omega$.

4. Capacity operators. This section is devoted to the derivation of explicit representations of the capacity operators $\mathcal{T}_S$ and $\mathcal{T}_S^{\mathrm{PML}}$. We first present series solutions for (3.3) and (3.12). We also give explicit representations of the capacity operators $\mathcal{T}_S$ and $\mathcal{T}_S^{\mathrm{PML}}$.

We start by representing the boundary data in terms of suitable vector basis functions on S. Following [8], let $Y_l^m(\theta,\varphi)$ be an orthonormal sequence of spherical harmonics on the unit sphere that satisfies (cf. [8])

$$\Delta_S Y_l^m + l(l+1)Y_l^m = 0, \tag{4.1}$$

where $\Delta_S = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\frac{\partial}{\partial\theta}) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}$ is the Laplace–Beltrami operator on S. Let $\nabla_S = \vec{e_\theta}\frac{\partial}{\partial\theta} + \vec{e_\varphi}\frac{1}{\sin\theta}\frac{\partial}{\partial\varphi}$ be the tangential gradient on S. Then an orthonormal basis for $TL^2(S)$ (the tangential fields on S) consists of functions of the form

$$V_l^m = \frac{1}{R\sqrt{l(l+1)}}\nabla_S Y_l^m \quad \text{and} \quad U_l^m = V_l^m \wedge n. \tag{4.2}$$

It follows that any tangential vector field $u \in TL^2(S)$ may be represented as

$$u = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}\left[c_l^m U_l^m + d_l^m V_l^m\right].$$

Using the series coefficients (see [14] or [1]), the norm on the space $TH^{-1/2}(\mathrm{curl},S)$ may be characterized by

$$\|u\|_{TH^{-1/2}(\mathrm{curl},S)}^2 = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}\left[\sqrt{1+l(l+1)}\,|c_l^m|^2 + \frac{1}{\sqrt{1+l(l+1)}}\,|d_l^m|^2\right]. \tag{4.3}$$

To represent the solutions, we need the first kind and second kind of spherical Hankel functions,

$$h_l^{(1)}(z) = (-\mathbf{i})^l \frac{e^{\mathbf{i}z}}{z} \sum_{m=0}^{l} \mathbf{i}^m \frac{(m+l)!}{m!(l-m)!2^m z^m}, \quad h_l^{(2)}(z) = \overline{h_l^{(1)}(\bar{z})}. \tag{4.4}$$

Define

$$\mathsf{z}_l^{(1)}(z) = z\frac{\frac{d}{dz}h_l^{(1)}(z)}{h_l^{(1)}(z)}, \quad \mathsf{z}_l^{(2)}(z) = \overline{\mathsf{z}_l^{(1)}(\bar{z})} = z\frac{\frac{d}{dz}h_l^{(2)}(z)}{h_l^{(2)}(z)}. \tag{4.5}$$

From the spherical harmonic expansion (cf. [8, Theorems 6.24 and 6.25]) of the radiating solution of (3.3), (2.6), (4.1), the definitions of ∇_S, Δ_S, and a simple calculation, we get

$$\begin{aligned} \mathcal{E} = &\sum_{l=1}^{\infty}\sum_{m=-l}^{l} \sqrt{\mu_0} h_l^{(1)}(kr) \\ &\times \left\{ u_l^m \nabla_S Y_l^m \wedge n + \frac{v_l^m}{\mathbf{i}kr}\left[\left(1+\mathsf{z}_l^{(1)}(kr)\right)\nabla_S Y_l^m + l(l+1)Y_l^m n\right]\right\}, \end{aligned} \tag{4.6}$$

$$\begin{aligned} \mathcal{H} = &\sum_{l=1}^{\infty}\sum_{m=-l}^{l} \sqrt{\varepsilon_0} h_l^{(1)}(kr) \\ &\times \left\{ -v_l^m \nabla_S Y_l^m \wedge n + \frac{u_l^m}{\mathbf{i}kr}\left[\left(1+\mathsf{z}_l^{(1)}(kr)\right)\nabla_S Y_l^m + l(l+1)Y_l^m n\right]\right\}. \end{aligned} \tag{4.7}$$

By using the definition of ∇_S, and noting that $n = \vec{e_r}$ and $\vec{e_\theta} \wedge \vec{e_r} = -\vec{e_\varphi}$, $\vec{e_\varphi} \wedge \vec{e_r} = \vec{e_\theta}$ (see (2.3)), we have on S,

$$\mathcal{E}_S = \sum_{l=1}^{\infty}\sum_{m=-l}^{l} \sqrt{\mu_0} h_l^{(1)}(kR) \left\{ u_l^m \nabla_S Y_l^m \wedge n + \frac{v_l^m}{\mathbf{i}kR}\left(1+\mathsf{z}_l^{(1)}(kR)\right)\nabla_S Y_l^m \right\}, \tag{4.8}$$

$$\mathcal{H}\wedge n = \sum_{l=1}^{\infty}\sum_{m=-l}^{l} \sqrt{\varepsilon_0} h_l^{(1)}(kR) \left\{ \frac{u_l^m}{\mathbf{i}kR}\left(1+\mathsf{z}_l^{(1)}(kR)\right)\nabla_S Y_l^m \wedge n + v_l^m \nabla_S Y_l^m \right\}. \tag{4.9}$$

Therefore, from definition (3.2) of $\mathcal{T}_S$ and definitions (4.2) of U_l^m, V_l^m, we obtain an explicit representation for the map $\mathcal{T}_S$: for any

$$u = \mathcal{E}_S = \sum_{l=1}^{\infty}\sum_{m=-l}^{l} \left[c_l^m U_l^m + d_l^m V_l^m\right],$$

$$\mathcal{T}_S u = \mathcal{H}\wedge n = \sum_{l=1}^{\infty}\sum_{m=-l}^{l} \left[\frac{c_l^m}{\mathbf{i}\omega\mu_0 R}\left(1+\mathsf{z}_l^{(1)}(kR)\right)U_l^m + \frac{\mathbf{i}\omega\varepsilon_0 R d_l^m}{1+\mathsf{z}_l^{(1)}(kR)} V_l^m\right]. \tag{4.10}$$

Next we derive an explicit representation of $\mathcal{T}_S^{\mathrm{PML}}$. Introduce a field $(\mathcal{E}^{sc}, \mathcal{H}^{sc})$:

$$\mathcal{E}^{sc}_{\hat r} = \frac{1}{s(\hat r)}\mathcal{E}^{\rm PML}_{\hat r}, \quad \mathcal{E}^{sc}_{\theta} = \frac{\hat r}{r}\mathcal{E}^{\rm PML}_{\theta}, \quad \mathcal{E}^{sc}_{\varphi} = \frac{\hat r}{r}\mathcal{E}^{\rm PML}_{\varphi},$$
$$\mathcal{H}^{sc}_{\hat r} = \frac{1}{s(\hat r)}\mathcal{H}^{\rm PML}_{\hat r}, \quad \mathcal{H}^{sc}_{\theta} = \frac{\hat r}{r}\mathcal{H}^{\rm PML}_{\theta}, \quad \mathcal{H}^{sc}_{\varphi} = \frac{\hat r}{r}\mathcal{H}^{\rm PML}_{\varphi},$$

where $(\mathcal{E}^{\rm PML}, \mathcal{H}^{\rm PML})$ is defined in (3.12). From the boundary condition in (3.12), it is obvious that

$$\mathcal{E}^{sc} \wedge n = 0 \quad \text{on } \widehat{S}. \tag{4.11}$$

It is easily verified, from the derivation of the PML formulation in section 2, that the field $(\mathcal{E}^{sc}, \mathcal{H}^{sc})$ satisfies Maxwell's equations (2.7), that is, the first two equations of (3.3). Note that the expressions of (4.6) and (4.7) form a class of solutions for (2.7), and if we replace the first kind of spherical Hankel functions in (4.6) and (4.7) with the second kind of spherical Hankel functions, then we get another class of solutions for Maxwell's equations (2.7). Let

$$\widetilde{R} = \int_0^{\widehat{R}} s(\tau)\, d\tau. \tag{4.12}$$

By choosing properly a linear combination of the two classes of solutions, we get the following solution satisfying the boundary condition (4.11):

$$\begin{aligned}
\mathcal{E}^{sc} = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}\sqrt{\mu_0}\Bigg\{ & u_l^m\left(\frac{h_l^{(1)}(kr)}{h_l^{(1)}(k\widetilde{R})} - \frac{h_l^{(2)}(kr)}{h_l^{(2)}(k\widetilde{R})}\right)\nabla_S Y_l^m \wedge n \\
&+ \frac{v_l^m}{\mathbf{i}kr}\Bigg[\left(\frac{h_l^{(1)}(kr)\big(1+\mathsf{z}_l^{(1)}(kr)\big)}{h_l^{(1)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(1)}(k\widetilde{R})\big)} - \frac{h_l^{(2)}(kr)\big(1+\mathsf{z}_l^{(2)}(kr)\big)}{h_l^{(2)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(2)}(k\widetilde{R})\big)}\right)\nabla_S Y_l^m \\
&+ \left(\frac{l(l+1)h_l^{(1)}(kr)}{h_l^{(1)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(1)}(k\widetilde{R})\big)} - \frac{l(l+1)h_l^{(2)}(kr)}{h_l^{(2)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(2)}(k\widetilde{R})\big)}\right)Y_l^m n\Bigg]\Bigg\},
\end{aligned}$$

$$\begin{aligned}
\mathcal{H}^{sc} = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}\sqrt{\varepsilon_0} & \\
\times\Bigg\{ & -v_l^m\left(\frac{h_l^{(1)}(kr)}{h_l^{(1)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(1)}(k\widetilde{R})\big)} - \frac{h_l^{(2)}(kr)}{h_l^{(2)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(2)}(k\widetilde{R})\big)}\right)\nabla_S Y_l^m \wedge n \\
&+ \frac{u_l^m}{\mathbf{i}kr}\Bigg[\left(\frac{h_l^{(1)}(kr)\big(1+\mathsf{z}_l^{(1)}(kr)\big)}{h_l^{(1)}(k\widetilde{R})} - \frac{h_l^{(2)}(kr)\big(1+\mathsf{z}_l^{(2)}(kr)\big)}{h_l^{(2)}(k\widetilde{R})}\right)\nabla_S Y_l^m \\
&+ \left(\frac{l(l+1)h_l^{(1)}(kr)}{h_l^{(1)}(k\widetilde{R})} - \frac{l(l+1)h_l^{(2)}(kr)}{h_l^{(2)}(k\widetilde{R})}\right)Y_l^m n\Bigg]\Bigg\}.
\end{aligned}$$

Similar to (4.8) and (4.9), since $\mathcal{E}^{\rm PML} = \mathcal{E}^{sc}$ and $\mathcal{H}^{\rm PML} = \mathcal{H}^{sc}$ on S, we have on S

$$
\begin{aligned}
\mathcal{E}_S^{\mathrm{PML}} = \mathcal{E}_S^{sc} = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}\sqrt{\mu_0}\Bigg\{ & u_l^m\left(\frac{h_l^{(1)}(kR)}{h_l^{(1)}(k\widetilde{R})}-\frac{h_l^{(2)}(kR)}{h_l^{(2)}(k\widetilde{R})}\right)\nabla_S Y_l^m\wedge n\\
&+\frac{v_l^m}{\mathbf{i}kR}\left(\frac{h_l^{(1)}(kR)\big(1+\mathsf{z}_l^{(1)}(kR)\big)}{h_l^{(1)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(1)}(k\widetilde{R})\big)}-\frac{h_l^{(2)}(kR)\big(1+\mathsf{z}_l^{(2)}(kR)\big)}{h_l^{(2)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(2)}(k\widetilde{R})\big)}\right)\nabla_S Y_l^m\Bigg\},\\
\mathcal{H}^{\mathrm{PML}}\wedge n = \mathcal{H}^{sc}\wedge n = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}\sqrt{\varepsilon_0}&\\
\times\Bigg\{&\frac{u_l^m}{\mathbf{i}kR}\left(\frac{h_l^{(1)}(kR)\big(1+\mathsf{z}_l^{(1)}(kR)\big)}{h_l^{(1)}(k\widetilde{R})}-\frac{h_l^{(2)}(kR)\big(1+\mathsf{z}_l^{(2)}(kR)\big)}{h_l^{(2)}(k\widetilde{R})}\right)\nabla_S Y_l^m\wedge n\\
&+v_l^m\left(\frac{h_l^{(1)}(kR)}{h_l^{(1)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(1)}(k\widetilde{R})\big)}-\frac{h_l^{(2)}(kR)}{h_l^{(2)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(2)}(k\widetilde{R})\big)}\right)\nabla_S Y_l^m\Bigg\}.
\end{aligned}
$$

Then from definition (3.11) of $\mathcal{T}_S^{\mathrm{PML}}$ and definitions (4.2) of U_l^m, V_l^m, we obtain an explicit representation for the map $\mathcal{T}_S^{\mathrm{PML}}$: for any

$$
u=\sum_{l=1}^{\infty}\sum_{m=-l}^{l}\big[c_l^m U_l^m + d_l^m V_l^m\big],
$$

$$
\begin{aligned}
\mathcal{T}_S^{\mathrm{PML}}u = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}\Bigg[&\frac{c_l^m}{\mathbf{i}\omega\mu_0 R}\frac{h_l^{(1)}(kR)h_l^{(2)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(1)}(kR)\big)-h_l^{(1)}(k\widetilde{R})h_l^{(2)}(kR)\big(1+\mathsf{z}_l^{(2)}(kR)\big)}{h_l^{(1)}(kR)h_l^{(2)}(k\widetilde{R})-h_l^{(1)}(k\widetilde{R})h_l^{(2)}(kR)}U_l^m\\
&+\mathbf{i}\omega\varepsilon_0 R d_l^m\\
&\times\frac{h_l^{(1)}(kR)h_l^{(2)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(2)}(k\widetilde{R})\big)-h_l^{(2)}(kR)h_l^{(1)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(1)}(k\widetilde{R})\big)}{h_l^{(1)}(kR)\big(1+\mathsf{z}_l^{(1)}(kR)\big)h_l^{(2)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(2)}(k\widetilde{R})\big)-h_l^{(1)}(k\widetilde{R})\big(1+\mathsf{z}_l^{(1)}(k\widetilde{R})\big)h_l^{(2)}(kR)\big(1+\mathsf{z}_l^{(2)}(kR)\big)}V_l^m\Bigg].
\end{aligned}
$$

By using the representations of $\mathcal{T}_S$ (4.10) and the above $\mathcal{T}_S^{\mathrm{PML}}$, we have

$$
\big(\mathcal{T}_S^{\mathrm{PML}}-\mathcal{T}_S\big)u=\sum_{l=1}^{\infty}\sum_{m=-l}^{l}\big(\rho_l c_l^m U_l^m+\sigma_l d_l^m V_l^m\big) \tag{4.13}
$$

with

$$
\rho_l=\frac{(\mathbf{i}\omega\mu_0 R)^{-1}\big[\mathsf{z}_l^{(1)}(kR)-\mathsf{z}_l^{(2)}(kR)\big]\cdot h_l^{(2)}(kR)\big[h_l^{(1)}(kR)\big]^{-1}}{h_l^{(2)}(k\widetilde{R})\big[h_l^{(1)}(k\widetilde{R})\big]^{-1}-h_l^{(2)}(kR)\big[h_l^{(1)}(kR)\big]^{-1}}, \tag{4.14}
$$

$$
\sigma_l=\frac{\mathbf{i}\omega\varepsilon_0 R\big[\mathsf{z}_l^{(2)}(kR)-\mathsf{z}_l^{(1)}(kR)\big]\big[1+\mathsf{z}_l^{(1)}(kR)\big]^{-2}h_l^{(2)}(kR)\big[h_l^{(1)}(kR)\big]^{-1}}{\frac{h_l^{(2)}(k\widetilde{R})\left(1+\mathsf{z}_l^{(2)}(k\widetilde{R})\right)}{h_l^{(1)}(k\widetilde{R})\left(1+\mathsf{z}_l^{(1)}(k\widetilde{R})\right)}-\frac{h_l^{(2)}(kR)\left(1+\mathsf{z}_l^{(2)}(kR)\right)}{h_l^{(1)}(kR)\left(1+\mathsf{z}_l^{(1)}(kR)\right)}}. \tag{4.15}
$$

Therefore, in order to estimate the error between $\mathcal{T}_S$ and $\mathcal{T}_S^{\mathrm{PML}}$, it is essential to derive upper bounds for ρ_l and σ_l. To do so, we need asymptotic results for the spherical Hankel functions.

Remark 4.1. In practice, it is important to have explicit error estimates between $\mathcal{T}_S$ and $\mathcal{T}_S^{\mathrm{PML}}$. For this reason, we choose not to use the usual uniform Airy-type asymptotic expansions of the spherical Hankel functions $h_l^{(1)}(z)$ and $h_l^{(2)}(z)$, which are efficient for $l \approx z$, because the Airy functions are implicit. Instead, we employ the explicit exponential-type expansions of the modified Bessel functions and derive explicit approximations for the spherical Hankel functions. It should be pointed out that the spherical Hankel functions and the modified Bessel functions are related by the connection formula given in section 5.2.

Remark 4.2. We comment on the validity of the exponential-type expansions of the spherical Hankel functions. Unlike the Airy-type expansions, the exponential-type expansions may be invalid for $l \leq z$ with real z. In this case, some special treatment must be used. A detailed discussion is given in subsequent sections.

5. Estimates on the spherical Hankel functions. Our goal in this section is to derive explicit estimates for the spherical Hankel functions. We first derive and estimate the asymptotic expansions of the modified Bessel functions in section 5.1. The connection formula between the spherical Hankel functions and the modified Bessel functions in section 5.2 can then be employed to establish the desirable estimates.

5.1. First approximations of the modified Bessel functions. Consider the asymptotic behavior of the modified Bessel functions $I_\nu(\nu z)$ and $K_\nu(\nu z)$ of large order ν. The order ν is always assumed to be real and positive throughout.

Introduce new variables

$$\xi = (1+z^2)^{1/2} + \ln \frac{z}{1+(1+z^2)^{1/2}}, \tag{5.1}$$

$$p = (1+z^2)^{-1/2}. \tag{5.2}$$

Define

$$U_1 = (3p - 5p^3)/24. \tag{5.3}$$

The following lemma was proved in sections 10.7–10.8 of [15].

LEMMA 5.1. *Denote by $b_1 = 0$ and $b_2 = +\infty$. Assume that z satisfies $|\arg z| < \pi/2$ or $|\arg z| = \pi/2$ but $|z| < 1$. Then*

$$I_\nu(\nu z) = \left(\frac{1}{2\pi\nu}\right)^{1/2} \frac{e^{\nu\xi}}{(1+z^2)^{1/4}} \frac{1+\eta_1(\nu,z)}{1+\eta_1(\nu,\infty)}, \tag{5.4}$$

$$K_\nu(\nu z) = \left(\frac{\pi}{2\nu}\right)^{1/2} \frac{e^{-\nu\xi}}{(1+z^2)^{1/4}} \big(1+\eta_2(\nu,z)\big), \tag{5.5}$$

$$I_\nu'(\nu z) = \left(\frac{1}{2\pi\nu}\right)^{1/2} \frac{e^{\nu\xi}(1+z^2)^{1/4}}{z} \frac{1+\eta_3(\nu,z)}{1+\eta_1(\nu,\infty)}, \tag{5.6}$$

$$K_\nu'(\nu z) = -\left(\frac{\pi}{2\nu}\right)^{1/2} \frac{e^{-\nu\xi}(1+z^2)^{1/4}}{z} (1+\eta_4(\nu,z)), \tag{5.7}$$

where

$$\eta_3(\nu,z) = \kappa_1(\nu,z) - \frac{z^2p^3}{2\nu}(1+\eta_1(\nu,z)), \tag{5.8}$$

$$\eta_4(\nu,z) = -\kappa_2(\nu,z) + \frac{z^2p^3}{2\nu}(1+\eta_2(\nu,z)). \tag{5.9}$$

Moreover, the error terms $\eta_j, \kappa_j, j = 1, 2$, are bounded by

$$|\eta_1(\nu, z)|, |\kappa_1(\nu, z)| \le \exp\left(\frac{2\,\mathcal{V}_{b_1,z}(U_1)}{\nu}\right)\frac{2\,\mathcal{V}_{b_1,z}(U_1)}{\nu}, \tag{5.10}$$

$$|\eta_2(\nu, z)|, |\kappa_2(\nu, z)| \le \exp\left(\frac{2\,\mathcal{V}_{b_2,z}(U_1)}{\nu}\right)\frac{2\,\mathcal{V}_{b_2,z}(U_1)}{\nu}, \tag{5.11}$$

where $\mathcal{V}_{b_j,z}(U_1)$ is the bounded variation from b_j to z, and the variational paths being taken by traveling in the ξ plane from $\xi(z)$ parallel to the imaginary axis until the real axis is reached, then proceeding along the real axis to $\xi(b_1)$ or $\xi(b_2)$.

Let

$$D_1 = \{z;\ |\arg z| \le \pi/4\}, \qquad D_2 = \{z;\ |\arg z| = \pi/2 \text{ and } |z| < 1\}. \tag{5.12}$$

Note that estimates (5.10) and (5.11) are not yet explicit in the sense that the right-hand sides remain to be estimated. This problem is resolved by our next lemma whose proof is given in the appendix.

LEMMA 5.2. *For $m = 1, 2, z \in D_m$,*

$$|\eta_1(\nu, z)|, |\eta_2(\nu, z)| \le \widehat{M}_m(\nu, z) := \exp\left(\frac{2M_m(z)}{\nu}\right)\frac{2M_m(z)}{\nu},$$
$$|\eta_1(\nu, \infty)| \le \widehat{M}_1(\nu, +\infty),$$
$$|\eta_3(\nu, z)|, |\eta_4(\nu, z)| \le \widehat{N}_m(\nu, z) := \frac{N_m(z)}{2\nu} + \left(1 + \frac{N_m(z)}{2\nu}\right)\widehat{M}_m(\nu, z),$$

where

$$M_1(z) = \frac{1}{12} + \frac{1}{6\sqrt{5}} + \frac{|\Im(z)|}{\Re(z)}\min\left\{\left(\frac{4}{27}\right)^{1/4}, \frac{1}{\Re(z)}\right\}, \tag{5.13}$$

$$N_1(z) = \min\left\{\left(\frac{4}{27}\right)^{1/4}, \frac{1}{|z|}\right\}, \tag{5.14}$$

$$M_2(z) = \frac{1}{12} + \frac{1}{6\sqrt{5}} + \frac{\pi\,|z|^2\,(4 + |z|^2)}{16(1 - |z|^2)^3}, \tag{5.15}$$

$$N_2(z) = \frac{|z|^2}{(1 - |z|^2)^{3/2}}. \tag{5.16}$$

Equipped with the estimates of the modified Bessel functions I_ν and K_ν, we are now ready to establish estimates of the spherical Hankel functions. The results are useful for the estimates of ρ_l and σ_l that are essential for the proof of Lemma 3.1.

5.2. Spherical Hankel functions. The spherical Hankel functions $h_l^{(j)}(z)$ ($j = 1, 2, l \ge 0$ integers) can also be defined by the Hankel functions of half-odd-integer order $H_{l+\frac{1}{2}}^{(j)}(z)$ (cf. (4.4) and [15, p. 238]):

$$h_l^{(1)}(z) = \mathbf{i}\sqrt{\frac{\pi}{2z}}H_{l+\frac{1}{2}}^{(1)}(z), \quad h_l^{(2)}(z) = -\mathbf{i}\sqrt{\frac{\pi}{2z}}H_{l+\frac{1}{2}}^{(2)}(z). \tag{5.17}$$

The Hankel functions and the modified Bessel functions satisfy the following connection formulas (see, e.g., [15, pp. 250–251]):

$$K_\nu(z) = \frac{\pi\mathbf{i}}{2}e^{\nu\pi\mathbf{i}/2}H_\nu^{(1)}\big(ze^{\pi\mathbf{i}/2}\big), \quad I_\nu(z) = \frac{1}{2}e^{-\nu\pi\mathbf{i}/2}\big[H_\nu^{(1)}\big(ze^{\pi\mathbf{i}/2}\big) + H_\nu^{(2)}\big(ze^{\pi\mathbf{i}/2}\big)\big],$$

hence

$$h_l^{(1)}(z) = \sqrt{\frac{2}{\pi z}}e^{-(l+\frac{1}{2})\pi\mathbf{i}/2}K_{l+\frac{1}{2}}\big(ze^{-\pi\mathbf{i}/2}\big), \tag{5.18}$$

$$h_l^{(2)}(z) = -\mathbf{i}\sqrt{\frac{2\pi}{z}}e^{(l+\frac{1}{2})\pi\mathbf{i}/2}I_{l+\frac{1}{2}}\big(ze^{-\pi\mathbf{i}/2}\big) + h_l^{(1)}(z). \tag{5.19}$$

It follows from the definition of $\mathsf{z}_l^{(j)}(z)$ that

$$\begin{aligned} h_l^{(1)}(z)\big[1+\mathsf{z}_l^{(1)}(z)\big] &= h_l^{(1)}(z) + zh_l^{(1)'}(z) \\ &= \sqrt{\frac{2}{\pi z}}e^{-(l+\frac{1}{2})\pi\mathbf{i}/2}\bigg(\frac{1}{2}K_{l+\frac{1}{2}}\big(ze^{-\pi\mathbf{i}/2}\big) - \mathbf{i}zK'_{l+\frac{1}{2}}\big(ze^{-\pi\mathbf{i}/2}\big)\bigg), \end{aligned} \tag{5.20}$$

$$\begin{aligned} h_l^{(2)}(z)\big[1+\mathsf{z}_l^{(2)}(z)\big] &= h_l^{(2)}(z) + zh_l^{(2)'}(z) \\ &= h_l^{(1)}(z)\big(1+\mathsf{z}_l^{(1)}(z)\big) - \mathbf{i}\sqrt{\frac{2\pi}{z}}e^{(l+\frac{1}{2})\pi\mathbf{i}/2}\bigg(\frac{1}{2}I_{l+\frac{1}{2}}\big(ze^{-\pi\mathbf{i}/2}\big) - \mathbf{i}zI'_{l+\frac{1}{2}}\big(ze^{-\pi\mathbf{i}/2}\big)\bigg). \end{aligned} \tag{5.21}$$

The following three lemmas present estimates that are crucial for estimating ρ_l and σ_l in (4.14) and (4.15), respectively.

Lemma 5.3. *Denote $\nu = l + 1/2$ and $\hat{z} = ze^{-\pi\mathbf{i}/2}$. For $m = 1, 2$, $\hat{z} \in D_m$, and $\nu \geq 3/2$, the following estimates hold:*

$$e^{2\nu\Re(\xi(\hat{z}))}\cdot C_m^-(\nu,\hat{z}) \leq \left|\frac{h_l^{(2)}(\nu z)}{h_l^{(1)}(\nu z)} - 1\right| \leq e^{2\nu\Re(\xi(\hat{z}))}\cdot C_m^+(\nu,\hat{z}), \tag{5.22}$$

$$e^{2\nu\Re(\xi(\hat{z}))}\cdot \widehat{C}_m^-(\nu,\hat{z}) \leq \left|\frac{h_l^{(2)}(\nu z)\big[1+\mathsf{z}_l^{(2)}(\nu z)\big]}{h_l^{(1)}(\nu z)\big[1+\mathsf{z}_l^{(1)}(\nu z)\big]} - 1\right| \leq e^{2\nu\Re(\xi(\hat{z}))}\cdot \widehat{C}_m^+(\nu,\hat{z}) \tag{5.23}$$

if $C_m^-(\nu,\hat{z})$ and $\widehat{C}_m^-(\nu,\hat{z})$ are positive. Here

$$C_m^{\pm}(\nu,\hat{z}) = \frac{1}{1\mp\widehat{M}_1(\nu,+\infty)}\cdot\frac{1\pm\widehat{M}_m(\nu,\hat{z})}{1\mp\widehat{M}_m(\nu,\hat{z})}, \tag{5.24}$$

$$\widehat{C}_m^{\pm}(\nu,\hat{z}) = \frac{1}{1\mp\widehat{M}_1(\nu,+\infty)}\cdot\frac{1\pm\Big\{\widehat{N}_m(\nu,\hat{z}) + \frac{|(1+\hat{z}^2)^{-1/2}|}{2\nu}\big[1+\widehat{M}_m(\nu,\hat{z})\big]\Big\}}{1\mp\Big\{\widehat{N}_m(\nu,\hat{z}) + \frac{|(1+\hat{z}^2)^{-1/2}|}{2\nu}\big[1+\widehat{M}_m(\nu,\hat{z})\big]\Big\}}. \tag{5.25}$$

Here ξ is defined by (5.1)*, D_m is defined by* (5.12)*, and $\widehat{M}_m, \widehat{N}_m$ are defined in Lemma* 5.2.

Proof. From (5.18) to (5.21) and (5.4) to (5.7), we obtain, after some simple calculations, that

$$\frac{h_l^{(2)}(\nu z)}{h_l^{(1)}(\nu z)} = 1 + \frac{(-1)^l e^{2\nu\xi(\hat{z})}}{1+\eta_1(\nu,\infty)}\cdot\frac{1+\eta_1(\nu,\hat{z})}{1+\eta_2(\nu,\hat{z})},$$

$$\frac{h_l^{(2)}(\nu z)\big[1+\mathsf{z}_l^{(2)}(\nu z)\big]}{h_l^{(1)}(\nu z)\big[1+\mathsf{z}_l^{(1)}(\nu z)\big]} = 1 + \frac{(-1)^l e^{2\nu\xi(\hat{z})}}{1+\eta_1(\nu,\infty)}\cdot\frac{\frac{(1+\hat{z}^2)^{-1/2}}{2\nu}\left[1+\eta_1(\nu,\hat{z})\right] + \left[1+\eta_3(\nu,\hat{z})\right]}{\frac{(1+\hat{z}^2)^{-1/2}}{2\nu}\left[1+\eta_2(\nu,\hat{z})\right] - \left[1+\eta_4(\nu,\hat{z})\right]}.$$

Estimates (5.22)–(5.23) then follow directly from Lemma 5.2 by noting that $\widehat{M}_1(\nu, +\infty) < 1$ for $v \geq 3/2$. □

Once again, for the representations ρ_l and σ_l of (4.14) and (4.15), an important term is $\mathsf{z}_l^{(1)} - \mathsf{z}_l^{(2)}$. The following result provides an estimate of the term.

LEMMA 5.4. *Denote $\nu = l + \frac{1}{2}$ and $\hat{z} = z e^{-\pi \mathbf{i}/2}$. Assume that z is real and $0 < z < 1$ and that $\widehat{M}_2(\nu, \hat{z}) < 1$. Then*

$$\left|\mathsf{z}_l^{(1)}(\nu z) - \mathsf{z}_l^{(2)}(\nu z)\right| \leq 2\nu e^{2\nu \Re(\xi(\hat{z}))}\left|(1+\hat{z}^2)^{1/2}\right|\left[1 - \widehat{M}_2(\nu, \hat{z})\right]^{-2}. \tag{5.26}$$

Proof. We need the following expression of $\mathsf{z}_l^{(1)}(z)$ for $z > 0$ [14, Theorem 2.6.1]:

$$\mathsf{z}_l^{(1)}(z) = -\frac{p_l}{q_l} + \mathbf{i}\frac{z}{q_l}, \tag{5.27}$$

where

$$\begin{aligned} q_l &= 1 + \alpha_1^l \frac{1}{z^2} + \cdots + \alpha_l^l \frac{1}{z^{2l}} = z^2 \left|h_l^{(1)}(z)\right|^2, \\ p_l &= 1 + 2\alpha_1^l \frac{1}{z^2} + \cdots + (l+1)\alpha_l^l \frac{1}{z^{2l}}, \quad \alpha_m^l = \frac{(m+l)!(2m)!}{m!^2(l-m)!4^m}. \end{aligned} \tag{5.28}$$

Hence,

$$\left|\mathsf{z}_l^{(1)}(z) - \mathsf{z}_l^{(2)}(z)\right| = \left|\mathsf{z}_l^{(1)}(z) - \overline{\mathsf{z}_l^{(1)}(z)}\right| = 2z/q_l = 2z^{-1}\left|h_l^{(1)}(z)\right|^{-2}. \tag{5.29}$$

Therefore, by (5.18), (5.5), and (5.29), we have

$$\left|\mathsf{z}_l^{(1)}(\nu z) - \mathsf{z}_l^{(2)}(\nu z)\right| = 2\nu\left|e^{2\nu\xi(\hat{z})}(1+\hat{z}^2)^{1/2}\right|\left|1 + \eta_2(\nu, \hat{z})\right|^{-2},$$

which implies (5.26). □

The result below presents additional estimates of the terms in the representations ρ_l and σ_l of (4.14) and (4.15).

LEMMA 5.5. *If z is real and positive, then*

$$\left|\frac{h_l^{(1)}(z)}{h_l^{(2)}(z)}\right| = 1, \qquad \left|\frac{h_l^{(1)}(z)\left[1 + \mathsf{z}_l^{(1)}(z)\right]}{h_l^{(2)}(z)\left[1 + \mathsf{z}_l^{(2)}(z)\right]}\right| = 1, \tag{5.30}$$

$$\left|\mathsf{z}_l^{(1)}(z) - \mathsf{z}_l^{(2)}(z)\right| \leq 2z\,, \tag{5.31}$$

$$\left|1 + \mathsf{z}_l^{(j)}(z)\right| \geq \frac{\sqrt{1 + l(l+1)}}{3z + 3/2}, \quad j = 1, 2, \;\; l \geq 2. \tag{5.32}$$

Proof. From (4.4) and (4.5), we know that (5.30) holds. By using (5.29) and $q_l \geq 1$, we get (5.31). Hence, it remains to prove (5.32). From (4.5) and (5.27), we have

$$\left|1 + \mathsf{z}_l^{(2)}(z)\right| = \left|1 + \mathsf{z}_l^{(1)}(z)\right| \geq \max\left\{\frac{p_l}{q_l} - 1, \; \frac{z}{q_l}\right\}.$$

From expressions (5.28) of p_l and q_l, some simple calculations yield

$$\frac{p_l}{q_l} - 1 \geq \frac{l(l+1)}{2z^2 + l + 1} \quad \text{and} \quad \frac{z}{q_l} \geq z - \frac{l(l+1)}{z},$$

which implies that $\frac{p_l}{q_l} - 1 > \frac{l+1}{3z+3/2}$ if $z < l+1$ and $\frac{z}{q_l} \geq 1 \geq \frac{l+1}{z}$ if $z \geq l+1$. The proof is now complete. □

6. Proof of the key lemma. Lemma 3.1 may be proved by estimating ρ_l and σ_l. For simplicity, we introduce the following notation:

$$\nu = l + 1/2, \quad z_\nu = kR/\nu, \quad \tilde{z}_\nu = k\widetilde{R}/\nu, \quad \hat{z}_\nu = z_\nu e^{-\pi \mathbf{i}/2}, \quad \text{and} \quad \hat{\tilde{z}}_\nu = \tilde{z}_\nu e^{-\pi \mathbf{i}/2}. \tag{6.1}$$

It follows from (4.12), (3.16), and $s(\tau) = 1 + \mathbf{i} s_I(\tau)$ that

$$\widetilde{R} = \widehat{R} + \mathbf{i}\widetilde{R}_I \quad \text{and} \quad \hat{\tilde{z}}_\nu = k\widetilde{R}_I/\nu - \mathbf{i}k\widehat{R}/\nu. \tag{6.2}$$

Note that the estimates in Lemmas 5.3 and 5.4 are no longer valid for $z = z_\nu$ if $z_\nu \geq 1$, that is, if $\nu = l + 1/2 \leq kR$. We estimate ρ_l and σ_l in (4.14) and (4.15) by considering two separate cases: $\nu \geq \nu_a$ and $\nu < \nu_a$, where

$$\nu_a = kR/a. \tag{6.3}$$

Here, recall from (3.16) that $a = \min\{1/2,\, kR/5\}$ is a positive number less than 1.

The following lemma gives uniform lower bounds for the exponential terms that appear in Lemmas 5.3 and 5.4.

LEMMA 6.1. *The following estimate holds:*

$$e^{2\nu \Re(\xi(\hat{\tilde{z}}_\nu))} \geq e^{\xi(a\widetilde{R}_I/R)\cdot 2kR/a} \quad \textit{for } 0 < \nu < \nu_a. \tag{6.4}$$

If, in addition,

$$2\frac{kR}{a} \ln \frac{(1 + \sqrt{1 - a^2}) \cdot \widetilde{R}_I/R}{1 + \sqrt{1 + a^2 \widetilde{R}_I^2/R^2}} \geq 1, \tag{6.5}$$

then

$$\frac{1}{2\nu} e^{2\nu[\Re(\xi(\hat{\tilde{z}}_\nu)) - \Re(\xi(\hat{z}_\nu))]} \geq \frac{a}{2kR} e^{\xi(a\widetilde{R}_I/R)\cdot 2kR/a} \quad \textit{for } \nu \geq \nu_a. \tag{6.6}$$

Proof. First from Lemma A.1(ii) and (6.2), we have

$$\Re(\xi(\hat{\tilde{z}}_\nu)) \geq \xi(k\widetilde{R}_I/\nu).$$

Let

$$g_1(\nu) = \nu \xi(k\widetilde{R}_I/\nu) = \sqrt{\nu^2 + k^2 \widetilde{R}_I^2} + \nu \ln \frac{k\widetilde{R}_I}{\nu + \sqrt{\nu^2 + k^2 \widetilde{R}_I^2}}.$$

Then

$$g_1'(\nu) = \ln \frac{k\widetilde{R}_I}{\nu + \sqrt{\nu^2 + k^2 \widetilde{R}_I^2}} < 0 \quad \text{for } \nu > 0,$$

which implies that $g_1(\nu) \geq g_1(\nu_a)$ for $0 < \nu < \nu_a$, and hence (6.4) holds.

Now we turn to prove (6.6). For $\nu \geq \nu_a$, define $g(\nu) = \nu[\xi(k\widetilde{R}_I/\nu) - \Re(\xi(\hat{z}_\nu))]$. Since $|\hat{z}_\nu| \leq kR/\nu_a = a < 1$, it follows from the definition (5.1) of ξ that

$$g(\nu) = g_1(\nu) - \left(\sqrt{\nu^2 - k^2 R^2} + \nu \ln \frac{kR}{\nu + \sqrt{\nu^2 - k^2 R^2}} \right),$$

$$g'(\nu) = g_1'(\nu) - \ln \frac{kR}{\nu + \sqrt{\nu^2 - k^2 R^2}}, \quad g''(\nu) = -\frac{1}{\sqrt{\nu^2 + k^2 \widetilde{R}_I^2}} + \frac{1}{\sqrt{\nu^2 - k^2 R^2}}.$$

By $g''(\nu) \geq 0$ and (6.5), we have, for $\nu \geq \nu_a$,

$$g'(\nu) \geq g'(\nu_a) = \ln \frac{(1+\sqrt{1-a^2}) \cdot \widetilde{R}_I/R}{1+\sqrt{1+a^2\widetilde{R}_I^2/R^2}} > 0,$$

which implies

$$\frac{d\left(\frac{1}{2\nu}e^{2g(\nu)}\right)}{d\nu} = \frac{e^{2g(\nu)}}{2\nu^2}(2\nu g'(\nu)-1) \geq \frac{e^{2g(\nu)}}{2\nu^2}(2\nu_a g'(\nu_a)-1) \geq 0.$$

Hence $\frac{1}{2\nu}e^{2g(\nu)} \geq \frac{1}{2\nu_a}e^{2g(\nu_a)}$ for $\nu \geq \nu_a$. Inequality (6.6) follows from the definition of $g(\nu)$ and the fact $g(\nu_a) \geq g_1(\nu_a)$. □

In addition, the next lemma gives estimates for the constants in Lemmas 5.3–5.4.

Lemma 6.2. *Under assumption* (3.17),

$$C_1^-(\nu, \hat{\tilde{z}}_\nu) > 1/5, \quad \widehat{C}_1^-(\nu, \hat{\tilde{z}}_\nu) > 1/5 \quad \textit{for } \nu \geq 5/2 \tag{6.7}$$

and

$$\begin{aligned} &C_1^-(\nu, \hat{\tilde{z}}_\nu) > 17/40, \quad \widehat{C}_1^-(\nu, \hat{\tilde{z}}_\nu) > 17/40, \quad \widehat{M}_2(\nu, \hat{z}_\nu) < 17/50, \\ &C_2^+(\nu, \hat{z}_\nu) < 91/25, \quad \widehat{C}_2^+(\nu, \hat{z}_\nu) < 91/25 \quad \textit{for } \nu \geq \nu_a. \end{aligned} \tag{6.8}$$

Furthermore inequality (6.5) *in the statement of Lemma* 6.1 *holds.*

Proof. Obviously, assumption (3.17) implies that

$$a \leq 1/2, \quad \nu_a \geq 5, \quad k\widetilde{R}_I \geq 17, \quad \text{and} \quad |\Im(\hat{\tilde{z}}_\nu)|/\Re(\hat{\tilde{z}}_\nu) = \widehat{R}/\widetilde{R}_I \leq 1. \tag{6.9}$$

Then by $\hat{\tilde{z}}_\nu = k\widetilde{R}_I/\nu - \mathbf{i}k\widehat{R}/\nu$, $\hat{z}_\nu = -\mathbf{i}kR/\nu$ (cf. (6.1) and (6.2)), and (5.13)–(5.16), we get

$$\begin{aligned} &M_1(\hat{\tilde{z}}_\nu) \leq \frac{1}{12} + \frac{1}{6\sqrt{5}} + 1 \cdot \frac{\nu}{k\widetilde{R}_I} \leq \frac{1}{12} + \frac{1}{6\sqrt{5}} + \frac{\nu}{17}, \\ &N_1(\hat{\tilde{z}}_\nu) \leq \frac{\nu}{k\widetilde{R}_I} \leq \frac{\nu}{17} \quad \text{for } \nu \geq \frac{5}{2} \end{aligned}$$

and

$$\begin{aligned} &M_2(\hat{z}_\nu) \leq \frac{1}{12} + \frac{1}{6\sqrt{5}} + \frac{\pi\,|a|^2\,(4+|a|^2)}{16(1-|a|^2)^3} \leq \frac{1}{12} + \frac{1}{6\sqrt{5}} + \frac{17\pi}{108}, \\ &N_2(\hat{z}_\nu) \leq \frac{|a|^2}{(1-|a|^2)^{3/2}} \leq \frac{2\sqrt{3}}{9} \quad \text{for } \nu \geq \nu_a. \end{aligned}$$

In addition

$$M_1(+\infty) = \frac{1}{12} + \frac{1}{6\sqrt{5}}, \quad (1+\hat{\tilde{z}}_\nu^2)^{-1/2} \leq 1, \quad (1+\hat{z}_\nu^2)^{-1/2} = (1-z_\nu^2)^{-1/2}.$$

By combining the above estimates and definitions (5.24)–(5.25), (6.1), and some direct calculations, it is straightforward to complete the proof of (6.7)–(6.8).

From (3.16) and (3.17), it follows that $kR/a \geq 5, \widetilde{R}_I/R \geq 7/5, a \leq 1/2$. Hence,

$$\text{the left-hand side of (6.5)} \geq 2 \times 5 \ln \frac{\left(1+\sqrt{1-(1/2)^2}\right) \cdot 7/5}{1+\sqrt{1+(1/2)^2(7/5)^2}} > 1,$$

which completes the proof of Lemma 6.2. □

By combining the above estimates, we can now estimate ρ_l and σ_l in (4.14) and (4.15), respectively.

LEMMA 6.3. *Under assumption* (3.17), *the following estimates hold for* $l = 1, 2, \dots$*:*

$$\omega\,|\rho_l| \le \mathcal{M}\sqrt{1+l(l+1)} \quad \textit{and} \quad \omega\,|\sigma_l| \le \frac{\mathcal{M}}{\sqrt{1+l(l+1)}}, \tag{6.10}$$

where

$$\mathcal{M} = \frac{4k(a\mu_0)^{-1}\max\left\{(kR)^2(3kR+3/2)^2, 1\right\}}{e^{k\widetilde{R}_I\left[2-\left(a\widetilde{R}_I/R\right)^{-2}+\left(a\widetilde{R}_I/R\right)^{-4}/19\right]}-10},$$

and the constant $\widetilde{R}_I$ *is defined in* (3.16).

Proof. Due to the validity consideration of the exponential-type estimates for the spherical Hankel functions, we divide the proof into three cases: $\nu = l + 1/2 \ge \nu_a$, $5/2 \le \nu < \nu_a$, and $\nu = 3/2$. These cases are proved by using different approaches. For the first case, we employ the exponential-type estimates for estimating ρ_l and σ_l. For the second case, we combine the exponential-type estimates and the estimates in Lemma 5.5. The definitions of the spherical Hankel functions are used to treat the third case.

Case I. $\nu = l + 1/2 \ge \nu_a$.

By using (4.14)–(4.15) and Lemmas 5.3–5.5, we get

$$\begin{aligned}|\rho_l| &= \frac{(\omega\mu_0 R)^{-1}|\mathsf{z}_l^{(1)}(\nu z_\nu) - \mathsf{z}_l^{(2)}(\nu z_\nu)|}{|h_l^{(2)}(\nu\tilde{z}_\nu)\left[h_l^{(1)}(\nu\tilde{z}_\nu)\right]^{-1} - 1 - \left(h_l^{(2)}(\nu z_\nu)\left[h_l^{(1)}(\nu z_\nu)\right]^{-1} - 1\right)|}\\ &\le \frac{(\omega\mu_0 R)^{-1}|1 - \widehat{M}_2(\nu, \hat{z}_\nu)|^{-2}}{\frac{1}{2\nu}e^{2\nu[\Re(\xi(\hat{\tilde{z}}_\nu)) - \Re(\xi(\hat{z}_\nu))]}C_1^-(\nu, \hat{\tilde{z}}_\nu) - \frac{1}{2\nu}C_2^+(\nu, \hat{z}_\nu)}\end{aligned}$$

and

$$\begin{aligned}|\sigma_l| &= \frac{\omega\varepsilon_0 R\cdot|\mathsf{z}_l^{(2)}(\nu z_\nu) - \mathsf{z}_l^{(1)}(\nu z_\nu)||1+\mathsf{z}_l^{(1)}(kR)|^{-2}}{\left|\frac{h_l^{(2)}(\nu\tilde{z}_\nu)\left(1+\mathsf{z}_l^{(2)}(\nu\tilde{z}_\nu)\right)}{h_l^{(1)}(\nu\tilde{z}_\nu)\left(1+\mathsf{z}_l^{(1)}(\nu\tilde{z}_\nu)\right)} - 1 - \left(\frac{h_l^{(2)}(\nu z_\nu)\left(1+\mathsf{z}_l^{(2)}(\nu z_\nu)\right)}{h_l^{(1)}(\nu z_\nu)\left(1+\mathsf{z}_l^{(1)}(\nu z_\nu)\right)} - 1\right)\right|}\\ &\le \frac{\omega\varepsilon_0 R|1 - \widehat{M}_2(\nu, \hat{z}_\nu)|^{-2}[1+l(l+1)]^{-1}(3kR+3/2)^2}{\frac{1}{2\nu}e^{2\nu[\Re(\xi(\hat{\tilde{z}}_\nu)) - \Re(\xi(\hat{z}_\nu))]}\widehat{C}_1^-(\nu, \hat{\tilde{z}}_\nu) - \frac{1}{2\nu}\widehat{C}_2^+(\nu, \hat{z}_\nu)}\ .\end{aligned}$$

Then by using Lemmas 6.1 and 6.2, we obtain

$$|\rho_l| \le \frac{11k(a\omega\mu_0)^{-1}[1+l(l+1)]^{-1/2}[1+l(l+1)]^{1/2}}{e^{\xi(a\widetilde{R}_I/R)\cdot 2kR/a} - 10}, \tag{6.11}$$

$$|\sigma_l| \le \frac{11k(a\omega\mu_0)^{-1}[1+l(l+1)]^{-1/2}[1+l(l+1)]^{-1/2}(kR)^2(3kR+3/2)^2}{e^{\xi(a\widetilde{R}_I/R)\cdot 2kR/a} - 10}. \tag{6.12}$$

Since $\nu_a \ge 5$ (cf. (6.9)) and the integer $l \ge \nu_a - \frac{1}{2}$, we have $l \ge 5$ and hence $11[1 + l(l+1)]^{-1/2} < 4$. Hence, estimate (6.10) follows from the inequality

$$\xi(a\widetilde{R}_I/R)\cdot 2kR/a > k\widetilde{R}_I\left[2 - \left(a\widetilde{R}_I/R\right)^{-2} + \left(a\widetilde{R}_I/R\right)^{-4}/19\right], \tag{6.13}$$

which can be derived by Lemma A.1(v) and $a\widetilde{R}_I/R = \min\{\widetilde{R}_I/(2R), k\widetilde{R}_I/5\} \geq 7/10$.

Case II. $5/2 \leq \nu = l + 1/2 < \nu_a$.

By using (4.14)–(4.15) and Lemmas 5.3–5.5, we have

$$|\rho_l| \leq \frac{(\omega\mu_0 R)^{-1}\left|\mathsf{z}_l^{(1)}(kR) - \mathsf{z}_l^{(2)}(kR)\right|}{\left|h_l^{(2)}(\nu\tilde{z}_\nu)/h_l^{(1)}(\nu\tilde{z}_\nu)\right| - 1} \leq \frac{(\omega\mu_0 R)^{-1} 2kR}{e^{2\nu\Re(\xi(\hat{\tilde{z}}_\nu))}C_1^-(\nu, \hat{\tilde{z}}_\nu) - 2},$$

$$|\sigma_l| \leq \frac{\omega\varepsilon_0 R\left|\mathsf{z}_l^{(2)}(kR) - \mathsf{z}_l^{(1)}(kR)\right|\left|1 + \mathsf{z}_l^{(1)}(kR)\right|^{-2}}{\left|\frac{h_l^{(2)}(\nu\tilde{z}_\nu)\left(1+\mathsf{z}_l^{(2)}(\nu\tilde{z}_\nu)\right)}{h_l^{(1)}(\nu\tilde{z}_\nu)\left(1+\mathsf{z}_l^{(1)}(\nu\tilde{z}_\nu)\right)}\right| - 1} \leq \frac{\omega\varepsilon_0 R \cdot 2kR\frac{(3kR+3/2)^2}{1+l(l+1)}}{e^{2\nu\Re(\xi(\hat{\tilde{z}}_\nu))}\widehat{C}_1^-(\nu, \hat{\tilde{z}}_\nu) - 2}.$$

From Lemmas 6.1 and 6.2, we further obtain that

$$|\rho_l| \leq \frac{10k(\omega\mu_0)^{-1}[1 + l(l+1)]^{-1/2}[1 + l(l+1)]^{1/2}}{e^{\xi(a\widetilde{R}_I/R)\cdot 2kR/a} - 10}, \tag{6.14}$$

$$|\sigma_l| \leq \frac{10k(\omega\mu_0)^{-1}[1 + l(l+1)]^{-1/2}[1 + l(l+1)]^{-1/2}(kR)^2(3kR + 3/2)^2}{e^{\xi(a\widetilde{R}_I/R)\cdot 2kR/a} - 10}. \tag{6.15}$$

Now estimate (6.10) follows from $10[1 + l(l+1)]^{-1/2} < 4$ and inequality (6.13).

Case III. $\nu = l + \frac{1}{2} = \frac{3}{2}$.

From (4.4)–(4.5), we have

$$h_1^{(1)}(z) = -\frac{\mathbf{i}e^{\mathbf{i}z}}{z}\left(1 + \frac{\mathbf{i}}{z}\right), \quad h_1^{(2)}(z) = \frac{\mathbf{i}e^{-\mathbf{i}z}}{z}\left(1 - \frac{\mathbf{i}}{z}\right),$$

$$\mathsf{z}_1^{(1)}(z) = \frac{-(z^2+2) + \mathbf{i}z^3}{z^2+1}, \quad \mathsf{z}_1^{(2)}(z) = \frac{-(z^2+2) - \mathbf{i}z^3}{z^2+1}.$$

It follows from definitions (4.14) and (4.15) of ρ_l and σ_l that

$$|\rho_1| \leq \frac{(\omega\mu_0 R)^{-1}\left|2\mathbf{i}(kR)^3[(kR)^2+1]^{-1}\right|}{\left|e^{-2\mathbf{i}k\widetilde{R}}(\mathbf{i} - k\widetilde{R})(\mathbf{i} + k\widetilde{R})^{-1}\right| - 1} \leq \frac{2k(\omega\mu_0)^{-1}}{e^{2k\widetilde{R}_I}(k\widetilde{R}_I - 1)(k\widetilde{R}_I + 1)^{-1} - 1},$$

$$|\sigma_1| \leq \frac{\left|\mathbf{i}\omega\varepsilon_0 R \cdot 2\mathbf{i}(kR)^3[(kR)^2+1][(kR)^6+1]^{-1}\right|}{\left|e^{-2\mathbf{i}k\widetilde{R}}\frac{\mathbf{i}-k\widetilde{R}}{\mathbf{i}+k\widetilde{R}}\frac{1+\mathbf{i}(k\widetilde{R})^3}{1-\mathbf{i}(k\widetilde{R})^3}\right| - 1} \leq \frac{2k(\omega\mu_0)^{-1}\cdot 4/3}{e^{2k\widetilde{R}_I}\frac{k\widetilde{R}_I-1}{k\widetilde{R}_I+1}\frac{(k\widetilde{R}_I)^3-1}{(k\widetilde{R}_I)^3+1} - 1}.$$

Since $k\widetilde{R}_I \geq 17$ (cf. (6.9)), it is easily seen that (6.10) holds for $l = 1$. The proof of Lemma 6.3 is complete. □

Now we return to the proof of Lemma 3.1. Let

$$v_S = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}[c_l^m U_l^m + d_l^m V_l^m] \quad \text{and} \quad \overline{\phi_S} = \sum_{l=1}^{\infty}\sum_{m=-l}^{l}[\tilde{c}_l^m U_l^m + \tilde{d}_l^m V_l^m].$$

From (4.13), we have from the orthogonality of the basis functions that

$$\left|\omega\int_S(\mathcal{T}_S^{\mathrm{PML}} - \mathcal{T}_S)v_S\cdot\phi_S\right| = \left|\omega\sum_{l=1}^{\infty}\sum_{m=-l}^{l}[\rho_l c_l^m\overline{\tilde{c}_l^m} + \sigma_l d_l^m\overline{\tilde{d}_l^m}]\right|$$
$$\leq \mathcal{M}\sum_{l=1}^{\infty}\sum_{m=-l}^{l}\left[\sqrt{1 + l(l+1)}\,|c_l^m|\,|\tilde{c}_l^m| + \frac{1}{\sqrt{1 + l(l+1)}}\,|d_l^m|\,|\tilde{d}_l^m|\right].$$

The proof of Lemma 3.1 may be completed by the Cauchy–Schwarz inequality, Lemma 6.3, and the definition (4.3) of the norm of $TH^{-1/2}(\mathrm{curl}, S)$.

7. Conclusion.

Under the assumption that there is a unique solution of the original three-dimensional electromagnetic problem and some proper assumptions on the PML medium parameter, it is shown that the truncated PML problem attains a unique solution in $H(\mathrm{curl})$. An explicit error estimate between the solution of the scattering problem and that of the truncated PML problem in the computational domain is obtained. The error decays exponentially as the product of the wave number and the integrated absorption across the layer goes to infinity. The error estimate implies particularly that the PML solution converges exponentially to the scattering solution by increasing either the PML medium parameter or the PML layer thickness.

Appendix. Proof of Lemma 5.2.

A.1. Properties of $\boldsymbol{\xi}$.

Let

$$z = x + \mathbf{i}y = re^{\mathbf{i}\theta}, \qquad r > 0 \text{ and } |\theta| < \pi/2 \quad \text{or} \quad 0 < r < 1 \text{ and } |\theta| = \pi/2,$$

$$1/p = (1+z^2)^{1/2} = r_1 e^{\mathbf{i}\theta_1}, \quad |\theta_1| \le \pi/2,$$

where

$$r_1 = (1 + 2r^2\cos 2\theta + r^4)^{1/4}, \quad r_1^2 \cos 2\theta_1 = 1 + r^2 \cos 2\theta, \quad r_1^2 \sin 2\theta_1 = r^2 \sin 2\theta.$$

Denote

$$1 + (1+z^2)^{1/2} = r_2 e^{\mathbf{i}\theta_2}, \quad |\theta_2| \le \pi/2,$$

where

$$r_2 = (1 + 2r_1\cos\theta_1 + r_1^2)^{1/2}, \quad r_2\cos\theta_2 = 1 + r_1\cos\theta_1, \quad r_2 \sin\theta_2 = r_1\sin\theta_1.$$

Then, by the definition (5.1) of ξ, we have

$$\Re(\xi) = r_1\cos\theta_1 + \ln\frac{r}{r_2}, \quad \Im(\xi) = r_1\sin\theta_1 + \theta - \theta_2. \tag{A.1}$$

LEMMA A.1.
(i) $\Re(\xi)$ *is decreasing in* $|\theta|$.
(ii) $\Re(\xi)$ *is increasing in* $|y|$ *and hence* $\Re(\xi) \ge \xi(x)$.
(iii) $\Re(\xi)$ *is increasing in* r.
(iv) $|\Im(\xi)| \le r_1 \sin|\theta| + |\theta|$.
(v) $x - \frac{1}{2x} + \frac{1}{24x^3} > \xi(x) > x - \frac{1}{2x} + \frac{1}{38x^3}$ *for* $x \ge \frac{7}{10}$.

Proof. From

$$r_1\cos\theta_1 = \sqrt{(r_1^2 + 1 + r^2\cos 2\theta)/2} \tag{A.2}$$

and a direct calculation, we deduce that

$$\frac{\partial\Re(\xi)}{\partial\theta} = -\frac{r^2\cos\theta\sin\theta}{r_1\cos\theta_1}, \tag{A.3}$$

which yields (i).

Let $u =: r^2 \cos 2\theta = x^2 - y^2$. From (A.1) and (A.2), we have

$$\frac{\partial \Re(\xi)}{\partial y} = \frac{y f(r,u)}{2r^2 r_1 \cos\theta_1 r_2^2}, \text{ where } f(r,u) =: r^4 - r^2 r_2^2 + 2r_1 \cos\theta_1 r_2^2. \tag{A.4}$$

Noting that $r_1^4 \geq (1-r^2)^2$, we conclude that

$$\frac{\partial f}{\partial u} = \frac{r_2^2}{2r_1^3 \cos\theta_1}\left(2r_1\cos\theta_1 + r_1^2 + 1 - r^2\right) > 0.$$

Consequently, statement (ii) may be proved by observing

$$f(r,u) \begin{cases} \geq f(r,-r^2) = 2(1-r^2)(1+\sqrt{1-r^2})^2 > 0 & \text{if } r < 1, \\ > f(r,-r^2) = 0 & \text{if } r \geq 1. \end{cases}$$

By

$$\frac{\partial \Re(\xi)}{\partial y} = \frac{\partial \Re(\xi)}{\partial r}\frac{\partial r}{\partial y} + \frac{\partial \Re(\xi)}{\partial \theta}\frac{\partial \theta}{\partial y} = \frac{\partial \Re(\xi)}{\partial r}\frac{y}{r} + \frac{\partial \Re(\xi)}{\partial \theta}\frac{\cos\theta}{r},$$

(A.3), and (A.4), we get

$$\frac{\partial \Re(\xi)}{\partial r} = \frac{f(r,u)}{2rr_1\cos\theta_1 r_2^2} + \frac{r\cos^2\theta}{r_1\cos\theta_1} > 0,$$

which implies that (iii) holds.

By $r_1^2 \leq 1 + r^2$, we have

$$1 + r^2\cos 2\theta \leq (1+r^2)\cos 2\theta_1 \leq 1 + r^2 \cos 2\theta_1,$$

which is $\cos 2\theta \leq \cos 2\theta_1$, and hence $|\theta_1| \leq |\theta|$. Furthermore, by $r_1^2 \sin 2\theta_1 = r^2 \sin 2\theta$, we know that $\theta\theta_1 \geq 0$. Similarly, we can prove that $|\theta_2| \leq |\theta_1|$ and $\theta_2\theta_1 \geq 0$. Then (iv) follows from (A.1).

By $\frac{d\xi(x)}{dx} = \frac{\sqrt{1+x^2}}{x}$, we know that $\xi(x) - (x - \frac{1}{2x} + \frac{1}{24x^3})$ increases for $x \geq \frac{7}{10}$ and $\xi(x) - (x - \frac{1}{2x} + \frac{1}{38x^3})$ first increases and then decreases for $x \geq \frac{7}{10}$. The estimate (v) follows then from the facts that $\xi(x) - (x - \frac{1}{2x} + \frac{1}{24x^3})$ and $\xi(x) - (x - \frac{1}{2x} + \frac{1}{38x^3})$ both approach 0 as x approaches infinity, and $\xi(x) - (x - \frac{1}{2x} + \frac{1}{38x^3}) > 0$ at $x = \frac{7}{10}$. The proof of Lemma A.1 is complete. □

A.2. Variations of U_1. In the ξ plane, the variations $\mathcal{V}_{b_j,z}(U_1)$ in (5.10) and (5.11) can be written as $\mathcal{V}_{\xi(b_j),\xi}(U_1)$. We denote by $\mathcal{P}(\xi)$ the first segment of the variational paths, i.e., $\tilde{\xi} = \Re(\xi) + \mathbf{i}\Im(\xi)(1-t)$, $0 \leq t \leq 1$. Denote by $\tilde{\xi} = \xi(\tilde{z})$ $(\tilde{z} = \tilde{r}e^{\mathbf{i}\tilde{\theta}})$ a point on $\mathcal{P}(\xi)$, and let $\tilde{p} = p(\tilde{z}) = (1+\tilde{z}^2)^{-1/2}$ (cf. (5.1) and (5.2)). It is clear that

$$\begin{aligned} \mathcal{V}_{b_j,z}(U_1) = \mathcal{V}_{\xi(b_j),\xi}(U_1) &= \mathcal{V}_{\xi(b_j),\Re(\xi)}(U_1) + \mathcal{V}_{\Re(\xi),\xi}(U_1) \\ &\leq \int_0^1 |U_1'(\tilde{p})|\, d\tilde{p} + \int_0^1 |U_1'(\tilde{p})\tilde{p}'_{\tilde{\xi}}\Im(\xi)|\, dt, \end{aligned} \tag{A.5}$$

where $U_1'(\tilde{p}) = \frac{1-5\tilde{p}^2}{8}$ and $\tilde{p}'_{\tilde{\xi}} = \tilde{p}^4 - \tilde{p}^2$ (cf. (5.1)–(5.3)).

It is obvious that

$$\int_0^1 |U_1'(\tilde{p})| \, d\tilde{p} = \frac{1}{12} + \frac{1}{6\sqrt{5}} \,. \tag{A.6}$$

Lemma A.2. *Suppose* $|\arg z| = |\theta| \le \pi/4$. *Then*

$$\mathcal{V}_{b_j,z}(U_1) \le \frac{1}{12} + \frac{1}{6\sqrt{5}} + \frac{|\Im(z)|}{\Re(z)} \min\left\{ \left(\frac{4}{27}\right)^{1/4}, \frac{1}{\Re(z)} \right\}, \quad j = 1, 2.$$

Proof. First by Lemma A.1, we have

$$r\cos\theta \le \tilde{r} = |\tilde{z}| \le r \quad \text{and} \quad |\tilde{\theta}| \le |\theta| \le \pi/4 \ \text{ for } \ \tilde{\xi} \ \text{ on } \ \mathcal{P}(\xi).$$

Thus,

$$|U_1'(\tilde{p})| = \frac{1}{8}\left|\frac{(\tilde{z}^2-4)}{1+\tilde{z}^2}\right| = \frac{1}{8}\Big(\frac{16-8\tilde{r}^2\cos 2\tilde{\theta}+\tilde{r}^4}{1+2\tilde{r}^2\cos 2\tilde{\theta}+\tilde{r}^4}\Big)^{1/2} \le \frac{1}{8}\Big(\frac{16+\tilde{r}^4}{1+\tilde{r}^4}\Big)^{1/2} \le \frac{1}{2}$$

and

$$\begin{aligned} \left|\tilde{p}^3 - \tilde{p}\right| = \left|\tilde{z}^2\tilde{p}^3\right| &= \tilde{r}^2(1+2\tilde{r}^2\cos 2\tilde{\theta}+\tilde{r}^4)^{-3/4} \\ &\le \tilde{r}^2(1+\tilde{r}^4)^{-3/4} \le \min\left\{(4/27)^{1/4}, 1/\tilde{r}\right\}. \end{aligned} \tag{A.7}$$

We also have from the definition of $\tilde{p}$ that

$$|\tilde{p}| = (1+2\tilde{r}^2\cos 2\tilde{\theta}+\tilde{r}^4)^{-1/4} \le (1+2r^2\cos^2\theta\cos 2\theta + r^4\cos^4\theta)^{-1/4} \le 1.$$

Hence, form Lemma A.1(iv), we get

$$|\tilde{p}|\,|\Im\xi| \le \Big(\frac{1+2r^2\cos 2\theta + r^4}{1+2r^2\cos^2\theta\cos 2\theta + r^4\cos^4\theta}\Big)^{1/4} \sin|\theta| + |\theta| \le \tan|\theta| + |\theta| \le 2\tan|\theta|.$$

Finally, the proof is complete by combining (A.5) and (A.6),

$$|U_1'(\tilde{p})\tilde{p}_{\tilde{\xi}}'\Im(\xi)| = |U_1'(\tilde{p})|\left|\tilde{p}^3 - \tilde{p}\right||\tilde{p}|\,|\Im\xi|$$

along with the above estimates. □

Lemma A.3. *Suppose* $|\arg z| = \pi/2$ *and* $|z| < 1$. *Then*

$$\mathcal{V}_{b_j,z}(U_1) \le \frac{1}{12} + \frac{1}{6\sqrt{5}} + \frac{\pi|z|^2(4+|z|^2)}{16(1-|z|^2)^3}, \quad j = 1, 2.$$

Proof. By Lemma A.1(i) and (iii), we have, for $\tilde{\xi}$ on $\mathcal{P}(\xi)$, $\tilde{r} = |\tilde{z}| \le |z|$. Hence

$$\begin{aligned} |U_1'(\tilde{p})\tilde{p}_{\tilde{\xi}}'| = \left|\frac{\tilde{z}^2(\tilde{z}^2-4)}{8(1+\tilde{z}^2)^3}\right| &= \frac{\tilde{r}^2(16-8\tilde{r}^2\cos 2\tilde{\theta}+\tilde{r}^4)^{1/2}}{8(1+2\tilde{r}^2\cos 2\tilde{\theta}+\tilde{r}^4)^{3/2}} \\ &\le \frac{\tilde{r}^2(4+\tilde{r}^2)}{8(1-\tilde{r}^2)^3} \le \frac{|z|^2(4+|z|^2)}{8(1-|z|^2)^3}. \end{aligned}$$

Then the proof follows from (A.5), (A.6) and the fact that $\Im(\xi) = \pi/2$. □

A.3. The proof of Lemma 5.2. From (A.7), we have the following lemma.

LEMMA A.4.

$$\left|z^2p^3\right| \leq \begin{cases} \min\left\{\left(4/27\right)^{1/4}, 1/|z|\right\} & \textit{if } z \in D_1, \\ |z|^2/(1-|z|^2)^{3/2} & \textit{if } z \in D_2, \end{cases}$$

where D_1 *and* D_2 *are defined in* (5.12).

By applying Lemmas A.2–A.4 and (5.8)–(5.11), Lemma 5.2 may be proved easily.

REFERENCES

[1] H. Ammari and J. Nédélec, *Low-frequency electromagnetic scattering*, SIAM J. Math. Anal., 31 (2000), pp. 836–861.

[2] I. Babuška and A. Aziz, *Survey lectures on the mathematical foundations of the finite element method,* in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic Press, New York, 1972, pp. 1–359.

[3] J. P. Bérenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[4] J. P. Bérenger, *Three-dimensional perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 127 (1996), pp. 363–379.

[5] Z.-M. Chen and H.-J. Wu, *An adaptive finite element method with perfectly matched absorbing layers for the wave scattering by periodic structures*, SIAM J. Numer. Anal., 41 (2003), pp. 799–826.

[6] F. Collino and P. Monk, *Optimizing the perfectly matched layer*, Comput. Methods Appl. Mech. Engrg., 164 (1998), pp. 157–171.

[7] F. Collino and P. Monk, *The perfectly matched layer in curvilinear coordinates*, SIAM J. Sci. Comput., 19 (1998), pp. 2061–2090.

[8] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1992.

[9] T. Hohage, F. Schmidt, and L. Zschiedrich, *Solving time-harmonic scattering problems based on the pole condition* II: *Convergence of the PML method*, SIAM J. Math. Anal., 35 (2003), pp. 547–560.

[10] E. Kashdan and E. Turkel, *Numerical solution of the time-dependent Maxwell's equations in spherical coordinates*, in 19th Annual Review of Progress in Applied Computational Electromagnetics, 2003, pp. 184–188.

[11] A. Kirsch and P. Monk, *A finite element/spectral method for approximating the time-harmonic Maxwell system in* $\mathbb{R}^3$, SIAM J. Appl. Math., 55 (1995), pp. 1324–1344.

[12] M. Lassas and E. Somersalo, *On the existence and convergence of the solution of PML equations*, Computing, 60 (1998), pp. 229–241.

[13] M. Lassas and E. Somersalo, *Analysis of the PML equations in general convex geometry*, Proc. Roy. Soc. Edinburgh Sect. A, 131 (2001), pp. 1183–1207.

[14] J. Nédélec, *Acoustic and Electromagnetic Equations*, Springer-Verlag, New York, 2001.

[15] F. W. J. Olver, *Asymptotics and Special Functions*, Academic Press, New York, 1974.

[16] L. Paquet, *Problèms mixtes pour le système de Maxwell*, Ann. Fac. Sci. Toulouse Math., 4 (1982), pp. 103–141.

[17] F. Teixeira and W. Chew, *Systematic derivation of anisotropic PML absorbing media in cylindrical and spherical coordinates*, IEEE Microwave Guided Wave Lett., 7 (1997), pp. 371–373.

[18] S. V. Tsynkov and E. Turkel, *A Cartesian perfectly matched layer for the Helmholtz equation*, in Absorbing Boundaries and Layers, Domain Decomposition Methods. Applications to Large Scale Computations, Nova Science, New York, 2001.

[19] E. Turkel and A. Yefet, *Absorbing PML boundary layers for wave-like equations*, Appl. Numer. Math., 27 (1998), pp. 533–557.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2144–2171

EXPONENTIALLY CONVERGENT ALGORITHMS FOR THE OPERATOR EXPONENTIAL WITH APPLICATIONS TO INHOMOGENEOUS PROBLEMS IN BANACH SPACES*

I. P. GAVRILYUK† AND V. L. MAKAROV‡

Abstract. New exponentially convergent algorithms for the operator exponential generated by a strongly positive operator A in a Banach space X are proposed. These algorithms are based on representations by a Dunford–Cauchy integral along paths enveloping the spectrum of A combined with a proper quadrature involving a short sum of resolvents where the choice of the integration path dramatically affects desired features of the algorithms. A parabola and a hyperbola are analyzed as the integration paths, and scales of estimates of dependence on the smoothness of initial data, i.e., of the initial vector and of the inhomogeneous right-hand side, are obtained. One of the algorithms possesses an exponential convergence rate for the operator exponential e^{-At} for all $t \geq 0$ including the initial point. This allows one to construct an exponentially convergent algorithm for inhomogeneous initial value problems. The algorithm is parallelizable. It turns out that the resolvent must be modified in order to get numerically stable algorithms near the initial point. The efficiency of the proposed method is demonstrated by numerical examples.

Key words. inhomogeneous evolution equation, operator exponential, exponentially convergent algorithms, sinc methods

AMS subject classifications. 65J10, 65M70, 35K90, 35L90

DOI. 10.1137/040611045

1. Introduction.

We consider the problem

$$\frac{du(t)}{dt} + Au(t) = f(t), \quad u(0) = u_0, \tag{1.1}$$

where A is a strongly positive operator in a Banach space X, $u_0 \in X$ is a given vector, and $f(t)$ is a given and $u(t)$ is the unknown vector-valued function. A simple example of a partial differential equation covered by the abstract setting (1.1) is the classical inhomogeneous heat equation

$$\frac{\partial u(t,x)}{\partial t} - \frac{\partial^2 u(t,x)}{\partial x^2} = f(t,x)$$

with corresponding boundary and initial conditions, where the operator A is defined by

$$\begin{aligned} D(A) &= \{v \in H^2(0,1) : \ v(0) = 0, v(1) = 0\}, \\ Av &= -\frac{d^2 v}{dx^2} \qquad \forall v \in D(A). \end{aligned}$$

The homogeneous equation

$$\frac{dT(t)}{dt} + AT(t) = 0, \quad T(0) = I, \tag{1.2}$$

*Received by the editors July 5, 2004; accepted for publication (in revised form) May 13, 2005; published electronically December 30, 2005.
http://www.siam.org/journals/sinum/43-5/61104.html

†Berufsakademie Thüringen, Staatliche Studienakademie, Am Wartenberg 2, D-99817 Eisenach, Germany (ipg@ba-eisenach.de).

‡Institute of Mathematics, National Academy of Sciences of Ukraine, 3 Tereshchenkivs'ka Str., Kyiv-4, 01601, Ukraine (makarov@imath.kiev.ua).

where I is the identity operator and $T(t)$ is an operator valued function, defines the semigroup of bounded operators $T(t) = e^{-At}$ generated by A (also called the operator exponential or the solution operator of the homogeneous equation (1.1)). Given the solution operator, the initial vector u_0, and the right-hand side $f(t)$, the solution of the homogeneous initial value problem (1.1) can be represented by

$$u(t) = u_o(t) = T(t)u_0 = e^{-At}u_0, \tag{1.3}$$

and the solution of the inhomogeneous problem can be represented by

$$u(t) = e^{-At}u_0 + u_p(t) \tag{1.4}$$

with

$$u_p(t) = \int_0^t e^{-A(t-\xi)} f(\xi)d\xi. \tag{1.5}$$

We can see that an efficient approximation of the operator exponential is needed in order to get an efficient discretization of both (1.3) and (1.4). Further, having in mind a discretization of the second summand in (1.4) by a quadrature sum, we need an efficient approximation of the operator exponential for all $t \geq 0$ including the point $t = 0$.

A convenient representation of the operator exponential is the one provided by the improper Dunford–Cauchy integral

$$e^{-At} = \frac{1}{2\pi i}\int_{\Gamma_I} e^{-tz}(zI - A)^{-1}dz, \tag{1.6}$$

where Γ_I is an integration path enveloping the spectrum of A. After parametrizing Γ we get an improper integral of the type

$$e^{-At} = \frac{1}{2\pi i}\int_{\Gamma_I} e^{-tz}(zI - A)^{-1}dz = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \mathcal{F}(t,\xi)d\xi. \tag{1.7}$$

The last integral can be discretized by a quadrature rule (desirable exponentially convergent) involving a short sum of resolvents. Such an algorithm inherits a two-level parallelism with respect to both the computation of resolvents and the treatment of different time values.

Two efficient methods for solving linear homogeneous parabolic problems based on the improper Dunford–Cauchy integrals along a path enveloping the spectrum of A were recently proposed in [12,14,18,33,34] where the boundary of a sector containing the spectrum of A or a parabola was used as the integration path. The method from [33] possesses a polynomial convergence rate. The method from [12,18,23] uses sinc quadratures [1,36,37] and possesses an exponential convergence rate for $t > 0$ and a polynomial convergence rate for $t = 0$ depending on the smoothness of the initial vector u_0 from a Hilbert space. An exponential convergence rate for all $t \geq 0$ was proved in [13,42] under assumptions that the initial function u_0 belongs to the domain of $D(A^\sigma)$ for some $\sigma > 1$, where the preliminary computation of $A^\sigma u_0$ is needed. Note that all these algorithms cannot be directly applied to inhomogeneous problems due to the inefficiency of computation of the operator exponential at $t = 0$. In [20] a hyperbola was used as the integration path which allows one to get the uniform exponential convergence rate with respect to $t \geq 0$ without preliminary computation of $A^\sigma u_0$. An exponentially convergent algorithm for the case of an operator family $A(t)$ depending on the parameter t was proposed in [21]. This algorithm uses an exponentially convergent algorithm for the operator exponential generated by a constant operator.

In contrast to various other approximation methods with a polynomial convergence rate for the problem (1.1) using finite differences [4, 5, 6, 7, 9, 24, 25, 26, 28, 30, 39, 40, 41] or the Padé fractions [3, 9, 30] (both discrete in time), the Cayley transform [2,16,17,21] (continuous in time), and other ideas (see, for example, [19,22,32,38] and the references therein), the present paper introduces and analyzes new efficient *exponentially convergent* algorithms for the operator exponential including $t = 0$ which are also applied to inhomogeneous problems with certain holomorphic right-hand sides. The algorithms under consideration are parallelizable in an evident way.

The paper is organized as follows. In section 2 we derive some preliminary results concerning estimates of the resolvent of strongly positive operators by fractional powers of these operators. In this section we also discuss the question of selecting an integration parabola enveloping the spectrum of strongly P-positive operators avoiding intersection with the spectral parabola. In section 3 we analyze a parabola different from that of [12, 18] as the integration path in the Dunford–Cauchy representation of the operator exponential generated by a strongly P-positive operator A in a Banach space [11]. Then we construct a sinc approximation of this representation and give a new unified estimate for all $t \geq 0$ which shows the exponential convergence for $t > 0$ and presents a scale of estimates with respect to σ for $t = 0$ provided that $u_o \in D(A^\sigma), \quad \sigma > 1/2$. Using a hyperbola as the integration path, in section 4 we justify a new algorithm (a quadrature sum with with a step-size h including $2N$ resolvents) for the operator exponential e^{-At} which is of the order $\mathcal{O}(e^{-c\sqrt{N}})$ uniformly in $t \geq 0$ provided that $h = \mathcal{O}(1/\sqrt{N})$ and of the order $\mathcal{O}\left(\max\left\{e^{-\pi dN/(c_1 \ln N)},\ e^{-c_1 a_I tN/2 - c_1 \alpha \ln N}\right\}\right)$ for each fixed $t \geq 0$ provided that $h = c_1 \ln N/N$. Note that this algorithm supposes $u_0 \in D(A^\sigma),\ \sigma > 0$, but does not need the computation of $A^\sigma u_0$. The algorithms of section 4 based on the integration along a hyperbola have much better convergence properties than the algorithms of section 3 based on the integration along a parabola. Nevertheless we consider it necessary to also include these results as an example of the application of estimates from section 2 and in order to complete the theory developed in [12, 15, 18, 23] which was not extended for the case $t \geq 0$.

Let A be a densely defined strongly positive operator and $u_0 \in D(A^\alpha),\ \alpha \in (0, 1)$. Then sinc quadrature (4.22) represents an approximate solution of the homogeneous initial value problem (1.1) (i.e., $u(t) = e^{-At}u_0$) and possesses a uniform, with respect to $t \geq 0$, exponential convergence rate with estimate (4.23) which is of the order $\mathcal{O}(e^{-c\sqrt{N}})$ uniformly in $t \geq 0$ provided that $h = \mathcal{O}(1/\sqrt{N})$ and of the order $\mathcal{O}\left(\max\left\{e^{-\pi dN/(c_1 \ln N)},\ e^{-c_1 a_I tN/2 - c_1 \alpha \ln N}\right\}\right)$ for each fixed $t \geq 0$ provided that $h = c_1 \ln N/N$.

Since the integrand in (1.7) is mainly concentrated on a finite interval and decreases very rapidly outside the interval, we truncate the integral to the one over the finite interval and implement a sinc quadrature to approximate it. Estimating the remainders and equalizing all estimates leads to another exponentially convergent approximation. A comparative analysis of both approximations is given in section 5 and shows that despite of lower asymptotic convergence rate this approach can be better for N not very large. Section 6 deals with the inhomogeneous problem (1.1). In order to approximate the inhomogeneous problem, we represent the second summand in

(1.4) in the following way:

$$
\begin{aligned}
u_p(t) &= \int_0^t e^{-A(t-\xi)} f(\xi) d\xi = \frac{1}{2\pi i} \int_{\Gamma_I} \int_0^t e^{-z(t-\xi)} f(\xi) d\xi dz \\
&= \int_{-\infty}^{\infty} \mathcal{F}_1(t,\eta) d\eta.
\end{aligned} \tag{1.8}
$$

We then replace this integral by a new quadrature. The latter uses our first algorithm to compute the integrand $\mathcal{F}_1(t,\eta)$ at each quadrature node. Note that one of the crucial points in our approach disinguishing the new algorithms from other ones cited above is the use of a modified resolvent which allows one to also get a numerical stable algorithm for small t. The theoretical results of this paper are confirmed by numerical examples. We establish the conditions on $f(t)$ under which this algorithm possesses an exponential convergence rate when solving the problem (1.2) on the infinite interval $[0,\infty)$. It turns out that the function $f(t)$ has to possess the analytical extension in a sector $\Sigma_f = \{z = \rho e^{i\psi} : \ \rho \in (0,\infty), \ \psi \in (-\phi_1, \phi_1), 0 < \phi_1 < \pi/2\}$ and has to possess the estimate $\|f(z)\| \le ce^{-c_1 \Re z}$ for all $z \in \Sigma_f$, where ϕ_1, c_1 are consistent with the spectral characteristics of the operator A.

2. Preliminaries.

2.1. Estimates of the resolvent through fractional powers of strongly positive operators.

Let A be a densely defined strongly positive (sectorial) operator in a Banach space X with the domain $D(A)$; i.e., its spectrum $\Sigma(A)$ lies in the sector

$$
\Sigma = \left\{ z = a_0 + re^{i\theta} : \ r \in [0,\infty), \ |\theta| < \varphi < \frac{\pi}{2} \right\} \tag{2.1}
$$

and on its boundary Γ_Σ, and outside the sector the following estimate for the resolvent holds true:

$$
\|(zI - A)^{-1}\| \le \frac{M}{1+|z|} \tag{2.2}
$$

with some positive constant M (compare with [17,27,29,35]). The angle φ is called the spectral angle of the operator A. A practically important example of strongly positive operators in $X = L_p(\Omega), \ 0 < p < \infty$, represents a strongly elliptic partial differential operator [10,11,12,17,21,29,31] where the parameters a_0, φ of the sector Σ are defined by its coefficients.

For an initial vector $u_0 \in D(A^{m+1})$ it holds

$$
\sum_{k=1}^{m+1} \frac{A^{k-1} u_0}{z^k} + \frac{1}{z^{m+1}} (zI - A)^{-1} A^{m+1} u_0 = (zI - A)^{-1} u_0. \tag{2.3}
$$

This equality together with

$$
A^{-(m+1)} v = \frac{1}{2\pi i} \int_{\Gamma_I} z^{-(m+1)} (zI - A)^{-1} v dz, \tag{2.4}
$$

by setting $v = A^{m+1} u_0$, yields the following representation:

$$
\begin{aligned}
u_0 &= A^{-(m+1)} A^{m+1} u_0 = \frac{1}{2\pi i} \int_{\Gamma_I} z^{-(m+1)} (zI - A)^{-1} A^{m+1} u_0 dz \\
&= \int_{\Gamma_I} \left[(zI - A)^{-1} - \sum_{k=1}^{m+1} \frac{A^{k-1}}{z^k} \right] u_0 dz
\end{aligned} \tag{2.5}
$$

with an integration path Γ_I situated in the right half-plane and enveloping Γ_Σ. Let us estimate the norm of the first integrand in (2.5) as a function of $|z|$ under the assumption $u_0 \in D(A^{m+\alpha})$, $m \in \mathbb{N}$, $\alpha \in [0,1]$. Since the operator A is strongly positive it holds on and outside the integration path:

$$\begin{aligned} &\|(zI-A)^{-1}w\| \le \frac{M}{1+|z|}\|w\|, \\ &\|A(zI-A)^{-1}w\| \le (1+M)\|w\|. \end{aligned} \tag{2.6}$$

These estimates yield (see, e.g., Theorem 4 of [27])

$$\|A^{1-\alpha}(zI-A)^{-1}w\| \le K\|A(zI-A)^{-1}w\|^{1-\alpha}\|(zI-A)^{-1}w\|^{\alpha}, \tag{2.7}$$

where the constant K depends on α and M only. This inequality while taking into account (2.6) implies

$$\|A^{1-\alpha}(zI-A)^{-1}\| \le \frac{K(1+M)}{(1+|z|)^{\alpha}}, \quad \alpha \in [0,1], \tag{2.8}$$

which leads to the estimate

$$\begin{aligned} &\left\| \left[(zI-A)^{-1} - \frac{1}{z}I\right]u_0 \right\| = \frac{1}{|z|}\|A^{1-\alpha}(zI-A)^{-1}A^{\alpha}u_0\| \\ &\le \frac{(1+M)K}{|z|(1+|z|)^{\alpha}}\|A^{\alpha}u_0\| \qquad \forall \alpha \in [0,1],\ u_0 \in D(A^{\alpha}). \end{aligned} \tag{2.9}$$

This estimate can be easily generalized to

$$\begin{aligned} &\left\| \left[(zI-A)^{-1} - \sum_{k=1}^{m+1}\frac{A^{k-1}}{z^k}\right]u_0 \right\| = \left\| \frac{1}{z^{m+1}}(zI-A)^{-1}A^{m+1}u_0 \right\| \\ &= \frac{1}{|z|^{m+1}}\|A^{1-\alpha}(zI-A)^{-1}A^{m+\alpha}u_0\| \le \frac{1}{|z|^{m+1}}\frac{(1+M)K}{(1+|z|)^{\alpha}}\|A^{m+\alpha}u_0\| \\ &\forall \alpha \in [0,1],\ u_0 \in D(A^{m+\alpha}). \end{aligned} \tag{2.10}$$

Thus, we get the following result, which we will need below.

THEOREM 2.1. *Let $u_0 \in D(A^{m+\alpha})$ for some $m \in \mathbb{N}$, and let $\alpha \in [0,1]$. Then the estimate* (2.10) *holds true.*

2.2. The integration parabola. There are many possibilities to define and to approximate functions of an operator A. Let Γ be the boundary of a domain Σ in the complex plane containing the spectrum of A, and let $\tilde{f}(z)$ be an analytical function in Σ. Then the Dunford–Cauchy integral

$$\tilde{f}(A) = \frac{1}{2\pi i}\int_{\Gamma}\tilde{f}(z)(zI-A)^{-1}dz \tag{2.11}$$

defines a function of A provided that the integral converges.

By a parametrizing of $\Gamma = \{z = \xi(s) + i\eta(s) :\ s \in (-\infty,\infty)\}$, one can translate the integral (2.11) into the integral

$$\tilde{f}(A) = \int_{-\infty}^{\infty}F(s)ds \tag{2.12}$$

with

$$F(s) = \frac{1}{2\pi i}\tilde{f}(z)(zI - A)^{-1}z'(s). \tag{2.13}$$

Choosing various integration paths and using various quadrature formulas, one can obtain approximations of $\tilde{f}(A)$ with desired properties (see, for example, [12, 13, 15, 17, 21, 23] where various functions of operators were investigated).

It was shown in [8, 11, 12, 18, 31] that the spectrum of a strongly elliptic operator in a Hilbert space lies in a domain enveloped by a parabola defined by the coefficients of the operator and that the resolvent on and outside of the parabola possesses the estimate

$$\|(zI - A)^{-1}\| \leq \frac{M}{1 + \sqrt{|z|}} \tag{2.14}$$

with some positive constant M. Such operators are called strongly P-positive operators. The paper [31] also contains examples of differential operators which are strongly P-positive in such genuine Banach spaces as $L_1(0,1)$ or $L_\infty(0,1)$. One of the natural choices of the integration path for these operators is a parabola which does not intersect the spectral parabola containing the spectrum of the operator.

Let

$$\Gamma_0 = \{z = \xi - i\eta :\ \xi = a_0\eta^2 + b_0,\ a_0 > 0,\ b_0 > 0,\ \eta \in (-\infty, \infty)\} \tag{2.15}$$

be the spectral parabola enveloping the spectrum of the operator A. [12, 18] showed how one can define the coefficients of an integration parabola by the coefficients of the spectral parabola so that the integrand in (2.12) can be analytically extended into a symmetric strip D_d of a width $2d$ around the real axes, but this choice was rather complicated.

Below we propose another (simpler) method to define the integration parabola through the spectral one.

We have to choose an integration parabola

$$\Gamma_I = \{z = \xi - i\eta :\ \xi = a_I\eta^2 + b_I,\ a_I > 0,\ b_I > 0,\ \eta \in (-\infty, \infty)\} \tag{2.16}$$

so that its top lies in $(0, b_0)$ and its opening is greater than the one of the spectral parabola, i.e., $a_I < a_0$. Moreover, by changing η to $\eta + i\nu$ the set of parabolas

$$\begin{aligned} &\Gamma(\nu) = \{z = \xi - i\eta : \xi = a_I\eta^2 + b_I - a_I\nu^2 + \nu - i\eta(1 - 2a_I\nu), \eta \in (-\infty, \infty)\} \\ &\quad = \left\{z = \xi - i\tilde{\eta} :\ \xi = \frac{a_I}{(1 - 2a_I\nu)^2}\tilde{\eta}^2 + b_I - a_I\nu^2 + \nu, \tilde{\eta} = (1 - 2a_I\nu)\eta \in (-\infty, \infty)\right\}, \end{aligned} \tag{2.17}$$

for $|\nu| < d$ must lie outside of the spectral parabola (only in this case can one guarantee that the resolvent of A remains bounded). Note that the substitution $\tilde{\eta} = (1 - 2a_I\nu)\eta$ must be nonsingular for all $|\nu| < d$, which yields $a_I < 1/(2d)$. We choose d so that the top of the integration parabola coincides with top of the spectral one and the opening of the integration parabola is greater than the opening of the spectral parabola for $\nu = d$. For $\nu = -d$ we demand that the integration parabola lies outside of the spectral parabola and that its top lies at the origin. Thus, it must be

$$\begin{cases} \frac{a_I}{(1-2a_Id)^2} = a_0, \\ b_I - a_Id^2 + d = b_0, \\ b_I - a_Id^2 - d = 0. \end{cases} \tag{2.18}$$

It follows immediately from the last two equations that $2d = b_0$. From the first equation

$$4d^2 a_0 a_I^2 - a_I(1 + 4a_0 d) + a_0 = 0. \tag{2.19}$$

After the substitution $d = b_0/2$ we get

$$a_I = \frac{1 + 2a_0 b_0 \pm \sqrt{1 + 4a_0 b_0}}{2a_0 b_0}, \tag{2.20}$$

but only the root

$$a_I = \frac{1 + 2a_0 b_0 - \sqrt{1 + 4a_0 b_0}}{2a_0 b_0} = \frac{2a_0}{1 + 2a_0 b_0 + \sqrt{1 + 4a_0 b_0}} \tag{2.21}$$

satisfies the condition $a_I < 1/(2d) = 1/b_0$. Thus, the parameters of the integration parabola from which the integrand can be analytically extended into the strip D_d of the width

$$d = b_0/2 \tag{2.22}$$

are

$$\begin{aligned} a_I &= \frac{1 + 2a_0 b_0 - \sqrt{1 + 4a_0 b_0}}{2a_0 b_0} = \frac{2a_0}{1 + 2a_0 b_0 + \sqrt{1 + 4a_0 b_0}}, \\ b_I &= \frac{a_I b_0^2}{4} + \frac{b_0}{2}. \end{aligned} \tag{2.23}$$

3. New algorithm with integration along a parabola and a scale of estimates. Let A be a strongly P-positive operator, and let

$$u_0 \in D(A^\alpha),\ \alpha > 0. \tag{3.1}$$

In this case due to (2.10) with $m = 0$ we have

$$\begin{aligned} &\left\| \left[(zI - A)^{-1} - \frac{1}{z} I \right] u_0 \right\| = \left\| \frac{1}{z} (zI - A)^{-1} A u_0 \right\| \\ &= \frac{1}{|z|} \| A^{1-\alpha} (zI - A)^{-1} A^\alpha u_0 \| \le \frac{1}{|z|} \| A^{1-\alpha} (zI - A)^{-1} \| \| A^\alpha u_0 \|. \end{aligned} \tag{3.2}$$

The resolvent of the strongly P-positive operator is bounded on and outside the spectral parabola; more precisely, we have

$$\begin{aligned} &\|(zI - A)^{-1} w\| \le \frac{M}{1 + \sqrt{|z|}} \|w\|, \\ &\|A(zI - A)^{-1} w\| \le \left(1 + \frac{M|z|}{1 + \sqrt{|z|}} \right) \|w\| \le (1 + M\sqrt{|z|}) \|w\|. \end{aligned} \tag{3.3}$$

We suppose that our operator A is at the same time strongly positive (note that a strongly elliptic operator is both strongly P-positive [11] and strongly positive). We can use Theorem 4 of [27] and get

$$\begin{aligned} &\|A^{1-\alpha} (zI - A)^{-1} w\| \le K(\alpha) (1 + M\sqrt{|z|})^{1-\alpha} \left(\frac{M}{1 + \sqrt{|z|}} \right)^\alpha \|w\| \\ &\le \max(1, M) K(\alpha) \frac{\|w\|}{(1 + \sqrt{|z|})^{2\alpha - 1}} \end{aligned} \tag{3.4}$$

with a constant $K(\alpha)$ independent of α where $K(1) = K(0) = 1$. The last inequality and (3.1) imply that

$$\left\| \left[(zI - A)^{-1} - \frac{1}{z} I \right] u_0 \right\| \le \max(1, M) K(\alpha) \frac{\|A^\alpha u_0\|}{|z|^{\alpha + \frac{1}{2}}} \tag{3.5}$$

which justifies for the integration path above the following representation of the solution of the homogeneous problem (1.1)

$$\begin{aligned} u(t) = e^{-At} u_0 &= \frac{1}{2\pi i} \int_{\Gamma_I} e^{-tz} (zI - A)^{-1} u_0 dz \\ &= \frac{1}{2\pi i} \int_{\Gamma_I} e^{-tz} \left[(zI - A)^{-1} - \frac{1}{z} I \right] u_0 dz, \end{aligned} \tag{3.6}$$

provided that $\alpha > 0$. After parametrizing the integral we get

$$u(t) = \int_{-\infty}^{\infty} F(t, \eta) d\eta \tag{3.7}$$

with

$$\begin{aligned} F(t, \eta) &= -\frac{1}{2\pi i} (2 a_I \eta - i) e^{-t(a_I \eta^2 + b_I - i\eta)} \\ &\times \left\{ [(a_I \eta^2 + b_I - i\eta) I - A]^{-1} - \frac{1}{a_I \eta^2 + b_I - i\eta} I \right\} u_0. \end{aligned} \tag{3.8}$$

Following to [36], we construct a quadrature rule for the integral in (2.12) by using the sinc approximation on $(-\infty, \infty)$. For $1 \le p \le \infty$, introduce the family $\mathbf{H}^p(D_d)$ of all vector-valued functions, which are analytic in the infinite strip D_d,

$$D_d = \{ z \in \mathbb{C} : -\infty < \Re z < \infty, |\Im z| < d \}, \tag{3.9}$$

such that if $D_d(\epsilon)$ is defined for $0 < \epsilon < 1$ by

$$D_d(\epsilon) = \{ z \in \mathbb{C} : |\Re z| < 1/\epsilon, |\Im z| < d(1 - \epsilon) \}, \tag{3.10}$$

then for each $\mathcal{F} \in \mathbf{H}^p(D_d)$ there holds $\|\mathcal{F}\|_{\mathbf{H}^p(D_d)} < \infty$ with

$$\|\mathcal{F}\|_{\mathbf{H}^p(D_d)} = \begin{cases} \lim\limits_{\epsilon \to 0} \left(\int_{\partial D_d(\epsilon)} \|\mathcal{F}(z)\|^p |dz| \right)^{1/p} & \text{if } 1 \le p < \infty, \\ \lim\limits_{\epsilon \to 0} \sup\limits_{z \in \partial D_d(\epsilon)} \|\mathcal{F}(z)\| & \text{if } p = \infty. \end{cases} \tag{3.11}$$

Let

$$S(k, h)(x) = \frac{\sin[\pi(x - kh)/h]}{\pi(x - kh)/h} \tag{3.12}$$

be the kth sinc function with step-size h, evaluated in x. Given $\mathcal{F} \in \mathbf{H}^p(D_d), h > 0,$

and positive integer N, let us use the notations

$$\begin{aligned}
&I(\mathcal{F}) = \int_{\mathbb{R}} \mathcal{F}(x)dx, \qquad T_N(\mathcal{F},h) = h\sum_{k=-N}^{N}\mathcal{F}(kh),\\
&T(\mathcal{F},h) = h\sum_{k=-\infty}^{\infty}\mathcal{F}(kh),\\
&C(\mathcal{F},h) = \sum_{k=-\infty}^{\infty}\mathcal{F}(kh)S(k,h),\\
&\eta_N(\mathcal{F},h) = I(\mathcal{F}) - T_N(\mathcal{F},h), \qquad \eta(\mathcal{F},h) = I(\mathcal{F}) - T(\mathcal{F},h).
\end{aligned}$$

Applying the quadrature rule T_N with the vector-valued function (3.8), we obtain for integral (3.7)

$$u(t) = \exp(-tA)u_0 \approx u_N(t) = \exp_N(-tA)u_0 = h\left(\sum_{k=-N}^{N}F(kh,t)\right)u_0. \tag{3.13}$$

Below we show that this Sinc-quadrature approximation with a proper choice of h converges exponentially provided that the integrand can be analytically extended into a strip D_d. This property of the integrand depends on the choice of the integration path.

Taking into account (3.5) we get

$$\|F(t,\eta)\| \le c\frac{e^{-t(a_I\eta^2+b_I)}}{(1+|\eta|)^{2\alpha}}\|A^\alpha u_0\| \qquad \forall t\ge 0,\ \alpha > 1/2 \tag{3.14}$$

(the inequality $\alpha > 1/2$ guarantees the convergence of the integral (3.7)). The analysis of the integration parabola above implies that the vector-valued function $F(\eta,t)$ can be analytically extended into the strip D_d and belongs to the class $\mathbf{H}^1(D_d)$ with respect to η with the estimate

$$\|F(t,z)\|_{\mathbf{H}^1(D_d)} \le c\frac{e^{-b_It}}{2\alpha-1}\|A^\alpha u_0\| \qquad \forall t\ge 0,\ \alpha > 1/2. \tag{3.15}$$

For our further analysis of the error $\eta_N(\mathcal{F},h) = \exp(-tA)u_0 - \exp_N(-tA)u_0$ of the quadrature rule (3.13), we use the following lemma from [23].

LEMMA 3.1. *For any vector-valued function $f \in \mathbf{H}^1(D_d)$, there holds*

$$\eta(\tilde{f},h) = \frac{i}{2}\int_{\mathbb{R}}\left\{\frac{\tilde{f}(\xi - id^-)e^{-\pi(d+i\xi)/h}}{\sin[\pi(\xi-id)/h]} - \frac{\tilde{f}(\xi + id^-)e^{-\pi(d-i\xi)/h}}{\sin[\pi(\xi+id)/h]}\right\}d\xi, \tag{3.16}$$

which yields the estimate

$$\|\eta(\tilde{f},h)\| \le \frac{e^{-\pi d/h}}{2\sinh(\pi d/h)}\|\tilde{f}\|_{\mathbf{H}^1(D_d)}. \tag{3.17}$$

If, in addition, $\tilde{f}$ satisfies on $\mathbb{R}$ the condition

$$\begin{aligned}&\|\tilde{f}(x)\| \le \frac{ce^{-\beta x^2}}{(1+x^2)^\sigma}, \qquad 1/2 < \sigma \le 1\\ &c,\beta > 0,\end{aligned} \tag{3.18}$$

then

$$\|\eta_N(\tilde{f},h)\| \le \frac{2c}{2\sigma-1}\left\{2\sigma\frac{\exp(-\pi d/h)}{\sinh(\pi d/h)}+\frac{\exp\left(-\beta(Nh)^2\right)}{(Nh)^{2\sigma-1}}\right\}. \tag{3.19}$$

Taking into account the estimates (3.14) and (3.15) and setting F for $\tilde{f}$, α for σ, and ta_I for β, we get the estimate

$$\begin{aligned}\|\eta_N(F,h)\| &= \|\exp(-tA)u_0-\exp_N(-tA)u_0\|\\ &\le c\frac{e^{-b_It}}{2\alpha-1}\left\{2\alpha\frac{\exp(-\pi d/h)}{\sinh(\pi d/h)}+\frac{\exp\left(-a_It(Nh)^2\right)}{(Nh)^{2\alpha-1}}\right\}\|A^\alpha u_0\|.\end{aligned} \tag{3.20}$$

Equalizing the exponents here by setting $\pi d/h = a_I(Nh)^2$, we get for the step-size of the quadrature

$$h=\sqrt[3]{\pi d/(a_IN^2)}. \tag{3.21}$$

Since $\sinh(\pi d/h)\ge e^{\pi d/h}/2$, $\pi d/h=(\sqrt{a_I}\pi dN)^{2/3}$, $Nh=\sqrt[3]{\pi dN/a_I}$, $(Nh)^{2\alpha-1}=(\pi dN/a_I)^{(2\alpha-1)/3}$, and $d=b_0$, we get with this step-size the following scale of estimates for the algorithm (3.13):

(3.22)
$$\begin{aligned}&\|\eta_N(\mathcal{F},h)\| = \|\exp(-tA)u_0-\exp_N(-tA)u_0\|\\ &\le c\frac{e^{-b_It}}{2\alpha-1}\left\{4\alpha\exp(-2(\sqrt{a_I}\pi b_0)^{2/3}N^{2/3})+\frac{\exp\left(-a_It(\pi b_0N/a_I)^{2/3}\right)}{(\pi b_0N/a_I)^{(2\alpha-1)/3}}\right\}\|A^\alpha u_0\|.\end{aligned}$$

Thus, we have proven the following statement.

THEOREM 3.2. *Let A be a strongly P-positive operator in a Banach space X with the resolvent satisfying* (2.14) *and with the spectral parabola given by* (2.15). *Then for the sinc approximation* (3.13) *we have the estimate* (3.22), *i.e.,*

$$\|\exp(-tA)u_0-\exp_N(-tA)u_0\| = \begin{cases}\mathcal{O}(e^{-c_1N^{2/3}}) & \text{if } t>0,\\ \mathcal{O}(N^{(2\alpha-1)/3}) & \text{if } t=0,\end{cases} \tag{3.23}$$

provided that $u_0\in D(A^\alpha)$, $\alpha>1/2$.

Remark 3.1. The above algorithm possesses two sequential levels of parallelism: First, one can compute all $\hat{u}(z_p)$ at step 2 in parallel, and second, the solution $u(t)=e^{-At}u_0$ at different time values $(t_1,t_2,\dots,t_M)$.

4. New algorithm for the operator exponential with an exponential convergence estimate including $t=0$. We consider the following representation of the operator exponential:

$$u(t)=\frac{1}{2\pi i}\int_{\Gamma_I}e^{-zt}(zI-A)^{-1}u_0dz. \tag{4.1}$$

Our aim is to approximate this integral by a quadrature with exponential convergence rate including $t=0$. It is of great importance to have in mind the representation of the solution of the nonhomogeneous initial value problem (1.1) by

$$u(t)=e^{-At}u_0+\int_0^t e^{-A(t-\xi)}\tilde{f}(\xi)d\xi, \tag{4.2}$$

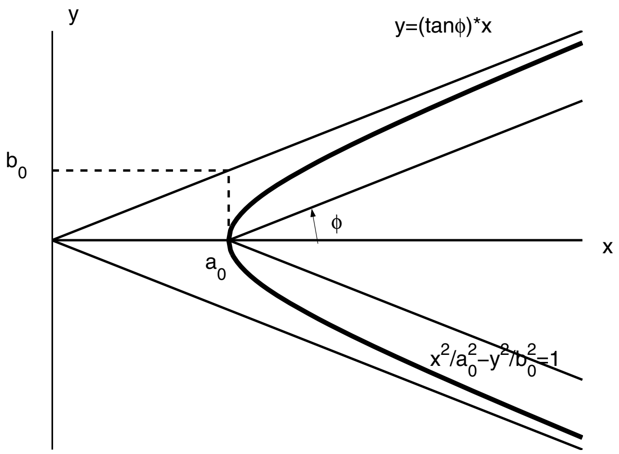


FIG. 4.1. *Spectral characteristics of the operator A.*

where the argument of the operator exponential under the integral becomes zero for $\xi = t$. Taking into account (2.5) for $m = 0$, one can see that we can represent

$$u(t) = \frac{1}{2\pi i}\int_{\Gamma_I} e^{-zt}\left[(zI - A)^{-1} - \frac{1}{z}I\right]u_0 dz \tag{4.3}$$

instead of (4.1) (for $t > 0$, the integral from the second summand is equal to zero due to the analyticity of the integrand inside of the integration path), and this integral represents the solution of the problem (1.1) for $u_0 \in D(A^\alpha),\ \alpha > 0$. We call the hyperbola

$$\Gamma_0 = \{z(\xi) = a_0\cosh\xi - ib_0\sinh\xi :\ \xi \in (-\infty,\infty),\ b_0 = a_0\tan\varphi\} \tag{4.4}$$

the spectral hyperbola, which has paths through the vertex $(a_0, 0)$ of the spectral angle and possesses asymptotes which are parallel to the rays of the spectral angle Σ (see Figure 4.1). We choose the following hyperbola as an integration path:

$$\Gamma_I = \{z(\xi) = a_I\cosh\xi - ib_I\sinh\xi :\ \xi \in (-\infty,\infty)\}. \tag{4.5}$$

After parametrizing the integral (4.3) by (4.5), we get

$$u(t) = \frac{1}{2\pi i}\int_{-\infty}^{\infty}\mathcal{F}(t,\xi)d\xi \tag{4.6}$$

with

$$\begin{aligned} &\mathcal{F}(t,\xi) = F_A(t,\xi)u_0,\\ &F_A(t,\xi) = e^{-z(\xi)t}(a_I\sinh\xi - ib_I\cosh\xi)\left[(z(\xi)I - A)^{-1} - \frac{1}{z(\xi)}I\right]. \end{aligned} \tag{4.7}$$

In order to estimate $\|\mathcal{F}(t,\xi)\|$ we need an estimate for $|z'(\xi)/z(\xi)| = \overline{(a_I\sinh\xi - ib_I\cosh\xi)/(a_I\cosh\xi - ib_I\sinh\xi)} = \sqrt{(a_I^2\tanh^2\xi + b_I^2)/(b_I^2\tanh^2\xi + a_I^2)}$. The quotient under the square root as a function of $v = \tanh^2\xi \in [0,1]$ takes its maximum at $v = 0$ since the sign of the first derivative coincides with the sign of

$a_I^4 - b_I^4 = -a_0^4 \sin\varphi/\cos^4\varphi$, i.e., we have

$$|z'(\xi)/z(\xi)| \le b_I/a_I. \tag{4.8}$$

Supposing $u_0 \in D(A^\alpha)$, $0 < \alpha < 1$, using (4.8) and Theorem 2.1, we can estimate the integrand on the real axis $\xi \in \mathbb{R}$ for each $t \ge 0$ by

$$\begin{aligned} \|\mathcal{F}(t,\xi)\| &\le e^{-a_I t\cosh\xi}\frac{(1+M)K\sqrt{a_I^2\sinh^2\xi + b_I^2\cosh^2\xi}}{(a_I^2\cosh^2\xi + b_I^2\sinh^2\xi)^{(1+\alpha)/2}}\|A^\alpha u_0\| \\ &\le (1+M)K\frac{b_I}{a_I}\frac{e^{-a_I t\cosh\xi}}{(a_I^2\cosh^2\xi + b_I^2\sinh^2\xi)^{\alpha/2}}\|A^\alpha u_0\| \\ &\le (1+M)K\frac{b_I}{a_I}\left(\frac{2}{a_I}\right)^\alpha e^{-a_I t\cosh\xi - \alpha|\xi|}\|A^\alpha u_0\|, \qquad \xi \in \mathbb{R},\ t \ge 0. \end{aligned} \tag{4.9}$$

Let us show that the function $\mathcal{F}(t,\xi)$ can be analytically extended with respect to ξ into a strip of a width d_1. After changing ξ to $\xi + i\nu$ the integration hyperbola Γ_I will be translated into the curve

$$\begin{aligned} \Gamma(\nu) &= \{z(w) = a_I\cosh(\xi+i\nu) - ib_I\sinh(\xi+i\nu) :\ \xi \in (-\infty,\infty)\} \\ &= \{z(w) = a(\nu)\cosh\xi - ib(\nu)\sinh\xi :\ \xi \in (-\infty,\infty)\} \end{aligned} \tag{4.10}$$

with

$$\begin{aligned} a(\nu) &= a_I\cos\nu + b_I\sin\nu = \sqrt{a_I^2+b_I^2}\sin(\nu+\phi/2), \\ b(\nu) &= b_I\cos\nu - a_I\sin\nu = \sqrt{a_I^2+b_I^2}\cos(\nu+\phi/2), \\ \cos\frac{\phi}{2} &= \frac{b_I}{\sqrt{a_I^2+b_I^2}}, \qquad \sin\frac{\phi}{2} = \frac{a_I}{\sqrt{a_I^2+b_I^2}}. \end{aligned} \tag{4.11}$$

The analyticity of the function $\mathcal{F}(t,\xi+i\nu)$, $|\nu| < d_1/2$, can be violated if the resolvent becomes unbounded. Thus, we must choose d_1 so that the hyperbola $\Gamma(\nu)$ for $\nu \in (-d_1/2, d_1/2)$ remains in the right half-plane of the complex plane, for $\nu = -d_1/2$ coincides with the imaginary axis, for $\nu = d_1/2$ coincides with the spectral hyperbola, and for all $\nu \in (-d_1/2, d_1/2)$ does not intersect the spectral sector. Then we choose the hyperbola $\Gamma(0)$ as the integration hyperbola.

This implies the following system of equations

$$\begin{cases} a_I\cos(d_1/2) + b_I\sin(d_1/2) = a_0, \\ b_I\cos(d_1/2) - a_I\sin(d_1/2) = a_0\tan\varphi, \\ a_I\cos(-d_1/2) + b_I\sin(-d_1/2) = 0, \end{cases} \tag{4.12}$$

from which we get

$$\begin{cases} 2a_I\cos(d_1/2) = a_0, \\ b_I = a_0\sin(d_1/2) + b_0\cos(d_1/2), \\ a_I = a_0\cos(d_1/2) - b_0\sin(d_1/2). \end{cases} \tag{4.13}$$

Eliminating a_I from the first and the third equations, we get $a_0\cos d_1 = b_0\sin d_1$, i.e., $d_1 = \pi/2 - \varphi$ with $\cos\varphi = \frac{a_0}{\sqrt{a_0^2+b_0^2}}$, $\sin\varphi = \frac{b_0}{\sqrt{a_0^2+b_0^2}}$. Thus, if we choose the parameters of the integration hyperbola by

$$\begin{aligned} a_I &= a_0\cos\left(\frac{\pi}{4}-\frac{\varphi}{2}\right) - b_0\sin\left(\frac{\pi}{4}-\frac{\varphi}{2}\right) \\ &= \sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right) = a_0\frac{\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)}{\cos\varphi}, \\ b_I &= a_0\sin\left(\frac{\pi}{4}-\frac{\varphi}{2}\right) + b_0\cos\left(\frac{\pi}{4}-\frac{\varphi}{2}\right) \\ &= \sqrt{a_0^2+b_0^2}\sin\left(\frac{\pi}{4}+\frac{\varphi}{2}\right) = a_0\frac{\sin\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)}{\cos\varphi}, \end{aligned} \tag{4.14}$$

then the vector-valued function $\mathcal{F}(t,w)$ is for all $t \geq 0$ analytic with respect to $w = \xi + i\nu$ in the strip

$$D_{d_1} = \{w = \xi + i\nu : \ \xi \in (-\infty,\infty), \ |\nu| < d_1/2\}. \tag{4.15}$$

Now, estimate (4.9) takes the form

$$\begin{aligned} \|\mathcal{F}(t,\xi)\| &\leq C(\varphi,\alpha)e^{-a_I t\cosh\xi - \alpha|\xi|}\|A^\alpha u_0\| \\ &\leq C(\varphi,\alpha)e^{-\alpha|\xi|}\|A^\alpha u_0\|, \qquad \xi\in\mathbb{R},\ t\geq 0, \end{aligned} \tag{4.16}$$

with

$$C(\varphi,\alpha) = (1+M)K\tan\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)\left(\frac{2\cos\varphi}{a_0\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)}\right)^\alpha. \tag{4.17}$$

Comparing (4.14) with (4.11) we get $\phi = \pi/2 - \varphi$ and

$$\begin{aligned} &a(\nu) = a_I\cos\nu + b_I\sin\nu = \frac{a_0\sin(\nu+\pi/4-\varphi/2)}{\cos\varphi} = \frac{a_0\cos(\pi/4+\varphi/2-\nu)}{\cos\varphi}, \\ &b(\nu) = b_I\cos\nu - a_I\sin\nu = \frac{a_0\sin(\pi/4+\varphi/2-\nu)}{\cos\varphi}, \\ &0 < a(\nu) < a_0, \quad a_0\tan\varphi < b(\nu) < \frac{a_0}{\cos\varphi}. \end{aligned} \tag{4.18}$$

We choose $d = d_1 - \delta$ for an arbitrarily small positive δ, and for $w \in D_d$ we get the estimate (compare with (4.9))

$$\begin{aligned} \|\mathcal{F}(t,w)\| &\leq e^{-a(\nu)t\cosh\xi}\frac{(1+M)K\sqrt{a^2(\nu)\sinh^2\xi + b^2(\nu)\cosh^2\xi}}{(a^2(\nu)\cosh^2\xi + b^2(\nu)\sinh^2\xi)^{(1+\alpha)/2}}\|A^\alpha u_0\| \\ &\leq (1+M)K\frac{b(\nu)}{a(\nu)}\frac{e^{-a(\nu)t\cosh\xi}}{(a^2(\nu)\cosh^2\xi + b^2(\nu)\sinh^2\xi)^{(\alpha/2)}}\|A^\alpha u_0\| \\ &\leq (1+M)K\frac{b(\nu)}{a(\nu)}\left(\frac{2}{a(\nu)}\right)^\alpha e^{-a(\nu)t\cosh\xi-\alpha|\xi|}\|A^\alpha u_0\| \\ &\leq (1+M)K\tan\left(\frac{\pi}{4}+\frac{\varphi}{2}-\nu\right)\left(\frac{2\cos\varphi}{a_0\cos(\pi/4+\varphi/2-\nu)}\right)^\alpha e^{-\alpha|\xi|}\|A^\alpha u_0\| \\ &\forall w \in D_d. \end{aligned} \tag{4.19}$$

Taking into account that the integrals over the vertical sides of the rectangle $D_d(\epsilon)$ vanish as $\epsilon \to 0$, this estimate implies

$$\|\mathcal{F}(t,\cdot)\|_{\mathbf{H}^1(D_d)} \le \|A^\alpha u_0\| [C_-(\varphi,\alpha,\delta) + C_+(\varphi,\alpha,\delta)] \int_{-\infty}^{\infty} e^{-\alpha|\xi|} d\xi = C(\varphi,\alpha,\delta)\|A^\alpha u_0\| \tag{4.20}$$

with

$$\begin{aligned} &C(\varphi,\alpha,\delta) = \frac{2}{\alpha}[C_+(\varphi,\alpha,\delta) + C_-(\varphi,\alpha,\delta)], \\ &C_\pm(\varphi,\alpha,\delta) = (1+M)K(\cos\varphi)^\alpha \tan\left(\frac{\pi}{4} + \frac{\varphi}{2} \pm \frac{d}{2}\right)\left(\frac{2}{a_0\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\pm\frac{d}{2}\right)}\right)^\alpha. \end{aligned} \tag{4.21}$$

Note that the constant $C(\varphi,\alpha,\delta)$ tends to ∞ if $\alpha \to 0$ or $\delta \to 0,\ \varphi \to \pi/2$.

We approximate integral (4.6) by the sinc quadrature

$$u_N(t) = \frac{h}{2\pi i}\sum_{k=-N}^{N} \mathcal{F}(t, z(kh)) \tag{4.22}$$

with the error

$$\begin{aligned} &\|\eta_N(\mathcal{F},h)\| = \|u(t) - u_N(t)\| \\ &\le \left\| u(t) - \frac{h}{2\pi i}\sum_{k=-\infty}^{\infty}\mathcal{F}(t,z(kh))\right\| + \left\|\frac{h}{2\pi i}\sum_{|k|>N}\mathcal{F}(t,z(kh))\right\| \\ &\le \frac{1}{2\pi}\frac{e^{-\pi d/h}}{2\sinh(\pi d/h)}\|\mathcal{F}\|_{\mathbf{H}^1(D_d)} + \frac{C(\varphi,\alpha)h\|A^\alpha u_0\|}{2\pi}\sum_{k=N+1}^{\infty}\exp[-a_I t\cosh(kh) - \alpha kh] \\ &\le \frac{c\|A^\alpha u_0\|}{\alpha}\left\{\frac{e^{-\pi d/h}}{\sinh(\pi d/h)} + \exp[-a_I t\cosh((N+1)h) - \alpha(N+1)h]\right\}, \end{aligned} \tag{4.23}$$

where the constant c does not depend on h, N, t. Equalizing both exponentials for $t = 0$ by

$$\frac{2\pi d}{h} = \alpha(N+1)h, \tag{4.24}$$

we get for the step-size

$$h = \sqrt{\frac{2\pi d}{\alpha(N+1)}}. \tag{4.25}$$

With this step-size the following error estimate holds true:

$$\|\eta_N(\mathcal{F},h)\| \le \frac{c}{\alpha}\exp\left(-\sqrt{\frac{\pi d\alpha}{2}(N+1)}\right)\|A^\alpha u_0\| \tag{4.26}$$

with a constant c independent of t, N. In the case $t > 0$ the first summand in the exponent of $\exp[-a_I t\cosh((N+1)h) - \alpha(N+1)h]$ in (4.23) contributes mainly to the error order. Setting in this case $h = c_1 \ln N/N$ with some positive constant c_1, we remain asymptotic for a fixed t with an error

$$\|\eta_N(\mathcal{F}, h)\| \le c \left[e^{-\pi d N/(c_1 \ln N)} + e^{-c_1 a_I t N/2 - c_1 \alpha \ln N} \right] \|A^\alpha u_0\|, \tag{4.27}$$

where c is a positive constant. Thus, we have proved the following result.

THEOREM 4.1. *Let A be a densely defined strongly positive operator, and let $u_0 \in D(A^\alpha)$, $\alpha \in (0,1)$. Then sinc quadrature (4.22) represents an approximate solution of the homogeneous initial value problem (1.1) (i.e., $u(t) = e^{-At}u_0$) and possesses a uniform, with respect to $t \ge 0$, exponential convergence rate with estimate (4.23) which is of the order $\mathcal{O}(e^{-c\sqrt{N}})$ uniformly in $t \ge 0$ provided that $h = \mathcal{O}(1/\sqrt{N})$ and of the order $\mathcal{O}\left(\max\left\{e^{-\pi d N/(c_1 \ln N)},\ e^{-c_1 a_I t N/2 - c_1 \alpha \ln N}\right\}\right)$ for each fixed $t \ge 0$ provided that $h = c_1 \ln N/N$.*

Remark 4.1. Two other algorithms of the convergence order $\mathcal{O}(e^{-c\sqrt{N}})$ uniformly in $t \ge 0$ were proposed in [15, Remark 4.3 and (2.41)]. One of them used a sum of resolvents applied to u_0 provided that the operator coefficient is bounded. Another one was based on the representation

$$u(t) = \int_\Gamma z^{-\sigma} e^{-zt} (zI - A)^{-1} A^\sigma u_0 \tag{4.28}$$

valid for $u_0 \in D(A^\sigma), \sigma > 1$. Approximating the integral (after parametrizing Γ) by a sinc quadrature, one gets a short sum of resolvents applied to $A^\sigma u_0$ (see [15, (2.41)] and [42]). The last vector must be computed as a preliminary where in practice $\sigma = 2$ is the first choice. It is easy to see that for $u_0 \in D(A^\sigma)$ both representations (4.28) and (4.3) are, in fact, equivalent although the orders of computational stages (i.e., the algorithms) are different depending on the integral representation in use. But in the case $\sigma < 1$ the convergence theory for (4.28) was not presented in [15, 42]. Our representation (4.3) produces a new approximation through a short sum of modified resolvents $(zI - A)^{-1} - z^{-1}I$ applied to u_0 with the convergence properties given by Theorem 4.1. An approximation of the accuracy order $\mathcal{O}(e^{-cN/\ln N})$ for each fixed $t > 0$ to the operator exponential generated by a strongly P-positive operator and using a short sum of the usual resolvents was recently proposed in [14].

Remark 4.2. Note that taking $(zI - A)^{-1}$ instead of $(zI - A)^{-1} - \frac{1}{z}I$ in (4.3) results in a difference given by

$$D_I(t) = -\frac{1}{2\pi i} \int_{\Gamma_I} e^{-zt} \frac{1}{z} u_0 dz. \tag{4.29}$$

For the integration path Γ_I and $t = 0$ this difference can be calculated analytically. Actually, taking into account that the real part is an odd function and the integral of it in the sense of Cauchy is equal to zero, we further get for the integral of the imaginary part

$$\begin{aligned} D_I(0) &= -\frac{1}{2\pi i} P.V. \int_{\Gamma_I} \frac{1}{z} u_0 dz \\ &= -\frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{a_I b_I d\xi}{a_I^2 \cosh^2 \xi + b_I \sinh^2 \xi} u_0 \\ &= \frac{a_I b_I}{2\pi} \int_{-\infty}^{\infty} \frac{d(\tanh \xi)}{a_I^2 + b_I^2 \tanh^2 \xi} u_0 \\ &= \frac{1}{\pi} \arctan \frac{b_I}{a_I} u_0 = \frac{1}{\pi} \left(\frac{\pi}{4} + \frac{\varphi}{2} \right) u_0, \end{aligned} \tag{4.30}$$

where the factor in the front of u_0 is less than 1/2. It means that one can expect a large error for t small enough when using $(zI-A)^{-1}$ instead of $(zI-A)^{-1}-\frac{1}{z}I$ in (4.3). This phenomena can be observed in the next example. Note that for $t > 0$ integral (4.29) is equal to 0 due to the analyticity of the integrand inside of the integration path.

Example 4.1. Let us choose $a_0 = \pi^2$, $\varphi = 0.8\pi/2$. Then Table 1 gives the values of $\|D_I(t)\|/\|u_0\|$ for various t.

TABLE 1

The unremovable error when using the resolvent instead of $(zI - A)^{-1} - \frac{1}{z}I$.

t	$\|D_I(t)\|/\|u_0\|$
0	0.45
$0.1 \cdot 10^{-8}$	0.404552
$0.1 \cdot 10^{-7}$	0.081008
$0.1 \cdot 10^{-6}$	0.000257
$0.1 \cdot 10^{-5}$	$0.147153 \cdot 10^{-6}$

5. Exponentially convergent algorithm II.

Figure 5.1 shows the behavior of the integrand $\mathcal{F}(t,\xi)$ in (4.6) with the operator A defined by $D(A) = \{v(x):\ v \in H^2(0,1),\ v(0) = v(1) = 0\},\ Au = -\frac{d^2u}{dx^2}$. One can observe that the integrand is concentrated on a small finite interval and decays very rapidly outside of this interval. This fact can be a reason for slow convergence of the above algorithm for N not large enough. In this section we construct another exponentially convergent quadrature which takes into account the behavior of the integrand.

Due to the fact that the integrand exponentially decays on the infinite interval, it is reasonable to use an exponentially convergent quadrature rule on a finite interval where the integrand is mostly concentrated and to estimate the residual part. We represent integral (4.6) in the form

$$u(t) = \frac{1}{2\pi i}\int_{-\infty}^{\infty} \mathcal{F}(t,\xi)d\xi = I_1(t) + I_2(t) \tag{5.1}$$

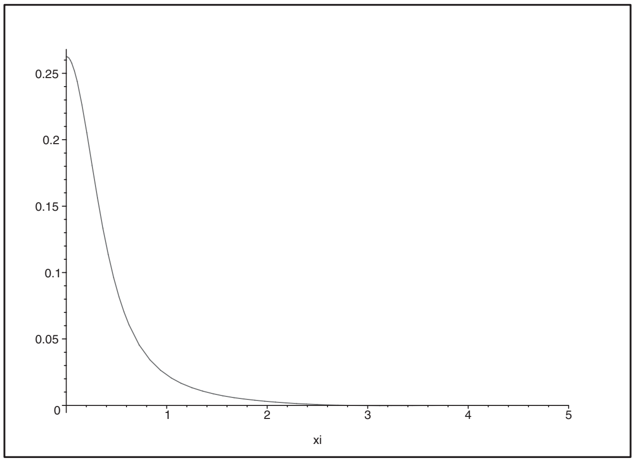


FIG. 5.1. *The behavior of the integrand* $\mathcal{F}(t,\xi)$ *in* (4.6).

with

$$
\begin{aligned}
I_1(t) &= \frac{1}{2\pi i}\int_{-\beta}^{\beta}\mathcal{F}(t,\xi)d\xi,\\
I_2(t) &= \frac{1}{2\pi i}\int_{-\infty}^{-\beta}\mathcal{F}(t,\xi)d\xi + \frac{1}{2\pi i}\int_{\beta}^{\infty}\mathcal{F}(t,\xi)d\xi.
\end{aligned} \tag{5.2}
$$

Using estimate (4.17) we get

(5.3)

$$
\begin{aligned}
&\left\|\frac{\|A^\alpha u_0\|}{2\pi i}\int_{\beta}^{\infty}\mathcal{F}(t,\xi)\right\| \le \frac{\|A^\alpha u_0\|}{2\pi}(1+M)K\tan\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)\left(\frac{2}{\sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)}\right)^\alpha\\
&\times\int_{\beta}^{\infty}e^{-\sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)t\cosh\xi-\alpha|\xi|}d\xi\\
&\le C_1(\varphi,\alpha)\|A^\alpha u_0\|e^{-\sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)t\cosh\beta}\int_{\beta}^{\infty}e^{-\alpha|\xi|}d\xi\\
&\le C_1(\varphi,\alpha)\|A^\alpha u_0\|e^{-\sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)t\cosh\beta}e^{-\alpha|\beta|}
\end{aligned}
$$

with the constant

$$
C_1(\varphi,\alpha) = \frac{(1+M)K}{2\pi\alpha}\tan\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)\left(\frac{2}{\sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)}\right)^\alpha
$$

independent of β. This constant tends to ∞ if $\alpha \to 0$ or $\varphi \to \pi/2$. Analogously one gets

$$
\left\|\frac{1}{2\pi i}\int_{-\infty}^{-\beta}\mathcal{F}(t,\xi)\right\| \le C_1(\varphi,\alpha)\|A^\alpha u_0\|e^{-\sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)t\cosh\beta}e^{-\alpha|\beta|}, \tag{5.4}
$$

which yields the estimate

$$
\|I_2\| \le 2C_1(\varphi,\alpha)\|A^\alpha u_0\|e^{-\sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)t\cosh\beta}e^{-\alpha|\beta|}. \tag{5.5}
$$

Following [36] let us define for $d \in (0,\pi)$ the eye-shaped region

$$
\mathcal{D} = D_d^2 = \left\{z \in \mathbb{C} : \left|\arg\left(\frac{z+\beta}{z-\beta}\right)\right| < d\right\} \tag{5.6}
$$

(see Figure 5.2) and the class $\mathbf{L}_{\kappa,\mu}(\mathcal{D})$ of all holomorphic in $\mathcal{D}$ vector-valued functions satisfying

$$
\|F(z)\| \le c|z+\beta|^{\kappa-1}|z-\beta|^{\mu-1} \tag{5.7}
$$

with some positive constants c, κ, μ.

In the previous section we have shown that $\mathcal{F}(t,\xi)$ can be analytically extended into the symmetric with respect to the real axis strip D_d of the width $2d$. The equation of the boundary of the eye-shaped region in cartesian coordinates is $\frac{2\beta y}{x^2+y^2-\beta^2} = \pm\tan d_1$. For $x = 0$ the maximal value of y, which still lies in the analyticity region, is $y = d$, and we get for the maximal d_1 the equation $\frac{2\beta d}{d^2-\beta^2} = \pm\tan d_1$, from which

$$
d_1 \asymp d/\beta \tag{5.8}
$$

for β large enough.

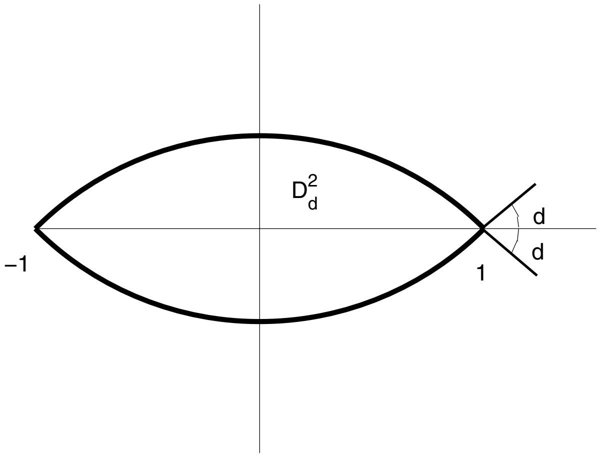


Fig. 5.2. *The eye-shaped region.*

Given N and a function $F(\xi) \in \mathbf{L}_{\kappa,\mu}$, which can be analytically extended into an eye-shaped domain $D^2_{d_1}$, let us define (see [36])

$$\begin{aligned} &\epsilon = \min(\kappa, \mu), \qquad \delta = \max(\kappa, \mu),\\ &h = \left(\frac{2\pi d}{\epsilon n}\right)^{1/2},\\ &M_l = \begin{cases} N & \text{if } \epsilon = \kappa,\\ [\mu N/\kappa] & \text{otherwise,}\end{cases} \qquad M_u = \begin{cases} [\kappa N/\mu] & \text{if } \epsilon = \kappa,\\ N & \text{otherwise.}\end{cases} \end{aligned} \tag{5.9}$$

Then

$$\left\| \int_{-\beta}^{\beta} F(\xi)d\xi - 2\beta h \sum_{-M_l}^{M_u} \frac{e^{kh}}{(1+e^{kh})^2} F(z_k) \right\| \le c e^{-\sqrt{2\pi d_1 \epsilon N}}, \tag{5.10}$$

where the nodes are $z_k = \frac{-\beta+\beta e^{kh}}{1+e^{kh}}$.

Using this quadrature and taking into account that $\mathcal{F}(t,\xi) \in \mathbf{L}_{1,1}(\mathcal{D})$ (with respect to ξ), we get the following sinc quadrature approximation for I_1:

$$\begin{aligned} &I_1(t) \approx I_{1,N}(t) = \frac{2\beta h}{2\pi i} \sum_{-N}^{N} \frac{e^{kh}}{(1+e^{kh})^2} \mathcal{F}(t, z_k),\\ &h = \left(\frac{2\pi d_1}{N}\right)^{1/2} \end{aligned} \tag{5.11}$$

with the approximation error

$$\|\eta_{N,1}(t)\| \le c\|A^\alpha u_0\| e^{-\sqrt{2\pi d_1 N}}. \tag{5.12}$$

Setting

$$u(t) = e^{-At}u_0 \approx I_1(t) \tag{5.13}$$

results in the full approximation error

$$\begin{aligned}\|u(t)-I_{1,N}\| &= \|e^{-At}u_0 - I_{1,N}\| \le \|\eta_{N,1}\| + \|I_2(t)\| \\ &\le c\|A^\alpha u_0\|(e^{-\sqrt{2\pi d_1 N}} + e^{-\sqrt{a_0^2+b_0^2}\cos\left(\frac{\pi}{4}+\frac{\varphi}{2}\right)t\cosh\beta}e^{-\alpha|\beta|}).\end{aligned} \tag{5.14}$$

Equalizing the exponents and taking into account (5.8), we get that $h = \left(\frac{2\pi d}{N^{4/3}}\right)^{1/2}$ and

$$\|e^{-At}u_0 - I_{1,N}(t)\| \le c\|A^\alpha u_0\| e^{-c_1 N^{1/3}} \tag{5.15}$$

provided that

$$\beta \asymp N^{1/3}. \tag{5.16}$$

Example 5.1. We consider problem (1.1) with $u_0 = (1-x)x^2$ and the operator A defined by $D(A) = \{v(x): \ v \in H^2(0,1), \ v(0) = v(1) = 0\}$, $Au = -\frac{d^2u}{dx^2}$. It is easy to see that $u_0 \in D(A^1)$, and the exact solution is given by $u(t,x) = -\frac{4}{\pi^3}\sum_1^\infty \frac{2(-1)^k+1}{k^3}e^{-\pi^2k^2t}\sin(\pi k x)$. One can show that

$$\begin{aligned}(zI-A)^{-1}u_0 - u_0/z &= \frac{1}{z}(zI-A)^{-1}Au_0 \\ &= \frac{6x-2}{z^2} - \frac{\cos[\sqrt{z}(1/2-x)]}{z^2\cos(\sqrt{z}/2)} + 3\frac{\sin[\sqrt{z}(1/2-x)]}{z^2\sin(\sqrt{z}/2)}.\end{aligned} \tag{5.17}$$

Table 2 gives the solutions computed by the algorithm (4.22) with $h = \sqrt{2\pi/N}$ (the first column) and by algorithm (5.13) with $h = \sqrt{2\pi/N^{4/3}}$ (the second column). The exact solution is $u(0,1/2) = u_0(1/2) = 1/8$. This example shows that although algorithm (4.22) is better for N large enough, algorithm (5.13) can be better for relatively small N. Besides the table confirms the exponential convergence of both algorithms.

TABLE 2
The solution for $t = 0$, $x = 1/2$ by the algorithms (4.22) (A1) *and* (5.13) (A2).

N	A1	A2
8	0.147319516168	0.121686777535
16	0.131006555144	0.124073586590
32	0.125894658654	0.124809057018
64	0.125055464496	0.124952849785
128	0.125000975782	0.124995882473
256	0.125000002862	0.124999802171

6. Inhomogeneous differential equation. In this section we consider the inhomogeneous problem (1.1) with the solution

$$u(t) = u_o(t) + u_p(t), \tag{6.1}$$

where

$$u_o(t) = e^{-At}u_0, \quad u_p(t) = \int_0^t e^{-A(t-s)}f(s)ds. \tag{6.2}$$

Note that there exist algorithms for convolution integrals of the same type as the ones from previous sections and also based on sinc quadratures [36]. Since these algorithms use the inverse Laplace transformation combined with Tikhonov's regularization, their justification is rather complicated and the convergence order is $\mathcal{O}(\sqrt{N}e^{-c\sqrt{N}})$. In order to shake off the factor $\sqrt{N}$ in the front of the exponential, we propose in this section a discretization different from [36].

Using representation (4.3) of the operator exponential we get

$$\begin{aligned} u_p(t) &= \int_0^t \frac{1}{2\pi i}\int_{\Gamma_I} e^{-z(t-s)}\left[(zI-A)^{-1}-\frac{1}{z}I\right]f(s)dzds \\ &= \frac{1}{2\pi i}\int_{\Gamma_I}\left[(z(\xi)I-A)^{-1}-\frac{1}{z(\xi)}I\right]\int_0^t e^{-z(\xi)(t-s)}f(s)dsz'(\xi)d\xi, \\ z(\xi) &= a_I\cosh\xi - ib_I\sinh\xi. \end{aligned} \tag{6.3}$$

Replacing here the first integral by quadrature (4.22) we get

$$u_p(t) \approx u_{ap}(t) = \frac{h}{2\pi i}\sum_{k=-N}^{N} z'(kh)\left[(z(kh)I-A)^{-1}-\frac{1}{z(kh)}I\right]f_k(t) \tag{6.4}$$

with

$$f_k(t) = \int_0^t e^{-z(kh)(t-s)}f(s)ds, \qquad k=-N,\dots,N. \tag{6.5}$$

In order to construct an exponentially convergent quadrature for these integrals, we change the variables by

$$\frac{t}{2}-s=\frac{t}{2}\tanh\xi \tag{6.6}$$

and get instead of (6.5)

$$f_k(t) = \int_{-\infty}^{\infty}\mathcal{F}_k(t,\xi)d\xi, \tag{6.7}$$

where

$$\mathcal{F}_k(t,\xi) = \frac{t}{2\cosh^2\xi}\exp[-z(kh)t(1+\tanh\xi)/2]f(t(1-\tanh\xi)/2). \tag{6.8}$$

Note that with the complex variables $z=\xi+i\nu$ and $w=u+iv$, equation (6.6) represents the conformal mapping $w=\psi(z)=t[1-\tanh z]/2$, $z=\phi(w)=\frac{1}{2}\ln\frac{t-w}{w}$, of the strip D_ν onto the domain $\mathcal{A}_\nu$ (compare with the domain D_ν^2 in [36]; also see Figure 6.1). The integrand can be estimated on the real axis by

$$\begin{aligned} \|\mathcal{F}_k(t,\xi)\| &\le \frac{t}{2\cosh^2\xi}\exp[-a_I\cosh(kh)t(1+\tanh\xi)/2]\|f(t(1-\tanh\xi)/2)\| \\ &\le 2te^{-2|\xi|}\|f(t(1-\tanh\xi)/2)\|. \end{aligned} \tag{6.9}$$

Lemma 6.1. *Let the right-hand side $f(t)$ in (1.1) for $t\in[0,\infty]$ be analytically extended into the sector $\Sigma_f=\{\rho e^{i\theta_1}:\ \rho\in[0,\infty],\ |\theta_1|<\varphi\}$, and for all complex $w\in\Sigma_f$ we have*

$$\|f(w)\| \le ce^{-\delta|\Re w|} \tag{6.10}$$

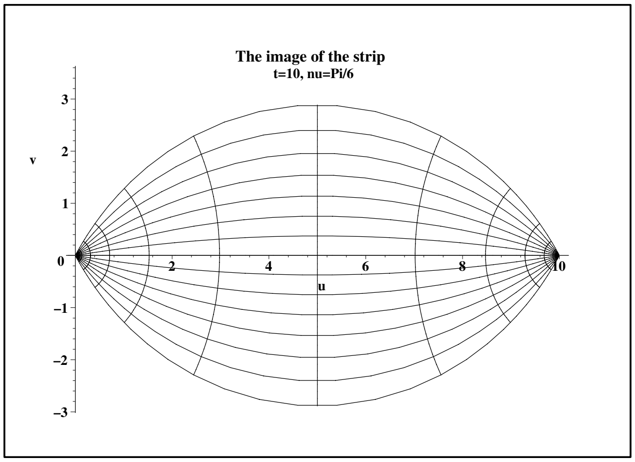


FIG. 6.1. *The image of the strip for* $t = 10,\ \nu = \pi/6$.

with $\delta \in (0, \sqrt{2}a_0]$*. Then the integrand* $\mathcal{F}_k(t,\xi)$ *can be analytically extended into the strip* D_{d_1}, $0 < d_1 < \varphi/2$, *and belongs to the class* $H^1(D_{d_1})$ *with respect to* ξ, *where* a_0, φ *are the spectral characterizations* (2.1) *of* A.

Proof. Let us investigate the domain in the complex plane to which the function $\mathcal{F}(t,\xi)$ can be analytically extended from the real axis $\xi \in \mathbb{R}$. Replacing in the integrand ξ to $\xi + i\nu$, $\xi \in (-\infty,\infty)$, $|\nu| < d_1$, we get in particularly for the argument of f

$$\begin{aligned} \tanh(\xi + i\nu) &= \frac{\sinh\xi\cos\nu + i\cosh\xi\sin\nu}{\cosh\xi\cos\nu + i\sinh\xi\sin\nu} \\ &= \frac{\sinh(2\xi) + i\sin(2\nu)}{2(\cosh^2\xi - \sin^2\nu)}, \\ 1 \pm \tanh(\xi + i\nu) &= q_r^{\pm} + iq_i^{\pm}, \end{aligned} \tag{6.11}$$

where

$$\begin{aligned} q_r^{\pm}(\xi,\nu) &= 1 \pm \frac{\sinh 2\xi}{2(\cosh^2\xi - \sin^2\nu)} = \frac{e^{\pm 2\xi} + \cos(2\nu)}{2(\cosh^2\xi - \sin^2\nu)}, \\ q_i^{\pm}(\xi,\nu) &= \pm\frac{\sin 2\nu}{2(\cosh^2\xi - \sin^2\nu)}. \end{aligned} \tag{6.12}$$

The denominator in (6.11) is not equal to zero for all $\xi \in (-\infty,\infty)$ provided that $\nu \in (-\pi/2, \pi/2)$. It is easy to see that for $\xi \in (-\infty,\infty)$ we have

$$\begin{aligned} 0 \le q_r^{\pm}(\xi,\nu) &\le 2, \\ |q_i^{\pm}(\xi,\nu)| &\le |\tan\nu|, \end{aligned} \tag{6.13}$$

i.e., for each fixed t, ν and for $\xi \in (-\infty,\infty)$ the parametric curve $\Gamma_{\mathcal{A}}(t)$ given by (in the coordinates μ, η)

$$\begin{aligned} \mu &= \frac{t}{2}q_r^{-}(\xi,\nu), \\ \eta &= \frac{t}{2}q_i^{-}(\xi,\nu) \end{aligned} \tag{6.14}$$

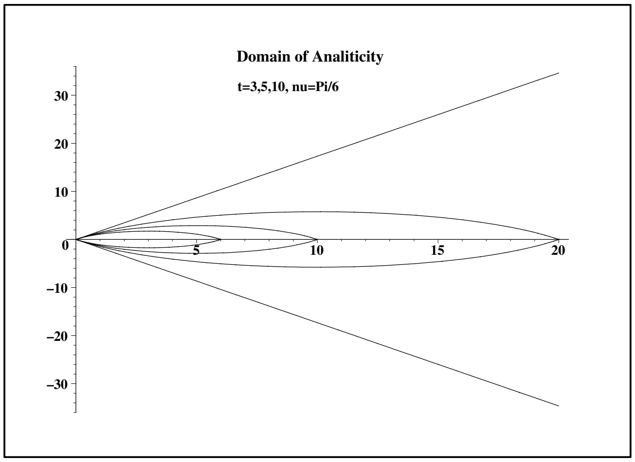


FIG. 6.2. *The domains of the analyticity of the integrand for* $t = 3, 5, 10,\ \nu = \pi/6$.

from (6.11) is closed and builds with the real axis at the origin the angle

$$\theta = \theta(\nu) = \arctan \left| \lim_{\xi \to \infty} q_i^-(\xi, \nu)/q_r^-(\xi, \nu) \right| = \arctan(\tan(2\nu)) = 2\nu. \tag{6.15}$$

For $\nu \in (-\pi/4, \pi/4)$ the domain $\mathcal{A}(t)$ inside of $\Gamma_{\mathcal{A}}(t)$ lies in the right half-plane, and for $t \to \infty$ fills the sector $\Sigma_f(\nu) = \{z = \rho e^{i\psi} :\ \rho \in (0, \infty),\ \psi \in (-\nu, \nu),\ \nu \in (0, \pi/4)\}$ (see Figure 6.2). Taking into account (4.14) we have

$$\begin{aligned} \|\mathcal{F}(t, \xi + i\nu)\| &\le \frac{t}{2(\cosh^2 \xi - \sin^2 \nu)} \\ &\times \left| \exp\left\{ -\frac{t[a_I \cosh(kh) - i b_I \sinh(kh)]}{2} \left[q_r^+ + i q_i^+\right] \right\} \right| \\ &\times \|f(t(1 - \tanh(\xi + i\nu))/2)\| \\ &\le \frac{t}{2(\cosh^2 \xi - \sin^2 \nu)} \\ &\times \exp\left\{ -\frac{t a_0 [\cosh(kh) \cos(\pi/4 + \varphi/2)(\cos(2\nu) + e^{2\xi})]}{2(\cosh^2 \xi - \sin^2 \nu)} \right. \\ &\left. -\frac{t a_0 [\sinh(kh) \cos(\pi/4 + \varphi/2) \sin(2\nu)]}{2(\cosh^2 \xi - \sin^2 \nu)} \right\} \\ &\times \|f(t(1 - \tanh(\xi + i\nu))/2)\| \end{aligned} \tag{6.16}$$

(note that $\nu \in (-\pi/2, \pi/2)$ provides that $\cosh^2 \xi - \sin^2 \nu > 0$ for all $\xi \in (-\infty, \infty)$). Since we suppose that

$$\|f(w)\| \le c e^{-\delta |\Re w|}, \qquad \delta > 0, \tag{6.17}$$

then by omitting the second summand in the argument of the exponential and replacing $\cosh(kh)$ by 1, we arrive at the inequality

$$\begin{aligned} \|\mathcal{F}(t, \xi + i\nu)\| &\le \frac{ct}{2(\cosh^2 \xi - \sin^2 \nu)} \\ &\times \exp\left\{ \frac{t[-\Delta e^{2\xi} - \delta e^{-2\xi}/2]}{2(\cosh^2 \xi - \sin^2 \nu)} \right\}, \end{aligned} \tag{6.18}$$

where

$$\frac{a_0}{2} \le \Delta = a_0 \frac{\cos(\varphi/2 + \pi/4)}{\cos\varphi} = \frac{a_0}{\sqrt{2}\sqrt{1+\sin\varphi}} \le \frac{a_0}{\sqrt{2}}. \tag{6.19}$$

Due to assumption $\delta \le \sqrt{2}a_0$ we have $\delta/2 \le \Delta$, and the last estimate yields

$$\|\mathcal{F}(t, \xi + i\nu)\| \le \frac{ct}{2(\cosh^2\xi - \sin^2\nu)} \exp\left\{ -\frac{t\delta\cosh(2\xi)}{2(\cosh^2\xi - \sin^2\nu)} \right\}. \tag{6.20}$$

Denoting $w = t\Delta\cosh(2\xi)/[2(\cosh^2\xi - \sin^2\nu)]$ and using (6.19) and the inequality $we^{-w} \le e^{-1}\ \forall\, w \ge 0$, we get

(6.21)
$$\begin{aligned}
&\int_{-\infty}^{\infty} \|\mathcal{F}(t, \xi + i\nu)\| d\xi \le \int_{-\infty}^{\infty} \frac{ct}{2(\cosh^2\xi - \sin^2\nu)} \exp\left\{ -\frac{t\delta\cosh(2\xi)}{2(\cosh^2\xi - \sin^2\nu)} \right\} d\xi \\
&= \int_{-\infty}^{\infty} \frac{1}{\Delta\cosh(2\xi)} we^{-w} d\xi \le \frac{c}{e\Delta} \int_{-\infty}^{\infty} \frac{1}{\cosh(2\xi)} d\xi \\
&\le \frac{2c}{e\Delta} \int_{-\infty}^{\infty} e^{-2|\xi|} d\xi = \frac{2c}{e\Delta} \le \frac{4c}{a_0 e}.
\end{aligned}$$

This estimate yields $\mathcal{F}_k(t,\xi) \in H^1(D_{d_1})$ with respect to ξ. □

The assumptions of Lemma 6.1 can be weakened if we consider problem (1.1) on some finite interval $(0, T]$.

LEMMA 6.2. *Let the right-hand side $f(t)$ in* (1.1) *for $t \in [0, T]$ be analytically extended into the domain $\mathcal{A}(T)$. Then the integrand $\mathcal{F}_k(t,\xi)$ can be analytically extended into the strip D_{d_1}, $0 < d_1 < \varphi/2$, and belongs to the class $H^1(D_{d_1})$ with respect to ξ.*

Proof. The proof is analogous to the proof of Lemma 6.1 but with constants depending on T. □

Let the assumptions of Lemma 6.1 hold. Then we can use the following quadrature rule to compute the integrals (6.7) (see [36], p. 144):

$$f_k(t) \approx f_{k,N}(t) = h \sum_{p=-N}^{N} \mu_{k,p}(t) f(\omega_p(t)), \tag{6.22}$$

where

$$\begin{aligned}
\mu_{k,p}(t) &= \frac{t}{2}\exp\{-\frac{t}{2}z(kh)[1 + \tanh(ph)]\}/\cosh^2(ph), \\
\omega_p(t) &= \frac{t}{2}[1 - \tanh(ph)], \qquad h = \mathcal{O}(1/\sqrt{N}), \\
z(\xi) &= a_I\cosh\xi - ib_I\sinh\xi.
\end{aligned} \tag{6.23}$$

Substituting (6.22) into (6.4) we get the following algorithm to compute an approach $u_{ap,N}(t)$ to $u_{ap}(t)$:

$$\begin{aligned}
u_{ap,N}(t) &= \frac{h}{2\pi i} \sum_{k=-N}^{N} z'(kh) \left[(z(kh)I - A)^{-1} - \frac{1}{z(kh)} I \right] \\
&\times h \sum_{p=-N}^{N} \mu_{k,p}(t) f(\omega_p(t)).
\end{aligned} \tag{6.24}$$

The next theorem characterizes the error of this algorithm.

THEOREM 6.3. *Let A be a densely defined strongly positive operator with the spectral characterization a_0, φ, and let the right-hand side $f(t) \in D(A^\alpha)$, $\alpha > 0$ for $t \in [0,\infty]$ be analytically extended into the sector $\Sigma_f = \{\rho e^{i\theta_1} : \rho \in [0,\infty], |\theta_1| < \varphi\}$ where the estimate*

$$\|A^\alpha f(w)\| \le c_\alpha e^{-\delta_\alpha |\Re w|}, \qquad w \in \Sigma_f \tag{6.25}$$

with $\delta_\alpha \in (0, \sqrt{2}a_0]$ holds. Then algorithm (6.24) converges with the error estimate

$$\|\mathcal{E}_N(t)\| = \|u_p(t) - u_{ap,N}(t)\| \le c e^{-c_1\sqrt{N}} \tag{6.26}$$

uniformly in t with positive constants c, c_1 depending on α, φ, a_0 and independent of N.

Proof. Let us denote

$$R_k(t) = f_k(t) - f_{k,N}(t). \tag{6.27}$$

Then we get for the error

$$\mathcal{E}_N(t) = u_p(t) - u_{ap,N}(t) = r_{1,N}(t) + r_{2,N}(t), \tag{6.28}$$

where

$$\begin{aligned} r_{1,N}(t) &= u_p(t) - u_{ap}(t), \\ r_{2,N}(t) &= u_{ap}(t) - u_{ap,N}(t). \end{aligned} \tag{6.29}$$

Using estimate (4.26) (see also Theorem 4.1) we get for $r_{1,N}(t)$ the estimate

$$\begin{aligned} &\|r_{1,N}(t)\| = \left\| \int_0^t \left\{ \frac{1}{2\pi i} \int_{-\infty}^{\infty} F_A(t-s,\xi)d\xi - \frac{h}{2\pi i} \sum_{k=-N}^{N} F_A(t-s,kh) \right\} f(s)ds \right\| \\ &\le \frac{c}{\alpha} \exp\left(-\sqrt{\frac{\pi d \alpha}{2}(N+1)} \right) \int_0^t \|A^\alpha f(s)\| ds, \end{aligned} \tag{6.30}$$

where $F_A(t,\xi)$ is the operator defined in (4.7). Due to (2.9) we have for the error $r_{2,N}(t)$

$$\begin{aligned} \|r_{2,N}(t)\| &= \left\| \frac{h}{2\pi i} \sum_{k=-N}^{N} z'(kh) \left[(z(kh)I - A)^{-1} - \frac{1}{z(kh)} I \right] R_k(t) \right\| \\ &\le \frac{h(1+M)K}{2\pi} \sum_{k=-N}^{N} \frac{|z'(kh)|}{|z(kh)|^{1+\alpha}} \|A^\alpha R_k(t)\|. \end{aligned} \tag{6.31}$$

The estimate (6.9) yields

$$\|A^\alpha \mathcal{F}(t,\xi)\| \le 2te^{-2|\xi|} \left\| A^\alpha f\left(\frac{t}{2}(1-\tanh \xi) \right) \right\|. \tag{6.32}$$

Due to Lemma 6.1 the assumption $\|A^{\alpha}f(w)\| \le c_{\alpha}e^{-\delta_{\alpha}|\Re w|}$ for all $w \in \Sigma_f$ guarantees that $A^{\alpha}f(w) \in H^1(D_{d_1})$ and $A^{\alpha}\mathcal{F}_k(t,w) \in H^1(D_{d_1})$. Then we are in the situation analogous to that of Theorem 3.2.1, p. 144, of [36] with $A^{\alpha}f(w)$ instead of f which implies

$$\begin{aligned}
&\|A^{\alpha}R_k(t)\| = \|A^{\alpha}(f_k(t) - f_{k,N}(t))\| \\
&= \left\| \int_{-\infty}^{\infty} A^{\alpha}\mathcal{F}_k(t,\xi)d\xi - h\sum_{k=-\infty}^{\infty} A^{\alpha}\mathcal{F}_k(t,kh)\right\| + \left\| h\sum_{|k|>N} A^{\alpha}\mathcal{F}_k(t,kh)\right\| \\
&\le \frac{e^{-\pi d_1/h}}{2\sinh(\pi d_1/h)} \|\mathcal{F}_k(t,w)\|_{H^1(D_{d_1})} \\
&+ h\sum_{|k|>N} 2te^{-2|kh|} \left\| A^{\alpha}f\left(\frac{t}{2}(1-\tanh kh)\right)\right\| \\
&\le ce^{-2\pi d_1/h}\|A^{\alpha}f(t,w)\|_{H^1(D_{d_1})} \\
&+ h\sum_{|k|>N} 2te^{-2|kh|}c_{\alpha}\mathrm{exp}\left\{-\delta_{\alpha}\frac{t}{2}(1-\tanh kh)\right\} \\
&\le ce^{-c_1\sqrt{N}},
\end{aligned} \tag{6.33}$$

where positive constants $c_{\alpha}, \delta_{\alpha}, c, c_1$ do not depend on $t,\ N,\ k$. Now, (6.31) takes the form

$$\begin{aligned}
\|r_{2,N}(t)\| &= \frac{h}{2\pi i}\sum_{k=-N}^{N} z'(kh)\left[(z(kh)I - A)^{-1} - \frac{1}{z(kh)}I\right] R_k(t) \\
&\le ce^{-c_1\sqrt{N}} S_N
\end{aligned} \tag{6.34}$$

with $S_N = \sum_{k=-N}^{N} h\frac{|z'(kh)|}{|z(kh)|^{1+\alpha}}$. Using the estimate (4.8) and

$$\begin{aligned}
|z(kh)| &= \sqrt{a_I^2\cosh^2(kh) + b_I^2\sinh^2(kh)} \\
&\ge a_I\cosh(kh) \ge a_I e^{|kh|}/2,
\end{aligned} \tag{6.35}$$

the last sum can be estimated by

$$|S_N| \le \frac{c}{\sqrt{N}}\sum_{k=-N}^{N} e^{-\alpha|k/\sqrt{N}|} \le c\int_{-\sqrt{N}}^{\sqrt{N}} e^{-\alpha t}dt \le c/\alpha. \tag{6.36}$$

Taking into account (6.33) and (6.36) we get from (6.34)

$$\|r_{2,N}(t)\| \le ce^{-c_1\sqrt{N}}. \tag{6.37}$$

The assertion of the theorem follows now from (6.28) and (6.30). □

Example 6.1. We consider the inhomogeneous problem (1.1) with the operator A defined by

$$\begin{aligned}
&D(A) = \{u(x) \in H^2(0,1):\ u(0) = u(1) = 0\}, \\
&Au = -u''(x) \qquad \forall u \in D(A).
\end{aligned} \tag{6.38}$$

The initial function is $u_0 = u(0,x) = 0$, and the right-hand side $f(t)$ is given by

$$f(t,x) = x^3(1-x)^3\frac{1-t^2}{(1+t^2)^2} - \frac{6t}{1+t^2}x(1-x)(5x^2-5x+1). \tag{6.39}$$

It is easy to see that the exact solution is $u(t,x) = x^3(1-x)^3\frac{t}{1+t^2}$. The algorithm (6.24) was implemented for $t = 1,\ x = 1/2$ in Maple 8 with Digits=16. Table 3 shows an exponential decay of the error $\varepsilon_N = |u(1,1/2) - u_{ap,N}(1)|$ with growing N.

TABLE 3
The error of algorithm (6.24) *for* $t = 0,\ x = 1/2$.

N	ε_N
8	0.485604499
16	0.184497471
32	0.332658314 e-1
64	0.196729786 e-2
128	0.236757688 e-4
256	0.298766899 e-7

Acknowledgments. The authors would like to acknowledge the support provided by the Deutsche Forschungsgemeinschaft (DFG). The valuable remarks of the anonymous referees are also gratefully acknowledged.

REFERENCES

[1] F. Stenger, B. Barkey, and R. Vakili, *About Sinc-approximations*, in Proceedings of Computation and Control III, K. Bowers and J. Lund, eds., Birkhäuser, Basel, Switzerland, 1993.

[2] D. Z. Arov, I. P. Gavrilyuk, and V. L. Makarov, *Representation and approximation of the solution of an initial value problem for a first order differential equation with an unbounded operator coefficient in Hilbert space based on the Cayley transform*, in Progress in Partial Differential Equations, C. Bandle et al., eds., Pitman Res. Notes Math. Sci. 1, C. Bandle et al., eds., Pitman, Boston, 1994, pp. 40–50.

[3] A. Ashyralyev and P. Sobolevskii, *Well-Posedness of Parabolic Difference Equations*, Birkhäuser, Basel, Switzerland, 1994.

[4] N. Yu. Bakaev, *Stability estimates for a general discretization method*, Soviet Math. Dokl., 40 (1990), pp. 11–15.

[5] N. Yu. Bakaev, *Maximum norm resolvent estimates for elliptic finite element operators*, BIT, 41 (2001), pp. 215–239.

[6] N. Yu. Bakaev, S. Larsson, and V. Thomée, *Long time behaviour of backward difference type methods for parabolic equations with memory in Banach space*, Numer. Math., 6 (1998), pp. 185–206.

[7] N. Yu. Bakaev and A. Ostermann, *Long time stability of variable stepsize approximations of semigroups*, Math. Comp., 71 (2002), pp. 1575–1567.

[8] T. Carleman, *Über die asymptotische verteilung der eigenwerte partieller differentialgleichungen*, Ber. der Sächs. Akad. Wiss. Leipzig, Math.-Nat. Kl., 88 (1936), pp. 119–132.

[9] M. Crouzeix, S. Larsson, S. Piskarev, and V. Thomée, *The stability of rational approximations of analytic semigroups*, BIT, 33 (1993), pp. 74–84.

[10] H. Fujita, N. Saito, and T. Suzuki, *Operator Theory and Numerical Methods*, Elsevier, Heidelberg, 2001.

[11] I. P. Gavrilyuk, *Strongly P-positive operators and explicit representation of the solutions of initial value problems for second order differential equations in Banach space*, J. Math. Anal. Appl., 236 (1999), pp. 327–349.

[12] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *$\mathcal{H}$-matrix approximation for the operator exponential with applications*, Numer. Math., 92 (2002), pp. 83–111.

[13] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Data-sparse approximation to the operator-valued functions of elliptic operator*, Math. Comp., 73 (2004), pp. 1297–1324.

[14] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Tensor-product approximation to elliptic and parabolic solution operators in higher dimensions*, Computing, 74 (2005), pp. 131–157.

[15] I. P. Gavrilyuk, W. Hackbusch, and B. N. Khoromskij, *Data-sparse approximation of a class of operator-valued functions*, Math. Comp., 74 (2005), pp. 681–708.

[16] I. P. Gavrilyuk and V. L. Makarov, *The Cayley transform and the solution of an initial value problem for a first order differential equation with an unbounded operator coefficient in Hilbert space*, Numer. Funct. Anal. Optim., 15 (1994), pp. 583–598.

[17] I. P. Gavrilyuk and V. L. Makarov, *Representation and approximation of the solution of an initial value problem for a first order differential equation in Banach space*, Z. Anal. Anwendungen, 15 (1996), pp. 495–527.

[18] I. P. Gavrilyuk and V. L. Makarov, *Exponentially convergent parallel discretization methods for the first order evolution equations*, Comput. Methods Appl. Math., 1 (2001), pp. 333–355.

[19] I. P. Gavrilyuk and V. L. Makarov, *Exponentially convergent parallel discretization methods for the first order differential equations*, Dokl. Ukrainian Acad. Sci., (2002), pp. 1–6.

[20] I. P. Gavrilyuk and V. L. Makarov, *Exponentially Convergent Algorithms for the Operator Exponential with Applications to Inhomogeneous Problems in Banach Spaces*, Jenaer Schriften zur Mathematik und Informatik, FSU Jena, 4 (2004), pp. 1–34 (available online at http://www.minet.uni-jena.de/Math-Net/reports/).

[21] I. P. Gavrilyuk and V. L. Makarov, *Algorithms without accuracy saturation for evolution equations in Hilbert and Banach spaces*, Math. Comp., 74 (2005), pp. 555–583.

[22] I. P. Gavrilyuk and V. L. Makarov, *An explicit boundary integral representation of the solution of the two-dimensional heat equation and its discretization*, J. Integral Equations Appl., 12 (2000), pp. 63–83.

[23] I. P. Gavrilyuk, V. L. Makarov, and V. Vasylyk, *A new estimate of the sinc method for linear parabolic problems including the initial point*, Comput. Methods Appl. Math., 4 (2004), pp. 1–27.

[24] C. González and C. Palencia, *Stability of time-stepping methods for abstract time-dependent parabolic problems*, SIAM J. Numer. Anal., 35 (1998), pp. 973–989.

[25] C. González and C. Palencia, *Stability of Runge-Kutta methods for abstract time-dependent parabolic problems: The Hölder case*, Math. Comp., 68 (1996), pp. 73–89.

[26] D. Guidetti, B. Karasozen, and S. I. Piskarev, *Approximation of Abstract Differential Equations*, Technical report, Moscow State University, Moscow, 2003; also available at http://ma1serv.mathematik.uni-karlsruhe.de/evolve-l/index.html and http://www.srcc.msu.su/nivc/english/about/home_pages/piskarev/obz1en.pdf.

[27] M. A. Krasnosel'skij and P. E. Sobolevskij, *Fractional powers of operators acting in Banach spaces*, Dokl. Akad. Nauk, 129 (1959), pp. 499–502 (in Russian).

[28] C. Palencia, *Maximum norm analysis of completely discrete finite element methods for parabolic problems*, SIAM J. Numer. Anal., 33 (1996), pp. 1654–1668.

[29] A. Pazy, *Semigroups of Linear Operator and Applications to Partial Differential Equations*, Springer-Verlag, New York, Berlin, Heidelberg, 1983.

[30] S. I. Piskarev, *Error estimates in the approximation of semigroups of operators by Padé fractions*, Izv. Vyssh. Uchebn. Zaved. Mat., 4 (1979), pp. 33–38 (in Russian).

[31] A. A. Samarskii, I. P. Gavrilyuk, and V. L. Makarov, *Stability and regularization of three-level difference schemes with unbounded operator coefficients in Banach spaces*, SIAM J. Numer. Anal., 39 (2001), pp. 708–723.

[32] A. H. Schatz, V. Thomée, and L. B. Wahlbin, *Stability, analyticity and almost best approximation in maximum norm for parabolic finite element equations*, Comm. Pure Appl. Math., 51 (1998), pp. 1349–1385.

[33] D. Sheen, I. H. Sloan, and V. Thomée, *A parallel method for time-discretization of parabolic problems based on contour integral representation and quadrature*, Math. Comp., 69 (2000), pp. 177–195.

[34] D. Sheen, I. H. Sloan, and V. Thomée, *A parallel method for time-discretization of parabolic equations based on Laplace transformation and quadrature*, IMA J. Numer. Anal., 23 (2003), pp. 269–299.

[35] M. Z. Solomjak, *Application of the semi-group theory to investigation of differential equations in Banach spaces*, Dokl. Akad. Nauk, 122 (1958), pp. 766–769 (in Russian).

[36] F. Stenger, *Numerical Methods Based on Sinc and Analytic Functions*, Springer-Verlag, New York, Berlin, Heidelberg, 1993.

[37] F. Stenger, *Collocating convolutions*, Math. Comp., 64 (1995), pp. 211–235.

[38] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer-Verlag, New York, Berlin, 1997.

[39] V. V. Vasil'ev and S. I. Piskarev, *Differential Equations in Banach Spaces* I. *Semigroup Theory*, Moscow State University Publishing House, Moscow, 1996 (in Russian).

[40] V. V. Vasil'ev and S. I. Piskarev, *Bibliography on Differential Equations in Abstract Spaces*, Moscow State University, Moscow, 2001; also available at http://www.srcc.msu.su/num_anal/list_wrk/page_5u.htm.

[41] V. V. Vasil'ev and S. I. Piskarev, *Differential Equations in Banach Spaces* II. *Theory of Cosine Operator Functions*, Moscow State University, Moscow, 2003; also available at http://www.srcc.msu.su/nivc/english/about/home_pages/piskarev/perek2.htm.

[42] V. Vasylyk, *Uniform exponentially convergent method for the first order evolution equation with unbounded operator coefficient*, J. Numer. Appl. Math., 1 (2003), pp. 99–104 (in Russian).

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2172–2199

A SECOND-ORDER MAXIMUM PRINCIPLE PRESERVING FINITE VOLUME METHOD FOR STEADY CONVECTION-DIFFUSION PROBLEMS*

ENRICO BERTOLAZZI† AND GIANMARCO MANZINI‡

Abstract. A cell-centered finite volume method is proposed to approximate numerically the solution to the steady convection-diffusion equation on unstructured meshes of d-simplexes, where $d \geq 2$ is the spatial dimension. The method is formally second-order accurate by means of a piecewise linear reconstruction within each cell and at mesh vertices. An algorithm is provided to calculate nonnegative and bounded weights. Face gradients, required to discretize the diffusive fluxes, are defined by a nonlinear strategy that allows us to demonstrate the existence of a maximum principle. Finally, a set of numerical results documents the performance of the method in treating problems with internal layers and solutions with strong gradients.

Key words. unstructured grids, finite volume methods, maximum principles, M-matrices

AMS subject classifications. 35J25, 76M12, 65N06, 35B50, 65N12

DOI. 10.1137/040607071

1. Introduction.

We are concerned with the finite volume approximation of the steady convection-diffusion boundary value problem:

Find a function $\mathtt{u}$ *satisfying*

$$\operatorname{div}(\boldsymbol{v}\mathtt{u} - \nu\nabla\mathtt{u}) = \mathtt{s} \qquad \text{in } \Omega, \tag{1a}$$

$$\mathtt{u} = \mathtt{g} \qquad \text{on } \Gamma, \tag{1b}$$

where $\Omega \in \mathbb{R}^d$, $d \geq 2$, is a polyhedral domain with boundary Γ. We assume that the unit vector almost everywhere orthogonal to Γ, denoted by $\boldsymbol{n}$, is always oriented outward of Ω. The model problem of (1a)–(1b) describes the advective transport of the scalar quantity $\mathtt{u}(\boldsymbol{x})$ by the velocity field $\boldsymbol{v}(\boldsymbol{x})$ and the diffusion process driven by the scalar viscosity field $\nu(\boldsymbol{x})$. A forcing term can be present on the right-hand side of (1a), namely, $\mathtt{s}(\boldsymbol{x})$. Dirichlet boundary conditions are set on Γ by using the scalar field $\mathtt{g}(\boldsymbol{x})$. Let the fields $\boldsymbol{v}(\boldsymbol{x})$, $\nu(\boldsymbol{x})$, $\mathtt{s}(\boldsymbol{x})$, and $\mathtt{g}(\boldsymbol{x})$ satisfy the constraints listed below:

$$\begin{array}{llll} & \text{(i)} & \nu(\boldsymbol{x}) \geq \beta > 0, & \nu \in \mathtt{C}^1(\overline{\Omega}); \\ & \text{(ii)} & \operatorname{div} \boldsymbol{v} \geq 0, & \boldsymbol{v} \in \mathtt{C}^1(\overline{\Omega})^d; \\ (2) & \text{(iii)} & \mathtt{s} \in \mathtt{L}^2(\Omega); & \\ & \text{(iv)} & \mathtt{g} \in \mathtt{H}^{1/2}(\Gamma) \bigcap \mathtt{C}(\Gamma) & \end{array}$$

for a suitable real constant β, and where

$$\mathtt{H}^m(\Omega) = \left\{\mathtt{u} \in \mathtt{L}^2(\Omega) \,:\, D^\alpha\mathtt{u} \in \mathtt{L}^2(\Omega),\, |\alpha| \leq m\right\},$$

*Received by the editors April 21, 2004; accepted for publication (in revised form) March 21, 2005; published electronically December 30, 2005.
http://www.siam.org/journals/sinum/43-5/60707.html

†Dip. Ingegneria Meccanica e Strutturale, Università di Trento, via Mesiano 77, I-38050 Trento, Italy (Enrico.Bertolazzi@ing.unitn.it).

‡Istituto di Matematica Applicata e Tecnologie Informatiche (IMATI)—CNR, via Ferrata 1, I-37100 Pavia, Italy (Marco.Manzini@ian.pv.cnr.it).

for $m \geq 0$, is the standard notation of a Sobolev space [1]. Under conditions (2) a weak reformulation of problem (1a)–(1b) is possible in terms of the $\mathtt{H}^1$-coercive bilinear form:

$$a(\mathsf{u}, \mathsf{v}) = \int_\Omega (\nu \nabla \mathsf{u} - \mathsf{u}\boldsymbol{v}) \cdot \nabla \mathsf{v}\, dV.$$

In view of the Lax–Milgram lemma, the weak problem has a unique solution in $\mathtt{H}^1(\Omega)$ [1].

Following Stampacchia [29], a function $\mathsf{u} \in \mathtt{H}^1(\Omega)$ is said to be *superelliptic* if $a(\mathsf{u}, \mathsf{v}) \leq 0$ for any $\mathsf{v} \in \mathtt{C}_0^\infty(\Omega)$ and $\mathsf{v} \geq 0$, where $\mathtt{C}_0^\infty(\Omega)$ is the space of infinitely differentiable functions with compact support in Ω. Under the above-stated regularity assumptions, a *maximum principle* for superelliptic weak solutions exists. If u is a superelliptic solution and is such that $\operatorname{ess\,sup}_{\boldsymbol{x} \in \Gamma} \mathsf{u}(\boldsymbol{x}) \leq k$, then $\operatorname{ess\,sup}_{\boldsymbol{x} \in \Omega} \mathsf{u}(\boldsymbol{x}) \leq \max\{0, k\}$; see [15]. A *minimum principle* can be stated as well by introducing the notion of *subelliptic* solution, which is a function $\mathsf{u} \in \mathtt{H}^1(\Omega)$ that satisfies the condition $a(\mathsf{u}, \mathsf{v}) \geq 0$ for any $\mathsf{v} \in \mathtt{C}_0^\infty(\Omega)$ and $\mathsf{v} \geq 0$. Then if u is a subelliptic weak solution and is such that $\operatorname{ess\,inf}_{\boldsymbol{x} \in \Gamma} \mathsf{u}(\boldsymbol{x}) \geq k$, then $\operatorname{ess\,inf}_{\boldsymbol{x} \in \Omega} \mathsf{u}(\boldsymbol{x}) \geq \min(0, k)$; see again [15].

From the analytical viewpoint, maximum and minimum principles are quite general but very significant due to their physical implications. For example, if u attains physically meaningful values in a specific range of real numbers, and is suitably bounded on Γ, then u must attain values in the same range (almost) everywhere in the interior of Ω. From the numerical viewpoint, it is widely recognized that maximum and minimum principles provide a valuable tool in proving solvability results (existence and uniqueness of discrete solutions), enforcing numerical stability, and deriving convergence results (a priori error estimates) for the sequence of approximate solutions; see [26]. Recent papers investigating the existence of maximum principles in discretization schemes for partial differential equations are cited in [9]. Other papers pertinent to the issue of discrete maximum principle preserving methods are found in [21, 22, 32].

A literature review of the many finite volume methods that ensure a maximum principle is beyond the scope of this paper. For this purpose, we refer the reader to the general introductions found, for example, in [13, 26] and to the references therein. However, we wish to emphasize a rather important fact that we realized after a careful and systematic inspection of the cell-centered finite volume methods that are available in the literature and capable of preserving a discrete maximum principle. Most of these methods are designed to show that the method gives rise to a *monotone matrix* or an *M-matrix* on uniform (or quasi-uniform) grids and their accuracy in approximating cell averages degenerates to first order on general unstructured grids. Let us recall that a monotone matrix is a nonsingular matrix with nonnegative inverse; a nonsingular M-matrix is a monotone matrix having nonpositive off-diagonal entries [2]. If a difference scheme is described by a monotone matrix or an M-matrix operator, a stability condition which is equivalent to a discrete maximum principle can be easily derived for the approximate solution. For example, we mention the schemes in [14, 18] for the steady model, and the ones proposed in [23, 31] to discretize the spatial terms of the time-dependent model.

As a matter of fact and against intuition, the major difficulty in obtaining second-order accurate schemes based on monotone or M-matrices is related to the discretization of the diffusive term and not of the advective term, the latter being properly controlled by limiters; see [6, 7, 19]. Owing to the negative result of [20], it is possible to show by appropriate counterexamples that *no finite difference approximation* *of second-order partial derivatives on general unstructured grids that is both linear and second- or higher-order accurate can be based on an M-matrix.* For instance, it is an enlightening fact that the second-order accurate cell-centered approximations developed in [31] satisfy a discrete maximum principle but only on uniform (or quasi-uniform) grids.

It is worth mentioning that the difficulty of obtaining M-matrices has also been related to the one of controlling the positivity and boundedness of the coefficients in the scheme required to ensure monotonicity [17]. A technique that surely suffers from this major fault is the *diamond scheme*; see [13] for a recent literature review. This approach defines the face gradients that are needed to discretize the numerical diffusive flux from the approximate cell averages. The solution gradient is provided at each internal face by a piecewise constant reconstruction of Gauss–Green type. This reconstruction takes place on the convex hull of the face vertices and the centers of the cells adjacent to the face. It is well known that this technique cannot ensure the correct sign of the coefficients in the scheme on general unstructured grids [11].

In this work, we propose the design of a finite volume approximation to the solution of (1a)–(1b) that is based on a *nonlinear extension* of the diamond scheme and that simultaneously satisfies these three conditions:

- it is based on a *conservative cell-centered formulation*;
- it provides *second order of accuracy in approximating cell averages of the solution on general unstructured grids*;
- it ensures a *maximum principle in some discrete form for the approximate solution.*

We emphasize that the nonlinearity is the crucial issue of the design of the scheme; this makes it possible to bypass the negative result of [20] and prove the existence of a discrete maximum principle for the numerical solution. Throughout the paper, we will refer to the standard technique as the *diamond scheme* and to this new method as the *nonlinear diamond scheme.*

The key steps of the derivation of the method are the following. First, (1a) is reformulated in the integral-conservative way on an unstructured mesh of d-dimensional simplices (triangles for $d = 2$ and tetrahedra for $d = 3$). As usual in finite volume methods, we relate the approximation of the solution-average on any mesh control volume to the discrete balance of the numerical advective and diffusive flux on the control volume boundary. The definition of the numerical flux uses the approximate cell averages, the values recovered at the mesh vertices by a piecewise linear reconstruction process, and the boundary data of (1b). The existence of a suitable set of nonnegative and bounded coefficients for the recovery of the vertex value is theoretically demonstrated by a constructive proof that provides an algorithm to compute them.

The numerical advective flux implements the usual first-order upstream formula, and second order of accuracy is formally achieved by the cellwise linear reconstruction from cell averages proposed and analyzed in [5].

The design of the numerical diffusive flux deserves much more attention. As pointed out in [5, 24], the face gradient provided by the diamond scheme can be written as the weighted average of two independent one-sided face gradients. Each one-sided gradient is calculated by linearly interpolating the values at the face vertices and the cell averages on the adjacent cells. The weights of the diamond scheme are constant and taken to be proportional to the measure of the portion of diamond cell shared by the corresponding adjacent cell.

In the nonlinear diamond scheme, the face gradient at any internal face is reformulated as a *nonlinear average* of the one-sided gradients *by suitably designing* *solution-dependent weights.* Under very general assumptions on the regularity of the mesh, we demonstrate that there always exists a solution to the discrete nonlinear problem and that all the numerical solutions (if more than one exist) satisfy a discrete maximum (or minimum) principle.

A similar nonlinear design was originally considered in [3] to build nonlinear M-matrix operators by relaxing conservation in difference approximations. These difference methods were also shown to possess second-order accurate solutions that preserve a discrete maximum principle.

The preservation of the maximum principle of the nonlinear diamond scheme solution has been experimentally verified by a comparative assessment with the behavior of the diamond scheme solution in the same critical situations. Numerical experiments also show the expected second-order convergence rate and a nonoscillatory behavior when the analytical solution has strong gradient regions such as internal layers.

The outline of the paper is as follows. In section 2, we introduce the notation adopted in this paper and other technical details. In section 3, we discuss the regularity assumption on the family of grids considered in the approximation process when the mesh size tends to zero. We also present the algorithm that recovers the vertex values from the cell averages. Then the derivation of the method is presented in section 4. In section 5, we investigate the theoretical properties concerning the solvability of the scheme and demonstrate the existence of a discrete maximum principle for the numerical solutions provided by the method. In section 6, a set of numerical results illustrates the performance of the method in treating problems with internal layers and solutions with strong gradients. Final remarks and conclusions are offered in section 7.

2. General setup and notation.

In this section, we introduce the notation adopted in this paper. For ease of reference, we also collect herein the definitions of the topological and geometrical entities and of the discrete function spaces, scalar product, and norms that are in use throughout the paper.

The polyhedral domain $\Omega \in \mathbb{R}^d$ is covered by a finite collection of nonoverlapping and nonempty d-dimensional simplices, namely, the *mesh*. These simplices are denoted by the letter "T" and labeled by a Latin index like i (j, k, ...); i.e., T_i is the ith control volume (cell) of the mesh. The set of all mesh control volumes is denoted by $\mathcal{T}_h = \{\mathsf{T}_i\}$; the control volumes are such that $\overline{\Omega} = \cup_{\mathsf{T}_i \in \mathcal{T}_h} \mathsf{T}_i$.

The mesh faces are denoted by the letter "f" and labeled by a couple of Latin indices, i.e., f_{ij}. It is useful to distinguish between *internal* and *boundary* faces. When f_{ij} is an internal face, there must exist two control volumes T_i and T_j such that $\mathsf{f}_{ij} = \mathsf{T}_i \cap \mathsf{T}_j$. When f_{ij} is a boundary face, i.e., $\mathsf{f}_{ij} \subseteq \Gamma$, the first index always refers to the unique control volume T_i to which the face belongs, while the second index is defined in accordance with a suitable boundary numbering system (such as a sort of fictitious *ghost cell*). The symbols $\mathcal{F}_h$, $\mathcal{F}_h^{\text{int}}$, and $\mathcal{F}_h^{\text{bnd}}$ denote, respectively, the set of all mesh faces, the set of the *internal faces*, and the set of the *boundary faces*. When dealing with internal faces, f_{ij} and f_{ji} are equivalent symbols that denote the same face; in expressions like $\mathsf{f}_{ij} \in \mathcal{F}_h$ (or $\mathcal{F}_h^{\text{int}}$) we assume that the face labeled by i and j is considered only once (for example, by taking the representative with $i < j$). Clearly, $\mathcal{F}_h = \mathcal{F}_h^{\text{int}} \cup \mathcal{F}_h^{\text{bnd}}$ and $\mathcal{F}_h^{\text{int}} \cap \mathcal{F}_h^{\text{bnd}} = \emptyset$.

The mesh vertices are denoted by the symbol "v" and labeled by Greek letters like α (β, γ, ...). The symbols $\mathcal{V}_h$, $\mathcal{V}_h^{\text{int}}$, and $\mathcal{V}_h^{\text{bnd}}$ denote, respectively, the set of all mesh vertices, the set of the *internal* vertices, and the set of the *boundary* vertices. We have that $\mathcal{V}_h = \mathcal{V}_h^{\text{int}} \cup \mathcal{V}_h^{\text{bnd}}$ and $\mathcal{V}_h^{\text{int}} \cap \mathcal{V}_h^{\text{bnd}} = \emptyset$.

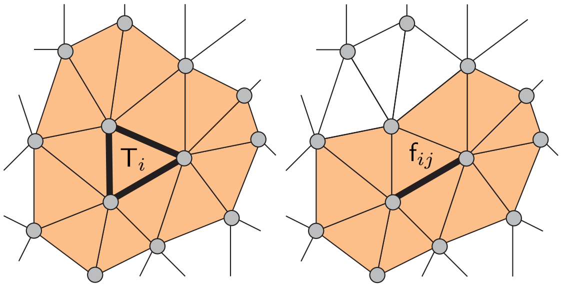


FIG. 1. *The stencils $\sigma\nu_i$ (left) and $\sigma\nu_{ij}$ (right) for the two-dimensional case.*

2.1. Topological quantities. Summations are taken over the index sets:

- σ_i, cells sharing a face with T_i;
- σ_α, cells surrounding the vertex v_α;
- σ'_i, "ghost" cells sharing a face with T_i;
- ν_i, vertices of the cell T_i;
- ν_α, vertices connected to the vertex v_α;
- ν_{ij}, vertices of the face f_{ij}; we also distinguish between $\nu_{ij}^{\rm int} = \{\alpha \in \nu_{ij} | \mathsf{v}_\alpha \in \mathcal{V}_h^{\rm int}\}$ and $\nu_{ij}^{\rm bnd} = \{\alpha \in \nu_{ij} | \mathsf{v}_\alpha \in \mathcal{V}_h^{\rm bnd}\}$;
- $\sigma\nu_i$, cells having a vertex in common with T_i;
- $\sigma\nu_{ij}$, cells having a vertex in common with f_{ij}.

We anticipate that the last two sets listed above, i.e., $\sigma\nu_i$ and $\sigma\nu_{ij}$, are the stencils of the discrete gradient $\boldsymbol{G}_i(\mathsf{u}_h)$ and of the numerical diffusive flux $\mathcal{G}_{ij}(\mathsf{u}_h)$ at $\mathsf{f}_{ij} \in \mathcal{F}_h$; see the definitions in section 4.2. Figure 1 illustrates the two-dimensional case.

Coherently with face notation, f_{ij} is a boundary face of T_i for every $j \in \sigma'_i$ and the internal face shared by T_i and T_j for $j \in \sigma_i$. Thus, the index $j \in \sigma_i \cup \sigma'_i$ labels all the faces forming the boundary of T_i.

2.2. Geometric quantities. The quantities related to T_i are consistently labeled by the same index i:

- $|\mathsf{T}_i|$, the d-dimensional Lebesgue measure of T_i (area in two dimensions, volume in three);
- $\partial\mathsf{T}_i$, the boundary of T_i;
- $\boldsymbol{x}_i$, the barycenter of T_i.

The quantities related to the mesh vertex v_α are consistently labeled by the same index α:

- $\boldsymbol{x}_\alpha$, the position vector of v_α;
- $\mathcal{B}_{\boldsymbol{x}_\alpha,r}$, the (closed) ball of center v_α and radius r; $\partial\mathcal{B}_{\boldsymbol{x}_\alpha,r}$ denotes the boundary of the ball.

The quantities related to the face f_{ij} are consistently labeled by the same couple of indices ij:

- $|\mathsf{f}_{ij}|$, the $(d-1)$-dimensional Lebesgue measure of $\mathsf{f}_{ij} \in \mathcal{F}_h$ (length in two dimensions, surface area in three);
- $\boldsymbol{x}_{ij}$, the position vector of the center of $\mathsf{f}_{ij} \in \mathcal{F}_h$;
- $\boldsymbol{n}_{ij}$, the unit vector orthogonal to $\mathsf{f}_{ij} \in \mathcal{F}_h$ and oriented from T_i to T_j if $\mathsf{f}_{ij} \in \mathcal{F}_h^{\rm int}$, outward of Ω if $\mathsf{f}_{ij} \in \mathcal{F}_h^{\rm bnd}$;
- $\tilde{\boldsymbol{x}}_{ij}$, the position vector of the orthogonal projection of $\boldsymbol{x}_i$ on the hyperplane containing the face $\mathsf{f}_{ij} \in \mathcal{F}_h$;

- $\tilde{\lambda}^{ij}_{\alpha}$, the barycentric coordinate of $\tilde{\boldsymbol{x}}_{ij}$ with respect to $\mathsf{v}_{\alpha} \in \mathsf{f}_{ij}$; hence,

$$\sum_{\alpha \in \nu_{ij}} \tilde{\lambda}^{ij}_{\alpha} = 1 \quad \text{and} \quad \sum_{\alpha \in \nu_{ij}} \tilde{\lambda}^{ij}_{\alpha} \boldsymbol{x}_{\alpha} = \tilde{\boldsymbol{x}}_{ij};$$

- $h_{ij} = (\tilde{\boldsymbol{x}}_{ij} - \boldsymbol{x}_i) \cdot \boldsymbol{n}_{ij}$, the distance between the center of T_i and the face $\mathsf{f}_{ij} \in \mathcal{F}_h$;
- $H_{ij} = (\boldsymbol{x}_j - \boldsymbol{x}_i) \cdot \boldsymbol{n}_{ij} = h_{ij} + h_{ji}$, the *effective distance* between the centers of T_i and T_j when $\mathsf{f}_{ij} \in \mathcal{F}^{\text{int}}_h$.

2.3. Discrete function spaces and norms. Let us introduce the set of piecewise constant functions,

$$\mathtt{T}_h = \Big\{ \mathsf{w}_h \in \mathtt{L}^2(\Omega), \text{ such that } \mathsf{w}_h|_{\mathsf{T}_i} = w_i \text{ for } \mathsf{T}_i \in \mathcal{T}_h \Big\},$$

that are defined on a given triangulation $\mathcal{T}_h$. The set $\mathtt{T}_h$ is clearly isomorphic to $\mathbb{R}^{\mathsf{card}\{\mathcal{T}_h\}}$. In this paper, we use the $\mathtt{L}^2$ mesh-dependent scalar product, its derived norm, and the $\mathtt{H}^1$-norm defined on $\mathtt{T}_h$ by the formulae

$$\begin{aligned}
(\mathsf{u}_h, \mathsf{w}_h)_{\mathcal{T}_h} &= \sum_{\mathsf{T}_i \in \mathcal{T}_h} |\mathsf{T}_i|\, u_i w_i, \\
\|\mathsf{w}_h\|_{\mathcal{T}_h} &= \sqrt{(\mathsf{w}_h, \mathsf{w}_h)_{\mathcal{T}_h}}, \\
\|\mathsf{w}_h\|_{\mathcal{T}_h,1} &= \Bigg(\sum_{\mathsf{f}_{ij} \in \mathcal{F}^{\text{int}}_h} \frac{|\mathsf{f}_{ij}|}{H_{ij}} (w_j - w_i)^2 + \sum_{\mathsf{f}_{ij} \in \mathcal{F}^{\text{bnd}}_h} \frac{|\mathsf{f}_{ij}|}{h_{ij}} w_i^2 \Bigg)^{1/2}.
\end{aligned}$$

3. Mesh regularity assumption and vertex reconstruction. In this section, we discuss the sense in which the mesh used to formulate the finite volume scheme is regular. We also describe the algorithm to recover the approximate vertex values.

3.1. Mesh regularity assumption. In accordance with the definition of [10], the parameter h that labels the mesh $\mathcal{T}_h$ is called the *mesh size*, and is formally given by the supremum of the mesh control volume diameters; i.e., $h = \max_{\mathsf{T}_i \in \mathcal{T}_h} h_i$ with $h_i = \mathsf{diam}\{\mathsf{T}_i\}$. Let us denote the maximum radius of the balls contained in the cell T_i and centered at $\boldsymbol{x}_i$ by ρ_i, and consider $\rho = \min_{\mathsf{T}_i \in \mathcal{T}_h} \rho_i$. The approximation method described in this paper is formulated on a family of d-dimensional conforming grids $\mathcal{T}_h$ that are *regular* in the following sense.

Assumption 1 (mesh regularity).

(i) There exists a *mesh regularity constant* $C_{\text{reg}} > 0$ that is independent of h and such that $(h/\rho)^d \leq C_{\text{reg}}$ for any $h \leq h_0$.

(ii) Let $\pi_{ij}\mathsf{v}_{\alpha}$ denote the orthogonal projection of the vertex v_{α} on the hyperplane containing the face f_{ij}. Then $\pi_{ij}\mathsf{v}_{\alpha} \in \mathsf{f}_{ij}$ for any $\alpha \in \nu_i$ and $j \in \sigma_i \cup \sigma'_i$.

(iii) The face f_{ij} is internal, i.e., $\mathsf{f}_{ij} \in \mathcal{F}^{\text{int}}_h$ iff $\mathsf{card}\big\{\nu^{\text{int}}_{ij}\big\} > 0$.

Note that the first condition is similar to but slightly stronger than the one that is normally met in the analysis of finite element methods; see [10]. The present formulation of item (i) is particularly useful to demonstrate the existence of the positive weights of the vertex reconstruction in the next subsection. A weaker assumption could be considered as well by imposing a local regularity constraint on h_j/ρ_j and taking the supremum on the set of cells surrounding any mesh vertex. Nonetheless, this would complicate the analysis that follows. The second condition is trivial when $\mathsf{v}_\alpha \in \mathsf{f}_{ij}$ because in this case $\pi_{ij}\mathsf{v}_\alpha = \mathsf{v}_\alpha$, and is satisfied by a wide family of grids, such as, for example, the ones based on regular acute and weakly acute two-dimensional triangulations and their suitable extensions to higher dimensions [10].

For the sake of reference, in the following proposition we list some properties of regular grids that will be useful in the paper.

Proposition 2.

(i) $\max\{h_{ij}/h_{ji},\, h_{ji}/h_{ij}\} \leq C_{\mathrm{reg}}$ *for any internal face* $\mathsf{f}_{ij} \in \mathcal{F}_h^{\mathrm{int}}$.

(ii) $\mathsf{card}\{\sigma_\alpha\} \leq C_{\mathrm{reg}}$ *for any vertex* $\mathsf{v}_\alpha \in \mathcal{V}_h$.

(iii) $\tilde{\lambda}_\alpha^{ij} \geq 1/(d+1)$ *for any face* $\mathsf{f}_{ij} \in \mathcal{F}_h$ *and any vertex* $\mathsf{v}_\alpha \in \nu_{ij}$.

The proof is omitted because these relations readily follow from Assumption 1.

3.2. Reconstruction of vertex values. The value at the mesh vertex v_α is reconstructed by

$$u_\alpha = \begin{cases} \sum_{k\in\sigma_\alpha} \mathrm{w}_k^\alpha u_k, & \mathsf{v}_\alpha \in \mathcal{V}_h^{\mathrm{int}}, \\ \mathsf{g}(\boldsymbol{x}_\alpha), & \mathsf{v}_\alpha \in \mathcal{V}_h^{\mathrm{bnd}}, \end{cases} \tag{3}$$

where $\{\mathrm{w}_k^\alpha,\ k \in \sigma_\alpha\}$ is the set of coefficients associated to $\mathsf{v}_\alpha \in \mathcal{V}_h^{\mathrm{int}}$, and $\mathsf{g}(\boldsymbol{x}_\alpha)$ is the Dirichlet boundary condition of the vertex $\mathsf{v}_\alpha \in \mathcal{V}_h^{\mathrm{bnd}}$. In Theorem 5 we demonstrate the existence of a special set of coefficients $\{\mathrm{w}_k^\alpha\}$ that are strictly positive, bounded from above by 1, and such that (3) is exact for linear polynomials defined on $\cup_{k\in\sigma_\alpha}\mathsf{T}_k \subseteq \overline{\Omega}$. The proof of Theorem 5 is based on the technical result of Lemma 4 that exploits the possibility of separating convex sets in $\mathbb{R}^n$. As this latter one is a standard result from convex analysis, it is given in Lemma 3 without proof for the sake of reference.

Lemma 3. *There exists a closed hyperplane that strictly separates any two nonempty and disjoint closed convex subsets of* $\mathbb{R}^n$*, provided that one of the two sets is also compact.*

Proof. See, e.g., Rockafellar [28, section 11]. □

Lemma 4. *Under the mesh regularity Assumption* 1*, for every* $\mathsf{v}_\alpha \in \mathcal{V}_h^{\mathrm{int}}$ *we have that* $\mathcal{B}_{\boldsymbol{x}_\alpha,\rho} \subseteq \mathcal{C}$*, where* $\mathcal{C} = \mathsf{conv}\{\boldsymbol{x}_k,\ k \in \sigma_\alpha\}$.

Proof. Let v_α be an internal vertex of the triangulation $\mathcal{T}_h$. We will demonstrate by contradiction that neither does $\mathcal{B}_{\boldsymbol{x}_\alpha,\rho}$ lie completely outside of the convex hull $\mathcal{C}$ nor is $\mathcal{B}_{\boldsymbol{x}_\alpha,\rho}$ partially (but not entirely) included in $\mathcal{C}$. Both contradiction arguments rely on the two following basic observations that are consequences of Assumption 1(ii): (a) any hyperplane for v_α divides the space $\mathbb{R}^d$ in two (closed) half-spaces that must entirely contain at least one d-simplex with index in σ_α; (b) the distance between each one of the barycenters of the two simplices of item (a) and the given hyperplane for v_α must be greater than or equal to ρ. This situation is illustrated by Figure 2(a).

Let us first consider the case in which the two closed and convex sets $\mathcal{B}_{\boldsymbol{x}_\alpha,\rho}$ and $\mathcal{C}$ are disjoint; a strictly separating hyperplane exists by Lemma 3, as shown by Figure 2(b). Thus, the barycenters of all the d-simplices surrounding v_α, which of course belong to $\mathcal{C}$, must lie in the same half-space of the two half-spaces defined by the hyperplane for v_α and parallel to the separating hyperplane. This fact contradicts the initial observation (a).

When $\mathcal{B}_{\boldsymbol{x}_\alpha,\rho}$ is partially but not fully included in $\mathcal{C}$, a point $\mathbf{y}$ exists in the interior of $\mathcal{B}_{\boldsymbol{x}_\alpha,\rho}\backslash\mathcal{C}$. Lemma 3 again provides a hyperplane separating $\mathbf{y}$ and $\mathcal{C}$, as shown by Figure 2(c). The distance between the separating hyperplane and v_α must be strictly less than ρ because $\mathbf{y}$ is an interior point of $\mathcal{B}_{\boldsymbol{x}_\alpha,\rho}$. Let us now consider the hyperplane

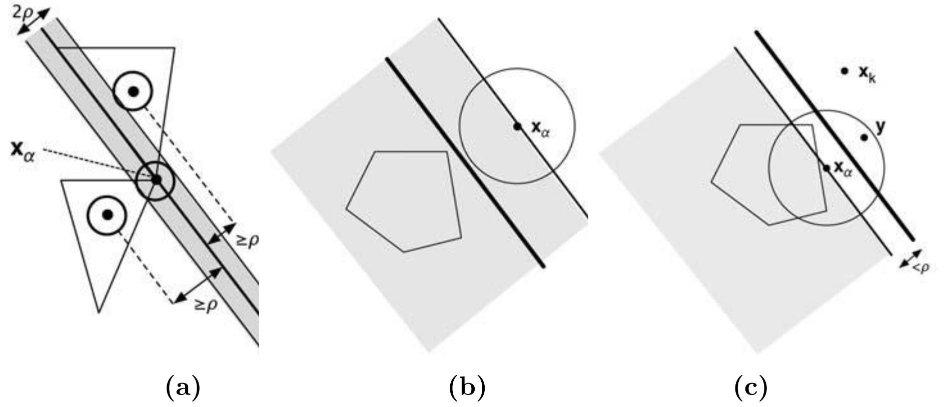


FIG. 2. *Proof of Lemma* 4*: the thicker lines in* (b) *and* (c) *are the separating hyperplane provided by Lemma* 3*; the polygonal body is the convex hull of the cell centers.*

for v_α, which is parallel to the separating hyperplane provided by Lemma 3, and the two half-spaces defined by it. Observation (a) implies that at least one d-simplex T_k exists in the same half-space containing $\mathbf{y}$, and observation (b) that its barycenter $\boldsymbol{x}_k$ has a distance greater than ρ from the hyperplane for v_α. As shown by Figure 2(c), the barycenter $\boldsymbol{x}_k$ must be located on the same side of $\mathbf{y}$ and beyond the separating hyperplane because the distance between the hyperplane for v_α and the separating hyperplane is strictly less than ρ. This last statement contradicts the fact that all the barycenters are in the half-space defined by the separating hyperplane and not containing $\mathbf{y}$. ▯

The following regularity constant is used in Theorem 5:

$$C_{\text{grid}} = \frac{1}{2} C_{\text{reg}}{}^{\frac{d-1}{d}}.$$

THEOREM 5. *Under the mesh regularity Assumption* 1*, for any internal vertex* $\mathsf{v}_\alpha \in \mathcal{V}_h^{\text{int}}$ *there exists a set of coefficients* $\{\mathrm{w}_k^\alpha\}$ *such that*

(i) $C_{\text{grid}} \leq \mathrm{w}_k^\alpha < 1$ *for* $k \in \sigma_\alpha$*;*

(ii) $\sum_{k\in\sigma_\alpha} \mathrm{w}_k^\alpha = 1$*;*

(iii) $\sum_{k\in\sigma_\alpha} \mathrm{w}_k^\alpha(\boldsymbol{x}_k - \boldsymbol{x}_\alpha) = 0$.

Proof. For any $k \in \sigma_\alpha$, let us define the two vectors

$$\sum_{j\in\sigma_\alpha} a_{kj}\boldsymbol{x}_j = \{\boldsymbol{x}_\alpha + t(\boldsymbol{x}_k - \boldsymbol{x}_\alpha),\ t \leq 0\} \cap \partial\mathcal{B}_{\boldsymbol{x}_\alpha,\rho}, \tag{4a}$$

$$\boldsymbol{x}_\alpha + b_k(\boldsymbol{x}_k - \boldsymbol{x}_\alpha) = \{\boldsymbol{x}_\alpha + t(\boldsymbol{x}_k - \boldsymbol{x}_\alpha),\ t \geq 0\} \cap \partial\mathcal{B}_{\boldsymbol{x}_\alpha,h} \tag{4b}$$

that are expressed by a suitable choice of the scalar coefficients a_{kj} and b_k. In view of Lemma 4, the right-hand side of (4a) defines a *convex* linear combination of the vectors $\boldsymbol{x}_j$ for $j \in \sigma_\alpha$. Thus, the coefficients $\{a_{kj}\}$ at the left-hand side of (4a) can be chosen nonnegative and such that $\sum_{j\in\sigma_\alpha} a_{kj} = 1$. Relation (4b) is valid with $b_k \geq 1$. By construction we have that

$$\rho b_k(\boldsymbol{x}_k - \boldsymbol{x}_\alpha) + h \sum_{j\in\sigma_\alpha} a_{kj}(\boldsymbol{x}_j - \boldsymbol{x}_\alpha) = 0 \qquad \text{for every } k \in \sigma_\alpha. \tag{5}$$

The proof terminates by verifying that the coefficients

$$\mathrm{w}_k^\alpha = \frac{1}{\mathsf{card}\{\sigma_\alpha\}\,(\rho b + h)}\Big[\rho b_k + h\sum_{j\in\sigma_\alpha} a_{jk}\Big], \qquad \text{with } b = \frac{1}{\mathsf{card}\{\sigma_\alpha\}}\sum_{k\in\sigma_\alpha} b_k, \tag{6}$$

actually satisfy the statements of the theorem. The left inequality of item (i) follows from

$$\mathrm{w}_k^\alpha \geq \frac{\rho b_k}{\mathsf{card}\{\sigma_\alpha\}\,(\rho b + h)} \geq \frac{\rho}{2\mathsf{card}\{\sigma_\alpha\}\,h} \geq \frac{\rho}{2hC_{\mathrm{reg}}} \geq C_{\mathrm{grid}}$$

because $b_k \geq 1$ and $b \leq h/\rho$. Items (ii) and (iii) are a straightforward consequence of the construction property (5) and definition (6). The proof of item (i) is finally completed by observing that none of the strictly positive coefficients w_k^α for $k \in \sigma_\alpha$ can be greater than or equal to 1 if item (ii) is true. □

Let us emphasize the constructive nature of the above proof that provides a practical method of computing the set of coefficients $\{\mathrm{w}_k^\alpha\}$. In view of the mesh regularity Assumption 1 and Lemma 4, for any cell center $\boldsymbol{x}_k$, $k \in \sigma_\alpha$, there always exists a subset of d indices $\{k_1, \ldots, k_d\} \in \sigma_\alpha$ such that $\mathsf{conv}\{\boldsymbol{x}_{k_1}, \ldots, \boldsymbol{x}_{k_d}\}$ has a nonempty intersection with the half-line starting from $\boldsymbol{x}_\alpha$ and having direction $\boldsymbol{x}_\alpha - \boldsymbol{x}_k$. As more than one choice of indices $\{k_1, \ldots, k_d\}$ may exist, we select the one that maximizes the distance between the intersection point and the vertex v_α. Let $\mathsf{p}_{\alpha k}$ denote the position vector of this intersection point. Replacing the ball $\mathcal{B}_{\boldsymbol{x}_\alpha,\rho}$ in the proof of Theorem 5 by the convex hull of this suitably chosen subset of cell centers surrounding v_α yields the following algorithm:

```
**foreach** $k \in \sigma_\alpha$ **do** let $\mathrm{w}_k^\alpha = 0$;
**foreach** $k \in \sigma_\alpha$ **do**
    choose the $d$ indices $\{\boldsymbol{x}_{k_1}, \ldots, \boldsymbol{x}_{k_d}\}$ maximizing $|\mathsf{p}_{\alpha k} - \boldsymbol{x}_\alpha|$;
    compute $0 \leq \beta \leq 1$ such that $\boldsymbol{x}_\alpha = (1-\beta)\mathsf{p}_{\alpha k} + \beta \boldsymbol{x}_k$;
    compute $\{\lambda_l\}$ such that $\mathsf{p}_{\alpha k} = \sum_{l=1}^d \lambda_l \boldsymbol{x}_{k_l}$;
    accumulate $\mathrm{w}_k^\alpha \leftarrow \mathrm{w}_k^\alpha + \beta$;
    **foreach** $l = 1, \ldots, d$ **do** accumulate $\mathrm{w}_{k_l}^\alpha \leftarrow \mathrm{w}_{k_l}^\alpha + (1-\beta)\lambda_l$;
**end**
**foreach** $k \in \sigma_\alpha$ **do** let $\mathrm{w}_k^\alpha \leftarrow \mathrm{w}_k^\alpha / \sum_{k\in\sigma_\alpha} \mathrm{w}_k^\alpha$.
```

4. Finite volume formulation.

The model problem (1a)–(1b) is reformulated by integrating on the generic control volume $\mathsf{T}_i \in \mathcal{T}_h$ and applying the Gauss–Green theorem as follows:

$$\frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i\,\cup\,\sigma_i'} \int_{\mathsf{f}_{ij}} \big(\mathsf{u}\boldsymbol{v} - \nu\nabla\mathsf{u}\big)\cdot \boldsymbol{n}_{ij}\, dS = \frac{1}{|\mathsf{T}_i|}\int_{\mathsf{T}_i} \mathsf{s}\, dV \quad \text{for } \mathsf{T}_i \in \mathcal{T}_h.$$

The finite volume approximation of the cell average of u on T_i is denoted by u_i. The vector that collects all the approximate cell averages is denoted by u_h; i.e., $\mathsf{u}_h|_i = u_i$. The finite volume scheme correlates the approximate cell-average vector u_h to the balance of the numerical fluxes across the bounding faces in $\partial\mathsf{T}_i$ by the set of relations

$$\frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i} |\mathsf{f}_{ij}|\,\Big[\mathcal{F}_{ij}(\mathsf{u}_h) + \mathcal{G}_{ij}(\mathsf{u}_h)\Big] + \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i'} |\mathsf{f}_{ij}|\,\mathcal{B}_{ij}(\mathsf{u}_h) = \mathsf{s}_i \text{ for } \mathsf{T}_i \in \mathcal{T}_h. \tag{7}$$

The terms $\mathcal{F}_{ij}(\mathsf{u}_h)$ and $\mathcal{G}_{ij}(\mathsf{u}_h)$ are the face integrals of the numerical advective and diffusive fluxes on $\mathsf{f}_{ij} \in \mathcal{F}_h^{\mathrm{int}}$; $\mathcal{B}_{ij}(\mathsf{u}_h)$ is the face integral of the numerical flux on $\mathsf{f}_{ij} \in \mathcal{F}_h^{\mathrm{bnd}}$ and collects boundary contributions from both advective and diffusive terms; s_i is the volume integral of the source term over T_i approximated by an appropriate quadrature formula.

4.1. Cell-centered gradient, slope limiter, and numerical advective flux. The piecewise constant approximation of the solution gradient within the control volume T_i is derived by applying the Gauss–Green theorem to the volume integral of $\nabla \mathsf{u}$ and by linearly approximating the boundary integrals on $\partial \mathsf{T}_i$.

Let f_α denote the face opposite to the vertex v_α and $\boldsymbol{n}_\alpha$ its unit normal vector. We thus obtain the formula

$$\boldsymbol{G}_i(\mathsf{u}_h) = -\frac{l_i(\mathsf{u}_h)}{d\,|\mathsf{T}_i|} \sum_{\alpha \in \nu_i} u_\alpha\, |\mathsf{f}_\alpha|\, \boldsymbol{n}_\alpha, \tag{8}$$

where $l_i(\mathsf{u}_h)$ is the limiter factor defined as follows. In view of (8), we first introduce the linear reconstruction of the solution approximation:

$$\mathcal{R}_i(\mathsf{u}_h)(\boldsymbol{x}) = \begin{cases} u_i + \boldsymbol{G}_i(\mathsf{u}_h) \cdot (\boldsymbol{x} - \boldsymbol{x}_i) & \text{if } \boldsymbol{x} \in \mathsf{T}_i, \\ 0 & \text{otherwise.} \end{cases}$$

The slope limiter factor $l_i(\mathsf{u}_h)$ in (8) is the largest real number in $[0, 1]$ such that

$$\min\{u_i, \min_{\alpha \in \nu_i} u_\alpha\} \leq \mathcal{R}_i(\mathsf{u}_h)(\boldsymbol{x}_{ij}) \leq \max\{u_i, \max_{\alpha \in \nu_i} u_\alpha\} \quad \text{for } j \in \sigma_i \cup \sigma_i', \tag{9}$$

and

$$\|\boldsymbol{G}(\mathsf{u}_h)\|_{\mathcal{T}_h} \leq C_{\mathrm{lim}}, \tag{10}$$

where C_{lim} is a suitable bound from above of the $\mathtt{T}_h$-norm of the finite volume approximation of the solution gradient $\nabla \mathsf{u}$. Note that condition (10) makes it possible to control a discrete counterpart of the total variation of u_h.

LEMMA 6. *If u_i is a local maximum (minimum), i.e., $u_i \geq \max_{\alpha \in \nu_i} u_\alpha$ ($u_i \leq \min_{\alpha \in \nu_i} u_\alpha$), then $\boldsymbol{G}_i(\mathsf{u}_h) = 0$.*

Proof. In view of the definition of the limiter and the assumption of the lemma, we have that $\mathcal{R}_i(\mathsf{u}_h)(\boldsymbol{x}_{ij}) \leq \max\{u_i, \max_{\alpha \in \nu_i} u_\alpha\} = u_i$; i.e., $\boldsymbol{G}_i(\mathsf{u}_h) \cdot (\boldsymbol{x}_{ij} - \boldsymbol{x}_i) \leq 0$ for $j \in \sigma_i \cup \sigma_i'$. Since $\boldsymbol{x}_i$ is the barycenter of T_i, we have that $(d+1)\boldsymbol{x}_i = \sum_{j \in \sigma_i \cup \sigma_i'} \boldsymbol{x}_{ij}$, and then $\sum_{j \in \sigma_i \cup \sigma_i'} \boldsymbol{G}_i(\mathsf{u}_h) \cdot (\boldsymbol{x}_{ij} - \boldsymbol{x}_i) = 0$. Consequently, d linearly independent directions $\boldsymbol{x}_{ij} - \boldsymbol{x}_i$ exist among the $(d+1)$-ones for $j \in \sigma_i \cup \sigma_i'$ such that the orthogonal projection of $\boldsymbol{G}_i(\mathsf{u}_h)$ onto them is the zero vector. □

The numerical advective flux at the internal face f_{ij} is derived from the standard upwind formula

$$\mathcal{F}_{ij}(\mathsf{u}_h) = v_{ij}^+ \mathcal{R}_i(\mathsf{u}_h)(\boldsymbol{x}_{ij}) + v_{ij}^- \mathcal{R}_j(\mathsf{u}_h)(\boldsymbol{x}_{ij}), \qquad j \in \sigma_i,$$

where

$$v_{ij} = \frac{1}{|\mathsf{f}_{ij}|} \int_{\mathsf{f}_{ij}} \boldsymbol{v}(\boldsymbol{x}) \cdot \boldsymbol{n}_{ij}\, dS, \qquad v_{ij}^\pm = \frac{v_{ij} \pm |v_{ij}|}{2}.$$

Since $v^-_{ij} + v^+_{ji} = 0$, the balance of the numerical advective flux of the internal face $\mathsf{f}_{ij} \in \mathcal{F}^{\rm int}_h$ of the cell T_i can be compactly written as

$$\frac{1}{|\mathsf{T}_i|} \sum_{j\in\sigma_i} |\mathsf{f}_{ij}|\, \mathcal{F}_{ij}(\mathsf{u}_h) = \big(\mathbf{F}\mathsf{u}_h - \mathbf{r}^{\rm int}(\mathsf{u}_h)\big)\Big|_i \tag{11}$$

by introducing the advective flux matrix

$$\mathsf{F}_{ij} = \frac{1}{|\mathsf{T}_i|} \begin{cases} \sum_{k\in\sigma_i} |\mathsf{f}_{ik}|\, v^+_{ik}, & j = i, \\ -\,|\mathsf{f}_{ij}|\, v^+_{ji}, & j \in \sigma_i, \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

and the advective flux vector

$$\mathsf{r}^{\rm int}_i(\mathsf{u}_h) = \frac{1}{|\mathsf{T}_i|} \sum_{j\in\sigma_i} |\mathsf{f}_{ij}| \big(v^+_{ij} \boldsymbol{G}_i(\mathsf{u}_h)\cdot(\boldsymbol{x}_i - \boldsymbol{x}_{ij}) - v^+_{ji} \boldsymbol{G}_j(\mathsf{u}_h)\cdot(\boldsymbol{x}_j - \boldsymbol{x}_{ji}) \big). \tag{13}$$

Note that the term $\mathbf{F}\mathsf{u}_h|_i$ in (11) is a first-order accurate discretization of the flux integral $\int_{\partial\mathsf{T}_i\setminus\partial\Omega} \mathsf{u}(\boldsymbol{x})\boldsymbol{v}(\boldsymbol{x})\cdot\boldsymbol{n}\, dS/\,|\mathsf{T}_i|$, while second-order accuracy is provided by the term $\mathsf{r}^{\rm int}_i(\mathsf{u}_h)$.

4.2. Face gradient and numerical diffusive flux. To define the numerical diffusive flux at the internal face $\mathsf{f}_{ij} \in \mathcal{F}^{\rm int}_h$, we introduce the face gradient by a special nonlinear average of two *one-sided* face gradients $\widetilde{\boldsymbol{G}}_{ij}(\mathsf{u}_h)$ and $\widetilde{\boldsymbol{G}}_{ji}(\mathsf{u}_h)$. The *one-sided* face gradient $\widetilde{\boldsymbol{G}}_{ij}(\mathsf{u}_h)$ is recovered from cell averages by applying the Gauss–Green theorem to the integral of $\nabla\mathsf{u}$ on the *half-diamond* delimited by the center of T_i and the face vertices v_α, $\alpha \in \nu_{ij}$. The linear approximation of the resulting boundary integrals gives the one-sided formula

$$\widetilde{\boldsymbol{G}}_{ij}(\mathsf{u}_h) = \frac{\tilde{u}_{ij} - u_i}{h_{ij}} \boldsymbol{n}_{ij} + \{\text{TANGENTIAL TERM}\}, \tag{14}$$

where $\tilde{u}_{ij}$ is the approximate solution at $\tilde{\boldsymbol{x}}_{ij} \in \mathsf{f}_{ij}$, and the tangential term is left unspecified because it does not contribute to the integral of the normal flux. The value $\tilde{u}_{ij}$ is approximated by linear interpolation of the values at the vertices of the face f_{ij} given by (3); we have the formula

$$\tilde{u}_{ij} = \sum_{\alpha\in\nu_{ij}} \tilde{\lambda}^{ij}_\alpha u_\alpha. \tag{15}$$

The other one-sided face gradient $\widetilde{\boldsymbol{G}}_{ji}(\mathsf{u}_h)$ is similarly defined on the opposite half-diamond (related formulae are readily available by simply interchanging i and j in the previous derivation).

Note that a *unique* definition of the face gradient is required to have a *conservative* formulation of the diffusive flux. The two one-sided face gradients previously built cannot be taken as possible candidates because selecting one of them would imply a loss of information from the discarded one. Thus, we define the face gradient at $\mathsf{f}_{ij} \in \mathcal{F}^{\rm int}_h$ by the nonlinear average

$$\boldsymbol{G}_{ij}(\mathsf{u}_h) = \omega_{ij}(\mathsf{u}_h)\widetilde{\boldsymbol{G}}_{ij}(\mathsf{u}_h) + \omega_{ji}(\mathsf{u}_h)\widetilde{\boldsymbol{G}}_{ji}(\mathsf{u}_h), \tag{16}$$

where $\omega_{ij}(\mathsf{u}_h)$ and $\omega_{ji}(\mathsf{u}_h)$ are two nonnegative solution-dependent weights satisfying $\omega_{ij}(\mathsf{u}_h)+\omega_{ji}(\mathsf{u}_h)=1$ for any vector $\mathsf{u}_h \in \mathtt{T}_h$.

As pointed out in [5], the usual face gradient of the diamond scheme is readily obtained by choosing the constant weights

$$\omega_{ij}(\mathsf{u}_h)=|\mathsf{T}_i|/(|\mathsf{T}_i|+|\mathsf{T}_j|). \tag{17}$$

When $d=2$, this choice leads to the numerical diffusive flux considered in [11].

It is well known that the diamond scheme cannot ensure a maximum principle on general grids; see, for instance, [11, 20]. Thus, a different design of the weights $\omega_{ij}(\mathsf{u}_h)$ and $\omega_{ji}(\mathsf{u}_h)$ must be envisaged that generalizes (17) in a nonlinear sense. To do that, we proceed as follows. We substitute the vertex-reconstructed values provided by (3) in (15) and use the summation rule

$$\sum_{\alpha\in\nu_{ij}^{\text{int}}}\sum_{k\in\sigma_\alpha}=\sum_{k\in\sigma\nu_{ij}}\sum_{\alpha\in\nu_k\cap\nu_{ij}^{\text{int}}} \tag{18}$$

to get

$$\tilde{u}_{ij}=\sum_{k\in\sigma\nu_{ij}}p_k^{ij}u_k+p^{ij}g_{ij}^{\text{bnd}},\qquad g_{ij}^{\text{bnd}}=\frac{1}{p^{ij}}\sum_{\alpha\in\nu_{ij}^{\text{bnd}}}\tilde{\lambda}_\alpha^{ij}\,\mathsf{g}(\boldsymbol{x}_\alpha),$$

where

$$p_k^{ij}=\sum_{\alpha\in\nu_k\cap\nu_{ij}^{\text{int}}}\tilde{\lambda}_\alpha^{ij}\mathrm{w}_k^\alpha\quad\text{for } k\in\sigma_\alpha,\qquad\text{and}\qquad p^{ij}=\sum_{\alpha\in\nu_{ij}^{\text{bnd}}}\tilde{\lambda}_\alpha^{ij}. \tag{19}$$

Note that the term g_{ij}^{bnd} collects the contributions to $\tilde{u}_{ij}$ from the boundary vertices that may belong to the internal face $\mathsf{f}_{ij}\in\mathcal{F}_h^{\text{int}}$. Next, we rewrite the orthogonal term in (14) as

$$\boldsymbol{n}_{ij}\cdot\widetilde{\boldsymbol{G}}_{ij}(\mathsf{u}_h)=D_{ij}\frac{u_j-u_i}{H_{ij}}+g_{ij}(\mathsf{u}_h) \tag{20}$$

by introducing the scalar quantity

$$D_{ij}=H_{ij}\min\{p_j^{ij}/h_{ij},\,p_i^{ji}/h_{ji}\} \tag{21}$$

and the remainder term

$$g_{ij}(\mathsf{u}_h)=\sum_{k\in\sigma\nu_{ij}}\left(\frac{p_k^{ij}}{h_{ij}}-\delta_{jk}\frac{D_{ij}}{H_{ij}}\right)(u_k-u_i)+\frac{p^{ij}}{h_{ij}}(g_{ij}^{\text{bnd}}-u_i), \tag{22}$$

with $\delta_{jk}=1$ if $j=k$, $\delta_{jk}=0$ otherwise.

The orthogonal component of $\widetilde{\boldsymbol{G}}_{ji}(\mathsf{u}_h)$ is similarly defined. The corresponding formulae are easily obtained by interchanging i and j in (20)–(22).

Finally, we define the nonlinear weights to be used in (16) by the formula

$$\omega_{ij}(\mathsf{u}_h)=\begin{cases}\frac{|g_{ji}(\mathsf{u}_h)|}{|g_{ij}(\mathsf{u}_h)|+|g_{ji}(\mathsf{u}_h)|} & \text{if } |g_{ij}(\mathsf{u}_h)|+|g_{ji}(\mathsf{u}_h)|>0,\\ 1/2 & \text{otherwise.}\end{cases} \tag{23}$$

Let $s_{ij}(\mathsf{u}_h)$ indicate the sign of $g_{ij}(\mathsf{u}_h)$; i.e., $|g_{ij}(\mathsf{u}_h)| = g_{ij}(\mathsf{u}_h)s_{ij}(\mathsf{u}_h)$. A direct calculation shows that

$$\omega_{ij}(\mathsf{u}_h)g_{ij}(\mathsf{u}_h) - \omega_{ji}(\mathsf{u}_h)g_{ji}(\mathsf{u}_h) = g_{ij}(\mathsf{u}_h)g_{ji}(\mathsf{u}_h)\frac{s_{ji}(\mathsf{u}_h) - s_{ij}(\mathsf{u}_h)}{|g_{ij}(\mathsf{u}_h)| + |g_{ji}(\mathsf{u}_h)|}.$$

Thus, the nonlinear coefficients introduced in (23) satisfy the relation

$$\omega_{ij}(\mathsf{u}_h)g_{ij}(\mathsf{u}_h) - \omega_{ji}(\mathsf{u}_h)g_{ji}(\mathsf{u}_h) = 2\,\omega^0_{ij}(\mathsf{u}_h)g_{ij}(\mathsf{u}_h), \tag{24}$$

where

$$\omega^0_{ij}(\mathsf{u}_h) = \begin{cases} \omega_{ij}(\mathsf{u}_h) & \text{if } g_{ij}(\mathsf{u}_h)g_{ji}(\mathsf{u}_h) < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{25}$$

This remark has a very important consequence: using (24)–(25) we can reformulate the orthogonal component of the nonlinearly weighted face gradient (16) in a one-sided form. To prove this fact, we first substitute into (16) the expression for $\boldsymbol{n}_{ij} \cdot \widetilde{\boldsymbol{G}}_{ij}(\mathsf{u}_h)$ given by (20)–(22) and the similar one for $\boldsymbol{n}_{ji} \cdot \widetilde{\boldsymbol{G}}_{ji}(\mathsf{u}_h)$ obtained by interchanging i and j. Then we consider the nonlinear weights (23) and use the property expressed by (24) together with (25). Finally, we introduce the nonlinear mapping $\boldsymbol{\psi}(\mathsf{u}_h) = \{w^{ij}_k(\mathsf{u}_h)\}$ whose components are

$$w^{ij}_k(\mathsf{u}_h) \;=\; 2\,\omega^0_{ij}(\mathsf{u}_h)\begin{cases} p^{ij}H_{ij}/h_{ij}, & k = i, \\ p^{ij}_j H_{ij}/h_{ij} - D_{ij}, & k = j, \\ p^{ij}_k H_{ij}/h_{ij}, & k \in \sigma\nu_{ij}\backslash\{i,\, j\}, \\ 0 & \text{otherwise.} \end{cases} \tag{26}$$

This straightforward calculation shows that (16) becomes

$$\begin{aligned} \boldsymbol{n}_{ij} \cdot \boldsymbol{G}_{ij}(\mathsf{u}_h) \;&=\; \frac{D_{ij}}{H_{ij}}(u_j - u_i) + 2\,\omega^0_{ij}(\mathsf{u}_h)g_{ij}(\mathsf{u}_h) \\ &=\; \frac{1}{H_{ij}}\Big(D_{ij}(u_j - u_i) + \sum_{k\in\sigma\nu_{ij}} w^{ij}_k(\mathsf{u}_h)(u_k - u_i) + w^{ij}_i(\mathsf{u}_h)(g^{\text{bnd}}_{ij} - u_i)\Big). \end{aligned} \tag{27}$$

Replacing $\boldsymbol{n} \cdot \nabla \mathsf{u}$ by $\boldsymbol{n}_{ij} \cdot \boldsymbol{G}_{ij}(\mathsf{u}_h)$ in the flux integral $\int_{\mathsf{f}_{ij}} \boldsymbol{n} \cdot \nabla \mathsf{u}\, dS / |\mathsf{f}_{ij}|$ and using the final expression of $\boldsymbol{G}_{ij}(\mathsf{u}_h)$ in (27) yield the average value of the numerical diffusive flux on $\mathsf{f}_{ij} \in \mathcal{F}^{\text{int}}_h$. Thus,

$$\mathcal{G}_{ij}(\mathsf{u}_h) = -\nu_{ij}\boldsymbol{G}_{ij}(\mathsf{u}_h) \cdot \boldsymbol{n}_{ij},$$

where $\nu_{ij} = \int_{\mathsf{f}_{ij}} \nu(\boldsymbol{x})\, dS / |\mathsf{f}_{ij}|$ is the average of the viscosity field $\nu(\boldsymbol{x})$ over f_{ij}.

Let us denote by $\mathbf{w}$ the generic instance of $\boldsymbol{\psi}(\mathsf{u}_h)$. The flux balance of the diffusive terms at the internal faces of T_i can be compactly written as

$$\frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i} |\mathsf{f}_{ij}|\, \mathcal{G}_{ij}(\mathsf{u}_h) = \big(\mathbf{G}(\mathbf{w})\mathsf{u}_h - \mathbf{g}(\mathbf{w})\big)\Big|_i \tag{28}$$

by taking into account the summation rule

$$\sum_{l\in\sigma_i}\;\sum_{k\in\sigma\nu_{il}} = \sum_{k\in\sigma\nu_i}\;\sum_{l\in\sigma_i\,\cap\,\sigma\nu_k}$$

and introducing the diffusive flux matrix and vector

$$
\begin{aligned}
\mathsf{G}_{ij}(\mathbf{w}) &= \frac{1}{|\mathsf{T}_i|}\begin{cases} \sum_{l\in\sigma_i} \frac{|\mathsf{f}_{il}|\nu_{il}}{H_{il}}(D_{il}+\sum_{k\in\sigma\nu_{il}} w_k^{il}) & \text{for } j=i,\\ -\sum_{l\in\sigma_i\cap\sigma\nu_j} \frac{|\mathsf{f}_{il}|\nu_{il}}{H_{il}}(w_j^{il}+\delta_{lj}D_{il}) & \text{for } j\in\sigma\nu_i\backslash\{i\},\\ 0 & \text{otherwise,}\end{cases} \\
\mathsf{g}_i(\mathbf{w}) &= \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i}\frac{|\mathsf{f}_{ij}|\,\nu_{ij}w_i^{ij}}{H_{ij}}g_{ij}^{\mathrm{bnd}}.
\end{aligned}
\tag{29}
$$

4.3. Treatment of boundary terms. The contribution to the flux balance of the cell T_i from the boundary faces $\mathsf{f}_{ij}\in\mathcal{F}_h^{\mathrm{bnd}}$ takes the vector form

$$
\frac{1}{\mathsf{T}_i}\sum_{j\in\sigma_i'}|\mathsf{f}_{ij}|\,\mathcal{B}_{ij}(\mathsf{u}_h) = \big(\mathbf{B}\mathsf{u}_h-\mathbf{b}-\mathsf{r}^{\mathrm{bnd}}(\mathsf{u}_h)\big)\Big|_i,
$$

where

$$
\begin{aligned}
\mathsf{B}_{ij} &= \frac{1}{|\mathsf{T}_i|}\begin{cases}\sum_{k\in\sigma_i'}|\mathsf{f}_{ij}|\,[v_{ij}^{+}+\frac{\nu_{ik}}{h_{ik}}], & j=i,\\ 0 & \text{otherwise,}\end{cases}\\
\mathsf{b}_i &= \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i'}|\mathsf{f}_{ij}|\left(v_{ij}^{+}g_{ij}^{D}+\nu_{ij}\frac{\widetilde{g}_{ij}^{D}}{h_{ij}}\right),\\
\mathsf{r}_i^{\mathrm{bnd}}(\mathsf{u}_h) &= \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i'}|\mathsf{f}_{ij}|\,v_{ij}^{+}\boldsymbol{G}_i(\mathsf{u}_h)\cdot(\boldsymbol{x}_i-\boldsymbol{x}_{ij}),
\end{aligned}
\tag{30}
$$

and the terms

$$
g_{ij}^{D}=\sum_{\alpha\in\nu_{ij}^{\mathrm{bnd}}}\mathsf{g}(\boldsymbol{x}_\alpha)/d \qquad \text{and} \qquad \widetilde{g}_{ij}^{D}=\sum_{\alpha\in\nu_{ij}^{\mathrm{bnd}}}\tilde{\lambda}_{\alpha}^{ij}\mathsf{g}(\boldsymbol{x}_\alpha)
\tag{31}
$$

are the linear interpolations at the face points $\boldsymbol{x}_{ij}$ and $\tilde{\boldsymbol{x}}_{ij}$, respectively, of the Dirichlet boundary data $\mathsf{g}(\boldsymbol{x}_\alpha)$ for $\mathsf{v}_\alpha\in\mathsf{f}_{ij}\in\mathcal{F}_h^{\mathrm{bnd}}$.

4.4. The finite volume approximation problem: Vector formulation. Let us collect all the numerical flux matrices and vectors in the terms

$$
\boldsymbol{\mathcal{A}}(\mathbf{w}) = \mathbf{F}+\mathbf{G}(\mathbf{w})+\mathbf{B}, \tag{32a}
$$

$$
\boldsymbol{b}(\mathsf{u}_h;\mathbf{w}) = \mathbf{s}+\mathbf{r}(\mathsf{u}_h)+\mathbf{g}(\mathbf{w})+\mathbf{b}. \tag{32b}
$$

The symbol $\mathbf{w}$—already introduced in (28)—will be considered in the rest of the paper as an independent variable to indicate the generic instance of the coefficients $\{w_k^{ij}\}$. The approximation problem in flux balance formulation (7) can be compactly reformulated as

$$
\boldsymbol{\mathcal{A}}(\mathbf{w})\mathsf{u}_h = \boldsymbol{b}(\mathsf{u}_h;\mathbf{w}), \tag{33a}
$$

$$
\mathbf{w} = \boldsymbol{\psi}(\mathsf{u}_h), \tag{33b}
$$

where $\mathsf{u}_h\in\mathtt{T}_h$ is the finite volume approximation of the solution cell averages, and $\mathbf{w}$ the vector collecting the solution-dependent weights for the face gradient definition.

5. Theoretical issues. In this section, we investigate the solvability of the nonlinear finite volume approximation problem (33a)–(33b) and we demonstrate the existence of at least one numerical solution. Furthermore, we prove that all the numerical solutions to (33a)–(33b) (if more than one exist) must satisfy a discrete maximum principle. Finally, we remark that a discrete minimum principle can be proved as well by suitably adapting the same arguments.

5.1. Technical lemmas. The following lemma formalizes the properties of the face gradient coefficients p^{ij} and $\{p^{ij}_k\}$ defined in (19), of D_{ij} defined in (21), and of the components of the nonlinear mapping $\boldsymbol{\psi}(\mathsf{u}_h) = \{w^{ij}_k(\mathsf{u}_h)\}$ in (26) used in the definition of the numerical diffusive flux (27)–(29).

LEMMA 7. *The coefficients p^{ij} and $\{p^{ij}_k,\, k \in \sigma\nu_{ij}\}$ are nonnegative and such that for any $\mathsf{f}_{ij} \in \mathcal{F}^{\rm int}_h$*

(i) $p^{ij} + \sum_{k\in\sigma\nu_{ij}} p^{ij}_k = 1$*;*

(ii) $C_{\rm grid}/(d+1) \le p^{ij}_j$*;*

(iii) $C_{\rm grid}/(d+1) \le D_{ij} \le 1 + C_{\rm reg}$*;*

(iv) $0 \le w^{ij}_k(\mathsf{u}_h) \le 2(1+C_{\rm reg})$ *for any* $\mathsf{u}_h \in \mathtt{T}_h$.

Proof. The nonnegativity of the coefficients p^{ij} and p^{ij}_k for $k \in \sigma\nu_{ij}$ follows from their definition in (19) by noting that both the vertex reconstruction weights $\{\mathrm{w}^{\alpha}_k\}$ (see Theorem 5(i)) and the coefficients $\{\tilde{\lambda}^{ij}_{\alpha}\}$ (see Proposition 2(iii)) are nonnegative real numbers.

Item (i) follows by reversing the summation rule (18) and using Theorem 5(ii) to get

$$\sum_{k\in\sigma\nu_{ij}} p^{ij}_k = \sum_{k\in\sigma\nu_{ij}} \sum_{\alpha\in\nu_k \cap \nu^{\rm int}_{ij}} \tilde{\lambda}^{ij}_{\alpha} \mathrm{w}^{\alpha}_k = \sum_{\alpha\in\nu^{\rm int}_{ij}} \tilde{\lambda}^{ij}_{\alpha} \sum_{k\in\sigma_\alpha} \mathrm{w}^{\alpha}_k = 1 - \sum_{\alpha\in\nu^{\rm bnd}_{ij}} \tilde{\lambda}^{ij}_{\alpha} = 1 - p^{ij}.$$

Item (ii) follows because $p^{ij}_j = \sum_{\alpha\in\nu^{\rm int}_{ij}} \tilde{\lambda}^{ij}_{\alpha} \mathrm{w}^{\alpha}_j$ from (19), $\tilde{\lambda}^{ij}_{\alpha} \ge 1/(d+1)$ from Proposition 2, $\mathrm{w}^{\alpha}_j \ge C_{\rm grid}$ from Theorem 5(i), and $\mathsf{card}\{\nu^{\rm int}_{ij}\} \ge 1$ from Assumption 1(iii).

Item (iii) and the left inequality of item (iv) readily follow by taking into account item (ii) in (21) and (26), respectively.

The right inequality of item (iv) follows by noting that

$$w^{ij}_k(\mathsf{u}_h) \;\le\; 2\,\omega^0_{ij}(\mathsf{u}_h)\frac{p^{ij}_k H_{ij}}{h_{ij}} \;\le\; 2(1+C_{\rm reg}),$$

by recalling that $H_{ij} = h_{ij} + h_{ji}$ (see the definition in section 2), and by applying Proposition 2(i). ▯

After Lemma 7(iv), it is natural to consider the coefficient vectors $\mathbf{w}$ for face gradient calculations that are ranging in the following compact convex set.

DEFINITION 8 (the set of coefficient vectors for face gradients). *Let*

$$\boldsymbol{\mathcal{K}} = \Big\{\mathbf{w} \textit{ such that } 0 \le w^{ij}_k \le 2(1+C_{\rm reg})\Big\}.$$

Similarly, it is possible to restrict the set of admissible solutions u_h of the nonlinear finite volume approximation problem (33a)–(33b). We proceed by demonstrating these two technical lemmas, the result of the first one being useful in proving the second one.

Lemma 9. *There exists a real constant $C_R > 0$, independent of h, and such that $\|\mathbf{r}(\mathsf{u}_h)\|_{\mathcal{T}_h} \leq C_R$.*

Proof. The proof of this lemma uses the same arguments of the proof of Lemma 6.2 in [5]. For this reason and for the sake of completeness, we shortly sketch the proof herein and refer the reader to that paper for a detailed discussion.

Let us first introduce the quantities

$$\Delta u_{ij} = (\boldsymbol{x}_{ij} - \boldsymbol{x}_i) \cdot \boldsymbol{G}_i(\mathsf{u}_h) \quad \text{and} \quad \Delta u_{ji} = (\boldsymbol{x}_{ji} - \boldsymbol{x}_j) \cdot \boldsymbol{G}_j(\mathsf{u}_h).$$

In view of (13), by applying the Cauchy–Schwarz inequality and by rewriting the summations over $\mathsf{f}_{ij} \in \mathcal{F}_h$ we get

$$\begin{aligned} \|\mathbf{r}(\mathsf{u}_h)\|^2_{\mathcal{T}_h} &= \sum_{\mathsf{T}_i \in \mathcal{T}_h} |\mathsf{T}_i| \left(\frac{1}{|\mathsf{T}_i|} \Big[\sum_{j \in \sigma_i} |\mathsf{f}_{ij}| \left(v^+_{ij} \Delta u_{ij} + v^-_{ij} \Delta u_{ji} \right) + \sum_{j \in \sigma'_i} |\mathsf{f}_{ij}| \, v^+_{ij} \Delta u_{ij} \Big] \right)^2 \\ &\leq 2(d+1) \, \|\boldsymbol{v}\|^2_{\mathrm{L}^\infty(\Omega)} \times (\star), \end{aligned}$$

where

$$(\star) = \sum_{\mathsf{f}_{ij} \in \mathcal{F}^{\mathrm{int}}_h} \left(\frac{|\mathsf{f}_{ij}|^2}{|\mathsf{T}_i|} + \frac{|\mathsf{f}_{ij}|^2}{|\mathsf{T}_j|} \right) \left(|\Delta u_{ij}|^2 + |\Delta u_{ji}|^2 \right) + \sum_{\mathsf{f}_{ij} \in \mathcal{F}^{\mathrm{bnd}}_h} \frac{|\mathsf{f}_{ij}|^2}{|\mathsf{T}_i|} \, |\Delta u_{ij}|^2 .$$

Then, taking into account that $|\mathsf{T}_i| = (d+1) h_{ij} \, |\mathsf{f}_{ij}| / d$, multiplying and dividing this expression by H_{ij}, and exploiting the fact that $H_{ij}/h_{ij} = 1 + h_{ji}/h_{ij} \leq 1 + C_{\mathrm{reg}}$ from Proposition 2(i) yield

$$(\star) \leq \frac{d(1 + C_{\mathrm{reg}})}{d+1} \Bigg(\sum_{\mathsf{f}_{ij} \in \mathcal{F}^{\mathrm{int}}_h} |\mathsf{f}_{ij}| \, H_{ij} \frac{|\Delta u_{ij}|^2 + |\Delta u_{ji}|^2}{H^2_{ij}} + \sum_{\mathsf{f}_{ij} \in \mathcal{F}^{\mathrm{bnd}}_h} |\mathsf{f}_{ij}| \, h_{ij} \frac{|\Delta u_{ij}|^2}{h^2_{ij}} \Bigg).$$

By using $|\Delta u_{ij}| \leq h \, |\boldsymbol{G}_i(\mathsf{u}_h)|$, we have

$$\begin{aligned} (\star) &\leq \frac{d(1 + C_{\mathrm{reg}})}{d+1} \Big(\sup_{\mathsf{f}_{ij} \in \mathcal{F}_h} \frac{h}{H_{ij}} \Big)^2 \Big(\sum_{\mathsf{T}_i \in \mathcal{T}_h} |\boldsymbol{G}_i(\mathsf{u}_h)|^2 \sum_{j \in \sigma_i} |\mathsf{f}_{ij}| \, H_{ij} \Big), \\ &\leq \left(2d^2 (1 + C_{\mathrm{reg}})^2 C^2 \right) / (d+1) \sum_{\mathsf{T}_i \in \mathcal{T}_h} |\mathsf{T}_i| \, |\boldsymbol{G}_i(\mathsf{u}_h)|^2 , \end{aligned}$$

where $h/H_{ij} \leq C$ is true from the mesh regularity; see Assumption 1. In view of (10), the lemma follows by setting $C_R = 2d \, \|\boldsymbol{v}\|_{\mathrm{L}^\infty(\Omega)} \, C_{\mathrm{lim}} C (1 + C_{\mathrm{reg}})$. ☐

Lemma 10. *The matrix $\boldsymbol{\mathcal{A}}(\mathbf{w}) = \mathbf{F} + \mathbf{G}(\mathbf{w}) + \mathbf{B}$ introduced in* (32a) *is a nonsingular M-matrix for any vector $\mathbf{w} \geq \mathbf{0}$. Moreover, there exists a nonnegative constant C_M independent of u_h and $\mathbf{w}$ such that*

$$\left\| \boldsymbol{\mathcal{A}}(\mathbf{w})^{-1} \boldsymbol{b}(\mathsf{u}_h; \mathbf{w}) \right\|_{\mathcal{T}_h} \leq C_M \qquad \textit{for any } \mathsf{u}_h \in \mathsf{T}_h \textit{ and } \mathbf{w} \in \boldsymbol{\mathcal{K}},$$

where $\boldsymbol{b}(\mathsf{u}_h; \mathbf{w})$ has been introduced in (32b).

Proof. In this proof, we use the symbol $\| \cdot \|_{\mathcal{T}_h}$ to denote both the vector norm and its induced matrix norm.

If $\mathbf{w} \geq \mathbf{0}$, the matrices $\mathbf{F}$ and $\mathbf{G}(\mathbf{w})$ are irreducible Z-matrices, all of whose rows have sum equal to zero by construction (see (12) and (29)), and the matrix $\mathbf{B}$ is a nonnegative diagonal matrix with some strictly positive diagonal elements (see (30)). Thus, $\boldsymbol{\mathcal{A}}(\mathbf{w})$ is a diagonally dominant Z-matrix.

The ijth entry of $\boldsymbol{\mathcal{A}}(\mathbf{w})$, namely, $\boldsymbol{\mathcal{A}}_{ij}(\mathbf{w})$, is strictly positive if the index pair ij refers to an internal face, i.e., $\mathsf{f}_{ij} \in \mathcal{F}_h^{\text{int}}$, because in such a case $\boldsymbol{\mathcal{A}}_{ij}(\mathbf{w}) \geq C_{\text{grid}}/(d+1)$ from Lemma 7(iii). Hence, the matrix $\boldsymbol{\mathcal{A}}(\mathbf{w})$ is irreducible and, consequently, a nonsingular M-matrix.

To demonstrate the inequality of the lemma, let us consider the inequality

$$\left\|\boldsymbol{\mathcal{A}}(\mathbf{w})^{-1}\boldsymbol{b}(\mathsf{u}_h;\mathbf{w})\right\|_{\mathcal{T}_h} \leq \left\|\boldsymbol{\mathcal{A}}(\mathbf{w})^{-1}\right\|_{\mathcal{T}_h}\left(\|\mathbf{s}+\mathbf{b}\|_{\mathcal{T}_h} + \|\mathbf{r}(\mathsf{u}_h)\|_{\mathcal{T}_h} + \|\mathbf{g}(\mathbf{w})\|_{\mathcal{T}_h}\right).$$

As $\boldsymbol{\mathcal{A}}(\mathbf{w})$ and $\mathbf{g}(\mathbf{w})$ are continuous functions of $\mathbf{w}$, the scalar quantities $\left\|\,\boldsymbol{\mathcal{A}}(\mathbf{w})^{-1}\,\right\|_{\mathcal{T}_h}$ and $\|\mathbf{g}(\mathbf{w})\|_{\mathcal{T}_h}$ are also continuous functions of this argument. Since $\mathbf{w}$ ranges over the compact set $\boldsymbol{\mathcal{K}}$, $\left\|\,\boldsymbol{\mathcal{A}}(\mathbf{w})^{-1}\,\right\|_{\mathcal{T}_h}$ and $\|\mathbf{g}(\mathbf{w})\|_{\mathcal{T}_h}$ must be uniformly bounded in view of the Weierstrass theorem. Thus,

$$\left\|\boldsymbol{\mathcal{A}}(\mathbf{w})^{-1}\boldsymbol{b}(\mathsf{u}_h;\mathbf{w})\right\|_{\mathcal{T}_h} \leq \sup_{\mathbf{w}\in\boldsymbol{\mathcal{K}}}\left\|\boldsymbol{\mathcal{A}}(\mathbf{w})^{-1}\right\|_{\mathcal{T}_h}\left(\|\mathbf{s}+\mathbf{b}\|_{\mathcal{T}_h}+C_R+\sup_{\mathbf{w}\in\boldsymbol{\mathcal{K}}}\|\mathbf{g}(\mathbf{w})\|_{\mathcal{T}_h}\right), \tag{34}$$

where we also used Lemma 9 and where C_R is the constant introduced therein. The inequality of the lemma follows by setting C_M equal to the right-hand side of (34). □

After Lemma 10, we can introduce the set $\boldsymbol{\mathcal{M}} \in \mathbb{R}^{\mathsf{card}\{\mathtt{T}_h\}}$ of the admissible finite volume solutions.

Definition 11 (the set of admissible finite volume solutions). *Let*

$$\boldsymbol{\mathcal{M}} = \Big\{\mathsf{u}_h \in \mathtt{T}_h \text{ such that } \|\mathsf{u}_h\|_{\mathcal{T}_h} \leq C_M\Big\},$$

where C_M is the constant of Lemma 10.

5.2. Existence of the finite volume approximation. The finite volume approximation problem whose vector formulation has been derived in section 4.4 can be formally restated by taking into consideration Definitions 8 and 11.

Definition 12 (the finite volume approximation problem).

$$\textit{Find } \mathsf{u}_h \in \boldsymbol{\mathcal{M}} \textit{ and } \mathbf{w} \in \boldsymbol{\mathcal{K}} \textit{ satisfying } \text{(33a)–(33b)}. \tag{35}$$

This problem can also be reformulated in the fixed-point form

$$\textit{find } (\mathsf{u}_h,\,\mathbf{w}) \in \boldsymbol{\mathcal{M}}\times\boldsymbol{\mathcal{K}} \textit{ such that } \begin{bmatrix} \mathsf{u}_h \\ \mathbf{w} \end{bmatrix} = \mathfrak{F}\begin{bmatrix} \mathsf{u}_h \\ \mathbf{w} \end{bmatrix}, \tag{36}$$

where the nonlinear mapping $\mathfrak{F} : \boldsymbol{\mathcal{M}}\times\boldsymbol{\mathcal{K}} \to \boldsymbol{\mathcal{M}}\times\boldsymbol{\mathcal{K}}$ is given by

$$\mathfrak{F}\begin{bmatrix} \mathsf{u}_h \\ \mathbf{w} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}(\mathsf{u}_h;\mathbf{w}) \\ \boldsymbol{\psi}(\mathsf{u}_h) \end{bmatrix},$$

and

$$\boldsymbol{\Phi}(\mathsf{u}_h;\mathbf{w}) = \boldsymbol{\mathcal{A}}(\mathbf{w})^{-1}\boldsymbol{b}(\mathsf{u}_h;\mathbf{w}).$$

Unfortunately, $\mathfrak{F}(\mathsf{u}_h,\,\mathbf{w})$ is not guaranteed to be a continuous function of its arguments because the mapping $\boldsymbol{\psi}(\mathsf{u}_h) = \{w_k^{ij}(\mathsf{u}_h)\}$ may be discontinuous at some internal face

$\mathsf{f}_{ij} \in \mathcal{F}_h^{\text{int}}$. For this reason, Brouwer's fixed-point theorem cannot be directly applied to demonstrate the existence of a solution to (36) and hence to (35).

We thus proceed in a different way. First, we define a special "regularization" mapping $\boldsymbol{\psi}^{\varepsilon}(\cdot)$ that is continuous in its argument and depends on the positive parameter ε. Replacing $\boldsymbol{\psi}(\mathsf{u}_h) = \{w_k^{ij}(\mathsf{u}_h)\}$ by $\boldsymbol{\psi}^{\varepsilon}(\mathsf{u}_h^{\varepsilon}) = \{w_k^{\varepsilon ij}(\mathsf{u}_h^{\varepsilon})\}$ in (27) regularizes both the face gradient formula and the numerical diffusive flux. Brouwer's fixed-point theorem is applicable to the regularized approximation problems that are defined for any $\varepsilon \geq 0$. Then we take the limit of $\varepsilon \to 0$, and a compactness argument proves that a converging subsequence can be extracted from the solutions of these regularized problems. Finally, we demonstrate that the limit of this subsequence of regularized solutions solves the finite volume approximation problem stated in (35).

Let us introduce, for any control volume T_i and any face $\mathsf{f}_{ij} \in \mathcal{F}_h$, $j \in \sigma_i \cup \sigma_i'$, the regularized continuous weights

$$\omega_{ij}^{\varepsilon}(\mathsf{u}_h) = \frac{|g_{ij}(\mathsf{u}_h)|}{|g_{ij}(\mathsf{u}_h)| + \varepsilon} \, \frac{|g_{ji}(\mathsf{u}_h)|}{|g_{ji}(\mathsf{u}_h)| + \varepsilon} \, \omega_{ij}^{0}(\mathsf{u}_h). \tag{37}$$

These weights are designed to satisfy the following important property.

LEMMA 13. *If* $\mathsf{u}_h^{\varepsilon} \to \mathsf{u}_h$ *for* $\varepsilon \to 0$*, then at each internal face* $\mathsf{f}_{ij} \in \mathcal{F}_h^{\text{int}}$ *we have*

$$\lim_{\varepsilon \to 0} \omega_{ij}^{\varepsilon}(\mathsf{u}_h^{\varepsilon}) g_{ij}(\mathsf{u}_h^{\varepsilon}) = \omega_{ij}^{0}(\mathsf{u}_h) g_{ij}(\mathsf{u}_h).$$

Proof. As $g_{ij}(\cdot)$ and $g_{ji}(\cdot)$ are continuous functions of their argument, $\mathsf{u}_h^{\varepsilon} \to \mathsf{u}_h$ for $\varepsilon \to 0$ implies $g_{ij}(\mathsf{u}_h^{\varepsilon}) \to g_{ij}(\mathsf{u}_h)$ and $g_{ji}(\mathsf{u}_h^{\varepsilon}) \to g_{ji}(\mathsf{u}_h)$. Furthermore, when $g_{ij}(\mathsf{u}_h) g_{ji}(\mathsf{u}_h) \neq 0$ a real number $\varepsilon' > 0$ must exist such that the signs of $g_{ij}(\mathsf{u}_h^{\varepsilon}) g_{ji}(\mathsf{u}_h^{\varepsilon})$ and $g_{ij}(\mathsf{u}_h) g_{ji}(\mathsf{u}_h)$ coincide for any $\varepsilon \leq \varepsilon'$. Let us distinguish among the following three cases.

(i) $g_{ij}(\mathsf{u}_h) g_{ji}(\mathsf{u}_h) < 0$.

Then $g_{ij}(\mathsf{u}_h^{\varepsilon}) g_{ji}(\mathsf{u}_h^{\varepsilon})$ is also negative for any $\varepsilon \leq \varepsilon'$, and from the definitions in (26) and (23)–(25) it follows that

$$\begin{aligned} \lim_{\varepsilon \to 0} \omega_{ij}^{\varepsilon}(\mathsf{u}_h^{\varepsilon}) g_{ij}(\mathsf{u}_h^{\varepsilon}) &= \lim_{\varepsilon \to 0} \frac{|g_{ji}(\mathsf{u}_h^{\varepsilon})|}{|g_{ij}(\mathsf{u}_h^{\varepsilon})| + \varepsilon} \, \frac{|g_{ji}(\mathsf{u}_h^{\varepsilon})|}{|g_{ji}(\mathsf{u}_h^{\varepsilon})| + \varepsilon} \, \frac{|g_{ji}(\mathsf{u}_h^{\varepsilon})|}{|g_{ij}(\mathsf{u}_h^{\varepsilon})| + |g_{ji}(\mathsf{u}_h^{\varepsilon})|} g_{ij}(\mathsf{u}_h^{\varepsilon}) \\ &= \frac{|g_{ji}(\mathsf{u}_h)|}{|g_{ij}(\mathsf{u}_h)| + |g_{ji}(\mathsf{u}_h)|} g_{ij}(\mathsf{u}_h) = \omega_{ij}^{0}(\mathsf{u}_h) g_{ij}(\mathsf{u}_h) \end{aligned}$$

due to the continuity of $g_{ij}(\cdot)$, $g_{ji}(\cdot)$ and because $\mathsf{u}_h^{\varepsilon} \to \mathsf{u}_h$.

(ii) $g_{ij}(\mathsf{u}_h) g_{ji}(\mathsf{u}_h) = 0$.

Definitions (23)–(25) yield $\omega_{ij}^{0}(\mathsf{u}_h) = 0$. As $g_{ij}(\cdot)$ and $g_{ji}(\cdot)$ are continuous functions, at least one of them approaches zero as $\mathsf{u}_h^{\varepsilon} \to \mathsf{u}_h$ for $\varepsilon \to 0$. The statement of the lemma follows by noting that whenever $g_{ij}(\mathsf{u}_h^{\varepsilon}) \neq 0$ we have

$$\left| \omega_{ij}^{\varepsilon}(\mathsf{u}_h^{\varepsilon}) g_{ij}(\mathsf{u}_h^{\varepsilon}) \right| \leq \left| \omega_{ij}^{0}(\mathsf{u}_h^{\varepsilon}) g_{ij}(\mathsf{u}_h^{\varepsilon}) \right| \leq \frac{|g_{ji}(\mathsf{u}_h^{\varepsilon})| \, |g_{ij}(\mathsf{u}_h^{\varepsilon})|}{|g_{ij}(\mathsf{u}_h^{\varepsilon})| + |g_{ji}(\mathsf{u}_h^{\varepsilon})|},$$

and taking the limit for $\varepsilon \to 0$.

(iii) $g_{ij}(\mathsf{u}_h) g_{ji}(\mathsf{u}_h) > 0$.

The statement of the lemma is trivially true because from definitions (23)–(25) and (37) it follows that $\omega_{ij}^{0}(\mathsf{u}_h) = 0$ and $\omega_{ij}^{\varepsilon}(\mathsf{u}_h^{\varepsilon}) = 0$ for $\varepsilon \leq \varepsilon'$. □

Let $\boldsymbol{\psi}^{\varepsilon}(\mathsf{u}_h)$ be the *regularized mapping* obtained by using $\omega_{ij}^{\varepsilon}(\mathsf{u}_h)$ in place of $\omega_{ij}(\mathsf{u}_h)$ in (26).

Theorem 14. *The regularized problem*

$$\begin{aligned}&\textit{find } \mathsf{u}_h^\varepsilon \in \boldsymbol{\mathcal{M}} \textit{ and } \mathbf{w}^\varepsilon \in \boldsymbol{\mathcal{K}} \textit{ such that}\\ &\boldsymbol{\mathcal{A}}(\mathbf{w}^\varepsilon)\mathsf{u}_h^\varepsilon = \boldsymbol{b}(\mathsf{u}_h^\varepsilon;\mathbf{w}^\varepsilon),\\ &\mathbf{w}^\varepsilon = \boldsymbol{\psi}^\varepsilon(\mathsf{u}_h^\varepsilon)\end{aligned} \tag{38}$$

admits at least one solution for any $\varepsilon > 0$.

Proof. Let us consider the mapping $\mathfrak{F}^\varepsilon : \boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{K}} \to \boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{K}}$ defined by

$$\mathfrak{F}^\varepsilon \begin{bmatrix} \mathsf{u}_h^\varepsilon \\ \mathbf{w}^\varepsilon \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}(\mathsf{u}_h^\varepsilon;\mathbf{w}^\varepsilon) \\ \boldsymbol{\psi}^\varepsilon(\mathsf{u}_h^\varepsilon) \end{bmatrix}.$$

Problem (38) can be reformulated as the fixed-point problem

$$\textit{find } (\mathsf{u}_h^\varepsilon,\, \mathbf{w}^\varepsilon) \in \boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{K}} \textit{ such that } \begin{bmatrix} \mathsf{u}_h^\varepsilon \\ \mathbf{w}^\varepsilon \end{bmatrix} = \mathfrak{F}^\varepsilon \begin{bmatrix} \mathsf{u}_h^\varepsilon \\ \mathbf{w}^\varepsilon \end{bmatrix}.$$

As $\mathfrak{F}^\varepsilon$ is a continuous mapping from the compact convex set $\boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{K}}$ into itself, this fixed-point problem admits at least one solution in view of Brouwer's fixed-point theorem. □

Theorem 15. *The finite volume approximation problem* (35) *admits at least one solution.*

Proof. Let us consider the set of regularized solutions of (38) for $\varepsilon > 0$. This set is a subset of the compact set $\boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{K}}$, and thus a subsequence exists that converges to a limit solution $(\mathsf{u}_h,\, \mathbf{w}) \in \boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{K}}$. Thus,

$$\begin{bmatrix} \mathsf{u}_h^{\varepsilon_k} \\ \mathbf{w}^{\varepsilon_k} \end{bmatrix} \to \begin{bmatrix} \mathsf{u}_h \\ \mathbf{w} \end{bmatrix} \in \boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{K}} \quad \text{for } k \to \infty, \tag{39}$$

where $\{\varepsilon_k\}$ is a suitable sequence of nonnegative "epsilons" converging to zero for $k \to \infty$.

The crucial fact to be proved is that the limit solution $(\mathsf{u}_h,\, \mathbf{w})$ in (39) actually solves (35). To achieve this task, we show that $\|\boldsymbol{\mathcal{A}}(\mathbf{w})\mathsf{u}_h - \boldsymbol{b}(\mathsf{u}_h;\mathbf{w})\|_{\mathcal{T}_h}$ is an infinitesimal quantity. First, we subtract the quantity $\boldsymbol{\mathcal{A}}(\mathbf{w}^{\varepsilon_k})\mathsf{u}_h^{\varepsilon_k} - \boldsymbol{b}(\mathsf{u}_h^{\varepsilon_k};\mathbf{w}^{\varepsilon_k}) = 0$ for $\varepsilon_k > 0$ to $\boldsymbol{\mathcal{A}}(\mathbf{w})\mathsf{u}_h - \boldsymbol{b}(\mathsf{u}_h;\mathbf{w})$, rearrange the terms, take the $\|\cdot\|_{\mathcal{T}_h}$-norm, and apply the triangle inequality to get

$$\begin{aligned}\|\boldsymbol{\mathcal{A}}(\mathbf{w})\mathsf{u}_h - \boldsymbol{b}(\mathsf{u}_h;\mathbf{w})\|_{\mathcal{T}_h} &\le \|\mathbf{F}(\mathsf{u}_h - \mathsf{u}_h^{\varepsilon_k})\|_{\mathcal{T}_h}\\ &+ \|\mathbf{G}(\boldsymbol{\psi}(\mathsf{u}_h))\mathsf{u}_h + \mathbf{g}(\boldsymbol{\psi}(\mathsf{u}_h)) - \mathbf{G}(\boldsymbol{\psi}^{\varepsilon_k}(\mathsf{u}_h^{\varepsilon_k}))\mathsf{u}_h^{\varepsilon_k} - \mathbf{g}(\boldsymbol{\psi}^{\varepsilon_k}(\mathsf{u}_h^{\varepsilon_k}))\|_{\mathcal{T}_h}\\ &+ \|\mathbf{B}(\mathsf{u}_h - \mathsf{u}_h^{\varepsilon_k})\|_{\mathcal{T}_h} + \|\mathbf{r}(\mathsf{u}_h) - \mathbf{r}(\mathsf{u}_h^{\varepsilon})\|_{\mathcal{T}_h}.\end{aligned}$$

Then we note that all of the terms in this inequality except the second one can be readily shown to be infinitesimal for $k \to \infty$ by a simple continuity argument. Instead, the second term has to be more carefully treated due to its nonlinear dependence on u_h and $\mathsf{u}_h^{\varepsilon_k}$ throughout the mappings $\boldsymbol{\psi}(\cdot)$ and $\boldsymbol{\psi}^{\varepsilon_k}(\cdot)$, respectively. Let us observe that this second term describes the difference between the numerical diffusive flux at any internal face calculated using the regularized solution $\mathsf{u}_h^{\varepsilon_k}$ and the one using the limit solution u_h. The ith component satisfies the inequality

$$\begin{aligned}
&\left| \mathbf{G}(\boldsymbol{\psi}(\mathsf{u}_h))\mathsf{u}_h\Big|_i - \mathbf{G}(\boldsymbol{\psi}^{\varepsilon_k}(\mathsf{u}_h^{\varepsilon_k}))\mathsf{u}_h^{\varepsilon_k}\Big|_i + \mathsf{g}_i(\boldsymbol{\psi}(\mathsf{u}_h)) - \mathsf{g}_i(\boldsymbol{\psi}^{\varepsilon_k}(\mathsf{u}_h^{\varepsilon_k})) \right| \\
&\qquad \leq \frac{1}{|\mathsf{T}_i|} \sum_{j\in\sigma_i} |\mathsf{f}_{ij}|\, \nu_{ij} D_{ij} \left| \frac{\tilde{u}_j^{\varepsilon_k} - u_i^{\varepsilon_k}}{H_{ij}} - \frac{\tilde{u}_j - u_i}{H_{ij}} \right| \\
&\qquad + 2\frac{1}{|\mathsf{T}_i|} \sum_{j\in\sigma_i} |\mathsf{f}_{ij}|\, \nu_{ij} \left| w_{ij}(\mathsf{u}_h) g_{ij}(\mathsf{u}_h) - w_{ij}^{\varepsilon_k}(\mathsf{u}_h^{\varepsilon_k}) g_{ij}(\mathsf{u}_h^{\varepsilon_k}) \right|.
\end{aligned}$$

The right-hand side of this inequality approaches zero when $k \to \infty$ because $u_i^{\varepsilon_k} \to u_i$, $\tilde{u}_{ij}^{\varepsilon_k} \to \tilde{u}_{ij}$ due to the continuity of the reconstruction process, and

$$\left| w_{ij}(\mathsf{u}_h) g_{ij}(\mathsf{u}_h) - w_{ij}^{\varepsilon_k}(\mathsf{u}_h^{\varepsilon_k}) g_{ij}(\mathsf{u}_h^{\varepsilon_k}) \right| \to 0$$

thanks to Lemma 13. □

5.3. The maximum principle for the finite volume approximation. In this section, we demonstrate the existence of a maximum principle for the solution of the finite volume approximation problem formulated in (35).

To achieve this task, we proceed as follows. We restrict the finite volume approximation problem (35) to $\boldsymbol{\mathcal{M}}$ by taking a constant weight vector $\mathbf{w} \in \boldsymbol{\mathcal{K}}$ in (33a), thus relaxing the dependence on the solution u_h through (33b). This restricted problem will be shown to admit at least one discrete solution in Theorem 16, and all of its solutions will be proved to satisfy a discrete maximum principle in Theorem 17. Finally, the solution of (35) will be shown to satisfy a discrete maximum principle as a consequence of Theorem 17 in Corollary 18.

THEOREM 16. *The nonlinear fixed-point problem that involves the mapping* $\boldsymbol{\Phi}(\cdot, \mathbf{w})$ *restricted to a constant* $\mathbf{w} \in \boldsymbol{\mathcal{K}}$,

$$\textit{find } \mathsf{u}_h \in \boldsymbol{\mathcal{M}} \textit{ such that } \mathsf{u}_h = \boldsymbol{\Phi}(\mathsf{u}_h; \mathbf{w}),$$

admits at least one solution.

Proof. The restriction $\boldsymbol{\Phi}(\cdot; \mathbf{w})$ to $\boldsymbol{\mathcal{M}}$ provided by any specific choice of the weight vector $\mathbf{w} \in \boldsymbol{\mathcal{K}}$ is a continuous mapping from the convex compact set $\boldsymbol{\mathcal{M}}$ into itself. The statement of the theorem follows by the application of Brouwer's fixed-point theorem. □

The following theorem states that the solution of this restricted problem satisfies a discrete maximum principle. The proof extends to unstructured finite volumes some standard arguments used for proving similar results for finite differences. Nonetheless, we report this proof in some detail to illustrate the role played by the positivity of the factors D_{ij} and the nonnegativity of the weights $\{w_k^{ij}(\mathsf{u}_h)\}$ in the definition of the face gradient.

THEOREM 17 (maximum principle for the solution of the nonlinear restricted problem). *Let us consider problem* (33a)–(33b) *with definitions* (32a)–(32b)*. Let* $\mathbf{s} \leq 0$ *in* (32b)*, and assume that* $\mathbf{w} \geq \mathbf{0}$ *be taken constant and that* $\mathsf{u}_h \in \boldsymbol{\mathcal{M}}$ *satisfies*

$$\boldsymbol{\mathcal{A}}(\mathbf{w})\mathsf{u}_h - \boldsymbol{b}(\mathsf{u}_h; \mathbf{w}) \leq \mathbf{0}. \tag{40}$$

Then u_h *satisfies the* discrete maximum principle *expressed in the inequality form*

$$\max_{\mathsf{T}_k\in\mathcal{T}_h} u_k \leq \max\Big[0, \max_{\mathsf{v}_\alpha\in\mathcal{V}_h^{\mathrm{bnd}}} \mathsf{g}(\boldsymbol{x}_\alpha)\Big].$$

Proof. Let us assume that there is a cell T_i such that

$$u_i = \max_{\mathsf{T}_k \in \mathcal{T}_h} u_k > \max\Big[0, \max_{\mathsf{v}_\alpha \in \mathcal{V}_h^{\mathrm{bnd}}} \mathsf{g}(\boldsymbol{x}_\alpha)\Big]. \tag{41}$$

We will demonstrate that this assumption contradicts the hypothesis of the theorem.

By rearranging (7) in order to collect the contribution from both internal and boundary faces, and using the definitions introduced in section 4, we reformulate the numerical flux balance for the cell T_i as the sum of the advective terms, labeled $\{\text{Adv. Terms}\}$, and the diffusive terms, labeled $\{\text{Diff. Terms}\}$:

$$\begin{aligned}\{\text{Adv. Terms}\} = {}& \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i} |\mathsf{f}_{ij}| \Big[v_{ij}^{+}\mathcal{R}_i(\mathsf{u}_h)(\boldsymbol{x}_{ij}) + v_{ij}^{-}\mathcal{R}_j(\mathsf{u}_h)(\boldsymbol{x}_{ij}) \Big] \\ &+ \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i'} |\mathsf{f}_{ij}| \Big[v_{ij}^{+}\mathcal{R}_i(\mathsf{u}_h)(\boldsymbol{x}_{ij}) + v_{ij}^{-} g_{ij}^{D} \Big]\end{aligned}$$

and

$$\begin{aligned}\{\text{Diff. Terms}\} = {}& \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i} \frac{|\mathsf{f}_{ij}|\,\nu_{ij} D_{ij}}{H_{ij}}(u_i - u_j) \\ &+ \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i} \frac{|\mathsf{f}_{ij}|\,\nu_{ij}}{H_{ij}} \Bigg(\sum_{k\in\sigma\nu_j} w_k^{ij}(u_i-u_k) + w_i^{ij}(u_i - g_{ij}^{\mathrm{bnd}}) \Bigg) \\ &+ \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i'} |\mathsf{f}_{ij}|\,\nu_{ij} \frac{u_i - \widetilde{g}_{ij}^{D}}{h_{ij}}.\end{aligned}$$

Both $\{\text{Adv. Terms}\}$ and $\{\text{Diff. Terms}\}$ are nonnegative quantities. Since from Lemma 6 $\mathcal{R}_i(\mathsf{u}_h)(\boldsymbol{x}) = u_i$ for $\boldsymbol{x} \in \mathsf{T}_i$, we can rewrite $\{\text{Adv. Terms}\}$ as

$$\begin{aligned}\{\text{Adv. Terms}\} = {}& \frac{u_i}{|\mathsf{T}_i|}\sum_{j\in\sigma_i\,\cup\,\sigma_i'} |\mathsf{f}_{ij}|\, v_{ij} + \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i} |\mathsf{f}_{ij}|\, v_{ij}^{-}\big(\mathcal{R}_j(\mathsf{u}_h)(\boldsymbol{x}_{ij}) - u_i\big) \\ &+ \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i'} |\mathsf{f}_{ij}|\, v_{ij}^{-}(g_{ij}^{D} - u_i),\end{aligned} \tag{42}$$

and note that the three terms on the right-hand side of (42) are nonnegative. The first term in the sum can be transformed as

$$\sum_{j\in\sigma_i\,\cup\,\sigma_i'} |\mathsf{f}_{ij}|\, v_{ij} = \sum_{j\in\sigma_i\,\cup\,\sigma_i'} \int_{\mathsf{f}_{ij}} \boldsymbol{n}_{ij}\cdot\boldsymbol{v}\, dS = \int_{\partial\mathsf{T}_i} \boldsymbol{n}\cdot\boldsymbol{v}\, dS = \int_{\mathsf{T}_i} \mathrm{div}\boldsymbol{v}\, dV \geq 0,$$

and is nonnegative because $\mathrm{div}\,\boldsymbol{v} \geq 0$ by the initial assumption (2)(ii) and $u_i > 0$ from the proof assumption in (41). The second term in the sum is nonnegative because $v_{ij}^{-} \leq 0$ by definition and $\mathcal{R}_j(\mathsf{u}_h)(\boldsymbol{x}_{ij}) \leq u_i$ in view of the limiter constraints (9). The third term in the sum is nonnegative because $v_{ij}^{-} \leq 0$ as before, and $g_{ij}^{D} < u_i$ by (31) and the assumption stated in (41).

Similarly, the entries in the sum comprising $\{\text{Diff. Terms}\}$ are all nonnegative quantities. Indeed, they contain scalar coefficients that are nonnegative by construction, such as ν_{ij}, H_{ij}, D_{ij}, and $\{w_k^{ij}\}$, which multiply difference terms like $(u_i - u_k)$

for $k \in \sigma\nu_i$ or $(u_i - g^{\rm bnd}_{ij})$ and $(u_i - \widetilde{g}^D_{ij})$ for $j \in \sigma_i \cap \sigma'_i$. All of these difference terms are nonnegative as a consequence of the assumption stated in (41).

Since $\{\textsc{Adv. Terms}\} + \{\textsc{Diff. Terms}\} - \mathsf{s}_i = (\boldsymbol{\mathcal{A}}(\mathbf{w})\mathsf{u}_h - \boldsymbol{b}(\mathsf{u}_h;\mathbf{w}))\big|_i \leq 0$, it follows that

$$\begin{aligned} \mathsf{s}_i &\geq \{\textsc{Adv. Terms}\} + \{\textsc{Diff. Terms}\} \\ &\geq \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma_i} |\mathsf{f}_{ij}|\,\nu_{ij}D_{ij}\frac{u_i-u_j}{H_{ij}} + \frac{1}{|\mathsf{T}_i|}\sum_{j\in\sigma'_i} |\mathsf{f}_{ij}|\,\nu_{ij}\frac{u_i-\widetilde{g}^D_{ij}}{h_{ij}} \geq 0, \end{aligned} \tag{43}$$

where the second inequality is obtained by canceling $\{\textsc{Adv. Terms}\} \geq 0$ and some other nonnegative terms in $\{\textsc{Diff. Terms}\}$. As $\mathbf{s} \leq \mathbf{0}$ by the hypothesis of the theorem, relation (43) implies that $u_j = u_i$ for *any* $j \in \sigma_i$ because the coefficients D_{ij} of the scheme are *all strictly positive* and $\widetilde{g}^D_{ij} = u_i$ for any $j \in \sigma'_i$. By repeating this argument on the control volumes adjacent to the cell T_i and exploiting also the connectivity of the mesh and the fact that $D_{ij} > 0$ for any $\mathsf{f}_{ij} \in \mathcal{F}^{\rm int}_h$, we eventually obtain that u_h must be constant over all the mesh cells and equal to the boundary data $\widetilde{g}^D_{ij}$.

We emphasize that any solution u_h consistent with the hypothesis (40) of the theorem and the assumption (41) must satisfy the condition proved above. Nonetheless, this result then contradicts (41) because $\widetilde{g}^D_{ij}$ is the convex linear combination of the vertex boundary values $\mathsf{g}(\boldsymbol{x}_\alpha)$ for $\mathsf{v}_\alpha \in \mathcal{V}^{\rm bnd}_h$ (see (31)), and this should imply $u_i \leq \max_{\mathsf{v}_\alpha\in\mathcal{V}^{\rm bnd}_h} \mathsf{g}(\boldsymbol{x}_\alpha)$. ∎

Corollary 18. *Let* $\mathbf{s}(\boldsymbol{x}) \leq 0$*, and let* $(\mathsf{u}_h, \mathbf{w}) \in \boldsymbol{\mathcal{M}} \times \boldsymbol{\mathcal{K}}$ *be a solution to the finite volume approximation problem* (35)*. Then* u_h *satisfies the maximum principle in the same discrete form as in Theorem* 17*; that is,*

$$\max_{\mathsf{T}_k\in\mathcal{T}_h} u_k \leq \max\left\{0,\ \max_{\mathsf{v}_\alpha\in\mathcal{V}^{\rm bnd}_h} \mathsf{g}(\boldsymbol{x}_\alpha)\right\}.$$

Proof. In view of the corollary's hypothesis, $\mathsf{u}_h \in \boldsymbol{\mathcal{M}}$ satisfies the fixed-point relation $\mathsf{u}_h = \boldsymbol{\Phi}(\mathsf{u}_h;\mathbf{w})$ with $\mathbf{w} = \boldsymbol{\psi}(\mathsf{u}_h) \in \boldsymbol{\mathcal{K}}$. Thus, u_h solves the restricted problem defined by taking these weights $\mathbf{w}$ constant, and the statement of the corollary is a straightforward consequence of Theorem 17. ∎

Remark 19. A minimum principle can be proved when $\mathbf{s}(\boldsymbol{x}) \geq 0$ and can be expressed in the form

$$\min_{\mathsf{T}_k\in\mathcal{T}_h} u_k \geq \min\left\{0,\ \min_{\mathsf{v}_\alpha\in\mathcal{V}^{\rm bnd}_h} \mathsf{g}(\boldsymbol{x}_\alpha)\right\}.$$

The proof essentially repeats the same arguments as in the proofs of Theorem 17 and Corollary 18, but the symbols max, min and "$\geq$", "$\leq$" must be consistently interchanged.

6. Numerical experiments. In this section we present the performance of the method on two different test cases that are taken from the recent literature. Both test cases require the resolution of an advection-diffusion problem that is strongly dominated by the convection term.

The first test case is taken from section 7.2 of [8]. A thorough comparison of the behavior of the diamond scheme of [5] and the nonlinear diamond scheme of this paper is carried out. In particular, we focus on the violation of the discrete maximum principle and the additional computational cost for solving the nonlinear diamond scheme. Our investigation reveals that the second-order accurate solution provided by the nonlinear diamond scheme exhibits a violation of the discrete maximum principle by an error less than 10^{-12} against about 10^{-6}–10^{-5} of the diamond scheme. The additional cost for ensuring the satisfaction of the discrete maximum principle is about 10–15% of the cost for solving the same problem by using the diamond scheme.

The second test case is taken from Examples 5.1 and 5.2 of [27] to investigate the order of accuracy of the nonlinear diamond scheme when numerically approximating strong gradient solutions. The strength of the gradients of the solution is controlled by varying the ratio between the viscosity coefficient and the absolute value of the velocity field; note that this ratio is proportional to the Péclet number [27]. As theoretically expected, second order of convergence is observed over a wide range of this parameter. However, deterioration may occur for the most difficult case, and second-order convergence is then lost. This behavior is similar to the behavior of the scheme presented in [27].

Both the diamond scheme and the nonlinear diamond scheme require the resolution of a nonlinear algebraic problem. To achieve this task, we consider the quasi-Newton iterative scheme GIANT described in [12]. This scheme is very efficient in solving general nonlinear problems and seems to be particularly suitable in linearizing the Jacobian matrix of the nonlinear functional (36). The nonlinear iterations are stopped when the residual is smaller than 10^{-10}; this threshold value is the same one that ensures the violation of the maximum principle of the order 10^{-12} in the test case of section 6.1. The linear algebraic problems arising from the Jacobian linearization at the quasi-Newton iterative steps are solved by applying the Bi-CGSTAB algorithm [30] preconditioned by an incomplete LU (ILU) factorization [16]. The mesh data structures are managed by P2MESH [4], which is a C^{++} public domain library designed for fast and efficient implementation of partial differential equation solvers.

6.1. Test case 1. We investigate a pure convection-diffusion problem with dominant convection. The problem is posed on the computational domain $\Omega = [0,\, 1]\times[0,\, 1]$ by using the constant velocity field $\boldsymbol{v} = -(\cos\theta, \sin\theta)$ with $\theta = 55°$, and the scalar viscosity coefficient $\nu = 10^{-5}$.

Dirichlet boundary conditions are set as follows: $\mathsf{u} = 0$ on the bottom edge; $\mathsf{u} = 1$ on the left and top edges; u steps from 0 to 1 at $y = 4/5$ on the right edge. The solution exhibits an internal layer across the domain and a boundary layer along a part of the bottom edge. As the discrete maximum principle is theoretically demonstrated by assuming that the mesh is weakly acute, we consider, as in [8], the mesh obtained by splitting into two triangles the square cells of a tensor product uniform mesh with 40 partitions per side of the computational domain. For this mesh, we solve the advection-diffusion problem by using the diamond scheme and the nonlinear diamond scheme. The two solutions are shown in Figure 3, where constant level curves are depicted for discrete values from 0.01 to 0.99. A comparison of the two pictures in Figure 3 shows that no additional smearing of the internal layer takes place in the solution approximated by the nonlinear diamond scheme. The width of the internal layer is approximately the same for both schemes because the satisfaction of the discrete maximum principle is not based on the addition of some form of artificial or cross-wind diffusion to the standard second-order version of the diamond scheme.

Similar results (not shown) were also obtained by repeating these calculations on a truly unstructured mesh with a comparable number of triangles and all angles smaller than 90°. The performance of the two schemes for the two different kinds of

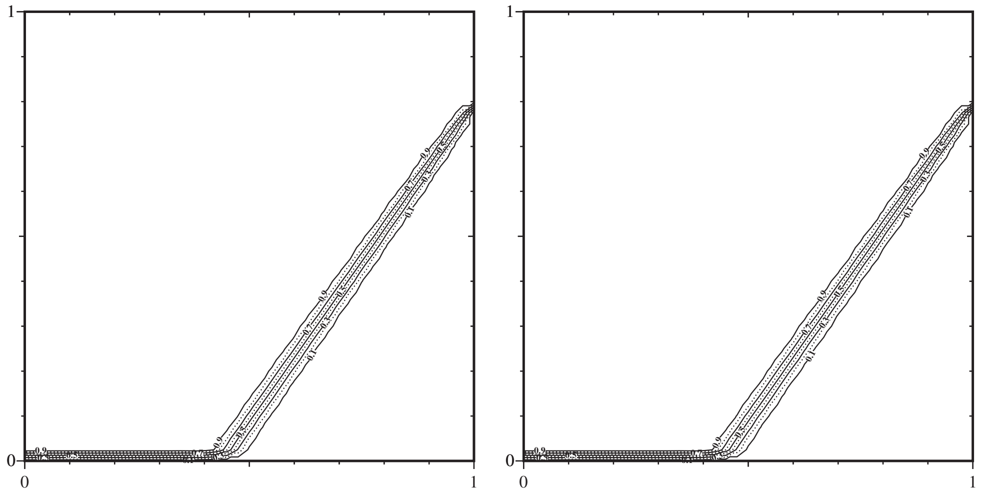


Fig. 3. *Test case* 1: *Linearly reconstructed vertex solution provided by the diamond scheme (left) and the nonlinear diamond scheme (right).*

Table 1

Test case 1: *Performance in terms of quality and computational cost for the calculation with a regular mesh violating the strictly acute mesh condition (top) and an unstructured mesh satisfying the strictly acute mesh condition (bottom).*

Regular mesh ($40 \times 40 \times 2$ triangles)				
Scheme	Overshoot	Undershoot	Layer width	Normalized cost
Diamond	2.42×10^{-4}	1.03×10^{-6}	0.1	1
NL diamond	5.65×10^{-12}	3.01×10^{-12}	0.1	1.12

Unstructured mesh (4352 triangles)				
Scheme	Overshoot	Undershoot	Layer width	Normalized cost
Diamond	8.87×10^{-5}	2.34×10^{-6}	0.1	1
NL diamond	9.48×10^{-11}	6.46×10^{-11}	0.1	1.13

meshes, i.e., the regular one and the unstructured one, is summarized by Table 1. The measure of the absolute values of the maximum overshoot and undershoot, which is a measure of the violation of the maximum and minimum principle, reveals that the nonlinear diamond scheme actually provides a better approximation on both meshes. Table 1 also gives the relative cost for the calculation using the two methods on the two grids and the layer width approximately taken at $y = 0.5$ on the two pictures. It is clear from this table that the nonlinear diamond scheme makes it possible to achieve a higher quality approximation at a very small additional cost, this latter one being no more than 15% of the cost of the diamond scheme for solving the same problem. An explanation of this fact is possible by comparing the convergence curves for the calculations that use the two schemes. The two convergence curves are illustrated in Figure 4. From this figure, we note that the number of quasi-Newton iterations does not significantly change when the nonlinear discretization is considered for the diffusive term. We observe that the maximum number of quasi-Newton iterations that

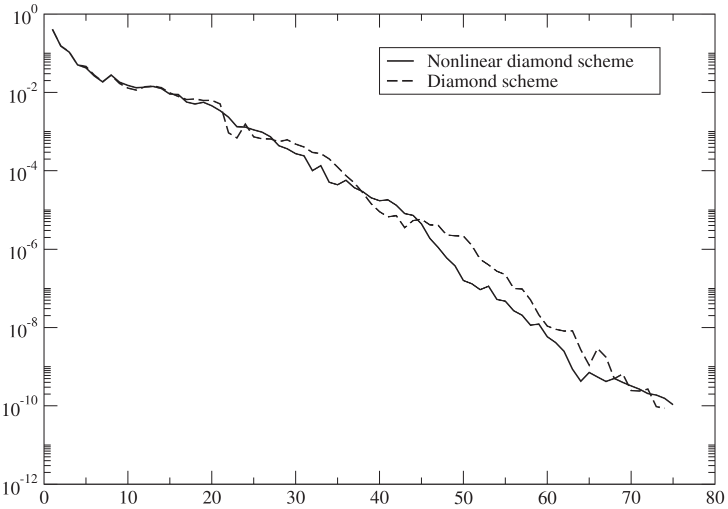


FIG. 4. *Test case* 1: *Comparison of the convergence curves for the quasi-Newton iterations.*

are necessary to attain a residual error under 10^{-10} is about the same: 75 iterations for the nonlinear diamond scheme and 74 iterations for the diamond scheme. The difference between the costs of the two calculations is mainly due to the different number of "internal" Bi-CGSTAB iterations that are required to solve the linearized problem at each step of the quasi-Newton "external" loop. The total number of internal Bi-CGSTAB iterations is 1970 for the diamond scheme and 2198 for the nonlinear diamond scheme.

From this numerical experiment we conclude that the nonlinear discretization of the diffusive term that is proposed in this paper effectively ensures the preservation of a discrete maximum principle. This linearity does not dramatically affect the computational cost compared to the cost of solving the same problem by the standard second-order version of the diamond scheme.

6.2. Test case 2. In this section, we investigate the performance of the nonlinear diamond scheme when numerically solving the two steady convection-diffusion problems of Examples 5.1 and 5.2 of [27], which will be respectively denoted "Problem A" and "Problem B." Both problems are defined on the computational domain $\Omega = [0,\,1] \times [0,\,1]$. The domain Ω is covered by a regular triangulation of 40×40 square-shaped cells, each one of these being divided into two triangles by using the same main diagonal. We run three different calculations by using $\nu \in \{1, 10^{-1}, 10^{-5}\}$. The source function s appearing on the right-hand side of the equation and the boundary data g are chosen so that the exact solution u is the one indicated in the following problem specifications. The velocity field $\boldsymbol{v}$ is also indicated below.

Problem A:

$$\boldsymbol{v} = (-1,-1)^T, \qquad \mathsf{u}(x,y) = x\cos(\pi y).$$

Problem B:

$$\boldsymbol{v} = (2,3)^T,$$
$$\mathsf{u}(x,y) = xy^2 - y^2\exp\Big(2\frac{x-1}{\nu}\Big) - x\exp\Big(3\frac{y-1}{\nu}\Big) + \exp\Big(\frac{2(x-1)+3(y-1)}{\nu}\Big).$$

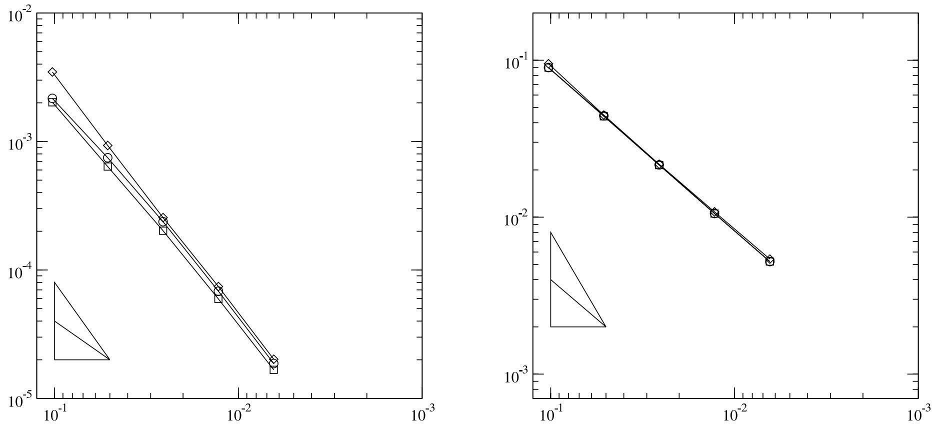


FIG. 5. *Test case 2, Problem* A: *Error convergence rates measured by using the* $\mathtt{L}^2(\Omega)$*-norm (left) and the* $\mathtt{H}^1(\Omega)$*-norm (right); the results concern the approximation of the solution for* $\nu = 1$ *(circles),* $\nu = 10^{-1}$ *(squares), and* $\nu = 10^{-5}$ *(diamonds); second-order and first-order convergence slopes are indicated in the bottom-left corner of both pictures.*

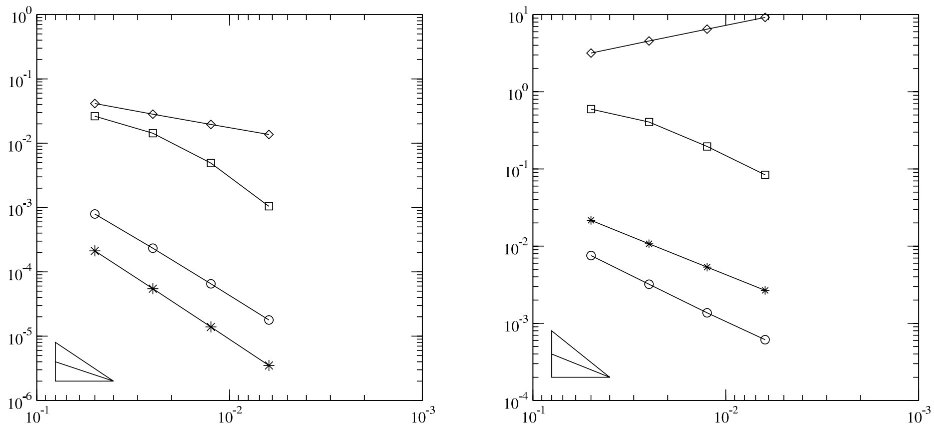


FIG. 6. *Test case 2, Problem* B: *Error convergence rates measured by using the* $\mathtt{L}^2(\Omega)$*-norm (left) and the* $\mathtt{H}^1(\Omega)$*-norm (right); the results concern the approximation of the solution on* $\Omega = [0,1]\times[0,1]$ *for* $\nu = 1$ *(circles),* $\nu = 10^{-1}$ *(squares),* $\nu = 10^{-5}$ *(diamonds) and on* $\Omega = [0,0.8]\times[0,0.8]$ *for* $\nu = 10^{-5}$ *(stars); second- and first-order convergence slopes are indicated in the bottom-left corner of both pictures.*

Problem A allows us to verify the convergence rate when approximating smooth solutions. Instead, the solution of Problem B is characterized by a boundary layer at the outflow near the edges defined by the equations $x = 1$ and $y = 1$. We mention, for comparison's sake, that this last test case is also solved in [27] by using a stabilized Lagrange multiplier method and in [25] by using a nonconforming finite element discretization. The results of this section are shown in Figures 5 and 6, both reporting the $\mathtt{H}^1(\Omega)$- and $\mathtt{L}^2(\Omega)$-norm error of the solution provided by the nonlinear diamond scheme for the above-mentioned values of ν. As shown by both figures, the formal second order of convergence is achieved when we approximate a smooth solution, as in Problem A, and a boundary layer on a mesh that is sufficiently fine to resolve it, as in Problem B, for $\nu = 1$ and $\nu = 10^{-1}$. In both cases, the convergence rate that is experimentally measured is 1 for the $\mathtt{H}^1(\Omega)$-norm error and 2 for the $\mathtt{L}^2(\Omega)$-norm error. When the layer is not resolved in Problem B, e.g., for the strongest advective case corresponding to $\nu = 10^{-5}$, the convergence rate of the $\mathtt{L}^2(\Omega)$-norm of the approximation error is only of order 1/2, while the $\mathtt{H}^1(\Omega)$-norm of the approximation error is not converging to zero on the sequence of meshes considered. Nonetheless, we still observe the expected convergence rate for both error norms away from the boundary layers, e.g., by restricting the error measurement to the domain $\Omega = [0, 0.8] \times [0, 0.8]$ as proposed in [27]. This local convergence behavior is perfectly in agreement with the one observed by the authors of [27], who used a completely different approximation method for this same problem.

7. Conclusions. A new finite volume method has been developed to numerically solve the steady multidimensional convection-diffusion equation. The method is designed on the conservative form of the equation and approximates the cell averages of the analytical solution on unstructured meshes of d-simplices, $d \geq 2$ being the spatial dimension. The formulation of the numerical advective fluxes requires a limiter on the reconstructed slope within each mesh cell in order to ensure monotonicity. Instead, the diffusive numerical fluxes are based on a nonlinear extension of the face gradients used in the standard version of the diamond scheme. Second-order accuracy has been achieved by using a piecewise linear reconstruction within each cell and at mesh vertices. The linear reconstruction relies on the weighted average of the cell averages of the solution. An algorithm was proposed to calculate nonnegative and bounded weights for the mesh vertex values. The nonlinear face gradients have some significant theoretical properties which allowed us to prove the solvability of the resulting scheme and the existence of a discrete maximum and minimum principle. The numerical results illustrate these features and the performance of the method in treating problems with both smooth solutions and solutions with internal and boundary layers. It turns out that second order of convergence of the sequence of approximate solutions is achievable with negligible violation of the discrete maximum and minimum principles. The discrete maximum and minimum principles are obtained at a relatively small additional cost with respect to the computational cost required for solving the same problem by the version of the method not preserving the discrete maximum and minimum principle. The relative additional cost is about 10–15%. Finally, we remark that the stencils of the diamond scheme and the nonlinear diamond scheme are identical; hence, the new method does not require additional memory storage.

Acknowledgment. The authors are grateful to the anonymous reviewers for their helpful suggestions.

REFERENCES

[1] S. Agmon, *Lectures on Elliptic Boundary Value Problems*, Mathematical Studies, Van Nostrand, New York, 1965.

[2] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, SIAM, Philadelphia, 1994.

[3] E. Bertolazzi, *Discrete conservation and discrete maximum principle for elliptic PDEs*, Math. Models Methods Appl. Sci., 8 (1998), pp. 685–711.

[4] E. Bertolazzi and G. Manzini, *Algorithm* 817 *P2MESH: Generic object-oriented interface between* 2*-D unstructured meshes and FEM/FVM-based PDE solvers*, ACM Trans. Math. Software, 28 (2002), pp. 101–132.

[5] E. Bertolazzi and G. Manzini, *A cell-centered second-order accurate finite volume method for convection-diffusion problems on unstructured meshes*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1235–1260.

[6] E. Bertolazzi and G. Manzini, *A finite volume method for transport of contaminants in porous media*, Appl. Numer. Math., 49 (2004), pp. 291–305.

[7] E. Bertolazzi and G. Manzini, *Limiting strategies for polynomial reconstructions in the finite volume approximation of the linear advection equation*, Appl. Numer. Math., 49 (2004), pp. 277–289.

[8] E. Burman and A. Ern, *Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection-diffusion-reaction equation*, Comput. Methods Appl. Mech. Engrg, 191 (2002), pp. 3833–3855.

[9] E. Burman and A. Ern, *A Discontinuous hp Finite Element Method for Convection-Diffusion Problems: Discrete Maximum Principle and Convergence*, Technical report, Institute of Analysis and Scientific Computing, EPFL, Lausanne, France, 2003.

[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1980; reprinted, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[11] Y. Coudière, J.-P. Vila, and P. Villedieu, *Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.

[12] P. Deuflhard, *Newton Methods for Nonlinear Problems: Affine Invariance and Adaptive Algorithms*, Springer Ser. Comput. Math. 35, Springer-Verlag, Berlin, 2004.

[13] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, North-Holland, Amsterdam, 2000, pp. 713–1020.

[14] T. Gallouët, R. Herbin, and M. H. Vignal, *Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions*, SIAM J. Numer. Anal., 37 (2000), pp. 1935–1972.

[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 2001; reprint of the 1998 edition.

[16] G. H. Golub and C. F. Van Loan, *Matrix Computation*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.

[17] B. Heinrich, *Finite Difference Methods on Irregular Networks*, Birkhäuser Verlag, Basel, 1987.

[18] R. Herbin, *An error estimate for a finite volume scheme for a diffusion-convection problem on a triangular mesh*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 165–173.

[19] M. E. Hubbard, *Multidimensional slope limiters for MUSCL-type finite volume schemes on unstructured grids*, J. Comput. Phys., 155 (1999), pp. 54–74.

[20] D. S. Kershaw, *Differencing of the diffusion equation in Lagrangian hydrodynamic codes*, J. Comput. Phys., 39 (1981), pp. 375–395.

[21] S. Korotov and M. Křížek, *Acute type refinements for tetrahedral partitions of polyhedral domains*, SIAM J. Numer Anal., 39 (2001), pp. 724–733.

[22] S. Korotov, M. Křížek, and P. Neittaanmaki, *Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle*, Math. Comp., 70 (2000), pp. 107–119.

[23] R. D. Lazarov, V. L. Makarov, and W. Weinelt, *On the convergence of difference schemes for the approximation of solutions $u \in W_2^m$ $(m > 0.5)$ of elliptic equations with mixed derivatives*, Numer. Math., 44 (1984), pp. 223–232.

[24] G. Manzini and S. Ferraris, *Mass-conservative finite-volumes on unstructured grids for the Richards' equation*, Adv. Water Resour., 27 (2004), pp. 1199–1215.

[25] G. Matthies and L. Tobiska, *The streamline-diffusion method form conforming and nonconforming finite elements of lowest order applied to convection-diffusion problems*, Computing, 66 (2001), pp. 343–364.

[26] K. W. Morton, *Numerical Solution of Convection-Diffusion Problems*, Chapman & Hall, London, 1996.

[27] G. Rapin and G. Lube, *A stabilized scheme for the Lagrange multiplier method for advection-diffusion equations*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1035–1060.

[28] R. T. Rockafellar, *Convex Analysis*, Princeton Landmarks Math., Princeton University Press, Princeton, NJ, 1997 (reprint).

[29] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre a coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

[30] H. A. van der Vorst, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 631–644.

[31] P. S. Vassilievski, S. I. Petrova, and R. D. Lazarov, *Finite difference schemes on triangular cell-centered grids with local refinement*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1287–1313.

[32] H. Yserentant, *Die maximale Konsistenzordnung von Differenzenapproximationen nichtnegativer Art*, Numer. Math., 42 (1983), pp. 119–123.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2200–2226

CONVERGENCE OF TIME-STEPPING METHOD FOR INITIAL AND BOUNDARY-VALUE FRICTIONAL COMPLIANT CONTACT PROBLEMS*

JONG-SHI PANG†, VIJAY KUMAR‡, AND PENG SONG§

Abstract. Beginning with a proof of the existence of a discrete-time trajectory, this paper establishes the convergence of a time-stepping method for solving continuous-time, boundary-value problems for dynamic systems with frictional contacts characterized by local compliance in the normal and tangential directions. Our investigation complements the analysis of the initial-value rigid-body model with one frictional contact encountering inelastic impacts by Stewart [Arch. Ration. Mech. Anal., 145 (1998), pp. 215–260] and the recent analysis by Anitescu [*Optimization-Based Simulation for Nonsmooth Rigid Multibody Dynamics*, Argonne National Laboratory, Argonne, IL, 2004] using the framework of measure differential inclusions. In contrast to the measure-theoretic approach of these authors, we follow a differential variational approach and address a broader class of problems with multiple elastic or inelastic impacts. Applicable to both initial and affine boundary-value problems, our main convergence result pertains to the case where the compliance in the normal direction is decoupled from the compliance in the tangential directions and where the friction coefficients are sufficiently small.

Key words. time-stepping methods, frictional contact problems, compliance models, differenctial complementarity problems, boundary-value problems

AMS subject classifications. 34B10, 90C33, 70E55, 70F40, 74H20

DOI. 10.1137/040612269

1. Introduction.

This paper investigates the limiting properties of time-stepping methods for rigid-body dynamics problems with multiple contacts characterized by friction and local compliance. Comprehensive reviews on rigid-body models and their applications can be found in the monographs [5, 13] and the excellent survey [20]. The benefits of introducing contact compliance for analysis and numerical simulation have been discussed in previous work [23]. In particular, a compliant model eliminates the static indeterminacy that is inherent in a rigid body dynamic model with multiple contacts and the need to make assumptions about linear independence of the columns of the Jacobian matrix [3, 10, 19]. Most important, even when one makes the requisite assumptions for uniqueness and existence, it is not possible to analyze the boundary-value problem in a fully rigid-body model because of the presence of discontinuities in velocities during impacts.

The present paper is closest in spirit to the work of Stewart [19], who analyzed the convergence of a time-stepping method [21] for initial-value rigid-body problems

*Received by the editors July 26, 2004; accepted for publication (in revised form) June 8, 2005; published electronically December 30, 2005. This work is based on research supported by the National Science Foundation under Focused Research Group grant DMS-0353216.

http://www.siam.org/journals/sinum/43-5/61226.html

†Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180-3590 (pangj@rpi.edu). The work of this author was partly supported by the National Science Foundation under grant CCR-0098013.

‡Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, PA 19104-6391 (kumar@grasp.cis.upenn.edu). The work of this author was partly supported by the National Science Foundation under grants IIS-0413138 and DMS-0139747.

§Department of Mechanical and Aerospace Engineering, Rutgers, State University of New Jersey, Piscataway, NJ 08854-8058 (pengsong@jove.rutgers.edu). The work of this author was partly supported by the National Science Foundation under grant IIS-0413138.

with frictional contact. Stewart's analysis is the first of its kind in the rigid-body dynamics literature. However, his analysis is somewhat limiting in several respects. In particular, the main result of the paper [19], Theorem 1, pertains essentially to the case of one inelastic contact. Even for such a simplified case, the analysis relies on a Radon–Nikodym derivative with several technical restrictions. It is difficult to fully extend Stewart's analysis because of the intrinsic analytical difficulties associated with the rigid-body paradigm. This difficulty is acknowledged in the recent paper by Anitescu [1], who established the convergence of a sequential quadratic programming method to a solution of a measure differential inclusion for nonsmooth rigid multibody dynamics.

Our previous work was concerned with several analytical aspects of dynamic models with compliant frictional contacts. Comparisons between results obtained with and without local compliance with a singular perturbation analysis are included in [15]. Uniqueness and existence results for the discrete-time problem are presented in [17] under a semi-implicit discretization that permits the use of linear complementarity theory [6]. In this paper, we analyze the convergence of a broad scheme of time-stepping methods for solving frictional compliant contact problems. In contrast to [17], the discretization scheme employed here is more general, allowing in particular for nonlinearities in the state variables, thus going well beyond the previous analysis of existence and uniqueness that is based on a linear theory. Unlike the analysis in [19, 1], our main convergence result is not in terms of measure differential inclusions. Most importantly, our analysis is carried out in a broad setting that includes both initial-value and boundary-value problems with affine constraints on the initial and final state (see (13)). It should be noted that although boundary-value problems arise naturally in the design of mechanical systems governed by dynamics, previous literature on this subject addresses only initial-value problems and ours is the first attempt to study contact problems subject to boundary conditions.

This paper addresses neither the numerical implementation nor the order of convergence of the time-stepping methods. For details on practical implementation and computational results, see [16, 18]; see also [4] for a part-insertion application of a boundary-value planar rigid-body problem. The order of convergence analysis for frictional contact problems is a very difficult topic, even for initial-value problems. The discontinuity of the friction forces as a function of the system states is a main cause for such difficulty.

The organization of the rest of the paper is as follows. In the next section, we summarize the formulation of the continuous-time frictional compliant contact problem and formally define a concept of a weak solution to the problem. A numerical time-stepping scheme for computing such a solution is described in section 3. The convergence analysis of the numerical scheme begins in section 4, where we first investigate in detail the normal and tangential frictional conditions in the discrete-time subproblems, establishing in particular the existence of a discrete-time trajectory of the normal and tangential contact forces that are continuous functions of the state. We also establish the uniqueness of such a trajectory under a "small coefficient of friction" assumption; see Propositions 6 and 7. With the aid of the machinery of differential variational inequalities [11], and under the small-friction-coefficient assumption, we complete the convergence analysis of the time-stepping method for a compliant-body frictional contact problem in section 5. There, an existence result, Theorem 8, for the discrete-time boundary-value problem is first proved, which is followed by the main convergence theorem of the paper, Theorem 9. The small-friction assumption is the artifact of the nonlinear friction law that is in turn a characteristic of the discretization scheme that we employ. Such a nonlinear analysis is in contrast to previous analysis by Stewart and Anitescu, which is based on a polygonal approximation of the quadratic Coulomb cone.

2. Model formulation.

The mathematical formulation of the frictional compliant contact problem has several components: (a) equations of motion, (b) compliance constitutive law, (c) contact and friction, and (d) boundary conditions. In what follows, we present only the essentials of the formulation and refer the reader to [14, 17] for the detailed explanation of the overall model.

Equations of motion. The dynamics equation of motion for a multibody system with frictional contacts can be written as

$$M(q)\dot{\nu} \;=\; f(t,q,\nu) + \boldsymbol{\Gamma}(q)^T \boldsymbol{\lambda}, \tag{1}$$

where q is the n_q-dimensional vector of generalized coordinates, ν is the n_ν-dimensional vector of the system velocities, $\dot{\nu}$ denotes the time derivative of ν (i.e., $\dot{\nu} = d\nu/dt$), $M(q)$ is the $n_\nu \times n_\nu$ symmetric positive definite mass-inertia matrix, $f(t,q,\nu)$ is the n_ν-dimensional external force vector (excluding contact forces),

$$\boldsymbol{\Gamma}(q)^T \;\equiv\; \left[\; \Gamma_{\rm n}(q)^T \;\; \Gamma_{\rm t}(q)^T \;\; \Gamma_{\rm o}(q)^T \;\right] \;\equiv\; G(q)^T \left[\; J\Psi_{\rm n}(q)^T \;\; J\Psi_{\rm t}(q)^T \;\; J\Psi_{\rm o}(q)^T \;\right]$$

is the transpose of the system Jacobian matrix, with $\Psi_{\rm n,t,o}(q)$ and $J\Psi_{\rm n,t,o}(q)$ being the constraint functions and their Jacobians for all possible contacts in the normal direction (labeled n) and the two tangential directions (labeled t and o), respectively, and $\boldsymbol{\lambda} \equiv (\lambda_{\rm n}, \lambda_{\rm t}, \lambda_{\rm o}) = \lambda_{\rm n,t,o}$ is the vector of contact forces in these directions. For compliant contact models, the dimensions of the contact forces and, accordingly, the orders of the associated Jacobian matrices, are related to the compliance constitutive model being used. The matrix $G(q)$ is a $n_q \times n_\nu$ parametrization matrix that allows us to use different parameterizations for the motion group via the the following kinematics equation:

$$\dot{q} \;=\; G(q)\nu, \tag{2}$$

where $\dot{q} \equiv dq/dt$ is the time-derivative of the system configuration. Together, (1) and (2) constitute the equations of motion governing the dynamics of the mechanical system.

Letting $T > 0$ be the terminal time of the problem, we postulate the following assumptions (A)–(C) on the above model functions. Notice that no rank assumption is imposed on $\boldsymbol{\Gamma}(q)$; this is a distinct advantage of a compliant model in that the number of contact points need not be restricted by the degrees of freedom of the bodies in contact.

(A) The function $f(t,q,\nu)$ is Lipschitz continuous on $[0,T] \times \Re^{n_q+n_\nu}$ with constant $L_f > 0$; thus,

$$\begin{gathered}\| f(t,q,\nu) - f(t',q',\nu') \| \leq L_f \left[|t-t'| + \|q-q'\| + \|\nu-\nu'\| \right] \\ \forall\,(t,q,\nu), (t',q',\nu') \in [0,T] \times \Re^{n_q+n_\nu}.\end{gathered}$$

(B) The functions $G(q)$ and $\boldsymbol{\Gamma}(q)$ are Lipschitz continuous and bounded on $\Re^{n_q}$; thus there exist positive constants L_G, L_W, η_G, and η_W such that for all q and q' in $\Re^{n_q}$,

$$\begin{aligned} &\|\, G(q) - G(q') \,\| \;\leq\; L_G \,\|\, q - q' \,\|, && \|\, \boldsymbol{\Gamma}(q) - \boldsymbol{\Gamma}(q') \,\| \;\leq\; L_W \,\|\, q - q' \,\|, \\ &\sup_{q \in \Re^{n_q}} \|\, G(q) \,\| \;\leq\; \eta_G, && \sup_{q \in \Re^{n_q}} \|\, \boldsymbol{\Gamma}(q) \,\| \;\leq\; \eta_W; \end{aligned}$$

moreover, the function $\Psi_{\rm n}(q)$ satisfies the limit condition

$$\lim_{\| q - q' \| \to 0} \frac{\| \Psi_{\rm n}(q) - \Psi_{\rm n}(q') - J\Psi_{\rm n}(q')(q - q') \|}{\| q - q' \|} = 0, \tag{3}$$

or, equivalently, for every scalar $\varepsilon > 0$, a scalar $\varsigma > 0$ exists such that

$$\| q - q' \| \leq \varsigma \Rightarrow \| \Psi_{\rm n}(q) - \Psi_{\rm n}(q') - J\Psi_{\rm n}(q')(q - q') \| \leq \varepsilon \| q - q' \|. \tag{4}$$

(C) The mass-inertia matrix $M(q)$ is Lipschitz continuous on $\Re^{n_q}$ with Lipschitz constant $L_M > 0$; moreover, positive constants σ_M and σ'_M exist such that

$$\inf_{q \in \Re^{n_q}} \min_{\| \nu \| = 1} \nu^T M(q) \nu \geq \sigma_M \quad \text{and} \quad \sup_{q \in \Re^{n_q}} \max_{\| \nu \| = 1} \nu^T M(q) \nu \leq 1/\sigma'_M.$$

Condition (3) is clearly satisfied if $J\Psi_{\rm n}(q)$ is Lipschitz continuous. Unlike the treatment in [1], $\Psi_{\rm n}(q)$ is not assumed to be twice differentiable. (The squared distance function to a closed convex set—the obstacle set—is an example of a (scalar) function that is continuously differentiable with a Lipschitz gradient but is not twice differentiable.) Conditions (A), (B), and (C) have several immediate consequences, which will be used freely throughout the paper where appropriate.

A constitutive model for compliance. While there are many compliance models, we employ the distributed model described in [14, 17], to which we refer the reader for details and references. Specifically, this model postulates that the contact forces are linearly dependent on the body deformations and on the deformation rates:

$$\boldsymbol{\lambda} = \mathbf{K}(q)\boldsymbol{\delta} + \mathbf{C}(q)\dot{\boldsymbol{\delta}} \tag{5}$$

where $\boldsymbol{\delta} \equiv (\delta_{\rm n}, \delta_{\rm t}, \delta_{\rm o}) = \delta_{\rm n,t,o}$ is the vector of body deformations in the normal (n) and the two tangential directions (t and o), $\dot{\boldsymbol{\delta}}$ denotes the vector of velocities of the deformations (i.e., $\dot{\boldsymbol{\delta}} = d\boldsymbol{\delta}/dt$); the stiffness matrix $\mathbf{K}(q)$ and the damping matrix $\mathbf{C}(q)$, which are partitioned as

$$\mathbf{K}(q) \equiv \begin{bmatrix} K_{\rm nn}(q) & K_{\rm nt}(q) & K_{\rm no}(q) \\ K_{\rm tn}(q) & K_{\rm tt}(q) & K_{\rm to}(q) \\ K_{\rm on}(q) & K_{\rm ot}(q) & K_{\rm oo}(q) \end{bmatrix} \text{ and } \mathbf{C}(q) \equiv \begin{bmatrix} C_{\rm nn}(q) & C_{\rm nt}(q) & C_{\rm no}(q) \\ C_{\rm tn}(q) & C_{\rm tt}(q) & C_{\rm to}(q) \\ C_{\rm on}(q) & C_{\rm ot}(q) & C_{\rm oo}(q) \end{bmatrix},$$

are each of order $3n_s^2 n_c$, with n_s^2 being the number of elements with lumped stiffness and damping properties that comprise a contact patch; each of the 18 block matrices (such as $K_{\rm nt}(q)$, etc.) in $\mathbf{K}(q)$ and $\mathbf{C}(q)$ is an $n_s^2 n_c$ block diagonal matrix with n_c diagonal blocks, one for each contact patch, and each such diagonal block is in turn a square matrix of order n_s^2. With $n_\delta \equiv n_s^2 n_c$, it follows that the vectors $\lambda_{\rm n}$, $\lambda_{\rm t}$, $\lambda_{\rm o}$, $\delta_{\rm n}$, $\delta_{\rm t}$, and $\delta_{\rm o}$ are each of dimension n_δ. We postulate the following condition:

(D) $\mathbf{K}(q)$ and $\mathbf{C}(q)$ are Lipschitz continuous symmetric positive definite matrix-valued functions of q; moreover, positive constants $\eta_K > 0$, σ_{KC} and η_{KC} exist such that $\sup_{q \in \Re^{n_q}} \| \mathbf{K}(q) \| \leq \eta_K$, and, for all scalars $h > 0$ sufficiently small,

$$\inf_{q \in \Re^{n_q}} \min_{\| \boldsymbol{\delta} \| = 1} \boldsymbol{\delta}^T [\, h\,\mathbf{K}(q) + \mathbf{C}(q)\,]^{-1} \boldsymbol{\delta} \geq \sigma_{KC} \text{ and } \sup_{q \in \Re^{n_q}} \left\| [\, h\,\mathbf{K}(q) + \mathbf{C}(q)\,]^{-1} \right\| \leq \eta_{KC}.$$

Notice that the above implies $\sup_{q \in \Re^{n_q}} \| \mathbf{C}(q) \| \leq 1/\sigma_{KC}$.

Contact and friction. Stated as a complementarity condition, the normal contact condition is

$$0 \leq \lambda_{\mathrm{n}} \perp \Psi_{\mathrm{n}}(q) + \delta_{\mathrm{n}} \geq 0, \tag{6}$$

where the notation $u \perp v$ means that the two vectors u and v are perpendicular. The tangential friction condition is expressed by a minimization principle over the cone of frictional forces: for each $i = 1, \ldots, n_\delta$,

$$(\lambda_{i\mathrm{t}}, \lambda_{i\mathrm{o}}) \in \operatorname{argmin} \left\{ s_{i\mathrm{t}}\tilde{\lambda}_{i\mathrm{t}} + s_{i\mathrm{o}}\tilde{\lambda}_{i\mathrm{o}} : (\tilde{\lambda}_{i\mathrm{t}}, \tilde{\lambda}_{i\mathrm{o}}) \in \mathcal{F}(\mu_i \lambda_{i\mathrm{n}}) \right\}, \tag{7}$$

where

$$\begin{aligned} s_{i\mathrm{t}} &\equiv \frac{d(\delta_{it} + \Psi_{it}(q))}{dt} = \dot{\delta}_{it} + \nabla\Psi_{it}(q)^T \dot{q}, \\ s_{i\mathrm{o}} &\equiv \frac{d(\delta_{i\mathrm{o}} + \Psi_{i\mathrm{o}}(q))}{dt} = \dot{\delta}_{i\mathrm{o}} + \nabla\Psi_{i\mathrm{o}}(q)\dot{q} \end{aligned} \tag{8}$$

are the tangential slip velocities at contact patch i, which depend on both the deformations of the compliant elements and the rigid body motions, and where $\mu_i \geq 0$ is the friction coefficient and

$$\mathcal{F}(\tau) \equiv \{ (a, b) \in \Re^2 : \sqrt{a^2 + b^2} \leq \tau \}, \quad \tau \geq 0,$$

is the standard Coulomb friction cone. From (7), it follows that

$$s_{it}\lambda_{it} + s_{i\mathrm{o}}\lambda_{i\mathrm{o}} = -\mu_i \lambda_{i\mathrm{n}} \sqrt{s_{i\mathrm{t}}^2 + s_{i\mathrm{o}}^2}. \tag{9}$$

Moreover, provided that $\mu_i\lambda_{i\mathrm{n}} > 0$, we have, with $r_i \equiv \sqrt{s_{i\mathrm{t}}^2 + s_{i\mathrm{o}}^2}$,

$$s_{i\mathrm{t}} + \frac{r_i \lambda_{i\mathrm{t}}}{\sqrt{\lambda_{i\mathrm{t}}^2 + \lambda_{i\mathrm{o}}^2}} = 0, \quad s_{\mathrm{o}} + \frac{r_i \lambda_{i\mathrm{o}}}{\sqrt{\lambda_{i\mathrm{t}}^2 + \lambda_{i\mathrm{o}}^2}} = 0,$$

$$0 \leq r_i \perp \mu_i \lambda_{i\mathrm{n}} - \sqrt{\lambda_{i\mathrm{t}}^2 + \lambda_{i\mathrm{o}}^2} \geq 0,$$

where we define 0/0 to be 1. If we use polar coordinates to represent the pair $(s_{i\mathrm{t}}, s_{i\mathrm{o}})$, say,

$$s_{i\mathrm{t}} = r_i \cos\psi_i \quad \text{and} \quad s_{i\mathrm{o}} = r_i \sin\psi_i,$$

then there exists a scalar $\phi_i \in [-1, 1]$ satisfying $r_i > 0 \Rightarrow \phi_i = 1$ such that

$$\lambda_{i\mathrm{t}} = -\mu_i \lambda_{i\mathrm{n}} \phi_i \cos\psi_i \quad \text{and} \quad \lambda_{i\mathrm{o}} = -\mu_i \lambda_{i\mathrm{n}} \phi_i \sin\psi_i.$$

The latter representation of $(\lambda_{it}, \lambda_{io})$ remains valid when $\mu_i\lambda_{in} = 0$, by letting $\phi_i = 0$.

More on the compliance model. The constitutive law (5) can be used to eliminate the slip velocities (s_{it}, s_{io}) in the friction law (7), resulting in an expression of the latter in terms of the state variables $(q, \nu, \delta_{\mathrm{n,t,o}})$ and the normal force λ_{n}. This reformulation of the friction law is significant because the slip velocities may behave discontinuously and lead to technical difficulties in the convergence analysis of a numerical method. From (5), we have $\dot{\boldsymbol{\delta}} = \mathbf{C}(q)^{-1}(\boldsymbol{\lambda} - \mathbf{K}(q)\boldsymbol{\delta})$. Writing

$$\mathbf{C}(q)^{-1} \equiv \begin{bmatrix} \widehat{C}_{\mathrm{nn}}(q) & \widehat{C}_{\mathrm{nt}}(q) & \widehat{C}_{\mathrm{no}}(q) \\ \widehat{C}_{\mathrm{tn}}(q) & \widehat{C}_{\mathrm{tt}}(q) & \widehat{C}_{\mathrm{to}}(q) \\ \widehat{C}_{\mathrm{on}}(q) & \widehat{C}_{\mathrm{ot}}(q) & \widehat{C}_{\mathrm{oo}}(q) \end{bmatrix},$$

we obtain

$$\begin{pmatrix} \dot{\delta}_{i\mathrm{t}} \\ \dot{\delta}_{i\mathrm{o}} \end{pmatrix} = \begin{bmatrix} \widehat{C}_{i\mathrm{tn}}(q) & \widehat{C}_{i\mathrm{tt}}(q) & \widehat{C}_{i\mathrm{to}}(q) \\ \widehat{C}_{i\mathrm{on}}(q) & \widehat{C}_{i\mathrm{ot}}(q) & \widehat{C}_{i\mathrm{oo}}(q) \end{bmatrix} (\boldsymbol{\lambda} - \mathbf{K}(q)\boldsymbol{\delta}),$$

where $\widehat{C}_{i\mathrm{tn}}(q)$ denotes the ith row of the (sub)matrix $\widehat{C}_{\mathrm{tn}}(q)$, and similarly for the other notation. Clearly, the friction condition (7) at contact i is equivalent to: for all $(\widetilde{\lambda}_{i\mathrm{t}}, \widetilde{\lambda}_{i\mathrm{o}}) \in \mathcal{F}(\mu_i \lambda_{i\mathrm{n}})$,

(10)
$$0 \leq \begin{pmatrix} \widetilde{\lambda}_{i\mathrm{t}} - \lambda_{i\mathrm{t}} \\ \widetilde{\lambda}_{i\mathrm{o}} - \lambda_{i\mathrm{o}} \end{pmatrix}^T \begin{pmatrix} s_{i\mathrm{t}} \\ s_{i\mathrm{o}} \end{pmatrix} = \begin{pmatrix} \widetilde{\lambda}_{i\mathrm{t}} - \lambda_{i\mathrm{t}} \\ \widetilde{\lambda}_{i\mathrm{o}} - \lambda_{i\mathrm{o}} \end{pmatrix}^T \left[\begin{pmatrix} \dot{\delta}_{i\mathrm{t}} \\ \dot{\delta}_{i\mathrm{o}} \end{pmatrix} + \begin{pmatrix} \Gamma_{i\mathrm{t}}(q) \\ \Gamma_{i\mathrm{o}}(q) \end{pmatrix} \nu \right]$$
$$= \begin{pmatrix} \widetilde{\lambda}_{i\mathrm{t}} - \lambda_{i\mathrm{t}} \\ \widetilde{\lambda}_{i\mathrm{o}} - \lambda_{i\mathrm{o}} \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{i\mathrm{tn}}(q) \; \widehat{C}_{i\mathrm{tt}}(q) \; \widehat{C}_{i\mathrm{to}}(q) \\ \widehat{C}_{i\mathrm{on}}(q) \; \widehat{C}_{i\mathrm{ot}}(q) \; \widehat{C}_{i\mathrm{oo}}(q) \end{bmatrix} (\boldsymbol{\lambda} - \mathbf{K}(q)\boldsymbol{\delta}) + \begin{pmatrix} \Gamma_{i\mathrm{t}}(q) \\ \Gamma_{i\mathrm{o}}(q) \end{pmatrix} \nu \right\}.$$

Proposition 1 shows that under the constitutive compliance law (5), the tangential friction forces in a frictional compliant model can be characterized by the solution to a convex quadratic program.

PROPOSITION 1. *Given q, ν, λ_{n}, and $\boldsymbol{\delta}$, under* (5)*, the tangential forces $(\lambda_{\mathrm{t}}, \lambda_{\mathrm{o}})$ satisfy the minimum principle* (7) *if and only if $(\lambda_{\mathrm{t}}, \lambda_{\mathrm{o}})$ is the optimal solution, which must necessarily be unique, of the convex quadratic program:*

(11)
$$\textit{minimize} \quad \frac{1}{2} \begin{pmatrix} \tilde{\lambda}_{\mathrm{t}} \\ \tilde{\lambda}_{\mathrm{o}} \end{pmatrix} \begin{bmatrix} \widehat{C}_{\mathrm{tt}}(q) & \widehat{C}_{\mathrm{to}}(q) \\ \widehat{C}_{\mathrm{ot}}(q) & \widehat{C}_{\mathrm{oo}}(q) \end{bmatrix} \begin{pmatrix} \tilde{\lambda}_{\mathrm{t}} \\ \tilde{\lambda}_{\mathrm{o}} \end{pmatrix}$$
$$+ \begin{pmatrix} \tilde{\lambda}_{\mathrm{t}} \\ \tilde{\lambda}_{\mathrm{o}} \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\mathrm{tn}}(q) \\ \widehat{C}_{\mathrm{on}}(q) \end{bmatrix} \lambda_{\mathrm{n}} + \begin{bmatrix} \Gamma_{\mathrm{t}}(q) \\ \Gamma_{\mathrm{o}}(q) \end{bmatrix} \nu - \begin{bmatrix} \widehat{C}_{\mathrm{tn}}(q) & \widehat{C}_{\mathrm{tt}}(q) & \widehat{C}_{\mathrm{to}}(q) \\ \widehat{C}_{\mathrm{on}}(q) & \widehat{C}_{\mathrm{ot}}(q) & \widehat{C}_{\mathrm{oo}}(q) \end{bmatrix} \mathbf{K}(q)\boldsymbol{\delta} \right\}$$
$$\textit{subject to} \quad (\tilde{\lambda}_{\mathrm{t}}, \tilde{\lambda}_{\mathrm{o}}) \in \prod_{i=1}^{n_\delta} \mathcal{F}(\mu_i \, \lambda_{i\mathrm{n}}).$$

Proof. It suffices to note that the first-order optimality conditions of (11) are equivalent to the variational conditions (10). □

Boundary conditions. To complete the description of the model, we postulate a set of boundary conditions that connect the state variable $\mathbf{x} \equiv (q, \nu, \boldsymbol{\delta}) \in \Re^n$, where $n \equiv n_q + n_\nu + 3n_\delta$, at the initial and terminal times: $t = 0$ and $t = T$, respectively. The most general such conditions would be expressed by a nonlinear functional relation of the form $F(\mathbf{x}(0), \mathbf{x}(T)) = 0$. Nevertheless, such generality would make the analysis extremely difficult, if not impossible. In general, the boundary conditions should be consistent with the constraints; such consistency would require the initial pair $(q^0, \delta_{\mathrm{n}}^0)$ to satisfy the feasibility condition:

(12) $$\Psi_{\mathrm{n}}(q^0) + \delta_{\mathrm{n}}^0 \geq 0.$$

Therefore, to both accommodate realistic applications and facilitate the mathematical analysis, we consider a class of boundary conditions where the initial configuration $q(0) = q^0$ and deformation $\boldsymbol{\delta}(0) = \boldsymbol{\delta}^0$ are known and satisfy (12); but the initial velocity $\nu(0)$ and terminal state $\mathbf{x}(T)$ are subject to a system of linear equations,

$$\mathbf{b} \,=\, \mathbf{M}_\nu \nu(0) + \mathbf{N}\mathbf{x}(T), \tag{13}$$

for some given vector $\mathbf{b} \in \Re^{n_\nu}$ and matrices $\mathbf{M}_\nu \in \Re^{n_\nu \times n_\nu}$ and $\mathbf{N} \in \Re^{n_\nu \times n}$. When $\mathbf{M}_\nu$ is the identity matrix and $\mathbf{N}$ is the zero matrix, we recover an initial-value problem with a known initial state $\mathbf{x}(0)$.

Weak solutions. The frictional compliant contact problem under study is to find a state trajectory $\mathbf{x} : [0,T] \to \Re^n$ and a force trajectory $\boldsymbol{\lambda} : [0,T] \to \Re^{3n_\delta}$ such that $q(0) = q^0$, $\boldsymbol{\delta}(0) = \boldsymbol{\delta}^0$, and the conditions (1), (2)–(8), (12), and (13) are satisfied. Ideally, we want these conditions to be satisfied at all times $t \in [0,T]$, but due to the possible discontinuity of the force trajectory $\boldsymbol{\lambda}$, this ideal goal is generally not attainable, especially when it pertains to the numerical solutions obtained by a time-stepping scheme, such as the one described in the next section; see [19, 20]. Therefore, we have to settle for a kind of weak solution that satisfies the dynamics equations and the contact and friction conditions in a weak sense. This is an inherent limitation of the model, particularly (5). It may be possible to get a strong solution by using a more sophisticated, nonlinear constitutive model. (See [15] for an example of such a model.) However, we refrain from pursuing such an extended consideration and restrict ourselves to the law (5), whose analysis is already fairly involved.

DEFINITION 2. *The pair of trajectories* $\mathbf{x} : [0,T] \to \Re^n$ *and* $\boldsymbol{\lambda} : [0,T] \to \Re^{3n_\delta}$ *is said to be a weak solution of the frictional compliant contact problem if*

(a) *(the state equations)* $\mathbf{x}(t)$ *is absolutely continuous on* $[0,T]$ *and satisfy for all* $\tau \leq \tau\,'$ *in* $[0,T]$,

$$\begin{aligned}
\nu(\tau\,') - \nu(\tau) &= \int_\tau^{\tau\,'} M(q(t))^{-1}[\, f(t,q(t),\nu(t)) + \boldsymbol{\Gamma}(q(t))^T\boldsymbol{\lambda}(t)\,]\, dt, \\
q(\tau\,') - q(\tau) &= \int_\tau^{\tau\,'} G(q(t))\nu(t)\, dt, \\
\boldsymbol{\delta}(\tau\,') - \boldsymbol{\delta}(\tau) &= \int_\tau^{\tau\,'} C(q(t))^{-1}[\,\boldsymbol{\lambda}(t) - \mathbf{K}(q(t))\boldsymbol{\delta}(t)\,]\, dt;
\end{aligned}$$

(b) *(the normal contact condition)* $\Psi_{\rm n}(q(t)) + \delta_{\rm n}(t) \geq 0$ *for all* $t \in [0,T]$, $\lambda_{\rm n}(t) \geq 0$ *for almost all* $t \in [0,T]$, *and*

$$\int_0^T \lambda_{\rm n}(t)^T[\,\Psi_{\rm n}(q(t)) + \delta_{\rm n}(t)\,]\, dt \,=\, 0;$$

(c) *(the friction condition) for every* $i = 1, \ldots, n_\delta$, $(\lambda_{it}(t), \lambda_{io}(t)) \in \mathcal{F}(\mu_i\lambda_{in}(t))$ *for almost all* $t \in [0,T]$ *and for every continuous function* $(\widetilde{\lambda}_{\rm t}, \widetilde{\lambda}_{\rm o}) : [0,T] \to \Re^{2n_\delta}$ *such that for every* $i = 1, \ldots, n_\delta$, $(\widetilde{\lambda}_{i\rm t}(t), \widetilde{\lambda}_{i\rm o}(t))$ *belongs to* $\mathcal{F}(\mu_i\lambda_{in}(t))$ *for almost all* $t \in [0,T]$, *it holds that*

$$\int_0^T \begin{pmatrix} \widetilde{\lambda}_{\rm t}(t) - \lambda_{\rm t}(t) \\ \widetilde{\lambda}_{\rm o}(t) - \lambda_{\rm o}(t) \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\rm tt}(q(t)) & \widehat{C}_{\rm to}(q(t)) \\ \widehat{C}_{\rm ot}(q(t)) & \widehat{C}_{\rm oo}(q(t)) \end{bmatrix} \left[\begin{pmatrix} \lambda_{\rm t}(t) \\ \lambda_{\rm o}(t) \end{pmatrix} \right.\right.$$
$$\left.\left. - \begin{bmatrix} K_{\rm tt}(q(t)) & K_{\rm to}(q(t)) \\ K_{\rm ot}(q(t)) & K_{\rm oo}(q(t)) \end{bmatrix} \begin{pmatrix} \delta_{\rm t}(t) \\ \delta_{\rm o}(t) \end{pmatrix} \right] + \begin{pmatrix} \Gamma_{\rm t}(q(t)) \\ \Gamma_{\rm o}(q(t)) \end{pmatrix} \nu(t) \right\} dt \geq 0,$$

(d) *(initial and boundary conditions)* $q(0) = q^0$, $\boldsymbol{\delta}(0) = \boldsymbol{\delta}^0$, *and* (13) *hold.*

We make a couple remarks about the above definition. First, the slip velocities do not enter in the above definition; second, the tangential friction condition is stipulated to hold in an integral form that is an aggregation over all contacts. This is in contrast to requiring the condition to hold at every contact. In the special case where compliance is decoupled among the contacts, then the aggregated condition indeed decouples into separate conditions at each individual contact.

3. A time-stepping scheme.

The kind of "semi-implicit" discretization methods described herein for computing a weak solution to the frictional compliant contact problems has been used extensively for solving initial-value rigid-body problems and, to a lesser extent, for compliant-body problems; see, e.g., [2, 3, 11, 21, 17, 18, 19]. Specifically, we divide the time interval $[0, T]$ into $N_h + 1$ subintervals each of equal length $h > 0$; thus $(N_h + 1)h = T$. The variables of the discrete-time system are

$$\begin{gathered}\{\, q^{h,0}, q^{h,1}, \ldots, q^{h,N_h+1} \,\},\ \{\, \nu^{h,0}, \nu^{h,1}, \ldots, \nu^{h,N_h+1} \,\},\ \{\, \delta^{h,0}_{\mathrm{n,t,o}}, \delta^{h,1}_{\mathrm{n,t,o}}, \ldots, \delta^{h,N_h+1}_{\mathrm{n,t,o}} \,\}, \\ \{\, \lambda^{h,1}_{\mathrm{n,t,o}}, \lambda^{h,2}_{\mathrm{n,t,o}}, \ldots, \lambda^{h,N_h+1}_{\mathrm{n,t,o}} \,\},\ \text{and}\ \{\, s^{h,1}_{\mathrm{t,o}}, \ldots, s^{h,N_h+1}_{\mathrm{t,o}} \,\}. \end{gathered} \tag{14}$$

We write $\mathbf{x}^{h,j} \equiv (q^{h,j}, \nu^{h,j}, \boldsymbol{\delta}^{h,j})$, $\boldsymbol{\delta}^{h,j} \equiv \delta^{h,j}_{\mathrm{n,t,o}}$, and $\boldsymbol{\lambda}^{h,j} \equiv \lambda^{h,j}_{\mathrm{n,t,o}}$. To derive the discrete-time system, we replace the time derivatives of the state variable $\mathbf{x} \equiv (q, \nu, \boldsymbol{\delta})$ by standard finite-difference quotients such as:

$$\dot{\mathbf{x}}(t) \approx \frac{\mathbf{x}(t+h) - \mathbf{x}(t)}{h}.$$

The right-hand expressions in the equation of motion (1) and in the kinematic equation (2) are approximated by a semi-implicit scheme that employs a θ-rule, whereby the differential variables q and ν are evaluated at some intermediate values in the respective subintervals determined by the scalar $\theta \in [0, 1]$. Specifically, with

$$q^{h,\theta_j} \equiv \theta\, q^{h,j} + (1-\theta)\, q^{h,j+1} \quad \text{and} \quad \nu^{h,\theta_j} \equiv \theta\, \nu^{h,j} + (1-\theta)\, \nu^{h,j+1},$$

the discrete-time dynamics and kinematics equations at time $t_{h,j+1}$ are

$$\begin{aligned} M(q^{h,j})(\nu^{h,j+1} - \nu^{h,j}) &= h\,[\, f(t_{h,j+1}, q^{h,\theta_j}, \nu^{h,\theta_j}) + \boldsymbol{\Gamma}(q^{h,j})^T \boldsymbol{\lambda}^{h,j+1}], \\ q^{h,j+1} - q^{h,j} &= h\, G(q^{h,j}) \nu^{h,\theta_j}. \end{aligned} \tag{15}$$

(More generally, we could use different θ-values in these two equations. For simplicity, we avoid this minor variation and use (15).) Since $s_{\mathrm{t,o}} = \dot{\delta}_{\mathrm{t,o}} + J\Psi_{\mathrm{t,o}}(q)\dot{q}$ by (8), we employ the following discrete-time approximation for the vector of tangential slip velocities $s_{\mathrm{t,o}}$:

$$s^{h,j+1}_{\mathrm{t,o}} = \frac{\delta^{h,j+1}_{\mathrm{t,o}} - \delta^{h,j}_{\mathrm{t,o}}}{h} + J\Psi_{\mathrm{t,o}}(q^{h,j}) \frac{q^{h,j+1} - q^{h,j}}{h} = \frac{\delta^{h,j+1}_{\mathrm{t,o}} - \delta^{h,j}_{\mathrm{t,o}}}{h} + \Gamma_{\mathrm{t,o}}(q^{h,j}) \nu^{h,\theta_j},$$

where we have used the discrete-time kinematics equation $q^{h,j+1} - q^{h,j} = hG(q^{h,j})\nu^{h,\theta_j}$ and the definition of $\Gamma_{\mathrm{t,o}}(q^{h,j}) \equiv J\Psi_{\mathrm{t,o}}(q^{h,j}) G(q^{h,j})$ to obtain the second equality. In deriving the discrete-time normal contact condition, we employ the first-order approximation

$$\Psi_{\mathrm{n}}(q(t+h)) \approx \Psi_{\mathrm{n}}(q(t)) + h\, J\Psi_{\mathrm{n}}(q(t))\, \dot{q}(t),$$

which holds for all $h > 0$ sufficiently small, and approximate $\dot{q}(t)$ similarly.

Putting together all the above approximations, we arrive at the following discrete-time frictional compliant contact problem: given $(q^0, \boldsymbol{\delta}^0)$ satisfying (12), compute (14) such that the conditions below are satisfied for all $j = 0, 1, \ldots, N_h$,

$$\begin{aligned}
&M(q^{h,j})(\,\nu^{h,j+1} - \nu^{h,j}\,) \,=\, h\,[\,f(t_{h,j+1}, q^{h,\theta_j}, \nu^{h,\theta_j}) + \boldsymbol{\Gamma}(q^{h,j})^T \boldsymbol{\lambda}^{h,j+1}],\\
&q^{h,j+1} - q^{h,j} \,=\, h\,G(q^{h,j})\nu^{h,\theta_j},\\
&\delta_{\rm t}^{h,j+1} - \delta_{\rm t}^{h,j} \,=\, h\,[s_{\rm t}^{h,j+1} - \Gamma_{\rm t}(q^{h,j})\nu^{h,\theta_j}\,],\\
&\delta_{\rm o}^{h,j+1} - \delta_{\rm o}^{h,j} \,=\, h\,[s_{\rm o}^{h,j+1} - \Gamma_{\rm o}(q^{h,j})\nu^{h,\theta_j}\,],\\
&0 \,\leq\, \lambda_{\rm n}^{h,j+1} \,\perp\, \Psi_{\rm n}(q^{h,j}) + h\,\Gamma_{\rm n}(q^{h,j})\nu^{h,\theta_j} + \delta_{\rm n}^{h,j+1} \,\geq\, 0,\\
&\boldsymbol{\lambda}^{h,j+1} \,=\, \mathbf{K}(q^{h,j})\boldsymbol{\delta}^{h,j+1} + \frac{\mathbf{C}(q^{h,j})}{h}\,(\,\boldsymbol{\delta}^{h,j+1} - \boldsymbol{\delta}^{h,j}\,),\\
&\begin{pmatrix} \lambda_{i\rm t}^{h,j+1} \\ \lambda_{i\rm o}^{h,j+1} \end{pmatrix} \in \underset{(\tilde{\lambda}_{i\rm t}, \tilde{\lambda}_{i\rm o}) \in \mathcal{F}(\mu_i\,\lambda_{i\rm n}^{h,j+1})}{\arg\min} \left\{ \begin{pmatrix} s_{i\rm t}^{h,j+1} \\ s_{i\rm o}^{h,j+1} \end{pmatrix}^T \begin{pmatrix} \tilde{\lambda}_{i\rm t} \\ \tilde{\lambda}_{i\rm o} \end{pmatrix} \right\},\\
&\\
&\mathbf{b} \,=\, \mathbf{M}_\nu \nu^{h,0} + \mathbf{N}\mathbf{x}^{h,N_h+1}, \quad \text{and} \quad (\,q^{h,0}, \boldsymbol{\delta}^{h,0}\,) \,=\, (\,q^0, \boldsymbol{\delta}^0\,).
\end{aligned} \tag{16}$$

The inclusion of the parameter θ in selected terms raises the question of why it is not used consistently throughout the constraints. An answer to this question can be traced to the paper [21], where the intention was to use a linear complementarity solver to solve the subproblems. As seen from the subsequent paper [19], excluding θ from the matrices $M(q^{h,j})$, $\boldsymbol{\Gamma}(q^{h,j})$, and $G(q^{h,j})$ simplifies the analysis significantly. The Ph.D. thesis [24] contains an analysis of a fully implicit time-stepping method for an initial-value rigid-body model, which leads to subproblems that are nonlinear complementarity problems. A computational comparison between a fully implicit scheme versus a semi-implicit scheme can be found in [25]. Presumably, the use of the parameter θ is to induce a higher order of convergence; yet such a goal is hard to substantiate formally. The analysis below does not address this issue of order of convergence.

Beginning in the next section, we will analyze two fundamental issues associated with the above discrete-time system: (a) the existence of a solution to each discrete-time boundary-value subproblem, and (b) the convergence of such a discrete-time trajectory to a weak solution of the frictional compliant contact problem. Part of the challenge in the convergence analysis lies in the coupled nature of the individual time-step subproblems, which are linked by the boundary equation $\mathbf{b} = \mathbf{M}_\nu \nu^{h,0} + \mathbf{N}\mathbf{x}^{h,N_h+1}$. Briefly, the analysis consists of two major tasks. First, we show that for an arbitrary triple $\mathbf{x}^{h,j}$, a unique friction triple $\boldsymbol{\lambda}^{h,j+1}$ exists that has some desirable continuity and boundedness properties, provided that the friction coefficients μ_i are sufficiently small. These properties of the friction forces allow us to apply an argument used in [11] for a class of boundary-value differential variational inequalities to complete the convergence analysis of the discrete-time trajectory as the time step tends to zero.

Naturally, there is an important computational issue associated with the above numerical scheme; namely, how can the discrete-time system (16) be efficiently solved in practice? The proof of Theorem 8 suggests a fixed-point method. Yet, specialized complementarity methods [8] may prove to be more effective. Nevertheless, there is presently no formal study on the applicability of the latter methods. The numerical experiments in [17] employed the complementarity solver PATH [7], which produced satisfactory results. Despite such practical experience, which is somewhat limited, there is an urgent need for the development of some robust algorithms for solving (16) along with a rigorous proof of applicability.

4. Preliminary analysis: Initial-value problems. The analysis in this section is best considered as one for an initial-value problem, where $\nu^{h,0}$, in addition to $(q^{h,0}, \boldsymbol{\delta}^{h,0})$, is assumed to be fixed but arbitrary. (This is the case where $\mathbf{M}_\nu$ is the identity matrix and $\mathbf{N}$ is the zero matrix.) This analysis will be the basis for extension to the boundary-value problem where $\nu^{h,0}$ has to be determined to satisfy the boundary equation defined by a more general pair of boundary matrices $(\mathbf{M}_\nu, \mathbf{N})$. As the first step in the convergence analysis, we show that the discrete-time dynamics and kinematics equations (15) have a unique solution $(q^{h,j+1}, \nu^{h,j+1})$ for any $(q^{h,j}, \nu^{h,j})$ and $\boldsymbol{\lambda}^{h,j+1}$; moreover, such a solution, for fixed $(q^{h,j}, \nu^{h,j})$, has several desirable properties in $\boldsymbol{\lambda}^{h,j+1}$.

PROPOSITION 3. *Under conditions* (A)–(C), *for any* $\theta \in [0,1]$, *positive constants* h_0, η_q, L_q, *and* σ_ν *exist such that for every tuple* $y \equiv (t, q^{\rm ref}, \nu^{\rm ref}) \in [0,T] \times \Re^{n_q+n_\nu}$ *and every* h *in* $(0, h_0]$, *a bounded continuous function* $(q^h(\cdot\,; y), \nu^h(\cdot\,; y)) : \Re^{n_\nu} \to \Re^{n_q+n_\nu}$ *exists satisfying the following properties:*

(a) *for every vector* $e \in \Re^{n_\nu}$, $(q^h(e;y), \nu^h(e;y))$ *is the unique pair* (q^h, ν^h) *satisfying*

$$M(q^{\rm ref})(\nu^h - \nu^{\rm ref}) = h[f(t, q^{\rm ref} + (1-\theta)(q^h - q^{\rm ref}), \nu^{\rm ref} + (1-\theta)(\nu^h - \nu^{\rm ref})) + e],$$
$$q^h - q^{\rm ref} = h\, G(q^{\rm ref})[\,\nu^{\rm ref} + (\,1-\theta\,)\,(\,\nu^h - \nu^{\rm ref}\,)\,];$$

moreover,

$$\|\, q^h - q^{\rm ref} \,\| + \|\, \nu^h - \nu^{\rm ref} \,\| \;\leq\; h\, \eta_q \,\left[\, 1 + \|\, q^{\rm ref} \,\| + \|\, \nu^{\rm ref} \,\| + \|\, e \,\| \,\right];$$

(b) $(q^h(\cdot\,; y), \nu^h(\cdot\,; y))$ *is Lipschitz continuous with constant* hL_q; *thus*

$$\|\, q^h(e^1; y) - q^h(e^2; y) \,\| + \|\, \nu^h(e^1; y) - \nu^h(e^2; y) \,\| \;\leq\; h\, L_q \,\|\, e^1 - e^2 \,\| \quad \forall\, e^1, e^2 \,\in\, \Re^{n_\nu};$$

(c) *the function* $\nu^h(\cdot\,; y) : \Re^{n_\nu} \to \Re^{n_\nu}$ *is strongly monotone with constant* $h\sigma_\nu$; *thus,*

$$(\, \nu^h(e^1; y) - \nu^h(e^2; y) \,)^T (\, e^1 - e^2 \,) \;\geq\; h\, \sigma_\nu \,\|\, e^1 - e^2 \,\|^2 \quad \forall\, e^1, e^2 \,\in\, \Re^{n_\nu}.$$

Proof. For a given vector $e \in \Re^{n_\nu}$, it is easily seen that the map

$$\begin{pmatrix} \nu \\ q \end{pmatrix} \mapsto \begin{pmatrix} \nu^{\rm ref} + hM(q^{\rm ref})^{-1}\,[f(t,\, q^{\rm ref} + (1-\theta)(q - q^{\rm ref}),\, \nu^{\rm ref} + (1-\theta)(\nu - \nu^{\rm ref})) + e] \\ q^{\rm ref} + hG(q^{\rm ref})[\nu^{\rm ref} + (1-\theta)(\nu - \nu^{\rm ref})] \end{pmatrix}$$

is a contraction with a modulus that can be made as small as we want by choosing $h > 0$ sufficiently small. Moreover, the constant h_0 depends only on the constants L_f, θ, η_G, and σ_M. Therefore, the above map has a unique fixed point, which yields the existence and uniqueness of the pair $(q^h(e;y), \nu^h(e;y))$. The proof of the bound on $\|q^h - q^{\rm ref}\| + \|\, \nu^h - \nu^{\rm ref}\|$ is similar to that of (b); for this reason, we prove only the latter. For any two vectors e^1 and e^2, we have

$$\begin{aligned} &\|\, \nu^h(e^1; y) - \nu^h(e^2; y) \,\| \\ &\leq \frac{h}{\sigma_M}\,[\, L_f\,(\,1-\theta\,)\,(\, \|q^h(e^1; y) - q^h(e^2; y)\| + \|\nu^h(e^1; y) - \nu^h(e^2; y)\| \,) + \|e^1 - e^2\| \,] \\ &\leq \frac{h}{\sigma_M}\,[\, L_f\,(\,1-\theta\,)\,(\,1 + h\,(\,1-\theta\,)\,\eta_G\,)\,\|\, \nu^h(e^1; y) - \nu^h(e^2; y) \,\|\,) + \|e^1 - e^2\| \,], \end{aligned}$$

which implies

$$\| \, \nu^h(e^1;y) - \nu^h(e^2;y) \, \| \; \leq \; \frac{h \, \sigma_{_M}^{-1} \, \| \, e^1 - e^2 \, \|}{1 - h \, \sigma_{_M}^{-1} \, [\, L_f \, (\, 1 - \theta \,) \, (\, 1 + h \, (\, 1 - \theta \,) \, \eta_{_G} \,) \,]}.$$

Hence,

$$\| \, q^h(e^1;y) - q^h(e^2;y) \, \| \; \leq \; \frac{h^2 \, (\, 1 - \theta \,) \, \eta_{_G} \, \sigma_{_M}^{-1} \, \| \, e^1 - e^2 \, \|}{1 - h \, \sigma_{_M}^{-1} \, [\, L_f \, (\, 1 - \theta \,) \, (\, 1 + h \, (\, 1 - \theta \,) \, \eta_{_G} \,) \,]},$$

establishing the desired Lipschitz continuity of $(q^h(\cdot;y), \nu^h(\cdot;y))$. To prove (c), note that

$$\begin{aligned} &\sigma_{_M} \, \| \, \nu^h(e^1;y) - \nu^h(e^2;y) \, \|^2 \; \leq \; h \, (\, \nu^h(e^1;y) - \nu^h(e^2;y) \,)^T (\, e^1 - e^2 \,) \\ &+ h L_f (1-\theta) \, \| \nu^h(e^1;y) - \nu^h(e^2;y) \| (\| q^h(e^1;y) - q^h(e^2;y) \| + \| \nu^h(e^1;y) - \nu^h(e^2;y) \|), \end{aligned}$$

which yields

$$\begin{aligned} &(\, \nu^h(e^1;y) - \nu^h(e^2;y) \,)^T (\, e^1 - e^2 \,) \\ &\qquad \geq \; \frac{[\, \sigma_{_M} - h \, L_f \, (\, 1 - \theta \,) \, (\, 1 + h \, (\, 1 - \theta \,) \, \eta_{_G} \,) \,]}{h} \, \| \, \nu^h(e^1;y) - \nu^h(e^2;y) \, \|^2. \end{aligned}$$

Furthermore,

$$\begin{aligned} h \, \| \, e^1 - e^2 \, \| \quad &\leq \quad \eta_{_M} \, \| \, \nu^h(e^1;y) - \nu^h(e^2;y) \, \| \\ &\qquad + h \, L_f \, (\, 1 - \theta \,) \, (\| q^h(e^1;y) - q^h(e^2;y) \| + \| \nu^h(e^1;y) - \nu^h(e^2;y) \|) \\ &\leq \quad [\, \eta_{_M} + h \, L_f \, (\, 1 - \theta \,) \, (\, 1 + h \, (\, 1 - \theta \,) \, \eta_{_G} \,) \,] \, \| \, \nu^h(e^1;y) - \nu^h(e^2;y) \|, \end{aligned}$$

which implies

$$\| \, \nu^h(e^1;y) - \nu^h(e^2;y) \| \; \geq \; \frac{h \, \| \, e^1 - e^2 \, \|}{\eta_{_M} + h \, L_f \, (\, 1 - \theta \,) \, (\, 1 + h \, (\, 1 - \theta \,) \, \eta_{_G} \,)}.$$

Consequently,

$$(\nu^h(e^1;y) - \nu^h(e^2;y))^T (e^1 - e^2) \geq \frac{h \, [\, \sigma_{_M} - h L_f (1-\theta) \, (1 + h(1-\theta)\eta_{_G}) \,] \, \| e^1 - e^2 \|^2}{\eta_{_M} + h L_f (1-\theta) \, (1 + h(1-\theta)\eta_{_G})},$$

which establishes the desired strong monotonicity of $\nu^h(\cdot;y)$. □

From the discrete-time compliance equation

$$\boldsymbol{\lambda}^{h,j+1} \; = \; \mathbf{K}(q^{h,j}) \boldsymbol{\delta}^{h,j+1} + \frac{\mathbf{C}(q^{h,j})}{h} \, (\, \boldsymbol{\delta}^{h,j+1} - \boldsymbol{\delta}^{h,j} \,),$$

we deduce

$$(17) \qquad \boldsymbol{\delta}^{h,j+1} - \boldsymbol{\delta}^{h,j} \; = \; h \, [\, h \, \mathbf{K}(q^{h,j}) + \mathbf{C}(q^{h,j}) \,]^{-1} \, [\, \boldsymbol{\lambda}^{h,j+1} - \mathbf{K}(q^{h,j}) \boldsymbol{\delta}^{h,j} \,].$$

Considering the expression in the normal direction,

$$\Psi_{\rm n}(q^{h,j}) + h \Gamma_{\rm n}(q^{h,j}) \nu^{h,\theta_j} + \delta_{\rm n}^{h,j+1} = \Psi_{\rm n}(q^{h,j}) + h \Gamma_{\rm n}(q^{h,j}) \nu^{h,\theta_j} + \delta_{\rm n}^{h,j} + \delta_{\rm n}^{h,j+1} - \delta_{\rm n}^{h,j},$$

we define the discrete-time normal slip velocity,

$$s_{\mathrm{n}}^{h,j+1} \equiv \frac{\delta_{\mathrm{n}}^{h,j+1} - \delta_{\mathrm{n}}^{h,j}}{h} + \Gamma_{\mathrm{n}}(q^{h,j})\,\nu^{h,\theta_j},$$

which is consistent with the corresponding expressions for the discrete-time tangential velocities $s_{\mathrm{t,o}}^{h,j+1}$. Writing $\mathbf{s} \equiv s_{\mathrm{n,t,o}}$, we have

$$\begin{aligned}\mathbf{s}^{h,j+1} &= \frac{\boldsymbol{\delta}^{h,j+1} - \boldsymbol{\delta}^{h,j}}{h} + \boldsymbol{\Gamma}(q^{h,j})\,[\,\theta\,\nu^{h,j} + (\,1-\theta\,)\,\nu^{h,j+1}\,] \\ &= [\,h\mathbf{K}(q^{h,j}) + \mathbf{C}(q^{h,j})\,]^{-1}[\,\boldsymbol{\lambda}^{h,j+1} - \mathbf{K}(q^{h,j})\boldsymbol{\delta}^{h,j}\,] \\ &\quad + \boldsymbol{\Gamma}(q^{h,j})\nu^{h,j} + (1-\theta)\boldsymbol{\Gamma}(q^{h,j})(\nu^{h,j+1} - \nu^{h,j}).\end{aligned}$$

In view of the latter expression, we define, for fixed $\mathbf{y}^{\mathrm{ref}} = (t, q^{\mathrm{ref}}, \nu^{\mathrm{ref}}, \boldsymbol{\delta}^{\mathrm{ref}})$, the following function in $\boldsymbol{\lambda}$:

$$\begin{aligned}\mathbf{s}(\boldsymbol{\lambda};\mathbf{y}^{\mathrm{ref}}) \equiv{}& [\,h\,\mathbf{K}(q^{\mathrm{ref}}) + \mathbf{C}(q^{\mathrm{ref}})\,]^{-1}[\boldsymbol{\lambda} - \mathbf{K}(q^{\mathrm{ref}})\boldsymbol{\delta}^{\mathrm{ref}}\,] + \boldsymbol{\Gamma}(q^{\mathrm{ref}})\nu^{\mathrm{ref}} \\ &+ (\,1-\theta\,)\,\boldsymbol{\Gamma}(q^{\mathrm{ref}})\,[\,\nu^{h,\mathrm{ref}}(\boldsymbol{\Gamma}(q^{\mathrm{ref}})^T\boldsymbol{\lambda}) - \nu^{\mathrm{ref}}\,],\end{aligned}$$

where $\nu^{h,\mathrm{ref}}(r) \equiv \nu^h(r;(t,q^{\mathrm{ref}},\nu^{\mathrm{ref}}))$. Since $\nu^{h,\mathrm{ref}}$ is strongly monotone (albeit nonlinear), it follows that the map $\boldsymbol{\lambda} \mapsto \boldsymbol{\Gamma}(q^{\mathrm{ref}})\,\nu^{h,\mathrm{ref}}(\boldsymbol{\Gamma}(q^{\mathrm{ref}})^T\boldsymbol{\lambda})$ is monotone. Consequently, by assumption (D), we deduce

$$(\,\boldsymbol{\lambda} - \boldsymbol{\lambda}'\,)^T(\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{y}^{\mathrm{ref}}) - \mathbf{s}(\boldsymbol{\lambda}';\mathbf{y}^{\mathrm{ref}})\,) \ge \sigma_{KC}\,\|\,\boldsymbol{\lambda} - \boldsymbol{\lambda}'\,\|^2, \quad \forall\,\boldsymbol{\lambda},\boldsymbol{\lambda}' \in \Re^{3n_\delta};$$

that is, the function $\mathbf{s}(\cdot;\mathbf{y}^{\mathrm{ref}})$ is strongly monotone with a modulus that is independent of $\mathbf{y}^{\mathrm{ref}}$. Moreover, $\mathbf{s}(\cdot;\mathbf{y}^{\mathrm{ref}})$ is Lipschitz continuous with a modulus that is also independent of $\mathbf{y}^{\mathrm{ref}}$; indeed, by assumption (D) and part (b) of Proposition 3, we have

$$\|\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{y}^{\mathrm{ref}}) - \mathbf{s}(\boldsymbol{\lambda}';\mathbf{y}^{\mathrm{ref}})\,\| \le [\,\eta_{KC} + h\,L_q\,(\,1-\theta\,)\,\eta_W^2\,]\,\|\,\boldsymbol{\lambda} - \boldsymbol{\lambda}'\,\| \quad \forall\,\boldsymbol{\lambda},\boldsymbol{\lambda}' \in \Re^{3n_\delta}.$$

Furthermore,

$$\mathbf{s}(0;\mathbf{y}^{\mathrm{ref}}) = -[\,h\mathbf{K}(q^{\mathrm{ref}}) + \mathbf{C}(q^{\mathrm{ref}})\,]^{-1}\mathbf{K}(q^{\mathrm{ref}})\boldsymbol{\delta}^{\mathrm{ref}} + \theta\,\boldsymbol{\Gamma}(q^{\mathrm{ref}})\nu^{\mathrm{ref}} + (1-\theta)\,\boldsymbol{\Gamma}(q^{\mathrm{ref}})\,\nu^{h,\mathrm{ref}}(0).$$

From part (a) of Proposition 3, we obtain

$$\|\,\nu^{h,\mathrm{ref}}(0)\,\| \le \|\,\nu^{\mathrm{ref}}\,\| + h\,\eta_q\,[\,1 + \|\,q^{\mathrm{ref}}\,\| + \|\,\nu^{\mathrm{ref}}\,\|\,].$$

Consequently, we deduce that a constant $c_s > 0$ exists such that

$$\|\,\mathbf{s}(0;\mathbf{y}^{\mathrm{ref}})\,\| \le c_s\,[\,1 + \|\,q^{\mathrm{ref}}\,\| + \|\,\nu^{\mathrm{ref}}\,\| + \|\,\boldsymbol{\delta}^{\mathrm{ref}}\,\|\,] \quad \forall\,\mathbf{y}^{\mathrm{ref}} \equiv (\,t, q^{\mathrm{ref}}, \nu^{\mathrm{ref}}, \boldsymbol{\delta}^{\mathrm{ref}}\,). \tag{18}$$

This inequality will be used later; see Lemma 4. It is important to remark that the above constant c_s and the strong modulus and the Lipschitz constant of the function $\mathbf{s}(\cdot;\mathbf{y}^{\mathrm{ref}})$ are all independent of $\mathbf{y}^{\mathrm{ref}}$ and of $h > 0$, provided that the latter is sufficiently small.

In terms of the function $\mathbf{s}(\cdot;\mathbf{y}^{h,j})$, where $\mathbf{y}^{h,j} \equiv (t_{h,j+1}, \mathbf{x}^{h,j})$, the discrete frictional compliant contact problem at time step $t_{h,j+1}$, without the boundary condition, can be stated simply as the quasi-variational inequality of finding a triple $\boldsymbol{\lambda} \equiv \lambda_{\mathrm{n,t,o}} \in \Re^{3n_\delta}$ such that

$$0 \le \lambda_{\mathrm{n}} \perp \frac{\Psi_{\mathrm{n}}(q^{h,j}) + \delta_{\mathrm{n}}^{h,j}}{h} + s_{\mathrm{n}}(\boldsymbol{\lambda};\mathbf{y}^{h,j}) \ge 0$$

and for all $i = 1, \ldots, n_\delta$,

$$(\,\lambda_{it}, \lambda_{io}\,) \,\in\, \operatorname{argmin}\,\left\{\, s_{it}(\boldsymbol{\lambda}; \mathbf{y}^{h,j})\tilde{\lambda}_{it} + s_{io}(\boldsymbol{\lambda}; \mathbf{y}^{h,j})\tilde{\lambda}_{io} \,:\, (\,\tilde{\lambda}_{it}, \tilde{\lambda}_{io}\,) \,\in\, \mathcal{F}(\mu_i\,\lambda_{in})\,\right\}.$$

We state and prove a lemma pertaining to the above contact and friction conditions. This lemma is the key to the entire convergence analysis of the time-stepping method.

LEMMA 4. *Let $\mathbf{s}(\boldsymbol{\lambda}; \mathbf{y})$ be a continuous function that is Lipschitz continuous and strongly monotone in $\boldsymbol{\lambda} \in \Re^{3n_\delta}$ uniformly in $\mathbf{y} \in \Re^m$; i.e., positive constants η_s and σ_s exist such that for all $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$ and $\mathbf{y}$,*

$$(\boldsymbol{\lambda} - \boldsymbol{\lambda}')^T(\,\mathbf{s}(\boldsymbol{\lambda}; \mathbf{y}) - \mathbf{s}(\boldsymbol{\lambda}'; \mathbf{y})\,) \geq \sigma_s\,\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|^2 \;\; \textit{and} \;\; \|\mathbf{s}(\boldsymbol{\lambda}; \mathbf{y}) - \mathbf{s}(\boldsymbol{\lambda}'; \mathbf{y})\| \leq \eta_s\,\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|.$$

Suppose further that a constant $c_s > 0$ exists such that $\|\mathbf{s}(0; \mathbf{y})\| \leq c_s\|\mathbf{y}\|$ for all $\mathbf{y} \in \Re^m$. There exists a positive scalar $\bar{\mu} > 0$ such that for every vector $\mu > 0$ satisfying $\max_{1 \leq i \leq n_\delta} \mu_i \leq \bar{\mu}$, a continuous function $\boldsymbol{\lambda}^\mu : \Re^m \to \Re^{3n_\delta}$ exists such that for every parameter $\mathbf{y}$, $\boldsymbol{\lambda}^\mu(\mathbf{y})$ is the unique triple $\lambda_{\mathrm{n,t,o}}$ satisfying

$$0 \,\leq\, \lambda_{\mathrm{n}} \,\perp\, s_{\mathrm{n}}(\lambda_{\mathrm{n,t,o}}; \mathbf{y}) \,\geq\, 0,$$

and, for every $i = 1, \ldots, n_\delta$,

$$\begin{pmatrix} \lambda_{it} \\ \lambda_{io} \end{pmatrix} \in \underset{(\tilde{\lambda}_{it}, \tilde{\lambda}_{io}) \in \mathcal{F}(\mu_i\,\lambda_{i\mathrm{n}})}{\arg\min} \left\{ \begin{pmatrix} \tilde{\lambda}_{it} \\ \tilde{\lambda}_{io} \end{pmatrix}^T \begin{pmatrix} s_{it}(\lambda_{\mathrm{n,t,o}}; \mathbf{y}) \\ s_{io}(\lambda_{\mathrm{n,t,o}}; \mathbf{y}) \end{pmatrix} \right\}.$$

Proof. There are several things to be proved: the existence of the scalar $\bar{\mu}$ and the existence, uniqueness, and continuity of $\lambda^\mu_{\mathrm{n,t,o}}(\mathbf{y})$ for all $\mu > 0$ as specified. Indeed, the existence of a triple $\boldsymbol{\lambda}$ satisfying the above friction conditions for every $\mu > 0$ is proved by invoking a general result from the theory of quasi-variational inequalities [8, Corollary 2.8.4], as done in several previous references, such as [12]. In what follows, we show the uniqueness of such a solution for all $\mu > 0$ sufficiently small.

Suppose that $\lambda^1_{\mathrm{n,t,o}}$ and $\lambda^2_{\mathrm{n,t,o}}$ are two solutions corresponding to a given $\mathbf{y}$. Write, for $j = 1, 2$, $s^j_{\mathrm{n,t,o}} \equiv s_{\mathrm{n,t,o}}(\lambda^j_{\mathrm{n,t,o}}; \mathbf{y})$. We may write, for every i,

$$s^j_{i\mathrm{t}} \,\equiv\, r^j_i\,\cos\psi^j_i \quad \text{and} \quad s^j_{i\mathrm{o}} \,\equiv\, r^j_i\,\sin\psi^j_i,$$

where $r^j_i \equiv \sqrt{(s^j_{i\mathrm{t}})^2 + (s^j_{i\mathrm{o}})^2}$. It then follows that $\phi^j_i \in [-1, 1]$ exist satisfying

$$r^j_i \,>\, 0 \,\Rightarrow\, \phi^j_i \,=\, 1 \tag{19}$$

(ϕ^j_i is not necessarily equal to 1 when $r^j_i = 0$) and

$$\lambda^j_{i\mathrm{t}} \,=\, -\mu_i\,\lambda^j_{i\mathrm{n}}\,\phi^j_i\,\cos\psi^j_i \quad \text{and} \quad \lambda^j_{i\mathrm{o}} \,=\, -\mu_i\,\lambda^j_{i\mathrm{n}}\,\phi^j_i\,\sin\psi^j_i.$$

We have

$$\begin{aligned} \lambda^1_{i\mathrm{t}} - \lambda^2_{i\mathrm{t}} &= -\mu_i\,\lambda^1_{i\mathrm{n}}\,\phi^1_i\,\cos\psi^1_i + \mu_i\,\lambda^2_{i\mathrm{n}}\,\phi^2_i\,\cos\psi^2_i \\ &= -(\,\mu_i\,\phi^1_i\,\cos\psi^1_i\,)\,(\,\lambda^1_{i\mathrm{n}} - \lambda^2_{i\mathrm{n}}\,) + \mu_i\,\lambda^2_{i\mathrm{n}}\,(\,\phi^2_i\,\cos\psi^2_i - \phi^1_i\,\cos\psi^1_i\,). \end{aligned}$$

Similarly,

$$\lambda^1_{i\mathrm{o}} - \lambda^2_{i\mathrm{o}} \;=\; -(\,\mu_i\,\phi^1_i\,\sin\psi^1_i\,)\,(\,\lambda^1_{i\mathrm{n}} - \lambda^2_{i\mathrm{n}}\,) + \mu_i\,\lambda^2_{i\mathrm{n}}\,(\,\phi^2_i\,\sin\psi^2_i - \phi^1_i\,\sin\psi^1_i\,).$$

Therefore, letting D_t and D_o be the diagonal matrices whose diagonal entries are $-\mu_i\phi^1_i\cos\psi^1_i$ and $-\mu_i\phi^1_i\sin\psi^1_i$, respectively, we can write

$$\begin{aligned}\lambda^1_\mathrm{t} - \lambda^2_\mathrm{t} &= D_t(\,\lambda^1_\mathrm{n} - \lambda^2_\mathrm{n}\,) + \mu\,\lambda^2_\mathrm{n}\,(\,\phi^2\cos\psi^2 - \phi^1\cos\psi^1\,),\\ \lambda^1_\mathrm{o} - \lambda^2_\mathrm{o} &= D_o(\,\lambda^1_\mathrm{n} - \lambda^2_\mathrm{n}\,) + \mu\,\lambda^2_\mathrm{n}\,(\,\phi^2\sin\psi^2 - \phi^1\sin\psi^1\,),\end{aligned}$$

where the notation in the second terms in the right-hand side of the above equations has an obvious componentwise meaning. Consequently,

$$\begin{pmatrix}\lambda^1_\mathrm{n} - \lambda^2_\mathrm{n}\\ \lambda^1_\mathrm{t} - \lambda^2_\mathrm{t}\\ \lambda^1_\mathrm{o} - \lambda^2_\mathrm{o}\end{pmatrix} = \begin{bmatrix} I & 0 & 0\\ D_t & I & 0\\ D_o & 0 & I\end{bmatrix}\begin{pmatrix}\lambda^1_\mathrm{n} - \lambda^2_\mathrm{n}\\ \mu\,\lambda^2_\mathrm{n}\,(\phi^2\cos\psi^2 - \phi^1\cos\psi^1)\\ \mu\,\lambda^2_\mathrm{n}\,(\phi^2\sin\psi^2 - \phi^1\sin\psi^1)\end{pmatrix}$$

or, equivalently,

$$\begin{aligned}\begin{pmatrix}\lambda^1_\mathrm{n} - \lambda^2_\mathrm{n}\\ \mu\lambda^2_\mathrm{n}(\phi^2\cos\psi^2 - \phi^1\cos\psi^1)\\ \mu\lambda^2_\mathrm{n}(\phi^2\sin\psi^2 - \phi^1\sin\psi^1)\end{pmatrix} &= \begin{bmatrix} I & 0 & 0\\ D_t & I & 0\\ D_o & 0 & I\end{bmatrix}^{-1}\begin{pmatrix}\lambda^1_\mathrm{n} - \lambda^2_\mathrm{n}\\ \lambda^1_\mathrm{t} - \lambda^2_\mathrm{t}\\ \lambda^1_\mathrm{o} - \lambda^2_\mathrm{o}\end{pmatrix}\\ &= \begin{bmatrix} I & 0 & 0\\ -D_t & I & 0\\ -D_o & 0 & I\end{bmatrix}\begin{pmatrix}\lambda^1_\mathrm{n} - \lambda^2_\mathrm{n}\\ \lambda^1_\mathrm{t} - \lambda^2_\mathrm{t}\\ \lambda^1_\mathrm{o} - \lambda^2_\mathrm{o}\end{pmatrix}.\end{aligned}$$

Writing

$$\mathbf{D}(\mu) \;\equiv\; \begin{bmatrix} I & 0 & 0\\ -D_t & I & 0\\ -D_o & 0 & I\end{bmatrix},$$

we claim that positive constants σ'_s and $\bar\mu$ exist such that for all $\mu > 0$ satisfying $\max\limits_{1\le i\le n_\delta}\mu_i \le \bar\mu$,

$$(\,\mathbf{D}(\mu)\boldsymbol{\lambda} - \mathbf{D}(\mu)\boldsymbol{\lambda}'\,)^T(\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{y}) - \mathbf{s}(\boldsymbol{\lambda}';\mathbf{y})\,) \;\ge\; \sigma'_s\,\|\,\boldsymbol{\lambda} - \boldsymbol{\lambda}'\,\|^2$$

for all $\boldsymbol{\lambda}$ and $\boldsymbol{\lambda}'$. To establish the claim, we write

$$\begin{aligned}&(\,\mathbf{D}(\mu)\boldsymbol{\lambda} - \mathbf{D}(\mu)\boldsymbol{\lambda}'\,)^T(\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{y}) - \mathbf{s}(\boldsymbol{\lambda}';\mathbf{y})\,)\\ &\quad= (\,\boldsymbol{\lambda} - \boldsymbol{\lambda}'\,)^T(\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{y}) - \mathbf{s}(\boldsymbol{\lambda}';\mathbf{y})\,) - [\,(\,I - \mathbf{D}(\mu)\,)\,(\,\boldsymbol{\lambda} - \boldsymbol{\lambda}'\,)\,]^T(\,\mathbf{s}(\boldsymbol{\lambda};\mathbf{y}) - \mathbf{s}(\boldsymbol{\lambda}';\mathbf{y})\,)\\ &\quad\ge [\,\sigma_s - \eta_s\,\|\,I - \mathbf{D}(\mu)\,\|\,]\,\|\,\boldsymbol{\lambda} - \boldsymbol{\lambda}'\,\|^2;\end{aligned}$$

clearly, we can choose $\bar\mu > 0$ sufficiently small such that for all $\mu > 0$ satisfying $\max\limits_{1\le i\le n_\delta}\mu_i \le \bar\mu$, we have $\sigma_s - \eta_s\|I - \mathbf{D}(\mu)\| \ge \frac{1}{2}\sigma_s \equiv \sigma'_s$. This establishes the claim. Next, we show that

$$0 \;\ge\; (\,\mathbf{D}(\mu)\boldsymbol{\lambda}^1 - \mathbf{D}(\mu)\boldsymbol{\lambda}^2\,)^T(\,\mathbf{s}(\boldsymbol{\lambda}^1;\mathbf{y}) - \mathbf{s}(\boldsymbol{\lambda}^2;\mathbf{y})\,). \tag{20}$$

The right-hand side of the above inequality is equal to

$$\left(\begin{array}{c} \lambda_{\rm n}^1 - \lambda_{\rm n}^2 \\ \mu\,\lambda_{\rm n}^2\,(\phi^2\cos\psi^2 - \phi^1\cos\psi^1) \\ \mu\,\lambda_{\rm n}^2\,(\phi^2\sin\psi^2 - \phi^1\sin\psi^1) \end{array} \right)^T \left(\begin{array}{c} s_{\rm n}^1 - s_{\rm n}^2 \\ s_{\rm t}^1 - s_{\rm t}^2 \\ s_{\rm o}^1 - s_{\rm o}^2 \end{array} \right).$$

By complementarity, we have $(\lambda_{\rm n}^1 - \lambda_{\rm n}^2)^T(s_{\rm n}^1 - s_{\rm n}^2) \leq 0$. Furthermore,

$$\begin{aligned} & (\,\phi_i^2\cos\psi_i^2 - \phi_i^1\cos\psi_i^1\,)\,(\,s_{i\rm t}^1 - s_{i\rm t}^2\,) + (\,\phi_i^2\sin\psi_i^2 - \phi_i^1\sin\psi_i^1\,)\,(\,s_{i\rm o}^1 - s_{i\rm o}^2\,) \\ =\ & (\,\phi_i^2\cos\psi_i^2 - \phi_i^1\cos\psi_i^1\,)\,(\,r_i^1\,\cos\psi_i^1 - r_i^2\,\cos\psi_i^2\,) \\ & \quad + (\,\phi_i^2\sin\psi_i^2 - \phi_i^1\sin\psi_i^1\,)\,(\,r_i^1\,\sin\psi_i^1 - r_i^2\,\sin\psi_i^2\,) \\ =\ & -r_i^1\,\phi_i^1 - r_i^2\,\phi_i^2 + (\,r_i^1\,\phi_i^2 + r_i^2\,\phi_i^1\,)\cos(\psi_i^1 - \psi_i^2\,) \\ =\ & -r_i^1 - r_i^2 + (\,r_i^1\,\phi_i^2 + r_i^2\,\phi_i^1\,)\cos(\psi_i^1 - \psi_i^2\,)\ \leq\ 0, \end{aligned}$$

where the last equality follows from (19) and the last inequality holds because $|\phi_j^{1,2}| \leq 1$. Consequently, the inequality (20) holds. In turn, this implies that $\lambda_{\rm n,t,o}^1 = \lambda_{\rm n,t,o}^2$. This establishes the uniqueness of $\boldsymbol{\lambda}^\mu(x)$ for all $\mu > 0$ sufficiently small.

In the rest of the proof, we fix an arbitrary $\mu > 0$ sufficiently small and drop the superscript μ in $\boldsymbol{\lambda}^\mu$. To show the continuity of $\lambda_{\rm n,t,o}(\mathbf{y})$, we first derive a bound for $\|\lambda_{\rm n,t,o}(\mathbf{y})\|$. We have

$$\begin{aligned} 0 \quad & \geq \quad \boldsymbol{\lambda}(\mathbf{y})^T\mathbf{s}(\boldsymbol{\lambda}(\mathbf{y});\mathbf{y})\ =\ \boldsymbol{\lambda}(\mathbf{y})^T(\,\mathbf{s}(\boldsymbol{\lambda}(\mathbf{y});\mathbf{y}) - \mathbf{s}(0;\mathbf{y})\,) + \boldsymbol{\lambda}(\mathbf{y})^T\mathbf{s}(0;\mathbf{y}) \\ & \geq \quad \sigma_s\,\|\,\boldsymbol{\lambda}(\mathbf{y})\,\|^2 - c_s\,\|\,\boldsymbol{\lambda}(\mathbf{y})\,\|\,\|\,\mathbf{y}\,\|, \end{aligned}$$

which implies $\|\boldsymbol{\lambda}(\mathbf{y})\| \leq c_s\|\mathbf{y}\|$. Let $\{\mathbf{y}^k\}$ be a sequence of parameters converging to $\mathbf{y}^\infty$. Write $\lambda_{\rm n,t,o}^k \equiv \lambda_{\rm n,t,o}(\mathbf{y}^k)$. Since the sequence $\{\lambda_{\rm n,t,o}^k\}$ is bounded, by what has just been shown, let $\lambda_{\rm n,t,o}^\infty$ be the limit of a convergent subsequence $\{\lambda_{\rm n,t,o}^k : k \in \kappa\}$, where κ is an infinite subset of $\{1, 2, \dots\}$. It suffices to show that $\lambda_{\rm n,t,o}^\infty$ is a solution to the limiting system

$$0\ \leq\ \lambda_{\rm n}^\infty\ \perp\ s_{\rm n}(\lambda_{\rm n,t,o}^\infty;\mathbf{y}^\infty)\ \geq\ 0 \tag{21}$$

and

$$\left(\begin{array}{c} \lambda_{i\rm t}^\infty \\ \lambda_{i\rm o}^\infty \end{array} \right) \in \underset{(\tilde{\lambda}_{it},\tilde{\lambda}_{io})\in\mathcal{F}(\mu_i\,\lambda_{i\rm n}^\infty)}{\arg\min} \left\{ \left(\begin{array}{c} \tilde{\lambda}_{it} \\ \tilde{\lambda}_{io} \end{array} \right)^T \left(\begin{array}{c} s_{it}(\lambda_{\rm n,t,o}^\infty;\mathbf{y}^\infty) \\ s_{io}(\lambda_{\rm n,t,o}^\infty;\mathbf{y}^\infty) \end{array} \right) \right\}.$$

Since $0 \leq \lambda_{\rm n}^k \perp s_{\rm n}(\boldsymbol{\lambda}^k;\mathbf{y}^k) \geq 0$ for all k, passing to the limit $k(\in\kappa) \to \infty$ yields (21). Similarly, since $(\lambda_{i\rm t}^k)^2 + (\lambda_{i\rm o}^k)^2 \leq \mu_i^2(\lambda_{i\rm n}^k)^2$ for all k, we deduce $(\lambda_{i\rm t}^\infty, \lambda_{i\rm o}^\infty) \in \mathcal{F}(\mu_i\lambda_{i\rm n}^\infty)$. Moreover, since

$$\lambda_{i\rm t}^k\,s_{i\rm t}(\boldsymbol{\lambda}^k;\mathbf{y}^k) + \lambda_{i\rm o}^k\,s_{i\rm o}(\boldsymbol{\lambda}^k;\mathbf{y}^k)\ =\ -\mu_i\,\lambda_{i\rm n}^k\,\sqrt{s_{i\rm t}(\boldsymbol{\lambda}^k;\mathbf{y}^k)^2 + s_{i\rm o}(\boldsymbol{\lambda}^k;\mathbf{y}^k)^2},$$

passing to the limit $k(\in\kappa) \to \infty$ easily completes the proof. □

Applying Lemma 4 to the friction and contact conditions, we conclude that for all $h > 0$ and sufficiently small and for all $\mu > 0$ not exceeding a certain upper bound $\bar{\mu}$, for a given triple $\mathbf{x}^{h,j} \equiv (q^{h,j}, \nu^{h,j}, \boldsymbol{\delta}^{h,j})$, a unique friction force triple $\lambda_{\rm n,t,o}^{h,j+1}$ exists at time step $t_{h,j+1}$ that is a continuous function of $\mathbf{x}^{h,j}$. In what follows, we derive a bound on $\|\boldsymbol{\lambda}^{h,j+1}\|$ that takes advantage of the normal contact condition at time step j. This improved bound is important for the subsequent analysis. (A straightforward application of the previous lemma would yield a bound of the order $1/h$, which tends to infinity as $h \downarrow 0$, and thus is not effective for small h. The bound obtained below stays finite as h tends to zero, as shown subsequently.)

LEMMA 5. *Let $\boldsymbol{\lambda}^{h,j+1}$ satisfy*

$$0 \,\le\, \lambda_{\rm n}^{h,j+1} \,\perp\, \frac{\Psi_{\rm n}(q^{h,j}) + \delta_{\rm n}^{h,j}}{h} + s_{\rm n}(\boldsymbol{\lambda}^{h,j+1};\mathbf{y}^{h,j}) \,\ge\, 0$$

and for all $i = 1, \dots, n_\delta$,

$$(\lambda_{it}^{h,j+1}, \lambda_{io}^{h,j+1}) \in \underset{(\tilde\lambda_{it},\tilde\lambda_{io})\in\mathcal{F}(\mu_i\lambda_{in}^{h,j+1})}{\arg\min} \left\{ s_{it}(\boldsymbol{\lambda}^{h,j+1};\mathbf{y}^{h,j})\tilde\lambda_{it} + s_{io}(\boldsymbol{\lambda}^{h,j+1};\mathbf{y}^{h,j})\tilde\lambda_{io} \right\}.$$

A constant $\eta_\lambda > 0$, which depends only the model functions, exists such that

$$\|\boldsymbol{\lambda}^{h,j+1}\| \,\le\, \eta_\lambda \left[\frac{\|\min(\,0, \Psi_{\rm n}(q^{h,j}) + \delta_{\rm n}^{h,j}\,)\,\|}{h} + 1 + \|\mathbf{x}^{h,j}\| \right].$$

Proof. As in the proof of Lemma 4, we have

$$\begin{aligned} 0 \,&\ge\, (\,\lambda_{\rm n}^{h,j+1}\,)^T \frac{\Psi_{\rm n}(q^{h,j}) + \delta_{\rm n}^{h,j}}{h} + (\,\boldsymbol{\lambda}^{h,j+1}\,)^T \mathbf{s}(\boldsymbol{\lambda}^{h,j+1};\mathbf{y}^{h,j}) \\ &\ge\, (\,\lambda_{\rm n}^{h,j+1}\,)^T \frac{\Psi_{\rm n}(q^{h,j}) + \delta_{\rm n}^{h,j}}{h} + \|\,\boldsymbol{\lambda}^{h,j+1}\,\|^2 - c_s\,\|\,\boldsymbol{\lambda}^{h,j+1}\,\|\,[\,1 + \|\,\mathbf{x}^{h,j}\,\|\,] \\ &\ge\, (\,\lambda_{\rm n}^{h,j+1}\,)^T \min\left(0,\, \frac{\Psi_{\rm n}(q^{h,j}) + \delta_{\rm n}^{h,j}}{h}\right) + \|\,\boldsymbol{\lambda}^{h,j+1}\,\|^2 - c_s\,\|\,\boldsymbol{\lambda}^{h,j+1}\,\|\,[\,1 + \|\,\mathbf{x}^{h,j}\,\|\,], \end{aligned}$$

where the last inequality holds because $\lambda_{\rm n}^{h,j+1} \ge 0$. Consequently, the desired bound on $\|\boldsymbol{\lambda}^{h,j+1}\|$ follows easily by rearranging terms and then applying the Cauchy–Schwartz inequality. □

Combining Proposition 3 and Lemmas 4 and 5, we obtain the following result, which brings us one step closer to the main existence and uniqueness for the discrete-time boundary value problem.

PROPOSITION 6. *Under conditions* (A)–(D)*, positive scalars $\bar\mu$, h_0, and η_x exist such that for every vector $\mu > 0$ satisfying $\max\limits_{1\le i\le n_\delta} \mu_i \le \bar\mu$, every scalar $h \in (0, h_0]$, and every tuple $(q^{h,0}, \nu^{h,0}, \boldsymbol{\delta}^{h,0})$, a unique discrete-time trajectory* (14) *exists satisfying* (16) *for every $j = 0, 1, \dots, N_h$ but not necessarily* (13)*; moreover,*

$$\|\,\mathbf{x}^{h,j+1} - \mathbf{x}^{h,j}\,\| \,\le\, h\,\eta_x\,[\,1 + \|\,\mathbf{x}^{h,j}\,\| + \|\,\boldsymbol{\lambda}^{h,j+1}\,\|\,]. \tag{22}$$

Finally, if $\Psi_{\rm n}(q^{h,0}) + \delta_{\rm n}^{h,0} \ge 0$, then, for any scalar $c_q > 0$, the implication below holds for all $j = 0, 1, \dots, N_h$, where $q^{h,-1} \equiv q^{h,0}$:

$$\begin{aligned} &\|\min(0, \Psi_{\rm n}(q^{h,j}) - \Psi_{\rm n}(q^{h,j-1}) - J\Psi_{\rm n}(q^{h,j-1})(q^{h,j} - q^{h,j-1}))\| \,\le\, c_q\,h \\ &\quad\Rightarrow\; \|\boldsymbol{\lambda}^{h,j+1}\| \,\le\, \eta_\lambda\,(\,1 + c_q + \|\mathbf{x}^{h,j}\|\,). \end{aligned} \tag{23}$$

Proof. The bound for $\|\boldsymbol{\delta}^{h,j+1} - \boldsymbol{\delta}^{h,j}\|$, which is part of (22), follows from (17). Since

$$\Psi_{\rm n}(q^{h,j}) + \Gamma_{\rm n}(q^{h,j-1})(q^{h,j} - q^{h,j-1}) + \delta_{\rm n}^{h,j} \,\geq\, 0,$$

we have

$$\begin{array}{rcl} 0 & \geq & \min(\,0, \Psi_{\rm n}(q^{h,j}) + \delta_{\rm n}^{h,j}\,) \\ & \geq & \min(\,0, \Psi_{\rm n}(q^{h,j}) - \Psi_{\rm n}(q^{h,j-1}) - \Gamma_{\rm n}(q^{h,j-1})(q^{h,j} - q^{h,j-1})\,) \\ & & \quad + \min(\,0, \Psi_{\rm n}(q^{h,j-1}) + \Gamma_{\rm n}(q^{h,j-1})(q^{h,j} - q^{h,j-1}) + \delta_{\rm n}^{h,j}\,) \\ & = & \min(\,0, \Psi_{\rm n}(q^{h,j}) - \Psi_{\rm n}(q^{h,j-1}) - \Gamma_{\rm n}(q^{h,j-1})(q^{h,j} - q^{h,j-1})\,). \end{array}$$

Taking norms, we obtain

$$\|\min(0, \Psi_{\rm n}(q^{h,j}) + \delta_{\rm n}^{h,j})\| \leq \|\min(\,0, \Psi_{\rm n}(q^{h,j}) - \Psi_{\rm n}(q^{h,j-1}) - \Gamma_{\rm n}(q^{h,j-1})(q^{h,j} - q^{h,j-1}))\|.$$

The bound (23) on $\|\boldsymbol{\lambda}^{h,j+1}\|$ follows readily from Lemma 5. ∎

So far, we have not used the limit condition (3) in proving the above results. This condition allows us to establish the boundedness of the state variables $\{\mathbf{x}^{h,j}\}$ and thus of the force variables $\{\boldsymbol{\lambda}^{h,j+1}\}$ also. We first state a technical fact, which can be proved by induction; see also [11, Lemma 7]. Namely, for every nonnegative integer $k \leq N_h$, if

$$(24) \qquad \|\,\mathbf{x}^{h,j+1} - \mathbf{x}^{h,j}\,\| \,\leq\, h\,\psi_x\,(\,1 + \|\,\mathbf{x}^{h,j}\|\,) \quad \forall\, j \,=\, 0, 1, \ldots, k,$$

then (recalling that $T = (N_h + 1)h$),

$$(25) \qquad \|\,\mathbf{x}^{h,j+1}\,\| \,\leq\, e^{T\,\psi_x}\,(\,1 + \|\,\mathbf{x}^{h,0}\,\|\,) - 1 \quad \forall\, j \,=\, 0, 1, \ldots, k.$$

PROPOSITION 7. *For any positive scalar c_q, let $\psi_x \equiv \eta_x(1 + \eta_\lambda(1 + c_q))$. For any scalar $R_0 > 0$, the scalar h_0 in Proposition 6 can be chosen such that (25) holds for $k = N_h$ for all $h \in (0, h_0]$ and for all $\mathbf{x}^{h,0}$ satisfying $\|\mathbf{x}^{h,0}\| \leq R_0$; moreover, for all $j = 0, 1, \ldots, N_h$,*

$$(26) \qquad \|\,\boldsymbol{\lambda}^{h,j+1}\,\| \,\leq\, \eta_\lambda\,[\,c_q + e^{T\,\psi_x}\,(\,1 + \|\,\mathbf{x}^{h,0}\,\|\,)].$$

Proof. Choose $\varepsilon > 0$ such that $\varepsilon\psi_x e^{T\psi_x}(1 + R_0) < c_q$. Corresponding to the chosen ε, let $\varsigma > 0$ be such that (3) holds. Let $h_0 > 0$ be sufficiently small such that $h_0\psi_x e^{T\psi_x}(1 + R_0) < \varsigma$. Let $\mathbf{x}^{h,0}$ be an arbitrary vector satisfying $\|\mathbf{x}^{h,0}\| \leq R_0$ and let $h \in (0, h_0]$ be arbitrary. It suffices to prove (24) for $k = N_h$. Clearly, (24) is valid for $k = 0$ because $\|\mathbf{x}^{h,1} - \mathbf{x}^{h,0}\| \leq h\eta_x(1 + \|\mathbf{x}^{h,0}\|) \leq h\psi_x(1 + \|\mathbf{x}^{h,0}\|)$. Assume that (24), and thus (25), holds for some $k \geq 0$. To complete the induction, we need to show

$$\|\,\mathbf{x}^{h,k+2} - \mathbf{x}^{h,k+1}\,\| \,\leq\, h\,\psi_x\,(\,1 + \|\,\mathbf{x}^{h,k+1}\,\|\,).$$

By the choice of h and $\|\mathbf{x}^{h,0}\|$, (24) with $j = k$ and (25) with $j = k - 1$ imply

$$\|\,\mathbf{x}^{h,k+1} - \mathbf{x}^{h,k}\,\| \,\leq\, h\,\psi_x\,e^{T\,\psi_x}\,(\,1 + \|\,\mathbf{x}^{h,0}\,\|\,) \,<\, \varsigma.$$

By (3) and the choice of c_q, it follows that

$$\begin{array}{l} \|\,\min(\,0, \Psi_{\rm n}(q^{h,k+1}) - \Psi_{\rm n}(q^{h,k}) - J\Psi_{\rm n}(q^{h,k})(q^{h,k+1} - q^{h,k})\,)\,\| \\ \quad \leq\, \varepsilon\,\|\,q^{h,k+1} - q^{h,k}\,\| \,\leq\, \varepsilon\,h\,\psi_x\,e^{T\,\psi_x}\,(\,1 + \|\,\mathbf{x}^{h,0}\,\|\,) \,\leq\, h\,c_q. \end{array}$$

Consequently, by the implication (23), we obtain

$$\|\boldsymbol{\lambda}^{h,k+2}\| \leq \eta_{\lambda}\,(1 + c_q + \|\mathbf{x}^{h,k+1}\|) \leq \eta_{\lambda}\,(1 + c_q)\,(1 + \|\mathbf{x}^{h,k+1}\|).$$

Substituting this into (22) with $j = k+1$ yields

$$\begin{aligned}\|\mathbf{x}^{h,k+2} - \mathbf{x}^{h,k+1}\| &\leq h\,\eta_x\,[\,1 + \|\mathbf{x}^{h,k+1}\| + \eta_{\lambda}\,(1+c_q)\,(1+\|\mathbf{x}^{h,k+1}\|)\,] \\ &= h\,\psi_x\,(1 + \|\mathbf{x}^{h,k+1}\|),\end{aligned}$$

completing the induction. The bound on $\|\boldsymbol{\lambda}^{h,j+1}\|$ holds by (23) and the bound on $\|\mathbf{x}^{h,j}\|$. □

Based on Proposition 7, we can establish the convergence of the time-stepping method for an initial-value frictional compliant contact problem where $\mathbf{x}(0)$ is completely known. Since our treatment of the boundary-value problem will cover this case, we proceed directly to the latter.

5. Boundary-value analysis.

Proposition 7 allows us to employ the line of proof in [11] to complete the convergence analysis of the time-stepping method. Needless to say, the boundary equation (13) will play a key role in this analysis. For this reason, we partition the boundary matrix $\mathbf{N}$ as

$$\mathbf{N} \equiv \left[\begin{array}{ccc} \mathbf{N}_q & \mathbf{N}_{\nu} & \mathbf{N}_{\delta} \end{array}\right],$$

where $\mathbf{N}_q \in \Re^{n_\nu \times n_q}$, $\mathbf{N}_\nu \in \Re^{n_\nu \times n_\nu}$, and $\mathbf{N}_\delta \in \Re^{n_\nu \times n_\delta}$, and we write the discrete-time boundary equation as

$$(27)\qquad (\mathbf{M}_\nu + \mathbf{N}_\nu)\nu^{h,0} = \mathbf{b} + \mathbf{N}_\nu \nu^{h,0} - \mathbf{N}\mathbf{x}^{h,N_h+1} = \widehat{\mathbf{b}} - \mathbf{N}\,(\mathbf{x}^{h,N_h+1} - \mathbf{x}^{h,0}),$$

where $\widehat{\mathbf{b}} \equiv \mathbf{b} - \mathbf{N}_q q^0 - \mathbf{N}_\delta \boldsymbol{\delta}^0$. We are now ready to formally state and prove the two main results of this paper: Theorems 8 and 9. While the former establishes the existence of a solution to the discrete-time boundary system (16), including the boundary condition, the latter proves the convergence to a continuous-time trajectory.

THEOREM 8. *Assume conditions* (A)–(D) *and that* $\mathbf{M}_\nu + \mathbf{N}_\nu$ *is nonsingular. Let* ψ_x *be the constant obtained in Proposition* 7. *If*

$$(28)\qquad e^{T\psi_x} < 1 + \frac{1}{\|(\mathbf{M}_\nu + \mathbf{N}_\nu)^{-1}\mathbf{N}\|},$$

positive scalars $\bar{\mu}$, h_0, *and* ψ_x *exist such that for every vector* $\mu > 0$ *satisfying* $\max_{1\leq i\leq n_\delta} \mu_i \leq \bar{\mu}$, *every scalar* $h \in (0, h_0]$, *and every pair* $(q^{h,0}, \boldsymbol{\delta}^{h,0})$ *satisfying* (12), *a discrete-time trajectory* (14) *exists satisfying* (16) *for every* $j = 0, 1, \ldots, N_h$. *Moreover,* (24) *holds for* $k = N_h$ *and* (26) *holds for all* $j = 0, 1, \ldots, N_h$.

Proof. Throughout the proof below, the scalars h and μ_i are taken to be sufficiently small so that the previous results can all be applied. More specifically, with the constant r_0 chosen at the end of the proof (cf. (31)), the upper limits h_0 and $\bar{\mu}$ are then guaranteed by Proposition 7. The derivation below emphasizes the process of how the constant r_0 is obtained.

For $\mathbf{x}^{\text{ref}} \equiv (q^{\text{ref}}, \nu^{\text{ref}}, \boldsymbol{\delta}^{\text{ref}})$ in $\Re^n$, let $\nu^h(\mathbf{x}^{\text{ref}})$ be the unique tuple $(q^h, \nu^h, \boldsymbol{\delta}^h)$, which, along with a (unique) triple of friction forces $\boldsymbol{\lambda}^h$, satisfies the following conditions:

$$
\begin{aligned}
& M(q^{\rm ref})(\,\nu^h - \nu^{\rm ref}\,) \,=\, h\,[\,f(t_{h,j+1}, q^{h,\theta_{\rm ref}}, \nu^{h,\theta_{\rm ref}}) + \boldsymbol{\Gamma}(q^{\rm ref})^T \boldsymbol{\lambda}^h\,],\\
& q^h - q^{\rm ref} \,=\, h\,G(q^{\rm ref})\nu^{h,\theta_{\rm ref}},\\
& \delta^h_{\rm t} - \delta^{\rm ref}_{\rm t} \,=\, h\,[s^h_{\rm t} - \Gamma_{\rm t}(q^{\rm ref})\nu^{h,\theta_{\rm ref}}\,],\\
& \delta^h_{\rm o} - \delta^{\rm ref}_{\rm o} \,=\, h\,[s^h_{\rm o} - \Gamma_{\rm o}(q^{\rm ref})\nu^{h,\theta_{\rm ref}}\,],\\
& 0 \,\le\, \lambda^h_{\rm n} \,\perp\, \Psi_{\rm n}(q^{\rm ref}) + h\,\Gamma_{\rm n}(q^{\rm ref})\nu^{h,\theta_{\rm ref}} + \delta^h_{\rm n} \,\ge\, 0,\\
& \boldsymbol{\lambda}^h \,=\, \mathbf{K}(q^{\rm ref})\boldsymbol{\delta}^h + \frac{\mathbf{C}(q^{\rm ref})}{h}\,(\,\boldsymbol{\delta}^h - \boldsymbol{\delta}^{\rm ref}\,),\\
& \begin{pmatrix} \lambda^h_{i{\rm t}} \\ \lambda^h_{i{\rm o}} \end{pmatrix} \in \underset{(\tilde\lambda_{it},\tilde\lambda_{io})\in\mathcal{F}(\mu_i\,\lambda^h_{i{\rm n}})}{\arg\min} \left\{ \begin{pmatrix} s^h_{i{\rm t}} \\ s^h_{i{\rm o}} \end{pmatrix}^T \begin{pmatrix} \tilde\lambda_{it} \\ \tilde\lambda_{io} \end{pmatrix} \right\},
\end{aligned}
$$

where

$$q^{h,\theta_{\rm ref}} \,\equiv\, \theta\,q^{\rm ref} + (\,1-\theta\,)\,q^h \quad \text{and} \quad \nu^{h,\theta_{\rm ref}} \,\equiv\, \theta\,\nu^{\rm ref} + (\,1-\theta\,)\,\nu^h.$$

The well-definedness of $\nu^h(\mathbf{x}^{\rm ref})$ is ensured by Proposition 3 and Lemma 4; moreover, this map is continuous. For $j = 0, 1, \ldots, N_h$, define the maps $\Lambda^{h,j} : \Re^n \to \Re^n$ recursively by $\Lambda^{h,j+1}(\mathbf{x}) \equiv \nu^h(\Lambda^{h,j}(\mathbf{x}))$, where $\Lambda^{h,0}$ is the identity map. Define the auxiliary map $\Phi : \Re^{n_\nu} \to \Re^n$ by $\Phi(\nu) \equiv (q^{h,0}, \nu, \boldsymbol{\delta}^{h,0})$. In terms of these maps we can write the boundary equation (27) as a fixed-point equation: $\nu^{h,0} = \Upsilon(\nu^{h,0})$, where $\Upsilon : \Re^{n_\nu} \to \Re^{n_\nu}$ is the map defined by

$$\Upsilon(\nu) \,=\, (\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}[\,\widehat{\mathbf{b}} - \mathbf{N} \circ (\,\Lambda^{h,N_h+1} - I\,) \circ \Phi(\nu)\,],$$

which is continuous. We claim that a constant $r_0 > 0$ exists such that Υ maps the closed Euclidean ball with center at the origin and radius r_0 into itself. Once this claim is established, Brouwer's fixed-point theorem then shows that the discrete-time boundary system (16) has a solution.

By Proposition 7, we have, for $j = 0, 1, \ldots, N_h$,

$$\| \Lambda^{h,j+1}(\mathbf{x}^{\rm ref}) \| \le R_{h,j+1} \quad \text{and} \quad \| \Lambda^{h,j+1}(\mathbf{x}^{\rm ref}) - \Lambda^{h,j}(\mathbf{x}^{\rm ref}) \| \le R_{h,j+1} - R_{h,j}, \tag{29}$$

where $R_{h,j+1}$ satisfies the recursion

$$R_{h,j+1} \,\equiv\, (\,1 + h\,\psi_x\,)\,R_{h,j} + h\,\psi_x, \quad j \,=\, 0, 1, \ldots, N_h, \tag{30}$$

with $R_{h,0} \ge \|\mathbf{x}^{\rm ref}\|$. Consequently, for any vector $\nu \in \Re^{n_\nu}$, letting $r_0 \ge \|\nu\|$ and $R_{h,0} \equiv r_0 + \|q^0\| + \|\boldsymbol{\delta}^0\|$, we have

$$
\begin{aligned}
\|\Upsilon(\nu)\| &\le \|\,(\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}\widehat{\mathbf{b}}\,\| + \|\,(\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}\mathbf{N}\,\|\,\|\Lambda^{h,N_h+1}(\Phi(\nu)) - \Lambda^{h,0}(\Phi(\nu))\|\\
&\le \|\,(\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}\widehat{\mathbf{b}}\,\| + \|\,(\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}\mathbf{N}\,\|\,(\,R_{h,N_h+1} - R_{h,0}\,)\\
&\le \|\,(\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}\widehat{\mathbf{b}}\,\| + \|\,(\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}\mathbf{N}\,\|\,(\,e^{T\psi_x} - 1\,)(\,1 + R_{h,0}\,),
\end{aligned}
$$

where the last inequality follows from (25), which gives $R_{h,N_h+1} \le e^{T\psi_x}(1 + R_{h,0}) - 1$. Consequently for any $r_0 \ge \|\nu\|$, we have

$$\|\Upsilon(\nu)\| \,\le\, \|\,(\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}\widehat{\mathbf{b}}\,\| + \|\,(\,\mathbf{M}_\nu + \mathbf{N}_\nu\,)^{-1}\mathbf{N}\,\|\,(\,e^{T\psi_x} - 1\,)(\,1 + \|\,q^0\,\| + r_0 + \|\,\boldsymbol{\delta}^0\,\|\,).$$

By (28), it follows that $1 > \|(\mathbf{M}_\nu + \mathbf{N}_\nu)^{-1}\mathbf{N}\|(e^{T\psi_x} - 1)$; hence if

$$r_0 > \frac{\|(\mathbf{M}_\nu + \mathbf{N}_\nu)^{-1}\widehat{\mathbf{b}}\| + \|(\mathbf{M}_\nu + \mathbf{N}_\nu)^{-1}\mathbf{N}\|(e^{T\psi_x} - 1)(1 + \|q^0\| + \|\boldsymbol{\delta}^0\|)}{1 - \|(\mathbf{M}_\nu + \mathbf{N}_\nu)^{-1}\mathbf{N}\|(e^{T\psi_x} - 1)} \tag{31}$$

then $\|\Upsilon(\nu)\| < r_0$. □

5.1. Final convergence. The remaining issue to be dealt with is the convergence of the discrete-time trajectory to a weak solution of the continuous-time frictional compliant contact problem. To deal with this issue, we use the discrete-time iterates $\{\mathbf{x}^{h,0}, \mathbf{x}^{h,1}, \ldots, \mathbf{x}^{h,N_h+1}\}$ to construct a continuous-time state trajectory by linear interpolation. Specifically, define the affine function $\widehat{\mathbf{x}}^h : [0,T] \to \Re^n$ as follows:

$$\widehat{\mathbf{x}}^h(t) \equiv \mathbf{x}^{h,j} + \frac{t - t_{h,j}}{h}(\mathbf{x}^{h,j+1} - \mathbf{x}^{h,j}) \quad \forall t \in [t_{h,j}, t_{h,j+1}].$$

Let $\widehat{\boldsymbol{\lambda}}^h(t)$ be the (possibly discontinuous) piecewise constant interpolants of the families $\{\boldsymbol{\lambda}^{h,j+1}\}$, i.e., $\widehat{\boldsymbol{\lambda}}^h(t) \equiv \boldsymbol{\lambda}^{h,j+1}$ for $t \in (t_{h,j}, t_{h,j+1}]$.

The following theorem is the main convergence result of this paper. Part (c) of the theorem assumes that the constitutive law of compliance for the normal forces is decoupled from that for the tangential forces. In this case, the submatrices $K_{\text{tn}}(q)$, $K_{\text{on}}(q)$, $\widehat{C}_{\text{tn}}(q)$ and $\widehat{C}_{\text{on}}(q)$ are zero, and the tangential friction QP becomes

$$\begin{aligned}
&\text{minimize} \quad \begin{pmatrix} \lambda_{\text{t}} \\ \lambda_{\text{o}} \end{pmatrix} \left\{ \frac{1}{2} \begin{bmatrix} \widehat{C}_{\text{tt}}(q) & \widehat{C}_{\text{to}}(q) \\ \widehat{C}_{\text{ot}}(q) & \widehat{C}_{\text{oo}}(q) \end{bmatrix} \begin{pmatrix} \lambda_{\text{t}} \\ \lambda_{\text{o}} \end{pmatrix} + \begin{bmatrix} \Gamma_{\text{t}}(q) \\ \Gamma_{\text{o}}(q) \end{bmatrix} \nu \right. \\
&\qquad \left. - \begin{bmatrix} \widehat{C}_{\text{tt}}(q) & \widehat{C}_{\text{to}}(q) \\ \widehat{C}_{\text{ot}}(q) & \widehat{C}_{\text{oo}}(q) \end{bmatrix} \begin{bmatrix} K_{\text{tt}}(q) & K_{\text{to}}(q) \\ K_{\text{ot}}(q) & K_{\text{oo}}(q) \end{bmatrix} \begin{pmatrix} \delta_{\text{t}} \\ \delta_{\text{o}} \end{pmatrix} \right\} \\
&\text{subject to} \quad (\lambda_{\text{t}}, \lambda_{\text{o}}) \in \prod_{i=1}^{n_\delta} \mathcal{F}(\mu_i \lambda_{i\text{n}}).
\end{aligned} \tag{32}$$

Theorem 9. *Under the setting of Theorem* 8, *the following statements hold:*

(a) *There is a sequence $\{h_\ell\} \downarrow 0$ such that $\widehat{\mathbf{x}}^{h_\ell}$ converges uniformly on $[0,T]$ to a Lipschitz function $\widehat{\mathbf{x}}$, and $\widehat{\boldsymbol{\lambda}}^{h_\nu}$ converge weakly to a function $\widehat{\boldsymbol{\lambda}}$ in $L^2(0,T)$; i.e.,*

$$\lim_{\ell\to\infty} \sup_{t\in[0,T]} \|\widehat{\mathbf{x}}(t) - \widehat{\mathbf{x}}^{h_\ell}(t)\| = 0$$

and, for any function $\varphi \in L^2(0,T)$,

$$\lim_{\ell\to\infty} \int_0^T \varphi(t)^T \lambda^{h_\ell}_{\text{n,t,o}}(t)\, dt = \int_0^T \varphi(t)^T \widehat{\lambda}_{\text{n,t,o}}(t)\, dt.$$

(b) *All such limits $(\widehat{\mathbf{x}}, \widehat{\boldsymbol{\lambda}})$ satisfy properties* (a), (b), *and* (d) *in Definition* 2 *of a weak solution of the frictional compliant contact problem.*

(c) *If $K_{\text{tn}}(q)$, $K_{\text{on}}(q)$, $C_{\text{tn}}(q)$, and $C_{\text{on}}(q)$ are equal to zero for all q, then $(\widehat{\mathbf{x}}, \widehat{\boldsymbol{\lambda}})$ also satisfies property* (c) *in Definition* 2 *and hence is a weak solution of the frictional compliant contact problem.*

Proof. Combining (24) and (25), we deduce

$$\| \widehat{\mathbf{x}}^h(t) - \mathbf{x}^{h,j} \| \le \| \mathbf{x}^{h,j+1} - \mathbf{x}^{h,j} \| \le h\, \psi_x\, e^{T\psi_x}\, (1 + r_0), \tag{33}$$

where r_0 satisfies (31). By the limit condition (3), we obtain

$$\lim_{h \downarrow 0} \max_{0 \le j \le N_h} \sup_{t \in [t_{h,j}, t_{h,j+1}]} \| \Psi_{\rm n}(\widehat{q}^h(t)) - \Psi_{\rm n}(q^{h,j}) - h\, \Gamma_{\rm n}(q^{h,j}) \nu^{h,\theta_j} \| = 0. \tag{34}$$

Moreover, the former inequalities show that the piecewise interpolants $\widehat{\mathbf{x}}^h$ are not only Lipschitz continuous on $[0,T]$, but the Lipschitz constant is independent of h. Hence there is a positive scalar h_0', which depends only on the model functions such that the family of functions $\{\widehat{\mathbf{x}}^h\}$ for h in $(0, h_0']$ is an equicontinuous family of functions. As in the proof of [11, Theorem 7.1], it follows from the Arzelá–Ascoli theorem (see, e.g., [22, p. 167] or [9, pp. 57–59]) that there is a sequence $\{h_\ell\} \downarrow 0$ such that $\{\widehat{\mathbf{x}}^{h_\ell}\}$ converges in the supremum (i.e., L^∞) norm to a Lipschitz function $\widehat{\mathbf{x}}$ on $[0,T]$. Since

$$\sup_{h \in (0,h_0']} \sup_{t \in [0,T]} \| \widehat{\mathbf{x}}^h(t) \| < \infty,$$

by (26), we deduce that

$$\sup_{h \in (0,h_0']} \sup_{t \in [0,T]} \| \widehat{\boldsymbol{\lambda}}^h(t) \| < \infty. \tag{35}$$

Moreover, by the same proof, it follows that, by working with an appropriate subsequence of $\{h_\ell\}$ if necessary and by invoking Alaoglu's theorem [9, pp. 71–72] and Mazur's theorem [9, p. 88], the sequence $\{\widehat{\boldsymbol{\lambda}}^{h_\ell}\}$ is weakly convergent with a weak* limit $\widehat{\boldsymbol{\lambda}}$, which satisfies $\widehat{\lambda}_{\rm n}(t) \ge 0$ and $(\widehat{\lambda}_{i{\rm t}}(t), \widehat{\lambda}_{i{\rm o}}(t)) \in \mathcal{F}(\mu_i \widehat{\lambda}_{i{\rm n}}(t))$ for almost all t. The proof of the latter frictional inclusion is based on the observation that a pair $(a,b) \in \mathcal{F}(\tau)$ if and only if the triple (a,b,τ) belongs to the closed convex graph of the friction map $\mathcal{F}$.

We need to verify the four properties (a)–(d) of a weak solution to the contact problem. The boundary equation (d) requires no verification, as it is a simple matter of passing to the limit in the discrete-time boundary equation (27). Hence we focus on the verification of (a)–(c). We first deal with the dynamics equations. We have

$$\begin{aligned} \nu^{h,j+1} - \nu^{h,j} &= h\, M(q^{h,j})^{-1} [f(t_{h,j+1}, q^{h,\theta_j}, \nu^{h,\theta_j}) + \boldsymbol{\Gamma}(q^{h,j})^T \boldsymbol{\lambda}^{h,j+1}] \\ &= \int_{t_{h,j}}^{t_{h,j+1}} M(\widehat{q}^h(t))^{-1} [\, f(t, \widehat{q}^h(t), \widehat{\nu}^h(t)) + \boldsymbol{\Gamma}(\widehat{q}^h(t))^T \widehat{\boldsymbol{\lambda}}^h(t) \,]\, dt + O(h^2). \end{aligned}$$

Hence for $0 \le \tau \le \tau' \le T$, we obtain

$$\widehat{\nu}^h(\tau') - \widehat{\nu}^h(\tau) = \int_\tau^{\tau'} M(\widehat{q}^h(t))^{-1} [\, f(t, \widehat{q}^h(t), \widehat{\nu}^h(t)) + \boldsymbol{\Gamma}(\widehat{q}^h(t))^T \widehat{\boldsymbol{\lambda}}^h(t) \,]\, dt + O(h).$$

Restricted to the subsequence $\{h_\ell\}$, we have

$$\lim_{\ell \to \infty} \int_\tau^{\tau'} M(\widehat{q}^{h_\ell}(t))^{-1} f(t, \widehat{q}^{h_\ell}(t), \widehat{\nu}^{h_\ell}(t))\, dt = \int_\tau^{\tau'} M(\widehat{q}(t))^{-1} f(t, \widehat{q}(t), \widehat{\nu}(t))\, dt$$

by the uniform convergence of $(\widehat{q}^{h_\ell}, \widehat{\nu}^{h_\ell}) \to (\widehat{q}, \widehat{\nu})$. We also have

$$\begin{aligned}
&\left\| \int_\tau^{\tau'} \left[M(\widehat{q}^{h_\ell}(t))^{-1}\mathbf{\Gamma}(\widehat{q}^{h_\ell}(t))^T \boldsymbol{\lambda}^{h_\ell}(t) - M(\widehat{q}(t))^{-1}\mathbf{\Gamma}(\widehat{q}(t))^T \widehat{\boldsymbol{\lambda}}(t) \right] \right\| \\
&\quad \le \int_\tau^{\tau'} \left\| M(\widehat{q}^{h_\ell}(t))^{-1}\mathbf{\Gamma}(\widehat{q}^{h_\ell}(t))^T - M(\widehat{q}(t))^{-1}\mathbf{\Gamma}(\widehat{q}(t))^T \right\| \| \widehat{\boldsymbol{\lambda}}^{h_\ell}(t) \| \, dt \\
&\qquad + \left\| \int_\tau^{\tau'} M(\widehat{q}(t))^{-1}\mathbf{\Gamma}(\widehat{q}(t))^T (\widehat{\boldsymbol{\lambda}}^{h_\ell}(t) - \widehat{\boldsymbol{\lambda}}(t)) \, dt \right\|
\end{aligned}$$

The first summand on the right-hand side converges to zero because $\{\widehat{q}^{h_\ell}\} \to \widehat{q}$ uniformly and $\widehat{\boldsymbol{\lambda}}^{h_\ell}$ is bounded; the second summand converges to zero because $\{\widehat{\boldsymbol{\lambda}}^{h_\nu}\}$ converges weakly in $L^2(0,T)$ to $\widehat{\boldsymbol{\lambda}}$. Consequently, we deduce

$$\widehat{\nu}(\tau') - \widehat{\nu}(\tau) = \lim_{\ell\to\infty} [\widehat{\nu}^{h_\ell}(\tau') - \widehat{\nu}^{h_\ell}(\tau)] = \int_\tau^{\tau'} M(\widehat{q}(t))^{-1}[f(t, \widehat{q}(t), \widehat{\nu}(t)) + \mathbf{\Gamma}(\widehat{q}(t))^T \widehat{\boldsymbol{\lambda}}(t)] dt.$$

Similarly, we can establish

$$\begin{aligned}
\widehat{q}(\tau') - \widehat{q}(\tau) &= \int_\tau^{\tau'} G(\widehat{q}(t))\widehat{\nu}(t)\, dt \quad \text{and} \\
\widehat{\boldsymbol{\delta}}(\tau') - \widehat{\boldsymbol{\delta}}(\tau) &= \int_\tau^{\tau'} \mathbf{C}(\widehat{q}(t))^{-1}[\, \widehat{\boldsymbol{\lambda}}(t) - \mathbf{K}(\widehat{q}(t))\widehat{\boldsymbol{\delta}}(t)\,]\, dt,
\end{aligned}$$

completing the proof of property (a) of a weak solution. We next address property (b). For t in $[t_{h,j}, t_{h,j+1}]$, we can write

$$\begin{aligned}
\Psi_{\mathrm{n}}(\widehat{q}^h(t)) + \widehat{\delta}^h_{\mathrm{n}}(t) = {}& \Psi_{\mathrm{n}}(\widehat{q}^h(t)) - \Psi_{\mathrm{n}}(q^{h,j}) + \widehat{\delta}^h_{\mathrm{n}}(t) - \delta^{h,j+1}_{\mathrm{n}} \\
&+ \Psi_{\mathrm{n}}(q^{h,j}) + h\,\Gamma_{\mathrm{n}}(q^{h,j})\nu^{h,\theta_j} + \delta^{h,j+1}_{\mathrm{n}} - h\Gamma_{\mathrm{n}}(q^{h,j})\nu^{h,\theta_j};
\end{aligned}$$

since $\Psi_{\mathrm{n}}(q^{h,j}) + h\,\Gamma_{\mathrm{n}}(q^{h,j})\nu^{h,\theta_j} + \delta^{h,j+1}_{\mathrm{n}} \ge 0$, we deduce

$$\Psi_{\mathrm{n}}(\widehat{q}^h(t)) + \widehat{\delta}^h_{\mathrm{n}}(t) \;\ge\; \Psi_{\mathrm{n}}(\widehat{q}^h(t)) - \Psi_{\mathrm{n}}(q^{h,j}) + \widehat{\delta}^h_{\mathrm{n}}(t) - \delta^{h,j+1}_{\mathrm{n}} - h\,\Gamma_{\mathrm{n}}(q^{h,j})\nu^{h,\theta_j}.$$

Letting $\Phi^h_{\mathrm{n}}(t)$ be the right-hand expression, we deduce from (34) and (33) that $\|\Phi^h_{\mathrm{n}}(t)\|$ is bounded by a constant for all $h > 0$ sufficiently small and all t and $\Phi^h_{\mathrm{n}}(t) \to 0$ for all t as $h \downarrow 0$. Restricted to the subsequence $\{h_\ell\}$, the left-hand side converges uniformly to $\Psi_{\mathrm{n}}(\widehat{q}(t)) + \widehat{\delta}_{\mathrm{n}}(t)$; therefore, $\Psi_{\mathrm{n}}(\widehat{q}(t)) + \widehat{\delta}_{\mathrm{n}}(t) \ge 0$ for all $t \in [0,T]$. Next we show that

$$\int_0^T \widehat{\lambda}_{\mathrm{n}}(t)^T[\, \Psi_{\mathrm{n}}(\widehat{q}(t)) + \widehat{\delta}_{\mathrm{n}}(t)\,]\, dt \;=\; 0. \tag{36}$$

The left-hand side is equal to the limit

$$\lim_{\ell\to\infty} \int_0^T \widehat{\lambda}^{h_\ell}_{\mathrm{n}}(t)^T[\, \Psi_{\mathrm{n}}(\widehat{q}(t)) + \widehat{\delta}_{\mathrm{n}}(t)\,]\, dt \;=\; \lim_{\ell\to\infty} \int_0^T \widehat{\lambda}^{h_\ell}_{\mathrm{n}}(t)^T[\, \Psi_{\mathrm{n}}(\widehat{q}^{h_\ell}(t)) + \widehat{\delta}^{h_\ell}_{\mathrm{n}}(t)\,]\, dt.$$

For each $h > 0$, we have

$$\begin{aligned}\int_0^T \widehat{\lambda}_{\rm n}^h(t)^T[\,\Psi_{\rm n}(\widehat{q}^{\,h}(t)) + \widehat{\delta}_{\rm n}^h(t)\,]\,dt &= \sum_{j=0}^{N_h}\int_{t_{h,j}}^{t_{h,j+1}} \widehat{\lambda}_{\rm n}^h(t)^T[\,\Psi_{\rm n}(\widehat{q}^{\,h}(t)) + \widehat{\delta}_{\rm n}^h(t)\,]\,dt \\ &= \sum_{j=0}^{N_h}\int_{t_{h,j}}^{t_{h,j+1}} (\,\lambda^{h,j+1}\,)^T\Phi_{\rm n}^h(t)\,dt. \end{aligned} \tag{37}$$

Since $\{\lambda^{h,j+1}\}$ is bounded, by letting $h \downarrow 0$ in (37) along the subsequence $\{h_\ell\}$, (36) follows readily from the dominated convergence theorem, thereby completing the proof of property (b) of weak solution.

Finally, to prove property (c), let $(\widetilde{\lambda}_{\rm t}, \widetilde{\lambda}_{\rm o}) : [0,T] \to \Re^{2n_\delta}$ be continuous functions such that $(\widetilde{\lambda}_{i{\rm t}}(t), \widetilde{\lambda}_{i{\rm o}}(t)) \in \mathcal{F}(\mu_i\widehat{\lambda}_{i{\rm n}}(t))$ for almost all $t \in [0,T]$ and all $i = 1,\dots,n_\delta$. We need to verify

$$\int_0^T \begin{pmatrix} \widetilde{\lambda}_{\rm t}(t) - \widehat{\lambda}_{\rm t}(t) \\ \widetilde{\lambda}_{\rm o}(t) - \widehat{\lambda}_{\rm o}(t) \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \left[\begin{pmatrix} \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}(t) \end{pmatrix} \right.\right.$$
$$\left.\left. - \begin{bmatrix} K_{\rm tt}(\widehat{q}(t)) & K_{\rm to}(\widehat{q}(t)) \\ K_{\rm ot}(\widehat{q}(t)) & K_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\delta}_{\rm t}(t) \\ \widehat{\delta}_{\rm o}(t) \end{pmatrix} \right] + \begin{pmatrix} \Gamma_{\rm t}(\widehat{q}(t)) \\ \Gamma_{\rm o}(\widehat{q}(t)) \end{pmatrix} \widehat{\nu}(t) \right\} dt \,\geq\, 0.$$

Since

$$\int_0^T \begin{pmatrix} \widehat{\lambda}_{\rm t}^{h_\ell}(t) \\ \widehat{\lambda}_{\rm o}^{h_\ell}(t) \end{pmatrix}^T \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\lambda}_{\rm t}^{h_\ell}(t) \\ \widehat{\lambda}_{\rm o}^{h_\ell}(t) \end{pmatrix} dt$$
$$= \int_0^T \begin{pmatrix} \widehat{\lambda}_{\rm t}^{h_\ell}(t) - \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}^{h_\ell}(t) - \widehat{\lambda}_{\rm o}(t) \end{pmatrix}^T \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\lambda}_{\rm t}^{h_\ell}(t) - \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}^{h_\ell}(t) - \widehat{\lambda}_{\rm o}(t) \end{pmatrix} dt$$
$$-2\int_0^T \begin{pmatrix} \widehat{\lambda}_{\rm t}^{h_\ell}(t) - \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}^{h_\ell}(t) - \widehat{\lambda}_{\rm o}(t) \end{pmatrix}^T \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}(t) \end{pmatrix} dt$$
$$+\int_0^T \begin{pmatrix} \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}(t) \end{pmatrix}^T \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}(t) \end{pmatrix} dt,$$

and since the first integral on the right-hand side is nonnegative (by the positive semidefiniteness of the quadratic form), the second integral converges to zero because $\widehat{\lambda}_{\rm t,o}^{h_\ell}$ converge to $\widehat{\lambda}_{\rm t,o}$ in $L^2(0,T)$, we deduce

$$\begin{aligned} \infty \;&>\; \liminf_{\ell\to\infty} \int_0^T \begin{pmatrix} \widehat{\lambda}_{\rm t}^{h_\ell}(t) \\ \widehat{\lambda}_{\rm o}^{h_\ell}(t) \end{pmatrix}^T \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\lambda}_{\rm t}^{h_\ell}(t) \\ \widehat{\lambda}_{\rm o}^{h_\ell}(t) \end{pmatrix} dt \\ &\geq\; \int_0^T \begin{pmatrix} \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}(t) \end{pmatrix}^T \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}(t) \end{pmatrix} dt, \end{aligned} \tag{38}$$

where the left-hand limit is finite by (35). Consequently, it follows that

$$\begin{aligned}
&\int_0^T \begin{pmatrix} \widetilde{\lambda}_{\rm t}(t) - \widehat{\lambda}_{\rm t}(t) \\ \widetilde{\lambda}_{\rm o}(t) - \widehat{\lambda}_{\rm o}(t) \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \left[\begin{pmatrix} \widehat{\lambda}_{\rm t}(t) \\ \widehat{\lambda}_{\rm o}(t) \end{pmatrix} \right.\right. \\
&\qquad \left.\left. - \begin{bmatrix} K_{\rm tt}(\widehat{q}(t)) & K_{\rm to}(\widehat{q}(t)) \\ K_{\rm ot}(\widehat{q}(t)) & K_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\delta}_{\rm t}(t) \\ \widehat{\delta}_{\rm o}(t) \end{pmatrix} \right] + \begin{pmatrix} \Gamma_{\rm t}(\widehat{q}(t)) \\ \Gamma_{\rm o}(\widehat{q}(t)) \end{pmatrix} \widehat{\nu}(t) \right\} dt \\
&\geq \limsup_{\ell\to\infty} \int_0^T \begin{pmatrix} \widetilde{\lambda}_{\rm t}(t) - \widehat{\lambda}^{h_\ell}_{\rm t}(t) \\ \widetilde{\lambda}_{\rm o}(t) - \widehat{\lambda}^{h_\ell}_{\rm o}(t) \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}(t)) & \widehat{C}_{\rm to}(\widehat{q}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}(t)) & \widehat{C}_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \left[\begin{pmatrix} \widehat{\lambda}^{h_\ell}_{\rm t}(t) \\ \widehat{\lambda}^{h_\ell}_{\rm o}(t) \end{pmatrix} \right.\right. \\
&\qquad \left.\left. - \begin{bmatrix} K_{\rm tt}(\widehat{q}(t)) & K_{\rm to}(\widehat{q}(t)) \\ K_{\rm ot}(\widehat{q}(t)) & K_{\rm oo}(\widehat{q}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\delta}_{\rm t}(t) \\ \widehat{\delta}_{\rm o}(t) \end{pmatrix} \right] + \begin{pmatrix} \Gamma_{\rm t}(\widehat{q}(t)) \\ \Gamma_{\rm o}(\widehat{q}(t)) \end{pmatrix} \widehat{\nu}(t) \right\} dt \quad \text{by (38)} \\
&\geq \limsup_{\ell\to\infty} \int_0^T \begin{pmatrix} \widetilde{\lambda}_{\rm t}(t) - \widehat{\lambda}^{h_\ell}_{\rm t}(t) \\ \widetilde{\lambda}_{\rm o}(t) - \widehat{\lambda}^{h_\ell}_{\rm o}(t) \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}^{h_\ell}(t)) & \widehat{C}_{\rm to}(\widehat{q}^{h_\ell}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}^{h_\ell}(t)) & \widehat{C}_{\rm oo}(\widehat{q}^{h_\ell}(t)) \end{bmatrix} \left[\begin{pmatrix} \widehat{\lambda}^{h_\ell}_{\rm t}(t) \\ \widehat{\lambda}^{h_\ell}_{\rm o}(t) \end{pmatrix} \right.\right. \\
&\qquad \left.\left. - \begin{bmatrix} K_{\rm tt}(\widehat{q}^{h_\ell}(t)) & K_{\rm to}(\widehat{q}^{h_\ell}(t)) \\ K_{\rm ot}(\widehat{q}^{h_\ell}(t)) & K_{\rm oo}(\widehat{q}^{h_\ell}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\delta}^{h_\ell}_{\rm t}(t) \\ \widehat{\delta}^{h_\ell}_{\rm o}(t) \end{pmatrix} \right] + \begin{pmatrix} \Gamma_{\rm t}(\widehat{q}^{h_\ell}(t)) \\ \Gamma_{\rm o}(\widehat{q}^{h_\ell}(t)) \end{pmatrix} \widehat{\nu}^{h_\ell}(t) \right\} dt,
\end{aligned}$$

where the second inequality holds because $\{(\widehat{q}^{h_\ell}, \widehat{\nu}^{h_\ell}, \widehat{\delta}^{h_\ell}_{\rm t,o})\}$ converges to $(\widehat{q}, \widehat{\nu}, \widehat{\delta}_{\rm t,o})$ uniformly. For each $h > 0$, we have

$$\begin{aligned}
&\int_0^T \begin{pmatrix} \widetilde{\lambda}_{\rm t}(t) - \widehat{\lambda}^{h}_{\rm t}(t) \\ \widetilde{\lambda}_{\rm o}(t) - \widehat{\lambda}^{h}_{\rm o}(t) \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}^{h}(t)) & \widehat{C}_{\rm to}(\widehat{q}^{h}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}^{h}(t)) & \widehat{C}_{\rm oo}(\widehat{q}^{h}(t)) \end{bmatrix} \left[\begin{pmatrix} \widehat{\lambda}^{h}_{\rm t}(t) \\ \widehat{\lambda}^{h}_{\rm o}(t) \end{pmatrix} \right.\right. \\
&\qquad \left.\left. - \begin{bmatrix} K_{\rm tt}(\widehat{q}^{h}(t)) & K_{\rm to}(\widehat{q}^{h}(t)) \\ K_{\rm ot}(\widehat{q}^{h}(t)) & K_{\rm oo}(\widehat{q}^{h}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\delta}^{h}_{\rm t}(t) \\ \widehat{\delta}^{h}_{\rm o}(t) \end{pmatrix} \right] + \begin{pmatrix} \Gamma_{\rm t}(\widehat{q}^{h}(t)) \\ \Gamma_{\rm o}(\widehat{q}^{h}(t)) \end{pmatrix} \widehat{\nu}^{h}(t) \right\} dt \\
&= \sum_{j=1}^{N_h} \int_{t_{h,j}}^{t_{h,j+1}} \begin{pmatrix} \widetilde{\lambda}_{\rm t}(t) - \lambda^{h,j+1}_{\rm t} \\ \widetilde{\lambda}_{\rm o}(t) - \lambda^{h,j+1}_{\rm o} \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\rm tt}(\widehat{q}^{h}(t)) & \widehat{C}_{\rm to}(\widehat{q}^{h}(t)) \\ \widehat{C}_{\rm ot}(\widehat{q}^{h}(t)) & \widehat{C}_{\rm oo}(\widehat{q}^{h}(t)) \end{bmatrix} \left[\begin{pmatrix} \lambda^{h,j+1}_{\rm t} \\ \lambda^{h,j+1}_{\rm o} \end{pmatrix} \right.\right. \\
&\qquad \left.\left. - \begin{bmatrix} K_{\rm tt}(\widehat{q}^{h}(t)) & K_{\rm to}(\widehat{q}^{h}(t)) \\ K_{\rm ot}(\widehat{q}^{h}(t)) & K_{\rm oo}(\widehat{q}^{h}(t)) \end{bmatrix} \begin{pmatrix} \widehat{\delta}^{h}_{\rm t}(t) \\ \widehat{\delta}^{h}_{\rm o}(t) \end{pmatrix} \right] + \begin{pmatrix} \Gamma_{\rm t}(\widehat{q}^{h}(t)) \\ \Gamma_{\rm o}(\widehat{q}^{h}(t)) \end{pmatrix} \widehat{\nu}^{h}(t) \right\} dt.
\end{aligned}$$

Since for almost all $t \in (t_{h,j}, t_{h,j+1}]$, we have $(\widetilde{\lambda}_{it}(t), \widetilde{\lambda}_{io}(t)) \in \mathcal{F}(\mu_i \lambda^{h,j+1}_{\rm in})$, it follows that

$$\begin{pmatrix} \widetilde{\lambda}_{\rm t}(t) - \lambda_{\rm t}^{h,j+1} \\ \widetilde{\lambda}_{\rm o}(t) - \lambda_{\rm o}^{h,j+1} \end{pmatrix}^T \left\{ \begin{bmatrix} \widehat{C}_{\rm tt}(q^{h,j}) & \widehat{C}_{\rm to}(q^{h,j}) \\ \widehat{C}_{\rm ot}(q^{h,j}) & \widehat{C}_{\rm oo}(q^{h,j}) \end{bmatrix} \left[\begin{pmatrix} \lambda_{\rm t}^{h,j+1} \\ \lambda_{\rm o}^{h,j+1} \end{pmatrix} \right.\right.$$

$$\left.\left. - \begin{bmatrix} K_{\rm tt}(q^{h,j}) & K_{\rm to}(q^{h,j}) \\ K_{\rm ot}(q^{h,j}) & K_{\rm oo}(q^{h,j}) \end{bmatrix} \begin{pmatrix} \delta_{\rm t}^{h,j} \\ \delta_{\rm o}^{h,j} \end{pmatrix} \right] + \begin{pmatrix} \Gamma_{\rm t}(q^{h,j}) \\ \Gamma_{\rm o}(q^{h,j}) \end{pmatrix} \nu^{h,\theta_j} \right\} \geq 0$$

for almost all $t \in (t_{h,j}, t_{h,j+1}]$. Since the state variables satisfy

$$\lim_{\substack{\ell \to \infty \\ t \in (t_{h_\ell,j}, t_{h_\ell,j+1}]}} (\widehat{\mathbf{x}}^{h_\ell}(t) - \mathbf{x}^{h_\ell,j}) = 0$$

uniformly on $[0,T]$, we easily derive the desired limiting friction property (c). □

6. Conclusion and discussion.

This paper provided an in-depth investigation of time-stepping methods for rigid body dynamics problems with multiple contacts characterized by friction and local compliance. The main results are (a) the existence of a discrete-time solution trajectory the boundary-value problem (Theorem 8), and (b) the convergence of such a solution to a weak solution of the corresponding continuous-time problem (Theorem 9). Whereas the convergence results obtained are in a sense stronger than those in [19, 1], it is worth noting that this is because of our choice of a phenomenologically correct model that explicitly characterizes the compliance at each contact. Even so, there are limitations in our investigation. First, the friction coefficients are required to be sufficiently small in the main results (this is the result of our discretization which respects the nonlinear friction conditions at all iterates $\boldsymbol{\lambda}^{h,\nu+1}$). Second, we are not able to establish convergence to a strong solution. This limitation begs the question of whether such a solution can be proved to exist in a continuous-time model under an appropriate compliance constitutive law. The key difficulty lies in the fact that the friction forces are not continuous functions of the system states with the model (5). This issue remains unresolved to date. Third, in our convergence analysis, the parameters of the compliance model (i.e., the stiffness and damping) are fixed. It would be very interesting to extend the analysis to allow these parameters to tend to infinity, with the goal of recovering a solution of some kind to a fully rigid-body model. Such an extended analysis is beyond the scope of this paper. In the previous paper [17], we considered, in a discrete-time framework with a fixed discretization step, the issue of convergence when the stiffness and damping both tend to infinity and obtained some positive results; nevertheless, such a convergence issue in a continuous-time model seems difficult and has not been studied.

In view of the unresolved issues associated with strong solutions, which are seemingly very difficult, our results are significant and provide a first step for a deeper analysis. Needless to say, we are interested in extending the analysis to models with nonlinear constitutive laws for which the existence of strong solutions to the continuous-time model could be shown and for which the convergence of a numerical time-stepping method to such a solution could be established. Our future work will address such extensions and the application of numerical methods for solving boundary value problems to the optimal design of manufacturing processes with frictional contacts.

Acknowledgements. This paper has benefitted from the constructive comments of two referees and the associate editor. The authors are also grateful to David Stewart who has offered some valuable insights on the properties of the compliant contact model relative to the rigid body model.

REFERENCES

[1] M. Anitescu, *Optimization-Based Simulation for Nonsmooth Rigid Multibody Dynamics*, Technical Report ANL/MCS-P1161-0504, Division of Mathematics and Computer Science, Argonne National Laboratory, Argonne, IL, 2004.

[2] M. Anitescu and F. Potra, *A time-stepping method for stiff multi-body dynamics with friction and contact*, Internat. J. Numer. Methods Engrg., 55 (2002), pp. 753–784.

[3] M. Anitescu, F. Potra, and D. Stewart, *Time-stepping for three-dimensional rigid body dynamics*, Methods Appl. Mech. Engrg., 17 (1999), pp. 183–197.

[4] D. Balkcom, E. J. Gottlieb, and J. C. Trinkle, *A sensorless insertion strategy for rigid planar parts*, in Proceedings of the IEEE International Conference on Robotics and Automation, 2002, pp. 882–887.

[5] B. Brogliato, *Nonsmooth Impact Mechanics—Models, Dynamics, and Control*, Lecture Notes in Control and Inform. Sci., vol. 220, Springer-Verlag, London, 1996.

[6] R. W. Cottle, J. S. Pang, and R. E. Stone, *The Linear Complementarity Problem*, Academic Press, Boston, 1992.

[7] S. P. Dirkse and M. C. Ferris, *The path solver: A non-monotone stabilization scheme for mixed complementarity problems*, Optim. Methods Soft., 5 (1995), pp. 123–156.

[8] F. Facchinei and J. S. Pang, *Finite-Dimensional Variational Inequalities and Complementarity Problems*, Springer-Verlag, New York, 2003.

[9] S. Lang, *Real and Functional Analysis*, 3rd ed., Springer, Berlin, 1993.

[10] J. S. Pang and D. E. Stewart, *A unified approach to discrete frictional contact problems*, Internat. J. Engrg. Sci., 37 (1999), pp. 1747–1768.

[11] J. S. Pang and D. E. Stewart, *Differential variational inequalities*, Math. Program., revision in review.

[12] J. S. Pang and J. C. Trinkle, *Complementarity formulations and existence of solutions of multi-rigid-body contact problems with Coulomb friction*, Math. Program., 73 (1996), pp. 199–226.

[13] F. Pfeiffer and Ch. Glocker, *Multibody Dynamics with Unilateral Contacts*, John Wiley, New York, 1996.

[14] P. Song, *Modeling, Analysis and Simulation of Multibody Systems with Contact and Friction*, Ph.D. thesis, Department of Mechanical Engineering and Applied Mechanics, University of Pennsylvania, Philadelphia, 2002.

[15] P. Song, P. Kraus, V. Kumar, and P. Dupont, *Analysis of rigid–body dynamic models for simulation of systems with frictional contacts*, J. Appl. Mech., 68 (2001), pp. 118–128.

[16] P. Song, V. Kumar, and J. S. Pang, *A two-point boundary-value approach for planning manipulation tasks*, in Proceedings of Robotics: Science and Systems, Cambridge, MA, 2005, http://www.roboticsproceedings.org/rss01.

[17] P. Song, J. S. Pang, and V. Kumar, *A semi-implicit time-stepping model for frictional compliant contact problems*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 2231–2261.

[18] P. Song, J.C. Trinkle, V. Kumar, and J.S. Pang, *Design of part feeding and assembly processes with dynamics*, in Proceedings of the 2004 IEEE International Conference on Robotics and Automation, 2004, pp. 39–44.

[19] D. Stewart, *Convergence of a time-stepping scheme for rigid-body dynamics and resolution of painlevé's problem*, Arch. Ration. Mech. Anal., 145 (1998), pp. 215–260.

[20] D. Stewart, *Rigid-body dynamics with friction and impact*, SIAM Rev., 42 (2000), pp. 3–39.

[21] D. Stewart and J. Trinkle, *An implicit time-stepping scheme for rigid-body dynamics with inelastic collisions and coulomb friction*, Internat. J. Numer. Methods Engrg., 39 (1996), pp. 2673–2691.

[22] K. R. Stromberg, *Introduction to Classical Real Analysis*, Wadsworth, Belmont, CA, 1981.

[23] Y. T. Wang and V. Kumar, *Simulation of mechanical systems with unilateral constraints*, J. Mech. Design, 116 (1994), pp. 571–580.

[24] J.A. Tzitzouris, *Numerical Resolution of Frictional Multi-Rigid-Body Systems via Fully-Implicit Time-Stepping and Nonlinear Complementarity*, Ph.D. thesis, Department of Mathematical Sciences, Johns Hopkins University, Baltimore, 2001.

[25] J. C. Trinkle, J. A. Tzitzouris, and J. S. Pang, *Dynamic multi-rigid-systems with concurrent distributed contacts*, Roy. Soc. Philos. Trans. Math. Physical Engrg. Sci., 359 (2001), pp. 2575–2593.

SIAM J. NUMER. ANAL.
Vol. 43, No. 5, pp. 2227–2249

A POSTERIORI ERROR ANALYSIS OF THE LINKED INTERPOLATION TECHNIQUE FOR PLATE BENDING PROBLEMS*

CARLO LOVADINA† AND ROLF STENBERG‡

Abstract. We develop a posteriori error estimates for the so-called linked interpolation technique to approximate the solution of plate bending problems. We show that the proposed (residual-based) estimator is both reliable and efficient.

Key words. Reissner–Mindlin plates, finite element methods, a posteriori error analysis.

AMS subject classifications. Primary: 65N30; Secondary: 74S05

DOI. 10.1137/040614645

1. Introduction.

In this paper we present an a posteriori error analysis for the so-called linked interpolation technique (cf. [2], [3], and [21], for instance) to approximate the solution of the Reissner–Mindlin plate problem.

It is worth noticing that the main effort concerning the finite element discretization of the plate bending problems has been focused on proposing and analyzing locking-free schemes. As a consequence, most of the mathematical literature on the subject is addressed to establish a priori error estimates. We mention here the works [1], [4], [6], [12], [13], [15], [18], [20], and the references therein, for example. On the contrary, when considering the a posteriori error analysis for plates, only very few results are available (see [7], [8], and [14]).

In this work we consider the so-called linked interpolation technique focusing on two triangular elements: the low-order element proposed in [21] (see also [22]), and the quadratic scheme proposed in [3]. An a priori error analysis has been developed for both the methods in [16], [17] and [3], respectively. We also remark that our a posteriori error analysis may be straightforwardly extended to other schemes taking advantage of the linked interpolation technique, such as the quadrilateral elements considered in [2] and [3], for example.

An outline of the paper is as follows. In section 2 we briefly recall the Reissner–Mindlin problem, together with a mixed variational formulation and some useful regularity results. The linked interpolation technique is described in section 3, where we develop an a priori analysis for the sake of completeness (see also [16] or [17]). Section 4 is devoted to the a posteriori error estimates. In particular we introduce our estimator, and we prove its reliability (section 4.1) and efficiency (section 4.2). We consider the case of a clamped plate only for simplicity. Indeed, both the a priori and the a posteriori error analysis can be easily adapted to cover other relevant boundary conditions.

Throughout the paper we use standard notations for Sobolev norms and seminorms (see [5], for example). Moreover, we denote with C a generic constant

*Received by the editors September 8, 2004; accepted for publication (in revised form) June 12, 2005; published electronically December 30, 2005. This work has been supported by European Project HPRN-CT-2002-00284.

http://www.siam.org/journals/sinum/43-5/61464.html

†Dipartimento di Matematica, Università di Pavia, and IMATI-CNR, Via Ferrata 1, Pavia I-27100, Italy (carlo.lovadina@unipv.it).

‡Institute of Mathematics, Helsinki University of Technology, P.O. Box 1100, 02015 HUT, Finland (stenberg@hut.fi).

independent of the mesh parameter h and the plate thickness t, which may take different values in different occurrences.

2. The Reissner–Mindlin problem. The Reissner–Mindlin equations for a clamped plate with polygonal midplane Ω require one to find $(\boldsymbol{\theta}, w, \boldsymbol{\gamma})$ such that

$$\begin{cases} -\operatorname{div} \mathbf{C}\,\varepsilon(\boldsymbol{\theta}) - \boldsymbol{\gamma} = 0 & \text{in } \Omega, \\ -\operatorname{div} \boldsymbol{\gamma} = g & \text{in } \Omega, \\ \boldsymbol{\gamma} = \mu t^{-2}(\boldsymbol{\nabla} w - \boldsymbol{\theta}) & \text{in } \Omega, \\ \boldsymbol{\theta} = 0, \ w = 0 & \text{on } \partial\Omega. \end{cases} \tag{2.1}$$

Here, $\mathbf{C}$ is the tensor of bending moduli, $\boldsymbol{\theta}$ represents the rotations, w the transversal displacement, $\boldsymbol{\gamma}$ the scaled shear stresses, and g a given transversal load. Moreover, ε is the usual symmetric gradient operator, μ is the shear modulus, and t is the thickness. The classical variational formulation of problem (2.1) is

$$\begin{cases} \text{Find } (\boldsymbol{\theta}, w, \boldsymbol{\gamma}) \in \boldsymbol{\Theta} \times W \times (L^2(\Omega))^2 : \\ a(\boldsymbol{\theta}, \boldsymbol{\eta}) + (\boldsymbol{\nabla} v - \boldsymbol{\eta}, \boldsymbol{\gamma}) = (g, v), & (\boldsymbol{\eta}, v) \in \boldsymbol{\Theta} \times W, \\ (\boldsymbol{\nabla} w - \boldsymbol{\theta}, \boldsymbol{\tau}) - \mu^{-1} t^2 (\boldsymbol{\gamma}, \boldsymbol{\tau}) = 0, & \boldsymbol{\tau} \in (L^2(\Omega))^2, \end{cases} \tag{2.2}$$

where $\boldsymbol{\Theta} = (H_0^1(\Omega))^2$, $W = H_0^1(\Omega)$, $(\cdot, \cdot)$ is the inner product in $L^2(\Omega)$ and

$$a(\boldsymbol{\theta}, \boldsymbol{\eta}) := \int_\Omega \mathbf{C}\,\varepsilon(\boldsymbol{\theta}) : \varepsilon(\boldsymbol{\eta}).$$

Following [9], we write the pair $(\boldsymbol{\theta}, w)$ as

$$(\boldsymbol{\theta}, w) = (\boldsymbol{\theta}_0 + \boldsymbol{\theta}_r, w_0 + w_r), \tag{2.3}$$

where the pair $(\boldsymbol{\theta}_0, , w_0)$ is the solution of the limit problem,

$$\begin{cases} \text{Find } (\boldsymbol{\theta}_0, w_0, \boldsymbol{\gamma}_0) \in \boldsymbol{\Theta} \times W \times \boldsymbol{\Gamma} : \\ a(\boldsymbol{\theta}_0, \boldsymbol{\eta}) + < \boldsymbol{\nabla} v - \boldsymbol{\eta}, \boldsymbol{\gamma}_0 >= (g, v), & (\boldsymbol{\eta}, v) \in \boldsymbol{\Theta} \times W, \\ < \boldsymbol{\nabla} w_0 - \boldsymbol{\theta}_0, \boldsymbol{\tau} >= 0, & \boldsymbol{\tau} \in \boldsymbol{\Gamma}, \end{cases} \tag{2.4}$$

and $(\boldsymbol{\theta}_r, w_r)$ can be thought of as a remainder. Furthermore, $\boldsymbol{\Gamma} = H^{-1}(\operatorname{div}, \Omega)$ and $< \cdot, \cdot >$ is the duality pairing between $H_0(\operatorname{rot}, \Omega)$ and $H^{-1}(\operatorname{div}, \Omega)$. One has the following proposition (cf. [9]).

PROPOSITION 2.1. *Suppose that Ω is convex and $g \in L^2(\Omega)$. Then it holds*

$$||w_0||_3 + ||\boldsymbol{\theta}||_2 + ||\boldsymbol{\gamma}||_0 + t\,||\boldsymbol{\gamma}||_1 \leq C(||g||_{-1} + t\,||g||_0), \tag{2.5}$$

$$||\boldsymbol{\theta}_r||_1 \leq Ct\,||g||_{-1}, \tag{2.6}$$

$$||w_r||_2 \leq Ct(||g||_{-1} + t\,||g||_0). \qquad \square \tag{2.7}$$

3. The linked interpolation scheme and an a priori analysis.

In this section we present the general idea of the linked interpolation technique (see [3] and [21], for instance), together with two examples of triangular elements. Furthermore, for the sake of completeness, we develop an a priori error analysis, focusing on the lowest-order element (see [16] and [17]).

3.1. The linked interpolation scheme.

Let $\{\mathcal{T}_h\}_{h>0}$ be a sequence of decompositions of Ω into triangular elements T, satisfying the usual compatibility conditions (see [11]). We also assume that the family $\{\mathcal{T}_h\}_{h>0}$ is regular, i.e., there exists a constant $\sigma > 0$ such that

$$h_T \leq \sigma \rho_T \qquad \forall T \in \mathcal{T}_h, \tag{3.1}$$

where h_T is the diameter of the element T and ρ_T is the maximum diameter of the circles contained in T. We recall (see [11], for instance) that regularity implies the minimum angle condition: there exists a constant $\alpha > 0$ such that

$$\alpha_T \geq \alpha \qquad \forall T \in \mathcal{T}_h, \tag{3.2}$$

where α_T denotes the smallest inner angle of T. Moreover, given the decomposition $\mathcal{T}_h$ we will denote with $\mathcal{E}_h$ the set of the edges e of the triangles $T \in \mathcal{T}_h$. We now select the finite element spaces $\boldsymbol{\Theta}_h \subset \boldsymbol{\Theta}$, $W_h \subset W$, $\boldsymbol{\Gamma}_h \subset L^2(\Omega)^2$, together with a suitable linear operator (the so-called linking operator)

$$L : \boldsymbol{\Theta}_h \longrightarrow H^1_0(\Omega). \tag{3.3}$$

We then form the finite dimensional subspace of $\boldsymbol{X} := \boldsymbol{\Theta} \times W$

$$\boldsymbol{X}_h = \big\{(\boldsymbol{\eta}_h, v_h^*) = (\boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) : \boldsymbol{\eta}_h \in \boldsymbol{\Theta}_h\,,\ v_h \in W_h\big\}, \tag{3.4}$$

and we finally consider the discrete problem

$$\begin{cases} \text{Find } (\boldsymbol{\theta}_h, w_h^*; \boldsymbol{\gamma}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h : \\ a(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) + (\boldsymbol{\gamma}_h, \boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h) = (g, v_h^*), \qquad (\boldsymbol{\eta}_h, v_h^*) \in \boldsymbol{X}_h, \\ (\boldsymbol{\nabla} w_h^* - \boldsymbol{\theta}_h, \boldsymbol{\tau}_h) - \mu^{-1} t^2 (\boldsymbol{\gamma}_h, \boldsymbol{\tau}_h) = 0, \qquad \boldsymbol{\tau}_h \in \boldsymbol{\Gamma}_h. \end{cases} \tag{3.5}$$

Remark 3.1. We point out that eliminating $\boldsymbol{\gamma}_h$ from system (3.5), our scheme is equivalent to the following problem involving only the rotations and the vertical displacements:

$$\begin{cases} \text{Find } (\boldsymbol{\theta}_h, w_h^*) \in \boldsymbol{X}_h : \\ a(\boldsymbol{\theta}_h, \boldsymbol{\eta}_h) + \mu t^{-2} \big(P_h(\boldsymbol{\nabla} w_h^* - \boldsymbol{\theta}_h), P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h)\big) = (g, v_h) \quad \forall (\boldsymbol{\eta}_h, v_h^*) \in \boldsymbol{X}_h, \end{cases} \tag{3.6}$$

where P_h denotes the L^2-projection operator onto $\boldsymbol{\Gamma}_h$.

We are now ready to present the following two elements. (For other methods based on the same strategy, see, e.g., [2], [3].)

3.1.1. The linear element. This element (see [21]) is described by the finite element spaces

$$\boldsymbol{\Theta}_h = \big\{\boldsymbol{\eta} \in \boldsymbol{\Theta} : \boldsymbol{\eta}_{|T} \in (P_1(T) \oplus B_3(T))^2\big\}, \tag{3.7}$$

$$W_h = \big\{v \in W : v_{|T} \in P_1(T)\big\}, \tag{3.8}$$

$$\boldsymbol{\Gamma}_h = \big\{\boldsymbol{\tau} \in L^2(\Omega)^2 : \boldsymbol{\tau}_{|T} \in P_0(T)^2\big\}, \tag{3.9}$$

where $P_k(T)$ is the space of polynomials of degree at most k defined on T and $B_3(T) = P_3(T) \cap H_0^1(T)$ is the space of cubic bubbles on T. The linking operator $L : \boldsymbol{\Theta}_h \longrightarrow H_0^1(\Omega)$ is defined as follows. For each $T \in \mathcal{T}_h$, we set

$$\varphi_i = \lambda_j \lambda_k \qquad \text{and} \qquad EB_2(T) = \operatorname{Span}\{\varphi_i\}_{1\le i\le 3}\,, \tag{3.10}$$

where $\{\lambda_i\}_{1\le i\le 3}$ are the barycentric coordinates of the triangle T and the indices (i, j, k) form a permutation of the set $(1, 2, 3)$. Then, the operator L is locally defined as

$$L\boldsymbol{\eta}_{h|T} = \sum_{i=1}^{3} \alpha_i \varphi_i \in EB_2(T), \tag{3.11}$$

where the coefficients α_i are determined by requiring that

$$(\boldsymbol{\nabla} L\boldsymbol{\eta}_h - \boldsymbol{\eta}_h) \cdot \mathbf{t} \quad \text{is constant on each } e. \tag{3.12}$$

Above, $\mathbf{t}$ denotes the tangential vector to the edge e. We recall that for the linking operator it holds (see [16] and [17])

$$||L\boldsymbol{\eta}_h||_{0,T} \le Ch_T||\boldsymbol{\nabla} L\boldsymbol{\eta}_h||_{0,T}, \qquad ||\boldsymbol{\nabla} L\boldsymbol{\eta}_h||_{0,T} \le Ch_T|\boldsymbol{\eta}_h|_{1,T}. \tag{3.13}$$

3.1.2. The quadratic element. This element (see [3]) is described by the finite element spaces

$$\boldsymbol{\Theta}_h = \big\{\boldsymbol{\eta} \in \boldsymbol{\Theta} : \boldsymbol{\eta}_{|T} \in P_2(T)^2 \oplus (P_1(T)^2 \oplus \boldsymbol{\nabla} B_3(T))b_T\big\}, \tag{3.14}$$

$$W_h = \big\{v \in W : v_{|T} \in P_2(T) \oplus B_3(T)\big\}, \tag{3.15}$$

$$\boldsymbol{\Gamma}_h = \big\{\boldsymbol{\tau} \in L^2(\Omega)^2 : \boldsymbol{\tau}_{|T} \in P_1(T)^2 \oplus \boldsymbol{\nabla} B_3(T)\big\}, \tag{3.16}$$

where $b_T = 27\lambda_1\lambda_2\lambda_3$. The linking operator $L : \boldsymbol{\Theta}_h \longrightarrow H_0^1(\Omega)$ is defined as follows. For each $T \in \mathcal{T}_h$, we set

$$\varphi_i = \lambda_j \lambda_k(\lambda_k - \lambda_j) \qquad \text{and} \qquad EB_3(T) = \operatorname{Span}\{\varphi_i\}_{1\le i\le 3}\,, \tag{3.17}$$

where the indices (i, j, k) form a permutation of the set $(1, 2, 3)$. Then, the operator L is locally defined as

$$L\boldsymbol{\eta}_{h|T} = \sum_{i=1}^{3} \alpha_i \varphi_i \in EB_3(T), \tag{3.18}$$

where the coefficients α_i are determined by requiring that

$$(\boldsymbol{\nabla} L\boldsymbol{\eta}_h - \boldsymbol{\eta}_h) \cdot \mathbf{t} \quad \text{is linear on each } e. \tag{3.19}$$

For this linking operator it holds (see [3])

$$||L\boldsymbol{\eta}_h||_{0,T} \le Ch_T||\boldsymbol{\nabla} L\boldsymbol{\eta}_h||_{0,T}, \qquad ||\boldsymbol{\nabla} L\boldsymbol{\eta}_h||_{0,T} \le Ch_T^2|\boldsymbol{\eta}_h|_{2,T} \le Ch_T|\boldsymbol{\eta}_h|_{1,T}. \tag{3.20}$$

3.2. A priori error estimates. Following the lines of [9], [10], [16], [18], [20], we prove a priori error estimates with respect to the norms

$$\|\!|(\boldsymbol{\eta}, v)\|\!|_h^2 := ||\boldsymbol{\eta}||_1^2 + ||v||_1^2 + \sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||\boldsymbol{\nabla} v - \boldsymbol{\eta}||_{0,T}^2 \qquad \forall(\boldsymbol{\eta}, v) \in \boldsymbol{\Theta}\times W \tag{3.21}$$

and

$$||\boldsymbol{\tau}||_{-1} + t\,||\boldsymbol{\tau}||_0 \qquad \forall\boldsymbol{\tau}\in L^2(\Omega)^2. \tag{3.22}$$

We will also use the following discrete norm:

$$||\boldsymbol{\tau}||_h^2 := \sum_{T\in\mathcal{T}_h} h_T^2||\boldsymbol{\tau}||_{0,T}^2 + t^2||\boldsymbol{\tau}||_0^2 \qquad \forall\boldsymbol{\tau}\in L^2(\Omega)^2. \tag{3.23}$$

Before proceeding, we need the following lemma, which establishes a suitable norm equivalence in the used finite element spaces.

Lemma 3.1. *Consider the finite element spaces and the linking operator detailed in section* 3.1.1 *(or in section* 3.1.2*), and let P_h denote the L^2-projection operator on $\boldsymbol{\Gamma}_h$. Then for each $(\boldsymbol{\eta}_h, v_h^*) \in \boldsymbol{X}_h$ it holds*

$$\left(||\boldsymbol{\eta}_h||_1^2 + \sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h)||_{0,T}^2\right)^{1/2} \leq \|\!|(\boldsymbol{\eta}_h, v_h^*)\|\!|_h \tag{3.24}$$

and

$$\|\!|(\boldsymbol{\eta}_h, v_h^*)\|\!|_h \leq C\left(||\boldsymbol{\eta}_h||_1^2 + \sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h)||_{0,T}^2\right)^{1/2}. \tag{3.25}$$

Proof. Since (3.24) is trivial, we consider only (3.25). Therefore, take $\boldsymbol{\eta}_h \in \boldsymbol{\Theta}_h$, $v_h \in W_h$ and form $(\boldsymbol{\eta}_h, v_h^*) = (\boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) \in \boldsymbol{X}_h$. We first notice that

$$\begin{aligned} ||\boldsymbol{\nabla} v_h^*||_0^2 &\leq 2\left(||\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h||_0^2 + ||\boldsymbol{\eta}_h||_0^2\right) \\ &\leq C\left(\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h||_{0,T}^2 + ||\boldsymbol{\eta}_h||_1^2\right), \end{aligned} \tag{3.26}$$

so that, by Poincaré's inequality, we have

$$||v_h^*||_1^2 \leq C\left(\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h||_{0,T}^2 + ||\boldsymbol{\eta}_h||_1^2\right), \tag{3.27}$$

Next, we write $\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h$ as

$$\begin{aligned} \boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h &= \boldsymbol{\nabla} v_h + \boldsymbol{\nabla} L\boldsymbol{\eta}_h - \boldsymbol{\eta}_h = P_h\boldsymbol{\nabla} v_h + \boldsymbol{\nabla} L\boldsymbol{\eta}_h - \boldsymbol{\eta}_h \\ &= P_h\boldsymbol{\nabla} v_h^* - (P_h\boldsymbol{\nabla} L\boldsymbol{\eta}_h - \boldsymbol{\nabla} L\boldsymbol{\eta}_h) - \boldsymbol{\eta}_h \\ &= P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h) - (P_h\boldsymbol{\nabla} L\boldsymbol{\eta}_h - \boldsymbol{\nabla} L\boldsymbol{\eta}_h) + (P_h\boldsymbol{\eta}_h - \boldsymbol{\eta}_h). \end{aligned} \tag{3.28}$$

Therefore, we have

$$\begin{aligned}||\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h||_{0,T} \leq ||P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h)||_{0,T} \\ + ||P_h \boldsymbol{\nabla} L\boldsymbol{\eta}_h - \boldsymbol{\nabla} L\boldsymbol{\eta}_h||_{0,T} + ||P_h\boldsymbol{\eta}_h - \boldsymbol{\eta}_h||_{0,T}.\end{aligned} \tag{3.29}$$

Since (see also (3.13) and (3.20))

$$||P_h \boldsymbol{\nabla} L\boldsymbol{\eta}_h - \boldsymbol{\nabla} L\boldsymbol{\eta}_h||_{0,T} \leq 2||\boldsymbol{\nabla} L\boldsymbol{\eta}_h||_{0,T} \leq Ch_T|\boldsymbol{\eta}_h|_{1,T} \tag{3.30}$$

and

$$||P_h\boldsymbol{\eta}_h - \boldsymbol{\eta}_h||_{0,T} \leq Ch_T|\boldsymbol{\eta}_h|_{1,T}, \tag{3.31}$$

from (3.29) we obtain

$$\begin{aligned}\frac{1}{h_T^2+t^2}||\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h||_{0,T}^2 &\leq C\left(\frac{1}{h_T^2+t^2}||P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h)||_{0,T}^2 + \frac{h_T^2}{h_T^2+t^2}|\boldsymbol{\eta}_h|_{1,T}^2\right) \\ &\leq C\left(\frac{1}{h_T^2+t^2}||P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h)||_{0,T}^2 + |\boldsymbol{\eta}_h|_{1,T}^2\right).\end{aligned} \tag{3.32}$$

Therefore, we get

$$\sum_{T\in\mathcal{T}_h}\frac{1}{h_T^2+t^2}||\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h||_{0,T}^2 \leq C\left(\sum_{T\in\mathcal{T}_h}\frac{1}{h_T^2+t^2}||P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h)||_{0,T}^2 + ||\boldsymbol{\eta}_h||_1^2\right). \tag{3.33}$$

Using (3.27) and (3.31) we deduce estimate (3.25). □

It is now useful to set

$$\begin{aligned}\mathcal{A}(\boldsymbol{\theta}, w, \boldsymbol{\gamma}; \boldsymbol{\eta}, v, \boldsymbol{\tau}) := a(\boldsymbol{\theta}, \boldsymbol{\eta}) + (\boldsymbol{\nabla} v - \boldsymbol{\eta}, \boldsymbol{\gamma}) \\ - (\boldsymbol{\nabla} w - \boldsymbol{\theta}, \boldsymbol{\tau}) + \mu^{-1}t^2(\boldsymbol{\gamma}, \boldsymbol{\tau}).\end{aligned} \tag{3.34}$$

Therefore, the continuous problem (2.2) reads

$$\begin{cases}\text{Find } (\boldsymbol{\theta}, w; \boldsymbol{\gamma}) \in \boldsymbol{X} \times L^2(\Omega)^2 \text{ such that} \\ \mathcal{A}(\boldsymbol{\theta}, w, \boldsymbol{\gamma}; \boldsymbol{\eta}, v, \boldsymbol{\tau}) = (g, v) \qquad \forall(\boldsymbol{\eta}, v; \boldsymbol{\tau}) \in \boldsymbol{X} \times L^2(\Omega)^2,\end{cases} \tag{3.35}$$

while the discrete problem (3.5) is

$$\begin{cases}\text{Find } (\boldsymbol{\theta}_h, w_h^*; \boldsymbol{\gamma}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h \text{ such that} \\ \mathcal{A}(\boldsymbol{\theta}_h, w_h^*, \boldsymbol{\gamma}_h; \boldsymbol{\eta}_h, v_h^*, \boldsymbol{\tau}_h) = (g, v_h^*) \qquad \forall(\boldsymbol{\eta}_h, v_h^*; \boldsymbol{\tau}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h.\end{cases} \tag{3.36}$$

We have the following stability result, for which we only sketch the proof, since it takes advantage of the same techniques detailed in [9] and [16].

PROPOSITION 3.2. *Given $(\boldsymbol{\beta}_h, z_h^*; \boldsymbol{\rho}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$ there exists $(\boldsymbol{\eta}_h, v_h^*; \boldsymbol{\tau}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$ such that*

$$\mathcal{A}(\boldsymbol{\beta}_h, z_h^*, \boldsymbol{\rho}_h; \boldsymbol{\eta}_h, v_h^*, \boldsymbol{\tau}_h) \geq C\left(|||(\boldsymbol{\beta}_h, z_h^*)|||_h^2 + ||\boldsymbol{\rho}_h||_h^2\right), \tag{3.37}$$

$$\|\!|(\boldsymbol{\eta}_h, v_h^*)|\!\|_h + ||\boldsymbol{\tau}_h||_h \leq C\left(\|\!|(\boldsymbol{\beta}_h, z_h^*)|\!\|_h + ||\boldsymbol{\rho}_h||_h\right). \tag{3.38}$$

Proof. Let us $(\boldsymbol{\beta}_h, z_h^*; \boldsymbol{\rho}_h)$ be given in $\boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$. Using exactly the same arguments of [9] and [16] we get that there exists $(\boldsymbol{\eta}_h, v_h^*; \boldsymbol{\tau}_h)$ in $\boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$ such that

$$\mathcal{A}(\boldsymbol{\beta}_h, z_h^*, \boldsymbol{\rho}_h; \boldsymbol{\eta}_h, v_h^*, \boldsymbol{\tau}_h) \geq C\left(||\boldsymbol{\beta}_h||_1^2 + \sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2 + t^2}||P_h(\boldsymbol{\nabla} z_h^* - \boldsymbol{\beta}_h)||_{0,T}^2 + ||\boldsymbol{\rho}_h||_h^2\right) \tag{3.39}$$

and

$$\begin{aligned}||\boldsymbol{\eta}_h||_1 &+ \left(\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2 + t^2}||P_h(\boldsymbol{\nabla} v_h^* - \boldsymbol{\eta}_h)||_{0,T}^2\right)^{1/2} + ||\boldsymbol{\tau}_h||_h \\ &\leq C\left(||\boldsymbol{\beta}_h||_1 + \left(\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2 + t^2}||P_h(\boldsymbol{\nabla} z_h^* - \boldsymbol{\beta}_h)||_{0,T}^2\right)^{1/2} + ||\boldsymbol{\rho}_h||_h\right).\end{aligned} \tag{3.40}$$

We now use Lemma 3.1 to infer that given $(\boldsymbol{\beta}_h, z_h^*; \boldsymbol{\rho}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$, there exists $(\boldsymbol{\eta}_h, v_h^*; \boldsymbol{\tau}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$ such that

$$\mathcal{A}(\boldsymbol{\beta}_h, z_h^*, \boldsymbol{\rho}_h; \boldsymbol{\eta}_h, v_h^*, \boldsymbol{\tau}_h) \geq C\left(\|\!|(\boldsymbol{\beta}_h, z_h^*)|\!\|_h^2 + ||\boldsymbol{\rho}_h||_h^2\right) \tag{3.41}$$

and

$$\|\!|(\boldsymbol{\eta}_h, v_h^*)|\!\|_h + ||\boldsymbol{\tau}_h||_h \leq C\left(\|\!|(\boldsymbol{\beta}_h, z_h^*)|\!\|_h + ||\boldsymbol{\rho}_h||_h\right). \qquad \square \tag{3.42}$$

We are now ready to prove our error estimate (see also [17] and [16]). We focus on the lowest-order element detailed in section 3.1.1, but a similar technique (together with the ideas developed in [18]) may be applied to appropriately treat the higher-order case of section 3.1.2.

PROPOSITION 3.3. *Suppose that Ω is a convex polygon and $g \in L^2(\Omega)$ and consider the element detailed in section* 3.1.1. *Let $(\boldsymbol{\theta}, w; \boldsymbol{\gamma}) \in \boldsymbol{X} \times L^2(\Omega)^2$ and $(\boldsymbol{\theta}_h, w_h^*; \boldsymbol{\gamma}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$ be the solutions of problem* (3.35) *and* (3.36), *respectively. Then the following a priori estimates holds:*

$$\|\!|(\boldsymbol{\theta} - \boldsymbol{\theta}_h, w - w_h^*)|\!\|_h + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{-1} + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_h \leq C\, h(||g||_{-1} + t\,||g||_0). \tag{3.43}$$

Proof. Since our method is consistent (cf. (3.35) and (3.36)) and stable (see Proposition 3.2), error estimates with respect to the norms in question can be established in the standard way. Hence, let

$$(\boldsymbol{\theta}_I, w_I^*; \boldsymbol{\gamma}_I) = (\boldsymbol{\theta}_I, w_I + L\boldsymbol{\theta}_I; \boldsymbol{\gamma}_I) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h \tag{3.44}$$

be a suitable interpolant (to be specified later) of the continuous solution $(\boldsymbol{\theta}, w^*; \boldsymbol{\gamma})$. Corresponding to $(\boldsymbol{\theta}_h - \boldsymbol{\theta}_I, w_h^* - w_I^*; \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$ there exists (see Proposition 3.2) $(\boldsymbol{\eta}_h, v_h^*; \boldsymbol{\tau}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$ such that

$$\mathcal{A}(\boldsymbol{\theta}_h - \boldsymbol{\theta}_I, w_h^* - w_I^*, \boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I; \boldsymbol{\eta}_h, v_h^*, \boldsymbol{\tau}_h) \geq C\big(\|\!|(\boldsymbol{\theta}_h - \boldsymbol{\theta}_I, w_h^* - w_I^*)|\!\|_h^2 + ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||_h^2\big), \tag{3.45}$$

and

$$\|\!|(\boldsymbol{\eta}_h, v_h^*)|\!\|_h + ||\boldsymbol{\tau}_h||_h \leq C\left(\|\!|(\boldsymbol{\theta}_h - \boldsymbol{\theta}_I, w_h^* - w_I^*)|\!\|_h + ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I||_h\right). \tag{3.46}$$

By consistency it holds

$$\begin{aligned}\mathcal{A}(\boldsymbol{\theta}_h-\boldsymbol{\theta}_I, w_h^*-w_I^*,\boldsymbol{\gamma}_h-\boldsymbol{\gamma}_I;\boldsymbol{\eta}_h,v_h^*,\boldsymbol{\tau}_h) &= \mathcal{A}(\boldsymbol{\theta}-\boldsymbol{\theta}_I, w-w_I^*,\boldsymbol{\gamma}-\boldsymbol{\gamma}_I;\boldsymbol{\eta}_h,v_h^*,\boldsymbol{\tau}_h)\\ &= a(\boldsymbol{\theta}-\boldsymbol{\theta}_I,\boldsymbol{\eta}_h)+(\boldsymbol{\nabla} v_h^*-\boldsymbol{\eta}_h,\boldsymbol{\gamma}-\boldsymbol{\gamma}_I)\\ &-\big(\boldsymbol{\nabla}(w-w_I^*)-(\boldsymbol{\theta}-\boldsymbol{\theta}_I),\boldsymbol{\tau}_h\big)+\mu^{-1}t^2(\boldsymbol{\gamma}-\boldsymbol{\gamma}_I,\boldsymbol{\tau}_h)\\ &= \text{(I)}+\text{(II)}+\text{(III)}+\text{(IV)}.\end{aligned} \tag{3.47}$$

To bound the four terms above, we first choose the interpolants $\boldsymbol{\theta}_I$, w_I^* and $\boldsymbol{\gamma}_I$ as follows. According to the splitting (2.3), $\boldsymbol{\theta}_I$ is given by

$$\boldsymbol{\theta}_I := \mathcal{I}\boldsymbol{\theta} = \mathcal{I}\boldsymbol{\theta}_0 + \mathcal{I}\boldsymbol{\theta}_r, \tag{3.48}$$

where $\mathcal{I}$ is the Lagrange interpolating operator. To define w_I^*, we need to specify w_I (cf. (3.44)). Again, the splitting (2.3) suggests to set

$$w_I := \mathcal{I}w = \mathcal{I}w_0 + \mathcal{I}w_r. \tag{3.49}$$

Therefore, w_I^* turns out to be $w_I^* = w_I + L\boldsymbol{\theta}_I = \mathcal{I}w + L(\mathcal{I}\boldsymbol{\theta})$. Finally, $\boldsymbol{\gamma}_I$ is simply the L^2-projection of $\boldsymbol{\gamma}$ onto $\boldsymbol{\Gamma}_h$.

Estimate for (I). Using the H^1-continuity of the bilinear form $a(\cdot,\cdot)$, standard approximation results, and estimate (2.5) we have

$$\text{(I)} = a(\boldsymbol{\theta}-\boldsymbol{\theta}_I,\boldsymbol{\eta}_h) \le Ch||\boldsymbol{\theta}||_2||\boldsymbol{\eta}_h||_1 \le Ch(||g||_{-1}+t\,||g||_0)||\boldsymbol{\eta}_h||_1. \tag{3.50}$$

Estimate for (II). We notice that

$$\begin{aligned}\text{(II)} &= (\boldsymbol{\nabla} v_h^*-\boldsymbol{\eta}_h,\boldsymbol{\gamma}-\boldsymbol{\gamma}_I)\\ &\le \left(\sum_{T\in\mathcal{T}_h}\frac{1}{h_T^2+t^2}||\boldsymbol{\nabla} v_h^*-\boldsymbol{\eta}_h||_{0,T}^2\right)^{1/2}\left(\sum_{T\in\mathcal{T}_h}(h_T^2+t^2)||\boldsymbol{\gamma}-\boldsymbol{\gamma}_I||_{0,T}^2\right)^{1/2},\end{aligned} \tag{3.51}$$

by which, using again (2.5) and standard approximation estimates, we get

$$\text{(II)} \le Ch(||g||_{-1}+t\,||g||_0)\left(\sum_{T\in\mathcal{T}_h}\frac{1}{h_T^2+t^2}||\boldsymbol{\nabla} v_h^*-\boldsymbol{\eta}_h||_{0,T}^2\right)^{1/2}. \tag{3.52}$$

Estimate for (III).

$$\begin{aligned}\text{(III)} &= -\big(\boldsymbol{\nabla}(w-w_I^*)-(\boldsymbol{\theta}-\boldsymbol{\theta}_I),\boldsymbol{\tau}_h\big)\\ &\le \left(\sum_{T\in\mathcal{T}_h}\frac{1}{h_T^2+t^2}||\boldsymbol{\nabla}(w-w_I^*)-(\boldsymbol{\theta}-\boldsymbol{\theta}_I)||_{0,T}^2\right)^{1/2}\left(\sum_{T\in\mathcal{T}_h}(h_T^2+t^2)||\boldsymbol{\tau}_h||_{0,T}^2\right)^{1/2}.\end{aligned} \tag{3.53}$$

We now notice that we have (see (2.3), (3.44) and (3.48)–(3.49))

$$\begin{aligned}\boldsymbol{\nabla}(w-w_I^*)-(\boldsymbol{\theta}-\boldsymbol{\theta}_I) &= \Big\{\boldsymbol{\nabla}\big(w_0-\mathcal{I}w_0-L(\mathcal{I}\boldsymbol{\theta}_0)\big)-(\boldsymbol{\theta}_0-\mathcal{I}\boldsymbol{\theta}_0)\Big\}\\ &+\Big\{\boldsymbol{\nabla}\big(w_r-\mathcal{I}w_r-L(\mathcal{I}\boldsymbol{\theta}_r)\big)-(\boldsymbol{\theta}_r-\mathcal{I}\boldsymbol{\theta}_r)\Big\}.\end{aligned} \tag{3.54}$$

In [16] it has been proved that

$$\big|\boldsymbol{\nabla}\big(w_0 - \mathcal{I}w_0 - L(\mathcal{I}\boldsymbol{\theta}_0)\big)\big|_{0,T} \le Ch_T^2|w_0|_{3,T}, \tag{3.55}$$

while standard approximation results give

$$|\boldsymbol{\theta}_0 - \mathcal{I}\boldsymbol{\theta}_0|_{0,T} \le Ch_T^2|\boldsymbol{\theta}_0|_{2,T}, \tag{3.56}$$

$$|\boldsymbol{\theta}_r - \mathcal{I}\boldsymbol{\theta}_r|_{0,T} \le Ch_T^2|\boldsymbol{\theta}_r|_{2,T}. \tag{3.57}$$

Furthermore, using also (3.13) it holds

$$\begin{aligned}\big|\boldsymbol{\nabla}\big(w_r - \mathcal{I}w_r - L(\mathcal{I}\boldsymbol{\theta}_r)\big)\big|_{0,T} &\le |\boldsymbol{\nabla}(w_r - \mathcal{I}w_r)|_{0,T} + |\boldsymbol{\nabla}L(\mathcal{I}\boldsymbol{\theta}_r)|_{0,T} \\ &\le |\boldsymbol{\nabla}(w_r - \mathcal{I}w_r)|_{0,T} + |\boldsymbol{\nabla}L(\mathcal{I}\boldsymbol{\theta}_r - \boldsymbol{\theta}_r)|_{0,T} + |\boldsymbol{\nabla}L(\boldsymbol{\theta}_r)|_{0,T} \\ &\le C\big(h_T|w_r|_{2,T} + h_T|\mathcal{I}\boldsymbol{\theta}_r - \boldsymbol{\theta}_r|_{1,T} + h_T|\boldsymbol{\theta}_r|_{1,T}\big) \\ &\le C\big(h_T|w_r|_{2,T} + h_T^2|\boldsymbol{\theta}_r|_{2,T} + h_T|\boldsymbol{\theta}_r|_{1,T}\big).\end{aligned} \tag{3.58}$$

From (3.54)–(3.58) we obtain

$$\begin{aligned}&\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||\boldsymbol{\nabla}(w - w_I^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_I)||_{0,T}^2 \\ &\qquad \le C\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}\left(h_T^4|w_0|_{3,T}^2 + h_T^4|\boldsymbol{\theta}|_{2,T}^2 + h_T^2|w_r|_{2,T}^2 + h_T^2|\boldsymbol{\theta}_r|_{1,T}^2\right) \\ &\qquad \le Ch^2\big(|w_0|_3^2 + |\boldsymbol{\theta}|_2^2\big) + \sum_{T\in\mathcal{T}_h}\frac{h_T^2}{h_T^2+t^2}\big(|w_r|_{2,T}^2 + |\boldsymbol{\theta}_r|_{1,T}^2\big) \\ &\qquad \le Ch^2\big(|w_0|_3^2 + |\boldsymbol{\theta}|_2^2\big) + \sum_{T\in\mathcal{T}_h} h_T^2\left(\frac{|w_r|_{2,T}^2}{t^2} + \frac{|\boldsymbol{\theta}_r|_{1,T}^2}{t^2}\right) \\ &\qquad \le Ch^2\left(|w_0|_3^2 + |\boldsymbol{\theta}|_2^2 + \frac{|w_r|_2^2}{t^2} + \frac{|\boldsymbol{\theta}_r|_1^2}{t^2}\right).\end{aligned} \tag{3.59}$$

Using (2.5)–(2.7), from (3.59) it follows that

$$\begin{aligned}&\left(\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||\boldsymbol{\nabla}(w - w_I^*) - (\boldsymbol{\theta} - \boldsymbol{\theta}_I)||_{0,T}^2\right)^{1/2} \\ &\qquad \le Ch\left(||w_0||_3 + ||\boldsymbol{\theta}||_2 + \frac{||w_r||_2}{t} + \frac{||\boldsymbol{\theta}_r||_1}{t}\right) \\ &\qquad \le Ch(||g||_{-1} + t\,||g||_0).\end{aligned} \tag{3.60}$$

Therefore, we obtain (see (3.53))

$$\text{(III)} \le Ch(||g||_{-1} + t\,||g||_0)\left(\sum_{T\in\mathcal{T}_h}(h_T^2+t^2)||\boldsymbol{\tau}_h||_{0,T}^2\right)^{1/2}. \tag{3.61}$$

Estimate for (IV). We simply notice that

$$(\mathrm{IV}) = \mu^{-1}t^2(\boldsymbol{\gamma} - \boldsymbol{\gamma}_I, \boldsymbol{\tau}_h) \leq Ct\,||\boldsymbol{\gamma} - \boldsymbol{\gamma}_I||_0 t\,||\boldsymbol{\tau}_h||_0 \leq Ch(||g||_{-1} + t\,||g||_0)t\,||\boldsymbol{\tau}_h||_0. \tag{3.62}$$

Collecting (3.50), (3.52), (3.61), and (3.62), from (3.47) we get

$$\begin{aligned}\mathcal{A}(\boldsymbol{\theta}_h - \boldsymbol{\theta}_I, w_h^* - w_I^*, &\boldsymbol{\gamma}_h - \boldsymbol{\gamma}_I; \boldsymbol{\eta}_h, v_h^*, \boldsymbol{\tau}_h)\\ &\leq Ch(||g||_{-1} + t\,||g||_0)\left(|||(\boldsymbol{\eta}_h, v_h^*)|||_h + ||\boldsymbol{\tau}_h||_h\right).\end{aligned} \tag{3.63}$$

From (3.45), (3.46), (3.63), and the triangle inequality, we infer

$$|||(\boldsymbol{\theta} - \boldsymbol{\theta}_h, w - w_h^*)|||_h + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_h \leq C\,h(||g||_{-1} + t\,||g||_0). \tag{3.64}$$

To obtain the error in the H^{-1}-norm for the shears, we use the Pitkäranta–Verfürth trick (cf. [19], [23], and [20]). Hence, we recall that

$$||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{-1} = \sup_{\boldsymbol{\eta} \in \boldsymbol{\Theta}} \frac{(\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\eta})}{||\boldsymbol{\eta}||_1}. \tag{3.65}$$

For a generic $\boldsymbol{\eta} \in \boldsymbol{\Theta}$ we consider its Clemént's interpolant $\boldsymbol{\eta}^c \in \boldsymbol{\Theta}_h$ (see [11], for instance), and we write

$$(\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\eta}) = (\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\eta} - \boldsymbol{\eta}^c) + (\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\eta}^c). \tag{3.66}$$

On the one hand, we have

$$\begin{aligned}(\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\eta} - \boldsymbol{\eta}^c) &\leq \left(\sum_{T \in \mathcal{T}_h} h_T^2 ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{0,T}^2\right)^{1/2} \left(\sum_{T \in \mathcal{T}_h} h_T^{-2} ||\boldsymbol{\eta} - \boldsymbol{\eta}^c||_{0,T}^2\right)^{1/2}\\ &\leq C||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_h ||\boldsymbol{\eta}||_1.\end{aligned} \tag{3.67}$$

On the other hand, recalling (2.2) and (3.5), we get

$$(\boldsymbol{\gamma} - \boldsymbol{\gamma}_h, \boldsymbol{\eta}^c) = a(\boldsymbol{\theta} - \boldsymbol{\theta}_h, \boldsymbol{\eta}^c) \leq C||\boldsymbol{\theta} - \boldsymbol{\theta}_h||_1 ||\boldsymbol{\eta}^c||_1 \leq C||\boldsymbol{\theta} - \boldsymbol{\theta}_h||_1 ||\boldsymbol{\eta}||_1. \tag{3.68}$$

From (3.65)–(3.68), we obtain

$$||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{-1} \leq C\left(||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_h + ||\boldsymbol{\theta} - \boldsymbol{\theta}_h||_1\right). \tag{3.69}$$

Estimate (3.43) now follows from (3.64) and (3.69). □

Using the technique in [9], one may also get the following improved estimates.

PROPOSITION 3.4. *Suppose that* Ω *is a convex polygon and* $g \in L^2(\Omega)$. *Then the following a priori estimates hold:*

$$||\boldsymbol{\theta} - \boldsymbol{\theta}_h||_0 \leq Ch^2(||g||_{-1} + t\,||g||_0), \tag{3.70}$$

$$||w - w_h^*||_1 \leq Ch(h + t)(||g||_{-1} + t\,||g||_0). \quad \square \tag{3.71}$$

4. A posteriori error estimates. The aim of this section is to introduce suitable error estimator for the elements based on the linked interpolation technique and to prove its reliability and efficiency. To begin, for each $T \in \mathcal{T}_h$ and $e \in \mathcal{E}_h$ we introduce the following quantities:

$$\begin{aligned}\widetilde{\eta}_T^2 := h_T^2 || \operatorname{div} \mathbf{C}\, \varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h ||_{0,T}^2 + h_T^2(h_T^2 + t^2) || \operatorname{div} \boldsymbol{\gamma}_h + g_h ||_{0,T}^2 \\ + \frac{\mu^2}{h_T^2 + t^2} || \mu^{-1} t^2\, \boldsymbol{\gamma}_h - (\boldsymbol{\nabla} w_h^* - \boldsymbol{\theta}_h) ||_{0,T}^2,\end{aligned} \tag{4.1}$$

$$\eta_e^2 := h_e || \left[\!\left[\mathbf{C}\, \varepsilon(\boldsymbol{\theta}_h) \mathbf{n} \right]\!\right] ||_{0,e}^2 + h_e(h_e^2 + t^2) || \left[\!\left[\boldsymbol{\gamma}_h \cdot \mathbf{n} \right]\!\right] ||_{0,e}^2, \tag{4.2}$$

where g_h is some approximation of the load g. Moreover, h_e is the length of the side e and $[\![\cdot]\!]$ denotes the jump operator. We then define a local indicator η_T as

$$\eta_T := \left(\widetilde{\eta}_T^2 + \sum_{e \subset \partial T} \eta_e^2 \right)^{1/2} \tag{4.3}$$

and a global indicator η as

$$\eta := \left(\sum_{T \in \mathcal{T}_h} \widetilde{\eta}_T^2 + \sum_{e \in \mathcal{E}_h} \eta_e^2 \right)^{1/2}. \tag{4.4}$$

Remark 4.1. When considering the element described in section 3.1.1, the expression in (4.1) becomes simpler, since we locally have $\operatorname{div} \boldsymbol{\gamma}_h = 0$ (see (3.9)).

We now introduce some useful notation. Given a generic $e \in \mathcal{E}_h$, we denote with ω_e the union of the triangles in $\mathcal{T}_h$ having e as a side. Furthermore, for $T \in \mathcal{T}_h$ we set ω_T as the union of the ω_e, with $e \subset \partial T$. We proceed with the following result.

LEMMA 4.1. *Given $e \in \mathcal{E}_h$, let $P_k(e)$ be the space of polynomials of degree at most k defined on e. There exists a linear operator*

$$\Pi_e \,:\, P_k(e) \longrightarrow H_0^2(\omega_e) \tag{4.5}$$

such that for all $p_k \in P_k(e)$ it holds

$$C_1 ||p_k||_{0,e}^2 \leq \int_e p_k \big(\Pi_e p_k\big) \leq ||p_k||_{0,e}^2, \tag{4.6}$$

$$||\Pi_e p_k||_{0,\omega_e} \leq C_2 h_e^{1/2} ||p_k||_{0,e}, \tag{4.7}$$

$$|\boldsymbol{\nabla}(\Pi_e p_k)|_{0,\omega_e} \leq C_3 h_e^{-1/2} ||p_k||_{0,e}, \tag{4.8}$$

$$|\boldsymbol{\nabla}(\Pi_e p_k)|_{1,\omega_e} \leq C_4 h_e^{-3/2} ||p_k||_{0,e}. \tag{4.9}$$

Above, the constants C_i depend only on k and on the minimum angle of the triangles in the meshes $\mathcal{T}_h$.

Proof. We consider only the case of an interior edge e: if e is a boundary edge (i.e., $e \subset \partial\Omega$), the required modifications are obvious. Due to the minimum angle condition, there exists a fixed reference rhomb $\widehat{D}$, as depicted in Figure 4.1, where, e.g., $\delta = \alpha/2$ (see (3.2)), with the following property: for each $e \in \mathcal{E}_h$ it is possible to determine a rhomb $D_e \subseteq \omega_e$ similar to $\widehat{D}$ (see Figure 4.2). According to Figure 4.2,

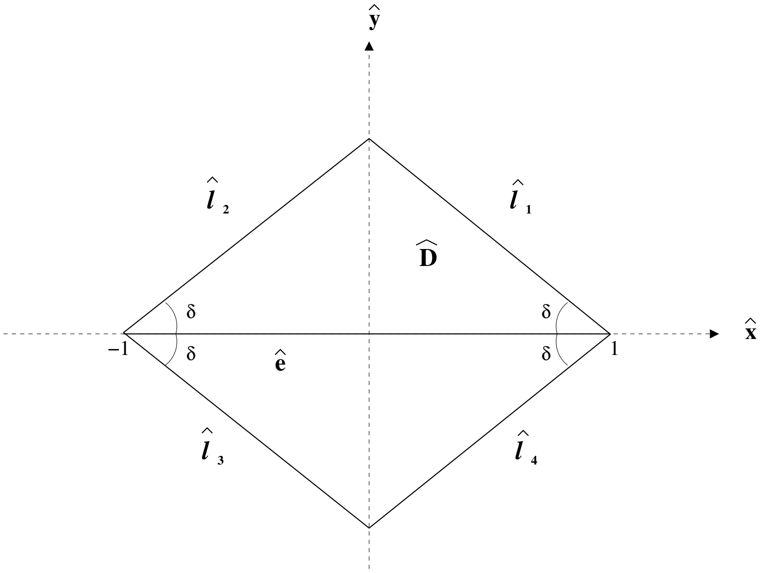


FIG. 4.1. *The reference rhomb $\widehat{D}$.*

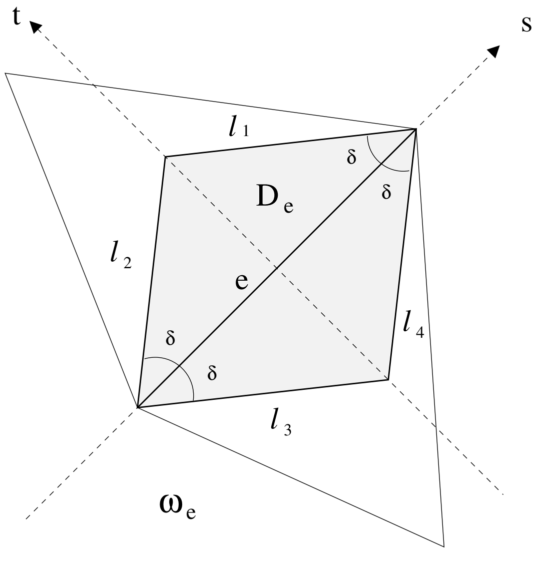


FIG. 4.2. *Relevant objects associated with the edge e.*

on ω_e we now introduce local Cartesian coordinates (s, t), as well as the functions

(4.10) $d_i(s, t) =$ "distance of (s, t) from the edge l_i", $i = 1, \dots, 4$ (see Figure 4.2).

Next, we define $\psi_e(s, t) : \omega_e \longrightarrow \mathbf{R}$ as

$$\psi_e(s, t) := \alpha_e\, \chi_{D_e}(s, t) \prod_{i=1}^{4} d_i(s, t)^2, \tag{4.11}$$

where $\chi_{D_e}(s, t)$ is the characteristic function of the set D_e, while α_e is a normalization constant in order to have $||\psi_e||_\infty = 1$. We also notice that in the coordinates (s,t) a generic polynomial $p_k \in P_k(e)$ can be simply written as $p_k(s)$. We are ready to define $\Pi_e : P_k(e) \longrightarrow H^2_0(\omega_e)$ by setting

$$(\Pi_e p_k)(s,t) := \psi_e(s,t)p_k(s), \qquad (s,t) \in \omega_e. \tag{4.12}$$

Estimates (4.6)–(4.9) easily follows from standard scaling arguments, using the fixed reference rhomb $\widehat{D}$. □

4.1. Upper bounds. We now prove that the indicator just introduced can be used as a reliable error estimator. We will prove our upper bounds for the linear element of section 3.1.1 by means of a saturation assumption involving its quadratic version. Therefore, to avoid confusion, we will denote all the quantities relative to the quadratic element described in section 3.1.2 by a tilde. For example, the approximation spaces and linking operator in (3.14)–(3.19) will be renamed as $\widetilde{\boldsymbol{\Theta}}_h$, $\widetilde{W}_h$, $\widetilde{\boldsymbol{\Gamma}}_h$, and $\widetilde{L}$, respectively. Accordingly, we define

$$\widetilde{\boldsymbol{X}}_h = \big\{(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}^*_h) = (\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h + \widetilde{L}\widetilde{\boldsymbol{\eta}}_h) : \widetilde{\boldsymbol{\eta}}_h \in \widetilde{\boldsymbol{\Theta}}_h\,,\ \widetilde{v}_h \in \widetilde{W}_h\big\}. \tag{4.13}$$

We need to make the following assumption.

Saturation assumption. Let $(\boldsymbol{\theta}_h, w^*_h, \boldsymbol{\gamma}_h) \in \boldsymbol{X}_h \times \boldsymbol{\Gamma}_h$ (resp., $(\widetilde{\boldsymbol{\theta}}_h, \widetilde{w}^*_h, \widetilde{\boldsymbol{\gamma}}_h) \in \widetilde{\boldsymbol{X}}_h \times \widetilde{\boldsymbol{\Gamma}}_h$) be the discrete solution using the linear (resp., quadratic) element. We assume that there exists $0 < \rho < 1$ such that

$$\begin{aligned} &|||(\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}_h, w - \widetilde{w}^*_h)|||_h + ||\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}}_h||_{-1} + t\,||\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}}_h||_0 \\ &\qquad \le \rho\Big(|||(\boldsymbol{\theta} - \boldsymbol{\theta}_h, w - w^*_h)|||_h + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{-1} + t\,||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_0\Big). \end{aligned} \tag{4.14} \quad \square$$

By using the saturation assumption (4.14), it is easily seen that one gets the reliability estimate

$$\begin{aligned} &|||(\boldsymbol{\theta} - \boldsymbol{\theta}_h, w - w^*_h)|||_h + ||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_{-1} + t\,||\boldsymbol{\gamma} - \boldsymbol{\gamma}_h||_0 \\ &\qquad \le C\left(\sum_{T\in\mathcal{T}_h}\Big(\eta_T^2 + h_T^2(h_T^2 + t^2)||g - g_h||^2_{0,T}\Big)\right)^{1/2}, \end{aligned} \tag{4.15}$$

provided one is able to bound

$$|||(\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h, \widetilde{w}^*_h - w^*_h)|||_h + ||\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h||_{-1} + t\,||\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h||_0. \tag{4.16}$$

To this aim, we need the next result, which states that $\boldsymbol{X}_h \subseteq \widetilde{\boldsymbol{X}}_h$, and that functions in $\widetilde{\boldsymbol{X}}_h$ can be approximated by functions in $\boldsymbol{X}_h$.

LEMMA 4.2. *It holds $\boldsymbol{X}_h \subseteq \widetilde{\boldsymbol{X}}_h$; moreover, given $(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}^*_h) \in \widetilde{\boldsymbol{X}}_h$, there exists $(\boldsymbol{\eta}_h, v^*_h) \in \boldsymbol{X}_h$ such that*

$$\begin{aligned} &\sum_{T\in\mathcal{T}_h} h_T^{-2}\left(||\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h||^2_{0,T} + \frac{1}{h_T^2 + t^2}||\widetilde{v}^*_h - v^*_h||^2_{0,T}\right) \\ &\quad + \sum_{e\in\mathcal{E}_h} h_e^{-1}\left(||\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h||^2_{0,e} + \frac{1}{h_e^2 + t^2}||\widetilde{v}^*_h - v^*_h||^2_{0,e}\right) \le C|||(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}^*_h)|||^2_h. \end{aligned} \tag{4.17}$$

Proof. First, we need to show that a generic $(\boldsymbol{\eta}_h, v_h^*) = (\boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) \in \boldsymbol{X}_h$ can be written as $(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*) = (\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h + \widetilde{L}\widetilde{\boldsymbol{\eta}}_h) \in \widetilde{\boldsymbol{X}}_h$ for suitable $\widetilde{\boldsymbol{\eta}}_h \in \widetilde{\boldsymbol{\Theta}}_h$ and $\widetilde{v}_h \in \widetilde{W}_h$. This forces $\widetilde{\boldsymbol{\eta}}_h = \boldsymbol{\eta}_h$, which is an admissible choice, since obviously $\boldsymbol{\Theta}_h \subseteq \widetilde{\boldsymbol{\Theta}}_h$. Noting that $v_h + L\boldsymbol{\eta}_h \in \widetilde{W}_h$, we set $\widetilde{v}_h = v_h + L\boldsymbol{\eta}_h$. We now observe (see (3.18)–(3.19)) that $\widetilde{L}\widetilde{\boldsymbol{\eta}}_h = \widetilde{L}\boldsymbol{\eta}_h = 0$. Indeed, given $\boldsymbol{\eta}_h \in \boldsymbol{\Theta}_h$, the equation

$$(\boldsymbol{\nabla}\widetilde{L}\boldsymbol{\eta}_h - \boldsymbol{\eta}_h) \cdot \mathbf{t} \quad \text{is linear on each } e \tag{4.18}$$

has unique solution $\widetilde{L}\boldsymbol{\eta}_h = 0$, since $\boldsymbol{\eta}_h$ is already linear on each edge e. Therefore we have

$$(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*) = (\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h + \widetilde{L}\widetilde{\boldsymbol{\eta}}_h) = (\boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h + \widetilde{L}\boldsymbol{\eta}_h) = (\boldsymbol{\eta}_h, v_h + L\boldsymbol{\eta}_h) = (\boldsymbol{\eta}_h, v_h^*),$$

which proves $\boldsymbol{X}_h \subseteq \widetilde{\boldsymbol{X}}_h$.

To prove estimate (4.17), let $(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*) = (\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h + \widetilde{\boldsymbol{\eta}}_h) \in \widetilde{\boldsymbol{X}}_h$ be given. We define (recalling that $\mathcal{I}$ denotes the Lagrange interpolating operator):

$$\boldsymbol{\eta}_h = \mathcal{I}\widetilde{\boldsymbol{\eta}}_h \in \boldsymbol{\Theta}_h, \qquad\qquad v_h = \mathcal{I}\widetilde{v}_h \in W_h. \tag{4.19}$$

Accordingly, we set

$$(\boldsymbol{\eta}_h, v_h^*) = \big(\mathcal{I}\widetilde{\boldsymbol{\eta}}_h, \mathcal{I}\widetilde{v}_h + L(\mathcal{I}\widetilde{\boldsymbol{\eta}}_h)\big) \in \boldsymbol{X}_h. \tag{4.20}$$

By standard approximation results and scaling arguments, we have

$$\sum_{T\in\mathcal{T}_h} h_T^{-2}||\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h||_{0,T}^2 + \sum_{e\in\mathcal{E}_h} h_e^{-1}||\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h||_{0,e}^2 \leq C||\widetilde{\boldsymbol{\eta}}_h||_1^2. \tag{4.21}$$

To continue, let us note that

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} \frac{h_T^{-2}}{h_T^2+t^2}||\widetilde{v}_h^* - v_h^*||_{0,T}^2 &\leq 2\sum_{T\in\mathcal{T}_h} \frac{h_T^{-2}}{h_T^2+t^2}||\widetilde{v}_h - v_h||_{0,T}^2 \\ &\quad + 2\sum_{T\in\mathcal{T}_h} \frac{h_T^{-2}}{h_T^2+t^2}||\widetilde{L}\widetilde{\boldsymbol{\eta}}_h - L\boldsymbol{\eta}_h||_{0,T}^2.\end{aligned} \tag{4.22}$$

From standard approximation theory we have

$$\begin{aligned}||\widetilde{v}_h - v_h||_{0,T}^2 &\leq Ch_T^4|\widetilde{v}_h|_{2,T}^2 = Ch_T^4|\boldsymbol{\nabla}\widetilde{v}_h|_{1,T}^2 \\ &\leq Ch_T^4\left(|\boldsymbol{\nabla}\widetilde{v}_h^* - \widetilde{\boldsymbol{\eta}}_h|_{1,T}^2 + |\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\nabla}\widetilde{L}\widetilde{\boldsymbol{\eta}}_h|_{1,T}^2\right).\end{aligned} \tag{4.23}$$

Using an inverse inequality and (3.20) we get

$$||\widetilde{v}_h - v_h||_{0,T}^2 \leq Ch_T^2||\boldsymbol{\nabla}\widetilde{v}_h^* - \widetilde{\boldsymbol{\eta}}_h||_{0,T}^2 + Ch_T^4|\widetilde{\boldsymbol{\eta}}_h|_{1,T}^2. \tag{4.24}$$

Therefore, we obtain

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} \frac{h_T^{-2}}{h_T^2+t^2}||\widetilde{v}_h - v_h||_{0,T}^2 &\leq C\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||\boldsymbol{\nabla}\widetilde{v}_h^* - \widetilde{\boldsymbol{\eta}}_h||_{0,T}^2 \\ &\quad + C\sum_{T\in\mathcal{T}_h} \frac{h_T^2}{h_T^2+t^2}|\widetilde{\boldsymbol{\eta}}_h|_{1,T}^2 \\ &\leq C\left(\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2}||\boldsymbol{\nabla}\widetilde{v}_h^* - \widetilde{\boldsymbol{\eta}}_h||_{0,T}^2 + ||\widetilde{\boldsymbol{\eta}}_h||_1^2\right).\end{aligned} \tag{4.25}$$

Furthermore, from (3.13), (3.20), and (4.19), we have

$$\begin{aligned}||\widetilde{L}\widetilde{\boldsymbol{\eta}}_h - L\boldsymbol{\eta}_h||^2_{0,T} &\le 2\left(||\widetilde{L}\widetilde{\boldsymbol{\eta}}_h||^2_{0,T} + ||L\boldsymbol{\eta}_h||^2_{0,T}\right)\\ &\le Ch_T^4\left(|\widetilde{\boldsymbol{\eta}}_h|^2_{1,T} + |\boldsymbol{\eta}_h|^2_{1,T}\right) \le Ch_T^4|\widetilde{\boldsymbol{\eta}}_h|^2_{1,T}.\end{aligned} \tag{4.26}$$

As a consequence, we get

$$\sum_{T\in\mathcal{T}_h}\frac{h_T^{-2}}{h_T^2+t^2}||\widetilde{L}\widetilde{\boldsymbol{\eta}}_h - L\boldsymbol{\eta}_h||^2_{0,T} \le C\sum_{T\in\mathcal{T}_h}\frac{h_T^2}{h_T^2+t^2}|\widetilde{\boldsymbol{\eta}}_h|^2_{1,T} \le C||\widetilde{\boldsymbol{\eta}}_h||^2_{1,T}. \tag{4.27}$$

Using (4.25) and (4.27), from (4.22) we have

$$\sum_{T\in\mathcal{T}_h}\frac{h_T^{-2}}{h_T^2+t^2}||\widetilde{v}_h^* - v_h^*||^2_{0,T} \le C|||(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*)|||^2_h. \tag{4.28}$$

The shape regularity of $\mathcal{T}_h$, scaling arguments, and estimate (4.28) show that

$$\sum_{e\in\mathcal{E}_h}\frac{h_e^{-1}}{h_e^2+t^2}||\widetilde{v}_h^* - v_h^*||^2_{0,e} \le C\sum_{T\in\mathcal{T}_h}\frac{h_T^{-2}}{h_T^2+t^2}||\widetilde{v}_h^* - v_h^*||^2_{0,T} \le C|||(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*)|||^2_h. \tag{4.29}$$

Collecting (4.21), (4.28) and (4.29), we infer estimate (4.17). □

We are now ready to prove the following proposition.

PROPOSITION 4.3. *We have*

$$\begin{aligned}&|||(\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h, \widetilde{w}_h^* - w_h^*)|||_h + ||\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h||_{-1} + t\,||\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h||_0\\ &\qquad\le C\left(\sum_{T\in\mathcal{T}_h}\left(\eta_T^2 + h_T^2(h_T^2+t^2)||g-g_h||^2_{0,T}\right)\right)^{1/2}.\end{aligned} \tag{4.30}$$

Proof. Consider $(\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h, \widetilde{w}_h^* - w_h^*; \widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h) \in \widetilde{\boldsymbol{X}}_h \times \widetilde{\boldsymbol{\Gamma}}_h$. Discrete stability for the quadratic element (see Proposition 3.2) implies that there exists $(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*; \widetilde{\boldsymbol{\tau}}_h)$ in $\widetilde{\boldsymbol{X}}_h \times \widetilde{\boldsymbol{\Gamma}}_h$ such that

$$|||(\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*)|||_h + ||\widetilde{\boldsymbol{\tau}}_h||_h \le 1 \tag{4.31}$$

and

$$\begin{aligned}&C\Big(|||(\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h, \widetilde{w}_h^* - w_h^*)|||_h + ||\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h||_h\Big)\\ &\qquad\le \Big\{a(\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h, \widetilde{\boldsymbol{\eta}}_h) + (\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h, \boldsymbol{\nabla}\widetilde{v}_h^* - \widetilde{\boldsymbol{\eta}}_h)\Big\}\\ &\qquad\quad + \Big\{-\big(\boldsymbol{\nabla}(\widetilde{w}_h^* - w_h^*) - (\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h), \widetilde{\boldsymbol{\tau}}_h\big) + \mu^{-1}t^2(\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h, \widetilde{\boldsymbol{\tau}}_h)\Big\}\\ &\qquad = \text{(I)} + \text{(II)}.\end{aligned} \tag{4.32}$$

On one hand, since $(\widetilde{\boldsymbol{\theta}}_h, \widetilde{w}_h^*; \widetilde{\boldsymbol{\gamma}}_h)$ (resp., $(\boldsymbol{\theta}_h, w_h^*; \boldsymbol{\gamma}_h)$) solves the higher-order (resp., low-order) discrete problem, we have

$$\begin{aligned}\text{(I)} &= a(\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h, \widetilde{\boldsymbol{\eta}}_h) + (\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h, \boldsymbol{\nabla}\widetilde{v}_h^* - \widetilde{\boldsymbol{\eta}}_h)\\ &= (g, \widetilde{v}_h^*) - a(\boldsymbol{\theta}_h, \widetilde{\boldsymbol{\eta}}_h) - (\boldsymbol{\gamma}_h, \boldsymbol{\nabla}\widetilde{v}_h^* - \widetilde{\boldsymbol{\eta}}_h)\\ &= (g, \widetilde{v}_h^* - v_h^*) - a(\boldsymbol{\theta}_h, \widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h) - \big(\boldsymbol{\gamma}_h, \boldsymbol{\nabla}(\widetilde{v}_h^* - v_h^*) - (\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h)\big),\end{aligned} \tag{4.33}$$

where we choose $(\boldsymbol{\eta}_h, v_h^*) \in \boldsymbol{X}_h$ satisfying estimate (4.17). An elementwise integration by parts gives

$$\text{(I)} = \sum_{T\in\mathcal{T}_h} \left\{ \int_T \big(\operatorname{div} \mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h\big)\cdot(\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h) - \int_{\partial T} \mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}\cdot(\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h) \right\}$$
$$+ \sum_{T\in\mathcal{T}_h} \left\{ \int_T \big(\operatorname{div}\boldsymbol{\gamma}_h + g\big)\,(\widetilde{v}_h^* - v_h^*) - \int_{\partial T} \boldsymbol{\gamma}_h\cdot\mathbf{n}\,(\widetilde{v}_h^* - v_h^*) \right\}, \tag{4.34}$$

by which

$$\text{(I)} = \sum_{T\in\mathcal{T}_h} \int_T \big(\operatorname{div} \mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h\big)\cdot(\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h) - \sum_{e\in\mathcal{E}_h} \int_e [\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!]\cdot(\widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h)$$
$$+ \sum_{T\in\mathcal{T}_h} \int_T \big(\operatorname{div}\boldsymbol{\gamma}_h + g\big)\,(\widetilde{v}_h^* - v_h^*) - \sum_{e\in\mathcal{E}_h} \int_e [\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!]\,(\widetilde{v}_h^* - v_h^*). \tag{4.35}$$

Hence, it holds

$$\text{(I)} \le C \Bigg(\Big(\sum_{T\in\mathcal{T}_h} h_T^2 || \operatorname{div} \mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h ||_{0,T}^2 \Big)^{1/2} \Big(\sum_{T\in\mathcal{T}_h} h_T^{-2} || \widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h ||_{0,T}^2 \Big)^{1/2}$$
$$+ \Big(\sum_{e\in\mathcal{E}_h} h_e || [\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!] ||_{0,e}^2 \Big)^{1/2} \Big(\sum_{e\in\mathcal{E}_h} h_e^{-1} || \widetilde{\boldsymbol{\eta}}_h - \boldsymbol{\eta}_h ||_{0,e}^2 \Big)^{1/2}$$
$$+ \Big(\sum_{T\in\mathcal{T}_h} h_T^2 (h_T^2 + t^2) || \operatorname{div}\boldsymbol{\gamma}_h + g ||_{0,T}^2 \Big)^{1/2} \Big(\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2 (h_T^2 + t^2)} || \widetilde{v}_h^* - v_h^* ||_{0,T}^2 \Big)^{1/2}$$
$$+ \Big(\sum_{e\in\mathcal{E}_h} h_e (h_e^2 + t^2) || [\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!] ||_{0,e}^2 \Big)^{1/2} \Big(\sum_{e\in\mathcal{E}_h} \frac{1}{h_e (h_e^2 + t^2)} || \widetilde{v}_h^* - v_h^* ||_{0,e}^2 \Big)^{1/2} \Bigg). \tag{4.36}$$

Using Lemma 4.2, we get

$$\text{(I)} \le C \Bigg(\Big(\sum_{T\in\mathcal{T}_h} h_T^2 || \operatorname{div} \mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h ||_{0,T}^2 \Big)^{1/2} + \Big(\sum_{e\in\mathcal{E}_h} h_e || [\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!] ||_{0,e}^2 \Big)^{1/2}$$
$$+ \Big(\sum_{T\in\mathcal{T}_h} h_T^2 (h_T^2 + t^2) || \operatorname{div}\boldsymbol{\gamma}_h + g ||_{0,T}^2 \Big)^{1/2} + \Big(\sum_{e\in\mathcal{E}_h} h_e (h_e^2 + t^2) || [\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!] ||_{0,e}^2 \Big)^{1/2} \Bigg)$$
$$\times |\!|\!| (\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*) |\!|\!|_h. \tag{4.37}$$

Therefore, one has

(4.38)

$$\begin{aligned}(\mathrm{I}) \leq C \Bigg(&\left(\sum_{T\in\mathcal{T}_h} h_T^2 || \operatorname{div} \mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h ||_{0,T}^2 \right)^{1/2} + \left(\sum_{e\in\mathcal{E}_h} h_e || [\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!] ||_{0,e}^2 \right)^{1/2} \\ &+ \left(\sum_{T\in\mathcal{T}_h} h_T^2(h_T^2+t^2) || \operatorname{div} \boldsymbol{\gamma}_h + g_h ||_{0,T}^2 \right)^{1/2} + \left(\sum_{T\in\mathcal{T}_h} h_T^2(h_T^2+t^2) || g - g_h ||_{0,T}^2 \right)^{1/2} \\ &+ \left(\sum_{e\in\mathcal{E}_h} h_e(h_e^2+t^2) || [\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!] ||_{0,e}^2 \right)^{1/2} \Bigg) |\!|\!| (\widetilde{\boldsymbol{\eta}}_h, \widetilde{v}_h^*) |\!|\!|_h.\end{aligned}$$

On the other hand, since $(\widetilde{\boldsymbol{\theta}}_h, \widetilde{w}_h^*; \widetilde{\boldsymbol{\gamma}}_h)$ solves the higher-order discrete problem, we have

(4.39)

$$\begin{aligned}(\mathrm{II}) &= -\big(\boldsymbol{\nabla}(\widetilde{w}_h^* - w_h^*) - (\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h), \widetilde{\boldsymbol{\tau}}_h\big) + \mu^{-1}t^2(\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h, \widetilde{\boldsymbol{\tau}}_h) \\ &= -\big(\mu^{-1}t^2\,\boldsymbol{\gamma}_h - (\boldsymbol{\nabla} w_h^* - \boldsymbol{\theta}_h), \widetilde{\boldsymbol{\tau}}_h\big) \\ &\leq \left(\sum_{T\in\mathcal{T}_h} \frac{1}{h_T^2+t^2} ||\mu^{-1}t^2\,\boldsymbol{\gamma}_h - (\boldsymbol{\nabla} w_h^* - \boldsymbol{\theta}_h)||_{0,T}^2 \right)^{1/2} \left(\sum_{T\in\mathcal{T}_h} (h_T^2+t^2) ||\widetilde{\boldsymbol{\tau}}_h||_{0,T}^2 \right)^{1/2}.\end{aligned}$$

As a consequence, from (4.32), (4.38), (4.39), using (4.31) and recalling definitions (4.1)–(4.3), we have

(4.40)

$$|\!|\!|(\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h, \widetilde{w}_h^* - w_h^*)|\!|\!|_h + ||\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h||_h \leq C \left(\sum_{T\in\mathcal{T}_h} \Big(\eta_T^2 + h_T^2(h_T^2+t^2)||g-g_h||_{0,T}^2 \Big) \right)^{1/2}.$$

The same arguments as in (3.65)–(3.69), applied to $\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h$, give

$$||\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h||_{-1} \leq C \left(||\widetilde{\boldsymbol{\gamma}}_h - \boldsymbol{\gamma}_h||_h + ||\widetilde{\boldsymbol{\theta}}_h - \boldsymbol{\theta}_h||_1 \right). \tag{4.41}$$

Combining (4.40) and (4.41) we infer estimate (4.30). The proof is complete. □

4.2. Lower bounds. We now prove the efficiency of our error estimator by establishing the following proposition.

PROPOSITION 4.4. *Let* $(\boldsymbol{\theta}, w; \boldsymbol{\gamma})$ *(resp.,* $(\boldsymbol{\theta}_h, w_h^*; \boldsymbol{\gamma}_h)$*) be the solution of the continuous (resp., discrete) problem. Given* $T \in \mathcal{T}_h$*, it holds*

(4.42)

$$\begin{aligned}\eta_T \leq C \Bigg(&\frac{1}{(h_T^2+t^2)^{1/2}} \big|\big| \boldsymbol{\nabla}(w_h^* - w) - (\boldsymbol{\theta}_h - \boldsymbol{\theta}) \big|\big|_{0,T} + ||\boldsymbol{\theta}_h - \boldsymbol{\theta}||_{1,\omega_T} \\ &+ ||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}||_{-1,\omega_T} + t\,||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}||_{0,\omega_T} + \left(\sum_{T'\subset\omega_T} h_{T'}^2(h_{T'}^2+t^2)||g-g_h||_{0,T'}^2 \right)^{1/2} \Bigg),\end{aligned}$$

where η_T *is defined by* (4.1)*–*(4.3)*.*

Proof. Fix $T \in \mathcal{T}_h$ and a generic edge $e \subset \partial T$. We proceed in three steps.

First step. Since

$$\mu^{-1}t^2\boldsymbol{\gamma} = \boldsymbol{\nabla} w - \boldsymbol{\theta}, \tag{4.43}$$

we get

$$\begin{aligned}
&\frac{1}{(h_T^2+t^2)^{1/2}}||\mu^{-1}t^2\,\boldsymbol{\gamma}_h - (\boldsymbol{\nabla} w_h^* - \boldsymbol{\theta}_h)||_{0,T}\\
&= \frac{1}{(h_T^2+t^2)^{1/2}}\big|\big|\mu^{-1}t^2\,(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}) - \big(\boldsymbol{\nabla}(w_h^*-w) - (\boldsymbol{\theta}_h-\boldsymbol{\theta})\big)\big|\big|_{0,T}\\
&\le C\left(t\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,T} + \frac{1}{(h_T^2+t^2)^{1/2}}\big|\big|\boldsymbol{\nabla}(w_h^*-w) - (\boldsymbol{\theta}_h-\boldsymbol{\theta})\big|\big|_{0,T}\right).
\end{aligned} \tag{4.44}$$

Second step. We choose

$$\boldsymbol{\eta}_T = h_T^2(\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h)\,b_T, \tag{4.45}$$

where b_T is the standard cubic bubble on T. We observe that

$$|\boldsymbol{\eta}_T|_{1,T} \le Ch_T||\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h||_{0,T}. \tag{4.46}$$

Taking advantage of the equilibrium equation

$$-\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}) - \boldsymbol{\gamma} = \mathbf{0}, \tag{4.47}$$

we get

$$\begin{aligned}
&h_T^2||\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h||_{0,T}^2\\
&\quad\le C\big(\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h, \boldsymbol{\eta}_T\big) = C\big(\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h-\boldsymbol{\theta}) + (\boldsymbol{\gamma}_h-\boldsymbol{\gamma}), \boldsymbol{\eta}_T\big)\\
&\quad= C\,(-a(\boldsymbol{\theta}_h-\boldsymbol{\theta},\boldsymbol{\eta}_T) + (\boldsymbol{\gamma}_h-\boldsymbol{\gamma},\boldsymbol{\eta}_T))\\
&\quad\le C\,(||\boldsymbol{\theta}_h-\boldsymbol{\theta}||_{1,T} + ||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,T})\;|\boldsymbol{\eta}_T|_{1,T}.
\end{aligned} \tag{4.48}$$

Using (4.46), from (4.48) we thus obtain

$$h_T||\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h) + \boldsymbol{\gamma}_h||_{0,T} \le C\,(||\boldsymbol{\theta}_h-\boldsymbol{\theta}||_{1,T} + ||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,T})\,. \tag{4.49}$$

Next, we choose

$$\boldsymbol{\eta}_e = h_e P([\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!])\,b_e, \tag{4.50}$$

where P is the prolongation operator introduced in [24] and b_e is the usual edge bubble on e. We observe that it holds

$$\left(\sum_{T\subset\omega_e} h_T^{-2}||\boldsymbol{\eta}_e||_{0,T}^2\right)^{1/2} \le C|\boldsymbol{\eta}_e|_{1,\omega_e} \le Ch_e^{1/2}||\,[\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!]\,||_{0,e}. \tag{4.51}$$

Integrating by parts and using again the equilibrium equation (4.47), we have

$$
\begin{aligned}
&h_e|| \,[\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!]\, ||_{0,e}^2 \\
&\quad \le C\int_e [\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!]\,\cdot\boldsymbol{\eta}_e = C\left(\int_{\omega_e}\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\cdot\boldsymbol{\eta}_e + \int_{\omega_e}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h):\varepsilon(\boldsymbol{\eta}_e)\right) \\
&\quad = C\Big(\big(\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)+\boldsymbol{\gamma}_h,\boldsymbol{\eta}_e\big) + a(\boldsymbol{\theta}_h-\boldsymbol{\theta},\boldsymbol{\eta}_e) - (\boldsymbol{\gamma}_h-\boldsymbol{\gamma},\boldsymbol{\eta}_e)\Big) \\
&\quad \le C\left(\left(\sum_{T\subset\omega_e} h_T^2||\operatorname{div}\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)+\boldsymbol{\gamma}_h||_{0,T}^2\right)^{1/2}\left(\sum_{T\subset\omega_e} h_T^{-2}||\boldsymbol{\eta}_e||_{0,T}^2\right)^{1/2}\right. \\
&\qquad \left. +\Big(||\boldsymbol{\theta}_h-\boldsymbol{\theta}||_{1,\omega_e} + ||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,\omega_e}\Big)\,|\boldsymbol{\eta}_e|_{1,\omega_e}\right).
\end{aligned}
\tag{4.52}
$$

Therefore, using (4.51) and (4.49), from (4.52) we get

$$
h_e^{1/2}||\,[\![\mathbf{C}\,\varepsilon(\boldsymbol{\theta}_h)\mathbf{n}]\!]\,||_{0,e} \le C\,(||\boldsymbol{\theta}_h-\boldsymbol{\theta}||_{1,\omega_e} + ||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,\omega_e})\,. \tag{4.53}
$$

Third step. We first define

$$
\varphi_T = (\operatorname{div}\boldsymbol{\gamma}_h + g_h)\,b_T^2. \tag{4.54}
$$

We observe that $\varphi_T \in H_0^2(T)$ and one has

$$
\begin{aligned}
|\varphi_T|_{1,T} &\le Ch_T^{-1}||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T}, \\
|\boldsymbol{\nabla}\varphi_T|_{1,T} &\le Ch_T^{-2}||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T}.
\end{aligned}
\tag{4.55}
$$

We then set

$$
v_T = h_T^2(h_T^2+t^2)\,\varphi_T. \tag{4.56}
$$

Using the equilibrium equation

$$
-\operatorname{div}\boldsymbol{\gamma} = g, \tag{4.57}
$$

we get

$$
\begin{aligned}
h_T^2(h_T^2+t^2)||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T}^2 &\le C\big(\operatorname{div}\boldsymbol{\gamma}_h + g_h, v_T\big) \\
&= C\Big(\big(\operatorname{div}(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}), v_T\big) + (g_h - g, v_T)\Big).
\end{aligned}
\tag{4.58}
$$

We now separately treat the two terms at the right-hand side of (4.58). Integrating by parts, recalling (4.54) and (4.56), and using (4.55), we have

$$
\begin{aligned}
&\big(\operatorname{div}(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}), v_T\big) = -(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}, \boldsymbol{\nabla}v_T) \\
&= -h_T^4(\boldsymbol{\gamma}_h-\boldsymbol{\gamma},\boldsymbol{\nabla}\varphi_T) - t^2h_T^2(\boldsymbol{\gamma}_h-\boldsymbol{\gamma},\boldsymbol{\nabla}\varphi_T) \\
&\le ||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,T}h_T^4|\boldsymbol{\nabla}\varphi_T|_{1,T} + t\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,T}h_T^2t\,||\boldsymbol{\nabla}\varphi_T||_{0,T} \\
&\le C\Big(||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,T}h_T^2||\operatorname{div}\boldsymbol{\gamma}_h+g_h||_{0,T} + t\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,T}h_Tt\,||\operatorname{div}\boldsymbol{\gamma}_h+g_h||_{0,T}\Big) \\
&\le C\Big(||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,T} + t\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,T}\Big)\,h_T(h_T^2+t^2)^{1/2}||\operatorname{div}\boldsymbol{\gamma}_h+g_h||_{0,T}.
\end{aligned}
\tag{4.59}
$$

Furthermore, it holds

$$
\begin{aligned}
(g_h - g, v_T) &\le h_T(h_T^2 + t^2)^{1/2}||g_h - g||_{0,T}\, h_T(h_T^2 + t^2)^{1/2}||\varphi_T||_{0,T} \\
&\le C h_T(h_T^2 + t^2)^{1/2}||g_h - g||_{0,T}\, h_T(h_T^2 + t^2)^{1/2}||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T}.
\end{aligned} \tag{4.60}
$$

Therefore, using (4.59) and (4.60), from (4.58) we infer

$$
\begin{aligned}
h_T(h_T^2 + t^2)^{1/2}||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T} \le C\Big(&||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}||_{-1,T} \\
&+ t\,||\boldsymbol{\gamma}_h - \boldsymbol{\gamma}||_{0,T} + h_T(h_T^2 + t^2)^{1/2}||g_h - g||_{0,T}\Big).
\end{aligned} \tag{4.61}
$$

Next, we define

$$
\varphi_e = \Pi_e([\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!]), \tag{4.62}
$$

where Π_e is the linear operator of Lemma 4.1. Therefore, we have

$$
||\,[\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!]\,||_{0,e}^2 \le C\int_e [\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!]\,\varphi_e, \tag{4.63}
$$

$$
||\varphi_e||_{0,\omega_e} \le C h_e^{1/2}||\,[\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!]\,||_{0,e}, \tag{4.64}
$$

$$
||\boldsymbol{\nabla}\varphi_e||_{0,\omega_e} \le C h_e^{-1/2}||\,[\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!]\,||_{0,e}, \tag{4.65}
$$

$$
|\boldsymbol{\nabla}\varphi_e|_{1,\omega_e} \le C h_e^{-3/2}||\,[\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!]\,||_{0,e}. \tag{4.66}
$$

We then set

$$
v_e = h_e(h_e^2 + t^2)\,\varphi_e. \tag{4.67}
$$

Integrating by parts using (4.63) and the equilibrium equation (4.57), we get

$$
\begin{aligned}
h_e(h_e^2 + t^2)||\,[\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!]\,||_{0,e}^2 &\le C\int_e [\![\boldsymbol{\gamma}_h \cdot \mathbf{n}]\!]\,v_e \\
&\le C\left(\int_{\omega_e} v_e \operatorname{div}\boldsymbol{\gamma}_h + \int_{\omega_e}\boldsymbol{\gamma}_h \cdot \boldsymbol{\nabla} v_e\right) \\
&= C\Big((\operatorname{div}\boldsymbol{\gamma}_h + g, v_e) + (\boldsymbol{\gamma}_h - \boldsymbol{\gamma}, \boldsymbol{\nabla} v_e)\Big) \\
&= C\Big((\operatorname{div}\boldsymbol{\gamma}_h + g_h, v_e) + (g - g_h, v_e) + (\boldsymbol{\gamma}_h - \boldsymbol{\gamma}, \boldsymbol{\nabla} v_e)\Big).
\end{aligned} \tag{4.68}
$$

We now estimate the three terms above. Recalling (4.67) and using (4.64), we obtain

(4.69)

$$\begin{aligned}(\operatorname{div}\boldsymbol{\gamma}_h + g_h, v_e) &= h_e(h_e^2+t^2)\big(\operatorname{div}\boldsymbol{\gamma}_h + g_h, \varphi_e\big)\\
&= \sum_{T\subset\omega_e}\int_T \left(h_e(h_e^2+t^2)^{1/2}(\operatorname{div}\boldsymbol{\gamma}_h + g_h)\right)\left((h_e^2+t^2)^{1/2}\varphi_e\right)\\
&\le \left(\sum_{T\subset\omega_e} h_e^2(h_e^2+t^2)||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T}^2\right)^{1/2}\left(\sum_{T\subset\omega_e}(h_e^2+t^2)||\varphi_e||_{0,T}^2\right)^{1/2}\\
&\le \left(\sum_{T\subset\omega_e} h_T^2(h_T^2+t^2)||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T}^2\right)^{1/2}\left(\sum_{T\subset\omega_e}(h_e^2+t^2)||\varphi_e||_{0,T}^2\right)^{1/2}\\
&\le C\left(\sum_{T\subset\omega_e} h_T^2(h_T^2+t^2)||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T}^2\right)^{1/2} h_e^{1/2}(h_e^2+t^2)^{1/2}||\,[\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!]\,||_{0,e}.\end{aligned}$$

The same argument shows that it holds

(4.70)

$$(g-g_h, v_e) \le C\left(\sum_{T\subset\omega_e} h_T^2(h_T^2+t^2)||g-g_h||_{0,T}^2\right)^{1/2} h_e^{1/2}(h_e^2+t^2)^{1/2}||\,[\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!]\,||_{0,e}.$$

We now notice that

$$(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}, \boldsymbol{\nabla} v_e) = h_e^3(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}, \boldsymbol{\nabla}\varphi_e) + h_e t^2(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}, \boldsymbol{\nabla}\varphi_e). \tag{4.71}$$

On one hand, using (4.66), we have

$$\begin{aligned}h_e^3(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}, \boldsymbol{\nabla}\varphi_e) &\le ||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,\omega_e} h_e^3|\boldsymbol{\nabla}\varphi_e|_{1,\omega_e}\\
&\le C||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,\omega_e} h_e^{3/2}||\,[\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!]\,||_{0,e}.\end{aligned} \tag{4.72}$$

On the other hand, from (4.65) we get

$$\begin{aligned}h_e t^2(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}, \boldsymbol{\nabla}\varphi_e) &\le t\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,\omega_e} h_e t\,||\boldsymbol{\nabla}\varphi_e||_{0,\omega_e}\\
&\le Ct\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,\omega_e} h_e^{1/2} t\,||\,[\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!]\,||_{0,e}.\end{aligned} \tag{4.73}$$

Therefore, using (4.72) and (4.73) from (4.71) we obtain

$$(\boldsymbol{\gamma}_h-\boldsymbol{\gamma}, \boldsymbol{\nabla} v_e) \le C\big(||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,\omega_e} + t\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,\omega_e}\big) h_e^{1/2}(h_e^2+t^2)^{1/2}||\,[\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!]\,||_{0,e}. \tag{4.74}$$

Collecting (4.69), (4.70) and (4.74), we infer from (4.68) that

(4.75)

$$\begin{aligned}h_e^{1/2}(h_e^2+t^2)^{1/2}||\,[\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!]\,||_{0,e} \le C\Bigg(&\left(\sum_{T\subset\omega_e} h_T^2(h_T^2+t^2)||\operatorname{div}\boldsymbol{\gamma}_h + g_h||_{0,T}^2\right)^{1/2}\\
&+ ||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,\omega_e} + t\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,\omega_e}\\
&+\left(\sum_{T\subset\omega_e} h_T^2(h_T^2+t^2)||g-g_h||_{0,T}^2\right)^{1/2}\Bigg).\end{aligned}$$

Hence, from (4.61) we get

(4.76)

$$h_e^{1/2}(h_e^2+t^2)^{1/2}||\,[\![\boldsymbol{\gamma}_h\cdot\mathbf{n}]\!]\,||_{0,e} \leq C\Bigg(||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{-1,\omega_e}+t\,||\boldsymbol{\gamma}_h-\boldsymbol{\gamma}||_{0,\omega_e} + \Bigg(\sum_{T\subset\omega_e} h_T^2(h_T^2+t^2)||g-g_h||_{0,T}^2\Bigg)^{1/2}\Bigg).$$

Estimate (4.42) now follows from (4.44), (4.49), (4.53), (4.61), and (4.76). □

Acknowledgments. The authors are grateful to L. Beirão da Veiga (University of Milan) for several suggestions about the manuscript and to G. Sangalli (IMATI–CNR, Pavia) for the useful discussions regarding Lemma 4.1.

REFERENCES

[1] D. N. Arnold and R. S. Falk, *A uniformly accurate finite element method for the Reissner-Mindlin plate*, SIAM J. Numer. Anal., 26 (1989), pp. 1276–1290.

[2] F. Auricchio and C. Lovadina, *Partial selective reduced integration schemes and kinematically linked interpolations for plate bending problems*, Math. Model Methods Appl. Sci., 9 (1999), pp. 693–722.

[3] F. Auricchio and C. Lovadina, *Analysis of kinematic linked interpolation methods for Reissner-Mindlin plate problems*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2465–2482.

[4] F. Brezzi, K. J. Bathe, and M. Fortin, *Mixed-interpolated elements for Reissner-Mindlin plates*, Internat. J. Numer. Methods Engrg., 28 (1989), pp. 1787–1801.

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

[6] F. Brezzi, M. Fortin, and R. Stenberg, *Error analysis of mixed-interpolated elements for Reissner-Mindlin plates*, Math. Models Methods Appl. Sci., 1 (1991), pp. 125–151.

[7] C. Carstensen, *Residual-based a posteriori error estimate for a nonconforming Reissner–Mindlin plate finite element*, SIAM J. Numer. Anal., 39 (2002), pp. 2034–2044.

[8] C. Carstensen and J. Schöberl, *Residual-Based a Posteriori Error Estimate for a Mixed Reissner–Mindlin Plate Finite Element*, preprint.

[9] D. Chapelle and R. Stenberg, *An optimal low-order locking-free finite element method for Reissner-Mindlin plates*, Math. Models Methods Appl. Sci., 8 (1998), pp. 407–430.

[10] D. Chapelle and R. Stenberg *Stabilized finite element formulations for shells in a bending dominated state*, SIAM J. Numer. Anal., 36 (1999), pp. 32–73.

[11] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[12] R. Duran and E. Liberman, *On mixed finite-element methods for the Reissner-Mindlin plate model*, Math. Comp., 58 (1992), pp. 561–573.

[13] R. S. Falk and T. Tu, *Locking-free finite elements for the Reissner-Mindlin plate*, Math. Comp., 69 (2000), pp. 911–928.

[14] E. Liberman, *A posteriori error estimator for a mixed finite element method for Reissner-Mindlin plate*, Math. Comp., 70 (2000), pp. 1383–1396.

[15] C. Lovadina, *A new class of mixed finite element methods for Reissner-Mindlin plates*, SIAM J. Numer. Anal., 33 (1996), pp. 2457–2467.

[16] C. Lovadina, *Analysis of a mixed finite element method for the Reissner-Mindlin plate problems*, Comput. Methods Appl. Mech. Engrg., 163 (1998), pp. 71–85.

[17] M. Lyly, *On the connection between some linear triangular Reissner-Mindlin plate bending elements*, Numer. Math., 85 (2000), pp. 77–107.

[18] M. Lyly and R. Stenberg, *Stabilized Finite Element Methods for Reissner-Mindlin Plates*, Forschungsbericht 4, Universität Innsbruck, Institut für Mathematik und Geometrie, (1999).

[19] J. Pitkäranta, *Boundary subspaces for the finite element method with Lagrange multipliers*, Numer. Math., 33 (1979), pp. 273–289.

[20] R. Stenberg, *A new finite element formulation for the plate bending problem*, in Asymptotic Methods for Elastic Structures, eds. P. G. Ciarlet, L. Trabucho, and J. Viaño, Walter de Gruyter & Co., Berlin, 1995.

[21] R. L. Taylor and F. Auricchio, *Linked interpolation for Reissner-Mindlin plate elements: Part II—A simple triangle*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 3057–3066.

[22] A. Tessler and T. J. R. Hughes, *A three-node Mindlin plate element with improved transverse shear*, Comput. Methods Appl. Mech. Engrg., 50 (1985), pp. 71–101.

[23] R. Verfürth, *Error estimates for a finite element approximation of the Stokes problem*, RAIRO Anal. Numer., 18 (1984), pp. 175–182.

[24] R. Verfürth, *A posteriori eror estimation and adaptive mesh-refinement techniques*, J. Comput. Appl. Math., 50 (1994), pp. 67–83.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2251–2271

COMPRESSION TECHNIQUES FOR BOUNDARY INTEGRAL EQUATIONS—ASYMPTOTICALLY OPTIMAL COMPLEXITY ESTIMATES[*]

WOLFGANG DAHMEN[†], HELMUT HARBRECHT[‡], AND REINHOLD SCHNEIDER[‡]

Abstract. Matrix compression techniques in the context of wavelet Galerkin schemes for boundary integral equations are developed and analyzed that exhibit optimal complexity in the following sense. The fully discrete scheme produces approximate solutions within discretization error accuracy offered by the underlying Galerkin method at a computational expense that is proven to stay proportional to the number of unknowns. Key issues are the second compression, which reduces the near field complexity significantly, and an additional a posteriori compression. The latter is based on a general result concerning an optimal work balance that applies, in particular, to the quadrature used to compute the compressed stiffness matrix with sufficient accuracy in linear time.

Key words. wavelets, norm equivalences, multilevel preconditioning, first and second compression, a posteriori compression, asymptotic complexity estimates

AMS subject classifications. 47A20, 65F10, 65F50, 65N38, 65R20

DOI. 10.1137/S0036142903428852

1. Introduction.

Many mathematical models concerning, e.g., field calculations, flow simulation, elasticity, or visualization are based on operator equations with *global operators*, especially *boundary integral* operators. Discretizing such problems will then lead in general to possibly very large linear systems with *densely populated* matrices. Moreover, the involved operator may have an order different from zero, which means that it acts on different length scales in a different way. This is well known to entail the linear systems to become more and more ill-conditioned when the level of resolution increases. Both features pose serious obstructions to the efficient numerical treatment of such problems to an extent that desirable realistic simulations are still beyond the current computing capacities.

This fact has stimulated enormous efforts to overcome these obstructions. The resulting significant progress made over the past 10 or 15 years manifests itself in several different approaches such as panel clustering (PC) [18], multipole expansions (FMM) [16], and wavelet compression (WC) [2]. Each of these methodologies has its recognized advantages and drawbacks whose balance may depend on the problem at hand. The first two (PC, FMM) are quite similar in spirit and exploit perhaps in the best way the (typical) smoothness of the potential kernel in the space rather than the integral kernel on the boundary manifold. As a consequence, they are fairly robust with regard to the shape and complexity of the boundary manifold. This is also the case for further developments like hierarchical matrices ($\mathcal{H}$-matrices) [17] and adaptive cross approximation (ACA) [1] based on a low rank approximation in the far

[*] Received by the editors May 28, 2003; accepted for publication (in revised form) March 29, 2005; published electronically January 6, 2006. This work was supported in part by the European Community's Human Potential Programme under contract HPRN-CT-202-00286 (BREAKING COMPLEXITY), and the SFB 393 "Numerical Simulation on Massive Parallel Computers" funded by the Deutsche Forschungsgemeinschaft.

http://www.siam.org/journals/sinum/43-6/42885.html

[†] RTWH Aachen, Institut für Geometrie und Praktische Mathematik, Aachen 52056, Germany (dahmen@igpm.rwth-aachen.de).

[‡] Christian-Albrechts-Universität zu Kiel, Institut für Informatik und Praktische Mathematik, Olshausenstr, 40, Kiel 24098, Germany (hh@numerik.uni-kiel.de, rs@numerik.uni-kiel.de).

field; see also [30]. Common experience seems to indicate that the third option (WC) depends in a more sensitive way on the underlying geometry, and its performance may suffer from strong domain anisotropies. On the other hand, WC allows one in a natural way to incorporate preconditioning techniques, which very much supports the fast solution of the resulting sparsified systems. Moreover, recent developments suggest a natural combination with adaptive discretizations to keep from the start, for a given target accuracy, the size of the linear systems as small as possible. Perhaps the main difference between PC, FMM on the one hand and WC on the other is that the former are essentially *agglomeration* techniques, while WC is more apt for refining a given coarse discretization. Since these methodologies are, in that sense, somewhat complementary in spirit, it is in our opinion less appropriate to contrapose them, but one should rather try to extract the best from each option.

As indicated before, a preference for any of the above-mentioned solution strategies will, in general, depend on the concrete application at hand. The objective of this paper is therefore to provide a complete analysis of the wavelet approach (WC) from the following perspectives. Recall that WC has been essentially initiated by the pioneering paper [2], where it was observed that certain operators have an almost sparse representation in wavelet coordinates. Discarding all entries below a certain threshold in a given principal section of the wavelet representation will then give rise to a sparse matrix that can be further processed by efficient linear algebra tools. This idea has since then initiated many subsequent studies. The diversity as well as the partly deceiving nature of the by now existing rich body of literature is one reason for us to take up this subject here again. Our attempt to provide a unified analysis is certainly based, to some extent, on previously used techniques. However, on the one hand we have extended these concepts essentially by several new analytical tools, to be detailed later below. On the other hand, the numerical implementation has been brought now to a state that makes the method applicable to practically relevant problems. The main objective of this paper is to present these new theoretical and practical developments centering on the following issues.

When dealing with large-scale problems, a sharp *asymptotic analysis* of the complexity is in our opinion ultimately essential for assessing its potential. It is important to clarify the meaning of “complexity” in this context. It is always understood as the *work/accuracy* rate of the method under consideration when the level of resolution increases, i.e., the overall accuracy of the computed approximate solution is to be tied to the computational work required to obtain it. There is no point in increasing the number of degrees of freedom, i.e., the size of the linear systems, without improving the accuracy of the resulting solutions. On the other hand, since one is ultimately interested in the “exact solution of the infinite-dimensional problem” it makes no sense to determine the solution to the discrete finite-dimensional problem with much more accuracy than that offered by the discretization. Thus, a reasonable target accuracy for the solutions of the discrete systems is *discretization error accuracy* which will guide all our subsequent considerations. A method will therefore be said to exhibit *asymptotically optimal complexity* if the discrete solution can be computed, for any discretization level, within discretization error accuracy at a computational expense that stays *proportional* to the number N of unknowns, i.e., when not even logarithmic factors are permitted. Obviously, computational optimality refers here to the given discretization framework based on the chosen hierarchy of trial spaces.

In the present context, this means that the solutions of the compressed systems should exhibit the same asymptotic convergence rates as the solutions to unperturbed discretizations. In connection with WC, this in turn means that any threshold parameters and truncation strategies have to be adapted to the current number N of degrees of freedom. Such an asymptotic analysis is missing in [2] and in many subsequent investigations. The program carried out in [11, 12, 24, 25, 27] aimed at determining exactly such work/accuracy rates for various types of boundary integral equations. Roughly speaking, it could be shown that discretization error accuracy can be realized for appropriately chosen wavelet bases at a computational work that stays bounded by $C\,N(\log N)^a$ for some constants C, a independent of the size N of the linear systems. Moreover, in [27] it was shown for the first time that, by incorporating a second compression, an overall optimal compression strategy can be devised that even avoids additional logarithmic factors, while the complexity estimates for a corresponding adaptive quadrature scheme were confined to collocation methods.

The purpose of the present paper can now be summarized as follows.

We present a complete, in several respects new and improved analysis of wavelet compression schemes for boundary integral equations based on Galerkin discretizations that exhibit overall asymptotically optimal complexity. This means that discretization error accuracy is obtained at a computational expense that stays proportional to the size N of the arising linear systems, uniformly in N. In contrast to the earlier treatments of the authors it is based on bilinear forms and Strang's lemma; see also [25].

The analysis significantly simplifies previous studies including the effect of the second compression. In fact, the analysis of the second compression is completely new, resulting in slightly different conclusions; see section 6. The complete work balance synchronizing quadrature and compression accuracy is based on new *balance estimates* given in section 11; see Theorem 11.1. In particular, it reveals the right work balance for the compression and the quadrature needed to compute the compressed matrices with sufficient accuracy, so as to realize asymptotically optimal computational complexity of the fully discretized scheme. Specifically, the computational work for computing and assembling the compressed stiffness matrix remains proportional to the number N of degrees of freedom.

This also lays the foundation for an additional new *a posteriori compression* whose analysis is again based on the above-mentioned balance estimate. This improves the quantitative performance of the scheme significantly as shown in [20, 22].

Our analysis concerns what is called the *standard* wavelet representation. A preference for using the so-called *nonstandard* form is frequently reported in the literature. The reason is that the entries in this latter form only involve scaling functions and wavelets on the *same* level. This indeed simplifies assembling the matrices and offers essential advantages when dealing with shift-invariant problems. However, aside from the problem of preconditioning in connection with operators of nonzero order, to our knowledge it has so far *not* been shown that, for a fixed order of vanishing moments, optimal computational complexity in the above sense can be obtained with the nonstandard form. In fact, for regular solutions approximate solutions with prescribed accuracy can be obtained at (asymptotically) a lower computational cost with the aid of the standard form when compared with the nonstandard form. This is backed by theory and confirmed by numerical experience; see [20, 22].

As mentioned before, it is important to employ the "right" wavelet bases. This question has been discussed extensively in previous work [3, 6, 13, 14]. The theory tells us that, depending on the order of the operator, a proper relation between the approximation order of the underlying multiresolution spaces and the order of vanishing moments matters, which often rules out orthonormal wavelets. Given the validity of this relation it is important to keep supports as small as possible; see Remark 7.1 in section 7. Moreover, our present analysis of the second compression refers exclusively to biorthogonal spline wavelets whose singular supports are well defined; see [21] for examples and graphical illustrations.

We shall frequently write $a \lesssim b$ to express that a is bounded by a constant multiple of b, uniformly with respect to all parameters on which a and b may depend. Then $a \sim b$ means $a \lesssim b$ and $b \lesssim a$.

2. Problem formulation and preliminaries. We consider boundary integral equations on a closed boundary surface Γ of an $(n+1)$-dimensional domain $\Omega \subset \mathbb{R}^{n+1}$

$$Au(\widehat{x}) = \int_\Gamma k(\widehat{x},\widehat{y})u(\widehat{y})d\Gamma_{\widehat{y}} = f(\widehat{x}), \quad \widehat{x} \in \Gamma, \tag{2.1}$$

where the boundary integral operator is assumed to be an operator of order $2q$, that is, $A : H^q(\Gamma) \to H^{-q}(\Gamma)$. The kernel functions under consideration are supposed to be smooth as functions in the variables $\widehat{x}, \widehat{y}$, apart from the diagonal $\{(\widehat{x},\widehat{y}) \in \Gamma\times\Gamma : \widehat{x} = \widehat{y}\}$ and may have a singularity on the diagonal. Such kernel functions arise, for instance, by applying a boundary integral formulation to a second-order elliptic problem. In general, they decay like a negative power of the distance of the arguments which depends on the spatial dimension n and the order $2q$ of the operator.

Throughout the remainder of this paper we shall assume that the boundary manifold Γ is given as a parametric surface consisting of smooth patches. More precisely, let $\square := [0,1]^n$ denote the unit n cube. The manifold $\Gamma \subset \mathbb{R}^{n+1}$ is partitioned into a finite number of *patches*,

$$\Gamma = \bigcup_{i=1}^{M} \Gamma_i, \qquad \Gamma_i = \gamma_i(\square), \qquad i = 1, 2, \ldots, M, \tag{2.2}$$

where each $\gamma_i : \square \to \Gamma_i$ defines a diffeomorphism of $\square$ onto Γ_i. We also assume that there exist smooth extensions $\Gamma_i \subset\subset \widetilde{\Gamma}_i$ and $\widetilde{\gamma}_i : \widetilde{\square} := [-1,2]^n \to \widetilde{\Gamma}_i$. The intersection $\Gamma_i \cap \Gamma_{i'}$, $i \neq i'$, of the patches Γ_i and $\Gamma_{i'}$ is supposed to be either $\emptyset$ or a lower-dimensional face.

A mesh of level j on Γ is induced by dyadic subdivisions of depth j of the unit cube into 2^{nj} cubes $C_{j,k} \subseteq \square$, where $k = (k_1, \ldots, k_n)$ with $0 \le k_m < 2^j$. This generates $2^{nj}M$ *elements* (or elementary domains) $\Gamma_{i,j,k} := \gamma_i(C_{j,k}) \subseteq \Gamma_i$, $i = 1, \ldots, M$.

In order to ensure that the collection of elements $\{\Gamma_{i,j,k}\}$ on the level j forms a regular mesh on Γ, the parametric representation is subjected to the following *matching condition*: for all $\widehat{x} \in \Gamma_i \cap \Gamma_{i'}$ there exists a bijective, affine mapping $\Xi : \square \to \square$ such that $\gamma_i(x) = (\gamma_{i'} \circ \Xi)(x) = \widehat{x}$ for $x = (x_1, \ldots, x_n) \in \square$ with $\gamma_i(x) = \widehat{x}$.

The first fundamental tensor of differential geometry is given by the matrix $K_i(x) \in \mathbb{R}^{n\times n}$ defined by

$$K_i(x) := \left[\left(\frac{\partial\gamma_i(x)}{\partial x_j}, \frac{\partial\gamma_i(x)}{\partial x_{j'}}\right)_{l^2(\mathbb{R}^{n+1})}\right]_{j,j'=1}^{n}. \tag{2.3}$$

Since γ_i is supposed to be a diffeomorphism, the matrix $K_i(x)$ is symmetric and positive definite. The canonical inner product in $L^2(\Gamma)$ is then given by

$$\langle u, v\rangle = \int_\Gamma u(x)v(x)d\Gamma_x = \sum_{i=1}^{M} \int_\square u\big(\gamma_i(x)\big)v\big(\gamma_i(x)\big)\sqrt{\det K_i(x)}dx. \tag{2.4}$$

The corresponding Sobolev spaces are denoted by $H^s(\Gamma)$, endowed with the norms $\|\cdot\|_s$, where for $s<0$ it is understood that $H^s(\Gamma)=(H^{-s}(\Gamma))'$. Of course, depending on the global smoothness of the surface, the range of permitted $s\in\mathbb{R}$ is limited to $s\in(-s_\Gamma,s_\Gamma)$. In the case of general Lipschitz domains we have at least $s_\Gamma=1$ since for all $0\le s\le 1$ the spaces $H^s(\Gamma)$ consist of traces of functions $\in H^{s+1/2}(\Omega)$; cf. [7].

We can now specify the kernel functions. To this end, we denote by $\alpha=(\alpha_1,\dots,\alpha_n)$ and $\beta=(\beta_1,\dots,\beta_n)$ multi-indices of dimension n and define $|\alpha|:=\alpha_1+\cdots+\alpha_n$. Recall that $\widehat{x}$ and $\widehat{y}$ are points on the surface, i.e., $\widehat{x}:=\gamma_i(x)$ and $\widehat{y}:=\gamma_{i'}(y)$ for some $1\le i,i'\le M$.

Definition 2.1. *A kernel $k(\widehat{x},\widehat{y})$ is called standard kernel of order $2q$ if the partial derivatives of the transported kernel function*

$$\widetilde{k}(x,y):=k\big(\gamma_i(x),\gamma_{i'}(y)\big)\sqrt{\det K_i(x)}\sqrt{\det K_{i'}(y)} \tag{2.5}$$

are bounded by

$$|\partial_x^\alpha\partial_y^\beta\widetilde{k}(x,y)|\le c_{\alpha,\beta}\ \operatorname{dist}(\widehat{x},\widehat{y})^{-(n+2q+|\alpha|+|\beta|)}, \tag{2.6}$$

provided that $n+2q+|\alpha|+|\beta|>0$.

We emphasize that this definition requires patchwise smoothness but *not* global smoothness of the geometry. The surface itself needs to be only Lipschitz. Generally, under this assumption, the kernel of a boundary integral operator A of order $2q$ is a standard kernel of order $2q$. Hence, we may assume this property in the following. We shall encounter further specifications below in connection with discretizations.

3. Galerkin scheme. We shall be concerned with the Galerkin method with respect to a hierarchy of conforming trial spaces $V_J\subset V_{J+1}\subset H^q(\Gamma)$: find $u_J\in V_J$ solving the variational problem

$$\langle Au_J,v_J\rangle=\langle f,v_J\rangle\quad\text{for all}\quad v_J\in V_J. \tag{3.1}$$

Here the index J reflects a meshwidth of the order 2^{-J}. Moreover, we say that the trial spaces have *(approximation) order* $d\in\mathbb{N}$ and *regularity* $\gamma>0$ if

$$\begin{aligned}\gamma&=\sup\{s\in\mathbb{R}:V_J\subset H^s(\Gamma)\},\\ d&=\sup\{s\in\mathbb{R}:\inf_{v_J\in V_J}\|v-v_J\|_0\lesssim 2^{-Js}\|v\|_s\ \text{ for all } v\in H^s(\Gamma)\}.\end{aligned} \tag{3.2}$$

Thus conformity requires, of course, that $\gamma>\max\{0,q\}$.

In order to ensure that (3.1) is well posed we shall make the following assumptions on the operator A throughout the remainder of the paper.

Assumptions:

1. A is strongly elliptic, i.e., there exists a symmetric compact operator $C: H^q(\Gamma)\to H^{-q}(\Gamma)$ such that $\langle(A+A^\star+C)u,u\rangle\gtrsim\|u\|_q^2$.
2. The operator $A:H^q(\Gamma)\to H^{-q}(\Gamma)$ is injective, i.e., $\operatorname{Ker}A=\{0\}$.

Remark. 1. Most boundary integral equations of the first kind, resulting from a direct approach, are known to be strongly elliptic, even if Γ is supposed to be the boundary of a Lipschitz domain [7]. In particular, this is the case for boundary integral equations of the first kind for the Laplacian, the system of Navier–Lamé equations, and the Stokes system. For integral equations of the second kind the condition is obvious if the double layer potential operator is compact, or in the case of smooth boundaries, since the principal symbol satisfies a Gårding inequality.

2. For several boundary integral operators like the single layer operator of the Stokes system, for operators associated with Neumann problems or multiply connected domains, the second assumption is not valid. But in these cases the kernel of the operator A is finite-dimensional and known a priori. The kernels can be factored out, i.e., $A : H^q(\Gamma)/\operatorname{Ker}(A) \to (H^q(\Gamma)/\operatorname{Ker}(A))'$. A standard approach uses constrained conditions and Lagrange multipliers. With a minor modification our method can be applied also to these cases. Therefore, the second assumption is only for the sake of simplicity and not a restriction of generality.

LEMMA 3.1 (see, e.g., [31]). *Under the above assumptions the Galerkin discretization is stable, i.e.,*

$$\langle (A + A^\star)v_J, v_J\rangle \gtrsim \|v_J\|_q^2, \quad v_J \in V_J, \tag{3.3}$$

for J sufficiently large, and

$$|\langle Av_J, w_J\rangle| \lesssim \|v_J\|_q \|w_J\|_q, \quad v_J, w_J \in V_J. \tag{3.4}$$

Furthermore, let u, u_J denote the solution of the original equation $Au = f$, respectively, of (3.1). *Then one has*

$$\|u - u_J\|_t \lesssim 2^{J(t-t')}\|u\|_{t'} \tag{3.5}$$

provided that $2q - d \le t < \gamma$, $t \le t'$, $q \le t' \le d$ and Γ is sufficiently regular.

Note that the best possible convergence rate is given by

$$\|u - u_J\|_{2q-d} \lesssim 2^{-2J(d-q)}\|u\|_d \tag{3.6}$$

provided that $u \in H^d(\Gamma)$, which is only possible when Γ is sufficiently regular. Since this case gives rise to the highest convergence rate, it will be seen later to impose the most stringent demands on the matrix compression.

4. Wavelets and multiresolution analysis.

The nested trial spaces $V_j \subset V_{j+1}$ that we shall employ in (3.1) are spanned by so-called *single-scale bases* $\Phi_j = \{\phi_{j,k} : k \in \Delta_j\}$, where Δ_j denote suitable index sets of cardinality $\dim V_j$. The elements of Φ_j are normalized in $L^2(\Gamma)$ and their compact supports scale like $\operatorname{diam}\operatorname{supp}\phi_{j,k} \sim 2^{-j}$. Associated with these collections are always *dual* bases $\widetilde{\Phi}_j = \{\widetilde{\phi}_{j,k} : k \in \Delta_j\}$, i.e., one has $\langle \phi_{j,k}, \widetilde{\phi}_{j,k'}\rangle = \delta_{k,k'}$, $k, k' \in \Delta_j$. For the current type of boundary surfaces Γ the $\Phi_j, \widetilde{\Phi}_j$ are generated by constructing first dual pairs of single-scale bases on the interval $[0, 1]$, using B-splines for the primal bases and the dual components from [5] adapted to the interval [10]. Tensor products yield corresponding dual pairs on $\square$. Using the parametric liftings γ_i and gluing across patch boundaries leads to globally continuous single-scale bases $\Phi_j, \widetilde{\Phi}_j$ on Γ [3, 6, 14, 19]. For B-splines of order d and duals of order $\widetilde{d} \ge d$ such that $d + \widetilde{d}$ is even, the $\Phi_j, \widetilde{\Phi}_j$ have approximation orders $d, \widetilde{d}$, respectively. It is known that the respective regularity indices $\gamma, \widetilde{\gamma}$ (inside each patch) satisfy $\gamma = d - 1/2$, while $\widetilde{\gamma} > 0$ is known to increase proportionally to $\widetilde{d}$. We refer the reader to [21] for a detailed description of the construction of wavelets on manifolds, including examples and figures.

In view of the biorthogonality of $\Phi_j, \widetilde{\Phi}_j$, it will be convenient to employ the canonical projectors

$$Q_j v := \sum_{k\in\Delta_j} \langle v, \widetilde{\phi}_{j,k}\rangle \phi_{j,k}, \qquad Q_j^\star v := \sum_{k\in\Delta_j} \langle v, \phi_{j,k}\rangle \widetilde{\phi}_{j,k}, \tag{4.1}$$

associated with the *multiresolution sequences* $\{V_j\}_{j>j_0}$, $\{\widetilde{V}_j\}_{j>j_0}$. Here and below j_0+1 always stands for some fixed coarsest level of resolution that may depend on Γ.

It follows from the L^2-boundedness of the Q_j that one has the following *Jackson-* and *Bernstein*-type estimates uniformly in j, namely,

$$\|v - Q_j v_j\|_s \lesssim 2^{-j(t-s)}\|v\|_t, \quad v \in H^t(\Gamma), \tag{4.2}$$

for all $-\widetilde{d} \leq s \leq t \leq d$, $s < \gamma$, $-\widetilde{\gamma} < t$ and

$$\|Q_j v\|_s \lesssim 2^{j(s-t)}\|Q_j v\|_t, \quad v \in H^t(\Gamma), \tag{4.3}$$

for all $t \leq s \leq \gamma$.

We introduce the index sets $\nabla_j := \Delta_{j+1} \setminus \Delta_j$. Given the single-scale bases $\Phi_j, \widetilde{\Phi}_j$, one can construct now biorthogonal *complement bases* $\Psi_j = \{\psi_{j,k} : k \in \nabla_j\}$, $\widetilde{\Psi}_j = \{\widetilde{\psi}_{j,k} : k \in \nabla_j\}$, i.e., $\langle \psi_{j,k}, \widetilde{\psi}_{j',k'}\rangle = \delta_{(j,k),(j',k')}$, such that

$$\operatorname{diam}\operatorname{supp}\psi_{j,k} \sim 2^{-j}, \quad j > j_0; \tag{4.4}$$

see, e.g., [3, 6, 13, 14] and [19] for particularly useful local representations of important construction ingredients. In fact, for these types of bases, the dual wavelets scale in the same way, but this will not be needed and does not hold for alternative constructions based on finite elements [15].

Denoting by $W_j, \widetilde{W}_j$ the span of Ψ_j, respectively, $\widetilde{\Psi}_j$, biorthogonality implies that

$$V_{j+1} = W_j \oplus V_j, \qquad \widetilde{V}_{j+1} = \widetilde{W}_j \oplus \widetilde{V}_j, \qquad \widetilde{V}_j \perp W_j, \qquad V_j \perp \widetilde{W}_j.$$

Hence V_J and $\widetilde{V}_J$ can be written as a direct sum of the complement spaces W_j, respectively, $\widetilde{W}_j$, $j_0 \leq j < J$ (using the convention $W_{j_0} := V_{j_0+1}$, $\widetilde{W}_{j_0} := \widetilde{V}_{j_0+1}$, $Q_{j_0} = Q^\star_{j_0} := 0$). In fact, one has for $v_J \in V_J$, $\widetilde{v}_J \in \widetilde{V}_J$

$$v_J = \sum_{j=j_0}^{J-1}(Q_{j+1} - Q_j)v_J, \qquad \widetilde{v}_J = \sum_{j=j_0}^{J-1}(Q^\star_{j+1} - Q^\star_j)\widetilde{v}_J,$$

where

$$(Q_{j+1} - Q_j)v = \sum_{k\in\nabla_j}\langle v, \widetilde{\psi}_{j,k}\rangle\psi_{j,k}, \qquad (Q^\star_{j+1} - Q^\star_j)v = \sum_{k\in\nabla_j}\langle v, \psi_{j,k}\rangle\widetilde{\psi}_{j,k}.$$

A biorthogonal or *dual* pair of wavelet bases is now obtained by taking the coarse single-scale basis and the union of the complement bases $\Psi = \bigcup_{j\geq j_0}\Psi_j$, $\widetilde{\Psi} = \bigcup_{j\geq j_0}\widetilde{\Psi}_j$, where we have set for convenience $\Psi_{j_0} := \Phi_{j_0+1}$, $\widetilde{\Psi}_{j_0} := \widetilde{\Phi}_{j_0+1}$. We will refer to Ψ and $\widetilde{\Psi}$ as the *primal*, respectively, *dual*, basis. Throughout the paper, all basis functions (scaling functions and wavelets) are normalized in $L^2(\Gamma)$.

From biorthogonality and the fact that the dual single-scale bases on $\square$ represent all polynomials of order $\widetilde{d}$ exactly, one infers vanishing polynomial moments of the primal wavelets on $\square$, which, on account of the locality (4.4), entails the first key feature of the primal wavelets, namely, *vanishing moments* or the *cancellation property*

$$|\langle v, \psi_{j,k}\rangle| \lesssim 2^{-j(\widetilde{d}+n/2)}|v|_{W^{\widetilde{d},\infty}(\operatorname{supp}\psi_{j,k})}. \tag{4.5}$$

Here $|v|_{W^{\tilde{d},\infty}(\Omega)} := \sup_{|\alpha|=\widetilde{d},\, x\in\Omega} |\partial^\alpha v(x)|$ denotes the seminorm in $W^{\widetilde{d},\infty}(\Omega)$. The fact that the concept of biorthogonality allows us to choose the order $\widetilde{d}$ of vanishing moments higher than the approximation order d will be essential for deriving optimal compression strategies that could not be realized by orthonormal bases.

Of course, in the infinite-dimensional case the notion of basis has to be made more specific. The second key feature of the basis Ψ is the fact that (properly scaled versions of) Ψ, $\widetilde{\Psi}$ are actually *Riesz bases* for a *whole range* of Sobolev spaces, i.e.,

$$\begin{aligned} \|v\|_t^2 &\sim \sum_{j\ge j_0}\sum_{k\in\nabla_j} 2^{2jt}|\langle v,\widetilde{\psi}_{j,k}\rangle|^2, \qquad t\in(-\widetilde{\gamma},\gamma), \\ \|v\|_t^2 &\sim \sum_{j\ge j_0}\sum_{k\in\nabla_j} 2^{2jt}|\langle v,\psi_{j,k}\rangle|^2, \qquad t\in(-\gamma,\widetilde{\gamma}). \end{aligned} \tag{4.6}$$

The validity of these norm equivalences hinges on the estimates (4.2) and (4.3) for *both* the primal and dual multiresolution sequences. The equivalences (4.6) will be essential for preconditioning.

5. Wavelet Galerkin schemes—preconditioning. As before let $A : H^q(\Gamma) \to H^{-q}(\Gamma)$ be a boundary integral operator of order $2q$. Since the wavelet basis Ψ is, in particular, a Riesz basis for $L^2(\Gamma)$, the associated system matrices

$$\mathbf{A}_J = [\langle A\psi_{j',k'},\psi_{j,k}\rangle]_{j_0\le j,j'<J,\, k\in\nabla_j,\, k'\in\nabla_{j'}}$$

become more and more ill-conditioned when J increases. In fact, one has $\operatorname{cond}_{l^2}\mathbf{A}_J \sim 2^{2J|q|}$. However, as a consequence of the stability of the Galerkin discretization under the given circumstances and the norm equivalences (4.6), the following simple *diagonal preconditioner* gives rise to uniformly bounded spectral condition numbers [8, 9, 11].

THEOREM 5.1. *Let the diagonal matrix* $\mathbf{D}_J^r$ *be defined by*

$$\left[\mathbf{D}_J^r\right]_{(j,k),(j',k')} = 2^{rj}\delta_{(j,k),(j',k')}, \quad k\in\nabla_j, \quad k'\in\nabla_{j'}, \quad j_0\le j,j'<J.$$

If $A : H^q(\Gamma)\to H^{-q}(\Gamma)$ *is a boundary integral operator of order* $2q$*, satisfying the assumptions* (1), (2) *from section* 3*, and if* $\widetilde{\gamma} > -q$*, the diagonal matrix* $\mathbf{D}_J^{2q}$ *defines an asymptotically optimal preconditioner for* $\mathbf{A}_J$*, i.e.,* $\operatorname{cond}_{l^2}(\mathbf{D}_J^{-q}\mathbf{A}_J\mathbf{D}_J^{-q}) \sim 1$.

Although the above scaling is asymptotically optimal, it should be stressed that the quantitative performance may vary significantly among different scalings with the same asymptotic behavior. In particular, since Ψ is, on account of the mapping properties of A and the norm equivalences (4.6), also a Riesz basis with respect to the energy norm, it would be natural to normalize the wavelets in the energy norm which would suggest the specific scaling $\langle A\psi_{j,k},\psi_{j,k}\rangle \sim 2^{2qj}$. In fact, this latter diagonal scaling improves and simplifies the wavelet preconditioning.

In view of the above simple preconditioning, the iterative solution of the Galerkin systems is feasible and its overall efficiency relies now on the cost of matrix/vector multiplications, which brings us to the central issue, namely, *matrix compression.*

6. Basic estimates. The basic ingredients in the analysis of the compression procedure are estimates for the matrix entries $\langle A\psi_{j',k'},\psi_{j,k}\rangle$ with $k\in\nabla_j$, $k'\in\nabla_{j'}$ and $j,j'\ge j_0$. The convex hulls of the supports of the wavelets will be denoted by

$$\Omega_{j,k} := \operatorname{conv\,hull}(\operatorname{supp}\psi_{j,k}). \tag{6.1}$$

A complete proof of the following estimates can be found, e.g., in [15, 27].

THEOREM 6.1. *Suppose $n + 2\widetilde{d} + 2q > 0$ and $j, j' > j_0$. Then one has*

$$|\langle A\psi_{j',k'}, \psi_{j,k}\rangle| \lesssim 2^{-(j+j')(\widetilde{d}+n/2)} \operatorname{dist}(\Omega_{j,k}, \Omega_{j',k'})^{-(n+2q+2\widetilde{d})}$$

uniformly with respect to J.

However, in order to arrive ultimately at solution schemes with linear complexity, the number of nonzero entries in the compressed matrices should remain proportional to their size while preserving discretization error accuracy. To achieve this, it is not sufficient to consider only coefficients where the supports of the involved wavelets do not overlap. There are still $\mathcal{O}(N_J \log N_J) = \mathcal{O}(2^{Jn}J)$ coefficients that would remain. To avoid the logarithmic term we propose an additional, so-called *second compression.* For this purpose, we require that our primal basis functions be piecewise polynomial, in the sense that $\psi_{j,k}\big|_{\Gamma_{i,j+1,l}} = p \circ \gamma_i^{-1}$, where p is a polynomial. By

$$\Omega'_{j,k} := \operatorname{sing\,supp} \psi_{j,k} \tag{6.2}$$

we denote the *singular support* of $\psi_{j,k}$, which is that subset of Γ where the function $\psi_{j,k}$ is not smooth. Thus the singular support of the wavelet $\psi_{j,k}$ consists of the boundaries of some of the elements $\Gamma_{i,j+1,l}$. The goal of the subsequent investigation is to estimate those matrix entries for which $\operatorname{dist}(\Omega_{j,k}, \Omega'_{j',k'})$, $j \geq j'$, is sufficiently large.

To this end, we require the following extension lemma which follows, e.g., immediately from the well-known extension theorem of Calderón [28].

LEMMA 6.2. *The function $f_{i,j,k,l}$, defined by*

$$f_{i,j,k,l} := \psi_{j,k}\big|_{\Gamma_{i,j+1,l}} \circ \gamma_i = (\psi_{j,k} \circ \gamma_i)\big|_{C_{j+1,l}} \in C^\infty(C_{j+1,l}),$$

can be extended to a function $\widetilde{f}_{i,j,k,l} \in C_0^\infty(\mathbb{R}^n)$ in such a way that $\operatorname{diam} \operatorname{supp} \widetilde{f}_{i,j,k,l} \lesssim 2^{-j}$*, $\widetilde{f}_{i,j,k,l} \equiv \psi_{j,k} \circ \gamma_i$ on $C_{j+1,l}$, and that for all $s \geq 0$ there holds $\big\|\widetilde{f}_{i,j,k,l}\big\|_{H^s(\mathbb{R}^n)} \lesssim 2^{js}$, independently of i, j, k, l.*

Proof. Suppose that $f_\square \in C^\infty(\square)$ with $\|f_\square\|_{H^s(\square)} \lesssim 1$. By virtue of Calderón's extension theorem, there exists an extension $f \in C_0^\infty(\mathbb{R}^n)$, i.e., $f(x) \equiv f_\square(x)$ on $\square$, satisfying $\|f\|_{H^s(\mathbb{R}^n)} \lesssim \|f_\square\|_{H^s(\square)}$. Let us consider an affine map κ with $\kappa(C_{j+1,l}) = \square$ and choose $f_{i,j,k,l}(x) := f_\square\big(\kappa(x)\big)$. The claim follows now from $|\partial_{x_i}\kappa| = 2^{j+1}$, $i = 1, \ldots, n$. $\square$

It is well known that boundary integral operators A of order $2q$, acting on smooth surfaces, are classical *pseudodifferential operators* [28]. Since the patches Γ_i are smooth and have smooth extensions $\widetilde{\Gamma}_i$, there exists for each i a pseudodifferential operator $A^\sharp : H^q(\mathbb{R}^n) \to H^{-q}(\mathbb{R}^n)$ such that

$$A^\sharp f(x) = \int_{\mathbb{R}^n} \chi(x)\chi(y) k\big(\widetilde{\gamma}_i(x), \widetilde{\gamma}_i(y)\big) f(y) \sqrt{\det \widetilde{K}_i(y)}dy, \tag{6.3}$$

where χ is a C^∞-cut-off function with respect to $\square$, i.e., $\chi(x) = 1$ on $\square$ and $\chi(x) = 0$ outside $[-1, 2]^n$. Therefore $A^\sharp f(x) = A(f \circ \gamma_i)(\gamma_i(x))$ for all $f \in C_0^\infty(\square)$, $x \in \square$, and $A^\sharp$ is compactly supported [29]. Moreover, it is well known [28] that the Schwartz kernel of pseudodifferential operators satisfies the standard estimate (2.6).

A compactly supported pseudodifferential operator $A^\sharp : H^s(\mathbb{R}^n) \to H^{s-2q}(\mathbb{R}^n)$ of order $2q$ acts continuously on Sobolev spaces [28, 29]. Therefore, for any function

$\widetilde{f}_{i,j,k,l} \in C_0^\infty(\mathbb{R}^n)$, satisfying $\operatorname{diam}\operatorname{supp}\widetilde{f}_{i,j,k,l} \sim 2^{-j}$ and $\|\widetilde{f}_{i,j,k,l}\|_{H^s(\mathbb{R}^n)} \lesssim 2^{js}$ for all $s \geq 0$, one has $A^\sharp \widetilde{f}_{i,j,k,l} \in C_0^\infty(\mathbb{R}^n)$ with

$$\|A^\sharp \widetilde{f}_{i,j,k,l}\|_{H^{s-2q}(\mathbb{R}^n)} \lesssim 2^{js}. \tag{6.4}$$

With these preparations at hand, we are able to formulate the following result.

THEOREM 6.3. *Suppose that $n + 2\widetilde{d} + 2q > 0$ and $j' > j \geq j_0$. Then, the coefficients $\langle A\psi_{j',k'}, \psi_{j,k}\rangle$ and $\langle A\psi_{j,k}, \psi_{j',k'}\rangle$ satisfy*

$$|\langle A\psi_{j',k'}, \psi_{j,k}\rangle|,\ |\langle A\psi_{j,k}, \psi_{j',k'}\rangle| \lesssim 2^{jn/2} 2^{-j'(\widetilde{d}+n/2)} \operatorname{dist}(\Omega'_{j,k}, \Omega_{j',k'})^{-(2q+\widetilde{d})},$$

uniformly with respect to j, provided that

$$\operatorname{dist}(\Omega'_{j,k}, \Omega_{j',k'}) \gtrsim 2^{-j'}. \tag{6.5}$$

Proof. We shall consider three cases.

(i) The first observation concerns an estimate for disjoint supports that will be applied several times.

LEMMA 6.4. *Suppose that $\Omega_{j,k} \cap \Omega_{j',k'} = \emptyset$ and that f is any function supported on $\Omega_{j,k}$ satisfying $|f(x)| \lesssim 2^{jn/2}$, $x \in \Omega_{j,k}$. Then one has*

$$|\langle A\psi_{j',k'}, f\rangle| \lesssim 2^{jn/2} 2^{-j'(\widetilde{d}+n/2)} \operatorname{dist}(\Omega_{j,k}, \Omega_{j',k'})^{-(2q+\widetilde{d})}. \tag{6.6}$$

To prove (6.6) note that our assumption implies $\operatorname{dist}(\Omega'_{j,k}, \Omega_{j',k'}) = \operatorname{dist}(\Omega_{j,k}, \Omega_{j',k'})$. On account of the cancellation property (4.5) of the wavelet bases and the decay property (2.6) of the kernel, we obtain

$$\begin{aligned}|A\psi_{j',k'}(x)| = |\langle k(x,\cdot), \psi_{j',k'}\rangle| &\lesssim 2^{-j'(\widetilde{d}+n/2)} |k(x,\cdot)|_{W^{\infty,\widetilde{d}}(\Omega_{j',k'})}\\ &\lesssim 2^{-j'(\widetilde{d}+n/2)} \operatorname{dist}(x, \Omega_{j',k'})^{-(n+2q+\widetilde{d})}\end{aligned}$$

for all $x \in \operatorname{supp}\psi_{j,k}$. Therefore, we conclude that

$$\begin{aligned}|\langle A\psi_{j',k'}, f\rangle| &\lesssim \|f\|_{L^\infty(\Gamma)} \int_{\Omega_{j,k}} |A\psi_{j',k'}(x)| d\Gamma_x\\ &\lesssim 2^{jn/2} 2^{-j'(\widetilde{d}+n/2)} \int_{\Omega_{j,k}} \operatorname{dist}(x, \Omega_{j',k'})^{-(n+2q+\widetilde{d})} d\Gamma_x\\ &\leq 2^{jn/2} 2^{-j'(\widetilde{d}+n/2)} \operatorname{dist}(\Omega_{j,k}, \Omega_{j',k'})^{-(2q+\widetilde{d})},\end{aligned}$$

which proves the lemma. Of course, the same reasoning applies to the adjoint boundary integral operator $A^\star$.

(ii) Next, we treat the case $\Omega_{j,k} \cap \Omega_{j',k'} \neq \emptyset$ and $\Omega_{j,k} \subset \Gamma_i$. By (6.5) we have $\Omega_{j',k'} \subset \Omega_{j,k}$ i.e., both wavelets are supported on the same patch. We infer from (6.5) that there exists an element $\Omega_{j',k'} \subset \Gamma_{i,j+1,l} \subset \Omega_{j,k}$ such that

$$f_{i,j,k,l} := \psi_{j,k}\big|_{\Gamma_{i,j+1,l}} \circ \gamma_i = (\psi_{j,k} \circ \gamma_i)\big|_{C_{j+1,l}}$$

is a $C^\infty(C_{j+1,l})$ function. On account of Lemma 6.2, we can choose an extension of $f_{i,j,k,l}$, denoted by $\widetilde{f}_{i,j,k,l}$. Decomposing $\psi_{j,k}\circ\gamma_i = \widetilde{f}_{i,j,k,l} + \widetilde{f}^C_{i,j,k,l}$, we obtain

$$\begin{aligned}|\langle A\psi_{j,k},\psi_{j',k'}\rangle| &= \left|\int_{\mathbb{R}^n} A^\sharp(\widetilde{f}_{i,j,k,l}+\widetilde{f}^C_{i,j,k,l})(x)(\psi_{j',k'}\circ\gamma_i)(x)dx\right| \\ &\le \left|\int_{\mathbb{R}^n} A^\sharp\widetilde{f}_{i,j,k,l}(x)(\psi_{j',k'}\circ\gamma_i)(x)dx\right| \\ &\quad + \left|\int_{\mathbb{R}^n} A^\sharp\widetilde{f}^C_{i,j,k,l}(x)(\psi_{j',k'}\circ\gamma_i)(x)dx\right|.\end{aligned}$$

The second term on the right-hand side can be treated analogously to (6.6), i.e.,

$$\left|\int_{\mathbb{R}^n} A^\sharp\widetilde{f}^C_{i,j,k,l}(x)(\psi_{j',k'}\circ\gamma_i)(x)dx\right| \lesssim 2^{jn/2}2^{-j'(\tilde{d}+n/2)}\operatorname{dist}(\Omega'_{j,k},\Omega_{j',k'})^{-(2q+\tilde{d})},$$

because $\operatorname{dist}\big(\operatorname{supp}\widetilde{f}^C_{i,j,k,l},\operatorname{supp}(\psi_{j',k'}\circ\gamma_i)\big)\sim\operatorname{dist}(\Omega'_{j,k},\Omega_{j',k'})$. Invoking (4.5) and (6.4), the first term can be estimated by

$$\left|\int_{\mathbb{R}^n} A^\sharp\widetilde{f}_{i,j,k,l}(x)(\psi_{j',k'}\circ\gamma_i)(x)dx\right| \lesssim 2^{-j'(\tilde{d}+n/2)}\|A^\sharp\widetilde{f}_{i,j,k,l}(x)\|_{W^{\infty,\tilde{d}}(\operatorname{supp}(\psi_{j',k'}\circ\gamma_i))}.$$

By virtue of Sobolev's embedding theorem, this implies, in view of (6.4),

$$\begin{aligned}\left|\int_{\mathbb{R}^n} A^\sharp\widetilde{f}_{i,j,k,l}(x)(\psi_{j',k'}\circ\gamma_i)(x)dx\right| &\lesssim 2^{-j'(\tilde{d}+n/2)}\|A^\sharp\widetilde{f}_{i,j,k,l}(x)\|_{H^{\tilde{d}+n/2}(\mathbb{R}^n)} \\ &\lesssim 2^{-j'(\tilde{d}+n/2)}2^{j(\tilde{d}+2q+n/2)}.\end{aligned}$$

Since $\Omega_{j,k}\cap\Omega_{j',k'}\neq\emptyset$, one has, in view of (6.5) and $j'\ge j$, that $\operatorname{dist}(\Omega'_{j,k},\Omega_{j',k'})\lesssim 2^{-j}$, so that we arrive at the desired estimate

$$\left|\int_{\mathbb{R}^n} A^\sharp\widetilde{f}_{i,j,k,l}(x)(\psi_{j',k'}\circ\gamma_i)(x)dx\right| \lesssim 2^{jn/2}2^{-j'(\tilde{d}+n/2)}\operatorname{dist}(\Omega'_{j,k},\Omega_{j',k'})^{-(2q+\tilde{d})}.$$

(iii) It remains to consider the case $\Omega_{j,k}\cap\Omega_{j',k'}\neq\emptyset$, where, however, $\psi_{j,k}$ is not supported completely in the patch Γ_i. In this case, we decompose $\psi_{j,k} = (\psi_{j,k} - \psi_{j,k}\big|_{\Gamma_i}) + \psi_{j,k}\big|_{\Gamma_i}$. Invoking (6.6), we derive

$$\big|\langle A(\psi_{j,k}-\psi_{j,k}\big|_{\Gamma_i}),\psi_{j',k'}\rangle\big| \lesssim 2^{jn/2}2^{-j'(\tilde{d}+n/2)}\operatorname{dist}(\Omega'_{j,k},\Omega_{j',k'})^{-(2q+\tilde{d})}$$

because we have again $\operatorname{dist}\big(\operatorname{supp}(\psi_{j,k}-\psi_{j,k}\big|_{\Gamma_i}),\Omega_{j',k'}\big)\ge\operatorname{dist}(\Omega'_{j,k},\Omega_{j',k'})$. Finally, estimating $\big|\langle A(\psi_{j,k}\big|_{\Gamma_i}),\psi_{j',k'}\rangle\big|$ as in step (ii) finishes the proof. □

Remark. We recall from [8, 27] that there is a general estimate which states that the matrix entries for wavelets with overlapping supports decay with increasing difference of scales. In fact, for each $0\le\delta<\gamma-q$ we have $|\langle A\psi_{j',k'},\psi_{j,k}\rangle|\lesssim 2^{-\delta|j-j'|}$. Since $\gamma<d$ this estimate is, however, not sufficient to achieve the optimal order of convergence within the desired linear complexity.

7. Matrix compression. The discretization of a boundary integral operator $A: H^q(\Gamma)\to H^{-q}(\Gamma)$ by wavelets with a sufficiently strong cancellation property (4.5) yields, in view of the above estimates, quasi-sparse matrices. In the first compression step all matrix entries, for which the distance of the supports of the corresponding trial and test functions is larger than a level depending cut-off parameter $\mathcal{B}_{j,j'}$, are set to zero. In the second compression step also some of those matrix entries are neglected, for which the corresponding trial and test functions have overlapping supports.

A priori compression. Let $\Omega_{j,k}$ and $\Omega'_{j,k}$ be given as in (6.1) and (6.2). Then, the compressed system matrix $\mathbf{A}_J^\epsilon$, corresponding to the boundary integral operator A, is defined by

$$[\mathbf{A}_J^\epsilon]_{(j,k),(j',k')} := \begin{cases} 0, & \operatorname{dist}(\Omega_{j,k},\Omega_{j',k'}) > \mathcal{B}_{j,j'} \text{ and } j,j' > j_0, \\ 0, & \operatorname{dist}(\Omega_{j,k},\Omega_{j',k'}) \lesssim 2^{-\min\{j,j'\}} \text{ and} \\ & \operatorname{dist}(\Omega'_{j,k},\Omega_{j',k'}) > \mathcal{B}'_{j,j'} \text{ if } j' > j \geq j_0, \\ & \operatorname{dist}(\Omega_{j,k},\Omega'_{j',k'}) > \mathcal{B}'_{j,j'} \text{ if } j > j' \geq j_0, \\ \langle A\psi_{j',k'},\psi_{j,k}\rangle, & \text{otherwise.} \end{cases} \tag{7.1}$$

Fixing

$$a, a' > 1, \qquad d < d' < \widetilde{d} + 2q, \tag{7.2}$$

the cut-off parameters $\mathcal{B}_{j,j'}$ and $\mathcal{B}'_{j,j'}$ are set as follows:

$$\begin{aligned} \mathcal{B}_{j,j'} &= a\ \max\left\{2^{-\min\{j,j'\}}, 2^{\frac{2J(d'-q)-(j+j')(d'+\widetilde{d})}{2(\widetilde{d}+q)}}\right\}, \\ \mathcal{B}'_{j,j'} &= a' \max\left\{2^{-\max\{j,j'\}}, 2^{\frac{2J(d'-q)-(j+j')d'-\max\{j,j'\}\widetilde{d}}{\widetilde{d}+2q}}\right\}. \end{aligned} \tag{7.3}$$

Remark 7.1. Relation (7.2) requires the order of vanishing moments, viz. the order of exactness of the dual basis, to exceed the order of the primal basis by an amount determined by the order of the operator $d < \widetilde{d} + 2q$. In the case of equality the matrices are still compressible, but additional log terms arise in the complexity estimates. One can find many pairs of biorthogonal wavelets from the family in [14, 21] satisfying this relation. To obtain quantitatively best performance we employ those with possibly small support, i.e., with possibly small $\widetilde{d}$ satisfying $d < \widetilde{d} + 2q$. This choice is confirmed by numerical experience.

The parameter a is a fixed constant which determines the bandwidth in the block matrices $\mathbf{A}_{j,j'}^\epsilon := [\mathbf{A}_J^\epsilon]_{(j,\nabla_j),(j',\nabla_{j'})}$, $j_0 \leq j, j' < J$. We emphasize that the parameters a and a' are independent of J.

When the entries of the compressed system matrix $\mathbf{A}_J^\epsilon$ have been computed, we apply an *a posteriori compression* by setting all entries to zero, which are smaller than a level-depending threshold. In this way, a matrix $\widetilde{\mathbf{A}}_J^\epsilon$ is obtained which has even less nonzero entries than the matrix $\mathbf{A}_J^\epsilon$. Although this does not accelerate the computation of the matrix coefficients, the amount of necessary memory for storing the system matrix is reduced considerably.

A posteriori compression. We define the a posteriori compression by

$$\left[\widetilde{\mathbf{A}}_J^\epsilon\right]_{(j,k),(j',k')} = \begin{cases} 0 & \text{if } \left|[\mathbf{A}_J^\epsilon]_{(j,k),(j',k')}\right| \leq \varepsilon_{j,j'}, \\ [\mathbf{A}_J^\epsilon]_{(j,k),(j',k')} & \text{if } \left|[\mathbf{A}_J^\epsilon]_{(j,k),(j',k')}\right| > \varepsilon_{j,j'}. \end{cases} \tag{7.4}$$

Here the level-dependent threshold $\varepsilon_{j,j'}$ is chosen as

$$\varepsilon_{j,j'} = a'' \min\left\{2^{-\frac{|j-j'|n}{2}}, 2^{-n(J-\frac{j+j'}{2})\frac{d'-q}{\widetilde{d}+q}}\right\} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})} \tag{7.5}$$

with $a'' < 1$ and $d' \in (d, \widetilde{d} + 2q)$ from (7.2).

8. Matrix estimates. In order to study the accuracy of the solutions to the compressed systems, we investigate the perturbation introduced by discarding specific matrix elements. The perturbation matrices are of scalewise blocks of the type $\mathbf{R}_{j,j'} := \mathbf{A}_{j,j'} - \mathbf{A}^{\epsilon}_{j,j'}$. By $\|\mathbf{R}_{j,j'}\|_p$ we denote the operator norm of the matrix $\mathbf{R}_{j,j'}$ with respect to the norm l^p.

In order to analyze the error introduced by our compression strategy, we decompose the complete compression into three subsequent steps.

THEOREM 8.1 (first compression). *We define the matrix* $\mathbf{A}^{\epsilon_1}_J$ *by*

$$[\mathbf{A}^{\epsilon_1}_J]_{(j,k),(j',k')} := \begin{cases} 0, & \operatorname{dist}(\Omega_{j,k}, \Omega_{j',k'}) > \mathcal{B}_{j,j'} \textit{ and } j, j' > j_0, \\ \langle A\psi_{j',k'}, \psi_{j,k}\rangle, & \textit{otherwise.} \end{cases}$$

Here the parameter $\mathcal{B}_{j,j'}$ *is given by* (7.2) *and* (7.3). *Then, one has for the perturbation matrix* $\mathbf{R}_{j,j'} := \mathbf{A}_{j,j'} - \mathbf{A}^{\epsilon_1}_{j,j'}$

$$\|\mathbf{R}_{j,j'}\|_2 \lesssim a^{-2(\tilde{d}+q)} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})}.$$

Proof. We proceed in two steps.

(i) We abbreviate $\mathbf{R}_{j,j'} := \left[r_{(j,k),(j',k')}\right]_{k\in\nabla_j, k'\in\nabla_{j'}}$. Invoking Theorem 6.1, we find for the column sum

$$\begin{aligned}\sum_{k\in\nabla_j} |r_{(j,k),(j',k')}| &= \sum_{\{k\in\nabla_j :\, \operatorname{dist}(\Omega_{j,k},\Omega_{j',k'})>\mathcal{B}_{j,j'}\}} |\langle A\psi_{j',k'}, \psi_{j,k}\rangle| \\ &\lesssim \sum_{\{k\in\nabla_j :\, \operatorname{dist}(\Omega_{j,k},\Omega_{j',k'})>\mathcal{B}_{j,j'}\}} \\ &\quad \times 2^{-(j+j')(\tilde{d}+n/2)} \operatorname{dist}\left(\Omega_{j,k}, \Omega_{j',k'}\right)^{-(n+2\tilde{d}+2q)}.\end{aligned}$$

Since $\mathcal{B}_{j,j'} \geq a \max\{2^{-j}, 2^{-j'}\}$, we can estimate this sum by an integral which yields

$$\begin{aligned}\sum_{k\in\nabla_j} |r_{(j,k),(j',k')}| &\lesssim 2^{-(j+j')(\tilde{d}+n/2)} 2^{jn} \int_{\|x\|>\mathcal{B}_{j,j'}} \|x\|^{-(n+2\tilde{d}+2q)} dx \\ &\lesssim 2^{-(j+j')(\tilde{d}+n/2)} 2^{jn} \mathcal{B}_{j,j'}^{-2(\tilde{d}+q)}.\end{aligned}$$

On the other hand, inserting the estimate $\mathcal{B}_{j,j'} \geq a 2^{\frac{2J(d'-q)-(j+j')(\tilde{d}+d')}{2(\tilde{d}+q)}}$ (see (7.3)), we arrive at

$$\sum_{k\in\nabla_j} |r_{(j,k),(j',k')}| \lesssim a^{-2(\tilde{d}+q)} 2^{\frac{(j-j')n}{2}} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})}.$$

In complete analogy, one proves an analogous estimate for the row sums,

$$\sum_{k'\in\nabla_{j'}} |r_{(j,k),(j',k')}| \lesssim a^{-2(\tilde{d}+q)} 2^{\frac{(j'-j)n}{2}} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})}.$$

(ii) From the estimate for the operator norms of matrices $\|\mathbf{R}_{j,j'}\|_2^2 \leq \|\mathbf{R}_{j,j'}\|_1 \|\mathbf{R}_{j,j'}\|_\infty$, it is easy to conclude the following version of the Schur lemma (see, e.g., [23, 27]):

$$\begin{aligned}\|\mathbf{R}_{j,j'}\|_2 &\leq \left[\max_{k\in\nabla_j} \sum_{k'\in\nabla_{j'}} 2^{\frac{(j-j')n}{2}} |r_{(j,k),(j',k')}|\right]^{1/2} \left[\max_{k'\in\nabla_{j'}} \sum_{k\in\nabla_j} 2^{\frac{(j'-j)n}{2}} |r_{(j,k),(j',k')}|\right]^{1/2} \\ &\lesssim a^{-2(\tilde{d}+q)} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})},\end{aligned}$$

which proves the assertion. □

The following so-called second compression concerns entries involving basis functions with overlapping supports. It is important that here the coarse scale basis function may be a scaling function which greatly affects the near field compression.

THEOREM 8.2 (second compression). *In addition to the first compression we apply the following second compression:*

$$[\mathbf{A}_J^{\epsilon_2}]_{(j,k),(j',k')} := \begin{cases} 0, & \operatorname{dist}(\Omega_{j,k},\Omega_{j',k'}) \lesssim 2^{-\min\{j,j'\}} \text{ and} \\ & \operatorname{dist}(\Omega'_{j,k},\Omega_{j',k'}) > \mathcal{B}'_{j,j'} \text{ if } j' > j \geq j_0, \\ & \operatorname{dist}(\Omega_{j,k},\Omega'_{j',k'}) > \mathcal{B}'_{j,j'} \text{ if } j > j' \geq j_0, \\ [\mathbf{A}_J^{\epsilon_1}]_{(j,k),(j',k')}, & \text{otherwise,} \end{cases}$$

where the parameter $\mathcal{B}'_{j,j'}$ is set in accordance with (7.2) *and* (7.3). *Then, the corresponding perturbation matrix $\mathbf{S}_{j,j'} := \mathbf{A}^{\epsilon_1}_{j,j'} - \mathbf{A}^{\epsilon_2}_{j,j'}$ satisfies*

$$\|\mathbf{S}_{j,j'}\|_2 \lesssim (a')^{-(\tilde{d}+2q)} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})}.$$

Proof. Abbreviating $\mathbf{S}_{j,j'} := [s_{(j,k),(j',k')}]_{k\in\nabla_j,k'\in\nabla_{j'}}$ and assuming without loss of generality that $j' > j$, we infer from Theorem 6.3 that

$$\begin{aligned} |s_{(j,k),(j',k')}| &\lesssim 2^{jn/2} 2^{-j'(\tilde{d}+n/2)} \mathcal{B}_{j,j'}^{-(2q+\tilde{d})} \\ &\lesssim (a')^{-(\tilde{d}+2q)} 2^{jn/2} 2^{-j'(\tilde{d}+n/2)} 2^{-2J(d'-q)+(j+j')d'+j'\tilde{d}} \\ &= (a')^{-(\tilde{d}+2q)} 2^{(j-j')n/2} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})}. \end{aligned}$$

The condition $\operatorname{dist}(\Omega_{j,k},\Omega_{j',k'}) \lesssim 2^{-\min\{j,j'\}}$ guarantees that in each row and column of $\mathbf{S}_{j,j'}$ we have set at most $\mathcal{O}(2^{(j'-j)n})$, respectively, $\mathcal{O}(1)$ entries to zero. Therefore, we obtain for the weighted row sums

$$\begin{aligned} \sum_{k\in\nabla_j} 2^{\frac{(j-j')n}{2}} |s_{(j,k),(j',k')}| &\lesssim \sum_{k'\in\nabla_{j'}} (a')^{-(\tilde{d}+2q)} 2^{j'n} 2^{(j-j')n} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})} \\ &\lesssim (a')^{-(\tilde{d}+2q)} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})}, \end{aligned}$$

and likewise for the weighted column sums

$$\sum_{k'\in\nabla_{j'}} 2^{\frac{(j'-j)n}{2}} |s_{(j,k),(j',k')}| \lesssim (a')^{-(\tilde{d}+2q)} 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})}$$

for all $j_0 \leq j < j' < J$. In complete analogy to the proof of Theorem 8.1 we conclude the assertion. □

THEOREM 8.3 (a posteriori compression). *Let the matrix $\mathbf{A}_J^{\epsilon}$ be compressed according to Theorems* 8.1 *and* 8.2. *Then the a posteriori compression defined by* (7.4) *with the level-dependent threshold $\varepsilon_{j,j'}$ from* (7.5) *causes a block perturbation $\mathbf{T}_{j,j'} := \widetilde{\mathbf{A}}^{\epsilon}_{j,j'} - \mathbf{A}^{\epsilon}_{j,j'}$ satisfying*

$$\|\mathbf{T}_{j,j'}\|_2 \lesssim a'' 2^{2Jq} 2^{-2d'(J-\frac{j+j'}{2})}.$$

Proof. We organize the proof in four steps.

(i) Abbreviating $\mathbf{T}_{j,j'} := [t_{(j,k),(j',k')}]_{k\in\nabla_j,k'\in\nabla_{j'}}$, one obviously has

$$\left|t_{(j,k),(j',k')}\right| \leq a'' \min\left\{2^{-\frac{|j-j'|n}{2}}, 2^{-n(J-\frac{j+j'}{2})\frac{d'-q}{\tilde{d}+q}}\right\} 2^{2Jq}2^{-2d'(J-\frac{j+j'}{2})}. \tag{8.1}$$

We shall use the first compression in order to derive from this inequality the desired. To this end, we find in each row and column of $\mathbf{T}_{j,j'}$ only $\mathcal{O}([\mathcal{B}_{j,j'}2^{j'}]^n)$, respectively, $\mathcal{O}([\mathcal{B}_{j,j'}2^{j}]^n)$ nonzero entries. Setting $M := \frac{d'+\tilde{d}}{2(\tilde{d}+q)}$, one has

$$2^{\frac{2J(d'-q)-(j+j')(d'+\tilde{d})}{2(\tilde{d}+q)}} = 2^{-J}2^{\frac{(J-j)(d'+\tilde{d})}{2(\tilde{d}+q)}}2^{\frac{(J-j')(d'+\tilde{d})}{2(\tilde{d}+q)}} = 2^{-J}2^{(J-j)M}2^{(J-j')M}.$$

Hence, by (7.3), the cut-off parameter for the first compression takes the form

$$\mathcal{B}_{j,j'} \sim \max\left\{2^{-\min\{j,j,'\}}, 2^{-J}2^{(J-j)M}2^{(J-j')M}\right\}. \tag{8.2}$$

From (7.2) and $q < d - \frac{1}{2}$, one concludes $\frac{1}{2} < M < 1$. Moreover, we shall make use of the identity

$$2^{-n(J-\frac{j+j'}{2})\frac{d'-q}{\tilde{d}+q}} = 2^{-2n(J-\frac{j+j'}{2})(M-\frac{1}{2})}. \tag{8.3}$$

Without loss of generality, we assume in the following that $j' \geq j$.

(ii) With these preparations at hand we shall first estimate the block matrices $\mathbf{T}_{j,j'}$ with $2^{-\frac{(j'-j)n}{2}} \leq 2^{-2n(J-\frac{j+j'}{2})(M-\frac{1}{2})}$. One readily verifies that this relation is equivalent to $2^{-j} \geq 2^{-J}2^{(J-j)M}2^{(J-j')M}$, which, by (8.2), implies that the cut-off parameter satisfies $\mathcal{B}_{j,j'} \sim 2^{-j}$. Thus, from (8.1) one infers the estimate

$$\begin{aligned}\sum_{k'\in\nabla_{j'}} 2^{\frac{(j-j')n}{2}}|t_{(j,k),(j',k')}| &\lesssim a''2^{\frac{(j+j')n}{2}}2^{-jn}2^{-\frac{(j'-j)n}{2}}2^{2Jq}2^{-2d'(J-\frac{j+j'}{2})}\\ &= a''2^{2Jq}2^{-2d'(J-\frac{j+j'}{2})}\end{aligned}$$

for the weighted row sums of $\mathbf{T}_{j,j'}$. Analogously, one derives

$$\sum_{k\in\nabla_{j}} 2^{\frac{(j'-j)n}{2}}|t_{(j,k),(j',k')}| \lesssim a''2^{2Jq}2^{-2d'(J-\frac{j+j'}{2})}$$

for the weighted column sums.

(iii) We still have to estimate the errors in the remaining blocks, where $2^{-\frac{(j'-j)n}{2}} > 2^{-2n(J-\frac{j+j'}{2})(M-\frac{1}{2})}$. Then, by (8.3), the cut-off parameter is given by $B_{j,j'} \sim 2^{-J}2^{(J-j)M}2^{(J-j')M}$. Therefore, we obtain for the weighted row sums

$$\begin{aligned}&\sum_{k'\in\nabla_{j'}} 2^{\frac{(j-j')n}{2}}|t_{(j,k),(j',k')}|\\ &\quad\lesssim a''2^{\frac{(j+j')n}{2}}2^{-Jn}2^{(J-j)Mn}2^{(J-j')Mn}2^{-2n(J-\frac{j+j'}{2})(M-\frac{1}{2})}2^{2Jq}2^{-2d'(J-\frac{j+j'}{2})}\\ &\quad= a''2^{2Jq}2^{-2d'(J-\frac{j+j'}{2})},\end{aligned}$$

and a similar estimate for the weighted column sums.

(iv) Combining the estimates in steps (ii) and (iii), we conclude that

$$\sum_{k'\in\nabla_{j'}} 2^{\frac{(j-j')n}{2}}|t_{(j,k),(j',k')}|, \sum_{k\in\nabla_{j}} 2^{\frac{(j'-j)n}{2}}|t_{(j,k),(j',k')}| \lesssim a''2^{2Jq}2^{-2d'(J-\frac{j+j'}{2})}$$

for all $j_0 \leq j, j' < J$. The proof can now be completed in complete analogy to the proof of Theorem 8.1. □

9. Consistency. We shall establish next the consistency of the compressed scheme with the original operator equation in the corresponding Sobolev norms. To this end, note that the operator $\widetilde{A}_J^\epsilon : H^s(\Gamma) \to H^{s-2q}(\Gamma)$, $-\widetilde{\gamma} < s < \widetilde{\gamma} + 2q$, defined by

$$A_J^\epsilon u = \sum_{j,j'=j_0}^{J-1} \sum_{k,k'} \widetilde{\psi}_{j,k} [\widetilde{\mathbf{A}}_J^\epsilon]_{(j,k),(j',k')} \langle \widetilde{\psi}_{j,k}, u \rangle,$$

is represented by the compressed system matrix, since apparently $\langle \widetilde{A}_J^\epsilon \psi_{j',k'}, \psi_{j,k} \rangle = [\widetilde{\mathbf{A}}_J^\epsilon]_{(j,k),(j',k')}$.

THEOREM 9.1. *Let $d < \widetilde{d}+2q$ and $\widetilde{\mathbf{A}}_J^\epsilon$ be the compressed matrix, defined according to section* 7. *Then, for $q \le t, t' \le d$ the estimate*

$$|\langle (A - \widetilde{A}_J^\epsilon) Q_J u, Q_J v \rangle| \lesssim \epsilon 2^{J(2q-t-t')} \|u\|_t \|v\|_{t'} \tag{9.1}$$

holds uniformly with respect to J, where

$$\epsilon := a^{-2(\widetilde{d}+q)} + (a')^{-(\widetilde{d}+2q)} + a'', \tag{9.2}$$

and a, a', a'' are the constants from (7.2) *and* (7.5).

Proof. By definition of the block perturbation matrices $\mathbf{R}_{j,j'}$, $\mathbf{S}_{j,j'}$, $\mathbf{T}_{j,j'}$, one has

$$|\langle (A - \widetilde{A}_J^\epsilon) \psi_{j',k'}, \psi_{j,k} \rangle| \le \big| [\mathbf{R}_{j,j'} + \mathbf{S}_{j,j'} + \mathbf{T}_{j,j'}]_{k,k'} \big|.$$

Hence, we can estimate

$$\begin{aligned} (9.3) \quad & |\langle (A - \widetilde{A}_J^\epsilon) Q_J u, Q_J v \rangle| \\ &= \left| \sum_{j,j'=j_0}^{J-1} \langle (A - \widetilde{A}_J^\epsilon)(Q_{j'+1} - Q_{j'}) u, (Q_{j+1} - Q_j) v \rangle \right| \\ &\le \sum_{j,j'=j_0}^{J-1} |\langle (A - \widetilde{A}_J^\epsilon)(Q_{j'+1} - Q_{j'}) u, (Q_{j+1} - Q_j) v \rangle| \\ &= \sum_{j,j'=j_0}^{J-1} \left| \sum_{k \in \nabla_j} \sum_{k \in \nabla_{j'}} \langle (A - \widetilde{A}_J^\epsilon) \psi_{j',k'}, \psi_{j,k} \rangle \langle u, \widetilde{\psi}_{j',k'} \rangle \langle v, \widetilde{\psi}_{j,k} \rangle \right| \\ &\le \sum_{j,j'=j_0}^{J-1} \big\| \mathbf{R}_{j,j'} + \mathbf{S}_{j,j'} + \mathbf{T}_{j,j'} \big\|_2 \big\| [\langle u, \widetilde{\psi}_{j',k'} \rangle]_{k' \in \nabla_{j'}} \big\|_2 \big\| [\langle v, \widetilde{\psi}_{j,k} \rangle]_{k \in \nabla_j} \big\|_2. \end{aligned}$$

Invoking the inverse estimate (4.3) and the approximation property (4.2) yields

$$\begin{aligned} \big\| [\langle u, \widetilde{\psi}_{j',k'} \rangle]_{k' \in \nabla_{j'}} \big\|_2 &\sim \|(Q_{j'+1} - Q_{j'}) u\|_0 \lesssim 2^{-j't} \|u\|_t, \\ \big\| [\langle v, \widetilde{\psi}_{j,k} \rangle]_{k \in \nabla_j} \big\|_2 &\sim \|(Q_{j+1} - Q_j) v\|_0 \lesssim 2^{-jt'} \|v\|_{t'}. \end{aligned}$$

Further, from Theorems 8.1, 8.2, and 8.3, we conclude

$$\|\mathbf{R}_{j,j'} + \mathbf{S}_{j,j'} + \mathbf{T}_{j,j'}\|_2 \lesssim \epsilon 2^{2Jq} 2^{-2d'(J - \frac{j+j'}{2})}.$$

Inserting these estimates in (9.3) provides

$$\begin{aligned}|\langle(A-\widetilde{A}_J^\epsilon)Q_J u, Q_J v\rangle| &\lesssim \epsilon 2^{J(2q-t-t')}\|u\|_t\|v\|_{t'}\sum_{j,j'=j_0}^{J-1}2^{-j'(d'-t)}2^{-j(d'-t')}\\ &\lesssim \epsilon 2^{J(2q-t-t')}\|u\|_t\|v\|_{t'}\end{aligned}$$

since $t,t'\le d<d'$. □

10. Convergence. With the estimates of section 8 at hand we can prove that the proposed compression strategy retains the optimal order of convergence of the underlying Galerkin scheme. In this context, we shall encounter conditions on the parameters a,a',a'' defining ϵ in (9.2). From Theorem 9.1 we deduce

$$|\langle(A-\widetilde{A}_J^\epsilon)u_J,u_J\rangle|\le\varepsilon\|u_J\|_q^2,$$

which implies the V_J ellipticity. Indeed, inserting this result into (3.3), we get for $J>J_0$ that

$$|\langle(\widetilde{A}_J^\epsilon+\widetilde{A}_J^{\epsilon\star})u_J,u_J\rangle|\ge(c-2\varepsilon)\|u_J\|_q^2\gtrsim\|u_J\|_q^2,$$

with $c>0$, if ϵ from (9.2) is sufficiently small.

THEOREM 10.1 (stability). *Let ϵ from* (9.2) *be sufficiently small. Then, the matrix $\widetilde{\mathbf{A}}_J^\epsilon$, which arises by the compression according to* (7.1) *and* (7.4), *defines a stable scheme, i.e., $\|\widetilde{A}_J^\epsilon u_J\|_{-q}\sim\|u_J\|_q$, uniformly in $J>J_0$.*

This theorem is an immediate consequence of Lemma 3.1 and the norm equivalences, which already requires that $\widetilde{\gamma}>-q$. In the limit case $\widetilde{\gamma}=-q$ a more sophisticated proof presented in [26] shows that Theorem 10.1 remains valid.

THEOREM 10.2 (convergence). *Let ϵ from* (9.2) *be sufficiently small to ensure uniform stability of $\widetilde{A}_J^\epsilon$. Then, the solution $u_J=\sum_{j=j_0}^{J-1}\sum_{k\in\nabla_j}u_{j,k}\psi_{j,k}$ of the compressed scheme $\widetilde{\mathbf{A}}_J^\epsilon\mathbf{u}_J=\mathbf{f}_J$, where $\mathbf{u}_J=[u_{j,k}]_{j_0\le j<J,\,k\in\nabla_j}$, differs from the exact solution u, satisfying $Au=f$, in the energy norm only by*

$$\|u-u_J\|_q\lesssim 2^{J(q-d)}\|u\|_d$$

uniformly in J.

Proof. Strang's first lemma [4] provides

$$\|u-u_J\|_q\lesssim\inf_{v_J\in V_J}\left\{\|u-v_J\|_q+\sup_{w_J\in V_J}\frac{|\langle(A-\widetilde{A}_J^\epsilon)v_J,w_J\rangle|}{\|w_J\|_q}\right\}.$$

The consistency (Theorem 9.1) implies that

$$|\langle(A-\widetilde{A}_J^\epsilon)Q_Ju,w_J\rangle|=|\langle(A-\widetilde{A}_J^\epsilon)Q_Ju,Q_Jw_J\rangle\lesssim 2^{J(q-d)}\|u\|_d\|w_J\|_q$$

for all $u\in H^d(\Gamma)$ and $w_J\in V_J$. Hence, choosing $v_J:=Q_Ju$, we arrive at

$$\begin{aligned}\|u-u_J\|_q&\lesssim\|u-Q_Ju\|_q+\sup_{w_J\in V_J}\frac{|\langle(A-\widetilde{A}_J^\epsilon)Q_Ju,Q_Jw_J\rangle|}{\|w_J\|_q}\\&\lesssim 2^{J(q-d)}\|u\|_d.\end{aligned}$$ □

Theorem 10.3 (Aubin–Nitsche). *In addition to the assumptions of Theorem* 10.2 *suppose that* $\|A^\star v\|_{t-q} \sim \|v\|_{t+q}$ *for all* $0 \le t \le d-q$, *i.e.,* $A^\star : H^{t+q}(\Gamma) \to H^{t-q}(\Gamma)$ *is an isomorphism. Then the error estimate*

$$\|u - u_J\|_{q-t} \lesssim 2^{J(q-d-t)}\|u\|_d$$

holds for all $0 \le t \le d-q$.

Proof. Recalling that

$$\|u - u_J\|_{q-t} = \sup_{g \in H^{t-q}(\Gamma)} \frac{\langle u - u_J, g\rangle}{\|g\|_{t-q}}.$$

we obtain for $v \in H^{t+q}(\Gamma)$ with $A^\star v = g$

$$\|u - u_J\|_{q-t} = \sup_{v \in H^{t+q}(\Gamma)} \frac{|\langle A(u - u_J), v\rangle|}{\|v\|_{t+q}}.$$

Utilizing the Galerkin orthogonality $\langle \widetilde{A}_J^\epsilon u_J, Q_J v\rangle = \langle Au, Q_J v\rangle$, we can decompose

$$\begin{aligned}\langle A(u - u_J), v\rangle &= \langle A(u - u_J), v - Q_J v\rangle + \langle A(u - u_J), Q_J v\rangle \\ &= \langle A(u - u_J), v - Q_J v\rangle - \langle (A - \widetilde{A}_J^\epsilon)u_J, Q_J v\rangle.\end{aligned}$$

The first term on the right-hand side is estimated by Theorem 10.2 in combination with the approximation property (4.2),

$$|\langle A(u - u_J), v - Q_J v\rangle| \lesssim \|u - u_J\|_q \|v - Q_J v\|_q \lesssim 2^{J(q-d-t)}\|u\|_d\|v\|_{t+q}.$$

For the second term we obtain, on account of Theorem 9.1,

$$\begin{aligned}|\langle (A - \widetilde{A}_J^\epsilon)u_J, Q_J v\rangle| &\le |\langle (A - \widetilde{A}_J^\epsilon)(u_J - Q_J u), Q_J v\rangle| + |\langle (A - \widetilde{A}_J^\epsilon)Q_J u, Q_J v\rangle| \\ &\lesssim 2^{-Jt}\|u_J - Q_J u\|_q \|v\|_{t+q} + 2^{J(q-d-t)}\|u\|_d\|v\|_{t+q}.\end{aligned}$$

Inserting $\|u_J - Q_J u\|_q \le \|u - u_J\|_q + \|u - Q_J u\|_q \lesssim 2^{J(q-d)}\|u\|_d$ yields

$$|\langle (A - \widetilde{A}_J^\epsilon)u_J, Q_J v\rangle| \lesssim 2^{J(q-d-t)}\|u\|_d\|v\|_{t+q}.$$

Therefore, we conclude

$$\|u - u_J\|_{q-t} = \sup_{v \in H^{t+q}(\Gamma)} \frac{\langle A(u - u_J), v\rangle}{\|v\|_{t+q}} \lesssim 2^{J(q-d-t)}\|u\|_d,$$

which finishes the proof. ▯

Note that in the extreme case $t = d - q$ we obtain the best possible convergence rate of the Galerkin scheme (3.6).

11. Complexity.

In this section, we present a general theorem which shows that the overall complexity of assembling the compressed system matrix with sufficient accuracy can be kept of the order $\mathcal{O}(N_J)$, even when a computational cost of logarithmic order is allowed for each entry. This theorem is used in [19, 22] as the essential ingredient to provide a quadrature strategy which scales linearly.

Theorem 11.1. *Assume that* $\mathbf{A}_J^\epsilon$ *is obtained by compressing the system matrix* $\mathbf{A}_J = [\langle A\psi_{j',k'}, \psi_{j,k}\rangle]_{j_0 \le j,j' < J,\ k \in \nabla_j,\ k' \in \nabla_{j'}}$ *according to* (7.1). *The complexity of* *computing this compressed matrix is* $\mathcal{O}(N_J)$ *provided that for some* $\alpha \geq 0$ *at most* $\mathcal{O}\big(\big[J - \frac{j+j'}{2}\big]^{\alpha}\big)$ *operations are spent on the approximate calculation of the nonvanishing entries* $\langle A\psi_{j',k'}, \psi_{j,k}\rangle$.

Proof. (i) We begin with some technical preparations. Recall from the proof of Theorem 8.3 that the cut-off parameter of the first compression is given by

$$\mathcal{B}_{j,j'} \sim \max\big\{2^{-\min\{j,j,'\}}, 2^{-J}2^{(J-j)M}2^{(J-j')M}\big\},$$

where, as in the proof of Theorem 8.3, $M = \frac{d'+\widetilde{d}}{2(\widetilde{d}+q)} < 1$. Moreover, we set $M' := \frac{2d'-2q}{\widetilde{d}+d'}$ and $N' := \frac{\widetilde{d}+d'}{\widetilde{d}+2q}$ with d' given by (7.2). Note that M' and N' satisfy the relations $0 < M' < 1$ and $0 < N'$. As one readily verifies, the cut-off parameter with respect to the second compression may now be rewritten as

$$\mathcal{B}'_{j,j'} \sim \max\big\{2^{-j}, 2^{-j'}2^{[JM'+(1-M')j'-j]N'}\big\}, \qquad j \geq j'. \tag{11.1}$$

Further, we make use of the inequality $x^{\alpha} \lesssim 2^{2\delta x}$ which holds for all $x > 0$ and any fixed $\alpha, \delta > 0$. Thus, it suffices to prove the claim for $\mathcal{O}\big(\big[J - \frac{j+j'}{2}\big]^{\alpha}\big)$ replaced by $\mathcal{O}\big(2^{\delta(J-j)}2^{\delta(J-j')}\big)$ where δ is chosen sufficiently small.

(ii) First, we determine now the complexity $\mathcal{C}^{(1)}$ of computing, within the above cost allowance, all matrix entries found in the block matrices $\mathbf{A}^{\epsilon}_{j,j'} = [\mathbf{A}^{\epsilon}_{J}]_{(j,\nabla_j),(j',\nabla_{j'})}$ with $\mathcal{B}_{j,j'} \sim 2^{-J}2^{(J-j)M}2^{(J-j')M}$. In such blocks, we have to process all coefficients $\langle A\psi_{j',k'}, \psi_{j,k}\rangle$ with

$$\operatorname{dist}(\Omega_{j,k}, \Omega_{j',k'}) \lesssim \operatorname{dist}^{(1)}_{j,j'} := 2^{-J}2^{(J-j)M}2^{(J-j')M}. \tag{11.2}$$

In each block, we find only $\mathcal{O}\big(\big[2^{j'}\operatorname{dist}^{(1)}_{j,j'}\big]^{n}\big)$ entries satisfying (11.2) per row, and hence a total of $\mathcal{O}\big(\big[2^{j+j'}\operatorname{dist}^{(1)}_{j,j'}\big]^{n}\big)$. Summing over all blocks yields

$$\begin{aligned}\mathcal{C}^{(1)} &\lesssim \sum_{j,j'=0}^{J} 2^{(j+j')n}2^{-Jn}2^{(J-j)(M+\delta)n}2^{(J-j')(M+\delta)n}\\ &= 2^{Jn}\sum_{j,j'=0}^{J} 2^{(J-j)(M+\delta-1)n}2^{(J-j')(M+\delta-1)n} \lesssim 2^{Jn},\end{aligned}$$

provided that δ is chosen so as to ensure $M + \delta < 1$.

(iii) It remains to show that the complexity for computing the omitted blocks is likewise $\mathcal{O}(N_J)$. Without loss of generality, we assume $j \geq j'$ in the remainder of this proof, since the roles of j and j' can be reversed. Observing that, because of $0 < M' < 1$, one has $0 < JM' + (1-M')j' \leq J$, we consider first the blocks $\mathbf{A}^{\epsilon}_{j,j'}$ with $(j,j') \in S$, where the index set S is given by

$$S := \{(j,j') : 0 \leq j' \leq J,\ JM' + (1-M')j' \leq j \leq J\}. \tag{11.3}$$

In these blocks, we estimate the complexity $\mathcal{C}^{(2)}$ required for the approximate computation of the matrix entries $\langle A\psi_{j',k'}, \psi_{j,k}\rangle$ satisfying the relation

$$\operatorname{dist}(\Omega'_{j,k}, \Omega_{j',k'}) \lesssim \operatorname{dist}^{(2)}_{j,j'} := 2^{-j'}2^{[JM'+(1-M')j'-j]N'}, \tag{11.4}$$

where we refer to expression (11.1) for $\mathcal{B}'_{j,j'}$. Since $\operatorname{dist}^{(2)}_{j,j'} \leq 2^{-j'}$ for all $(j,j') \in S$, in each block one finds only $\mathcal{O}([2^{jn}2^{-j'(n-1)}\operatorname{dist}^{(2)}_{j,j'}])$ nontrivial matrix entries per column with (11.4), and thus a total of $\mathcal{O}([2^{jn}2^{j'}\operatorname{dist}^{(2)}_{j,j'}])$. Therefore, noting that the set S is equivalent to $S = \{(j,j') : JM' \le j \le J,\ 0 \le j' \le \frac{j-JM'}{1-M'}\}$, the complexity is bounded by

$$\begin{aligned}\mathcal{C}^{(2)} &\lesssim \sum_{j=JM'}^{J} \sum_{j'=0}^{\frac{j-JM'}{1-M'}} 2^{jn} 2^{[JM'+(1-M')j'-j]N'} 2^{\delta(J-j)} 2^{\delta(J-j')} \\ &= \sum_{j=JM'}^{J} 2^{jn} 2^{[JM'-j]N'} 2^{\delta(J-j)} 2^{\delta J} \sum_{j'=0}^{\frac{j-JM'}{1-M'}} 2^{j'[(1-M')N'-\delta]} \\ &\lesssim 2^{\delta J \frac{2-M'}{1-M'}} \sum_{j=0}^{J} 2^{j(n-\delta\frac{2-M'}{1-M'})} \lesssim 2^{Jn}.\end{aligned}$$

$\mathcal{C}^{(2)}$ estimates the complexity for those blocks with $(j,j') \in S$ when $\mathcal{B}'_{j,j'} \sim \operatorname{dist}^{(2)}_{j,j'}$. But according to (11.1), the cut-off parameter $\mathcal{B}'_{j,j'}$ is bounded from below by 2^{-j}. In the case of $\mathcal{B}'_{j,j'} \sim 2^{-j}$ we find $\mathcal{O}\big(\big[2^{jn}2^{j'}\operatorname{dist}^{(3)}_{j,j'}\big]\big)$ matrix entries $\langle A\psi_{j',k'}, \psi_{j,k}\rangle$ with $\operatorname{dist}(\Omega'_{j,k}, \Omega_{j',k'}) \lesssim \operatorname{dist}^{(3)}_{j,j'} := 2^{-j}$. Arguing analogously as above, summing over all blocks with $(j,j') \in S$, one obtains

$$\begin{aligned}\mathcal{C}^{(3)} &\lesssim \sum_{j=JM'}^{J} \sum_{j'=0}^{\frac{j-JM'}{1-M'}} 2^{j(n-1)} 2^{j'} 2^{\delta(J-j)} 2^{\delta(J-j')} = \sum_{j=JM'}^{J} 2^{j(n-1)} 2^{\delta(2J-j)} \sum_{j'=0}^{\frac{j-JM'}{1-M'}} 2^{j'(1-\delta)} \\ &\lesssim 2^{\delta J \frac{2-M'}{1-M'}} \sum_{j=0}^{J} 2^{j(n-\delta\frac{2-M'}{1-M'})} \lesssim 2^{Jn}.\end{aligned}$$

(iv) Finally, we consider the blocks $\mathbf{A}^{\epsilon}_{j,j'}$ with $j \ge j'$ and $(j,j') \notin S$. In view of step (ii), it suffices to consider all entries $\langle A\psi_{j',k'}, \psi_{j,k}\rangle$ which fulfill

$$\operatorname{dist}(\Omega_{j,k}, \Omega_{j',k'}) \lesssim \operatorname{dist}^{(4)} := 2^{-\min\{j',j\}} = 2^{-j'}. \tag{11.5}$$

Each block $\mathbf{A}^{\epsilon}_{j,j'}$ consists of only $\mathcal{O}\big(\big[2^{j}2^{j'}\operatorname{dist}^{(4)}\big]^{n}\big)$ entries with (11.5). Hence, according to (11.3), the complexity $\mathcal{C}^{(4)}$ for the computation of these entries is

$$\begin{aligned}\mathcal{C}^{(4)} &\lesssim \sum_{j'=0}^{J} \sum_{j=j'}^{JM'+(1-M')j'} 2^{jn} 2^{\delta(J-j)} 2^{\delta(J-j')} \lesssim \sum_{j'=0}^{J} 2^{2\delta(J-j')} 2^{j'n} \sum_{j=0}^{(J-j')M'} 2^{j(n-\delta)} \\ &\lesssim 2^{Jn} \sum_{j'=0}^{J} 2^{(J-j')((M'-1)(n-\delta)+\delta)} \lesssim 2^{Jn},\end{aligned}$$

since $(M'-1)(n-\delta)+\delta < 0$. This completes this proof. □

REFERENCES

[1] M. Bebendorf and S. Rjasanow, *Adaptive low-rank approximation of collocation matrices*, Computing, 70 (2003), pp. 1–24.

[2] G. Beylkin, R. Coifman, and V. Rokhlin, *The fast wavelet transform and numerical algorithms*, Comm. Pure Appl. Math., 44 (1991), pp. 141–183.

[3] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method, part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 1–52.

[4] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[5] A. Cohen, I. Daubechies, and J.-C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Pure Appl. Math., 45 (1992), pp. 485–560.

[6] A. Cohen and R. Masson, *Wavelet adaptive method for second order elliptic problems—boundary conditions and domain decomposition*, Numer. Math., 86 (2000), pp. 193–238.

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

[8] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228.

[9] W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Numer. Math., 63 (1992), pp. 315–344.

[10] W. Dahmen, A. Kunoth, and K. Urban, *Biorthogonal spline-wavelets on the interval—stability and moment conditions*, Appl. Comp. Harmon. Anal., 6 (1999), pp. 259–302.

[11] W. Dahmen, S. Prössdorf, and R. Schneider, *Wavelet approximation methods for periodic pseudodifferential equations. Part* II. *Fast solution and matrix compression*, Adv. Comput. Math., 1 (1993), pp. 259–335.

[12] W. Dahmen, S. Prössdorf, and R. Schneider, *Multiscale methods for pseudodifferential equations on smooth manifolds*, in Wavelets: Theory, Algorithms, and Applications (Taormina, 1993), Wavelet Anal. Appl. 5, Academic Press, San Diego, CA, 1994, pp. 385–424.

[13] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comput., 68 (1999), pp. 1533–1567.

[14] W. Dahmen and R. Schneider, *Wavelets on manifolds* I. *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.

[15] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.

[16] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulation*, J. Comput. Phys., 73 (1987), pp. 325–348.

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

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

[19] H. Harbrecht, *Wavelet Galerkin Schemes for the Boundary Element Method in Three Dimensions*, Ph.D. thesis, Technische Universität Chemnitz, Germany, 2001.

[20] H. Harbrecht, M. Konik, and R. Schneider, *Fully discrete wavelet Galerkin schemes*, Eng. Anal. Bound. Elem., 27 (2003), pp. 423–437.

[21] H. Harbrecht and R. Schneider, *Biorthogonal wavelet bases for the boundary element method*, Math. Nachr., 167–188 (2004), pp. 269–270.

[22] H. Harbrecht and R. Schneider, *Wavelet Galerkin schemes for boundary integral equations—implementation and quadrature*, SIAM J. Sci. Comput., to appear.

[23] Y. Meyer, *Ondelettes et Opérateurs* 2: *Opérateur de Caldéron-Zygmund*, Hermann, Paris, 1990.

[24] T. von Petersdorff, R. Schneider, and C. Schwab, *Multiwavelets for second kind integral equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2212–2227.

[25] T. von Petersdorff and C. Schwab., *Fully discretized multiscale Galerkin BEM*, in Multiscale Wavelet Methods for PDEs, W. Dahmen, A. Kurdila, and P. Oswald, eds., Academic Press, San Diego, CA, 1997, pp. 287–346.

[26] A. Rathsfeld, *A quadrature algorithm for wavelet Galerkin methods*, in Über Waveletalgorithmen für die Randelementmethode, Habilitation thesis, Technische Universität Chemnitz, Germany, 2001.

[27] R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur Lösung großer vollbesetzter Gleichungssysteme*, B.G. Teubner, Stuttgart, 1998.

[28] E.M. Stein, *Harmonic Analysis*, Princeton University Press, Princeton, NJ, 2002.

[29] M.E. Taylor, *Pseudodifferential Operators*, Princeton University Press, Princeton, NJ, 1981.

[30] S. Goreinov, E. Tyrtischnikov, and Y. Yeremin, *Matrix-free iterative solution strategies for large dense linear systems*, Numer. Linear Algebra Appl., 4 (1997), pp. 273–294.

[31] W. Wendland, *Boundary element methods and their asymptotic convergence*, in Theoretical Acoustics and Numerical Techniques, CISM Courses and Lectures 277, P. Filippi, ed., Springer, New York, 1983, pp. 135–216.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2272–2293

ARTIFICIAL BOUNDARY CONDITIONS FOR ONE-DIMENSIONAL CUBIC NONLINEAR SCHRÖDINGER EQUATIONS[*]

XAVIER ANTOINE[†], CHRISTOPHE BESSE[†], AND STÉPHANE DESCOMBES[‡]

Abstract. This paper addresses the construction of nonlinear integro-differential artificial boundary conditions for one-dimensional nonlinear cubic Schrödinger equations. Several ways of designing such conditions are provided and a theoretical classification of their accuracy is given. Semidiscrete time schemes based on the method developed by Durán and Sanz-Serna [*IMA J. Numer. Anal.* 20 (2000), pp. 235–261] are derived for these unusual boundary conditions. Stability results are stated and several numerical tests are performed to analyze the capacity of the proposed approach.

Key words. nonlinear cubic Schrödinger equation, artificial boundary conditions, pseudodifferential operators, stable semidiscrete schemes, solitons interaction

AMS subject classifications. 35Q55, 35Q51, 47G30, 26A33, 65M12

DOI. 10.1137/040606983

1. Introduction.

In many physical and technological domains of interest, the numerical solution to a one-dimensional cubic nonlinear Schrödinger (NLS) equation of the form

$$
\begin{aligned}
i\partial_t u + \partial_x^2 u + q|u|^2 u = 0,\ x \in \mathbb{R},\ t > 0, \\
u(x,0) = u_0(x),\ x \in \mathbb{R},
\end{aligned}
\tag{1.1}
$$

is required. The real parameter q corresponds to a focusing ($q > 0$) or defocusing ($q < 0$) effect of the cubic nonlinearity. One example of such an equation is given in nonlinear optic for laser beam propagation where the polarization of the material has a cubic nonlinearity according to the electric field. Using the slowly varying envelope assumption of the electric field and several approximations, it can be shown that the problem reduces to (1.1). The solution is then the unknown amplitude of the electric field. Other applications come from plasma physics or quantum mechanics [11].

The numerical treatment of (1.1) is often realized by restricting the computational domain to a finite one. More precisely, let us assume that the initial datum u_0 is compactly supported in a finite domain $\Omega_i =]x_l, x_r[\subset \mathbb{R}$, with $x_r > x_l$. The Dirichlet boundary condition is usually imposed on the boundary $\Sigma = \{x_l, x_r\}$ of Ω_i. However, the wave reflects back into the computational domain when u strikes Σ. If the Neumann boundary condition is preferred, then reflection still occurs. The problem of the choice of a suitable boundary condition is linked to the construction of a nonreflecting (also called transparent) boundary condition which models the propagation of the solution into the complementary unbounded domain $\mathbb{R}/\overline{\Omega_i}$.

[*] Received by the editors April 19, 2004; accepted for publication (in revised form) April 22, 2005; published electronically January 6, 2006. This work was supported by the TMR HYKE (EU financed network HPRN-CT-2002-00282) and the GDR CNRS 2103 project.

http://www.siam.org/journals/sinum/43-6/60698.html

[†] Laboratoire de Mathématiques pour l'Industrie et la Physique, CNRS UMR 5640, Université Paul Sabatier, 118 Route de Narbonne, 31062 Toulouse Cedex 4, France (antoine@mip.ups-tlse.fr-besse@mip.ups-tlse.fr).

[‡] Unité de Mathématiques Pures et Appliquées, CNRS UMR 5669, Ecole Normale Supérieure de Lyon, 46, Allée d'Italie, 69364 Lyon Cedex 07, France (sdescomb@umpa.ens-lyon.fr).

Such conditions have been widely studied in the linear case ($q = 0$) where the exact boundary condition is given through the Dirichlet–Neumann (DN) operator [8]

$$\partial_{\mathbf{n}} u + e^{-i\frac{\pi}{4}} \partial_t^{1/2} u = 0 \quad \text{on } \Sigma \times \mathbb{R}^{*+}. \tag{1.2}$$

The operator $\partial_t^{1/2}$ designates the fractional time derivative operator of half-order given by the Riemann–Liouville nonlocal integral representation

$$\partial_t^{1/2} f(t) = \frac{1}{\sqrt{\pi}} \partial_t \int_0^t \frac{f(s)}{\sqrt{t-s}} ds. \tag{1.3}$$

Another equivalent continuous representation is given by the Neumann–Dirichlet (ND) operator

$$u + e^{i\frac{\pi}{4}} I_t^{1/2} \partial_{\mathbf{n}} u = 0 \quad \text{on } \Sigma \times \mathbb{R}^{*+},$$

where the fractional integral operator of half-order is given by the convolution operator

$$I_t^{1/2} f(t) = \frac{1}{\sqrt{\pi}} \int_0^t \frac{f(s)}{\sqrt{t-s}} ds.$$

Even if the above two boundary conditions are transparent in the linear case, it appears that they still generate a lot of reflection at the boundary when the nonlinear perturbation is added [20]. This is finally quite natural since the nonlinear term generally compensates the dispersion due to the linear part of the NLS equation. Therefore, we cannot expect that a linear boundary condition simulates the nonlinear phenomenon. To the best of the authors' knowledge, no nonlinear artificial boundary conditions (NLABCs) have been derived and studied until now for a nonlocal NLS equation. The goal of this paper is to propose some efficient NLABCs for the model problem (1.1) (section 2), to construct some suitable and (if possible) stable schemes for their discretization (sections 3 and 4), and finally to numerically test them (section 4) in some interesting situations (e.g., interaction of two solitons).

2. A first construction of NLABCs for the cubic NLS equation.

This section is devoted to the construction of NLABCs for the one-dimensional cubic NLS equation. Two ways of designing such conditions are proposed and energy bounds are derived.

2.1. Construction of NLABCs.

The adopted strategy for constructing some artificial boundary conditions for the NLS equation is issued from linear analysis. It consists first in designing some suitable artificial boundary conditions for the linear Schrödinger equation being given a potential V

$$i\partial_t u + \partial_x^2 u + Vu = 0, \tag{2.1}$$

and next in making the formal substitution $V(x,t) = q|u(x,t)|^2$ to deduce some NLABCs.

Let us assume that V is a sufficiently smooth potential. First, we consider a time-dependent potential, V: $V(x,t) = V(t)$. Then, considering the new unknown $v(x,t) = e^{-i\mathcal{V}(t)} u(x,t)$ in (2.1), where the phase function $\mathcal{V}$ is given by

$$\mathcal{V}(t) = \int_0^t V(s) ds,$$

we reduce the initial equation to a linear Schrödinger equation without potential,

$$i\partial_t v + \partial_x^2 v = 0, \tag{2.2}$$

where $v(x,0) = u_0(x)$. Now, we assume that the initial datum u_0 is compactly supported in an open bounded computational domain $\Omega_i =]x_l, x_r[$ of boundary $\Sigma = \{x_l, x_r\}$. We can therefore directly show that the DN transparent boundary condition for this latter equation is [3, 6, 16]

$$\partial_{\mathbf{n}} v(x,t) + e^{-i\frac{\pi}{4}} \partial_t^{1/2} v(x,t) = 0 \qquad \text{on } \Sigma \times \mathbb{R}^+,$$

which can be rewritten according to the initial unknown u as the *exact* boundary condition

$$\partial_{\mathbf{n}} u(x,t) + e^{-i\frac{\pi}{4}} e^{i\mathcal{V}(t)} \partial_t^{1/2} (e^{-i\mathcal{V}(t)} u(x,t)) = 0 \qquad \text{on } \Sigma \times \mathbb{R}^+. \tag{2.3}$$

The operator $\partial_t^{1/2}$ stands for the Riemann–Liouville fractional derivative operator of order 1/2 defined by (1.3). Following the proposed approach, we consider the formal NLABC

$$\partial_{\mathbf{n}} u(x,t) + e^{-i\frac{\pi}{4}} e^{i\mathbb{U}(x,t)} \partial_t^{1/2} (e^{-i\mathbb{U}(x,t)} u(x,t)) = 0 \qquad \text{on } \Sigma \times \mathbb{R}^+, \tag{2.4}$$

with

$$\mathbb{U}(x,t) = q \int_0^t |u(x,s)|^2 ds. \tag{2.5}$$

The boundary condition (2.4)–(2.5) is denoted by NLABC_1^1 in the rest of the paper.

We have seen that we can explicitly write the transparent boundary condition for a potential which depends only on the time variable. Now a question is: How can we extend this approach to the case of a potential which also depends on the spatial variable x? To give a possible answer to this problem, let us denote by u the solution to (2.1) and by v the new unknown defined by the relation $v(x,t) = e^{-i\mathcal{V}(x,t)} u(x,t)$. We straightforwardly remark that $v_0(x) = u_0(x)$ and then the initial solutions coincide. Moreover, the time and spatial derivatives of u are given in terms of derivatives of v by

$$i\partial_t u = e^{i\mathcal{V}} (i\partial_t - V) v$$

and

$$\partial_x^2 u = ie^{i\mathcal{V}} (\partial_x^2 v + 2i\partial_x \mathcal{V} \partial_x v + i\partial_x^2 \mathcal{V} v - (\partial_x \mathcal{V})^2 v).$$

As a consequence, the function v satisfies the Schrödinger equation

$$L(x,t,\partial_x,\partial_t) v = i\partial_t v + \partial_x^2 v + A\partial_x v + Bv = 0, \tag{2.6}$$

where we have defined the two functions A and B by the relations $A = 2i\partial_x \mathcal{V}$ and $B = (i\partial_x^2 \mathcal{V} - (\partial_x \mathcal{V})^2)$. For a potential which depends only on the time variable t, the operators A and B vanish and the approach coincides with the previous case. However, in the general situation, (2.6) is a variable coefficients Schrödinger equation. For this reason, we propose to develop a constructive approach of the artificial boundary conditions based on the theory of pseudodifferential operators as proposed by Engquist and Majda [14, 15] in the seventies. Many authors have extended and improved this technique to various equations and systems. In the particular case of the Schrödinger equation, a fractional pseudodifferential operators calculus [2] is required to include the inhomogeneity appearing between the time and space derivatives. Using a Nirenberg-like factorization theorem [2, 18] for the Schrödinger operator L, we can compute an asymptotic expansion in inhomogeneous symbols of the transparent operator for (2.6).

The inhomogeneous pseudodifferential operator calculus used here has been introduced by Lascar [17]. We just give the main results adapted to our situation and refer to [17] for further details. Let us consider a real number α and Ω an open subset of the space $\mathbb{R}^n$. Then (see, e.g., [18]), $S^\alpha(\Omega\times\Omega)$ denotes the linear space of $\mathcal{C}^\infty$ functions $a(x,t,\tau)$ in $\Omega\times\Omega\times\mathbb{R}^n$ such that for each $K\subseteq\Omega\times\Omega$ and for any multiindices β, δ, γ there exists a constant $C_{\beta,\delta,\gamma}(K)$ such that we get

$$|\partial_\tau^\beta\partial_t^\delta\partial_x^\gamma a(x,t,\tau)| \leq C_{\beta,\delta,\gamma}(K)(1+|\tau|^2)^{\alpha-|\beta|}$$

for all $(x,t)\in K$ and $\tau\in\mathbb{R}^n$. Here, $|\beta|$ denotes the length of a multi-index β and τ is the time Fourier symbol. Let us introduce now some specific notation and definitions for our problem (setting $n=1$). A function f is said to be inhomogeneous of degree m if it fulfills $f(x,t,\mu^2\tau)=\mu^m f(x,t,\tau)$ for any $\mu>0$. Then, a pseudodifferential operator $P=P(x,t,\partial_t)$ is called inhomogeneous and classical of order M, $M\in\mathbb{Z}/2$, if its total symbol, denoted by $p=\sigma(P)$, admits an asymptotic expansion in inhomogeneous symbols $\{p_{M-j/2}\}_{j=0}^{+\infty}$ under the form

$$p(x,t,\tau)\sim\sum_{j=0}^{+\infty}p_{M-j/2}(x,t,\tau),$$

where functions $p_{M-j/2}$ are inhomogeneous of degree $2M-j$ for $j\in\mathbb{N}$. The sense to give to this last approximation is that

$$\forall\widetilde{m}\in\mathbb{N},\quad p-\sum_{j=0}^{\widetilde{m}}p_{M-j/2}\in S^{M-(\widetilde{m}+1)/2}.$$

A symbol p satisfying the above property is quoted by $p\in S_S^M$ and the associated operator $P=Op(p)$ by inverse Fourier transform by $P\in OPS_S^M$. For instance, if we consider the fractional derivative operator $P=e^{-i\frac{\pi}{4}}\partial_t^{1/2}$, a direct calculation shows that its symbol is equal to $\sqrt{\tau}$, where τ is the time covariable. Notation $\sqrt{z}$ designates the standard principal determination of the complex square root of a complex number z for a branch-cut along the negative real axis. Function $\sqrt{\tau}$ is inhomogeneous of degree 1 and is an element of $S_S^{1/2}$. So, P is a pseudodifferential operator of $OPS_S^{1/2}$. Similarly, the fractional integration operators $i^{-\alpha/2}I_t^{\alpha/2}$ defined by the relations

$$I_t^{\alpha/2}f(t)=\frac{1}{\Gamma(\alpha/2)}\int_0^t(t-s)^{\alpha/2-1}f(s)ds\quad\text{for }\alpha\in\mathbb{N}$$

have $\tau^{-\alpha/2}$ as symbol, where Γ stands for the Gamma function. This symbol is inhomogeneous of degree $-\alpha$ and is an element of $S_S^{-\alpha/2}$. Therefore, $i^{-\alpha/2}I_t^{\alpha/2}$ defines a pseudodifferential operator of $OPS_S^{-\alpha/2}$.

Under the previous notation, the following proposition holds.

PROPOSITION 2.1. *Let L be the variable coefficients Schrödinger operator defined by* (2.6). *There exist two inhomogeneous and classical pseudodifferential operators $\Lambda^{\pm} = \Lambda^{\pm}(x, t, \partial_t) \in OPS_S^{1/2}$, regular with respect to the spatial variable x and such that*

$$L = (\partial_x + i\Lambda^-)(\partial_x + i\Lambda^+) + R, \tag{2.7}$$

where R is a smoothing operator of $OPS_S^{-\infty}$ and the principal symbols $\lambda^{\pm}_{1/2}$ of the operators $\Lambda^{\pm}$ are given by $\lambda^{\pm}_{1/2} = \mp\sqrt{-\tau}$. Furthermore, the total symbol $\lambda^{\pm} = \sigma(\Lambda^{\pm})$ of $\Lambda^{\pm}$ admits an asymptotic expansion in inhomogeneous symbols as

$$\lambda^{\pm} = \sigma(\Lambda^{\pm}) \sim \sum_{j=0}^{+\infty} \lambda^{\pm}_{1/2-j/2}. \tag{2.8}$$

Proof. Expanding the factorization (2.7) (of Nirenberg-type [18]) and using similar calculations to those in [1, 2], we get

$$(\partial_x + i\Lambda^-)(\partial_x + i\Lambda^+) = \partial_x^2 + i(\Lambda^+ + \Lambda^-)\partial_x + iOp(\partial_x\lambda^+) - \Lambda^-\Lambda^+.$$

By identification with the terms appearing in front of the spatial derivatives ∂_x in the expression (2.6) of L, we deduce the system of operators

$$\begin{aligned} i(\Lambda^+ + \Lambda^-) &= A, \\ iOp(\partial_x\lambda^+) - \Lambda^-\Lambda^+ &= i\partial_t + B, \end{aligned} \tag{2.9}$$

which yields the symbolic system of equations

$$\begin{aligned} li(\lambda^+ + \lambda^-) &= a, \\ i\partial_x\lambda^+ - \sum_{\alpha=0}^{+\infty} \frac{(-i)^\alpha}{\alpha!}\partial_\tau^\alpha\lambda^-\partial_t^\alpha\lambda^+ &= -\tau + b, \end{aligned} \tag{2.10}$$

setting $a = A$ and $b = B$. These two functions correspond to zero order operators. If we identify the terms of order 1/2 in the first relation of system (2.10), we obtain $\lambda^-_{1/2} = -\lambda^+_{1/2}$. Using now the second equation, we deduce that

$$\lambda^+_{1/2} = \pm\sqrt{-\tau}.$$

For a potential which depends only on the time variable, the DN operator corresponds to the choice $\lambda^+_{1/2} = -\sqrt{-\tau}$ which can be extended to the space-dependent case. If we consider now the next order of identification, the first equation of the symbolic system gives the zero order term $\lambda^-_0 = -\lambda^+_0 - ia$. Substituting this expression into the second equation of system (2.10), we get

$$i\partial_x\lambda^+_{1/2} - (\lambda^-_0\lambda^+_{1/2} + \lambda^+_0\lambda^-_{1/2}) = 0.$$

But since $\partial_x\lambda^+_{1/2} = 0$, the previous relation yields

$$\lambda^+_0 = -i\frac{a}{2}.$$

Let us now consider the next term. Developing the computations and using the fact that $\lambda^+_{-1/2} = -\lambda^-_{-1/2}$, the zero order identification gives

$$i\partial_x\lambda_0^+ - (\lambda^-_{-1/2}\lambda^+_{1/2} + \lambda_0^-\lambda_0^+ + \lambda^+_{-1/2}\lambda^-_{1/2}) = b$$

since $\partial_t^\alpha\lambda^\pm_{-1/2} = 0$ and $\partial_\tau^\alpha\lambda_0^\pm = 0$ for any $\alpha \in \mathbb{N}$. After some simplifications, we conclude that

$$\lambda^+_{-1/2} = \frac{1}{2\lambda^+_{1/2}}(b - i\partial_x\lambda_0^+ + (\lambda_0^+)^2 + ia\lambda_0^+),$$

with $b = i\partial_x^2\mathcal{V} - (\partial_x\mathcal{V})^2$. An explicit computation yields $\lambda^+_{-1/2} = 0$. Following the process initialized above, we obtain that the next symbol is given by

$$\lambda^+_{-1} = i\frac{\partial_x V}{4\tau}.$$

The computation of the four first symbols shows how to construct the symbolic asymptotic expansion. Using the same inductive arguments as in [2], we end the proof of the proposition. □

The factorization (2.7) yields a splitting of the Schrödinger operator into two parts: one corresponding to an outgoing wave to the computational domain and another one yielding the reflected part of the wave. Then we can show that a condition for having a vanishing reflected wave is given by the DN transparent boundary condition applied to the unknown field v

$$(\partial_{\mathbf{n}} + i\Lambda^+)v = 0 \quad \text{on } \Sigma \times \mathbb{R}^*.$$

Since the operator Λ^+ has an infinite expansion in inhomogeneous symbols, we choose to approximate the above condition by retaining the M first terms yielding the following definition.

DEFINITION 2.2. *Let M be a nonnegative integer. We consider the operator $\Lambda^+_{M/2}$ approximating $i\Lambda^+$ in $OPS_S^{1/2-M/2}$ and given by*

$$\Lambda^+_{M/2} = iOp\left(\sum_{j=0}^{M-1}\lambda^+_{1/2-j/2}\right).$$

Then, the artificial boundary condition of order $M/2$ (acting on u) is defined by

$$\partial_{\mathbf{n}}u + e^{i\mathcal{V}}\Lambda^+_{M/2}(e^{-i\mathcal{V}}u) = 0 \quad \text{on } \Sigma \times \mathbb{R}^*,$$

where we have set

$$\mathcal{V}(x,t) = \int_0^t V(x,s)ds.$$

Using the computation of the symbols obtained during the proof of Proposition 2.1, we show that the first order boundary condition is given by

$$\partial_{\mathbf{n}}u + e^{-i\frac{\pi}{4}}e^{i\mathcal{V}}\partial_t^{1/2}(e^{-i\mathcal{V}}u) = 0 \ \text{ on } \Sigma \times \mathbb{R}^*, \tag{2.11}$$

and the second order one by

$$\partial_{\mathbf{n}} u + e^{-i\frac{\pi}{4}} e^{i\mathcal{V}} \partial_t^{1/2}(e^{-i\mathcal{V}} u) + i\frac{\partial_x V}{4} e^{i\mathcal{V}} I_t(e^{-i\mathcal{V}} u) = 0 \text{ on } \Sigma \times \mathbb{R}^*. \tag{2.12}$$

In the case of an x-independent potential, we straightforwardly observe that all these conditions coincide with the transparent boundary condition (2.3). Moreover, the conditions of order 1 and 3/2 are exactly the same. This indicates that, even for a space varying potential, the first order artificial boundary condition should be quite efficient.

Before writing the formal transposition to the nonlinear operator, we emphasize the fact that, in the linear situation, the function $\partial_x V$ is seen as an operator of order zero since it just multiplies u. In the nonlinear case, the situation is completely different if we make the substitution $V \to q|u|^2$ since we have $\partial_x V \to \partial_x(q|u|^2)$. For this reason, we must specify this aspect in the deduced boundary conditions by adding the dependence of the nonlinear operators according to ∂_x. Following the proposed strategy, we formally deduce the two NLABCs of order $M/2$ ($\text{NLABC}_1^{M/2}$)

$$\partial_{\mathbf{n}} u + \Lambda_{M/2}^{NLS}(x, t, \partial_x, \partial_t, |u|)u = 0 \text{ on } \Sigma \times \mathbb{R}^*, \tag{2.13}$$

where the nonlinear operators are defined by

$$\begin{aligned} &\Lambda_1^{NLS}(x, t, \partial_x, \partial_t, |u|) = e^{-i\frac{\pi}{4}} e^{i\mathbb{U}} \partial_t^{1/2}(e^{-i\mathbb{U}} u), \\ &\Lambda_2^{NLS}(x, t, \partial_x, \partial_t, |u|) = \Lambda_1^{NLS}(x, t, \partial_x, \partial_t, |u|) + i\frac{q}{4}\partial_{\mathbf{n}}(|u|^2) e^{i\mathbb{U}} I_t(e^{-i\mathbb{U}} u). \end{aligned} \tag{2.14}$$

The function $\mathbb{U}$ is defined by the relation (2.5).

Remark 2.3. The particular form of the asymptotic expansion (2.8) for the operators $\Lambda^{\pm}$ suggests that we use a high-frequency assumption on the solution to the linear Schrödinger equation. This should also be the case for the formal extension to the nonlinear case. We will see during the numerical experiments that these NLABCs are particularly accurate for a "fast" soliton.

2.2. An energy bound on the solution to the approximate initial boundary value problems. In the case of the linear Schrodinger equation sets in the whole space with a smooth real-valued time-dependent potential V, the conservation of the L^2-norm of the solution can be proved. When the infinite domain is truncated by the transparent boundary condition (2.3), the conservation of the L^2-norm becomes an energy bound; the $L^2(\Omega_i)$-norm, denoted by $\|u\|_{0,\Omega_i}$, of the solution u at a given time is bounded by $\|u_0\|_{0,\Omega_i}$. This implies the uniqueness of the solution to the bounded initial boundary value problem. In the case of the cubic NLS equation defined on the whole space, both the L^2-norm of the solution and its Hamiltonian defined by

$$\mathcal{H}(t) = \|\partial_x u\|_0^2(t) - \frac{1}{2}\left\||u|^2\right\|_0^2(t)$$

are conserved. The question which naturally arises is to know if the $L^2(\Omega_i)$-norm of the solution to the truncated NLS is bounded by $\|u_0\|_{0,\Omega_i}$. In the case of the first-order condition, the following results hold.

PROPOSITION 2.4. *Let $u_0 \in L^2(\Omega_i)$ be a compactly supported initial datum such that* $\text{Supp}(u_0) \subset \Omega_i$ *and u is a solution to the initial boundary value problem*

$$\begin{aligned} &i\partial_t u + \partial_x^2 u + q|u|^2 u = 0 \quad \text{in } \Omega_i \times \mathbb{R}^+, \\ &\partial_{\mathbf{n}} u + \Lambda_1^{NLS}(x, t, \partial_x, \partial_t, |u|)u = 0 \quad \text{on } \Sigma \times \mathbb{R}^+, \\ &u(x, 0) = u_0(x) \quad \forall x \in \Omega_i, \end{aligned} \tag{2.15}$$

with the function $\mathbb{U}$ *given by* (2.5)*. Then, u fulfills the energy bound*

$$\forall t > 0, \forall u_0 \in L^2(\Omega_i),\ \ \|u(t)\|_{0,\Omega_i} \leq \|u_0\|_{0,\Omega_i}\,. \tag{2.16}$$

Proof. Let us multiply the NLS equation, given by the first equation of system (2.15), by $-i\overline{u}$, where $\overline{u}$ designates the conjugate complex value of u. Integrating by parts on Ω_i and next taking the real part of the resulting equation, we obtain the following expression after integration on an arbitrary time interval $[0, T]$, with $T > 0$:

$$\frac{1}{2}(\|u\|^2_{0,\Omega_i}(T) - \|u_0\|^2_{0,\Omega_i}) = \Re\left(i\int_0^T [\overline{u}\partial_x u]_{x_l}^{x_r}\,dt\right). \tag{2.17}$$

To get the result of the proposition, let us prove the positiveness of the term involved in the right-hand side of (2.17). To simplify the presentation, we consider only the term at point x_r; the other one at x_l can be treated in a similar way. Using the expression of the first order artificial boundary condition and considering the extension $\widetilde{u}$ of u by zero for any time $t > T$, we have

$$\Re\left(i\int_0^T \overline{u}(x_r,t)\partial_{\mathbf{n}}u(x_r,t)dt\right) = -\Re\left(e^{i\frac{\pi}{4}}\int_0^\infty \partial_t^{1/2}(e^{-i\widetilde{\mathbb{U}}(x_r,t)}\widetilde{u})\overline{(e^{-i\widetilde{\mathbb{U}}(x_r,t)}\widetilde{u})}dt\right). \tag{2.18}$$

The control of the sign of the term on the right-hand side is a consequence of the property that the operator $e^{i\frac{\pi}{4}}\partial_t^{1/2}$ is a positive memory-type operator [5, 7]. The proof is based on the Plancherel theorem of the Laplace transform.

LEMMA 2.5. *Let* $\varphi \in H^{1/4}(0,T)$ *be a function extended by zero for any time* $s > t$*. Then, we have the inequality*

$$\Re\left(e^{i\frac{\pi}{4}}\int_0^\infty \partial_s^{1/2}\varphi(s)\overline{\varphi(s)}ds\right) \geq 0.$$

Moreover, the following estimate holds for the imaginary part:

$$\Im\left(e^{i\frac{\pi}{4}}\int_0^\infty \partial_s^{1/2}\varphi(s)\overline{\varphi(s)}ds\right) \geq 0.$$

Applying this lemma to (2.18) and using (2.17), we prove the needed inequality. ☐

If we develop the same approach for the second order artificial boundary condition, it does not seem possible to control the sign of the corrective term. This is essentially due to the quantity $\partial_{\mathbf{n}}(|u|^2)$ which does not have a well-defined sign. Another point is that the Hamiltonian $\mathcal{H}$ of the nonlinear cubic Schrodinger equation defined on $\mathbb{R}_x \times \mathbb{R}_t^+$ is conserved. A similar bound as (2.16) for the Hamiltonian of the solution to (2.15) cannot be obtained by direct arguments.

2.3. Other asymptotic NLABCs. We consider now the artificial boundary conditions (2.11) and (2.12) derived for a potential V. These boundary conditions involve some time fractional derivative and integration operators applied to the product of two functions. To expand these quantities, we can use the Leibnitz derivation rule for fractional operators. To this end, let us recall that if f and g are two functions, where f is $\mathcal{C}^\infty$ on $[0,t]$ and g is continuous, then the fractional derivative of fg is given by

$$\partial_t^p(fg) = \sum_{k=0}^{\infty} \frac{\Gamma(p+1)}{k!\Gamma(p-k+1)} f^{(k)} \partial_t^{p-k} g. \tag{2.19}$$

(A similar formula with a finite number of terms and an integral rest also holds [22].) Using formula (2.19), we can expand the fractional derivative operator as

$$\partial_t^{1/2}(e^{-i\mathcal{V}}u) = \sum_{k=0}^{\infty} \frac{\Gamma(3/2)}{k!\Gamma(3/2-k)} \partial_t^k(e^{-i\mathcal{V}})\partial_t^{1/2-k}u.$$

Considering the relation $\Gamma(z+1) = z\Gamma(z)$ for the Gamma function, we have $\Gamma(z) = \Gamma(z-k)\prod_{j=1}^{k}(z-j)$ for any $k \geq 1$. We then deduce the expression

$$\partial_t^{1/2}(e^{-i\mathcal{V}}u) = e^{-i\mathcal{V}}\partial_t^{1/2}u + \sum_{k=1}^{\infty} \frac{\prod_{j=1}^{k}(3/2-j)}{k!} \partial_t^k(e^{-i\mathcal{V}}) I_t^{k-1/2}u.$$

Truncating this series to a finite number M of terms, we get some new approximations of the artificial boundary conditions (2.11) and (2.12). For instance, keeping three terms in the above expression yields the approximation of (2.11)

$$\partial_{\mathbf{n}}u + e^{-i\pi/4}\partial_t^{1/2}u - \frac{e^{i\pi/4}}{2}VI_t^{1/2}u + \frac{e^{i\pi/4}}{8}\partial_t V I_t^{3/2}u + e^{-i\pi/4}\frac{V^2}{8}I_t^{3/2}u = 0. \tag{2.20}$$

Let us notice that the condition (2.11) has been constructed modulo some operators of OPS^{-1} for the modified Schrödinger equation defined by the operator (2.6). As a consequence, we should keep only the first two time operators of the previous expression. However, if we try to prove an energy bound for the $L^2(\Omega_i)$-norm of the solution to the truncated nonlinear initial boundary value problem, it appears that we cannot derive an a priori estimate to control the operator of order $-1/2$. There is a lack of symmetry of this latter term. To circumvent this drawback, we propose a modification of the conditions. To this end, let us recall that we have from the Leibnitz derivation rule for any strictly positive real number ν and real analytic functions v and u on $[0;+\infty[$

$$I_t^\nu(vu) = \sum_{k=0}^{\infty} \frac{\Gamma(-\nu+1)}{k!\Gamma(-\nu-k+1)} \partial_t^k v I_t^{\nu+k}u.$$

In particular, the first two coefficients are 1 and $-1/2$ for $\nu = 1/2$. If we now consider $v = \sqrt{V}$, the following approximation holds:

$$\sqrt{V}I_t^{1/2}(\sqrt{V}u) = VI_t^{1/2}u - \frac{\sqrt{V}}{2}\partial_t\sqrt{V}I_t^{3/2}u \mod(OPS^{-5/2}).$$

Moreover, we also have the symmetrization

$$\frac{e^{-i\pi/4}}{8}V^2 I_t^{3/2}u = \frac{e^{-i\pi/4}}{8}VI_t^{3/2}(Vu) \mod(OPS^{-5/2}). \tag{2.21}$$

Since the artificial boundary condition (2.11) is designed modulo an operator of OPS^{-1}, we can modify it by using the above relation and considering the new boundary condition

$$\partial_{\mathbf{n}}u + e^{-i\pi/4}\partial_t^{1/2}u - \frac{e^{i\pi/4}}{2}\sqrt{V}I_t^{1/2}(\sqrt{V}u) + \frac{e^{-i\pi/4}}{8}VI_t^{3/2}(Vu) = 0 \quad \text{on } \Sigma \times \mathbb{R}^+. \tag{2.22}$$

If we now come back to the NLS, we can deduce some symmetric NLABCs by using the formal substitution: $Vu = q|u|^2u$. Following this process, one gets different alternative boundary conditions of order $M \in \mathbb{N}$. For instance, the modification of the NLABC$_1^2$ artificial boundary condition (2.13)–(2.14) leads to

$$\partial_{\mathbf{n}}u + e^{-i\pi/4}\partial_t^{1/2}u - q\frac{e^{i\pi/4}}{2}|u|I_t^{1/2}(|u|u) + q^2\frac{e^{-i\pi/4}}{8}|u|^2I_t^{3/2}(|u|^2u) + i\frac{q}{4}\partial_{\mathbf{n}}(|u|^2)I_tu = 0. \tag{2.23}$$

More generally, one gets the following definition.

Definition 2.6. *The symmetric approximate NLABC of order $M/2$ (denoted by NLABC$_2^{M/2}$) for $M \in \mathbb{N}^*$ is given by*

$$\partial_{\mathbf{n}}u + T_{M/2}^{NLS}(x, t, \partial_x, \partial_t, |u|)u = 0 \quad \textit{on } \Sigma \times \mathbb{R}^{*+}, \tag{2.24}$$

defining the different nonlinear fractional artificial boundary operators $T_{M/2}$ as

$$\begin{aligned} T_1^{NLS}u &= e^{-i\pi/4}\partial_t^{1/2}u, \\ T_2^{NLS}u &= T_1^{NLS}u - q\frac{e^{i\pi/4}}{2}|u|I_t^{1/2}(|u|u) + i\frac{q}{4}\partial_{\mathbf{n}}(|u|^2)I_tu, \\ T_{5/2}^{NLS}u &= T_2^{NLS}u + q^2\frac{e^{-i\pi/4}}{8}|u|^2I_t^{3/2}(|u|^2u). \end{aligned} \tag{2.25}$$

Once again and similarly to the artificial boundary conditions (2.11) and (2.12), it seems impossible to control the sign of $\partial_{\mathbf{n}}(|u|^2)I_tu$. However, if we do not consider this term in the definition of the above boundary conditions, we can prove that the $L^2(\Omega_i)$-norm of the solution to the truncated nonlinear initial boundary value problem is bounded by the norm of the initial datum. The proof (not detailed here) uses some lemmas similar to Lemma 2.5 to treat the quantities issued from the integral operators $I_t^{1/2}$ [3] and $I_t^{3/2}$. Moreover, in the particular case of the first order condition where only the time fractional operator of order 1/2 appears, one can show that the Hamiltonian of the solution at a given instant t is bounded by the Hamiltonian at $t = 0$.

3. Semidiscrete approximation of the cubic NLS equation and the NLABC.

We investigate the construction of stable time discretization schemes of the NLABC associated to the NLS equation. The developments are based on the method of Dúran and Sanz-Serna [13] and on works for treating the linear transparent boundary condition [3].

3.1. Preliminary results.

Several time discretization schemes can be developed for solving the cubic NLS equation. The most widely used approach is based on the second order Strang splitting formula for the time discretization and on the application of the FFT for the spatial part [10, 19]. This method requires the application of some periodic boundary conditions to bound the computational domain.

Since these conditions do not reproduce the real physical propagation phenomenon at the boundary, this limits the applicability of this method (for instance, to simulate the evolution of the interaction of two solitons). Furthermore, this technique does not take into account more complicated boundary conditions such as the one that we propose here. Another possibility consists in using an interior Crank–Nicolson scheme. The Schrödinger equation is then discretized at time $t_{n+1/2} = (t_{n+1} + t_n)/2$ by a second order approximation. In what follows, if δt designates the time step, then $t_n = n\delta t$ stands for the nth time step, where $n \in \mathbb{N}$. The first approximation introduced by Delfour, Fortin, and Payne [12] consists in approximating both u and $|u|^2$ by the midpoint formula. More recently, Durán and Sanz-Serna [13] proposed another second order scheme based on the only discretization of u by the midpoint rule. The usual Crank–Nicolson scheme is given by

$$i\frac{u^{n+1}-u^n}{\delta t} + \partial_x^2 \frac{u^{n+1}+u^n}{2} + q\left(\frac{|u^{n+1}|^2+|u^n|^2}{2}\right)\frac{u^{n+1}+u^n}{2} = 0 \tag{3.1}$$

and the Durán–Sanz-Serna scheme by

$$i\frac{u^{n+1}-u^n}{\delta t} + \partial_x^2 \frac{u^{n+1}+u^n}{2} + q\left|\frac{u^{n+1}+u^n}{2}\right|^2\frac{u^{n+1}+u^n}{2} = 0. \tag{3.2}$$

We recall, following [13], that this scheme is very well adapted for computing soliton-like solutions. We denote here by u^n the approximate value of u at time t_n. Let us notice that the scheme (3.2) has a lower computational cost than (3.1). Indeed, if we set $2v^{n+1} = u^{n+1} + u^n$, the scheme reads for $n \geq 0$

$$2i\frac{v^{n+1}-u^n}{\delta t} + \partial_x^2 v^{n+1} + q|v^{n+1}|^2 v^{n+1} = 0, \tag{3.3}$$

imposing $v^0 = 0$. Moreover, the simple form of the scheme (3.2) leads to an easier implementation of the NLABCs. In the developments below, we focus our attention on the presentation of this latter scheme.

Remark 3.1. The approximation of the artificial boundary conditions for the Crank–Nicolson scheme (3.1) is however possible. We do not present the results here since this approach has proved to be less accurate than for the Durán–Sanz-Serna scheme.

In the linear case, the fractional operators defining the boundary conditions need to be discretized by the trapezoidal rule to yield the stability of the Crank–Nicolson scheme coupled to a discrete transparent artificial boundary condition. This is still the case for the NLS equation. To prove it, we begin by recalling the main results stated in [3].

PROPOSITION 3.2. *If $\{f_n\}_{n\in\mathbb{N}}$ is a sequence of complex values approximating $\{f(t_n)\}_{n\in\mathbb{N}}$, then the approximations of $\partial_t^{1/2} f(t_n)$ and $I_t^{1/2} f(t_n)$ are given by the numerical quadrature formulas*

$$I_t^{1/2} f(t_n) \approx \frac{\sqrt{2\delta t}}{2}\sum_{k=0}^{n} \alpha_k f^{n-k} \tag{3.4}$$

and

$$\partial_t^{1/2} f(t_n) \approx \frac{2}{\sqrt{2\delta t}}\sum_{k=0}^{n} \beta_k f^{n-k}, \tag{3.5}$$

where $(\alpha_k)_{k\in\mathbb{N}}$ *and* $(\beta_k)_{k\in\mathbb{N}}$ *designate the sequences defined by*

$$\begin{cases} (\alpha_0,\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5,\dots) = \left(1,1,\frac{1}{2},\frac{1}{2},\frac{1\cdot 3}{2\cdot 4},\frac{1\cdot 3}{2\cdot 4},\dots\right), \\ \beta_k = (-1)^k\alpha_k \ \forall k\geq 0. \end{cases}$$

3.2. Approximation and stability result for the Dúran–Sanz-Serna-type scheme and the boundary conditions NLABC$_1^j$, $j = 1, 2$. Let us recall that the initial boundary value problem in the truncated domain with a NLABC$_1^j$-type condition is

$$\text{(3.6)} \qquad \begin{cases} i\partial_t u + \partial_x^2 u + q|u|^2 u = 0,\ (x,t)\in\Omega_i\times\mathbb{R}^{*+}, \\ \partial_{\mathbf{n}} u + \Lambda_j^{NLS}(x,t,\partial_x,\partial_t,|u|)u = 0 \ \text{ on } \Sigma\times\mathbb{R}^{*+} \text{ for } j = 1/2,\ 1, \text{ or } 2, \\ u(x,0) = u_0(x),\ x\in\Omega_i, \end{cases}$$

where the artificial boundary operators are given by the expressions (2.14).

To define the semidiscretization in the time of the boundary conditions, we introduce the approximation $\mathbb{U}^p$ of

$$\mathbb{U}(x,t) = q\int_0^t |u(x,s)|^2 ds$$

by the trapezoidal formula for $p\geq 2$

$$\text{(3.7)} \qquad \mathbb{U}^p = q\delta t\left(\sum_{l=1}^{p-1}\left|u^l\right|^2 + \frac{1}{2}\left|u^p\right|^2\right),$$

with $\mathbb{U}^0 = 0$ and $\mathbb{U}^1 = q\delta t\left|u^1\right|^2/2$. Let $\mathbb{E}^p$ and $\widetilde{\mathbb{E}^{p-1}}$ be the quantities defined by

$$\text{(3.8)}\quad \mathbb{E}^p = \exp(i\mathbb{U}^p) = \exp\left(iq\delta t\sum_{l=1}^{p-1}\left|u^l\right|^2\right)\exp\left(iq\frac{\delta t}{2}\left|u^p\right|^2\right) = \widetilde{\mathbb{E}^{p-1}}\exp\left(iq\frac{\delta t}{2}\left|u^p\right|^2\right),$$

setting $\mathbb{E}^0 = 1$ and $\mathbb{E}^1 = \exp(i\mathbb{U}^1)$. Using relations (3.5), we can define the semidiscrete approximations, denoted by $\Lambda_{j,n+1}^{NLS}$, of the continuous artificial operators Λ_j^{NLS} by

$$\text{(3.9)} \qquad \Lambda_{1,n+1}^{NLS}(x,\partial_x,|u^{n+1}|)u^{n+1} = e^{-i\frac{\pi}{4}}\sqrt{\frac{2}{\delta t}}\mathbb{E}^{n+1}\sum_{k=0}^{n+1}\beta_k\overline{\mathbb{E}^{n+1-k}}u^{n+1-k}$$

and

$$\begin{aligned} \Lambda_{2,n+1}^{NLS}(x,\partial_x,|u^{n+1}|)u^{n+1} &= \Lambda_{1,n+1}^{NLS}(x,|u^{n+1}|)u^{n+1} \\ \text{(3.10)} \qquad &\quad + i\frac{q}{4}\partial_{\mathbf{n}}\left(|u^{n+1}|^2\right)\mathbb{E}^{n+1}\delta t\sum_{k=0}^{n+1}\gamma_k\overline{\mathbb{E}^{n+1-k}}u^{n+1-k}, \end{aligned}$$

where $(\gamma_k)_{k\in\mathbb{N}}$ is the sequence defined by $(\gamma_0,\gamma_1,\gamma_2,\gamma_3,\dots) = \left(\frac{1}{2},1,1,1,\dots\right)$.

Proposition 3.3. *The Durán–Sanz-Serna semidiscrete scheme for the initial boundary value problem* (3.6) *is given by*

$$\begin{aligned} &2i\frac{v^{n+1}-u^n}{\delta t}+\partial_x^2 v^{n+1}+q|v^{n+1}|^2v^{n+1}=0,\ x\in\Omega_i,\\ &\partial_{\mathbf{n}}v^{n+1}+\Lambda^{NLS}_{j,n+1}(x,\partial_x,|v^{n+1}|)v^{n+1}=0 \quad \text{on } \Sigma \text{ for } j=1 \text{ or } 2,\\ &u^0=u_0,\ x\in\Omega_i, \end{aligned} \tag{3.11}$$

where $v^{n+1}=\frac{u^{n+1}+u^n}{2}$. *Furthermore, the following energy inequality holds for* $j=1$*:*

$$\|u^{N+1}\|_{L^2(\Omega_i)}\le\|u_0\|_{L^2(\Omega_i)}\ \forall N\ge 0, \tag{3.12}$$

implying the stability of the scheme.

Proof. The construction of the Durán–Sanz-Serna semidiscrete scheme is immediate in view of the preliminary results. We just sketch the proof of the stability result. Essentially, the arguments are adapted from the analysis developed in [3] for the one-dimensional linear Schrödinger equation (the techniques are mainly based on the $\mathcal{Z}$-transform). The particular symmetrical form of the continuous NLABCs and the property of dissipation of the fractional operators used in Proposition 2.4 extend in a certain way at the semidiscrete level. This is finally quite natural since the semidiscretization of the boundary conditions has been written a priori to be consistent with the interior scheme. □

Remark 3.4. Like in the continuous case, we cannot prove the estimate (3.12) for the second order condition. This is due to a lack of control of the sign of the corrective term.

3.3. Approximation and stability result for the Dúran–Sanz-Serna-type scheme and the boundary conditions NLABC$_2^j$, $j=1,2$. The initial boundary value problem in the bounded computational domain for an NLABC$_2^j$ condition is given by

$$\left\{\begin{array}{l} i\partial_t u+\partial_x^2u+q|u|^2u=0,\ (x,t)\in\Omega_i\times\mathbb{R}^+,\\ \partial_{\mathbf{n}}u+T_j^{NLS}(x,t,\partial_x,\partial_t,|u|)u=0 \quad \text{on } \Sigma\times\mathbb{R}^{*+} \text{ for } j=1 \text{ or } 2,\\ u(x,0)=u_0(x)\ x\in\Omega_i. \end{array}\right. \tag{3.13}$$

Using the semidiscrete versions (3.4) and (3.5) of the half-order integral and derivative operators, we introduce the following semidiscretizations of the artificial operators T_1^{NLS} and T_2^{NLS}, respectively, given by

$$T^{NLS}_{1,n+1}(x,\partial_x,|v^{n+1}|)v^{n+1}=e^{-i\frac{\pi}{4}}\sqrt{\frac{2}{\delta t}}\sum_{k=0}^{n+1}\beta_k v^{n+1-k} \tag{3.14}$$

and

$$\begin{aligned} T^{NLS}_{2,n+1}(x,\partial_x,|v^{n+1}|)v^{n+1}&=T^{NLS}_{1,n+1}(x,\partial_x,|v^{n+1}|)v^{n+1}\\ &\quad-q\frac{e^{i\frac{\pi}{4}}}{2}|v^{n+1}|\frac{\sqrt{2\delta t}}{2}\sum_{k=0}^{n+1}\alpha_k|v^{n+1-k}|v^{n+1-k}\\ &\quad+i\frac{q}{4}\partial_{\mathbf{n}}(|v^{n+1}|^2)\delta t\left(\sum_{k=1}^{n}v^k+\frac{v^{n+1}}{2}\right). \end{aligned} \tag{3.15}$$

As for the NLABCs of the first type, the following proposition holds.

Proposition 3.5. *Let us define* $v^{n+1} = \frac{u^{n+1}+u^n}{2}$*. Then, the Durán–Sanz-Serna semidiscrete scheme for the initial boundary value problem* (3.13) *reads*

$$\begin{aligned} &2i\frac{v^{n+1}-u^n}{\delta t} + \partial_x^2 v^{n+1} + q|v^{n+1}|^2 v^{n+1} = 0, \;\; x \in \Omega_i, \\ &\partial_{\mathbf{n}} v^{n+1} + T_{j,n+1}^{NLS}(x, \partial_x, |v^{n+1}|)v^{n+1} = 0 \quad \text{on } \Sigma \text{ for } j = 1 \text{ or } 2, \\ &u^0 = u_0, \quad x \in \Omega_i. \end{aligned} \tag{3.16}$$

Moreover, we have the inequality

$$\|u^{N+1}\|_{L^2(\Omega_i)} \leq \|u_0\|_{L^2(\Omega_i)} \; \forall N \geq 0,$$

for $j = 1$ *or for* $j = 2$ *by neglecting the last term of the expression* (3.15)*. In this case, this implies the stability of the scheme* (3.16)*.*

Proof. Once again, we do not detail the proof of Proposition 3.5 which is obtained by some arguments close to the ones used in the linear case [3]. □

4. Numerical implementation and simulations.

4.1. Some aspects of the numerical implementation. Since the Jacobian of the map associated to the nonlinear problems (3.11) and (3.16) is complicated to obtain, we propose to rather use a classical fixed point method. To this end, let us begin by writing the artificial boundary conditions as some nonlinear Fourier–Robin-type boundary conditions. We choose to develop only the calculations for the conditions $\Lambda_{j,n+1}^{NLS}$. We get the relation

$$\partial_{\mathbf{n}} v^{n+1} + e^{-i\frac{\pi}{4}}\sqrt{\frac{2}{\delta t}}\left(v^{n+1} + \widetilde{\mathbb{E}^n}\exp\left(iq\delta t\frac{\left|v^{n+1}\right|^2}{2}\right)\sum_{k=1}^{n}\beta_{n+1-k}\overline{\mathbb{E}^k}v^k\right) = 0 \tag{4.1}$$

for the operator $\Lambda_{1,n+1}^{NLS}$. The corrective term involved in the definition of $\Lambda_{2,n+1}^{NLS}$ can be rewritten as

$$i\frac{q}{4}\partial_{\mathbf{n}}(|v^{n+1}|^2)\left(\frac{\delta t}{2}v^{n+1} + \delta t\widetilde{\mathbb{E}^n}\exp\left(iq\delta t\frac{\left|v^{n+1}\right|^2}{2}\right)\sum_{k=1}^{n}\overline{\mathbb{E}^k}v^k\right). \tag{4.2}$$

The fixed point algorithm is applied for treating the nonlinearities appearing both in the Schrödinger equation and in the artificial boundary conditions. The resulting scheme is summarized in Table 4.1.

At each iteration of the algorithm, we incorporate the linear Fourier–Robin boundary condition by using the weak formulation

$$\begin{aligned} \int_{\Omega_i}\frac{2i}{\delta t}w^{s+1}\psi dx - \int_{\Omega_i}\partial_x w^{s+1}\partial_x\psi dx - \int_{\Sigma}e^{-i\frac{\pi}{4}}\sqrt{\frac{2}{\delta t}}w^{s+1}\psi d\Sigma \\ = -\int_{\Omega_i}q|w^s|^2 w^s\psi dx + \int_{\Omega_i}\frac{2i}{\delta t}u^n\psi dx - \int_{\Sigma}g^s\psi d\Sigma, \end{aligned}$$

where ψ designates a sufficiently smooth function. The spatial discretization is performed by a conform linear Galerkin finite element method for u, $|u|^2$, and ψ providing hence the stability of the whole scheme. This variational approach leads to a tridiagonal banded matrix. The solution to the associated linear system is therefore simple and realized by a direct LU solver. The involvement of the other NLABCs follows the same approach.

Table 4.1

Fixed point algorithm for solving the cubic NLS equation using the NLABC$_1^2$-type nonlinear artificial boundary condition.

```
let $w^0 = u^n$
$s = 0$
while $\|w^{s+1} - w^s\|_{L^2(\Omega_i)} > \varepsilon$ do
        solve the linear boundary-value problem
        $\begin{cases} \dfrac{2i}{\delta t} w^{s+1} + \partial_x^2 w^{s+1} = -q|w^s|^2 w^s + \dfrac{2i}{\delta t} u^n \text{ in } \Omega_i, \\ \partial_{\mathbf{n}} w^{s+1} + e^{-i\frac{\pi}{4}} \sqrt{\dfrac{2}{\delta t}} w^{s+1} = g^s \text{ on } \Sigma, \end{cases}$
        setting
        $g^s = -e^{-i\frac{\pi}{4}} \sqrt{\dfrac{2}{\delta t}} \left( \widetilde{\mathbb{E}^n} \exp\left( iq\delta t \dfrac{|w^s|^2}{2} \right) \sum_{k=1}^{n} \beta_{n+1-k} \overline{\mathbb{E}^k} v^k \right)$
              $-i\dfrac{q}{4} \partial_{\mathbf{n}}(|w^s|^2) \left( \dfrac{\delta t}{2} w^s + \delta t \widetilde{\mathbb{E}^n} \exp\left( iq\delta t \dfrac{|w^s|^2}{2} \right) \sum_{k=1}^{n} \overline{\mathbb{E}^k} v^k \right)$
end while
$v^{n+1} = w^{s+1}$
$u^{n+1} = 2v^{n+1} - u^n$
```

4.2. Numerical results. The one-dimensional cubic NLS equation is integrable by using the inverse scattering theory [21]. This approach yields the so-called exact *soliton* solution given by

$$u_{\text{ex}}(x,t) = \sqrt{\frac{2a}{q}} \operatorname{sech}(\sqrt{a}(x-ct)) \exp\left(i\frac{c}{2}(x-ct) \right) \exp\left(i\left(a + \frac{c^2}{4} \right) t \right). \tag{4.3}$$

From now on, we fix the focusing parameter q to 1. The real parameter a gives the amplitude of the wavefield. Finally, c is the velocity of the soliton. Since the derivation of the NLABCs has been constructed under a high-frequency assumption (see Remark 2.3), we can expect that our approach will be more efficient for a high-speed soliton. Throughout the computations, we have taken $\varepsilon = 10^{-6}$ in the fixed point algorithm (4.1).

To perform an exhaustive study of the proposed artificial boundary conditions, we compare the nonlinear conditions NLABC$_i^j$ for $1 \leq i,j \leq 2$ to the linear artificial boundary condition (LABC) (1.2) for the soliton defined by $a = 2$ and $c = 15$. The numerical parameters are $\delta t = 10^{-3}$ for a final time $T_f = 2$. The finite computational spatial domain is $\Omega_i = [-10, 10]$ discretized by 4000 equally spaced points. To focus on the spurious reflections link to the different methods, we plot the contour of $\log_{10}(|u|)$ in Figures 1–5. The curves are presented with respect to the increasing accuracy of the artificial boundary conditions. We see in Figure 1 that the maximal reflection is approximately equal to 10^{-2} for an initial amplitude of 2 and the LABC. For Figures 2–5, the reflection attains a maximal value around 5×10^{-3}. The reflection occurring at the right boundary decreases according to the order of the different conditions NLABC$_1^j$ or NLABC$_2^j$. Moreover, the most accurate results are obtained for the condition NLABC$_1^2$ with a minimal region of maximal reflection. Unlike the LABC, the reflection at the left boundary has an amplitude inferior to 10^{-4}.

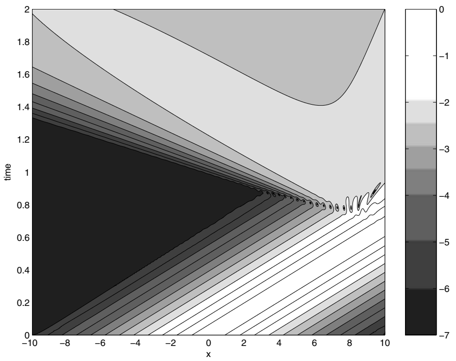


FIG. 1. *Contour plot of* $\log_{10}(|u|)$ *for the linear artificial boundary condition (LABC).*

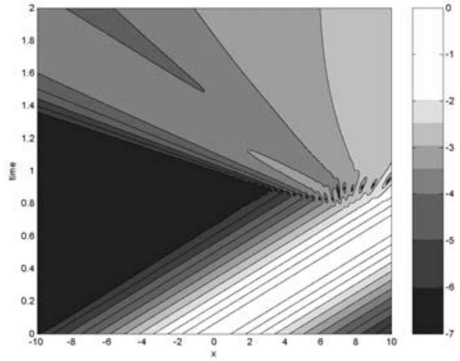


FIG. 2. *Contour plot for the nonlinear artificial boundary condition* $NLABC_2^1$.

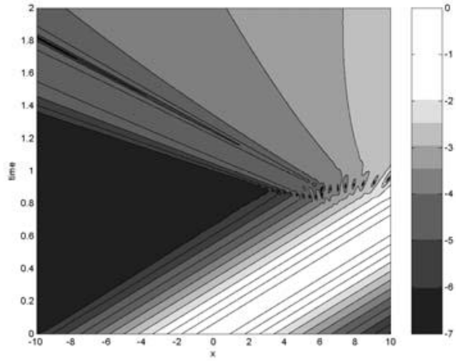


FIG. 3. *Contour plot for the nonlinear artificial boundary condition* $NLABC_2^2$.

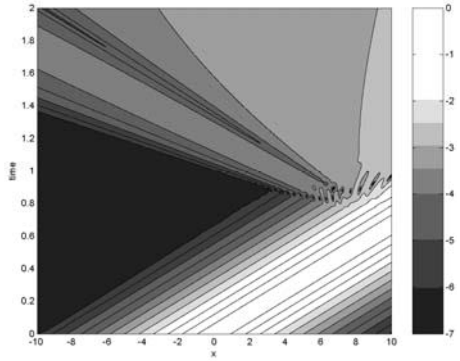


FIG. 4. *Contour plot for the nonlinear artificial boundary condition* $NLABC_1^1$.

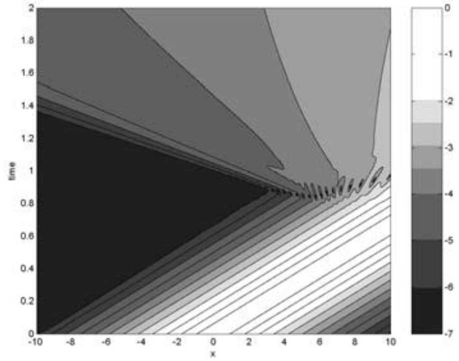


FIG. 5. *Contour plot for the nonlinear artificial boundary condition* $NLABC_1^2$.

To specify these results, we plot in Figure 6 the relative error for the $L^2(\Omega_i)$-norm

$$\frac{\|u_{\mathrm{ex}} - u_{\mathrm{num}}\|_{0,\Omega_i}}{\|u_{\mathrm{num}}\|_{0,\Omega_i}},$$

where u_{num} denotes the numerical solution. For the linear case, the error is about 2%,

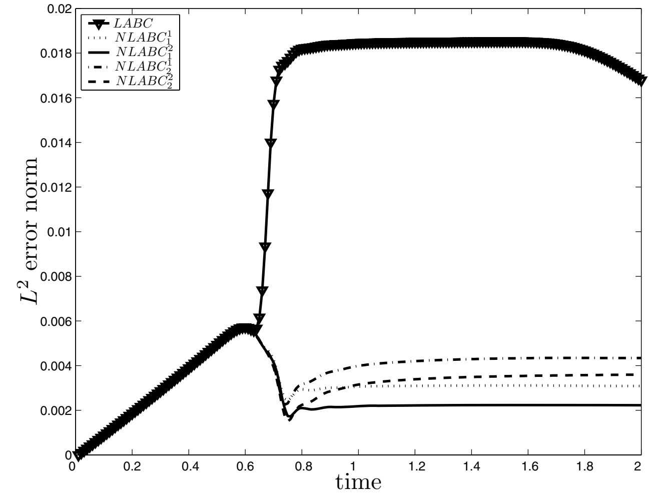


Fig. 6. *Relative error for the different linear and nonlinear artificial boundary conditions.*

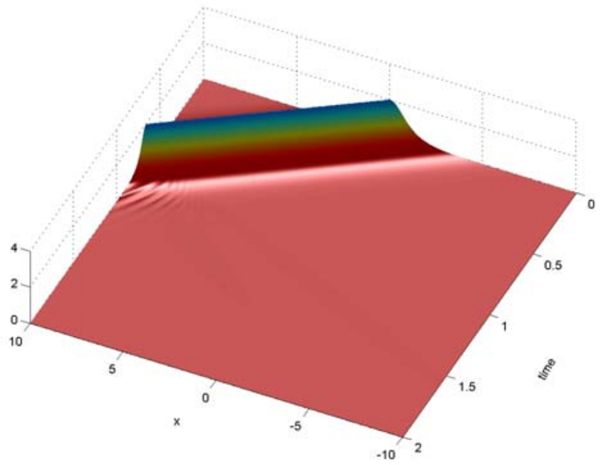


Fig. 7. *Representation of the amplitude of the computed "fast" soliton.*

whereas the best result is obtained for the NLABC_1^2 for a final error of 0.2%. This last error is less than the intrinsic phase error of the Sanz-Serna scheme generally linked to the Crank–Nicolson-type schemes (see, for instance, [9]). The most accurate NLABC does not require any additional cost. The numerical classification of the artificial boundary conditions coincides with the theoretical one. To end with this test case, we depict in Figure 7 the evolution of this "fast" soliton. We emphasize the very small spurious reflections by adding a light to the figure. This allows to visualize it by showing the associated shadow zones. Without this brightness, the reflections are too small to be seen in the representation.

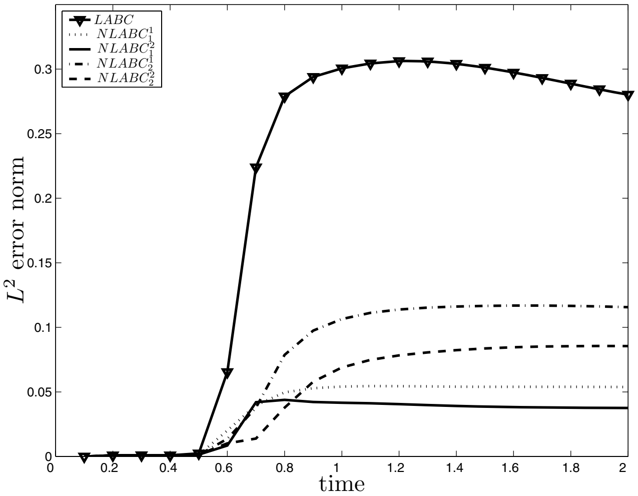


Fig. 8. *Relative error for the different linear and nonlinear artificial boundary conditions.*

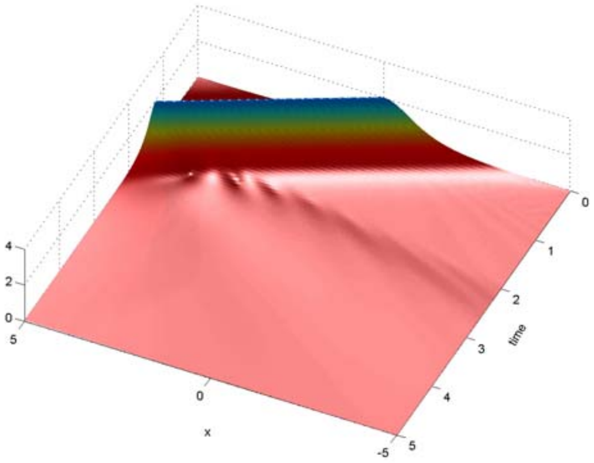


Fig. 9. *Representation of the amplitude of the computed "slow" soliton.*

The next experiment concerns a "slow" propagative soliton defined by a velocity $c = 4$ and an amplitude $a = 2$. The finite domain of computation is reduced to $\Omega_i = [-5, 5]$ and discretized with 4000 points. The time step is now $\delta t = 5 \times 10^{-3}$ and the final time is fixed to $T_f = 5$. Let us recall that all the conditions have been derived under an assumption of high frequency. Since we consider a slower soliton, the reflection should be larger. We get an acceptable error of 5% (see Figure 8) for the NLABC_1^2. This is not the case of the LABC which yields a large error of 30%. Even if the error occurring in this situation for the NLABC_1^1 and NLABC_1^2 is larger than for the "fast" soliton, this always allows us to reproduce the behavior of the solution as can be seen in Figure 9. Once again, an artificial light has been added to show the numerical reflection involving in the approximation. Finally, from all the above tests,

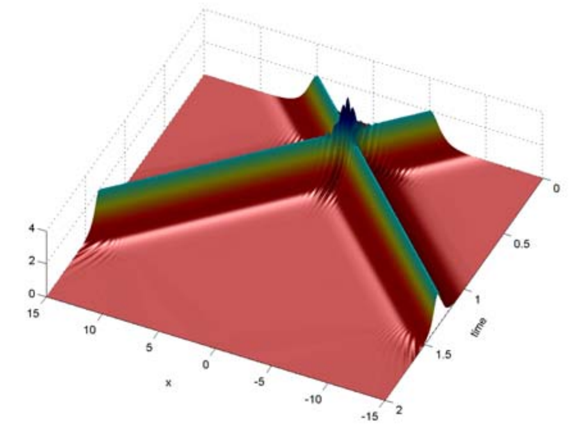


Fig. 10. *Interaction of two fast solitons with opposite directions.*

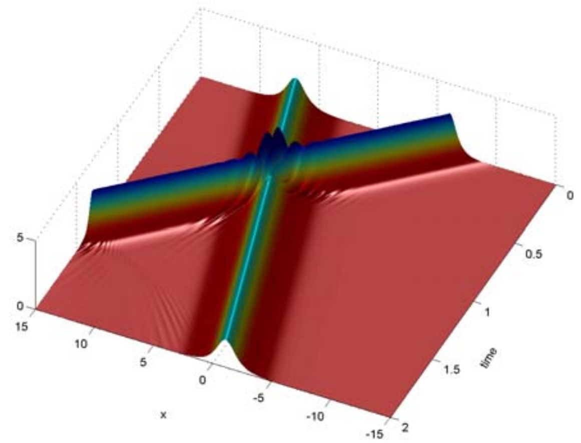


Fig. 11. *Interaction of slow and fast solitons with opposite directions.*

we conclude that the most accurate results are obtained with the NLABC_1^2, which is from now on taken as the reference.

We consider now the problem of simulating the interaction of two solitons. The domain of computation $\Omega_i = [-15, 15]$ is discretized with 6000 points. The final time is $T_f = 2$ for a time step $\delta t = 10^{-3}$. We consider in Figure 10 two fast solitons evolving in two opposite directions and centered at $x = -5$ and $x = 5$ at the initial time. The two velocities are $c = 12$ and $c = -12$. As for one soliton, we observe a small reflection which is made visible by the added artificial light.

We now consider in Figure 11 the interaction of a slow soliton centered at $x = 7$ for a velocity $c = -4$ and a fast soliton of velocity $c = 16$ and centered at a point $x = -6$. We see that some small reflections can be visualized with the help of the light. The last example, presented in Figure 12, consists in the same interaction as in

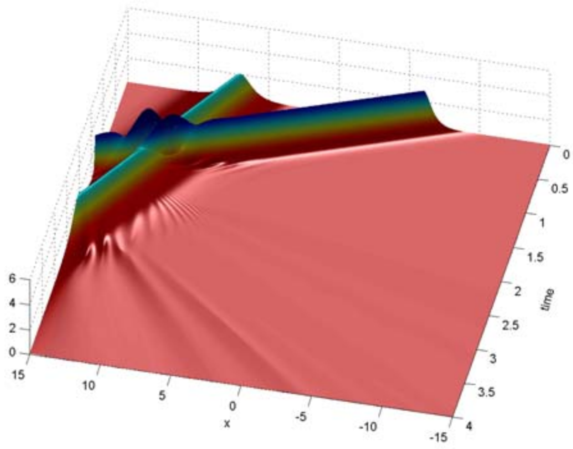


FIG. 12. *Interaction of slow and fast solitons with same directions.*

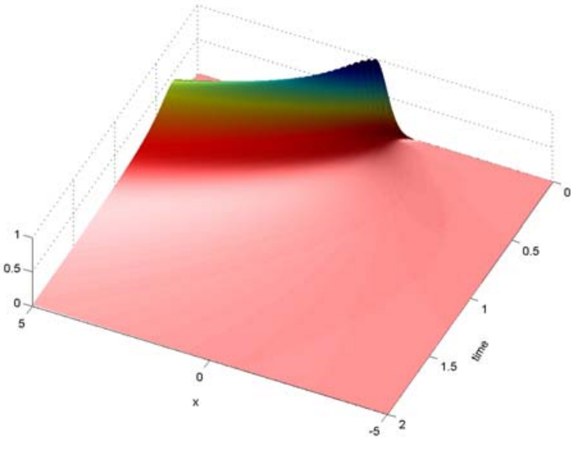


FIG. 13. *Propagation of the solution for the gaussian initial datum.*

the previous case but with the velocity $c = 4$ for the slowest soliton. In this situation, the two solitons interact near the right boundary. Once again, the artificial boundary condition reveals a satisfactory behavior and generates some relatively low-amplitude reflections (however more important than in the previous case).

To end the numerical experiments, we consider a gaussian initial datum

$$u(x, 0) = \exp(icx/2) \exp(-5x^2)$$

to observe the pure dispersion phenomenons involved in the NLS equation. The velocity is taken to $c = 15$. The finite spatial domain $\Omega_i = [-5, 5]$ is discretized with 4000 points, the time discretization being unchanged. As can be noticed in Figures 13 and 14, no reflection occurs and the dispersion is not affected by any wave reflected back into the computational domain.

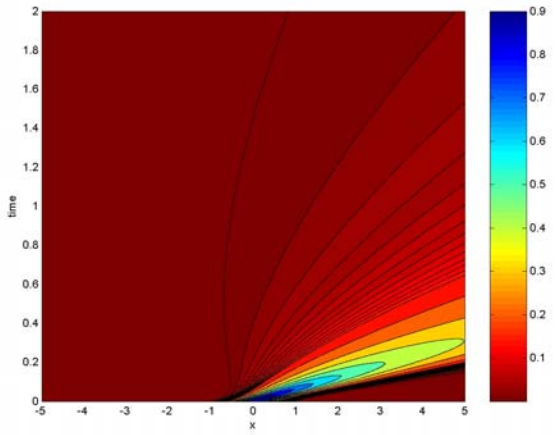


FIG. 14. *Contour plot of the solution for the gaussian initial datum.*

5. Conclusion. We have introduced different kinds of NLABCs for the one-dimensional nonlinear cubic Schrödinger equation. They are constructed with the help of some general pseudodifferential techniques usually involved in the derivation of artificial boundary conditions associated to linear operators. Stable and accurate semidiscretizations in time have been proposed. These conditions appear to be efficient for simulating the propagation of sufficiently fast solitons (simple soliton, interaction of two solitons). A loss of accuracy occurs for slower solutions but the artificial reflection is always much lower than for the LABC without any additional cost. These results are currently being extended for two-dimensional problems using the approach analyzed in [4].

REFERENCES

[1] X. Antoine, H. Barucq, and A. Bendali, *Bayliss-Turkel-like radiation condition on surfaces of arbitrary shape*, J. Math. Anal. Appl., 229 (1999), pp. 184–211.

[2] X. Antoine and C. Besse, *Construction, structure and asymptotic approximations of a microdifferential transparent boundary condition for the linear Schrödinger equation*, J. Math. Pures Appl., 80 (2001), pp. 701–738.

[3] X. Antoine and C. Besse, *Unconditionally stable discretization schemes of non-reflecting boundary conditions for the one-dimensional Schrödinger equation*, J. Comput. Phys., 181 (2003), pp. 157–175.

[4] X. Antoine, C. Besse, and V. Mouysset, *Numerical schemes for the simulation of the two-dimensional Schrödinger equation using non-reflecting boundary conditions*, Math. Comp., 73 (2004), pp. 1779–1799.

[5] A. Arnold, *Numerically absorbing boundary conditions for quantum evolution equations*, VLSI Design, 6 (1998), pp. 313–319.

[6] A. Arnold, *Mathematical concepts of open quantum boundary conditions*, Transport Theory Statist. Phys., 30 (2001), pp. 561–584.

[7] A. Arnold and M. Erhardt, *Discrete transparent boundary conditions for the Schrödinger equation*, Riv. Mat. Univ. Parma, 6 (2001), pp. 57–108.

[8] V.A. Baskakov and A.V. Popov, *Implementation of transparent boundaries for the numerical solution of the Schrödinger equation*, Wave Motion, 14 (1991), pp. 123–128.

[9] C. Besse and C.H. Bruneau, *Numerical study of elliptic-hyperbolic Davey-Stewartson system: Dromions simulation and blow-up*, Math. Models Methods Appl. Sci., 8 (1998), pp. 1363–1386.

[10] C. Besse, B. Bidégaray, and S. Descombes, *Order estimates in time of splitting method for the nonlinear Schrödinger equation*, SIAM J. Numer. Anal., 40 (2002), pp. 26–40.

[11] C.-H. Bruneau, L. Di Menza, and T. Lerhner, *Numerical resolution of some nonlinear Schrödinger-like equations in plasmas*, Numer. Methods Partial Differential Equations (1999), pp. 672–696.

[12] M. Delfour, M. Fortin, and G. Payre, *Finite-difference solutions of a nonlinear Schrödinger equation*, J. Comput. Phys., 44 (1981), pp. 277–288.

[13] A. Durán and J. M. Sanz-Serna, *The numerical integration of relative equilibrium solutions. The nonlinear Schrödinger equation*, IMA J. Numer. Anal., 20 (2000), pp. 235–261.

[14] B. Engquist and A. Majda, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (1977), pp. 629–651.

[15] B. Engquist and A. Majda, *Radiation boundary conditions for acoustic and elastic wave calculations*, Comm. Pure Appl. Math., 32 (1979), pp. 313–357.

[16] T. Friese, F. Schmidt, and D. Yevick, *A comparison of transparent boundary conditions for the Fresnel equation*, J. Comput. Phys., 168 (2001), pp. 433–444.

[17] R. Lascar, *Propagation des singularités des solutions d'équations pseudo-différentielles quasi-homogènes*, Ann. Inst. Fourier (Grenoble), 27 (1977), pp. 79–123.

[18] L. Nirenberg, *Pseudodifferential operators and some applications*, in Lectures on Linear Partial Differential Equations, CBMS Reg. Conf. Ser. Math. 17, AMS, Providence, RI, 1973, pp. 19–58.

[19] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[20] J. Szeftel, *Design of absorbing boundary conditions for Schrödinger equations in $\mathbb{R}^d$*, SIAM J. Numer. Anal., 42, (2004), pp. 1527–1551.

[21] V. E. Zakharov, *Theory of solitons, the inverse scattering method*, Contemp. Soviet Math. (1868).

[22] A. I. Zayed, *Handbook of Function and Generalized Function Transformation*, CRC Press, Boca Raton, FL, 1996.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2294–2319

A POSTERIORI ERROR ESTIMATES FOR FINITE ELEMENT APPROXIMATION OF PARABOLIC p-LAPLACIAN*

CARSTEN CARSTENSEN†, WENBIN LIU‡, AND NINGNING YAN§

Abstract. In this paper, we derive a posteriori error estimates in the quasi-norm for the finite element approximation of the parabolic p-Laplacian. We obtain a posteriori error bounds for the semidiscrete scheme and the fully backward Euler discretization. We show that the new a posteriori error estimators provide both upper and lower bounds on the discretization error.

Key words. finite element approximation, backward Euler discretization, parabolic p-Laplacian, a posteriori error estimators, quasi-norm error bounds

AMS subject classifications. 65N30, 49J40

DOI. 10.1137/040611008

1. Introduction.

In this paper, we derive a posteriori error estimates for the finite element approximation of the parabolic p-Laplacian with homogeneous Dirichlet data

$$
\begin{aligned}
u_t(x,t) - \operatorname{div}(|\nabla u(x,t)|^{p-2}\nabla u(x,t)) &= f(x,t), && x \in \Omega,\ t \in [0,T],\\
u(x,t) &= 0, && x \in \partial\Omega,\ t \in [0,T],\\
u(x,0) &= u_0(x), && x \in \Omega,
\end{aligned}
\tag{1.1}
$$

where $1 < p < \infty$ and Ω is a bounded open subset of R^2 with a Lipschitz boundary $\partial\Omega$. Assumptions on the data f and u_0 will be specified later. This equation is viewed as one of the typical examples of a large class of nonlinear problems—parabolic degenerate nonlinear systems, where many existing techniques (such as the linearization or deformation procedure) in the finite element method do not seem to work well.

Finite element approximations of the p-Laplacian have been extensively studied in the literature; see [Ci, GM, Ch] for some previous work. Sharp a priori error bounds were obtained in [BL1] and [LB] via the quasi-norm techniques; see [BL2] for an overview of some recent work. The quasi-norm approach, which has proved successful in deriving sharper a priori error bounds for the conforming finite element approximation of the degenerate systems, was summarized in [LY1, LY2] with a review of relevant recent work. Furthermore, sharp a priori error bounds in the space variable approximation were derived in [BL4] for the parabolic p-Laplacian, although the error bounds in time variable approximation are only suboptimal there. In [BB],

*Received by the editors July 5, 2004; accepted for publication (in revised form) February 9, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sinum/43-6/61100.html

†Institute of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, Office: Rudower Chaussee 25, 10099 Berlin, Germany (cc@math.hu-berlin.de). The work of this author was supported by the DFG Research Center MATHEON "Mathematics for Key Technologies" in Berlin and the DFG-Schwerpunktprogramm "Multi-Scale Problems" (SPP 1095).

‡Department of Mathematics, Xiangtan University, Xiangtan, Hunan, People's Republic of China (w.b.liu@ukc.ac.uk).

§LSEC, Institute of Systems Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing (100080), People's Republic of China (ynn@amss.ac.cn). The work of this author was supported by EPSRC research grant GR/R31980 and the Special Funds for Major State Basic Research Projects (G2000067102), and by the National Natural Science Foundation of China (60474027).

sharp a priori error bounds were obtained for both space and time approximation of the parabolic p-Laplacian. Another important area is a posteriori error estimation of the p-Laplacian. The work in this area seems to date back to [ODSD], and some of the recent work can be found in [BA, BL2, P, GS, V1], where among other things, a posteriori error estimates on the conforming and nonconforming discretization errors were derived with both upper and lower bounds. In these contributions, however, there are *gaps* in the power between the established upper and lower estimates. Recently in [LY1, LY2], the quasi-norm techniques were further developed, and improved a posteriori error estimates of residual type were derived for the p-Laplacian. Initial analysis and numerical tests indicate that the new estimators are sharper than the existing ones and, indeed, lead to more efficient computational meshes [CK, LY1].

It is the purpose of this work to extend our a posteriori error estimates to the finite element approximation of the parabolic p-Laplacian by combining the quasi-norm techniques developed in [LY1, LY2] and the weighted Clement-type interpolation introduced in [Ca, CF2, CF3, CF4] and further modified in [CLY].

The plan of this paper is as follows. In section 2 we state some important inequalities. In section 3 we give the parabolic p-Laplacian a variational formulation. We then set up the finite element approximation for the equation. We also introduce some quasi-norms and related results. In section 4, we derive a posteriori upper and lower error estimates in quasi-norm for the semidiscrete finite element approximation. In section 5, we derive quasi-norm a posteriori error estimates for the fully discrete scheme-backward Euler discretization. In the appendix, we introduce the weighted Clement-type interpolator and prove interpolation error estimates in the quasi-norm.

Let Ω be a bounded open set in R^2 with a Lipschitz boundary $\partial\Omega$. In this paper we adopt the standard notation $W^{m,q}(\Omega)$ for Sobolev spaces on Ω with norm $\|\cdot\|_{W^{m,q}(\Omega)}$ (or $\|\cdot\|_{m,q,\Omega}$ as a simplification) and seminorm $|\cdot|_{W^{m,q}(\Omega)}$ (or $|\cdot|_{(m,q,\Omega)}$). We set $W_0^{m,q}(\Omega) \equiv \{w \in W^{m,q}(\Omega) : w|_{\partial\Omega} = 0\}$. We denote $W^{m,2}(\Omega)$ by $H^m(\Omega)$ with norm $\|\cdot\|_{m,\Omega}$ and seminorm $|\cdot|_{m,\Omega}$. We also denote by $L^s(0,T;W^{m,q}(\Omega))$ the Banach space of L^s functions from $(0,T)$ into $W^{m,q}(\Omega)$ with norm $\|v\|_{L^s(0,T;W^{m,q}(\Omega))} = (\int_0^T \|v\|^s_{W^{m,q}(\Omega)} dt)^{\frac{1}{s}}$ for $s \in [1,\infty)$ with the standard modification for $s = \infty$. Similarly, one can define $H^1(0,T;W^{m,q}(\Omega))$ and $C^k(0,T;W^{m,q}(\Omega))$. In addition, c or C denotes a general positive constant independent of h, and $A \le CB$ is abbreviated as $A \lesssim B$. The generic constant C is allowed to depend only on p, Ω, and the aspect ratio of the finite elements.

2. Preliminaries.

The following inequalities play an essential role in our error analysis. Therein, the generic positive constant C in $A \le CB$ (abbreviated $A \lesssim B$) depends only on p. The first two lemmas have been used in our work on a priori quasi-norm error bounds for the finite element approximation of degenerate nonlinear PDEs [BL1, LB, BL3].

LEMMA 2.1 (see [BL3]). *For all $p > 1$, ξ, $\eta \in \mathbb{R}^n$, there holds*

$$||\xi|^{p-2}\xi - |\eta|^{p-2}\eta\,| \lesssim |\xi - \eta|(|\xi| + |\eta|)^{p-2}, \tag{2.1}$$

$$|\xi - \eta|^2(|\xi| + |\eta|)^{p-2} \lesssim (|\xi|^{p-2}\xi - |\eta|^{p-2}\eta, \xi - \eta). \tag{2.2}$$

LEMMA 2.2 (see [BL3, LY1, LY2]). *For all a, σ_1, $\sigma_2 \ge 0$, $p > 1$, $\theta > 0$, there holds*

$$(a + \sigma_1)^{p-2}\sigma_1\sigma_2 \le \theta^{-\gamma}(a + \sigma_1)^{p-2}\sigma_1^2 + \theta(a + \sigma_2)^{p-2}\sigma_2^2,$$

where

$$\gamma = \begin{cases} 1 & 1 < p \le 2,\ \theta \in [1, \infty) \quad \textit{or} \quad 2 < p < \infty,\ \theta \in (0, 1), \\ \dfrac{1}{p-1} & 1 < p \le 2,\ \theta \in (0, 1) \quad \textit{or} \quad 2 < p < \infty,\ \theta \in [1, \infty). \end{cases}$$

The following generalization of the Young inequality (for $a = 0$) is essential for estimating a bilinear form via the quasi-norms.

LEMMA 2.3 (see [LY1, LY2]). *For all $a, \sigma_1, \sigma_2 \ge 0$, $p > 1$, and $\delta > 0$, there holds*

$$\sigma_1 \sigma_2 \le \delta^{-\beta}(a^{p-1} + \sigma_1)^{p'-2}\sigma_1^2 + \delta(a + \sigma_2)^{p-2}\sigma_2^2,$$

where β is such that $\delta^{-\beta} = \max\{\delta^{-1}, \delta^{-\frac{1}{p-1}}\}$, and p' is such that $\frac{1}{p} + \frac{1}{p'} = 1$.

The following two lemmas will be used to prove some triangle-inequality like results for a power of the quasi-norms.

LEMMA 2.4 (see [LY1, LY2]). *For all $1 < p < \infty$, $\sigma_1, \sigma_2 \in \mathbb{R}^n$, and $a \ge 0$, there holds*

$$(a + |\sigma_1 + \sigma_2|)^{p-2}|\sigma_1 + \sigma_2|^2 \lesssim (a + |\sigma_1|)^{p-2}|\sigma_1|^2 + (a + |\sigma_2|)^{p-2}|\sigma_2|^2.$$

LEMMA 2.5 (see [LY1, LY2]). *For all $1 < p < \infty$ and $\sigma, \sigma_1, \sigma_2 \in \mathbb{R}^n$, there holds*

$$(|\sigma_1| + |\sigma_2|)^{p-2}|\sigma_1 - \sigma_2|^2 \lesssim (|\sigma| + |\sigma - \sigma_1|)^{p-2}|\sigma - \sigma_1|^2 + (|\sigma| + |\sigma - \sigma_2|)^{p-2}|\sigma - \sigma_2|^2.$$

3. Discretization of the parabolic p-Laplacian. In this section we consider the weak formulation of the parabolic p-Laplacian and its semidiscrete and full-discrete finite element approximation. We also introduce some quasi-norms which naturally arise in degenerate problems of this type.

The weak form (WP) of the p-Laplacian reads as follows: Given $f \in C(0, T; L^2(\Omega))$ and $u_0 \in W_0^{1,p}(\Omega)$, find $u \in L^\infty(0, T; W_0^{1,p}(\Omega)) \cap H^1(0, T; L^2(\Omega))$ such that

$$(u_t, v) + a(u, v) = (f, v) \qquad \forall v \in W_0^{1,p}(\Omega), \tag{3.1}$$

$$u(x, 0) = u_0(x),$$

where

$$a(u, v) = \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \quad \text{and} \quad (w, v) = \int_\Omega wv.$$

It can be shown that WP has a unique solution $u \in C([0, T], L^2(\Omega))$; see [BL4], for instance. There has been a great deal of work on the regularity of the solution u to WP. For sufficiently regular data, global $C^{1,\alpha}$ regularity was established in [DeB].

Let Ω^h be a polygonal approximation to Ω with boundary $\partial\Omega^h$. Let T^h be a partitioning of Ω^h into disjoint open regular triangles K, so that $\bar{\Omega}^h = \bigcup_{K \in T^h} \bar{K}$. Each element has at most one edge on $\partial\Omega^h$, and $\bar{K}$ and $\bar{K}'$ have either only one common vertex or a whole edge if K and $K' \in T^h$. We further require that $P_i \in \partial\Omega^h \Rightarrow P_i \in \partial\Omega$, where $\{P_i\}(i = 1, \dots, J)$ is the vertex set associated with the partitioning T^h. Let h_K denote the maximum diameter of the element K in T^h and let ρ_K denote the diameter of the largest ball contained in K. We assume that there is a regularity constant R of T^h, independent of h, such that $1 \le \max_{K \in T^h}(h_K/\rho_K) \le R$. Let $h = \max_{K \in T^h} h_K$.

Because of limited regularity for the solution of the p-Laplacian, we shall discuss only the conforming piecewise linear elements in this paper.

Associated with T^h is a finite-dimensional subspace V^h of $H_0^1(\Omega^h)$ such that $\chi|_K$ are linear functions for all $\chi \in V^h$ and $K \in T^h$. For ease of exposition we will assume that $\Omega^h = \Omega \subset R^2$, though all the results can be extended to the more general case, where $\Omega^h \subset \Omega$. Let

$$u_0^h \in V_0^h = \{v \in V^h : v = 0 \text{ on } \partial\Omega\}$$

be an approximation to u_0 and let $W^h = L^\infty(0,T;V^h) \cap H^1(0,T;V^h)$, $W_0^h = L^\infty(0,T;V_0^h) \cap H^1(0,T;V_0^h)$.

The weak form of the semidiscrete finite element approximation (WP^h) for (3.1) reads as follows: Find $u_h \in W_0^h$ such that

$$\left(\frac{\partial u_h}{\partial t}, v_h\right) + a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_0^h, \tag{3.2}$$

$$u_h(x,0) = u_0^h(x).$$

Furthermore, we can consider fully discrete approximation of WP. In this paper we consider the following the backward Euler discretization applied to (WP^h): (WP^{hk}).

Let $0 = t_0 < t_1 < t_2 < \cdots < t_{N-1} < t_N = T$, $k_n = t_n - t_{n-1}$, $I_n = (t_{n-1}, t_n]$, $n = 1, 2, \ldots, N$, $k = \max_n\{k_n\}$. Then for $n = 1, 2, \ldots, N$, find $U^n \in V^h$ such that

$$\left(\frac{U^n - U^{n-1}}{k_n}, v_h\right) + a(U^n, v_h) = (f(x,t_n), v_h) \qquad \forall v_h \in V_0^h, \tag{3.3}$$

$$U^0(x) = u_0^h(x).$$

For the purposes of the error analysis it is convenient to introduce the fact that for $t \in (t_{n-1}, t_n]$, $n = 1, 2, \ldots, N$,

$$U(x,t) = \frac{t - t_{n-1}}{k_n} U^n(x) + \frac{t_n - t}{k_n} U^{n-1}(x),$$

$$\hat{U}(x,t) = U(x,t_n), \quad \hat{f}(x,t) = f(x,t_n).$$

Then (WP^{hk}) can be restated as follow: For almost every $t \in (0,T]$ there holds

$$\left(\frac{\partial U}{\partial t}, v_h\right) + a(\hat{U}, v_h) = (\hat{f}, v_h) \quad \forall v_h \in V_0^h, \tag{3.4}$$

$$U(x,0) = u_0^h(x).$$

One of the key ideas in our approach is to introduce some quasi-norms to handle the degeneracy of the p-Laplacian in order to obtain sharp error bounds. We briefly introduce a quasi-norm and some relations between it and the standard Sobolev norms. Given $v, w \in W^{1,p}(\Omega)$, set

$$|v|_{(w,p)}^2 \equiv \int_\Omega |\nabla v|^2 (|\nabla w| + |\nabla v|)^{(p-2)}. \tag{3.5}$$

We shall simply write $|\cdot|_{(u,p)}$ as $|\cdot|_{(p)}$ when doing so causes no confusion.

PROPOSITION 3.1. (i) *There holds* $|v|_{(w,p)} \geq 0$ *and, when* $v \in W_0^{1,p}(\Omega)$, $|v|_{(w,p)} = 0$ *if and only if* $v = 0$.

(ii) *There holds* $|v_1 + v_2|_{(w,p)} \lesssim |v_1|_{(w,p)} + |v_2|_{(w,p)}$ *for any* $v_1, v_2 \in W^{1,p}(\Omega)$.

(iii) *Furthermore, for* $1 < p \leq 2$*, there holds*

$$|v|_{W^{1,p}(\Omega)} \lesssim (|w|_{W^{1,p}(\Omega)}, |v|_{W^{1,p}(\Omega)})|v|_{(w,p)} \quad \textit{and} \quad |v|^2_{(w,p)} \leq |v|^p_{W^{1,p}(\Omega)}. \tag{3.6}$$

(iv) *For* $2 \leq p < \infty$*,* $s \in [2, p]$*,* $r = s(2-p)/(2-s)$*, there holds*

$$|v|^p_{W^{1,p}(\Omega)} \leq |v|^2_{(w,p)} \leq C(|w|_{W^{1,r}(\Omega)}, |v|_{W^{1,r}(\Omega)})|v|^2_{W^{1,s}(\Omega)}. \tag{3.7}$$

Proof. Conclusion (ii) can be proved with Lemma 2.4. The rest of the proposition can be shown as in [BL3]. We, therefore, omit the details. □

The essential relations between the quasi-norm and the equation are reflected in the following inequalities. If u solves WP and $v \in W^{1,p}(\Omega)$, then it follows from Lemma 2.1 that

$$|u - v|^2_{(u,p)} \lesssim a(u, u - v) - a(v, u - v). \tag{3.8}$$

For any $\theta > 0$, $v, w \in W^{1,p}(\Omega)$, it follows from Lemmas 2.1 and 2.2 that there exists a $\gamma > 0$ such that

$$|a(u, w) - a(v, w)| \lesssim \theta^\gamma |u - v|^{\tilde{2}}{}_{(u,p)} + \theta |w|^2_{(u,p)}. \tag{3.9}$$

Then it follows from (3.8)–(3.9) that for any $u, v \in W^{1,p}(\Omega)$,

$$a(u, u - v) - a(v, u - v) \lesssim |u - v|^2_{(u,p)} \lesssim a(u, u - v) - a(v, u - v).$$

Thus the quasi-norm is naturally related to the total energy difference.

The relations (3.8)–(3.9) are important in proving the following optimal a priori error bound in the quasi-norm [BL1, LB] for the finite element approximation of the p-Laplacian:

$$|u - u_h|^2_{(p)} \lesssim \min_{v_h \in V_0^h} |u - v_h|^2_{(p)}.$$

Explicit error bounds can then be obtained. For example, if $1 < p \leq 2$, one has the optimal a priori error bound in $W^{1,p}$ ([BL1]),

$$\|u - u_h\|_{W^{1,p}} \lesssim h,$$

provided u is smooth enough. Furthermore, (3.8) and (3.9) are among the keys to the proof of the optimal a priori error bound for the parabolic p-Laplacian in [BL4]. For the semidiscrete finite element approximation of the parabolic p-Laplacian, for example, one can show that for sufficiently smooth u and almost all $s \in (0, T]$

$$\|(u - u_h)(s)\|^2_{L^2(\Omega)} + \int_0^s |u - u_h|^2_{(p)}\, dt \leq Ch^2.$$

Remark 3.1. *In the a priori error analysis, the error* $|u - u_h|_{(w,p)}$ *is considered with* $w = u$*. To make this norm computable within an a posteriori error analysis, one considers* $w = u_h$*. The triangle inequality shows equivalence of the two error terms, that is,*

$$|u - u_h|_{(u,p)} \lesssim |u - u_h|_{(u_h,p)} \lesssim |u - u_h|_{(u,p)}.$$

Recently, the quasi-norm techniques have been further developed towards a posteriori error estimates for the p-Laplacian [LY1, LY2]. For instance, let u_h be the finite element approximation of the p-Laplacian and let $1/p + 1/p' = 1$. Then

$$\eta_1^2 + \eta_2^2 - \epsilon_1 \lesssim |u - u_h|_{(p)}^2 \lesssim (\eta_1^2 + \eta_2^2) + \epsilon_2,$$

where ε_1 and ε_2 are higher order terms and

$$\begin{aligned} |u - u_h|_{(p)}^2 &= \int_\Omega (|\nabla u| + |\nabla(u - u_h)|)^{p-2} |\nabla(u - u_h)|^2, \\ \eta_1^2 &= \sum_K \int_K (|\nabla u_h|^{p-1} + h_K|f|)^{p'-2} h_K^2 |f|^2, \\ \eta_2^2 &= \sum_l \int_{K_l} (|\nabla u_h|^{p-1} + |A_l|)^{p'-2} A_l^2, \end{aligned}$$

where A_l is the jump of the A-normal derivative of $u_h \in V^h$ on the edge $l = \bar{K}_l^1 \cap \bar{K}_l^2$,

$$A_l = ((|\nabla u_h|^{p-2} \nabla u_h)_{K_l^1} - (|\nabla u_h|^{p-2} \nabla u_h)_{K_l^2})n$$

with n being the unit normal vector on $l = \bar{K}_l^1 \cap \bar{K}_l^2$ outwards K_l^1.

In the next section, we shall derive such a posteriori error estimates for the finite element approximation of the parabolic p-Laplacian. To this end let us introduce a weighted Clement-type interpolator in the finite element space V_0^h introduced in [Ca].

DEFINITION 3.1. *Let D be the set of nodes,*

$$\Lambda = \{z \in D : z \in \partial\Omega\}.$$

Given the nodal basis function φ_z of z in V^h, set $\omega_z = \{x \subset \Omega : \varphi_z(x) > 0\}$,

$$\psi_z = \varphi_z / \psi \quad \textit{and} \quad \psi = \sum_{z \in \Lambda} \varphi_z.$$

Then, for all $v \in W_0^{1,p}(\Omega)$, define the interpolation of v by

$$\pi v = \sum_{z \in \Lambda} v_z \varphi_z \in V_0^h, \quad v_z = \left(\int_\Omega \psi_z v \right) \Big/ \left(\int_\Omega \varphi_z \right).$$

Some interpolation error estimates for the interpolator in the quasi-norm will be given in the appendix.

4. A posteriori error estimates for the semidiscrete scheme.

In the following sections we derive a posteriori error estimates for the semidiscrete finite element approximation of WP. First we need some further notation. Let l be an edge of an element $K \in T^h$. If l is on the boundary of Ω^h, then we define the element $K_{max}^l = K_{min}^l = K$. Otherwise let $l = \bar{K}_l^1 \cap \bar{K}_l^2$, where K_l^1, K_l^2 are the two elements sharing the common edge l. Then we define the element $K_{max}^l(K_{min}^l) = K_l^i$ (i = 1 or 2) be such that

$$|\nabla u_h|_{K_{max}^l}^{p-2} = \max_{i=1,2} \{|\nabla u_h|_{K_l^i}^{p-2}\} \quad \text{and} \quad |\nabla u_h|_{K_{min}^l}^{p-2} = \min_{i=1,2} \{|\nabla u_h|_{K_l^i}^{p-2}\}.$$

We will take $K_{max}^l = K_{min}^l = K_l^1$ just for fixing the idea if $|\nabla u_h|_{K_l^1}^{p-2} = |\nabla u_h|_{K_l^2}^{p-2}$. Let $[w]_l = w|_{K_l^1} - w|_{K_l^2}$. The purpose of introducing K_{min} and K_{max} is to make the estimators η and η_2 below sharper. The necessity of introducing them will be discussed after Lemmas 4.1 and 4.2.

THEOREM 4.1. *Let u and u_h be solutions of (3.1) and (3.2), respectively. Let $p > 1$ and $p' > 1$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$. Assume that $f \in L^1(0,T; W^{1,p'}(\Omega))$. Then there exists a $\delta_0 > 0$ such that for all $s \in [0,T]$, $0 < \delta \le \delta_0$, there exist C, C_1 (where only C_1 depends on δ) such that*

$$\|(u-u_h)(s)\|_{0,\Omega}^2 + \int_0^s |u-u_h|_{(p)}^2 \, dt \le C_1\eta^2 + C\eta_0^2 + C_1\eta_1^2 + C\delta\eta_2^2, \tag{4.1}$$

with

$$\begin{aligned}
\eta^2 &= \int_0^s \sum_{l\cap\partial\Omega=\emptyset} \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |A_l|)^{p'-2}|A_l|^2 \, dx\, dt,\\
\eta_0^2 &= \|u_0 - u_0^h\|_{0,\Omega}^2,\\
\eta_1^2 &= \int_0^s \sum_K \int_K \Big(|\nabla u_h|^{p-1} + h_K^2\Big|\nabla\Big(f - \frac{\partial u_h}{\partial t}\Big)\Big|\Big)^{p'-2} h_K^4 \Big|\nabla\Big(f-\frac{\partial u_h}{\partial t}\Big)\Big|^2 \, dx\, dt,\\
\eta_2^2 &= \int_0^s \sum_{l\cap\partial\Omega=\emptyset} \int_{K^l_{min}} \Big(|\nabla u_h| + \Big|\Big[\frac{\partial u_h}{\partial n}\Big]_l\Big|\Big)^{p-2}\Big|\Big[\frac{\partial u_h}{\partial n}\Big]_l\Big|^2,
\end{aligned}$$

where $[\frac{\partial u_h}{\partial n}]_l$ is the jump of the normal derivative of u_h on the edge l and A_l is the jump of the A-normal derivative of $u_h \in V^h$ on the edge $l = \bar{K}_l^1 \cap \bar{K}_l^2$,

$$A_l = \big((|\nabla u_h|^{p-2}\nabla u_h)_{K_l^1} - (|\nabla u_h|^{p-2}\nabla u_h)_{K_l^2}\big)n,$$

with n being the unit normal vector on $l = \bar{K}_l^1 \cap \bar{K}_l^2$ outwards K_l^1.

Proof. Let $e = u - u_h$, $e_I(x,t) = \pi e(x,t) \in V_0^h$ be the weighted Clement-type interpolation of $e(x,t)$ defined in Definition 3.1 for almost all $t \in [0,T]$. It follows from Lemma 2.1 and (3.1), (3.2) that

$$\begin{aligned}
&\frac{1}{2}\|(u-u_h)(s)\|_{0,\Omega}^2 + c\int_0^s |u-u_h|_{(p)}^2 dt\\
&\le \frac{1}{2}\|(u-u_h)(0)\|_{0,\Omega}^2 + \int_0^s\int_\Omega \frac{\partial}{\partial t}(u-u_h)e + \int_0^s\int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla u_h|^{p-2}\nabla u_h)\nabla e\\
&= \frac{1}{2}\eta_0^2 + \int_0^s\int_\Omega \frac{\partial}{\partial t}(u-u_h)(e-e_I) + \int_0^s\int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla u_h|^{p-2}\nabla u_h)\nabla(e-e_I)\\
&= \frac{1}{2}\eta_0^2 + \int_0^s\int_\Omega \Big(f - \frac{\partial u_h}{\partial t}\Big)(e-e_I) - \int_0^s \sum_K \int_{\partial K} |\nabla u_h|^{p-2}\frac{\partial u_h}{\partial n}(e-e_I)\\
&= \frac{1}{2}\eta_0^2 + I_1 + I_2,
\end{aligned}$$

where the constant c results from that in Lemma 2.1. It follows from Lemma 6.4 that for any $\delta_1 > 0$ there exist constants C and C_1 (where only C_1 depends on δ_1) such that

$$\begin{aligned}
I_1 &= \int_0^s \int_\Omega \Big(f - \frac{\partial u_h}{\partial t}\Big)(e - e_I) \\
&\leq C_1 \int_0^s \sum_K \int_K \Big(|\nabla u_h|^{p-1} + h_K^2 \Big|\nabla \Big(f - \frac{\partial u_h}{\partial t}\Big)\Big|\Big)^{p'-2} h_K^4 \Big|\nabla\Big(f - \frac{\partial u_h}{\partial t}\Big)\Big|^2 \\
&\quad + C\delta_1 \int_0^s \sum_K \int_K (|\nabla u_h| + |\nabla e|)^{p-2} |\nabla e|^2 \\
&\quad + C\delta_1 \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K^l_{min}} \Big(|\nabla u_h| + \Big|\Big[\frac{\partial u_h}{\partial n}\Big]_l\Big|\Big)^{p-2} \Big|\Big[\frac{\partial u_h}{\partial n}\Big]_l\Big|^2 \\
&\leq C_1 \eta_1^2 + C\delta_1 \bigg(\int_0^s |u - u_h|^2_{(p)} + \eta_2^2 \bigg).
\end{aligned}$$

Similarly, by Lemmas 2.3, 6.3, and 6.5, for any $\delta_2 > 0$

$$\begin{aligned}
I_2 &= -\int_0^s \sum_K \int_{\partial K} |\nabla u_h|^{p-2} \frac{\partial u_h}{\partial n}(e - e_I) = -\int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_l A_l (e - e_I) \\
&\lesssim \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K^l_{max}} |A_l| (h^{-1}_{K^l_{max}} |e - e_I| + |\nabla(e - e_I)|) \\
&\lesssim \delta_2^{-\beta} \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |A_l|)^{p'-2} A_l^2 \\
&\quad + \delta_2 \int_0^s \sum_K \int_K (|\nabla u_h| + h_K^{-1} |e - e_I|)^{p-2} h_K^{-2} |e - e_I|^2 \\
&\quad + \delta_2 \int_0^s \sum_K \int_K (|\nabla u_h| + |\nabla(e - e_I)|)^{p-2} |\nabla(e - e_I)|^2 \\
&\lesssim \delta_2^{-\beta} \eta^2 + \delta_2 \bigg(\int_0^s |u - u_h|^2_{(p)} + \eta_2^2 \bigg).
\end{aligned}$$

Hence, by letting $\delta_0 = \frac{c}{4C}$, we have that for all $0 < \delta \leq \delta_0$, there exist C, C_1 (where only C_1 depends on δ) such that

$$\|(u - u_h)(s)\|^2_{2,\Omega} + \int_0^s |u - u_h|^2_{(p)}\, dt \leq C\eta_0^2 + C_1(\eta_1^2 + \eta^2) + C\delta\eta_2^2\ .$$

This proves (4.1). □

Remark 4.1. *Let P_k denote the space of k-degree polynomials and set*

$$V_h^k = \{v \in C^1(\bar{\Omega}) \cap H_0^1(\Omega) : v|_K \in P_k\, \forall K \in T^h\}.$$

Lemmas 2.5 *and* 6.6 *imply, for all $v_h^k \in V_h^k$, that*

$$\begin{aligned}
\eta_2^2 &= \sum_{l \cap \partial\Omega = \emptyset} \int_{K^l_{min}} \Big(|\nabla u_h| + \Big|\Big[\frac{\partial u_h}{\partial n}\Big]_l\Big|\Big)^{p-2} \Big|\Big[\frac{\partial u_h}{\partial n}\Big]_l\Big|^2 \\
&\lesssim \sum_{l \cap \partial\Omega = \emptyset} h_{K^l_{min}} \int_l \Big(|\nabla u_h|_{K^l_{min}} + \Big|\Big[\frac{\partial u_h}{\partial n}\Big]_l\Big|\Big)^{p-2} \Big|\Big[\frac{\partial u_h}{\partial n}\Big]_l\Big|^2
\end{aligned}$$

$$
\begin{aligned}
&\lesssim \sum_{l\cap\partial\Omega=\emptyset} h_{K^l_{min}} \int_l \Big(|\nabla u_h|_{K^l_{min}} + \Big|\Big[\frac{\partial u_h}{\partial n} - \frac{\partial v_h^k}{\partial n}\Big]_l\Big|\Big)^{p-2} \Big|\Big[\frac{\partial u_h}{\partial n} - \frac{\partial v_h^k}{\partial n}\Big]_l\Big|^2 \\
&\lesssim \sum_{l\cap\partial\Omega=\emptyset} \int_{K_l^1\cup K_l^2} \big(|\nabla u_h|_{K^l_{min}} + |\nabla(u_h - v_h^k)|\big)^{p-2} |\nabla(u_h - v_h^k)|^2 \\
&\lesssim \sum_K \int_K \big(|\nabla u_h| + |\nabla(u_h - v_h^k)|\big)^{p-2} |\nabla(u_h - v_h^k)|^2 \\
&\lesssim |u-u_h|^2_{(p)} + |u - v_h^k|^2_{(p)}.
\end{aligned}
$$

Let v_h^k be the Hermite interpolation of u in V_h^k. Then, if u is smooth enough, say $u \in W^{1+\frac{2}{p},p}(\Omega)$ when $1 < p \leq 2$ or $u \in W^{2,p}(\Omega)$ when $p > 2$, $|u - v_h^k|^2_{(p)} = o(h^2)$. Moreover, note that $\eta_0^2 = \|u_0 - u_0^h\|^2_{0,\Omega} = O(h^4)$, $\eta_1^2 \leq Ch^{2s}$, $s = \min\{p', 2\}$ if $f, \frac{\partial u}{\partial t} \in L^1(0,T; W^{1,p'}(\Omega))$.

REMARK 4.2. *It follows from the proofs of Theorem 4.1 and Lemma 6.4 that for the case where f is not so smooth, we still have the a posteriori error estimates if we replace η_1^2 by*

$$\bar{\eta}_1^2 = \int_0^s \sum_K \int_K \Big(|\nabla u_h|^{p-1} + \Big|f - \frac{\partial u_h}{\partial t}\Big|\Big)^{p'-2} h_K^2 \Big|f - \frac{\partial u_h}{\partial t}\Big|^2 \, dx\, dt.$$

To derive a lower bound we need two lemmas.

LEMMA 4.1. *Let l be an edge shared by two elements $K_l^1, K_l^2 \in T^h$: $l = \bar{K}_l^1 \cap \bar{K}_l^2$. Let K^l_{max} be defined as in the beginning of this section. Then for any constant A, there exists a function $w_l \in H_0^1(\Omega)$ such that there holds $w_l|_{\Omega\setminus(\bar{K}_l^1\cup\bar{K}_l^2)} = 0$,*

$$\int_l A w_l = \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |A|)^{p'-2} A^2, \tag{4.2}$$

$$\|w_l\|_{0,\infty} \lesssim h_{K^l_{max}} \Big(|\nabla u_h|^{p-1}_{K^l_{max}} + |A|\Big)^{p'-2} |A|, \tag{4.3}$$

$$\int_{K_l^1\cup K_l^2} (|\nabla u_h| + |\nabla w_l|)^{p-2} |\nabla w_l|^2 \lesssim \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |A|)^{p'-2} A^2. \tag{4.4}$$

LEMMA 4.2. *For any element $K \in T^h$ and any constant f, there exists a polynomial w_K on K such that there holds $w_K|_{\partial K} = 0$,*

$$\int_K f w_K = \int_K (|\nabla u_h|^{p-1} + h_K|f|)^{p'-2} h_K^2 |f|^2,$$

$$\|w_K\|_{0,\infty} \lesssim (|\nabla u_h|^{p-1} + h_K|f|)^{p'-2} h_K^2 |f|,$$

$$\int_K (|\nabla u_h| + |\nabla w_K|)^{p-2} |\nabla w_K|^2 \lesssim \int_K \big(|\nabla u_h|^{p-1} + h_K|f|\big)^{p'-2} h_K^2 |f|^2.$$

Lemmas 4.1 and 4.2 have been proved in [LY1, LY2]. In the fact, let w_l in Lemma 4.1 be such that $w_l = \alpha_l \lambda_1 \lambda_2$, where λ_1, λ_2 are the base functions of linear triangular elements (barycentric co-ordinates) for the two vertices on l, and

$$\alpha_l = \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |A|)^{p'-2} |A|^2 \Big/ \int_l A \lambda_1 \lambda_2.$$

Let w_K in Lemma 4.2 be such that $w_K = \alpha_K \lambda_1 \lambda_2 \lambda_3$, where λ_1, λ_2, and λ_3 are the base functions of linear triangular elements on three vertices of K,

$$\alpha_K = \frac{\int_K (|\nabla u_h|^{p-1} + h_K |f|)^{p'-2} h_K^2 |f|^2}{\int_K f \lambda_1 \lambda_2 \lambda_3}.$$

Then one can check that all the conclusions in the Lemmas 4.1 and 4.2 hold. It can be shown that the first two inequalities in Lemma 4.1 still hold without using K_{max}. However, there are counterexamples for the third inequality. Thus it is necessary to introduce K_{max} (K_{min}) in this sense.

THEOREM 4.2. *Let u and u_h be the solutions of* (3.1) *and* (3.2)*, respectively. Let $p > 1$ and $p' > 1$ be such that $\frac{1}{p} + \frac{1}{p'} = 1$. Assume that $f, \frac{\partial u}{\partial t} \in L^1(0,T; L^{p'}(\Omega))$. Then*

$$\eta^2 \lesssim \int_0^s |u - u_h|_{(p)}^2 \, dt + \epsilon,$$

with $\bar{f}|_K = \int_K f/|K|$, $\overline{\frac{\partial u}{\partial t}}|_K = \int_K \frac{\partial u}{\partial t}/|K|$, and

$$\begin{aligned}\epsilon = &\int_0^s \sum_K \int_K \left(|\nabla u_h|^{p-1} + h_K |f - \bar{f}|\right)^{p'-2} h_K^2 |f - \bar{f}|^2 \\ &+ \int_0^s \sum_K \int_K \left(|\nabla u_h|^{p-1} + h_K \left|\frac{\partial u}{\partial t} - \overline{\frac{\partial u}{\partial t}}\right|\right)^{p'-2} h_K^2 \left|\frac{\partial u}{\partial t} - \overline{\frac{\partial u}{\partial t}}\right|^2,\end{aligned}$$

Proof. It follows from Lemmas 2.1, 2.2, and 4.1 that, for any $\theta_1, \theta_2 > 0$,

$$\begin{aligned}\eta^2 &= \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K^l_{max}} (|\nabla u_h|^{p-1} + |A_l|)^{p'-2} A_l^2 = \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_l A_l w_l \\ &= \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_l \left[|\nabla u_h|^{p-2} \frac{\partial u_h}{\partial n} - |\nabla u|^{p-2} \frac{\partial u}{\partial n}\right]_l w_l \\ &= \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{\partial K_l^1 \cup \partial K_l^2} \left(|\nabla u_h|^{p-2} \frac{\partial u_h}{\partial n} - |\nabla u|^{p-2} \frac{\partial u}{\partial n}\right) w_l \\ &= \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \left(\int_{K_l^1 \cup K_l^2} (|\nabla u_h|^{p-2} \nabla u_h - |\nabla u|^{p-2} \nabla u) \nabla w_l + \int_{K_l^1 \cup K_l^2} \left(f - \frac{\partial u}{\partial t}\right) w_l\right) \\ &\lesssim \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K_l^1 \cup K_l^2} (|\nabla u_h| + |\nabla (u - u_h)|)^{p-2} |\nabla (u_h - u)| \, |\nabla w_l| \\ &\quad + \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K_l^1 \cup K_l^2} \left|f - \frac{\partial u}{\partial t}\right| h_{K^l_{max}} \left(|\nabla u_h|^{p-1}_{K^l_{max}} + |A_l|\right)^{p'-2} |A_l| \\ &\lesssim \theta_1^{-\gamma} \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K_l^1 \cup K_l^2} (|\nabla u_h| + |\nabla (u_h - u)|)^{p-2} |\nabla (u_h - u)|^2 \\ &\quad + \theta_1 \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K_l^1 \cup K_l^2} (|\nabla u_h| + |\nabla w_l|)^{p-2} |\nabla w_l|^2 \\ &\quad + \theta_2^{-\gamma} \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K_l^1 \cup K_l^2} \left(|\nabla u_h|^{p-1}_{K^l_{max}} + |A_l|\right)^{p'-2} A_l^2\end{aligned}$$

$$+\theta_2\int_0^s \sum_{l\cap\partial\Omega=\emptyset}\int_{K_l^1\cup K_l^2}\Big(|\nabla u_h|^{p-1}_{K^l_{max}}+h_{K^l_{max}}\Big|f-\frac{\partial u}{\partial t}\Big|\Big)^{p'-2}h^2_{K^l_{max}}\Big|f-\frac{\partial u}{\partial t}\Big|^2$$
$$\lesssim \theta_1^{-\gamma}\int_0^s |u-u_h|^2_{(p)}+\theta_2 L+C(\theta_1+\theta_2^{-\gamma})\eta^2,$$

where γ is defined in Lemma 2.2, and

$$L=\int_0^s\sum_K\int_K\Big(|\nabla u_h|^{p-1}+h_K\Big|f-\frac{\partial u}{\partial t}\Big|\Big)^{p'-2}h_K^2\Big|f-\frac{\partial u}{\partial t}\Big|^2.$$

Let $\theta_1=\frac{1}{4C}$, $\theta_2=(4C)^{\frac{1}{\gamma}}$. Then we have

$$\eta^2\lesssim\int_0^s|u-u_h|^2_{(p)}\,dt+L. \tag{4.5}$$

Let $F=f-\frac{\partial u}{\partial t}$. It follows from Lemma 2.4 that

$$\begin{aligned}L&=\int_0^s\sum_K\int_K(|\nabla u_h|^{p-1}+h_K|F|)^{p'-2}h_K^2|F|^2\\&\lesssim\int_0^s\sum_K\int_K(|\nabla u_h|^{p-1}+h_K|\bar F|)^{p'-2}h_K^2|\bar F|^2\\&\quad+\int_0^s\sum_K\int_K(|\nabla u_h|^{p-1}+h_K|f-\bar f|)^{p'-2}h_K^2|f-\bar f|^2\\&\quad+\int_0^s\sum_K\int_K\Big(|\nabla u_h|^{p-1}+h_K\Big|\frac{\partial u}{\partial t}-\overline{\frac{\partial u}{\partial t}}\Big|\Big)^{p'-2}h_K^2\Big|\frac{\partial u}{\partial t}-\overline{\frac{\partial u}{\partial t}}\Big|^2\\&\lesssim I+\epsilon,\end{aligned} \tag{4.6}$$

where we have simply written $\bar f|_K$ as $\bar f$ and $\overline{\frac{\partial u}{\partial t}}|_K$ as $\overline{\frac{\partial u}{\partial t}}$. It follows from Lemma 4.2 that

$$\begin{aligned}I&=\int_0^s\sum_K\int_K(|\nabla u_h|^{p-1}+h_K|\bar F|)^{p'-2}h_K^2|\bar F|^2\\&=\int_0^s\sum_K\int_K\bar F w_K=\int_0^s\sum_K\int_K F w_K+\int_0^s\sum_K\int_K(\bar F-F)w_K\\&=I_1+I_2.\end{aligned} \tag{4.7}$$

It follows from Lemmas 2.1, 2.2, and 4.2 that, for any $\theta>0$,

$$\begin{aligned}I_1&=\int_0^s\sum_K\int_K F w_K=-\int_0^s\sum_K\int_K div\big(|\nabla u|^{p-2}\nabla u-|\nabla u_h|^{p-2}\nabla u_h\big)w_K\\&=\int_0^s\sum_K\int_K(|\nabla u|^{p-2}\nabla u-|\nabla u_h|^{p-2}\nabla u_h)\nabla w_K\\&\le C\int_0^s\sum_K\int_K\big(|\nabla u_h|+|\nabla(u-u_h)|\big)^{p-2}|\nabla(u-u_h)|\,|\nabla w_K|\\&\lesssim\theta_1^{-\gamma}\int_0^s\sum_K\int_K(|\nabla u_h|+|\nabla(u-u_h)|)^{p-2}|\nabla(u-u_h)|^2\end{aligned} \tag{4.8}$$

$$\begin{aligned}
&+\theta_1 \int_0^s \sum_K \int_K (|\nabla u_h| + |\nabla w_K|)^{p-2} |\nabla w_K|^2 \\
&\lesssim \theta_1^{-\gamma} \int_0^s |u-u_h|_{(p)}^2 + \theta_1 \int_0^s \sum_K \int_K (|\nabla u_h|^{p-1} + h_K |\bar{F}|)^{p'-2} h_K^2 |\bar{F}|^2 \\
&\approx \theta_1^{-\gamma} \int_0^s |u-u_h|_{(p)}^2 + \theta_1 I.
\end{aligned}$$

It follows from Lemmas 4.2 and 2.2 that, for any $\theta_2 > 0$,

$$\begin{aligned}
I_2 &= \int_0^s \sum_K \int_K (\bar{F} - F) w_K \leq \sum_K \int_K |F - \bar{F}| \, \|w_K\|_{0,\infty} \\
&\lesssim \int_0^s \sum_K \int_K |F - \bar{F}| (|\nabla u_h|^{p-1} + h_K |\bar{F}|)^{p'-2} h_K^2 |\bar{F}| \\
&\lesssim \theta_2^{-\gamma} \int_0^s \sum_K \int_K (|\nabla u_h|^{p-1} + h_K |\bar{F}|)^{p'-2} h_K^2 |\bar{F}|^2 \qquad (4.9) \\
&\quad + \theta_2 \int_0^s \sum_K \int_K (|\nabla u_h|^{p-1} + h_K |F - \bar{F}|)^{p'-2} h_K^2 |F - \bar{F}|^2 \\
&\lesssim \theta_2^{-\gamma} I + \theta_2 \epsilon.
\end{aligned}$$

From (4.7)–(4.9),

$$I \lesssim \theta_1^{-\gamma} \int_0^s |u-u_h|_{(p)}^2 + (\theta_1 + \theta_2^{-\gamma}) I + \theta_2 \epsilon.$$

Let $\theta_1 + \theta_2^{-\gamma} = \frac{1}{2C}$. Then

$$I \lesssim \int_0^s |u-u_h|_{(p)}^2 + \epsilon. \qquad (4.10)$$

It follows from (4.6) and (4.10) that

$$L \lesssim \int_0^s |u-u_h|_{(p)}^2 dt + \epsilon. \qquad (4.11)$$

Therefore, Theorem 4.2 follows from (4.5) and (4.11). □

REMARK 4.3. *Note that when* $f, \frac{\partial u}{\partial t} \in L^1(0,T;W^{1,p'}(\Omega))$, $\epsilon \lesssim h^{2s}$, $s = \min\{2, p'\}$. *Then, combined with the results of Remark* 4.1, *we have that when the solution is smooth enough,*

$$\int_0^s |u-u_h|_{(p)}^2 \, dt - \epsilon^* \lesssim \eta^2 \lesssim \int_0^s |u-u_h|_{(p)}^2 \, dt + \epsilon \quad \textit{with} \quad \epsilon^*, \epsilon = o(h^2).$$

5. A posteriori error estimates for the full discrete scheme.

THEOREM 5.1. *Let* u *and* U *be solutions of* (3.1) *and* (3.4), *respectively. Let* $p > 1$ *and* $p' > 1$ *be such that* $\frac{1}{p} + \frac{1}{p'} = 1$. *Assume that* $f \in L^1(0,T;W^{1,p'}(\Omega))$. *Then there exists a* $\delta_0 > 0$ *such that for all* $s \in [0,T]$, $0 < \delta \leq \delta_0$,

$$\|(u-U)(s)\|_{0,\Omega}^2 + \int_0^s |u - \hat{U}|_{(p)}^2 \, dt \lesssim \eta_0^2 + C(\delta)\big(\hat{\eta}_1^2 + \hat{\eta}_2^2\big) + \hat{\eta}_3^2 + \hat{\eta}_4^2 + \delta \hat{\eta}_5^2 \qquad (5.1)$$

with $\hat{A}_l = ((|\nabla\hat{U}|^{p-2}\nabla\hat{U})_{K_l^1} - (|\nabla\hat{U}|^{p-2}\nabla\hat{U})_{K_l^2})n$ *and* $\eta_0 = \|u_0 - u_0^h\|_{0,\Omega}$,

$$\hat{\eta}_1^2 = \int_0^s \sum_K \int_K \left(|\nabla\hat{U}|^{p-1} + h_K^2 \left|\nabla\left(\hat{f} - \frac{\partial U}{\partial t}\right)\right|\right)^{p'-2} h_K^4 \left|\nabla\left(\hat{f} - \frac{\partial U}{\partial t}\right)\right|^2 dx\,dt,$$

$$\hat{\eta}_2^2 = \int_0^s \sum_{l\cap\partial\Omega=\emptyset} \int_{K^l_{max}} (|\nabla\hat{U}|^{p-1} + |\hat{A}_l|)^{p'-2}|\hat{A}_l|^2,$$

$$\hat{\eta}_3^2 = \int_0^s \|f - \hat{f}\|^2_{L^2(\Omega)}\,dt, \qquad \hat{\eta}_4^2 = \int_0^s |U - \hat{U}|^2_{(\hat{U},p)}\,dt,$$

$$\hat{\eta}_5^2 = \int_0^s \sum_{l\cap\partial\Omega=\emptyset} \int_{K^l_{min}} \left(|\nabla\hat{U}| + \left|\left[\frac{\partial\hat{U}}{\partial n}\right]_l\right|\right)^{p-2} \left|\left[\frac{\partial\hat{U}}{\partial n}\right]_l\right|^2.$$

Proof. Let $\hat{e} = u - U$, $\hat{e}_I(x,t) = \pi_h\hat{e}(x,t) \in V_0^h$ be the interpolation of $\hat{e}(x,t)$ defined in Definition 3.1 for almost all $t \in [0,T]$. It follows from Lemma 2.1 and (3.1), (3.4) that

$$\begin{aligned}
&\frac{1}{2}\|(u-U)(s)\|^2_{0,\Omega} + \alpha\int_0^s |u - \hat{U}|^2_{(p)}\,dt \\
&\le \frac{1}{2}\|(u-U)(0)\|^2_{0,\Omega} + \int_0^s\int_\Omega \frac{\partial}{\partial t}(u-U)(u-U) \\
&\quad + \int_0^s\int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla\hat{U}|^{p-2}\nabla\hat{U})\nabla(u-\hat{U}) \\
&= \frac{1}{2}\hat{\eta}_0^2 + \int_0^s\int_\Omega \frac{\partial}{\partial t}(u-U)(\hat{e}-\hat{e}_I) + \int_0^s\int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla\hat{U}|^{p-2}\nabla\hat{U})\nabla(\hat{e}-\hat{e}_I) \\
&\quad + \int_0^s\int_\Omega \frac{\partial}{\partial t}(u-U)\hat{e}_I + \int_0^s\int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla\hat{U}|^{p-2}\nabla\hat{U})\nabla\hat{e}_I \\
&\quad + \int_0^s\int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla\hat{U}|^{p-2}\nabla\hat{U})\nabla(U-\hat{U}) \\
&= \frac{1}{2}\hat{\eta}_0^2 + \int_0^s\int_\Omega \left(f - \frac{\partial U}{\partial t}\right)(\hat{e}-\hat{e}_I) - \int_0^s\sum_K\int_{\partial K} |\nabla\hat{U}|^{p-2}\frac{\partial\hat{U}}{\partial n}(\hat{e}-\hat{e}_I) \\
&\quad + \int_0^s (f-\hat{f})\hat{e}_I + \int_0^s\int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla\hat{U}|^{p-2}\nabla\hat{U})\nabla(U-\hat{U}) \\
&= \frac{1}{2}\hat{\eta}_0^2 + \int_0^s\int_\Omega \left(\hat{f} - \frac{\partial U}{\partial t}\right)(\hat{e}-\hat{e}_I) - \int_0^s\sum_K\int_{\partial K} |\nabla\hat{U}|^{p-2}\frac{\partial\hat{U}}{\partial n}(\hat{e}-\hat{e}_I) \\
&\quad + \int_0^s (f-\hat{f})\hat{e} + \int_0^s\int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla\hat{U}|^{p-2}\nabla\hat{U})\nabla(U-\hat{U}) \\
&= \frac{1}{2}\eta_0^2 + I_1 + I_2 + I_3 + I_4.
\end{aligned}$$

Similarly, as in Theorem 4.1, it follows from Lemma 6.4 that

$$\begin{aligned}
I_1 &= \int_0^s \int_\Omega \Big(\hat{f} - \frac{\partial U}{\partial t}\Big)(\hat{e} - \hat{e}_I) \\
&\lesssim C(\delta_1) \int_0^s \sum_K \int_K \Big(|\nabla \hat{U}|^{p-1} + h_K^2 \Big|\nabla\Big(\hat{f} - \frac{\partial U}{\partial t}\Big)\Big|\Big)^{p'-2} h_K^4 \Big|\nabla\Big(\hat{f} - \frac{\partial U}{\partial t}\Big)\Big|^2 \\
&\quad + \delta_1 \int_0^s \sum_K \int_K (|\nabla \hat{U}| + |\nabla \hat{e}|)^{p-2} |\nabla \hat{e}|^2 \\
&\quad + \delta_1 \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K^l_{min}} \Big(|\nabla \hat{U}| + \Big|\Big[\frac{\partial \hat{U}}{\partial n}\Big]_l\Big|\Big)^{p-2} \Big|\Big[\frac{\partial \hat{U}}{\partial n}\Big]_l\Big|^2 \\
&\lesssim C(\delta_1)\hat{\eta}_1^2 + \delta_1 \Big(\int_0^s |u - U|^2_{(\hat{U},p)} + \hat{\eta}_5^2\Big).
\end{aligned}$$

By Lemmas 2.3, 6.3, and 6.5, we similarly have

$$\begin{aligned}
I_2 &= -\int_0^s \sum_K \int_{\partial K} |\nabla \hat{U}|^{p-2} \frac{\partial \hat{U}}{\partial n}(\hat{e} - \hat{e}_I) = -\int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_l \hat{A}_l (\hat{e} - \hat{e}_I) \\
&\lesssim \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K^l_{max}} |A_l| \Big(h^{-1}_{K^l_{max}} |\hat{e} - \hat{e}_I| + |\nabla(\hat{e} - \hat{e}_I)|\Big) \\
&\lesssim \delta_2^{-\beta} \int_0^s \sum_{l \cap \partial\Omega = \emptyset} \int_{K^l_{max}} (|\nabla \hat{U}|^{p-1} + |\hat{A}_l|)^{p'-2} \hat{A}_l^2 \\
&\quad + \delta_2 \int_0^s \sum_K \int_K (|\nabla \hat{U}| + h_K^{-1} |\hat{e} - \hat{e}_I|)^{p-2} h_K^{-2} |\hat{e} - \hat{e}_I|^2 \\
&\quad + \delta_2 \int_0^s \sum_K \int_K (|\nabla \hat{U}| + |\nabla(\hat{e} - \hat{e}_I)|)^{p-2} |\nabla(\hat{e} - \hat{e}_I)|^2 \\
&\lesssim \delta_2^{-\beta} \hat{\eta}_2^2 + \delta_2 \Big(\int_0^s |u - U|^2_{(\hat{U},p)} + \hat{\eta}_5^2\Big).
\end{aligned}$$

Then, by Lemma 2.4,

$$\begin{aligned}
I_1 + I_2 &\lesssim C(\delta_1, \delta_2)(\hat{\eta}_1^2 + \hat{\eta}_2^2) + (\delta_1 + \delta_2)\Big(\int_0^s |u - U|^2_{(\hat{U},p)} + \hat{\eta}_5^2\Big) \\
&\lesssim C(\delta_1, \delta_2)(\hat{\eta}_1^2 + \hat{\eta}_2^2) + (\delta_1 + \delta_2)\Big(\int_0^s |u - \hat{U}|^2_{(p)} + \hat{\eta}_4^2 + \hat{\eta}_5^2\Big).
\end{aligned}$$

It follows from the Schwarz inequality that

$$I_3 = \int_0^s \int_\Omega (f - \hat{f})\hat{e} \lesssim \int_0^s \|f - \hat{f}\|^2_{0,\Omega}\, dt + \int_0^s \|\hat{e}\|^2_{0,\Omega}\, dt \approx \hat{\eta}_3^2 + \int_0^s \|u - U\|^2_{0,\Omega}\, dt.$$

It follows from Lemmas 2.1 and 2.2 that

$$\begin{aligned}
I_4 &= \int_0^s \int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla \hat{U}|^{p-2}\nabla \hat{U})\nabla(U - \hat{U}) \\
&\lesssim \theta^{-\gamma} \int_0^s (|\nabla \hat{U}| + |\nabla(u - \hat{U})|)^{p-2} |\nabla(u - \hat{U})|^2 \\
&\quad + \theta \int_0^s (|\nabla \hat{U}| + |\nabla(U - \hat{U})|)^{p-2} |\nabla(U - \hat{U})|^2 \\
&\approx \theta^{-\gamma} \int_0^s |u - \hat{U}|^2_{(p)} + \theta \hat{\eta}_4^2.
\end{aligned}$$

Hence, letting $2\delta_0 + \theta^{-\gamma} = \frac{\alpha}{2C}$, we have that

$$\|(u-U)(s)\|_{0,\Omega}^2 + \int_0^s |u-\hat{U}|_{(p)}^2\,dt \lesssim \hat{\eta}_0^2 + C(\delta)(\hat{\eta}_1^2 + \hat{\eta}_2^2) + \hat{\eta}_3^2 + \hat{\eta}_4^2 + \delta\hat{\eta}_5^2 + \int_0^s \|u-U\|_{0,\Omega}^2 dt.$$

Then the estimate (5.1) follows from a Gronwall-type inequality. □

Theorem 5.2. *Let u and U be the solutions of* (3.1) *and* (3.4), *respectively. Let $p>1$ and $p'>1$ be such that $\frac{1}{p}+\frac{1}{p'}=1$. Assume that $f, \frac{\partial u}{\partial t} \in L^1(0,T;L^{p'}(\Omega))$, and that $c \le k_j/k_{j-1} \le C$. Then*

$$\hat{\eta}_2^2 \lesssim \int_0^s |u-\hat{U}|_{(p)}^2 dt + \hat{\epsilon}, \tag{5.2}$$

with

$$\hat{\epsilon} = \int_0^s \sum_K \int_K (|\nabla\hat{U}|^{p-1} + h_K|f-\bar{f}|)^{p'-2} h_K^2 |f-\bar{f}|^2$$
$$+ \int_0^s \sum_K \int_K \left(|\nabla\hat{U}|^{p-1} + h_K\left|\frac{\partial u}{\partial t} - \overline{\frac{\partial u}{\partial t}}\right|\right)^{p'-2} h_K^2 \left|\frac{\partial u}{\partial t} - \overline{\frac{\partial u}{\partial t}}\right|^2,$$

where $\bar{f}|_K = \int_K f/|K|$, $\overline{\frac{\partial u}{\partial t}}|_K = \int_K \frac{\partial u}{\partial t}/|K|$. Furthermore, for each $V \in C\,(0,T;W^{1,2}(\Omega))$ which is affine in time on each time interval $(t_{j-1},t_j]$, $j=1,\dots,N$, there holds

$$\hat{\eta}_4^2 \lesssim \int_0^s |u-\hat{U}|_{(p)}^2\,dt + \int_0^s |u-V|_{(p)}^2\,dt. \tag{5.3}$$

Proof. Similarly as in Theorem 4.2, it can be proved that

$$\hat{\eta}_2^2 \lesssim \int_0^s |u-\hat{U}|_{(p)}^2\,dt + L,$$

where

$$L = \int_0^s \sum_K \int_K \left(|\nabla\hat{U}|^{p-1} + h_K|f - \frac{\partial u}{\partial t}|\right)^{p'-2} h_K^2 \left|f - \frac{\partial u}{\partial t}\right|^2.$$

Similarly, as in Theorem 4.2, it also can be proved that

$$L \lesssim \int_0^s |u-\hat{U}|_{(p)}^2\,dt + \hat{\epsilon}.$$

Then (5.2) follows. For $x,y\ge 0$ and $1<p<\infty$, let

$$G(x,y) := \begin{cases} y^2(x+y)^{p-2} & \text{if } x+y>0, \\ 0 & \text{if } x=y=0. \end{cases} \tag{5.4}$$

Then

$$\hat{\eta}_4^2 := \int_0^s |U-\hat{U}|_{(\hat{U},p)}^2\,dt = \int_0^s \int_\Omega G(|\nabla\hat{U}|, |\nabla U - \nabla\hat{U}|)\,dx\,dt$$
$$= \sum_K \sum_{j=1}^n \int_K \int_{t_{j-1}}^{t_j} G(|\nabla\hat{U}|, |\nabla U - \nabla\hat{U}|)\,dt\,dx.$$

We consider one contribution

$$\int_K \int_{t_{j-1}}^{t_j} G(|\nabla \hat{U}|, |\nabla U - \nabla \hat{U}|)\, dt\, dx \le k_j \int_K G(|\nabla U_j|, |\nabla U_j - \nabla U_{j-1}|)\, dx. \tag{5.5}$$

The inequality follows from $(U - \hat{U})(t) = -(t_j - t)(U_j - U_{j-1})/k_j$ for $t \in I_j := (t_{j-1}, t_j)$ and $|(U - \hat{U})(t)| \le |U_j - U_{j-1}|$ combined with the monotonicity of $G(x, y)$ in y. It follows from Lemma 2.5 that for all $a, b, c \in R^n$,

$$\begin{aligned} G(|a|, |a-b|) &\lesssim (|a| + |b|)^{p-2}|b-a|^2 \lesssim G(|c|, |a-c|) + G(|c|, |b-c|) \\ &\lesssim G(|a|, |a-c|) + G(|b|, |b-c|). \end{aligned}$$

Given $V \in C(I_{j-1} \cup I_j; W^{1,2}(K))$, let $a := \nabla U_j$, $b := \nabla U_{j-1}$, and $c := \nabla V(t_{j-1}) = \nabla V_{j-1}$. Then

$$\begin{aligned} &\int_K G(|\nabla U_j|, |\nabla U_j - \nabla U_{j-1}|)\, dx \\ &\lesssim \int_K G(|\nabla U_{j-1}|, |\nabla U_{j-1} - \nabla V_{j-1}|)\, dx + \int_K G(|\nabla U_j|, |\nabla U_j - \nabla V_{j-1}|)\, dx. \end{aligned}$$

A direct calculation shows the estimate

$$\frac{k}{2} = \min_{m \in R} \int_0^k |1 + mt|\, dt$$

for all $k > 0$. Therefore, for any $b \in R$ and $k > 0$,

$$\frac{k}{2}\, |b| = \min_{m \in R} \int_0^k |b + mt|\, dt.$$

From this and a decomposition of $b \in R^n$ and $m \in R^n$ in the direction b and its orthogonal complement, one infers for all $b \in R^n$ and $k > 0$

$$\frac{k}{2}\, |b| = \min_{m \in R^n} \int_0^k |b + mt|\, dt = \min_{m \in R^n} \int_{-k}^0 |b + mt|\, dt.$$

Therefore, for all $b, m \in R^n$,

$$|b| \le \frac{2}{k_j} \int_{t_{j-1}}^{t_j} |b + m(t_j - s)|\, ds, \quad |b| \le \frac{2}{k_j} \int_{t_{j-1}}^{t_j} |b + m(t_{j-1} - s)|\, ds.$$

Since $G(|a|, \cdot)$ is monotone increasing and convex in $|\cdot|$,

$$G(|a|, |b|) \le G\left(|a|, \frac{2}{k_j} \int_{t_{j-1}}^{t_j} |b + m(t_{j-1} - s)|\, ds\right) \le \frac{1}{k_j} \int_{t_{j-1}}^{t_j} G(|a|, 2|b + m(t_{j-1} - s)|)\, ds.$$

The latter inequality follows Jensen's inequality for $G(|a|, |\cdot|)$ convex and so the value $G(|a|, 2\overline{b})$ for the mean value $\overline{b} = \frac{1}{k_j} \int_{t_{j-1}}^{t_j} |b + m(t_{j-1} - s)|\, ds$ of $|b + m(t_{j-1} - s)|$ is smaller than or equal to the mean value of $G(|a|, 2|b + m(t_{j-1} - s)|)$. Note that

$$\hat{U} - V = U_j - V_{j-1} + \frac{V_j - V_{j-1}}{k_j}(t_{j-1} - t).$$

The last inequality is employed to bound

$$\text{(5.6)}\quad k_j\int_K G(|\nabla U_j|,|\nabla U_j-\nabla V_{j-1}|)\,dx \le \int_{t_{j-1}}^{t_j}\int_K G(|\nabla U_j|,2|\nabla(\hat U-V)|)\,dxdt$$
$$\lesssim \int_K\int_{t_{j-1}}^{t_j} G(|\nabla U_j|,|\nabla\hat U-\nabla V)|\,dt\,dx.$$

Similarly (with another interval but the same argument), we have

(5.7)
$$k_j\int_K G(|\nabla U_{j-1}|,|\nabla U_{j-1}-\nabla V_{j-1}|)\,dx \lesssim \int_K\int_{t_{j-2}}^{t_{j-1}} G(|\nabla U_{j-1}|,|\nabla\hat U-\nabla V|)\,dtdx,$$

using the condition $c\le k_j/k_{j-1}\le C$ and

$$\hat U - V = U_{j-1}-V_{j-1}+\frac{V_{j-1}-V_{j-2}}{k_{j-1}}(t_{j-1}-t).$$

The combination of (5.5)–(5.7) proves that for each $V\in C(0,T;W^{1,2}(\Omega))$ which is affine in time on each time interval $(t_{j-1},t_j]$, $j=1,\dots,N$, there holds

$$\hat\eta_4^2\lesssim\int_0^s|\hat U-V|^2_{(\hat U,p)}\,dt.$$

From this estimate and the aforementioned estimate

$$G(|a|,|b-a|)\lesssim G(|c|,|a-c|)+G(|c|,|b-c|)\quad\forall\,a,b,c\in R^n,$$

we deduce

$$\hat\eta_4^2\lesssim\int_0^s\int_\Omega G(|\nabla\hat U|,|\nabla\hat U-\nabla V|)\,dt\,dx$$
$$\lesssim\int_0^s\int_\Omega G(|\nabla u|,|\nabla u-\nabla V|)\,dt\,dx+\int_0^s\int_\Omega G(|\nabla u|,|\nabla u-\nabla\hat U|)\,dt\,dx.$$

Then the estimate (5.3) follows. □

6. Appendix. In this section, we state some results on the interpolation error in the quasi-norm from [CLY]. For the readers' convenience, we include the proofs here. First, we prove a lemma which is a quasi-norm version of quotient theorem. We take a general approach here so that the results obtained can be applied to a class of degenerate systems.

Recall the definition of $G(x,y)$ from (5.4). Without further (explicit) notice, we shall use the fact that $G(x,y)$ is monotone increasing and convex with respect to the variable y.

First, we prove a quasi-norm version of the quotient theorem.

LEMMA 6.1. *Let Ω be a bounded connected open set in R^2. Let $1<p<\infty$ and $f\in(W^{1,p}(\Omega))^*$ with $P_0(\Omega)\cap\mathrm{Ker}(f)=\{0\}$. Then there exists a constant $c_1=c(f,p,\Omega)$ such that, for all $a\in R$, $a\ge0$, and $v\in W^{1,p}(\Omega)$,*

$$\int_\Omega G(a,|v|)\,dx\le c_1\,G(a,|f(v)|)+c_1\int_\Omega G(a,|\nabla v|)\,dx.$$

Proof. We argue by contradiction and suppose the lemma is false. Then there would exist a sequence v_j in $W^{1,p}(\Omega)$ with $\delta_j := \|v_j\|_{1,q} > 0$, $q = \min\{2,p\}$, and a sequence a_j of nonnegative real numbers such that

$$G(a_j, |f(v_j)|) + \int_\Omega G(a_j, |\nabla v_j|)\,dx \le 1/j \int_\Omega G(a_j, |v_j|)\,dx \tag{6.1}$$

for all $j \in N$. We observe in any case there exists a $u \in W^{1,q}(\Omega)$ with

$$u_j := v_j/\delta_j \text{ satisfies } \|u_j\|_{1,q} = 1, \quad u_j \rightharpoonup u \text{ in } W^{1,q}(\Omega). \tag{6.2}$$

Here we have chosen a weak convergent subsequence with Banach Alaoglu's theorem. In the first case we suppose that there exists a constant γ, $0 < \gamma < \infty$, with

$$a_j \le \gamma\,\delta_j \quad \text{for} \quad j = 1, 2, 3, \dots . \tag{6.3}$$

At least we suppose (6.3) for a subsequence we have not relabeled. If $1 < p \le 2$, then $G(a, x) \le x^p$ for all $x \ge 0$. Therefore, even without (6.3),

$$\int_\Omega G(a_j/\delta_j, |u_j|)\,dx \le \|u_j\|_p^p \le 1.$$

If $2 \le p$, then $G(\cdot, |u_j|)$ is monotone increasing. Hence, (6.2)–(6.3) yields

$$\int_\Omega G(a_j/\delta_j, |u_j|)\,dx \le \int_\Omega (\gamma + |u_j|)^{p-2}|u_j|^2\,dx \le \|\gamma + |u_j|\|_p^p \le (1 + \gamma|\Omega|^{1/p})^p.$$

Hence, for all $1 < p < \infty$, $\int_\Omega G(a_j/\delta_j, |u_j|)\,dx$ is bounded. A scaling of (6.1) then shows

$$\lim_{j\to\infty} \int_\Omega G(a_j/\delta_j, |\nabla u_j|)\,dx = \lim_{j\to\infty} G(a_j/\delta_j, |f(u_j)|) = 0. \tag{6.4}$$

If $1 < p \le 2$, a Hölder inequality with exponents $2/p$ and $2/(2-p)$ leads to

$$\begin{aligned} \|\nabla u_j\|_p^p &= \int_\Omega |\nabla u_j|^p (a_j/\delta_j + |\nabla u_j|)^{p(p-2)/2} (a_j/\delta_j + |\nabla u_j|)^{p(2-p)/2}\,dx \\ &\le \left(\int_\Omega G(a_j/\delta_j, |\nabla u_j|)\,dx\right)^{p/2} \left(\int_\Omega (a_j/\delta_j + |\nabla u_j|)^p\,dx\right)^{1-p/2}. \end{aligned} \tag{6.5}$$

The last factor is bounded as $j \to \infty$ by (6.2)–(6.3) and the second last tends to zero by (6.4). Again, for $1 < p \le 2$ (when $G(\cdot, |f(u_j)|)$ is monotone decreasing), (6.4) shows that $G(\gamma, |f(u_j)|)$ tends to zero and, hence, so does $|f(u_j)|$. Consequently,

$$\lim_{j\to\infty} \|\nabla u_j\|_q = \lim_{j\to\infty} |f(u_j)| = 0. \tag{6.6}$$

So far we have established (6.6) for $1 < p \le 2$. For $2 < p < \infty$, $|\nabla u_j|^p \le G(a_j/\delta_j, |\nabla u_j|)$ and $|f(u_j)|^p \le G(a_j/\delta_j, |f(u_j)|)$ and so (6.4) implies (6.6) directly. From (6.6) we deduce a contradiction to (6.2): Since $W^{1,q}(\Omega)$ is compactly embedded in $L^q(\Omega)$ we have $u_j \to u$ in $L^q(\Omega)$. With (6.6), $u_j \to u$ in $W^{1,q}(\Omega)$ and so $\|u\|_{1,q} = 1$. Conversely, u is constant (as $\nabla u_j \to 0$ in $L^q(\Omega)$). Since f is a bounded linear form, $f(u_j) \to f(u)$ and $f(u) = 0$. Since $u \in P_0(\Omega) \cap \operatorname{Ker}(f)$, we have $u = 0$. This contradiction with $\|u\|_{1,q} = 1$ concludes the proof in case (6.3).

In the remaining second case we suppose that a_j/δ_j is not bounded (not even for a subsequence). Hence, $\lim_{j\to\infty} a_j/\delta_j = +\infty$. One can assume that

$$\delta_j \le \gamma\, a_j \quad \text{for } q = \min\{2,p\} \quad \text{and} \quad \text{for } j = 1,2,3,\dots \tag{6.7}$$

for a constant γ (and at least for sufficiently large j which we have not relabeled). If $1 < p \le 2$, we use $(1+\delta_j/a_j|u_j|)^{p-2} \le 1$. If $2 \le p < \infty$, we use $\delta_j/a_j \le \gamma$. This leads to

$$\frac{1}{j}\int_\Omega (1+\delta_j/a_j|u_j|)^{p-2}|u_j|^2\,dx \le \begin{cases} \|u_j\|_2^2/j & \text{if } 1 < p \le 2, \\ \|u_j\|_p^2(\gamma\|u_j\|_p + |\Omega|^{1/p})^{p-2}/j & \text{if } 2 \le p < \infty. \end{cases} \tag{6.8}$$

Since $q = \min\{2,p\}$ and $\|u_j\|_{1,q} = 1$, we conclude that (6.8) tends to zero as $j \to \infty$ from embedding. A scaling of (6.1) therefore yields

$$\lim_{j\to\infty}\int_\Omega (1+\delta_j/a_j\,|\nabla u_j|)^{p-2}|\nabla u_j|^2\,dx = 0 \tag{6.9}$$

and

$$\lim_{j\to\infty}(1+\delta_j/a_j\,|f(u_j)|)^{p-2}|f(u_j)|^2 = 0. \tag{6.10}$$

If $2 \le p < \infty$, we directly deduce (6.6) for $q = 2$ and finish the proof as in the first case since $\|u_j\|_{1,2} = 1$. If $1 < p \le 2$, we argue with a Hölder inequality analogy to (6.5) and infer

$$\|\nabla u_j\|_p^2 \le \int_\Omega (1+\delta_j/a_j|\nabla u_j|)^{p-2}|\nabla u_j|^2\,dx \left(\int_\Omega (1+\delta_j/a_j|\nabla u_j|)^p\,dx\right)^{\frac{2-p}{p}}.$$

The last factor is bounded according to (6.7) and $\|u_j\|_{1,p} = 1$. This and (6.9) show (6.6) with $p = q \le 2$. The proof is then finished as in the first case. □

REMARK 6.1. *Lemma* 6.1 *is employed in connection with a scaling argument. If we scale the domain* Ω *from a reference size* 1 *to a patch-size* h*, the first term obtains the factor* h^2 *from a change of variables while the last term in values* $h\,|\nabla v|$ *instead of* $|\nabla v|$. *With a different* $a > 0$*, this yields*

$$\int_\Omega G(a,|v|)\,dx \lesssim \int_\Omega G(a,h|\nabla v|)\,dx$$

for all $v \in W^{1,p}(\Omega) \cap \mathrm{Ker(f)}$ *and* $h = \mathrm{diam}(\Omega)$*; the generic constant depends on the shape of* Ω *but is h-independent.*

Now, let us recall a weighted Clement-type interpolation on the finite element space V_0^h defined in Definition 3.1. It is essential for later analysis to establish approximation error estimates in the quasi-norm for the operator π.

LEMMA 6.2. *For any* $1 < p < \infty$ *and positive integers* d *and* n *there exists a constant* $c_2 = c(p,d,n)$ *such that, for all* $a_1, a_2, \dots, a_n \in R^d$*, there holds*

$$\sum_{j=1}^{n}\sum_{k=1}^{j-1} G(|a_j|,|a_j-a_k|) \le c_2 \sum_{\ell=1}^{n-1}\min_{m=1,\dots,n} G(|a_m|,|a_{\ell+1}-a_\ell|).$$

Proof. Let $\alpha := (a_1+\cdots+a_n)/n \in R^d$ and $b_j := a_j - \alpha \in R^d$ so that $b_1+\cdots+b_n = 0$. Define

$$f(\alpha; b_1, \ldots, b_n) := \sum_{j=1}^{n} \sum_{k=1}^{j-1} G(|\alpha + b_j|, |b_j - b_k|),$$
$$g(\alpha; b_1, \ldots, b_n) := \sum_{\ell=1}^{n-1} \min_{m=1,\ldots,n} G(|\alpha + b_m|, |b_{\ell+1} - b_\ell|).$$

Observe that $g(\alpha, \cdot)$ is positive for nonzero arguments on

$$X := \{(b_1, \ldots, b_n) \in R^{d\times n} : \, b_1 + \cdots + b_n = 0\},$$

since $g(\alpha; b_1, \ldots, b_n) = 0$ implies $b_1 = b_2 = \cdots = b_n$. Let

$$B := \{(b_1, \ldots, b_n) \in X : \, |b_1|^2 + \cdots + |b_n|^2 = 1\}$$

denote the unit ball surface in X. Then, for any $\beta \in R^d$,

$$c(\beta) := \max_{(b_1,\ldots,b_n)\in B} f(\beta; b_1, \ldots, b_n)/g(\beta; b_1, \ldots, b_n) < \infty,$$

since the denominator is positive and $f(\alpha; \cdot), g(\alpha; \cdot)$ are continuous on the compact set B. The same argument shows

$$c_\infty := \max_{(b_1,\ldots,b_n)\in X\setminus\{0\}} \sum_{j=1}^{n} \sum_{k=1}^{j-1} |b_j - b_k|^2 \Big/ \sum_{\ell=1}^{n-1} |b_{\ell+1} - b_\ell|^2 < \infty.$$

Note that

$$\limsup_{|\beta|\to\infty} c(\beta) \le c_\infty < \infty,$$

and so $c(\beta)$ is a bounded continuous function in $\beta \in R^d$. For all $a_1, \ldots, a_n \in R^d$, we have $\alpha \in R^d$ and $(b_1, \ldots, b_n) \in X$ as above. Since f and g are positively homogeneous functions we have, for $\lambda := (|b_1|^2 + \cdots + |b_n|^2)^{1/2} > 0$,

$$\begin{aligned} f(\alpha; b_1, \ldots, b_n) &= \lambda^p f(\alpha/\lambda; b_1/\lambda, \ldots, b_n/\lambda) \\ &\lesssim \lambda^p g(\alpha/\lambda, b_1/\lambda, \ldots, b_n/\lambda) \lesssim g(\alpha; b_1, \ldots, b_n). \qquad \Box \end{aligned}$$

Then we have following interpolation error estimates for the interpolator π in the quasi-norm.

LEMMA 6.3. *Let π be the operator of Definition* 3.1. *For any $1 < p < \infty$, $u_h \in V_h$, $v \in W_0^{1,p}(\Omega)$, and $K \in T^h$, there holds*

$$\begin{aligned} &\int_K G(|\nabla u_h|, |v - \pi v|/h_K)\, dx + \int_K G(|\nabla u_h|, |\nabla(v - \pi v)|)\, dx \\ &\quad \lesssim \sum_{z\in\Lambda\cap K} \left(\int_{\omega_z} G(|\nabla u_h|, |\nabla v|)\, dx + \min_{K'\in T_K} \int_{\cup E_K} G(|\nabla u_h|_{K'}|, |[\partial u_h/\partial n_\varepsilon]|)\, ds \right). \end{aligned} \tag{6.11}$$

Here, $T_K := \{K' \in T^h : \, \bar{K}' \cap \bar{K} \ne \emptyset\}$, $\cup E_K := \cup\{l \subset \partial K' : \, K' \in T_K, \, l \cap \partial\Omega = \emptyset\}$, and $[\partial u_h/\partial n_\varepsilon]$ denotes the jump of the discrete normal fluxes across inner element *boundaries. Consequently,*

$$\begin{aligned}&\sum_K\Big(\int_K G(|\nabla u_h|,|v-\pi v|/h_K)\,dx+\int_K G(|\nabla u_h|,|\nabla(v-\pi v)|)\,dx\Big)\\ &\qquad\lesssim\sum_K\int_K G(|\nabla u_h|,|\nabla v|)\,dx+\tilde\eta^2,\end{aligned}\tag{6.12}$$

where

$$\tilde\eta^2=\sum_{l\cap\partial\Omega=\emptyset}\int_{K^l_{min}}\left(|\nabla u_h|+\left|\left[\frac{\partial u_h}{\partial n}\right]_l\right|\right)^{p-2}\left|\left[\frac{\partial u_h}{\partial n}\right]_l\right|^2.$$

Proof. In the first step, we show that the first term of the left-hand side of (6.11) is bounded by the right-hand side. Fix $a:=\nabla u_h|_K$ and set $v_z:=(\pi v)(z)$ for all $z\in D$. Since $(\psi_z)_{z\in\Lambda\cap K}$ is a partition of unity on K, and G satisfies a triangle inequality in the sense of Lemma 2.4, we have

$$\begin{aligned}\int_K G(|\nabla u_h|,|v-\pi v|/h_K)\,dx&=\int_K G(|a|,|\sum_{z\in\Lambda\cap K}(v\psi_z-v_z\varphi_z)|/h_K)dx\\ &\lesssim\sum_{z\in\Lambda\cap K}\int_K G(|a|,|v\psi_z-v_z\varphi_z|/h_K)\,dx.\end{aligned}\tag{6.13}$$

Let ω_z be the support of ϕ_z. Since $K\subseteq\overline{\omega}_z$, we have for any fixed $z\in\Lambda\cap\bar K$,

$$\int_K G(|a|,|v\psi_z-v_z\varphi_z|/h_K)\,dx\le\int_{\omega_z}G(|a|,|v\psi_z-v_z\varphi_z|/h_K)\,dx.\tag{6.14}$$

In the first case, we suppose $\psi_z=\varphi_z$, i.e., all nodes in $\overline{\omega}_z$ are free nodes. A scaled version of Lemma 6.1 with $w:=(v-v_z)/h_K$, $f(w)=\int_{\omega_z}\varphi_z w\,dx$, and $\psi_z=\varphi_z$ and $|\varphi_z|\le 1$ yield

$$\begin{aligned}\int_{\omega_z}G(|a|,|v\psi_z-v_z\varphi_z|/h_K)\,dx&\le\int_{\omega_z}G(|a|,|v-v_z|/h_K)\,dx\\ &\lesssim\int_{\omega_z}G(|a|,|\nabla v|)\,dx.\end{aligned}\tag{6.15}$$

In the remaining second case, $\psi\not\equiv 1$ on ω_z and so $\partial\omega_z\cap\partial\Omega$ includes at least one outer age E. With $f(w)=\int_E w\,dx$, we deduce from Lemma 6.1 that

$$\int_{\omega_z}G(|a|,|v\psi_z|/h_K)dx\le\int_{\omega_z}G(|a|,|v|/h_K)dx\lesssim\int_{\omega_z}G(|a|,|\nabla v|)dx.\tag{6.16}$$

Since $|v_z|\lesssim|⨍_{\omega_z}\psi_z v\,dx|$ and $G(|a|,\cdot/h_K)$ is convex, Jensen's inequality shows

$$\begin{aligned}\int_{\omega_z}G(|a|,|v_z|/h_K)\,dx&\lesssim\int_{\omega_z}G(|a|,\Big|⨍_{\omega_z}\psi_z v\,dx\Big|/h_K)\,dy\\ &\le\int_{\omega_z}⨍_{\omega_z}G(|a|,|\psi_z v|/h_K)\,dx\,dy=\int_{\omega_z}G(|a|,|\psi_z v|/h_K)\,dx,\end{aligned}\tag{6.17}$$

where the sign $\fint_{\omega_z}$ represents the integral average over ω_z. It follows from the triangle-like inequality of Lemma 2.4, (6.16)–(6.17), monotonicity in the second argument of G, and the inequalities $0 \le \varphi_z \le 1, 0 \le \psi \le 1$ that

$$\begin{aligned}(6.18)\qquad \int_{\omega_z} G(|a|, |v\psi_z - v_z\varphi_z|/h_K)\,dx &\lesssim \int_{\omega_z} G(|a|, |v\psi_z|/h_K)\,dx \\ + \int_{\omega_z} G(|a|, |v_z\varphi_z|/h_K)\,dx &\lesssim \int_{\omega_z} G(|a|, |\nabla v|)\,dx.\end{aligned}$$

Notice that $a + |\nabla u_h - a| \approx a + |\nabla u_h| \approx |\nabla u_h| + |a - \nabla u_h|$ and so

$$\begin{aligned}G(|a|, |\nabla v|) &\le G(|a|, |\nabla v| + |a - \nabla u_h|) \\ &= (|a| + |a - \nabla u_h| + |\nabla v|)^{p-2}(|\nabla v| + |a - \nabla u_h|)^2 \\ (6.19)\qquad &\lesssim (|\nabla u_h| + |\nabla v| + |a - \nabla u_h|)^{p-2}(|\nabla v| + |a - \nabla u_h|)^2 \\ &= G(|\nabla u_h|, |\nabla v| + |a - \nabla u_h|).\end{aligned}$$

Then, the triangle inequality of Lemma 2.4 shows

$$(6.20)\qquad G(|a|, |\nabla v|) \lesssim G(|\nabla u_h|, |\nabla v|) + G(|\nabla u_h|, |a - \nabla u_h|).$$

This, (6.15), and (6.18) result in

$$\begin{aligned}(6.21)\qquad &\int_{\omega_z} G(|a|, |v\psi_z - v_z\varphi_z|/h_K)\,dx \\ &\lesssim \int_{\omega_z} G(|\nabla u_h|, |\nabla v|)\,dx + \int_{\omega_z} G(|\nabla u_h|, |a - \nabla u_h|)\,dx.\end{aligned}$$

This and Lemma 6.2 with $a_j = \nabla u_h|_{K_j}$ for $K_1, \dots, K_n \in T^h$ with $K_1 \cup \dots \cup K_n = \overline{\omega}_z$ and $\{K_1 \cap K_2, \dots, K_{n-1} \cap K_n\} \subset \omega_z$ proves that (6.13) is bounded by the right-hand side of (6.11). The second step is to show that the second term on the left-hand side of (6.11) is bounded in this way as well. To this end, we let c be the integral mean of v on K. The triangle like inequality of Lemma 2.4 shows

$$\begin{aligned}(6.22)\qquad &\int_K G(|\nabla u_h|, |\nabla(v - \pi v)|)\,dx \\ &\lesssim \int_K G(|\nabla u_h|, |\nabla v|)\,dx + \int_K G(|\nabla u_h|, |\nabla(\pi v - c)|\,dx.\end{aligned}$$

The first term on the right-hand side of (6.22) is already bounded as asserted. To estimate the second term, note that $\pi v - c$ is an affine function on K. Then an inverse estimate shows

$$(6.23)\qquad |\nabla(\pi v - c)| \lesssim \fint_K |\pi v - c|\,dx/h_K.$$

It follows from the Jensen's inequality that

$$\begin{aligned}\int_K G(|\nabla u_h|, |\nabla(\pi v - c)|)\,dx &\lesssim \int_K \fint_K G(\nabla u_h|, |\pi v - c|/h_K)\,dxdy \\ &= \int_K G(|\nabla u_h|, |\pi v - c|/h_K)\,dx.\end{aligned}$$

The triangle like inequality of Lemma 2.4 yields

$$\int_K G(|\nabla u_h|, |\pi v - c|/h_K)\,dx \lesssim \int_K G(|\nabla u_h|, |v - \pi v|/h_K)\,dx + \int_K G(|\nabla u_h|, |v - c|/h_K)\,dx. \tag{6.24}$$

The first term on the right-hand side of (6.24) is already shown to be bounded by the right side of (6.11). The same conclusion for the second term follows from Lemma 6.1 with $f(w) = \int_K w\,dx$ and $\Omega = K$ as in (6.15):

$$\int_K G(|\nabla u_h|, |v - c|/h_K)\,dx \lesssim \int_K G(|\nabla u_h|, |\nabla v|)\,dx. \tag{6.25}$$

Then it follows from (6.22)–(6.25) that

$$\int_K G(|\nabla u_h|, |\nabla(v - \pi v)|)\,dx \lesssim \int_K G(|\nabla u_h|, |\nabla v|)\,dx + \int_K G(|\nabla u_h|, |v - \pi v|/h_K)\,dx. \tag{6.26}$$

Hence the desired estimate of the second term on the left-hand side of (6.11) follows from (6.26) and the first step of the proof. □

The next lemma establishes a quasi-norm estimate for the inner produce of a function and an interpolation error.

Lemma 6.4. *For any $\delta > 0, 1 < p < \infty$, $u_h \in V_h$, $v \in W_0^{1,p}(\Omega)$, and $f \in W^{1,p'}(\Omega)$, where $1/p + 1/p' = 1$, there exist constants C and C_1 (where only C_1 depends on δ) such that*

$$\begin{aligned}\int_\Omega f(v - \pi v)\,dx \le C\delta \int_\Omega G(|\nabla u_h|, |\nabla v|)\,dx \\ + C_1 \int_\Omega \left(|\nabla u_h|^{p-1} + h_z^2 |\nabla f|\right)^{p'-2} h_z^4 |\nabla f|^2\,dx \\ + C\delta \sum_K \min_{K' \in T_K} \int_{\cup E_K} G(|\nabla u_h|_{K'}|, |[\partial u_h/\partial n_\varepsilon]|)\,ds,\end{aligned} \tag{6.27}$$

where $G(\cdot,\cdot)$ is defined by (5.4). Therefore, with $\tilde\eta$ defined in Lemma 6.3, there holds

$$\begin{aligned}\int_\Omega f(v - \pi v) \le C_1 \sum_K \int_K (|\nabla u_h|^{p-1} + h_K^2 |\nabla f|)^{p'-2} h_K^4 |\nabla f|^2 \\ + C\delta \sum_K \int_K (|\nabla u| + |\nabla v|)^{p-2} |\nabla v|^2 + C\delta \tilde\eta^2.\end{aligned} \tag{6.28}$$

Proof. First note that $\int (v\psi_z - v_z\varphi_z)\,dx = 0$. Thus, with $f_z := ⨍_{\omega_z} f(x)\,dx$,

$$\begin{aligned}\int_\Omega f(v - \pi v)\,dx &= \sum_{z\in\Lambda} \int_\Omega f(v\psi_z - v_z\varphi_z)\,dx \\ &= \sum_{z\in\Lambda} \int_{\omega_z} (f - f_z) h_z\,(v\psi_z - v_z\varphi_z)/h_z\,dx.\end{aligned} \tag{6.29}$$

We use Lemma 2.3 to estimate the product inside the integral. This yields

$$\begin{aligned}&\int_{\omega_z}(f-f_z)h_z\,(v\psi_z-v_z\varphi_z)/h_z\,dx\\ &\le \delta^{-\beta}\int_{\omega_z}(|a|^{p-1}+|f-f_z|h_z)^{p'-2}h_z^2|f-f_z|^2\,dx\\ &+\delta\int_{\omega_z}G(|a|,|v\psi_z-v_z\varphi_z)|/h_z)\,dx.\end{aligned}\tag{6.30}$$

Here a is one of the discrete gradients $|\nabla u_h|$ on ω_z. Lemma 6.1 will be employed for $f-f_z$ and the functional $g(w)=\int_{\omega_z}w$ (so it vanishes for $w:=f-f_z$). Notice that a is replaced by $|a|^{p-1}$ and p is replaced by p'. Then we obtain

$$\begin{aligned}&\int_{\omega_z}(|a|^{p-1}+|f-f_z|h_z)^{p'-2}h_z^2|f-f_z|^2\,dx\\ &\quad\lesssim\int_{\omega_z}(|a|^{p-1}+h_z^2|\nabla f|)^{p'-2}h_z^4|\nabla f|^2\,dx.\end{aligned}\tag{6.31}$$

Arguing as we had done in proving (6.14)–(6.18), we deduce from (6.30)–(6.31) that

$$\begin{aligned}\int_{\omega_z}(f-f_z)h_z\,(v\psi_z-v_z\varphi_z)/h_z\,dx&\lesssim\delta\int_{\omega_z}G(|a|,|\nabla v|)\,dx\\ +\delta^{-\beta}\int_{\omega_z}&(|a|^{p-1}+h_z^2|\nabla f|)^{p'-2}h_z^4|\nabla f|^2\,dx.\end{aligned}\tag{6.32}$$

So far, a is a constant vector on ω_z. Depending on p', we choose a so that $|a|$ is minimal or maximal among $(|\nabla u_h|_{K'}:K'\in T_K)$, and thus

$$\int_{\omega_z}(|a|^{p-1}+h_z^2|\nabla f|)^{p'-2}h_z^4|\nabla f|^2\,dx\le\int_{\omega_z}(|\nabla u_h|^{p-1}+h_z^2|\nabla f|)^{p'-2}h_z^4|\nabla f|^2\,dx.$$

Arguing as in the first step of the proof of Lemma 6.3, we have

$$\begin{aligned}\int_{\omega_z}G(|a|,|\nabla v|)\,dx&\lesssim\int_{\omega_z}G(|\nabla u_h|,|\nabla v|)\,dx\\ &\quad+\sum_{K\in\omega_z}\min_{K'\in T_K}\int_{\cup E_K}G(|\nabla u_h|_{K'}|,|[\partial u_h/\partial n_\varepsilon]|)\,ds.\end{aligned}$$

Thus the desired estimate follows from (6.32) and the above two inequalities. □

REMARK 6.2. *It follows from the above proofs that Lemmas* 6.1–6.4 *hold for any continuous function* $G(\cdot,\cdot)$ *such that it is increasing* (*decreasing*) *as* $p\ge 2$ ($p\le 2$) *in the first argument, and is convex and increasing in the second argument.*

Next is a well-known trace theorem [KJF].

LEMMA 6.5. *For all* $v\in W^{1,q}(K)$, $1\le q<\infty$,

$$\|v\|_{0,q,\partial K}\lesssim h_K^{-\frac{1}{q}}\|v\|_{0,q,K}+h_K^{1-\frac{1}{q}}|v|_{1,q,K}.\tag{6.33}$$

We need a quasi-norm version of the trace theorem for polynomials.

LEMMA 6.6. *Let* $K\in T^h$ *and* v *be a polynomial of degree* $s\le k$. *Then*

$$h_K\int_{\partial K}(|\nabla u_h|+|\nabla v|)^{p-2}|\nabla v|^2\lesssim\int_K(|\nabla u_h|+|\nabla v|)^{p-2}|\nabla v|^2.\tag{6.34}$$

The generic constant depends only on k and the aspect ratio of the finite elements.

Proof. Because v is a polynomial in K, by an inverse inequality, we have that, for any $x \in K$,

$$|\nabla v(x)| \leq |v|_{1,\infty,K} \lesssim h_K^{-2} \int_K |\nabla v| \, .$$

Therefore, it follows from Jensen's inequality that

$$\begin{aligned} h_K \int_{\partial K} (|\nabla u_h| + |\nabla v|)^{p-2} |\nabla v|^2 &\lesssim h_K^2 \left(|\nabla u_h| + \int_K h_K^{-2} |\nabla v| \right)^{p-2} \left(\int_K h_K^{-2} |\nabla v| \right)^2 \\ &\leq \int_K (|\nabla u_h| + |\nabla v|)^{p-2} |\nabla v|^2. \qquad \square \end{aligned}$$

REFERENCES

[BA] J. Baranger and H. El-Amri, *Estimateurs a posteriori d'erreur pour le calcul adaptatif d'ecoulements quasi-Newtoniens,* RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 31–48.

[BL1] J.W. Barrett and W.B. Liu, *Finite element approximation of the p-Laplacian,* Math. Comp., 61 (1993), pp. 523–537.

[BL2] J.W. Barrett and W.B. Liu, *Finite Element Approximation of Some Degenerate Quasi-Linear Problems,* Lecture Notes in Math. 303, Pitman, Boston, 1994, pp. 1–16.

[BL3] J.W. Barrett and W.B. Liu, *Quasi-norm error bounds for finite element approximation of quasi-Newtonian flows,* Numer. Math., 68 (1994), pp. 437–456.

[BL4] J.W. Barrett and W.B. Liu, *Finite element approximation of the parabolic p-Laplacian,* SIAM J. Numer. Anal., 31 (1994), pp. 413–428.

[BB] J.W. Barrett and R. Bermejo, *An improved error bound for the discretization of the parabolic p-Laplacian and related degenerate quasilinear equations and variational inequalities,* to appear.

[Ca] C. Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element method,* RAIRO Modél. Math. Anal. Numér., 33 (1999), pp. 1187–1202.

[CF2] C. Carstensen and S.A. Funken, *Averaging technique for FE—a posteriori error control in elasticity.* I. *Conforming FEM,* Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 2483–2498.

[CF3] C. Carstensen and S.A. Funken, *Averaging technique for FE—a posteriori error control in elasticity.* II. *λ-independent estimates,* Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4663–4675.

[CF4] C. Carstensen and S.A. Funken, *Averaging technique for a posteriori error control in elasticity.* III. *Locking-free nonconforming FEM,* Comput. Methods Appl. Mech. Engrg., 191 (2001), pp. 861–877.

[CK] C. Carstensen and R. Klose, *A posteriori finite element error control for the p-Laplace problem,* SIAM J. Sci. Comput., 25 (2003), pp. 792–814.

[CLY] C. Carstensen, W.B. Liu, and N. Yan, *A posteriori error estimators based on gradient recovery for finite element approximation of p-Laplacian,* Math. Comp., to appear.

[Ch] S.S. Chow, *Finite element error estimates for nonlinear elliptic equations of monotone type*, Numer. Math., 54 (1988), pp. 373–393.

[Ci] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems,* North-Holland, Amsterdam, 1978.

[DeB] E. DiBenedetto, *Degenerate Parabolic Equations,* Springer-Verlag, New York, 1994.

[GM] R. Glowinski and A. Marrocco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une class de problèmes de Dirichlet non linéaires,* RAIRO Anal. Numer., 2 (1975), pp. 41–76.

[KJF] A. Kufner, O. John, and S. Fucik, *Function Spaces,* Nordhoff, Leyden, The Netherlands, 1977.

[LB] W.B. Liu and J.W. Barrett, *Finite element approximation of some degenerate monotone quasi-linear elliptic systems,* SIAM J. Numer. Anal., 33 (1996), pp. 88–106.

[LY1] W.B. Liu and N.N. Yan, *Quasi-norm local error estimates for p-Laplacian,* SIAM J. Numer. Anal., 39 (2001) pp. 100–127.

[LY2] W.B. Liu and N.N. Yan, *Quasi-norm a posteriori error estimates for nonconforming FEM of p-Laplacian,* Numer. Math., 89 (2001), pp. 341–378.

[ODSD] J.T. Oden, L. Demkovicz, T. Strouboulis, and P. Devloo, *Accuracy Estimates and Adaptive Refinements in Finite Element Computations,* John Wiley & Sons, New York, 1986.

[P] C. Padra, *A posteriori error estimators for nonconforming approximation of some quasi-Newtonian flows,* SIAM J. Numer. Anal., 34 (1997), pp. 1600–1615.

[GS] G. Simms, *Finite Element Approximation of Some Nonlinear Elliptic and Parabolic Problems,* thesis, Imperial College, University of London, 1995.

[V1] R. Verfürth, *A posteriori error estimates for nonlinear problems,* Math. Comp., 62 (1994), pp. 445–475.

SIAM J. Numer. Anal.
Vol. 43, No. 6, pp. 2320–2344

SOME NEW ERROR ESTIMATES OF A SEMIDISCRETE FINITE VOLUME ELEMENT METHOD FOR A PARABOLIC INTEGRO-DIFFERENTIAL EQUATION WITH NONSMOOTH INITIAL DATA*

RAJEN K. SINHA[†], RICHARD E. EWING[‡], AND RAYTCHO D. LAZAROV[§]

Abstract. A semidiscrete finite volume element (FVE) approximation to a parabolic integro-differential equation (PIDE) is analyzed in a two-dimensional convex polygonal domain. An optimal-order L^2-error estimate for smooth initial data and nearly the same optimal-order L^2-error estimate for nonsmooth initial data are obtained. More precisely, for homogeneous equations, an elementary energy technique and a duality argument are used to derive an error estimate of order $O\left(t^{-1}h^2 \ln h\right)$ in the L^2-norm for positive time when the given initial function is only in L^2.

Key words. parabolic equation, integro-differential equation, optimal-order error estimate, smooth and nonsmooth initial data

AMS subject classifications. 65M12, 65M60, 65N40

DOI. 10.1137/040612099

1. Introduction.

The aim of this paper is to analyze a semidiscrete finite volume element (FVE) method for solving initial-boundary value problems for an integro-differential equation of the form

$$\begin{aligned} (1.1) \qquad u_t - \nabla\cdot(\mathcal{A}\nabla u) &= -\int_0^t \nabla\cdot(\mathcal{B}\nabla u(s))ds + f(x,t) \ \text{ in } \ \Omega\times J,\\ u &= 0 \ \text{ on } \ \partial\Omega\times J,\\ u(\cdot,0) &= u_0 \ \text{ in } \ \Omega. \end{aligned}$$

Here, $\Omega \subset \mathbb{R}^2$ is a bounded convex polygonal domain with boundary $\partial\Omega$, $J=(0,T]$ with $T<\infty$, and $u_t = \partial u/\partial t$. Further, $\mathcal{A}=\{a_{i,j}(x)\}$ is a symmetric and uniformly positive definite matrix of size 2×2 in Ω and $\mathcal{B}=\{b_{i,j}(x,t,s)\}$ is a 2×2 matrix. The nonhomogeneous term $f=f(x,t)$ and the coefficients $a_{ij}(x)$, $b_{ij}(x;t,s)$ are assumed to be smooth for our purpose. For the sake of simplicity, we shall denote $Au = -\nabla\cdot(\mathcal{A}\nabla u)$ and $B(t,s)u(s) = -\nabla\cdot(\mathcal{B}\nabla u(s))$. For references to studies regarding existence, uniqueness, and regularity of such problems, one may refer to [33].

Parabolic integro-differential equations (PIDEs) of the above type arise naturally in many applications, such as, for instance, heat conduction in materials with memory [27], nonlocal reactive flows in porous media [10, 11], and non-Fickian flow of fluid in porous media [15]. One very important characteristic of these models is that they all

*Received by the editors July 22, 2004; accepted for publication (in revised form) May 20, 2005; published electronically January 6, 2006. The work of the first author was supported by the DST, Government of India under BOYSCAST fellowship. The work of the second and third authors was partially supported by the NSF-EIA grant 0218229.

http://www.siam.org/journals/sinum/43-6/61209.html

[†] Department of Mathematics, Indian Institute of Technology Guwahati, Guwahati - 781039, India (rajen@iitg.ernet.in).

[‡] Institute for Scientific Computation, Texas A&M University, College Station, TX 77843-3404 (ewing@isc.tamu.edu).

[§] Department of Mathematics, Texas A&M University, College Station, TX 77843-3404 (lazarov @math.tamu.edu).

express the conservation of a certain quantity (mass, momentum, heat, etc.) in any moment for any subdomain. This in many applications is the most desirable feature of the approximation method when it comes to numerical solution of the corresponding initial-boundary value problem. For references to studies of existence, uniqueness, and regularity of such problems, one may refer to [33].

To put our work into proper perspective, we first give a brief account of the development of the finite element methods for such problems. Over the last decade, various numerical methods based on finite element approximations in space and special quadrature in time have been developed and studied for this type of problem [20, 24, 25, 29, 31, 32, 34]. The crucial tools used in the analysis are the Ritz and Ritz–Volterra projections which are instrumental in deriving optimal-order error estimates in various Sobolev norms [5, 6, 20]. In [31], the authors studied this type of problem for both smooth and nonsmooth initial data cases. In particular, for a homogeneous equation with nonsmooth initial data, an optimal-order L^2-error estimate is proved via a semigroup theoretic approach. Subsequently, using the energy method, the authors of [25] derived convergence of order $O\big(\frac{h^2}{t}\big)$ for the L^2-norm and $O\big(\frac{h^2}{t}\log(\frac{1}{h})\big)$ for the L^∞-norm for the homogeneous equation when the initial function is in $H_0^1(\Omega)\cap H^2(\Omega)$. Recently, in [26], the analysis from [21] of the case $B(t,s)=0$ was carried over to a time dependent PIDE. An optimal-order error estimate by energy techniques and a duality argument for the homogeneous equation with both smooth and nonsmooth initial data were carried over. In both [21] and [26], negative norm estimates are used in a crucial way in their analyses. In the absence of the memory term, i.e., when $B(t,s)\equiv 0$, the error estimates for finite element methods for both smooth and nonsmooth data cases are described in [2, 18, 28, 30] and the references cited therein.

In recent years, the numerical methods for problem (1.1) by means of FVE discretizations were considered in [13] and [14]. The interest in such methods is due to certain conservation features of FVE methods that are desirable in many applications. In [13] and [14], the authors studied FVE approximation of such a problem in the framework of the standard Petrov–Galerkin formulation and obtained L^2-error estimate of the form (cf. [14, p. 305])

$$\begin{aligned}\|u(t)-u_h(t)\| \le{}& Ch^2(\|u_0\|_{3,p}+\|u(t)\|_{3,p}\\ &+\int_0^t(\|u(s)\|_{3,p}+\|u_t(s)\|_{3,p})ds), \quad p>1,\end{aligned} \tag{1.2}$$

where u and u_h represent the solution of (1.1) and its FVE approximation, respectively. Note that the estimate (1.2) is optimal with respect to the approximation property, but its regularity requirement on the exact solution seems to be too high when compared with that for finite element methods. This is primarily due to the fact that the bounds in the L^2-norm of a new variant of the Ritz–Volterra projection (the so-called Petrov–Volterra projection introduced in [13, 14]) are not optimal with respect to the regularity of the solution.

In this paper, we analyze the FVE method for the problem (1.1) and derive optimal-order L^2-error estimates for both smooth and nonsmooth initial data. For the homogeneous problem with smooth initial data, we are able to show an L^2-error estimate which is optimal with respect to the order of convergence as well as the regularity of the solution. This is exactly the result known for finite element methods (cf. [25]). More precisely, we prove an optimal-order L^2-error estimate for $f=0$ and initial data $u_0\in H^2(\Omega)\cap H_0^1(\Omega)$. This technique, quite new and promising, is based on improved estimates for a new variant of the Ritz–Volterra projection (see

Theorems 3.1 and 3.2).

The main concern of this paper is to prove an L^2-error estimate for the homogeneous equation ($f = 0$) with nonsmooth initial data. This is motivated by the fact that the solutions of a homogeneous linear parabolic equation have the so-called *smoothing property.* That is, the solution is smooth for positive time t, even when the initial data are not. In quantitative form, this may be expressed by the inequality

$$\|u(t)\|_\alpha \leq Ct^{-\alpha/2}\|u_0\|, \quad t \in J, \tag{1.3}$$

which is valid for any $\alpha \geq 0$. Here $\|\cdot\|_\alpha$ is a Sobolev norm. However, this is not the case with PIDEs as they have a limited smoothing property. This fact is proved in [31], where the inequality (1.3) is shown to be valid only for $\alpha \leq 2$. Since the smoothing property plays a significant role in the error analysis in the semidiscrete solution, an attempt has been made in this paper to derive an L^2-error estimate for the FVE method when the initial data u_0 is only in $L^2(\Omega)$. More importantly, our analysis uses only energy techniques and a duality argument.

The proposed techniques have several attractive features. Unlike the analyses of [21] and [26], we do not require error estimates in negative-indexed Sobolev norms while dealing with L^2-error estimates with nonsmooth initial data. Thus, these results hold for convex polygonal domains with corners, unlike [21] and [26]. Since the FVE method is thought of as a perturbation of the Galerkin finite element method, the proposed technique can easily be adopted to the finite element method as well. However, to the best of our knowledge the error estimates for nonsmooth initial data using the FVE method were not established earlier.

The previous work on the theoretical framework and the basic tools for the analysis of the FVE methods for elliptic and parabolic problems are described in [3, 4, 9, 7, 8, 12, 16, 17, 19, 22, 23] and references therein.

The outline of this paper is as follows. In section 2, we introduce some notation, formulate FVE approximations for piecewise linear finite element spaces defined on a triangulation, and recall some basic estimates from the literature. Further, the Ritz–Volterra projection is introduced and related estimates are obtained in section 3. Section 4 is devoted to the error estimates for smooth initial data. Finally, error estimates with nonsmooth initial data are carried out in section 5.

Throughout this paper C denotes a generic positive constant which does not depend on the mesh parameter h but may depend on T.

2. Notation and preliminaries.

Let $H_0^1(\Omega) = \{\phi \in H^1(\Omega) \mid \phi = 0 \text{ on } \partial\Omega\}$. Further, let $A(\cdot,\cdot)$ and $B(t,s;\cdot,\cdot)$ be the bilinear forms on $H_0^1(\Omega) \times H_0^1(\Omega)$ given by

$$A(u,v) = \int_\Omega \mathcal{A}(x)\nabla u \cdot \nabla v dx; \quad B(t,s;u(s),v) = \int_\Omega \mathcal{B}(x,t,s)\nabla u(s) \cdot \nabla v dx. \tag{2.1}$$

For the purpose of FVE approximations we now consider the following weak formulation: Find $u : \bar{J} \to H_0^1(\Omega)$ such that

$$(u_t, v) + A(u,v) = \int_0^t B(t,s;u(s),v)ds + (f,v) \quad \forall v \in H_0^1(\Omega), \quad t \in J, \tag{2.2}$$

with $u(0) = u_0$.

Here and below, $(\cdot,\cdot)$ and $\|\cdot\|$ denote the L^2 inner product and the induced norm on $L^2(\Omega)$. Further, we shall use the standard notation for Sobolev spaces $W^{m,p}(\Omega)$

with $1 \leq p \leq \infty$. The norm on $W^{m,p}(\Omega)$ is defined by

$$\|u\|_{m,p,\Omega} = \|u\|_{m,p} = \left(\int_{\Omega} \sum_{|\alpha| \leq m} |D^{\alpha} u|^p dx \right), \quad 1 \leq p < \infty,$$

with the standard modification for $p = \infty$. When $p = 2$, we write $W^{m,2}(\Omega)$ by $H^m(\Omega)$ and denote the norm by $\| \cdot \|_m$. Further, $H^{-1}(\Omega)$ denotes the space of all bounded linear functionals on $H_0^1(\Omega)$. For a functional $f \in H^{-1}(\Omega)$, its action on a function $u \in H_0^1(\Omega)$ is denoted by (f, u), which represents the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. To simplify notation, we use $(\cdot, \cdot)$ to denote both the $L^2(\Omega)$ inner product and the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$.

2.1. A priori estimates. In the following lemmas, we state some a priori bounds for the solution u satisfying (1.1) under appropriate regularity assumptions on the initial function u_0. For a proof, one may refer to [25, 26, 21].

LEMMA 2.1. *Let u satisfy* (1.1). *If $u_0 \in L^2(\Omega)$ and $f \in L^2(\Omega)$, then*

$$\|u(t)\|^2 + \int_0^t \|u(s)\|_1^2 ds \leq C \left(\|u_0\|^2 + \int_0^t \|f(s)\|^2 ds \right).$$

Moreover, when $u_0 \in H_0^1(\Omega)$ and $f \in L^2(\Omega)$, we have

$$\|u(t)\|_1^2 + \int_0^t \{\|u_s(s)\|^2 + \|u(s)\|_2^2\} ds \leq C \left(\|u_0\|_1^2 + \int_0^t \|f(s)\|^2 ds \right).$$

LEMMA 2.2. *Let u satisfy* (1.1). *If $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$ and $f \in L^2(\Omega)$, then*

$$\|u_t(t)\|^2 + \int_0^t \|u_s(s)\|_1^2 ds \leq C \left(\|u_t(0)\|^2 + \int_0^t \|f(s)\|^2 ds \right).$$

LEMMA 2.3. *Let u satisfy* (1.1) *with $f = 0$, and let $0 \leq i, j, k \leq 2$. If $0 \leq k + 2j - i \leq 2$, then*

$$t^i \left\| \frac{\partial^j u}{\partial t^j}(t) \right\|_k^2 \leq C \|u_0\|_{k+2j-i}^2.$$

Further, if $0 \leq k + 2j - i - 1 \leq 2$, then

$$\int_0^t s^i \left\| \frac{\partial^j u}{\partial s^j}(s) \right\|_k^2 ds \leq C \|u_0\|_{k+2j-i-1}^2.$$

2.2. FVE approximation. Let T_h be a quasi-uniform family of triangulations of Ω such that $\bar{\Omega} = \cup_{K \in T_h} K$, where K is a closed triangle element. Let N_h be the set of all nodes or vertices of T_h, i.e.,

$$N_h = \{p \ : \ p \text{ is a vertex of element } K \in T_h \text{ and } p \in \bar{\Omega}\}.$$

Further, we denote $N_h^0 = N_h \cap \Omega$. For a vertex $x_i \in N_h$, let $\Pi(i)$ be the index set of those vertices that, along with x_i, are in some element of T_h.

For a given triangulation T_h, we now introduce a dual mesh T_h^* as follows: In each element $K \in T_h$ with vertices x_i, x_j, and x_k, select a point $q \in K$, select a point

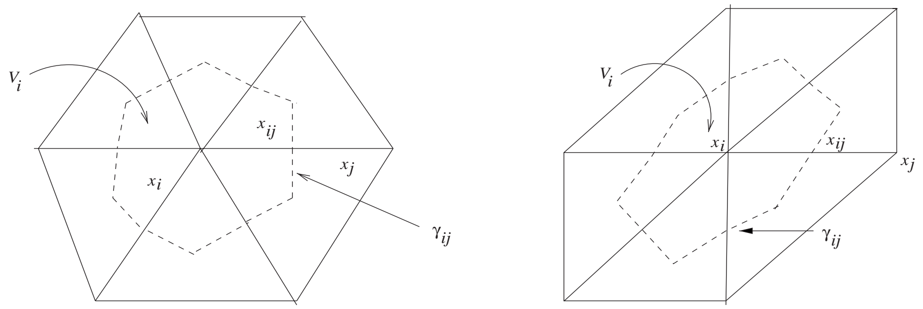


FIG. 2.1. *Control volumes with barycenter as internal point and interface γ_{ij} of V_i and V_j.*

x_{ij} on the edge connecting x_i and x_j, and connect q with x_{ij} by straight lines $\gamma_{ij,K}$. Then for a vertex x_i we let V_i be the polygon whose edges are $\gamma_{ij,K}$ in which x_i is a vertex of the element K. We call this V_i a *control volume* centered at x_i. Further, we note that $\cup_{x_i \in N_h} V_i = \bar{\Omega}$. Thus, the dual mesh T_h^* is then defined as the collection of these *control volumes.* A *control volume* centered at a vertex x_i is given in Figure 2.1.

We call the control volume mesh T_h^* regular or quasi-uniform if there exists a positive constant $C > 0$ such that

$$C^{-1}h^2 \leq \text{meas}(V_i) \leq Ch^2 \quad \forall\, V_i \in T_h^*,$$

where h is the maximum diameter of all elements $K \in T_h$.

There are various ways to introduce a regular dual mesh T_h^* depending on the choices of the point q in an element $K \in T_h$ and the points x_{ij} on its edges. In this paper, we choose q to be the barycenter of an element $K \in T_h$, and the points x_{ij} are chosen to be the midpoints of the edges of K. Obviously, if T_h is regular, i.e., there is a constant C such that

$$Ch_K^2 \leq \text{meas}(K) \leq h_K^2,$$

where $h_K = \text{diam}(K)$ for all elements $K \in T_h$, then the dual mesh T_h^* is also regular. For the purpose of FVE approximation, let S_h be the standard linear finite element space defined on the triangulation T_h,

$$S_h = \{v \in C(\Omega) \;:\; v|_K \text{ is linear } \forall\ K \in T_h \text{ and } v|_{\partial\Omega} = 0\},$$

and its dual volume element space S_h^*,

$$S_h^* = \{v \in L^2(\Omega) \;:\; v|_V \text{ is constant } \forall\ V \in T_h^* \text{ and } v|_{\partial\Omega} = 0\}.$$

Obviously, $S_h = \text{span}\{\phi_i(x) \;:\; x_i \in N_h^0\}$ and $S_h^* = \text{span}\{\psi_i(x) \;:\; x_i \in N_h^0\}$, where ϕ_i are the standard nodal basis functions associated with the node x_i, and ψ_i are the characteristic functions of the volume V_i. Let $I_h \;:\; C(\Omega) \to S_h$ and $I_h^* \;:\; C(\Omega) \to S_h^*$ be the usual interpolation operators, i.e.,

$$I_h u(x) = \sum_{x_i \in N_h} u_i \phi_i(x) \quad \text{and} \quad I_h^* u(x) = \sum_{x_i \in N_h} u_i \psi_i(x),$$

where $u_i = u(x_i)$.

The FVE approximation is then defined to be the function $u_h : \bar{J} \to S_h$ such that

$$(u_{ht}, I_h^*\chi) + A(u_h, I_h^*\chi) = \int_0^t B(t, s; u_h(s), I_h^*\chi)ds + (f, I_h^*\chi) \quad \forall \chi \in S_h. \tag{2.3}$$

Here $u_h(0) = \tilde{P}_h u_0$, where $\tilde{P}_h u_0$ is the L^2-projection of u_0 onto S_h defined by

$$(\tilde{P}_h u_0, I_h^*\chi) = (u_0, I_h^*\chi) \quad \forall \chi \in S_h, \tag{2.4}$$

the bilinear forms $A(\cdot,\cdot)$ and $B(t, s; \cdot, \cdot)$ in (2.3) are defined by

$$A(u, v) = -\sum_{x_i \in N_h} v_i \int_{\partial V_i} \mathcal{A}(x)\nabla u \cdot \mathbf{n} dS_x,$$

$$B(t, s; u, v) = -\sum_{x_i \in N_h} v_i \int_{\partial V_i} \mathcal{B}(x, t, s)\nabla u \cdot \mathbf{n} dS_x$$

for $(u, v) \in ((H_0^1 \cap H^2) \cup S_h) \times S_h^*$, and $\mathbf{n}$ is the outer-normal vector of the involved integration domain. Note that when $(u, v) \in H_0^1(\Omega) \times H_0^1(\Omega)$, the bilinear forms $A(\cdot,\cdot)$ and $B(t, s; \cdot, \cdot)$ are given by (2.1).

In order to describe features of the bilinear forms defined in (2.2) and (2.3), we use some discrete norms on S_h and S_h^*,

$$|u_h|_{0,h}^2 = (u_h, u_h)_{0,h}, \quad |u_h|_{1,h}^2 = \sum_{x_i \in N_h} \sum_{x_j \in \Pi(i)} \text{meas}(V_i)((u_{hi} - u_{hj})/d_{ij}^2,$$

$$\|u_h\|_{1,h}^2 = |u_h|_{0,h}^2 + |u_h|_{1,h}^2, \qquad |||u_h|||^2 = (u_h, I_h^* u_h),$$

where $(u_h, v_h)_{0,h} = \sum_{x_i \in N_h} \text{meas}(V_i) u_{hi} v_{hi} = (I_h^* u_h, I_h^* v_h)$ and $d_{ij} = d(x_i, x_j)$ is the Euclidean distance between x_i and x_j.

The discrete norms $|\cdot|_{0,h}$ and $\|\cdot\|_{1,h}$ are equivalent to the usual norms $\|\cdot\|$ and $\|\cdot\|_1$, respectively, on S_h. Some properties of the bilinear forms are stated below without proof. For a proof, see, e.g., [1, 12, 14].

Lemma 2.4. *There exist positive constants C_1 and C_2 such that for all $v_h \in S_h$, we have*

$$C_1 |v_h|_{0,h} \le \|v_h\| \le C_2 |v_h|_{0,h},$$

$$C_1 |||v_h||| \le \|v_h\| \le C_2 |||v_h|||,$$

$$C_1 \|v_h\|_{1,h} \le \|v_h\|_1 \le C_2 \|v_h\|_{1,h}.$$

Lemma 2.5. *There exist positive constants C and c such that for all $\phi_h, \psi_h \in S_h$, we have*

$$|A(\phi_h, I_h^* \psi_h)| \le C\|\phi_h\|_1 \|\psi_h\|_1,$$

$$|B(t, s; \phi_h, \psi_h)| \le C\|\phi_h\|_1 \|\psi_h\|_1,$$

and

$$A(\phi_h, I_h^* \phi_h) \ge c\|\phi_h\|_1^2.$$

Lemma 2.6. *If the matrix $\mathcal{A}(x)$ is constant over each element $K \in T_h$, then we have*

$$A(u_h, \chi) = A(u_h, I_h^*\chi) \quad \forall u_h, \chi \in S_h.$$

Following the arguments of Lemma 2.3 on the discrete level, it is easy to derive the following stability estimates for the FVE solution u_h satisfying (2.3).

LEMMA 2.7. *Let u_h satisfy (2.3) with $f = 0$. Then we have*

$$\|u_h(t)\|^2 + \int_0^t \|u_h(s)\|_1^2 ds \le C\|u_h(0)\|^2,$$
$$\int_0^t s\|u_{hs}(s)\|^2 + t\|u_h(t)\|_1^2 ds \le C\|u_h(0)\|^2,$$
$$t^2\|u_{ht}(t)\|^2 + \int_0^t s^2\|u_{hs}(s)\|_1^2 ds \le C\|u_h(0)\|^2.$$

The following lemma gives the key feature of the bilinear forms in the FVE method. For a proof, see [12] or [7].

LEMMA 2.8. *Let $\phi \in H_0^1(\Omega)$. Then we have*

$$\begin{aligned} A(\phi,\chi) = A(\phi, I_h^*\chi) &+ \sum_{K\in T_h}\int_{\partial K}(A\nabla\phi\cdot\mathbf{n})(\chi - I_h^*\chi)dS \\ &- \sum_{K\in T_h}\int_K(\nabla\cdot A\nabla\phi)(\chi - I_h^*\chi)dx \quad \forall\chi\in S_h. \end{aligned}$$

The above identity holds true when $A(\cdot,\cdot)$ is replaced by $B(t,s;\cdot,\cdot)$.

Remark 2.9. We note that the above identity is proved in [12, 7] for $\phi, \chi \in S_h$. In fact, identities in Lemma 2.8 holds true even if $\phi \in H_0^1(\Omega)$.

3. Ritz–Volterra projection and related estimates. Following Lin et al. [20], we define the Ritz–Volterra projection $W_h u$ of a function $u(x,t)$ defined on $\Omega \times \bar{J}$ in the context of the FVE method and obtain bounds for the error in H^1 and L^2 norms. The Ritz–Volterra projection $W_h : L^\infty(H_0^1 \cap H^2) \to L^\infty(S_h)$ is defined by

$$A((u - W_h u)(t), I_h^*\chi) = \int_0^t B(t,s;(u - W_h u)(s), I_h^*\chi)ds \;\; \forall\,\chi\in S_h,\; t\in\bar{J}. \tag{3.1}$$

Below, we shall prove a lemma which is frequently used in our subsequent analysis.

LEMMA 3.1. *For any function $\phi \in H^r(\Omega)(r = 0, 1)$, we have*

$$|(\phi, \chi - I_h^*\chi)| \le Ch^{1+r}\|\phi\|_r\|\chi\|_1 \;\; \forall\,\chi\in S_h. \tag{3.2}$$

Further, for $\phi \in H_0^1(\Omega)$, we have

$$|A(\phi, \chi - I_h^*\chi)| \le Ch\|\phi\|_1\|\chi\|_1 \quad \forall\,\chi\in S_h. \tag{3.3}$$

The second inequality also holds true when $A(\cdot,\cdot)$ is replaced by $B(t,s;\cdot,\cdot)$.

Proof. We borrow the proof of (3.2) from [8]. To show (3.3), we have [7]

$$\begin{aligned} A(\phi, \chi - I_h^*\chi) = &- \sum_{K\in T_h}\int_K(\nabla\cdot\mathcal{A}\nabla\phi)(\chi - I_h^*\chi)dx \\ &+ \sum_{K\in T_h}\int_{\partial K}((\mathcal{A} - \bar{\mathcal{A}}_K)\nabla\phi\cdot\mathbf{n})(\chi - I_h^*\chi)dS. \end{aligned} \tag{3.4}$$

Here, $\bar{\mathcal{A}}_K$ is a function designed in a piecewise manner such that for any edge E of a triangle $K \in T_h$ and $x \in E$, $\bar{\mathcal{A}}_K(x) = \mathcal{A}(x_c)$, where x_c is the midpoint of E. Applying the Cauchy–Schwarz inequality and using the fact that $\|\chi - I_h^*\chi\| \le Ch\|\chi\|_1$ and $|\mathcal{A}(x) - \bar{\mathcal{A}}_K| \le h\|\mathcal{A}\|_{1,\infty}$, we obtain

$$|A(\phi, \chi - I_h^*\chi)| \le Ch\|\phi\|_1\|\chi\|_1,$$

and this completes the proof. □

Set $\rho = u - W_h u$. We now establish the H^1-error estimate for ρ and its temporal derivative.

THEOREM 3.1. *Let $W_h u$ be defined by* (3.1). *Then we have*

$$\|\rho(t)\|_1 \le Ch\left(\|u(t)\|_2 + \int_0^t \|u(s)\|_2 ds\right),$$
$$\|\rho_t(t)\|_1 \le Ch\left(\|u(t)\|_2 + \|u_t(t)\|_2 + \int_0^t \|u(s)\|_2 ds\right).$$

Proof. With $\phi_h = I_h u - W_h u$, we have

$$\begin{aligned} c\|\rho\|_1^2 \le A(\rho,\rho) &= A(\rho, u - I_h u) + A(\rho, I_h u - W_h u) \\ &= A(\rho, u - I_h u) + A(\rho, \phi_h - I_h^*\phi_h) + \int_0^t B(t,s;\rho(s), I_h^*\phi_h)ds. \end{aligned}$$

An application of (3.3) yields

$$c\|\rho\|_1^2 \le Ch(\|u\|_2 + \|u\|_1)\|\rho\|_1 + C\left(\int_0^t \|\rho\|_1 ds\right)(\|\rho\|_1 + h\|u\|_2),$$

where for the last term on the right we have used the fact that $\|\phi_h\|_1 \le C(\|\rho\|_1 + h\|u\|_2)$. Kicking back $\|\rho\|_1$, we get

$$\|\rho\|_1^2 \le C\left(h^2\|u\|_2^2 + \int_0^t \|\rho\|_1^2 ds\right).$$

Now applying Gronwall's lemma, we obtain the first inequality. To estimate $\|\rho_t\|_1$, we differentiate (3.1) with respect to time t to get

$$A(\rho_t, I_h^*\chi) = B(t,t,\rho(t), I_h^*\chi) + \int_0^t B_t(t,s;\rho(s), I_h^*\chi)ds. \tag{3.5}$$

As before, with $\phi_h = I_h u_t - W_h u_t$ we obtain

$$\begin{aligned} c\|\rho_t\|_1^2 \le A(\rho_t,\rho_t) = A(\rho_t, u_t - I_h u_t) + A(\rho_t, \phi_h - I_h^*\phi_h) + B(t,t,\rho(t), I_h^*\phi_h) \\ + \int_0^t B_t(t,s;\rho(s), I_h^*\phi_h)ds. \end{aligned}$$

Now apply (3.3), the estimate of $\|\rho\|_1$, and the standard kickback argument to obtain the second inequality. □

Next, we derive L^2 estimates for $\rho = u - W_h u$ and its temporal derivative in the following theorem.

Theorem 3.2. *Let $W_h u$ be defined by* (3.1). *Then we have*

$$\|\rho(t)\| \le Ch^2 \left(\|u(t)\|_2 + \int_0^t \|u(s)\|_2 ds \right),$$
$$\|\rho_t(t)\| \le Ch^2 \left(\|u(t)\|_2 + \|u_t(t)\|_2 + \int_0^t \|u(s)\|_2 ds \right).$$

Proof. The proof will proceed by the duality argument. For $t \in (0,T)$ let $\psi(t) \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution of

$$\begin{aligned} A\psi &= \rho \quad \text{in } \Omega, \\ \psi &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{3.6}$$

satisfying the following regularity estimate (recall that Ω is convex):

$$\|\psi\|_2 \le C\|\rho\|. \tag{3.7}$$

Multiplying (3.6) by ρ and then integrating over Ω, we obtain

$$\begin{aligned} \|\rho\|^2 &= A(\rho, \psi - I_h\psi) + A(\rho, I_h\psi - I_h^*(I_h\psi)) \\ &\quad + \int_0^t B(t,s;\rho(s), I_h^*(I_h\psi) - I_h\psi) ds + \int_0^t B(t,s;\rho(s), I_h\psi - \psi) ds \\ &\quad + \int_0^t B(t,s;\rho(s),\psi) ds = I_1 + I_2 + I_3 + I_4 + I_5. \end{aligned}$$

In view of Theorem 3.1, I_1 and I_4 are bounded as

$$|I_1| + |I_4| \le Ch^2 \left(\|u\|_2 + \int_0^t \|u\|_2 ds \right) \|\psi\|_2.$$

For I_2 and I_3, an application of Lemma 3.3 and Theorem 3.1 yields

$$|I_2| + |I_3| \le Ch \left(\|\rho\|_1 + \int_0^t \|\rho\|_1 ds \right) \|\psi\|_1 \le Ch^2 \left(\|u\|_2 + \int_0^t \|u\|_2 ds \right) \|\psi\|_1.$$

Finally, I_5 is estimated as

$$|I_5| \le \left| \int_0^t (\rho(s), B^*(t,s)\psi) \right| ds \le C \left(\int_0^t \|\rho\| ds \right) \|\psi\|_2,$$

where $B^*(t,s)$ is the adjoint of $B(t,s)$. Now putting these estimates together and with an aid of (3.7) we obtain

$$\|\rho\| = Ch^2 \left(\|u\|_2 + \int_0^t \|u\|_2 ds \right) + C \int_0^t \|\rho\| ds.$$

Finally, an application of Gronwall's lemma yields the first estimate. To estimate $\|\rho_t\|$, we again use the duality argument, (3.5), and the estimate of $\|\rho\|$ to complete the proof. □

Remark 3.2. (i) The estimates in Theorem 3.2 are optimal with respect to the order of convergence as well as the regularity requirement on the solution. This improves upon the earlier result of [13] and [14] by requiring less regularity on the solution.

(ii) In the absence of an integral term (when $B(t,s)=0$), as a consequence of Theorems 3.1 and 3.2, error estimates associated with the Petrov–Ritz projection $R_h : H_0^1 \to S_h$ defined by

$$A(R_h u - u, I_h^* \chi) = 0 \ \ \forall \, \chi \in S_h, \ \ t \in \bar{J},$$

can easily be obtained. Thus, we immediately have

$$\|R_h u - u\| + h\|R_h u - u\|_1 \le Ch^j \|u\|_j, \ \ u \in H_0^1(\Omega) \cap H^j(\Omega), \ \ j = 1, 2. \tag{3.8}$$

Below, we shall prove a lemma which is crucial for the error estimate in the case of nonsmooth initial data to be discussed in section 5.

Define $\hat{\rho}(t) = \int_0^t \rho(\tau) d\tau$. Then, integrating by parts we rewrite (3.1) as

$$\begin{aligned} A(\hat{\rho}_t(t), I_h^* \chi) &= \int_0^t B(t,s;\hat{\rho}_s(s), I_h^* \chi) ds \\ &= B(t,t,\hat{\rho}, I_h^* \chi) - \int_0^t B_s(t,s;\hat{\rho}(s), I_h^* \chi) ds. \end{aligned}$$

Integrate from 0 to t to obtain

$$A(\hat{\rho}(t), I_h^* \chi) = \int_0^t B(s,s,\hat{\rho}(s), I_h^* \chi) ds - \int_0^t \int_0^s B_s(s,\tau;\hat{\rho}(\tau), I_h^* \chi) d\tau ds. \tag{3.9}$$

LEMMA 3.2. *Let u be the solution of the initial value problem* (1.1) *with* $f = 0$ *and* $\hat{\rho}(t) = \int_0^t (u - W_h u)(s) ds$. *Then we have*

$$\|\hat{\rho}\| + h\|\hat{\rho}\|_1 \le Ch^2 \|u_0\|.$$

Proof. With $\phi_h = I_h \hat{u} - W_h \hat{u}$, we have

$$\begin{aligned} c\|\hat{\rho}\|_1^2 &\le A(\hat{\rho},\hat{\rho}) = A(\hat{\rho}, \hat{u} - I_h \hat{u}) + A(\hat{\rho}, I_h \hat{u} - W_h \hat{u}) \\ &\le A(\rho, \hat{u} - I_h \hat{u}) + A(\hat{\rho}, \phi_h - I_h^* \phi_h) + \int_0^t B(s,s;\hat{\rho}(s), I_h^* \phi_h) ds \\ &\quad - \int_0^t \int_0^s B_s(s,\tau;\hat{\rho}(\tau), I_h^* \phi_h) d\tau ds, \end{aligned}$$

where $\hat{u}(t) = \int_0^t u(s) ds$. Then proceeding as in the estimate of $\|\rho\|_1$ in Theorem 3.1 we obtain

$$\|\hat{\rho}\|_1 \le Ch \left(\|\hat{u}\|_2 + \int_0^t \|\hat{u}\|_2 ds \right). \tag{3.10}$$

Now it remains to estimate $\|\hat{u}\|_2$. From (1.1) with $f = 0$, we have

$$Au = -u_t + \int_0^t B(t,s)\hat{u}_s(s) ds = -u_t + B(t,t)\hat{u}(t) - \int_0^t B_s(t,s)\hat{u}(s) ds.$$

Integrating from 0 to t and then using elliptic regularity and Lemma 2.1, we obtain

$$\|\hat{u}\|_2 \le \|u_0\| + \|u(t)\| + C\int_0^t \|\hat{u}\|_2 ds \le C\|u_0\| + C\int_0^t \|\hat{u}\|_2 ds.$$

Now an application of Gronwall's lemma yields

$$\|\hat{u}\|_2 \le C\|u_0\|. \tag{3.11}$$

Combine (3.10) and (3.11) to obtain $\|\hat{\rho}\|_1$. Next, using (3.9), the proof technique of $\|\rho\|$ in Theorem 3.2, and (3.11), the estimate of $\|\hat{\rho}\|$ can be easily obtained. This completes the rest of the proof. □

4. Error estimates for problems with smooth initial data. In this section, we estimate the error of the semidiscrete FVE method for problems with smooth initial data. In particular, an optimal-order L^2-error estimate is obtained when $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$.

As usual we write the error $e(t) = u(t) - u_h(t)$ as a sum of two terms $e(t) = (u - W_h u) + (W_h u - u_h) = \rho + \theta$. The estimate of $\|\rho\|$ is already established, so it is enough to estimate $\|\theta\|$. Using (2.3), an equation of the form (2.3) with u_h replaced by u, and (3.1), it is easy to verify that θ satisfies an error equation of the form

$$(\theta_t, I_h^*\chi) + A(\theta, I_h^*\chi) = \int_0^t B(t,s;\theta(s), I_h^*\chi)ds - (\rho_t, I_h^*\chi) \ \ \forall \chi \in S_h. \tag{4.1}$$

Analogously, integrating (2.3) from 0 to t and then using the resulting equation with u_h replaced by u, (3.9), and $u_h(0) = \tilde{P}_h u_0$, we obtain an error equation in $\hat{\theta}$ as

$$\begin{aligned}(\hat{\theta}_t, I_h^*\chi) + A(\hat{\theta}, I_h^*\chi) = &\int_0^t B(s,s;\hat{\theta}(s), I_h^*\chi)ds \\ &- \int_0^t \int_0^s B_\tau(s,\tau;\hat{\theta}(\tau), I_h^*\chi)d\tau ds - (\rho, I_h^*\chi), \ \ \chi \in S_h,\end{aligned} \tag{4.2}$$

where $\hat{\theta}(t) = \int_0^t \theta(s)ds$. Below, we shall prove a sequence of lemmas that will lead us to the desired result.

LEMMA 4.1. *There is a positive constant C such that*

$$\|\hat{\theta}(t)\|^2 + \int_0^t \|\hat{\theta}(s)\|_1^2 ds \le C \int_0^t \|\rho(s)\|^2 ds.$$

Proof. Choose $\chi = \hat{\theta}$ in (4.2) to have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}(\hat{\theta}, I_h^*\hat{\theta}) + A(\hat{\theta}, I_h^*\hat{\theta}) = &\int_0^t B(s,s;\hat{\theta}(s), I_h^*\hat{\theta}(t))ds \\ &- \int_0^t \int_0^s B_\tau(s,\tau;\hat{\theta}(\tau), I_h^*\hat{\theta})d\tau ds - (\rho, I_h^*\hat{\theta}).\end{aligned}$$

Integrating from 0 to t and using the standard kickback argument yield

$$\|\hat{\theta}(t)\|^2 + \int_0^t \|\hat{\theta}(s)\|_1^2 ds \le C \int_0^t \|\rho(s)\|^2 ds + C \int_0^t \int_0^s \|\hat{\theta}(\tau)\|_1^2 d\tau ds.$$

Finally, apply Gronwall's lemma to complete the rest of the proof. □

LEMMA 4.2. *There is a positive constant C such that*

$$\int_0^t \|\theta(s)\|^2 ds + \|\hat{\theta}(t)\|_1^2 \le C \int_0^t \|\rho(s)\|^2 ds.$$

Proof. Take $\chi = \theta$ in (4.2) and integrate from 0 to t to have

$$\begin{aligned}\int_0^t (\theta, I_h^*\theta)ds + \frac{1}{2}A(\hat{\theta}, I_h^*\hat{\theta}) &= \int_0^t \int_0^s B(\tau,\tau;\hat{\theta}(\tau), I_h^*\theta(s))d\tau ds \\ &\quad - \int_0^t \int_0^s \int_0^\tau B_{\tau'}(\tau,\tau';\hat{\theta}(\tau'), I_h^*\theta(s))d\tau' d\tau ds - (\rho, I_h^*\theta) \\ &= I_1 + I_2 + I_3.\end{aligned}$$

For I_1, we note that

$$\begin{aligned}I_1 &= \int_0^t \int_\tau^t B(\tau,\tau;\hat{\theta}(\tau), I_h^*\hat{\theta}_s(s))dsd\tau \\ &= \int_0^t B(\tau,\tau;\hat{\theta}(\tau), I_h^*\hat{\theta}(t))d\tau ds - \int_0^t B(\tau,\tau;\hat{\theta}(\tau), I_h^*\hat{\theta}(\tau))d\tau.\end{aligned}$$

Similarly, we rewrite the term I_2. Now use the standard kickback argument to obtain

$$\int_0^t \|\theta(s)\|^2 ds + \|\hat{\theta}\|_1^2 \le C\int_0^t \|\hat{\theta}\|_1^2 ds + C\int_0^t \|\rho\|^2 ds.$$

Finally, an application of Lemma 4.1 completes the rest of the proof. □

LEMMA 4.3. *There is a positive constant C such that*

$$t\|\theta(t)\|^2 + \int_0^t s\|\theta(s)\|_1^2 ds \le C\int_0^t \{\|\rho(s)\|^2 + s^2\|\rho_s(s)\|^2\}ds.$$

Proof. Take $\chi = t\theta$ in (4.1) and integrate by parts to have

$$\begin{aligned}\frac{1}{2}\frac{d}{dt}\{t(\theta, I_h^*\theta)\} + tA(\theta, I_h^*\theta) &= \frac{1}{2}(\theta, I_h^*\theta) + tB(t,t;\hat{\theta}(t), I_h^*\theta(t)) \\ &\quad - \int_0^t tB_s(t,s;\hat{\theta}(s), I_h^*\theta(t))ds - t(\rho_t, I_h^*\theta).\end{aligned}$$

Integrating from 0 to t and applying the standard kickback argument, we obtain

$$\begin{aligned}t\|\theta(t)\|^2 + c\int_0^t s\|\theta(s)\|_1^2 ds &\le C\int_0^t \|\hat{\theta}(s)\|_1^2 ds + C\int_0^t \int_0^s \|\hat{\theta}(\tau)\|_1^2 d\tau ds \\ &\quad + C\int_0^t \{\|\theta\|^2 + s^2\|\rho_s\|^2\}ds.\end{aligned}$$

Then use Lemmas 4.1 and 4.2 to complete the proof. □

The main result of this section is given in the following theorem.

THEOREM 4.1. *Let u and u_h, respectively, satisfy* (1.1) *and* (2.3) *with* $f = 0$. *Then for $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$ and $u_h(0) = \tilde{P}_h u_0$, we have*

$$\|e(t)\| \le Ch^2\|u_0\|_2.$$

Proof. By the triangle inequality, we write

$$t^{1/2}\|e(t)\| \le t^{1/2}\|\rho(t)\| + t^{1/2}\|\theta(t)\|.$$

By Theorem 3.2 and a priori estimates in Lemma 2.3, the first term on the right is bounded by

$$t^{1/2}\|\rho(t)\| \le Ch^2 t^{1/2}\|u_0\|_2.$$

For the second term, we use Lemma 4.3, Theorem 3.2, and a priori estimates in Lemma 2.3 to have

$$t^{1/2}\|\theta\| \le C\left(\int_0^t \{\|\rho(s)\|^2 + s^2\|\rho_s(s)\|^2\}ds\right)^{1/2} \le Ch^2 t^{1/2}\|u_0\|_2.$$

Altogether these estimates yield the desired result and this completes the proof. □

Remark 4.4. Note that the result presented in Theorem 4.1 is optimal with respect to the approximation property as well as the regularity of the solution. Similar result for finite element methods is established in [31, 25, 26].

5. Error estimates for nonsmooth initial data. In this section we establish one of the main results of the paper, namely, an error estimate for problems with nonsmooth initial data. More precisely, an almost optimal-order L^2-error estimate is obtained when $u_0 \in L^2(\Omega)$.

The following lemma is useful in our subsequent analysis.

LEMMA 5.1. *For all $\chi_1, \chi_2 \in S_h$, we have*

$$A(\chi_1, \chi_2 - I_h^*\chi_2)| \le Ch^2(\|\chi_1\|_1 + h^{-1/2}\|\hat{u} - \chi_1\|_1)\|\chi_2\|_1,$$

where $\hat{u}(t) = \int_0^t u(s)ds$. The above estimate also holds true when $A(\cdot,\cdot)$ is replaced by $B(t,s;\cdot,\cdot)$.

Proof. From (3.4), we have

$$\begin{aligned} A(\chi_1, \chi_2 - I_h^*\chi_2) = &- \sum_{K\in T_h}\int_K (\nabla\cdot\mathcal{A}\nabla\chi_1)(\chi_2 - I_h^*\chi_2)dx \\ &+ \sum_{K\in T_h}\int_{\partial K}((\mathcal{A} - \bar{\mathcal{A}}_K)\nabla\chi_1\cdot\mathbf{n})(\chi_2 - I_h^*\chi_2)dS = I_1 + I_2. \end{aligned}$$

Since the dual mesh is formed by the barycenters, we have

$$\int_K (\chi - I_h^*\chi)dx = 0 \quad \forall\chi \in T_h,$$

and hence, we apply the Cauchy–Schwarz inequality to have

$$|I_1| = \left|\sum_{K\in T_h}\int_K \left(\nabla\cdot\mathcal{A}\nabla\chi_1 - (\overline{\nabla\cdot\mathcal{A}\nabla\chi_1})_K\right)(\chi_2 - I_h^*\chi_2)\,dx\right| \le Ch^2\|\chi_1\|_1\|\chi_2\|_1,$$

where $(\overline{\nabla\cdot\mathcal{A}\nabla\chi_1})_K = \frac{1}{\text{area}(K)}\int_K \nabla\cdot A\nabla\chi_1\,dx$. Since $\nabla\hat{u}\cdot\mathbf{n}$ is continuous across any edge $E \in T_h$, we may rewrite I_2 as

$$I_2 = \sum_{K\in T_h}\int_{\partial K}\left((\mathcal{A} - \bar{\mathcal{A}}_K)\nabla(\hat{u} - \chi_1)\cdot\mathbf{n}\right)(\chi_2 - I_h^*\chi_2)dS,$$

and hence using the fact that $|\mathcal{A}(x) - \bar{\mathcal{A}}_K| \le h\|\mathcal{A}\|_{1,\infty}$, the Cauchy–Schwarz inequality, and trace results, we obtain

$$|I_2| \le Ch\sum_{K\in T_h}\|\nabla(\hat{u} - \chi_1)\|_{L^2(\partial K)}\|\chi_2 - I_h^*\chi_2\|_{L^2(\partial K)} \le Ch^{3/2}\|\hat{u} - \chi_1\|_1\|\chi_2\|_1.$$

Combining these estimates we complete the proof. □

Below, we shall prove several lemmas which will be used to derive error estimates for problems with nonsmooth initial data.

LEMMA 5.2. *Let u and u_h be the solution of* (1.1) *and* (2.3)*, respectively. Then for $u_0 = 0$, we have*

$$\int_0^t \|u(s) - u_h(s)\|_1^2 ds \leq Ch^2 \left(\|f(0)\|^2 + \int_0^t \|f\|^2 ds \right).$$

Proof. Set $\chi = R_h e$ in the error equation

$$(e_t, I_h^* \chi) + A(e, I_h^* \chi) = \int_0^t B(t, s; e(s), I_h^* \chi) ds \tag{5.1}$$

to get

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}|||e|||^2 + A(e,e) &= (e_t, u - I_h^*(R_h u)) + A(e, u - I_h^*(R_h u)) \\ &\quad - \int_0^t B(t, s; e(s), I_h^*(R_h u) - I_h^* u_h) ds + (e_t, I_h^* u_h - u_h) \\ &\quad + A(e, I_h^* u_h - u_h) = I_1 + I_2 + I_3 + I_4 + I_5. \end{aligned}$$

By (3.8), (3.2), and (3.3), we have

$$\begin{aligned} |I_1| + |I_2| &\leq |(e_t, u - R_h u)| + |(e_t, R_h u - I_h^*(R_h u))| \\ &\quad + |A(e, u - R_h u)| + |A(e, R_h u - I_h^*(R_h u))| \\ &\leq Ch^2(\|e_t\|\|u\|_2 + \|e_t\|_1\|u\|_1) + Ch\|e\|_1(\|u\|_2 + \|u\|_1). \end{aligned}$$

Again, in view of (3.2) and (3.3), I_4 and I_5 can be estimated as

$$|I_4| + |I_5| \leq C(h^2\|e_t\|_1 + h\|e\|_1)\|u_h\|_1.$$

To estimate I_3, we first rewrite it as

$$\begin{aligned} I_3 = &\int_0^t B(t, s; e(s), I_h^*(R_h u) - R_h u) ds + \int_0^t B(t, s; e(s), R_h u - u) ds \\ &+ \int_0^t B(t, s; e(s), e) ds + \int_0^t B(t, s; e(s), u_h - I_h^* u_h) ds. \end{aligned}$$

Apply (3.3) and (3.8) to obtain

$$|I_3| \leq Ch \left(\int_0^t \|e(s)\|_1 ds \right) (\|u\|_1 + \|u\|_2 + \|u_h\|_1) + C \left(\int_0^t \|e(s)\|_1 ds \right) \|e\|_1.$$

Combining these estimates now leads to

$$\begin{aligned} \frac{1}{2}\frac{d}{dt}\|e\|^2 + A(e,e) &\leq Ch^2\|e_t\|\|u\|_2 + Ch^2\|e_t\|_1(\|u\|_1 + \|u_h\|_1) + Ch\|e\|_1\|u\|_1 \\ &\quad + h\left(\int_0^t \|e\|_1 ds \right)(\|u_1\| + \|u\|_2 + \|u_h\|_1) + C\left(\int_0^t \|e(s)\|_1 ds \right)\|e\|_1. \end{aligned}$$

Integrate from 0 to t, use the fact that $e(0) = 0$, and then apply the standard kickback argument to obtain

$$\int_0^t \|e(s)\|_1^2 ds \le Ch^2 \left[\int_0^t (\|u_h\|_1^2 + \|u\|_2^2 + \|e_t\|^2 + \|e_t\|_1^2) ds\right] + C\int_0^t \int_0^s \|e(\tau)\|_1^2 d\tau ds.$$

The desired estimate now easily follows from Lemmas 2.1 and 2.2, its discrete analogue, and Gronwall's lemma. □

Define $\hat{e}(t) = \int_0^t e(s)ds$. Then, as a consequence of Lemmas 4.2 and 3.2 and a priori estimates we have the following lemma.

LEMMA 5.3. *Assume that $u_0 \in L^2(\Omega)$ and $f = 0$. Then there is a positive constant C independent of h such that*

$$\|\hat{e}(t)\|_1 \le Ch\|u_0\|.$$

In order to obtain optimal L^2-error estimate for problems with nonsmooth data, it is convenient to prove an estimate of $\|\hat{e}\|$. For this purpose, we now consider the following backward problems. For fixed time $t > 0$ and given any $\bar{f} \in L^2(\Omega)$, let $v(s) \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution of the backward problem

$$v_s - Av = -\int_s^t B^*(\tau, s)v(\tau)d\tau + \bar{f}, \quad s \le t, \tag{5.2}$$

with $v(t) = g$, where $B^*(\tau, s)$ is the adjoint of $B(\tau, s)$.

The associated weak solution is then defined to be the function $v : [0, t) \to H_0^1(\Omega)$ such that

$$(\phi, v_s) - A(\phi, v) = -\int_s^t B(\tau, s; \phi, v(\tau))d\tau + (\phi, \bar{f}) \quad \forall \phi \in H_0^1(\Omega), \ s \le t, \tag{5.3}$$

with $v(t) = g$. Analogous to (2.3), the FVE approximation is then defined to be the function $v_h : [0, t) \to S_h$ such that

$$(I_h^*\chi, v_{hs}) - A(I_h^*\chi, v_h) = -\int_s^t B(\tau, s; I_h^*\chi, v_h(\tau))d\tau + (I_h^*\chi, \bar{f}) \tag{5.4}$$

for all $\chi \in S_h$, $s \le t$, with $v_h(t) = g_h$, where g_h is a suitable approximation of g in S_h to be defined later.

Remark 5.4. With a simple change of variables in the proofs of Lemmas 2.1–2.3 and using the backward Gronwall lemma, it is easy to obtain a priori bounds for the backward solutions v and v_h.

LEMMA 5.4. *Assume that $u_0 \in L^2(\Omega)$ and $f = 0$. Then there is a generic constant C such that*

$$\|\hat{e}(t)\| \le Ch^2\|u_0\| \quad \forall t > 0. \tag{5.5}$$

Proof. Let $w(s) \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution of the backward problem (5.2) with $\bar{f} = \hat{e}$ and $g = 0$. Then, with a change of variables in the proofs of Lemmas 2.1–2.3 and its discrete analogue, Lemma 5.2 and using the backward Gronwall lemma, it is an easy exercise to check that the solution $w(s)$ and its FVE solution $w_h(s)$, which may be stated in a manner similar to (5.3)–(5.4), satisfy the following estimate:

$$\int_0^t \{\|w_s - w_{hs}\|^2 + \|w_s - w_{hs}\|_1^2 + h^{-2}\|w - w_h\|_1^2 + \|w\|_2^2\}ds \le C\int_0^t \|\hat{e}\|^2 ds. \tag{5.6}$$

We take an L^2 inner product of (5.2) with e and use (5.1) to obtain

$$\begin{aligned}\frac{1}{2}\frac{d}{ds}\|\hat{e}(s)\|^2 = &\frac{d}{ds}(e, I_h^* w_h) + (e, w_s - w_{hs}) - (e, w_{hs} - I_h^* w_{hs}) - A(e, w - w_h)\\ &-A(e, w_h - I_h^* w_h) + \int_s^t B(\tau, s; e(s), (w - w_h)(\tau))d\tau\\ &+\int_s^t B(\tau, s; e(s), (w_h - I_h^* w_h)(\tau))d\tau + \int_s^t B(\tau, s; e(s), I_h^* w_h(\tau))d\tau\\ &-\int_0^s B(s, \tau; e(\tau), I_h^* w_h(s))d\tau.\end{aligned}$$

With $\eta = u - I_h^*(R_h u)$ and $\zeta = I_h^* u_h - u_h$, we rewrite the above equation as

$$\begin{aligned}\frac{1}{2}\frac{d}{ds}\|\hat{e}(s)\|^2 = &\frac{d}{ds}(e, I_h^* w_h) + (\eta, w_s - w_{hs}) - A(\eta, w - w_h)\\ &+\int_s^t B(\tau, s; \eta(s), (w - w_h)(\tau))d\tau + (e, w_{hs} - I_h^* w_{hs}) - A(e, w_h - I_h^* w_h)\\ &+\int_s^t B(\tau, s; e(s), w_h(\tau) - I_h^* w_h(\tau))d\tau + (\zeta, w_s - w_{hs}) - A(\zeta, w - w_h)\\ &+\int_s^t B(\tau, s; \zeta(s), (w - w_h)(\tau))d\tau.\end{aligned}$$

Multiply both sides by s and integrate from 0 to t to have

$$\begin{aligned}\frac{1}{2}t\|\hat{e}(t)\|^2 = &\frac{1}{2}\int_0^t \|\hat{e}(s)\|^2 ds + \int_0^t (e, I_h^* w_h) ds + \int_0^t s(\eta, w_s - w_{hs}) ds\\ &-\int_0^t sA(\eta, w - w_h) ds + \int_0^t \int_s^t sB(\tau, s; \eta, (w - w_h))(\tau) d\tau ds\\ &+\int_0^t s(e, w_{hs} - I_h^* w_{hs}) ds - \int_0^t sA(e, w_h - I_h^* w_h) ds\\ &-\int_0^t \int_0^s sB(s, \tau; e(\tau), I_h^* w_h(s) - w_h(s)) d\tau ds + \int_0^t s(\zeta, w_s - w_{hs}) ds\\ &-\int_0^t sA(\zeta, w - w_h) ds + \int_0^t \int_s^t sB(\tau, s; \zeta(s), (w - w_h)(\tau)) d\tau ds\\ = &\frac{1}{2}\int_0^t \|\hat{e}(s)\|^2 ds + \sum_{i=1}^{10} I_i.\end{aligned}$$

Since $\hat{e}(0) = 0 = I_h^* w(t)$, we obtain using (5.6)

$$\begin{aligned}|I_1| = &\left|-\int_0^t (\hat{e}, I_h^* w_{hs}) ds\right| \le C\int_0^t \|\hat{e}\|\|w_{hs}\| ds\\ \le &\left(\int_0^t \|\hat{e}\|^2 ds\right)^{1/2}\left(\int_0^t \|w_{hs}\|^2 ds\right)^{1/2} \le C\int_0^t \|\hat{e}(s)\|^2 ds.\end{aligned}$$

For I_2, an application of (3.2), (3.8), (5.6), and a priori estimates yield

$$\begin{aligned}|I_2| &= \left| \int_0^t s(u - R_h u, w_s - w_{hs})ds + \int_0^t s(R_h u - I_h^*(R_h u), w_s - w_{hs})ds \right| \\ &\le Ch^4 \int_0^t (s^2\|u\|_2^2 + s\|u\|_1^2)ds + C\int_0^t \|w_s - w_{hs}\|_1^2 ds \\ &\le Ch^4 t\|u_0\|^2 + C\int_0^t \|\hat{e}\|^2 ds.\end{aligned}$$

Similarly, for I_3 and I_4, using (3.8), (3.3), and (5.6), we obtain

$$\begin{aligned}|I_3| + |I_4| &\le C\int_0^t s\|u - R_h u\|_1 \|w - w_h\|_1 ds + h\int_0^t s\|R_h u - I_h^*(R_h u)\|_1 \|w - w_h\|_1 ds \\ &\le Ch^4 \int_0^t (s^2\|u\|_2^2 + s^2\|u\|_1^2)ds + Ch^{-2}\int_0^t \|w - w_h\|_1^2 ds \\ &\le Ch^4 t\|u_0\|^2 + C\int_0^t \|\hat{e}\|^2 ds.\end{aligned}$$

Apply (3.2), (5.6), and a priori estimates to have

$$|I_5| = Ch^4 \int_0^t s\|e\|_1^2 ds + C\int_0^t \|w_{hs}\|_1^2 ds \le Ch^4 t\|u_0\|^2 + C\int_0^t \|\hat{e}\|^2 ds.$$

For I_7, with a change of variables and integration by parts, we note that

$$\begin{aligned}\int_0^t \int_s^t sB(\tau, s; e(s), (w_h - I_h^* w_h)(\tau))d\tau ds &= \int_0^t \int_0^s \tau B(s, \tau; \hat{e}_\tau(\tau), (w_h - I_h^* w_h)(s))d\tau ds \\ &= \int_0^t sB(s, s; \hat{e}(s), (w_h - I_h^* w_h)(s))ds \\ &\quad - \int_0^t \int_0^s \tau B_\tau(s, \tau; \hat{e}(\tau), (w_h - I_h^* w_h)(s))d\tau ds \\ &\quad - \int_0^t \int_0^s B(s, \tau; \hat{e}(\tau), (w_h - I_h^* w_h)(s))d\tau ds.\end{aligned}$$

Similarly, we rewrite the term I_6 as

$$I_6 = \int_0^t sA(\hat{e}, w_{hs} - I_h^* w_{hs})ds + \int_0^t A(\hat{e}, w_h - I_h^* w_h)ds,$$

where we have used the fact that $w_h(t) = 0 = \hat{e}(0)$. Thus, applying (3.3), Lemma 5.3, and (5.6), I_6 and I_7 are bounded by

$$|I_6| + |I_7| \le Cth^4\|u_0\|^2 + C\int_0^t (\|w_h\|_1^2 + \|w_{hs}\|_1^2)ds \le Cth^4\|u_0\|^2 + C\int_0^t \|\hat{e}\|^2 ds.$$

Finally, using (3.2), (3.3), and (5.6), we obtain

$$\begin{aligned}|I_8| + |I_9| + |I_{10}| &\le Ch^4 \int_0^t s\|u_h\|_1^2 ds + C\int_0^t (\|w_s - w_{hs}\|_1^2 + h^{-2}\|w - w_h\|_1^2)ds \\ &\le Ch^4 t\|u_0\|^2 + C\int_0^t \|\hat{e}\|^2 ds.\end{aligned}$$

Altogether this now leads to

$$t\|\hat{e}(t)\| \le Ch^4 t\|u_0\|^2 + C\int_0^t \|\hat{e}(s)\|^2 ds. \tag{5.7}$$

It now remains to estimate $\int_0^t \|\hat{e}\|^2 ds$. Multiply (5.2) by $\hat{e}$ and integrate by parts with respect to x to have

$$\begin{aligned}\|\hat{e}(s)\|^2 &= \frac{d}{ds}(\hat{e}, I_h^* w_h) + (\hat{e}, w_s - w_{hs}) - A(\hat{e}, w - w_h)\\ &\quad + \int_s^t B(\tau, s; \hat{e}(s), w(\tau) - w_h(\tau))d\tau + (\hat{e}, w_{hs} - I_h^* w_{hs}) - A(\hat{e}, w_h - I_h^* w_h)\\ &\quad - \int_0^s B(\tau, \tau; \hat{e}(\tau), I_h^* w_h(s))d\tau + \int_0^s \int_0^\tau B_{\tau'}(\tau, \tau'; \hat{e}(\tau'), I_h^* w_h(s))d\tau' d\tau\\ &\quad + \int_s^t B(\tau, s; \hat{e}(s), w_h(\tau))d\tau. \end{aligned} \tag{5.8}$$

Here, we have used the relation

$$(e, I_h^*\chi) + A(\hat{e}, I_h^*\chi) = \int_0^t B(s, s; \hat{e}(s), I_h^*\chi)ds - \int_0^t \int_0^s B_\tau(s, \tau; \hat{e}(\tau), I_h^*\chi)d\tau ds,$$

which is obtained by integrating (5.1) from 0 to t and using (2.4). Now integrate (5.8) from 0 to t and use the fact that $\hat{e}(0) = 0 = I_h^* w_h(t)$ to have

$$\begin{aligned}\int_0^t \|\hat{e}(s)\|^2 ds &= \int_0^t (\hat{\eta}, w_s - w_{hs})ds - \int_0^t A(\hat{\eta}, w - w_h)ds\\ &\quad - \int_0^t \int_s^t B(\tau, s; \hat{\eta}(s), (w - w_h)(\tau))d\tau ds\\ &\quad - \int_0^t \int_0^s B(\tau, \tau; \hat{e}(\tau), I_h^* w_h(s))d\tau ds\\ &\quad + \int_0^t \int_0^s \int_0^\tau B_{\tau'}(\tau, \tau'; \hat{e}(\tau'), I_h^* w_h(s))d\tau' d\tau ds\\ &\quad + \int_0^t (\hat{e}(s), w_{hs} - I_h^* w_{hs})ds - \int_0^t A(\hat{e}, w_h - I_h^* w_h)ds\\ &\quad + \int_0^t \int_s^t B(\tau, s; \hat{e}(s), w_h(\tau))d\tau ds + \int_0^t (\hat{\zeta}, w_s - w_{hs})ds\\ &\quad - \int_0^t A(\hat{\zeta}, w - w_h)ds - \int_0^t \int_s^t B(\tau, s; \hat{\zeta}(s), (w - w_h)(\tau))d\tau ds\\ &= \sum_{i=1}^{11} J_i,\end{aligned}$$

where $\hat{\eta} = \hat{u} - I_h^*(R_h\hat{u})$ and $\hat{\zeta} = I_h^*\hat{u}_h - \hat{u}_h$. Let us estimate each term separately. For J_1, use of (3.8), (3.11), (5.1), and (5.6) yields

$$\begin{aligned}|J_1| &\le \int_0^t \{|(\hat{u} - R_h\hat{u}, w_s - w_{hs})| + |(R_h\hat{u} - I_h^*(R_h\hat{u}), w_s - w_{hs})|\}ds\\ &\le C(\epsilon)h^4 \int_0^t \|\hat{u}\|_2^2 ds + \epsilon \int_0^t \|w_s - w_{hs}\|^2 ds \le C(\epsilon)th^4\|u_0\|^2 + \epsilon C\int_0^t \|\hat{e}\|^2 ds.\end{aligned}$$

Similarly,

$$|J_2| + |J_3| \le C(\epsilon) t h^4 \|u_0\|^2 + \epsilon h^{-2} \int_0^t \|w - w_h\|_1^2 ds \le C(\epsilon) t h^4 \|u_0\|^2 + \epsilon C \int_0^t \|\hat{e}\|^2 ds.$$

For J_4, we note that

$$\begin{aligned} J_4 = &- \int_0^t \int_0^s B(\tau, \tau; \hat{e}(\tau), I_h^* w_h(s) - w_h(s)) d\tau ds \\ &- \int_0^t \int_0^s B(\tau, \tau; \hat{e}(\tau), w_h(s) - w(s)) d\tau ds - \int_0^t \int_0^s B(\tau, \tau; \hat{e}(\tau), w(s)) d\tau ds, \end{aligned}$$

and hence, using (3.3), Lemma 5.3, and (5.6), we obtain

$$\begin{aligned} |J_4| &\le C \int_0^t \int_0^s \{h \|\hat{e}(\tau)\|_1 (\|w_h(s)\|_1 + h^{-1} \|w(s) - w_h(s)\|_1) + \|\hat{e}(\tau)\| \|w(s)\|_2\} d\tau ds \\ &\le C t h^4 \|u_0\|^2 + C \int_0^t \int_0^s \|\hat{e}(\tau)\|^2 d\tau ds. \end{aligned}$$

Similarly,

$$|J_5| \le C t h^4 \|u_0\|^2 + C \int_0^t \int_0^s \|\hat{e}(\tau)\|^2 d\tau ds.$$

Using (3.1)–(3.3), Lemma 5.3, and (5.6), we obtain

$$\begin{aligned} |J_6| + |J_7| &\le C(\epsilon) h^2 \int_0^t \|\hat{e}\|_1^2 ds + \epsilon \int_0^t \{\|w_{hs}\|_1^2 + \|w_h\|_1^2\} ds \\ &\le C t h^4 \|u_0\|^2 + \epsilon C \int_0^t \|\hat{e}\|^2 ds. \end{aligned}$$

By changing the order of integration, rewrite the term J_8 as

$$J_8 = \int_0^t \int_0^s B(\tau, s; \hat{e}(s), (w_h - w)(\tau)) ds d\tau + \int_0^t \int_0^s B(\tau, s; \hat{e}(s), w(\tau)) ds d\tau.$$

In view of Lemma 5.3 and (5.6), we obtain

$$|J_8| \le C t h^4 \|u_0\|^2 + C \int_0^t \int_0^s \|\hat{e}(\tau)\|^2 d\tau ds.$$

Finally, using (3.2), (3.3), and (5.6), we have

$$\begin{aligned} |J_9| + |J_{10}| + |J_{11}| &\le C(\epsilon) t h^4 \|u_0\|^2 + \epsilon \int_0^t (\|w_s - w_{hs}\|_1^2 + h^{-2} \|w - w_h\|_1^2) ds \\ &\le C(\epsilon) t h^4 \|u_0\|^2 + \epsilon C \int_0^t \|\hat{e}(s)\|^2 ds. \end{aligned}$$

Putting these estimates together and choosing ϵ appropriately, we arrive at

$$\int_0^t \|\hat{e}(s)\|^2 ds \le C t h^4 \|u_0\|^2 + C \int_0^t \int_0^s \|\hat{e}(\tau)\|^2 d\tau ds.$$

An application of Gronwall's lemma yields

$$\int_0^t \|\hat{e}\|^2 ds \le C t h^4 \|u_0\|^2,$$

and this combined with (5.7) completes the rest of the proof. □

Remark 5.5. Defining the error $\bar{e} = v - v_h$ associated with the backward problem (5.3) and its FVE approximation (5.4), set $\tilde{\bar{e}}(s) = -\int_s^t \bar{e}(\tau)d\tau,\ s \le t$. Then, for $g \in L^2(\Omega)$ and $\bar{f} = 0$, analogous to Lemmas 5.3 and 5.4, it is easy to show that

$$\|\tilde{\bar{e}}\|_j \le Ch^{2-j}\|g\|, \ \ j = 0, 1. \tag{5.9}$$

We conclude this section by showing our main result in the following theorem.

THEOREM 5.1. *Let u and u_h be solutions of* (1.1) *and* (1.4), *respectively, with $f = 0$. Assume that $u_0 \in L^2(\Omega)$ and the matrix $\mathcal{A}$ is constant over each element $K \in T_h$. Then there is a generic positive constant C independent of h such that*

$$\|e(t)\| \le Ct^{-1}h^2 \ln h \|u_0\|, \quad t \in J.$$

Proof. Using (5.3) and (5.4) with $\bar{f} = 0$ and Lemma 2.6, we first note that

$$\begin{aligned}
&\frac{d}{ds}\left\{s^2[(u,v) - (I_h^* u_h, v_h)]\right\} \\
&\quad = 2s\left\{(u,v) - (I_h^* u_h, v_h)\right\} + \int_0^s s^2 B(s,\tau; u(\tau), v(s))d\tau \\
&\qquad - \int_s^t s^2 B(\tau, s; u(s), v(\tau))d\tau - \int_0^s s^2 B(s,\tau; u_h(\tau), I_h^* v_h(s) - v_h(s))d\tau \\
&\qquad + \int_s^t s^2 B(\tau, s; I_h^* u_h(s) - u_h(s), v_h(\tau))d\tau - \int_0^s s^2 B(s,\tau; u_h(\tau), v_h(s))d\tau \\
&\qquad + \int_s^t s^2 B(\tau, s; u_h(s), v_h(\tau))d\tau - s^2(u_{hs}, v_h - I_h^* v_h) - s^2(I_h^* u_{hs} - u_{hs}, v_h).
\end{aligned}$$

Integrate the above equation from 0 to t. Then, with $g_h = L_h g$, where $L_h : L^2(\Omega) \to S_h$ defined by $(L_h g, I_h^* \chi) = (g, \chi),\ \chi \in S_h$, we have

$$\begin{aligned}
t^2(e(t), g) = {} & 2\int_0^t s\left\{(u(s), v(s)) - (u_h(s), v_h(s))\right\}ds \\
& - \int_0^t \int_0^s s^2 B(s,\tau; u_h(\tau), I_h^* v_h(s) - v_h(s))d\tau \\
& + \int_0^t \int_s^t s^2 B(\tau, s; I_h^* u_h(s) - u_h(s), v_h(\tau))d\tau - \int_0^t s^2(u_{hs}, v_h - I_h^* v_h)ds \\
& - \int_0^t s^2(u_{hs} - I_h^* u_{hs}, v_h)ds = 2I_1 + I_2 + I_3 + I_4 + I_5.
\end{aligned} \tag{5.10}$$

For the term I_2, with $\hat{u}_h(t) = \int_0^t u_h(s)ds$, we integrate by parts to have

$$\begin{aligned}
I_2 &= -\int_0^t \int_0^s s^2 B(s,\tau; \hat{u}_{h\tau}(\tau), (I_h^* v_h - v_h)(s))d\tau ds \\
&= -\int_0^t s^2 B(s,s; \hat{u}_h(s), (I_h^* v_h - v_h)(s))ds \\
&\quad + \int_0^t \int_0^s s^2 B_\tau(s,\tau; \hat{u}_h(\tau), (I_h^* v_h - v_h)(s))d\tau ds \\
&= I_2^1 + I_2^2.
\end{aligned}$$

For I_2^1, apply Lemma 5.1 with $\chi_1 = \hat{u}_h$ and $\chi_2 = v_h$, Lemma 5.3, and a priori estimates to obtain

$$|I_2^1| \leq Ch^2 \int_0^t s^2(\|\hat{u}_h\|_1 + h^{-1/2}\|\hat{e}\|_1)\|v_h(s)\|_1 ds \leq Cth^2\|u_0\|\|g\|.$$

The term I_2^2 is treated in a similar manner and hence

$$|I_2| \leq Cth^2\|u_0\|\|g\|.$$

Similarly, defining $\tilde{v}_h(s) = -\int_s^t v_h(\tau)d\tau,\ s \leq t$, we rewrite the term I_3 as

$$\begin{aligned} I_3 &= \int_0^t \int_s^t s^2 B(\tau, s; I_h^* u_h(s) - u_h(s), \tilde{v}_{h,\tau}(\tau))d\tau ds \\ &= -\int_0^t s^2 B(s, s; I_h^* u_h(s) - u_h(s), \tilde{v}_h(s))ds \\ &\quad - \int_0^t \int_s^t s^2 B_\tau(\tau, s; I_h^* u_h(s) - u_h(s), \tilde{v}_h(\tau))d\tau ds \\ &= I_3^1 + I_3^2. \end{aligned}$$

As before, again an application of Lemma 5.1 (analogous result for the backward problem), (5.9), and a priori bounds for the discrete solution yield

$$|I_3^1| \leq Ch^2 \int_0^t s^2(\|v_h\|_1 + h^{-1/2}\|\tilde{\bar{e}}\|_1)\|u_h(s)\|_1 ds \leq Cth^2\|u_0\|\|g\|.$$

The term I_3^2 is treated in a similar fashion and hence

$$|I_3| \leq Cth^2\|u_0\|\|g\|.$$

For I_4 and I_5, apply (3.2) and a priori estimates to have

$$|I_4| + |I_5| \leq Cth^2 \left(\int_0^t \|v_h\|_1^2 ds\right)^{1/2} \left(\int_0^t s^2\|u_{hs}\|_1^2 ds\right)^{1/2} \leq Cth^2\|u_0\|\|g\|.$$

It now remains to estimate the term I_1. We first rewrite I_1 as

$$\begin{aligned} I_1 &= \int_0^t s(e(s), v)ds - \int_0^t s(e(s), \bar{e}(s))ds + \int_0^t s(u, \bar{e}(s))ds - \int_0^t s(I_h^* u_h - u_h, v_h)ds \\ &= I_1^1 + I_1^2 + I_1^3 + I_1^4. \end{aligned}$$

To estimate I_1^1, we note that

$$I_1^1 = \int_0^{t-h^2} s(e(s), v(s))ds + \int_{t-h^2}^t s(e(s), v(s))ds = II_1 + II_2.$$

For II_1, we integrate by parts and use the fact that $\hat{e}(0) = 0$ to have

$$\begin{aligned} II_1 = \int_0^{t-h^2} s(\hat{e}_s, v)ds &= (t - h^2)(\hat{e}(t - h^2), v(t - h^2)) \\ &\quad - \int_0^{t-h^2} (\hat{e}, v)ds - \int_0^{t-h^2} s(\hat{e}, v_s)ds, \end{aligned}$$

and hence, by the Cauchy–Schwarz inequality, Lemma 5.4, and Lemma 2.3 (with time reversed), we obtain

$$\begin{aligned} |II_1| &\leq t\|\hat{e}(t-h^2)\|\|v(t-h^2)\| + \int_0^{t-h^2} \|\hat{e}(s)\|\|v(s)\|ds + \int_0^{t-h^2} s\|\hat{e}(s)\|\|v_s(s)\|ds \\ &\leq Cth^2\|u_0\|\|g\| + Cth^2\|u_0\|\|g\| \int_0^{t-h^2} \frac{1}{(t-s)}ds \\ &\leq \left(Cth^2\|u_0\|\|g\| + Cth^2 \ln h\|u_0\|\|g\|\right) \leq Cth^2 \ln h\|u_0\|\|g\|. \end{aligned} \tag{5.11}$$

By Lemma 2.3, its semidiscrete analogue, and further using a priori estimates for the backward solution v, we obtain

$$|II_2| \leq Ct\int_{t-h^2}^{t} \|e(s)\|\|v(s)\|ds \leq Cth^2\|u_0\|\|g\|,$$

which together with (5.11) yields

$$|I_1^1| \leq Cth^2 \ln h\|u_0\|\|g\|.$$

Since $\tilde{\tilde{e}}(t) = 0$, integrate I_1^2 by parts to have

$$I_1^2 = -\int_0^t s(e, \tilde{\tilde{e}}_s)ds = \int_0^t (e, \tilde{\tilde{e}})ds + \int_0^t s(e_s, \tilde{\tilde{e}}(s))ds.$$

Apply the Cauchy–Schwarz inequality, (5.9), and a priori estimates in Lemma 2.3 to obtain

$$|I_1^2| \leq \int_0^t \|e\|\|\tilde{\tilde{e}}(s)\|ds + \int_0^t s\|e_s\|\|\tilde{\tilde{e}}\|ds \leq Cth^2\|u_0\|\|g\|.$$

Similarly, using (5.9) and Lemma 2.3 we estimate I_1^3 as

$$|I_1^3| \leq \int_0^t \|\tilde{\tilde{e}}\|\|u\|ds + \int_0^t s\|\tilde{\tilde{e}}\|\|u_s\|ds \leq Cth^2\|u_0\|\|g\|.$$

Finally, for I_1^4, apply (3.2) and a priori estimates to have

$$|I_1^4| \leq Cth^2 \left(\int_0^t \|u_h\|_1^2\right)^{1/2} \left(\int_0^t \|v_h\|_1^2 ds\right)^{1/2} \leq Cth^2\|u_0\|\|g\|.$$

Altogether these estimates yield the desired result and this completes the proof. □

Acknowledgments. The authors wish to thank the referees for their valuable comments and suggestions which led to the improvement of the presentation of this paper. This first author thanks the Institute for Scientific Computation and the Department of Mathematics, at Texas A&M University for providing necessary facilities for carrying out the research.

REFERENCES

[1] R. E. Bank and D. J. Rose, *Some error estimates for the box method*, SIAM J. Numer. Anal., 24 (1987), pp. 777–787.

[2] J. H. Bramble, A. H. Schatz, V. Thomée, and L. B. Wahlbin, *Some convergence estimates for semidiscrete Galerkin type approximations for parabolic equations*, SIAM J. Numer. Anal., 14 (1977), pp. 218–241.

[3] Z. Cai, *On the finite volume element method*, Numer. Math., 58 (1991), pp. 713–735.

[4] Z. Cai and S. McCormick, *On the accuracy of the finite volume element method for diffusion equations on composite grids*, SIAM J. Numer. Anal., 27 (1990), pp. 636–655.

[5] J. R. Cannon and Y. Lin, *Nonclassical H^1 projections and Galerkin methods for nonlinear parabolic integro-differential equations*, Calcolo, 25 (1988), pp. 187–201.

[6] J. R. Cannon and Y. Lin, *A priori L^2 error estimates for finite-element methods for nonlinear diffusion equations with memory*, SIAM J. Numer. Anal., 27 (1990), pp. 595–607.

[7] P. Chatzipantelidis, *Finite volume methods for elliptic PDE's: A new approach, M2AN*, Math. Model. Numer. Anal., 36 (2002), pp. 307–324.

[8] P. Chatzipantelidis, R. D. Lazarov, and V. Thomée, *Error estimates for the finite volume element method for parabolic equations in convex polygonal domains*, Numer. Methods Partial Differential Equations, 20 (2004), pp. 650–674.

[9] S. H. Chou and Q. Li, *Error estimates in L^2, H^1 and L^∞ in covolume methods for elliptic and parabolic problems: A unified approach*, Math. Comp., 69 (2000), pp. 103–120.

[10] J. H. Cushman and T. R. Glinn, *Nonlocal dispersion in media with continuously evolving scales of heterogeneity*, Trans. Porous Media, 13 (1993), pp. 123–138.

[11] G. Dagan, *The significance of heterogeneity of evolving scales to transport in porous formations*, Water Resour. Res., 30 (1994), pp. 3327–3336.

[12] R. E. Ewing, T. Lin, and Y. Lin, *On the accuracy of the finite volume element method based on piecewise linear polynomials*, SIAM J. Numer. Anal., 39 (2002), pp. 1865–1888.

[13] R. E. Ewing, R. D. Lazarov, and Y. Lin, *Finite volume element approximations of nonlocal in time one-dimensional flows in porous media*, Computing, 64 (2000), pp. 157–182.

[14] R. E. Ewing, R. D. Lazarov, and Y. Lin, *Finite volume element approximations of nonlocal reactive flows in porous media*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 285–311.

[15] R. E. Ewing, Y. Lin, and J. Wang, *A numerical approximation of non-Fickian flows with mixing length growth in porous media*, Acta. Math. Univ. Comenian., 70 (2001), pp. 75–84.

[16] W. Hackbusch, *On first and second order box schemes*, Computing, 41 (1989), pp. 277–296.

[17] H. Jianguo and X. Shitong, *On the finite volume element method for general self-adjoint elliptic problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1762–1774.

[18] M. Huang and V. Thomée, *Some convergence estimates for semidiscrete type schemes for time-dependent nonselfadjoint parabolic equations*, Math. Comp., 37 (1981), pp. 327–346.

[19] R. Li, Z. Chen, and W. Wu, *Generalized Difference Methods for Differential Equations: Numerical Analysis of Finite Volume Methods*, Monogr. Textbooks Pure Appl. Math. 226, Marcel Dekker, New York, 2000.

[20] Y. Lin, V. Thomée, and L. B. Wahlbin, *Ritz–Volterra projections to finite-element spaces and applications to integrodifferential and related equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1047–1070.

[21] M. Luskin and R. Rannacher, *On the smoothing property of the Galerkin method for parabolic equations*, SIAM J. Numer. Anal., 19 (1982), pp. 93–113.

[22] I. D. Mishev, *Finite Volume and Finite Volume Element Methods for Non-Symmetric Problems*, Ph.D. thesis, Technical report ISC-96-04-MATH, Institute for Scientific Computation, Texas A&M University, College Station, TX, 1997.

[23] I. D. Mishev, *Finite volume methods on Voronoi meshes*, Numer. Methods Partial Differential Equations, 14 (1998), pp. 193–212.

[24] A. K. Pani, V. Thomée, and L. B. Wahlbin, *Numerical methods for hyperbolic and parabolic integro-differential equations*, J. Integral Equations Appl., 4 (1992), pp. 533–584.

[25] A. K. Pani and T. E. Peterson, *Finite element methods with numerical quadrature for parabolic integrodifferential equations*, SIAM J. Numer. Anal., 33 (1996), pp. 1084–1105.

[26] A. K. Pani and R. K. Sinha, *Error estimates for semidiscrete Galerkin approximation to a time dependent parabolic integro-differential equation with nonsmooth dada*, Calcolo, 37 (2000), pp. 181–205.

[27] M. Renardy, W. Hrusa, and J. Nohel, *Mathematical Problems in Viscoelasticity*, Pitman Monographs and Surveys in Pure and Applied Mathematics 35, Wiley, New York, 1987.

[28] P. H. Sammon, *Convergence estimates for semidiscrete parabolic equation approximations*, SIAM J. Numer. Anal., 19 (1982), pp. 68–92.

[29] I. H. Sloan and V. Thomée, *Time discretization of an integro-differential equation of parabolic type*, SIAM J. Numer. Anal., 23 (1986), pp. 1052–1061.

[30] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Lecture Notes in Math. 1054, Springer-Verlag, New York, 1984.

[31] V. Thomée and N.-Y. Zhang, *Error estimates for semidiscrete finite element methods for parabolic integro-differential equations*, Math. Comp., 53 (1989), pp. 121–139.

[32] V. Thomée and N.-Y. Zhang, *Backward Euler type methods for parabolic integro-differential equations with nonsmooth data*, in Contributions in Numerical Mathematics, World Sci. Ser. Appl. Anal. 2, World Scientific, River Edge, NJ, 1993, pp. 373–388.

[33] E.G. Yanik and G. Fairweather, *Finite element methods for parabolic and hyperbolic partial integro-differential equations*, Nonlinear Anal. 12 (1988), pp. 785–809.

[34] N.-Y. Zhang, *On fully discrete Galerkin approximations for partial integro-differential equations of parabolic type*, Math. Comp., 60 (1993), pp. 133–166.

SIAM J. Numer. Anal.
Vol. 43, No. 6, pp. 2344–2370

ANALYTICAL AND NUMERICAL STUDY OF A MODEL OF EROSION AND SEDIMENTATION*

ROBERT EYMARD† AND THIERRY GALLOUËT‡

Abstract. We consider the following problem, arising within a geological model of sedimentation-erosion: For a given vector field g and a given nonnegative function F defined on a one- or two-dimensional domain Ω, find a vector field under the form $\tilde{g} = ug$, with $0 \leq u(x) \leq 1$ for a.e. $x \in \Omega$, such that $\text{div}\tilde{g} + F \geq 0$ and $(u-1)(\text{div}\tilde{g} + F) = 0$ in Ω. We first give a weak formulation of this problem, and we prove a comparison principle on a weak solution of the problem. Thanks to this property, we get the proof of the uniqueness of the weak solution. The existence of a solution results from the proof of the convergence of an original scheme. Numerical examples show the efficiency of this scheme and illustrate its convergence properties.

Key words. hyperbolic inequalities, doubling variable technique, process solutions, finite volume methods, erosion and sedimentation models

AMS subject classifications. 65N12, 65N30, 35R45

DOI. 10.1137/040605874

1. Introduction.

In the framework of the petroleum industry, geological simulations are used more and more in order to get a better knowledge of the history of the sedimentary basins. Among them, the computation of the sedimentation and erosion processes leads to a better knowledge of the geometry of the layers and of their lithological nature (see, for example, [15], [11], or [5]). An unknown of such models is the thickness $H(x,t)$ of the sediments at a point $(x,t) \in \Omega \times (0,T)$, where Ω describes the horizontal extension of the basin (the magnitude of the diameter of Ω can be about several hundreds of kilometers) and T is the age of the basin (between 0 and 10^7 years for example). The simplest model is a diffusion equation

$$H_t(x,t) - \text{div}[\Lambda(x)\nabla H(x,t)] = 0 \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{1}$$

where $\Lambda(x)$ is a matrix in the general case, reducing in most of the cases to a scalar function. But the model (1) is not sufficient for actual applications, in particular, because it does not account for the assymetry between the erosion process (due to the action of the weather) and the sedimentation process. Indeed, more realistic models (see [1] or [6] and references therein) are based on the introduction in (1) of a multiplier $\overline{u}(x,t)$ on the fluxes of sediments:

$$H_t(x,t) - \text{div}[\Lambda(x)\overline{u}(x,t)\nabla H(x,t)] = 0 \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{2}$$

in order to satisfy the following constraints on $(\overline{u}, H)$,

$$H_t(x,t) \geq -F(x) \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{3}$$

$$0 \leq \overline{u}(x,t) \leq 1 \text{ for a.e. } (x,t) \in \Omega \times (0,T), \tag{4}$$

*Received by the editors March 29, 2004; accepted for publication (in revised form) June 1, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sinum/43-6/60587.html

†Université de Marne-la-Vallée, 77454 Marne-la-Vallée, France (Robert.Eymard@univ-mlv.fr).
‡Université de Provence, 13453 Marseille cedex 13, France (gallouet@cmi.univ-mrs.fr).

and

$$(\overline{u}(x,t)-1)\ (H_t(x,t)+F(x))=0 \text{ for a.e. } (x,t)\in\Omega\times(0,T). \tag{5}$$

In (3) and (5), we denote by $F(x)\geq 0$ the maximum erosion rate at point x.

In practical situations, F is estimated by the geological study of the sedimental history, and may be improved by solving an inverse problem (which is quite complicated, by the way), using (2)–(5) as the direct problem. In large parts of the simulation domain, the transport of sediments is due mainly to gravity effects, taken into account by a scalar value for the matrix $\Lambda(x)$. In a same way as above, this scalar value can be estimated by some geological studies or by solving an inverse problem. However, the main mechanism for the transport of sediments is the action of surface water flows. These flows, located in river basins, can be represented by introducing anisotropic values for this matrix $\Lambda(x)$. The determination of realistic values for these parameters is not an easy task and is still a challenging subject of research for the simulation of the sedimentary basins. The function $\overline{u}$ is a complete unknown factor, reducing the flux of sediments in order to respect the constraint (3). Despite these difficulties of data identification, this model is considered interesting enough to be actually implemented in an industrial simulator (see [11], [6]).

Existence and uniqueness for the full problem (2)–(5) is an open problem (some partial results can be found in [10] or [2]). Thus we consider a semidiscretization in time of this system of equations. We define a time step $\delta t>0$, and for an integer n such that $n\delta t<T$, we assume that the function $H^{(n)}$ is an approximation of $H(\cdot,n\delta t)$. We then look for the functions $H^{(n+1)}$ and $\overline{u}^{(n+1)}$, respective approximations of $H(\cdot,(n+1)\delta t)$ and $\overline{u}(\cdot,(n+1)\delta t)$, which are solutions of the system of equations

$$\frac{1}{\delta t}(H^{(n+1)}(x)-H^{(n)}(x))-\operatorname{div}[\Lambda(x)\overline{u}^{(n+1)}(x)\nabla H^{(n)}(x)]=0 \text{ for a.e. } x\in\Omega, \tag{6}$$

under the constraints

$$\frac{1}{\delta t}(H^{(n+1)}(x)-H^{(n)}(x))\geq -F(x) \text{ for a.e. } x\in\Omega, \tag{7}$$

$$0\leq \overline{u}^{(n+1)}(x)\leq 1 \text{ for a.e. } x\in\Omega, \tag{8}$$

and

$$(\overline{u}^{(n+1)}(x)-1)\ \left(\frac{1}{\delta t}(H^{(n+1)}(x)-H^{(n)}(x))+F(x)\right)=0 \text{ for a.e. } x\in\Omega. \tag{9}$$

Denoting by $g(x)=\Lambda(x)\nabla H^{(n)}(x)$ and reporting in (7)–(9) the expression of $\frac{1}{\delta t}(H^{(n+1)}-H^{(n)})$ taken from (6), the unknown function $\overline{u}^{(n+1)}$ is then a solution u of the following system of equations:

$$\begin{aligned}&\operatorname{div}[u(x)g(x)]+F(x)\geq 0 \text{ for a.e. } x\in\Omega,\\ &0\leq u(x)\leq 1 \text{ for a.e. } x\in\Omega,\end{aligned} \tag{10}$$

and

$$(u(x)-1)\ (\operatorname{div}[u(x)g(x)]+F(x))=0 \text{ for a.e. } x\in\Omega. \tag{11}$$

Hence, if we are able to prove that problem (10)–(11) has one and only one solution $\tilde{g} = u(\cdot)g(\cdot)$, the function $H^{(n+1)}$ is then given by the relation $H^{(n+1)}(x) = H^{(n)}(x) + \delta t\, \mathrm{div}\tilde{g}(x)$ for a.e. $x \in \Omega$.

Remark 1.1. If there exist some regions where $g = 0$ and $F = 0$ simultaneously, it is clear that any value in $[0, 1]$ is possible for u. Nevertheless, $\tilde{g}$ is uniquely defined by the value 0 in such a region.

A fully implicit version of this method (namely $\nabla H^{(n)}(x)$ is replaced by $\nabla H^{(n+1)}(x)$ in (6)) in addition to a finite volume space discretization are used in an industrial simulator (see [11], [6]).

The aim of this paper is to focus on both the analytical and the numerical aspects of the subproblem (10)–(11). Although it is not clear that the resolution of this subproblem yields the complete theoretical resolution of the fully coupled problem (2)–(5), we emphasize that it leads to the key points of a correct numerical implementation.

In this paper, the following hypotheses, denoted Hypotheses (H), are assumed.

Hypotheses (H).

1. Ω is a bounded open subset of $\mathbb{R}^d$, $d \in \mathbb{N}^\star = \mathbb{N} \setminus \{0\}$ (in applications, $d = 2$), with a Lipschitz continuous boundary $\partial\Omega$ (this gives the existence, for a.e. $x \in \partial\Omega$, of the unit outward vector $\mathbf{n}(x)$ normal to the boundary).
2. There exist two functions, $h \in C^1(\overline{\Omega})$ and $\Lambda\ : \Omega \longrightarrow \mathcal{M}_d$ (the set of bounded, symmetric, definite positive, $d \times d$ matrices) such that the function $g\ : \Omega \to \mathbb{R}^d$, defined by $g(x) = \Lambda(x)\nabla h(x)$ for all $x \in \Omega$, is Lipschitz continuous on $\overline{\Omega}$ and satisfies $g(x) \cdot \mathbf{n}(x) = 0$ for a.e. $x \in \partial\Omega$.
3. $F \in L^\infty(\Omega)$ is such that $F(x) \geq 0$ for a.e. $x \in \Omega$.

As we see below in section 2, there does not always exist a continuous function $u\ : \Omega \to \mathbb{R}$ such that (10)–(11) are satisfied, and the regularity of $\tilde{g} = ug$ in the general case is an open problem. Therefore we first look for a weak formulation of problem (10)–(11). For this purpose, let $\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+)$, and let $\xi \in C^1(\mathbb{R})$ be such that $\xi'(1) \geq 0$. We multiply the first inequality of (10) by $\xi'(u(x))\varphi(x)$, and we integrate on Ω. We get

(12)
$$\int_\Omega \xi'(u(x))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x = \int_\Omega \xi'(1)\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x$$
$$+ \int_\Omega (\xi'(u(x)) - \xi'(1))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x.$$

The second term of the right-hand side vanishes, using (11), and the first one is nonnegative. This leads to

$$\int_\Omega \xi'(u(x))\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x \geq 0. \tag{13}$$

We remark that, for any function ξ which is such that $\xi'(1) \geq 0$ and ξ' is decreasing, we can get (13) from (12) for any function u which only verifies (10). For this reason, we now assume that ξ is convex (in the sense that ξ' is nondecreasing, this terminology is used in the sequel of this paper), and we develop equation (13), integrating by parts. We then derive the following weak sense for a solution to problem (10)–(11).

Definition 1.1 (weak solution to problem (10)–(11)). *Under Hypotheses* (H), *we say that a function $\tilde{g} \in L^\infty(\Omega)^d$ is a weak solution to problem* (10)–(11) *if there* *exists $u \in L^\infty(\Omega)$ such that $\tilde{g}(x) = u(x)g(x)$ for a.e. $x \in \Omega$, and u satisfies the following inequalities: $0 \le u(x) \le 1$ for a.e. $x \in \Omega$ and*

$$\begin{aligned}(14)\qquad \int_\Omega \big(\xi(u(x))(-g(x)\cdot\nabla\varphi(x)) + [\xi'(u(x))u(x) - \xi(u(x))]\varphi(x)\mathrm{div}g(x)\\ + \xi'(u(x))\varphi(x)F(x)\big)\mathrm{d}x \ge 0\\ \forall \xi \in C^1(\mathbb{R}) \textit{ convex such that } \xi'(1) \ge 0 \qquad \forall\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+).\end{aligned}$$

The following proposition expresses that any weak solution in the above sense satisfies (10) in a weak sense, and the next one shows that any regular weak solution satisfies (10)–(11), thus completing the justification of Definition 1.1.

PROPOSITION 1.2. *Under Hypotheses* (H), *let $\tilde{g} : \Omega \to \mathbb{R}^d$ be a weak solution to problem* (10)–(11) *in the sense of Definition* 1.1. *Then*

$$(15)\qquad \int_\Omega (-\tilde{g}(x)\cdot\nabla\varphi(x)\mathrm{d}x + F(x)\varphi(x))\mathrm{d}x \ge 0 \qquad \forall\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+).$$

Proof. Let us assume that $u \in L^\infty(\Omega)$ is such that $\tilde{g}(x) = u(x)g(x)$ and $0 \le u(x) \le 1$ for a.e. $x \in \Omega$, and (14) is satisfied. Let us take $\xi : s \mapsto s$ in (14). We then obtain (15). □

PROPOSITION 1.3. *Under Hypotheses* (H), *let $\tilde{g} : \Omega \to \mathbb{R}^d$ be a Lipschitz continuous function. Then $\tilde{g}$ is a weak solution to problem* (10)–(11) *in the sense of Definition* 1.1 *if and only if there exists a function $u \in L^\infty(\Omega)$ with $\tilde{g}(x) = u(x)g(x)$ and $0 \le u(x) \le 1$ for a.e. $x \in \Omega$ such that* (10) *and* (11) *are satisfied by the function u.*

Proof. Let us assume that $\tilde{g}$ is a weak solution to problem (10)–(11) in the sense of Definition 1.1. Then there exists $u \in L^\infty(\Omega)$ such that $\tilde{g}(x) = u(x)g(x)$ and $0 \le u(x) \le 1$ for a.e. $x \in \Omega$, and (14) is satisfied. Proposition 1.2 shows that (10) is satisfied by the function u for a.e. $x \in \Omega$. In order to prove that (11) is satisfied for a.e. $x \in \Omega$ by the function u, we shall separate the cases $x \in \Omega_0 := \{x \in \Omega, g(x) = 0\}$ and $x \in \Omega \setminus \Omega_0$. Let us take in (14) a test function φ whose support is included in the open set $\Omega \setminus \Omega_0$. Since the function u verifies $u(x) = |\tilde{g}(x)|/|g(x)|$ for a.e. $x \in \Omega \setminus \Omega_0$, u is Lipschitz continuous on the support of φ; we can thus integrate by parts, which produces, from (14), that (13) is satisfied by u. Let us now prove that u verifies (11).

Choosing $\xi : s \mapsto (s-1)^2$, we get that $\int_\Omega (u(x)-1)\varphi(x)(\mathrm{div}[u(x)g(x)] + F(x))\mathrm{d}x \ge 0$ holds. This implies that $(u(x) - 1)(\mathrm{div}[u(x)g(x)] + F(x)) \ge 0$ for a.e. $x \in \Omega$ such that $g(x) \ne 0$. But on one hand, $u(x) \le 1$ for a.e. $x \in \Omega$, and on the other hand, (10) is satisfied for a.e. $x \in \Omega$. Therefore, u verifies (11) for a.e. $x \in \Omega \setminus \Omega_0$.

Let us now obtain the same conclusion for a.e. $x \in \Omega_0$. Let $\eta \in C^1(\mathbb{R})$ be a function such that $0 \le \eta(x) \le 1$ for all $x \in \mathbb{R}$, $\eta(0) = 1$ and support$(\eta) \subset [-1, 1]$. For all $n \in \mathbb{N}^\star$, let us define the Lipschitz continuous function $\varphi_n : x \mapsto \eta(n|g(x)|)$. On one hand, we have that for a.e. $x \in \Omega_0$, $g(x)\cdot\nabla\varphi_n(x) = 0$ holds. On the other hand, for all $x \in \Omega \setminus \Omega_0$, we get that $g(x)\cdot\nabla\varphi_n(x)$ tends to 0 as $n \to \infty$ and remains bounded (indeed, it suffices to consider the cases $|g(x)| \le 1/n$ and $|g(x)| \ge 1/n$ and to use the property $\nabla g_i \in L^\infty(\Omega)^d$, where g_i, $i = 1, \dots, d$ are the components of g).

We then introduce $\xi : s \to (s-1)^2$ and $\varphi = \varphi_n$ in (14) (this is possible, taking regularizations in $C^1(\overline{\Omega}, \mathbb{R}_+)$ of φ_n). We then get

$$(16)\qquad T_1^{(n)} + T_2^{(n)} + T_3^{(n)} \ge 0,$$

with $T_1^{(n)} = \int_\Omega (u(x)-1)^2(-g(x)\cdot\nabla\varphi_n(x))\mathrm{d}x$, $T_2^{(n)} = \int_\Omega (u(x)^2-1)\varphi_n(x)\mathrm{div}g(x)\mathrm{d}x$, and $T_3^{(n)} = 2\int_\Omega (u(x)-1)F(x)\varphi_n(x)\mathrm{d}x$. Thus, thanks to the convergence properties of $g\cdot\nabla\varphi_n$ and to the dominated convergence theorem, we get that $T_1^{(n)}$ tends to 0 as $n\to\infty$.

Since $\varphi_n(x)$ tends to 0 for all $x\in\Omega\setminus\Omega_0$ and to 1 for all $x\in\Omega_0$, we get that $T_2^{(n)}$ tends to $\int_{\Omega_0}(u(x)^2-1)\mathrm{div}g(x)\mathrm{d}x$. Since $g(x)=0$ for all $x\in\Omega_0$, then $\partial_i g(x)=0$ for a.e. $x\in\Omega_0$ and all $i=1,\dots,d$ (this classical property has been shown, for example, in [17]), which produces $\int_{\Omega_0}(u(x)^2-1)\mathrm{div}g(x)\mathrm{d}x=0$.

We finally get that $T_3^{(n)}$ tends to $2\int_{\Omega_0}(u(x)-1)F(x)\mathrm{d}x$ as $n\to\infty$.

We thus get, passing to the limit $n\to\infty$ in (16), $\int_{\Omega_0}(u(x)-1)F(x)\mathrm{d}x\ge 0$, which proves that $u(x)=1$ for a.e. $x\in\Omega_0$ such that $F(x)>0$.

Therefore, for a.e. $x\in\Omega_0$, either $F(x)>0$ and $u(x)=1$, or $F(x)=0$ and $\mathrm{div}(\tilde g(x))+F(x)=0$, since $\tilde g(x)=0$ for a.e. $x\in\Omega_0$. Thus (11) is satisfied for a.e. $x\in\Omega_0$.

Reciprocally, let us assume that (10) and (11) are satisfied a.e. by the function u. We then get that (13) is satisfied, and therefore equation (14) is satisfied. This proves that $\tilde g$ is a weak solution to problem (10)–(11) in the sense of Definition 1.1. □

This paper is organized as follows. We first give, in section 2, the analytical expression of the weak solution in the one-dimensional case (the uniqueness result, proved in section 3, indeed holds in this case). In section 3, we first give a characterization of the set $\mathcal{C}(g,F)$ of functions which weakly satisfy (10). We prove a comparison result between a weak process solution to problem (10)–(11) (defined in Definition 3.3) and any element of $\mathcal{C}(g,F)$. This result suffices to prove the uniqueness of the weak solution to problem (10)–(11) in the sense of Definition 1.1. We then present a numerical scheme in section 4. The existence and uniqueness of a discrete solution is itself a nontrivial problem, which we solve by proving the convergence of an iterative method. This scheme is then proven to converge to a weak process solution to problem (10)–(11) in the sense of Definition 3.3. Thanks to the uniqueness result of the weak solution, we deduce the strong convergence result of the numerical scheme to this weak solution. We then give some numerical results in section 5 and conclude with some open problems.

2. Weak solutions in the one-dimensional case.

We have the following result.

PROPOSITION 2.1 (expression of the weak solution in the one-dimensional case). *Let $(a,b)\in\mathbb{R}^2$ be such that $a<b$, let $F\in L^\infty((a,b))$ be a nonnegative function, and let $g\in C^0([a,b])$ be a Lipschitz continuous function with $g(a)=g(b)=0$.*

Then, the function $\tilde g\ :[a,b]\to\mathbb{R}$ defined by

$$\tilde g(x) = \min_{y\in[x,b]}\left(g^+(y)+\int_x^y F(t)\mathrm{d}t\right) - \min_{y\in[a,x]}\left(g^-(y)+\int_y^x F(t)\mathrm{d}t\right) \quad \forall x\in[a,b], \tag{17}$$

where for all $s\in\mathbb{R}$ we denote $s^+=\max(s,0)$ and $s^-=\max(-s,0)$, is the unique weak solution to problem (10)–(11) *in the sense of Definition* 1.1.

Proof. Let us first remark that $\tilde g$ defined as such verifies that for all $x\in[a,b]$, $\tilde g^+(x)=\min_{y\in[x,b]}\left(g^+(y)+\int_x^y F(t)\mathrm{d}t\right)$ and $\tilde g^-(x)=\min_{y\in[a,x]}(g^-(y)+\int_y^x F(t)\mathrm{d}t)$ with $0\le\tilde g^+(x)\le g^+(x)$ and $0\le\tilde g^-(x)\le g^-(x)$. Then the function $\tilde g^+$ satisfies

$\tilde{g}^+(x) = \min_{y\in[a,b]} G_p(x,y)$ for all $x \in [a,b]$ with

$$G_p(x,y) = g^+(\max(x,y)) + \int_x^{\max(x,y)} F(t)\mathrm{d}t \qquad \forall (x,y) \in [a,b]^2.$$

Similarly, we have $\tilde{g}^-(x) = \min_{y\in[a,b]} G_m(x,y)$ for all $x \in [a,b]$ with

$$G_m(x,y) = g^-(\min(x,y)) + \int_{\min(x,y)}^{x} F(t)\mathrm{d}t \qquad \forall (x,y) \in [a,b]^2.$$

It is then clear that the functions G_p and G_m are Lipschitz continuous on $[a,b]^2$ with any Lipschitz constant M such that M is a bound of $F + |g'|$ in $L^\infty((a,b))$. Let $(x,\bar{x}) \in [a,b]^2$ be given, and let $(Y,\bar{Y}) \in [a,b]^2$ be such that $\tilde{g}^+(x) = G_p(x,Y)$ and $\tilde{g}^+(\bar{x}) = G_p(\bar{x},\bar{Y})$. Since we have

$$\tilde{g}^+(x) - \tilde{g}^+(\bar{x}) \le G_p(x,\bar{Y}) - G_p(\bar{x},\bar{Y}) \le M|x-\bar{x}|,$$

and, inverting the roles of x and $\bar{x}$,

$$\tilde{g}^+(\bar{x}) - \tilde{g}^+(x) \le G_p(x,Y) - G_p(\bar{x},Y) \le M|x-\bar{x}|,$$

we thus get that $\tilde{g}^+$ is Lipschitz continuous. Since the same proof holds for $\tilde{g}^-$, we thus get that $\tilde{g} = \tilde{g}^+ - \tilde{g}^-$ is Lipschitz continuous as well. We thus define the function $u\ :\ [a,b] \to [0,1]$ by $u(x) = 1$ for all $x \in \Omega$ such that $g(x) = 0$ and $u(x) = \tilde{g}(x)/g(x)$ for all $x \in [a,b]$ such that $g(x) \neq 0$. Let us prove that u satisfies (10)–(11) (from Proposition 1.3, since Hypotheses (H) are satisfied, this is sufficient to conclude). Since for all $x \in [a,b]$ such that $g(x) = 0$, $\tilde{g}(x) = 0$ holds, $\tilde{g}'(x) + F(x) \ge 0$ for a.e. $x \in [a,b]$ such that $g(x) = 0$ [17]. Let $x \in [a,b]$ be such that $g(x) > 0$. Then there exists $\alpha > 0$ such that $x+\alpha \le b$ and $g(y) > 0$ for all $y \in (x, x+\alpha)$. For $\bar{x} \in (x, x+\alpha)$, let $\bar{Y} \in [\bar{x}, b]$ be such that $\tilde{g}(\bar{x}) = G_p(\bar{x},\bar{Y})$. We have

$$\tilde{g}(x) - \tilde{g}(\bar{x}) \le G_p(x,\bar{Y}) - G_p(\bar{x},\bar{Y}) = \int_x^{\bar{x}} F(t)\mathrm{d}t. \tag{18}$$

The above inequality proves that $\tilde{g}'(x)+F(x) \ge 0$ for a.e. $x \in [a,b]$ such that $g(x) > 0$. Similarly, we obtain that $\tilde{g}'(x) + F(x) \ge 0$ for a.e. $x \in [a,b]$ such that $g(x) < 0$. This proves that (10) is satisfied. Let $x \in (a,b)$ such that $u(x) < 1$. Let us assume that $g(x) > 0$. Again, there exists $\alpha > 0$ such that $x + \alpha \le b$ and $g(y) > 0$ for all $y \in (x, x+\alpha)$, and again, for all $\bar{x} \in (x, x+\alpha)$, (18) holds. Since we have $0 \le \tilde{g}(x) < g(x)$, there exists $Y \in (x,b)$ such that $\tilde{g}(x) = G_p(x,Y)$. Therefore, for all $\bar{x} \in (x,Y)$, since $Y > \bar{x}$, we get

$$\tilde{g}(x) - \tilde{g}(\bar{x}) \ge G_p(x,Y) - G_p(\bar{x},Y) = \int_x^{\bar{x}} F(t)\mathrm{d}t.$$

Thus, for all $\bar{x} \in (x, \min(Y, x+\alpha))$, we get $\tilde{g}(x) - \tilde{g}(\bar{x}) = \int_x^{\bar{x}} F(t)\mathrm{d}t$, which implies that $\tilde{g}'(x) = -F(x)$ for a.e. $x \in \Omega$ such that $u(x) < 1$ and $g(x) > 0$. The case $u(x) < 1$ and $g(x) < 0$ can be similarly handled. Therefore $\tilde{g}$ is Lipschitz continuous and (10)–(11) are satisfied. Thanks to Proposition 1.3, this completes the proof that $\tilde{g}$ is a weak solution to problem (10)–(11) in the sense of Definition 1.1.

Since, within the hypotheses of the above proposition, Hypotheses (H) are satisfied (in particular, $g = h'$ with $h\ :\ x \mapsto \int_a^x g(t)\mathrm{d}t$), we can apply Proposition 3.5, which implies the uniqueness of the weak solution to problem (10)–(11) in the sense of Definition 1.1. □

Let us take two simple examples (one can find some examples inspired by geological problems in [6]). We consider a one-dimensional case (see Figure 1 below), with $\Omega = (-1, 1)$, $g : x \mapsto x^3 - x$, and $F : x \mapsto 1/2$. In this case, it is easy to verify that the function $\tilde{g}$ defined by (17) is such that $\tilde{g} = ug$, where the function u is such that $u : x \mapsto 1$ for all $x \in (-1, -\sqrt{1/2}) \cup (\sqrt{1/2}, 1)$ and $u : x \mapsto 1/(2(1 - x^2))$ for all $x \in (-\sqrt{1/2}, \sqrt{1/2})$. We thus obtain that the function u is continuous over Ω, but this is not always the case.

Indeed, let us consider the case $\Omega = (-1, 1)$, $g : x \mapsto x^3 - x$ for all $x \in [-1, 0]$, $g : x \mapsto \frac{1}{2}(x^3 - x)$ for all $x \in [0, 1]$, and $F : x \mapsto 1/2$. In such a case, g is only Lipschitz continuous, and the function $\tilde{g} = ug$ given by (17) is such that $u : x \mapsto 1$ for all $x \in (-1, -\sqrt{1/2}) \cup (0, 1)$ and $u : x \mapsto 1/(2(1 - x^2))$ for all $x \in (-\sqrt{1/2}, 0)$. This function u is therefore discontinuous in 0, although the function $\tilde{g} = ug$ remains Lipschitz continuous.

3. Uniqueness results.

3.1. Properties of the set of functions which satisfy (10). We consider in this section the set $\mathcal{C}(g, F)$ of functions which satisfy (10) in the sense of distributions. We shall prove below that the weak solution $\tilde{g}$ to problem (10)–(11) in the sense of Definition 1.1 is the projection of g in $L^2(\Omega)^d$ on $\mathcal{C}(g, F)$, and it is an extremal point of $\mathcal{C}(g, F)$ in the sense that $|\tilde{g}| \geq |\gamma|$ for all $\gamma \in \mathcal{C}(g, F)$ (see Proposition 3.5). The proof of this property is obtained thanks to the characterization of $\mathcal{C}(g, F)$ given by Proposition 3.2.

Definition 3.1 (the set $\mathcal{C}(g, F)$). *Under Hypotheses* (H), *we define the set $\mathcal{C}(g, F)$ of functions $\gamma \in L^2(\Omega)^d$ such that there exists $v \in L^\infty(\Omega)$, with $\gamma(x) = v(x)g(x)$ and $0 \leq v(x) \leq 1$, for a.e. $x \in \Omega$ and*

$$\int_{\mathbb{R}_+} \int_\Omega ([-\gamma(x) \cdot \nabla\varphi(x)] + \varphi(x)F(x))\,\mathrm{d}x \geq 0 \qquad \forall\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+). \tag{19}$$

Remark 3.1 (some properties of $\mathcal{C}(g, F)$). The set $\mathcal{C}(g, F)$ is nonempty (because $0 \in \mathcal{C}(g, F)$), convex (since the left-hand side of (19) is linear with respect to γ), and closed (in $L^2(\Omega)^d$).

Remark 3.2 (weak solutions and $\mathcal{C}(g, F)$). Thanks to Proposition 1.2, any weak solution to problem (10)–(11) in the sense of Definition 1.1 belongs to $\mathcal{C}(g, F)$.

We have the following proposition, which gives a characterization of the functions of $\mathcal{C}(g, F)$.

Proposition 3.2 (characterization of $\mathcal{C}(g, F)$). *Under Hypotheses* (H), *let $v \in L^\infty(\Omega)$, such that $0 \leq v(x) \leq 1$ for a.e. $x \in \Omega$, and let $\gamma(x) = v(x)g(x)$. Then $\gamma \in \mathcal{C}(g, F)$ (defined in Definition* 3.1*) holds if and only if the following property holds:*

$$\begin{aligned} &\int_\Omega \big(\xi(v(x))[-g(x) \cdot \nabla\varphi(x)] + [\xi'(v(x))v(x) - \xi(v(x))]\,\varphi(x)\ \mathrm{div} g(x) \\ &\qquad\qquad + \xi'(v(x))\varphi(x)F(x)\big)\mathrm{d}x \geq 0 \\ &\qquad \forall\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+),\ \forall\xi \in C^1(\mathbb{R}) \textit{ s.t. } \forall\kappa \in [0, 1],\ \xi'(\kappa) \geq 0. \end{aligned} \tag{20}$$

Proof. Under the hypotheses of the above proposition, let us assume that $\gamma \in \mathcal{C}(g, F)$. We introduce a sequence of mollifiers in $\mathbb{R}^d$. Let $\rho \in C_c^\infty(\mathbb{R}^d, \mathbb{R}_+)$ (the set of smooth functions with a compact support) be such that

$$\{x \in \mathbb{R}^d;\ \rho(x) \neq 0\} \subset \{x \in \mathbb{R}^d;\ |x| \leq 1\} \tag{21}$$

and

$$\int_{\mathbb{R}^d} \rho(x)\mathrm{d}x = 1. \tag{22}$$

For $n \in \mathbb{N}^\star$, we define

$$\rho_n(x) = n^d \rho(nx) \qquad \forall x \in \mathbb{R}^d. \tag{23}$$

We then define the functions $v_n(y) = \int_\Omega v(x)\rho_n(x-y)\mathrm{d}x$. Let $\psi \in C^1(\overline{\Omega}, \mathbb{R}_+)$ be given. For a given $y \in \Omega$, we introduce the function $\varphi \ :\ x \to \xi'(v_n(y))\psi(y)\rho_n(y-x) \in C^1(\overline{\Omega}, \mathbb{R}_+)$ in (19), and we integrate with respect to y. We thus get $T_4^{(n)} + T_5^{(n)} \geq 0$ with

$$T_4^{(n)} = -\int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)v(x)g(x) \cdot \nabla\rho_n(y-x)\mathrm{d}x\mathrm{d}y \tag{24}$$

and

$$T_5^{(n)} = \int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)F(x)\rho_n(y-x)\mathrm{d}x\mathrm{d}y. \tag{25}$$

The limit of the last term, as $n \to \infty$, satisfies

$$\lim_{n\to\infty} T_5^{(n)} = \int_\Omega \left(\xi'(v(y))\psi(y)F(y)\right)\mathrm{d}y.$$

We then turn to the study of $T_4^{(n)}$ as $n \to \infty$. We have $T_4^{(n)} = T_6^{(n)} + T_7^{(n)} + T_8^{(n)}$ with

$$T_6^{(n)} = \int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)v(x)g(y) \cdot \nabla\rho_n(y-x)\mathrm{d}x\mathrm{d}y,$$

$$T_7^{(n)} = \int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)v(y)(g(x)-g(y)) \cdot \nabla\rho_n(y-x)\mathrm{d}x\mathrm{d}y,$$

and

$$T_8^{(n)} = \int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)(v(x)-v(y))(g(x)-g(y)) \cdot \nabla\rho_n(y-x)\mathrm{d}x\mathrm{d}y.$$

We then have

$$T_6^{(n)} = \int_\Omega \xi'(v_n(y))\psi(y)g(y) \cdot \nabla v_n(y)\mathrm{d}y = \int_\Omega \psi(y)g(y) \cdot \nabla\xi(v_n)(y)\mathrm{d}y,$$

which delivers, thanks to an integration by parts with respect to y,

$$T_6^{(n)} = -\textstyle\int_\Omega \xi(v_n(y))\mathrm{div}[\psi(y)g(y)]\mathrm{d}y.$$

This leads to

$$\lim_{n\to\infty} T_6^{(n)} = \int_\Omega \xi(v(y))\mathrm{div}[\psi(y)g(y)]\mathrm{d}y.$$

We also have, thanks to an integration by parts with respect to x,

$$T_7^{(n)} = -\int_\Omega \int_\Omega \xi'(v_n(y))\psi(y)v(y)\rho_n(y-x)\mathrm{div}g(x)\mathrm{d}x\mathrm{d}y,$$

which produces

$$\lim_{n\to\infty} T_7^{(n)} = \int_\Omega \xi'(v(y))v(y)\psi(y)\mathrm{div}g(y)\mathrm{d}y.$$

Finally, we get

$$\lim_{n\to\infty} T_8^{(n)} = 0$$

thanks to the continuity in means of v and to the fact that $x \mapsto (g(x)-g(y))\cdot\nabla\rho_n(y-x)$ belongs to $L^1(\Omega)$. Then (20) is obtained by gathering all the results obtained above by passing to the limit $n \to \infty$.

Conversely, it suffices to choose the function $\xi \; : \; s \mapsto s$ in (20), for obtaining (19). □

3.2. Weak process solutions. Since we consider below the convergence of numerical schemes, on which the only estimate that we obtain in this case is an $L^\infty(\Omega)$ estimate, we must therefore consider weaker solutions than that defined in Definition 1.1, namely, weak process solutions. This notion of a weak process solution, introduced in [7], is related to the notion of Young measure first used by [3] in the nonlinear scalar hyperbolic framework. Young measures are extensively used in optimal control, nonconvex variational problems, phase transitions, microstructure problems, ... (see, e.g., [14], [16]).

The uniqueness result proven below leads to the uniqueness of such a weak process solution and to the fact that any weak process solution is indeed a weak solution. We then obtain the uniqueness of the weak solution to problem (10)–(11) in the sense of Definition 1.1. Moreover, this result is mainly used in the study of the numerical scheme in order to prove its strong convergence.

Definition 3.3 (weak process solutions to problem (10)–(11)). *Under Hypotheses* (H), *we say that a function $\hat{g}$ is a weak process solution to problem* (10)–(11) *if there exists $u \in L^\infty(\Omega\times(0,1))$ such that $\hat{g} \; : \; (x,\alpha) \mapsto u(x,\alpha)g(x)$ for a.e. $(x,\alpha) \in \Omega\times(0,1)$. And u satisfies the following inequalities: $0 \le u(x,\alpha) \le 1$ for a.e. $(x,\alpha) \in \Omega\times(0,1)$ and*

(26)
$$\int_\Omega\int_0^1 \big(\xi(u(x,\alpha))(-g(x)\cdot\nabla\varphi(x)) + [\xi'(u(x,\alpha))u(x,\alpha) - \xi(u(x,\alpha))]\varphi(x)\mathrm{div}g(x) \\ + \xi'(u(x,\alpha))\varphi(x)F(x)\big)\mathrm{d}\alpha\mathrm{d}x \ge 0 \\ \forall\xi\in C^1(\mathbb{R}),\ \text{convex s.t. } \xi'(1)\ge 0,\ \forall\varphi\in C^1(\overline{\Omega},\mathbb{R}_+).$$

We first prove the following property, which at the same time, gives some elements to conclude to the uniqueness of the weak process solution but also helps to prove that this solution is an extremal point of $\mathcal{C}(g,F)$.

Proposition 3.4 (comparison of a weak process solution and an element of $\mathcal{C}(g,F)$). *Under Hypotheses* (H), *let $\gamma \in \mathcal{C}(g,F)$ be given, where $\mathcal{C}(g,F)$ is defined* *in Definition* 3.1, *and let* $v \in L^\infty(\Omega)$, *such that* $\gamma(x) = v(x)g(x)$ *and* $0 \le v(x) \le 1$ *for a.e.* $x \in \Omega$. *Let* $\hat{g}$ *be a weak process solution to problem* (10)–(11) *in the sense of Definition* (3.3). *Let* $u \in L^\infty(\Omega\times(0,1))$ *be such that* $0 \le u(x,\alpha) \le 1$ *and* $\hat{g} \; : \; (x,\alpha) \mapsto u(x,\alpha)g(x)$ *for a.e.* $(x,\alpha) \in \Omega\times(0,1)$ *and such that* u *satisfies* (27). *Then the following inequality holds:*

$$\int_\Omega \int_0^1 (v(x) - u(x,\alpha))^+ \left[-g(x) \cdot \nabla\varphi(x)\right] \mathrm{d}\alpha \mathrm{d}x \ge 0 \qquad \forall \varphi \in C^1(\overline{\Omega}, \mathbb{R}_+). \tag{27}$$

Proof. This proof uses the method of doubling variables (first introduced by Krushkov [12]) adapted to weak process solutions [8].

Let us assume the hypotheses of the proposition. Let $\eta \in C^1(\mathbb{R}^2, \mathbb{R})$ be given such that $\eta(\cdot, b)$ is convex for all $b \in (-\infty, 1]$. We also assume that $\partial_1\eta$, the derivative of η with respect to its first argument, is such that $\partial_1\eta(1,b) \ge 0$ for all $b \in [0,1]$, and that $\partial_2\eta$, the derivative of η with respect to its second argument, is such that $\partial_2\eta(a,b) \ge 0$ for all $a, b \in [0,1]$. Let $\psi \in C^1(\mathbb{R}^d \times \mathbb{R}^d, \mathbb{R}_+)$ be given.

Then, for all $x \in \Omega$, we have $\psi(x,\cdot) \in C^1(\overline{\Omega}, \mathbb{R}_+)$ and for all $y \in \Omega$, $\psi(\cdot, y) \in C^1(\overline{\Omega}, \mathbb{R}_+)$. We introduce $\xi(\cdot) = \eta(\cdot, v(y))$ and $\varphi = \psi(\cdot, y)$ in (27) for $y \in \Omega$, and we integrate the result on Ω. This produces

$$\begin{aligned} &\int_\Omega\int_\Omega \int_0^1 \Big(\eta(u(x,\alpha), v(y)) \left[-g(x) \cdot \nabla_x\psi(x,y)\right] \\ &+ \left[\partial_1\eta(u(x,\alpha), v(y))u(x,\alpha) - \eta(u(x,\alpha), v(y))\right] \psi(x,y) \mathrm{div} g(x) \\ &\qquad + \partial_1\eta(u(x,\alpha), v(y))\psi(x,y)F(x)\Big) \mathrm{d}\alpha \mathrm{d}x \mathrm{d}y \ge 0. \end{aligned} \tag{28}$$

We now consider (20) for v, with $\xi(\cdot) = \eta(u(x,\alpha), \cdot)$ and $\varphi = \psi(x, \cdot)$, and we integrate the result on $\Omega \times (0,1)$. We thus get

$$\begin{aligned} &\int_\Omega\int_\Omega \int_0^1 \Big(\eta(u(x,\alpha), v(y)) \left[-g(y) \cdot \nabla_y\psi(x,y)\right] \\ &+ \left[\partial_2\eta(u(x,\alpha), v(y))v(y) - \eta(u(x,\alpha), v(y))\right] \psi(x,y) \mathrm{div} g(y) \\ &\qquad + \partial_2\eta(u(x,\alpha), v(y))\psi(x,y)F(y)\Big) \mathrm{d}\alpha \mathrm{d}x \mathrm{d}y \ge 0. \end{aligned} \tag{29}$$

We now add (28) and (29). This delivers

$$T_9 + T_{10} + T_{11} \ge 0, \tag{30}$$

where

$$\begin{aligned} T_9 = -\int_\Omega\int_\Omega \int_0^1 \quad &\eta(u(x,\alpha), v(y)) \\ &\times \big(g(x) \cdot \nabla_x\psi(x,y) + g(y) \cdot \nabla_y\psi(x,y)\big) \mathrm{d}\alpha \mathrm{d}x \mathrm{d}y, \end{aligned} \tag{31}$$

$$\begin{aligned} T_{10} = \int_\Omega\int_\Omega \int_0^1 &\Big(\big(\partial_1\eta(u(x,\alpha), v(y))u(x,\alpha) - \eta(u(x,\alpha), v(y))\big)\psi(x,y) \mathrm{div} g(x) \\ &+ \big(\partial_2\eta(u(x,\alpha), v(y))v(y,\beta) - \eta(u(x,\alpha), v(y))\big)\psi(x,y) \mathrm{div} g(y)\Big) \mathrm{d}\alpha \mathrm{d}x \mathrm{d}y, \end{aligned} \tag{32}$$

and

$$T_{11} = \int_\Omega\int_\Omega\int_0^1 \big(\partial_1\eta(u(x,\alpha),v(y))F(x) + \partial_2\eta(u(x,\alpha),v(y))F(y)\big)\psi(x,y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y. \tag{33}$$

We again use the sequence of mollifiers in $\mathbb{R}$ and $\mathbb{R}^d$, defined by (21)–(23). Let $\phi \in C^1(\mathbb{R}^d,\mathbb{R}_+)$ and $n \in \mathbb{N}^\star$ be given. We then take $\psi(x,y) = \phi(x)\rho_n(x-y)$ in (28) and (29), which gives $\psi \in C^1(\mathbb{R}^d\times\mathbb{R}^d,\mathbb{R}_+)$. We thus get, from (30),

$$T_9^{(n)} + T_{10}^{(n)} + T_{11}^{(n)} \geq 0, \tag{34}$$

with

$$\begin{aligned} T_9^{(n)} &= -\int_\Omega\int_\Omega\int_0^1 \eta(u(x,\alpha),v(y)) \\ &\quad\times\Big(\rho_n(x-y)g(x)\cdot\nabla\phi(x) + \phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\Big)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{35}$$

$$\begin{aligned} T_{10}^{(n)} = \int_\Omega\int_\Omega\int_0^1 \Big(&[\partial_1\eta(u(x,\alpha),v(y))u(x,\alpha) - \eta(u(x,\alpha),v(y))]\,\mathrm{div}g(x) \\ &+ [\partial_2\eta(u(x,\alpha),v(y))v(y) - \eta(u(x,\alpha),v(y))]\,\mathrm{div}g(y)\Big)\phi(x)\rho_n(x-y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{36}$$

$$\begin{aligned} T_{11}^{(n)} = \int_\Omega\int_\Omega\int_0^1 \quad &\big(\partial_1\eta(u(x,\alpha),v(y))F(x) + \partial_2\eta(u(x,\alpha),v(y))F(y)\big) \\ &\times\phi(x)\rho_n(x-y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{37}$$

We have $T_9^{(n)} = T_{12}^{(n)} + T_{13}^{(n)} + T_{14}^{(n)}$, with

$$T_{12}^{(n)} = -\int_\Omega\int_\Omega\int_0^1 \eta(u(x,\alpha),v(y))\rho_n(x-y)g(x)\cdot\nabla\phi(x)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \tag{38}$$

$$\begin{aligned} T_{13}^{(n)} = -\int_\Omega\int_\Omega\int_0^1 \quad &\eta(u(x,\alpha),v(x)) \\ &\times\phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y, \end{aligned} \tag{39}$$

$$\begin{aligned} T_{14}^{(n)} = -\int_\Omega\int_\Omega\int_0^1 \quad &\Big(\eta(u(x,\alpha),v(y)) - \eta(u(x,\alpha),v(x))\Big) \\ &\times\phi(x)(g(x)-g(y))\cdot\nabla\rho_n(x-y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y. \end{aligned} \tag{40}$$

The limit of $T_{12}^{(n)}$ as $n \longrightarrow \infty$ is given by

$$\lim_{n\to\infty} T_{12}^{(n)} = -\int_\Omega\int_0^1 \eta(u(x,\alpha),v(x))g(x)\cdot\nabla\phi(x)\mathrm{d}\alpha\mathrm{d}x.$$

Thanks to an integration by parts with respect to y and to Hypotheses (H), we get $T_{13}^{(n)} = T_{15}^{(n)} + T_{16}^{(n)}$, where

$$T_{15}^{(n)} = \int_\Omega\int_{\partial\Omega}\int_0^1 \eta(u(x,\alpha),v(x))\phi(x)\rho_n(x-y)g(x)\cdot\mathbf{n}(y)\mathrm{d}\alpha\mathrm{d}y\mathrm{d}x \tag{41}$$

and

$$T_{16}^{(n)} = \int_\Omega \int_\Omega \int_0^1 \eta(u(x,\alpha), v(x))\phi(x)\rho_n(x-y)\mathrm{div}g(y)\mathrm{d}\alpha\mathrm{d}x\mathrm{d}y. \tag{42}$$

We have, for a.e. $y \in \partial\Omega$,

$$\lim_{n\to\infty} \int_\Omega \int_0^1 \eta(u(x,\alpha), v(x))\phi(x)\rho_n(x-y)g(x)\cdot\mathbf{n}(y)\mathrm{d}\alpha\mathrm{d}x = 0,$$

which produces

$$\lim_{n\to\infty} T_{15}^{(n)} = 0,$$

and therefore

$$\lim_{n\to\infty} T_{13}^{(n)} = \lim_{n\to\infty} T_{16}^{(n)} = \int_\Omega \int_0^1 \eta(u(x,\alpha), v(x))\phi(x)\mathrm{div}g(x)\mathrm{d}\alpha\mathrm{d}x.$$

Thanks to the theorem of continuity in means applied to the function v and thanks to the fact that $(x,y) \mapsto (g(x) - g(y)) \cdot \nabla\rho_n(x-y)$ vanishes for $|x-y| > 1/n$ and belongs to $L^1(\Omega)$ since g is regular, we get

$$\lim_{n\to\infty} T_{14}^{(n)} = 0.$$

We have, again using the Lebesgue dominated convergence theorem,

$$\lim_{n\to\infty} T_{10}^{(n)} = \int_\Omega \int_0^1 \big(\partial_1\eta(u(x,\alpha), v(x))u(x,\alpha) + \partial_2\eta(u(x,\alpha), v(x))v(x) \\ - 2\eta(u(x,\alpha), v(x))\big)\phi(x)\mathrm{div}g(x)\mathrm{d}\alpha\mathrm{d}x$$

and

$$\lim_{n\to\infty} T_{11}^{(n)} = \int_\Omega \int_0^1 \left(\partial_1\eta(u(x,\alpha), v(x)) + \partial_2\eta(u(x,\alpha), v(x))\right) F(x)\phi(x)\mathrm{d}\alpha\mathrm{d}x.$$

We thus get, passing to the limit $n \to \infty$ in (34),

(43)

$$\int_\Omega \int_0^1 \Big(\eta(u(x,\alpha), v(x))\left[-g(x)\cdot\nabla\phi(x)\right] \\ + \big(\partial_1\eta(u(x,\alpha), v(x))u(x,\alpha) + \partial_2\eta(u(x,\alpha), v(y))v(x) - \eta(u(x,\alpha), v(x))\big)\phi(x)\mathrm{div}g(x) \\ + \big(\partial_1\eta(u(x,\alpha), v(x)) + \partial_2\eta(u(x,\alpha), v(x))\big)F(x)\phi(x)\Big)\mathrm{d}\alpha\mathrm{d}x \geq 0.$$

We now consider, for a given $\varepsilon > 0$, the function $S_\varepsilon \in C^1(\mathbb{R})$ defined by

$$\begin{array}{ll} S_\varepsilon(s) = 0 & \forall s \in (-\infty, 0], \\ S_\varepsilon(s) = s^2(3\varepsilon - 2s)/\varepsilon^3 & \forall s \in [0,\varepsilon], \\ S_\varepsilon(s) = 1 & \forall s \in [\varepsilon, +\infty). \end{array} \tag{44}$$

We define $\xi_\varepsilon(s) = \int_0^s S_\varepsilon(\tau)d\tau$, and we set, for all $(a,b) \in \mathbb{R}^2$, $\eta_\varepsilon(a,b) = \xi_\varepsilon(b-a)$. We then easily get that this function η_ε satisfies $\partial_1\eta_\varepsilon(1,b) = -S_\varepsilon(b-1) = 0 \geq 0$ for all

$b \le 1$, $\eta_\varepsilon(\cdot,b)$ is convex for all $b \le 1$, and $\partial_2\eta_\varepsilon(a,b) = S_\varepsilon(b-a) \ge 0$ for all $(a,b) \in \mathbb{R}^2$. We can then use this function in (44). We remark that, for all $(a,b) \in \mathbb{R}^2$,

$$a\partial_1\eta_\varepsilon(a,b) + b\partial_2\eta_\varepsilon(a,b) - \eta_\varepsilon(a,b) = (b-a)S_\varepsilon(b-a) - \eta_\varepsilon(a,b)$$

leads to

$$\lim_{\varepsilon\to 0}(a\partial_1\eta_\varepsilon(a,b) + b\partial_2\eta_\varepsilon(a,b) - \eta_\varepsilon(a,b)) = 0,$$

and we also remark that

$$\partial_1\eta_\varepsilon(a,b) + \partial_2\eta_\varepsilon(a,b) = 0.$$

Thus, using the Lebesgue dominated convergence theorem, we can let $\varepsilon \to 0$ in (44) which produces

$$\int_\Omega \int_0^1 (v(x) - u(x,\alpha))^+ \left[-g(x)\cdot\nabla\phi(x)\right] \mathrm{d}\alpha \mathrm{d}x \ge 0, \tag{45}$$

which is (27) and thus concludes the proof of the proposition. □

The above result is now used to yield the uniqueness of the weak process solution, and thus to obtain that this weak process solution is in fact a weak solution. Note that, in the proof of all the above propositions, the hypothesis that g can be written under the form $g(x) = \Lambda(x)\nabla h(x)$ for all $x \in \Omega$ is not used (g being Lipschitz continuous is sufficient). A uniqueness result for the weak solution could then be obtained assuming that $F > 0$ a.e. in addition to g being Lipschitz continuous, but the uniqueness result for the weak process solution remains an open problem under such hypotheses. The proof of the uniqueness result given below explicitly uses the hypothesis $g(x) = \Lambda(x)\nabla h(x)$ for all $x \in \Omega$, which fortunately holds in the physical problem.

Proposition 3.5 (uniqueness of the weak process solution). *Under Hypotheses* (H), *there exists at most one weak process solution $\hat g$ to problem* (10)–(11) *in the sense of Definition* 3.3. *Moreover, if $\hat u \in L^\infty(\Omega\times(0,1))$ is such that* $0 \le u(x,\alpha) \le 1$ *and $\hat g : (x,\alpha) \mapsto u(x,\alpha)g(x)$ for a.e. $(x,\alpha) \in \Omega\times(0,1)$ and if u satisfies* (27), *then $u(x,\alpha)$ does not depend on α on a.e. $x \in \Omega$ such that $g(x) \ne 0$ ($g(x) = 0$ and $F(x) > 0$). Then the function $\tilde g$ defined by $\tilde g(x) = u(x,\alpha)g(x)$ for a.e. $x \in \Omega$ and $\alpha \in (0,1)$ is the unique weak solution to problem* (10)–(11) *in the sense of Definition* 1.1. *Moreover, this function $\tilde g$ is an extremal point of $\mathcal{C}(g,F)$ in the sense that $|\tilde g| \ge |\gamma|$ for all $\gamma \in \mathcal{C}(g,F)$ (the set $\mathcal{C}(g,F)$ is defined in Definition* 3.1*), and it is also the projection in $L^2(\Omega)^d$ of g on the convex set $\mathcal{C}(g,F)$.*

Proof. Let us assume that $\hat g$ is a weak process solution to problem (10)–(11) in the sense of Definition 3.3. Let $u \in L^\infty(\Omega\times(0,1))$ correspond to $\hat g$ in Definition 3.3. We again denote $\Omega_0 = \{x \in \Omega,\ g(x) = 0\}$ and we remark that (27), proven in Proposition 3.4, gives for all $\gamma \in \mathcal{C}(g,F)$, letting $v \in L^\infty(\Omega)$ be such that $\gamma(x) = v(x)g(x)$ and $0 \le v(x) \le 1$ for a.e. $x \in \Omega$, that

$$\int_{\Omega\setminus\Omega_0} \int_0^1 (v(x) - u(x,\alpha))^+ \left[-g(x)\cdot\nabla\varphi(x)\right] \mathrm{d}\alpha \mathrm{d}x \ge 0.$$

Thanks to Hypotheses (H), we can define the nonnegative function φ by $\varphi(x) = h(x) - \min_{y\in\Omega} h(y)$ for all $x \in \Omega$, where $h \in C^1(\overline{\Omega})$ is such that $g(x) = \Lambda(x)\nabla h(x)$ for all

$x \in \Omega$. We then get that, for all $x \in \Omega \setminus \Omega_0$, $-g(x) \cdot \nabla\varphi(x) = -\Lambda(x)\nabla h(x).\nabla h(x) < 0$. This produces

$$(v(x) - u(x,\alpha))^+ = 0 \text{ for a.e. } (x,\alpha) \in \Omega \setminus \Omega_0 \times (0,1). \tag{46}$$

We then remark that the function $\gamma \ : \ x \mapsto \int_0^1 u(x,\alpha)\mathrm{d}\alpha g(x)$ belongs to the convex set $\mathcal{C}(g,F)$. Therefore, setting $v = \int_0^1 u(\cdot,\alpha)\mathrm{d}\alpha$ in (46), we get that for a.e. $x \in \Omega \setminus \Omega_0$, $\int_0^1 (\int_0^1 u(x,\beta)d\beta - u(x,\alpha))^+\mathrm{d}\alpha = 0$, which proves that $u(x,\alpha)$ does not depend on α for a.e. $x \in \Omega \setminus \Omega_0$. We define $T(u) \in L^\infty(\Omega)$ by $T(u)(x) = u(x,\alpha)$ for a.e. $x \in \Omega \setminus \Omega_0$ and $\alpha \in (0,1)$ and by $T(u)(x) = 1$ for a.e. $x \in \Omega_0$. We then get that the function $\tilde{g} \ : \ \Omega \to \mathbb{R}^d$ such that $\tilde{g} = T(u)g$ is such that $\tilde{g}(x) = \hat{g}(x,\alpha)$ for a.e. $x \in \Omega$ and $\alpha \in (0,1)$.

Let us assume that $\hat{g}$ and $\hat{\hat{g}}$ are two weak process solutions to problem (10)–(11) in the sense of Definition 3.3. Let u and $\hat{u}$ be some elements of $L^\infty(\Omega \times (0,1))$ which correspond to $\hat{g}$ and $\hat{\hat{g}}$, respectively, in Definition 3.3. We then get, setting $v = T(\hat{u})$ in (46), that $(T(\hat{u})(x) - T(u)(x))^+ = 0$ for a.e. $x \in \Omega \setminus \Omega_0$ and, inverting the roles of u and $\hat{u}$, $(T(u)(x) - T(\hat{u})(x))^+ = 0$. This suffices to prove that $T(\hat{u})(x) = T(u)(x)$ for a.e. $x \in \Omega \setminus \Omega_0$, which completes the proof of uniqueness of the weak process solution.

Let us prove that the function $\tilde{g} = T(u)g$ is a weak solution to problem (10)-(11) in the sense of Definition 1.1. We introduce in (27) the functions $\xi \ : \ s \to (s-1)^2$ and, for all $n \in \mathbb{N}^\star$, $\varphi = \varphi_n$, as defined in the proof of Proposition 1.3. The same analysis as that which is done in the proof of Proposition 1.3 delivers that, passing to the limit $n \to \infty$, $\int_{\Omega_0} \int_0^1 (u(x,\alpha) - 1)\mathrm{d}\alpha F(x)\mathrm{d}x \geq 0$. This proves that $u(x,\alpha) = 1 = u(x)$ for a.e. $\alpha \in (0,1)$ and a.e. $x \in \Omega_0$ such that $F(x) > 0$. Since all the terms of (27) under the symbols $\int$ vanish a.e. on $\{x \in \Omega,\ g(x) = 0 \text{ and } F(x) = 0\}$, we get that (27) with u implies (14) with $T(u)$. Thus the function $\tilde{g}$ is a weak solution to problem (10)–(11) in the sense of Definition 1.1. Since it is obvious that any weak solution is a weak process solution, we thus deduce, from the uniqueness of the weak process solution, that of this weak solution.

Let us now show that $\tilde{g}$ is an extremal point of $\mathcal{C}(g,F)$. Let $\gamma \in \mathcal{C}(g,F)$, and let $v \in L^\infty(\Omega)$ such that $\gamma(x) = v(x)g(x)$ and $0 \leq v(x) \leq 1$ for a.e. $x \in \Omega$. Thanks to (46), we get that, for a.e. $x \in \Omega \setminus \Omega_0$, $v(x) \leq T(u)(x)$. This proves that, for a.e. $x \in \Omega$, $|\gamma(x)| \leq |\tilde{g}(x)|$. This property implies that $\int_\Omega (g(x) - \tilde{g}(x)) \cdot (\tilde{g}(x) - \gamma(x))\mathrm{d}x = \int_\Omega |g(x)|^2 (1 - T(u)(x))(T(u)(x) - v(x))\mathrm{d}x \geq 0$ for all $\gamma \in \mathcal{C}(g,F)$, which shows that $\tilde{g}$ is the projection of g on $\mathcal{C}(g,F)$ in $L^2(\Omega)^d$. □

4. Passing to the limit in numerical schemes. We now start the study of the convergence of a numerical scheme, which is based on finite volume methods. Such methods proved their efficiency for various nonlinear problems such as, for instance, nonlinear degenerate problems (see, e.g., [9], [13] and references therein) and nonlinear hyperbolic problems (see [8], but there exists a huge literature on this subject). The main additional difficulty of the present problem is due to the introduction of the limiter $\overline{u}$ in (2) in order to satisfy the constraints (3)–(5) (recall that (2)–(5) lead to (10) and (11) using a time discretization). The "equation" on this unknown $\overline{u}$ seems to lead to a new type of problem which is unexpectedly not really related to variational inequalities but has some similarity with a scalar conservation law, leading to a nonlinear hyperbolic equation. From the numerical point of view, this similarity may be viewed in the upwinding choice for u in (50) (more precisely, the choice of u_K or u_L, on the interface between the control volumes K and L, depends on the sign of $g_{K,L}$). This upwinding is crucial, for instance, in order to have a solution

u taking values in $[0,1]$ (which is a constraint given by (10)). Numerical simulations using a centered choice of u often lead to troubles (such as oscillations) and the simulation has to stop (this is also true in the industrial framework). Another similarity with scalar conservation laws appears in the choice of the convex function ξ in Definition 1.1.

Let us first define the notion of admissible mesh of $\mathbb{R}^d$ (this definition is inspired by [8]).

DEFINITION 4.1 (admissible meshes). *An admissible finite volume mesh of Ω, denoted by $\mathcal{T}$, is given by a finite family of disjoint polygonal (one uses here the two space dimensions terms for the setting of the general space dimension) connected subsets of $\mathbb{R}^d$ such that Ω is the union of the closure of the elements of $\mathcal{T}$ (which are called control volumes in the following) and such that the common interface of any pair of neighboring control volumes is included in a hyperplane of $\mathbb{R}^d$ (this is not necessary but is introduced in order to simplify the formulation). We denote by* $\text{size}(\mathcal{T}) := \sup\{\text{diam}(K), K \in \mathcal{T}\}$*, by m_K the measure of K for all $K \in \mathcal{T}$, and by $\mathcal{N}_K$ the subset of $\mathcal{T}$ of all the control volumes having a common interface with K. We then denote by $\mathcal{E}$ one set of pairs of neighbors $(K,L) \in \mathcal{T}^2$ such that, if $(K,L) \in \mathcal{E}$, $(L,K) \notin \mathcal{E}$, and for all $K \in \mathcal{T}$ and $L \in \mathcal{N}_K$, $(K,L) \in \mathcal{E}$ or $(L,K) \in \mathcal{E}$. For $K \in \mathcal{T}$ and $L \in \mathcal{N}_K$, we denote by m_{KL} the measure of the common interface between K and L. We measure the regularity of the mesh by means of the following expression:*

$$\text{regul}(\mathcal{T}) := \max\left\{\sum_{L \in \mathcal{N}_K} m_{KL}\text{diam}(K)/m_K,\ K \in \mathcal{T}\right\}.$$

Let $\mathcal{T}$ be an admissible mesh of Ω. Let $g_{\mathcal{T}} := (g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real numbers such that

$$g_{K,L} = -g_{L,K} \qquad \forall K \in \mathcal{T},\ \forall L \in \mathcal{N}_K \tag{47}$$

and

$$\sum_{L\in\mathcal{N}_K} g_{K,L} = \int_K \text{div}g(x)\text{d}x := G_K \qquad \forall K \in \mathcal{T}. \tag{48}$$

Denoting

$$F_K = \int_K F(x)\text{d}x, \tag{49}$$

the finite volume scheme, in order to approximate problem (10)–(11), is given by

$$\begin{aligned} &\sum_{L\in\mathcal{N}_K} (g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K = 0 \text{ and } u_K \le 1 \text{ or} \\ &\sum_{L\in\mathcal{N}_K} (g^+_{K,L}u_L - g^-_{K,L}u_K) + F_K \ge 0 \text{ and } u_K = 1. \end{aligned} \tag{50}$$

We define the function $u_{\mathcal{T}}$ by

$$u_{\mathcal{T}}(x) = u_K \qquad \forall x \in K,\ \forall K \in \mathcal{T}. \tag{51}$$

We then define the following value, which measures the consistency of the approximation $g_{\mathcal{T}}$ of the fluxes by means of a discrete $L^2(\Omega)^d$ norm and which is expected to tend to 0 with size($\mathcal{T}$):

$$\text{cons}(g_{\mathcal{T}}) := \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}_K} \frac{\text{diam}(K)}{m_{KL}} \left(g_{K,L} - \bar{g}_{K,L}\right)^2, \tag{52}$$

where

$$\bar{g}_{K,L} = \int_{K|L} g(x)\cdot \mathbf{n}_{K,L} \mathrm{d}s(x) \qquad \forall K \in \mathcal{T},\ \forall L \in \mathcal{N}_K. \tag{53}$$

Different choices are possible for $g_{\mathcal{T}}$. We can propose the following, for example.

- The choice $g_{K,L} = \bar{g}_{K,L}$ for all $K \in \mathcal{T}$ and $L \in \mathcal{N}_K$ is the simplest one which satisfies that cons($g_{\mathcal{T}}$) tends to 0 as size($\mathcal{T}$) tends to 0. Unfortunately, it demands in the general case the knowledge of g.
- In the framework of the coupled problem given in the introduction to this paper, the field $g = \Lambda\nabla h$ is not analytically known, and it must be approximated. This can be achieved, assuming that Λ is scalar (this is the case in some of the geological applications), using for example the finite volume method (see [8]). The notion of admissible meshes must then be restricted to the case where there exists, for all $K \in \mathcal{T}$, a point x_K in the control volume K such that, for a pair of two neighboring grid blocks K and L, the line (x_K, x_L) is orthogonal to the interface $\bar{K}\cap\bar{L}$ between these grid blocks. One then defines $\tau_{KL} = \int_{\bar{K}\cap\bar{L}} \Lambda(x)\mathrm{d}s(x)/d(x_K, x_L)$, where we denote by $\mathrm{d}s(x)$ the $d-1$ Lebesgue measure at point $x \in \bar{K}\cap\bar{L}$. One can then compute the family $(h_K)_{K\in\mathcal{T}}$ of reals such that (48) holds under the condition

  $$g_{K,L} = \tau_{KL}(h_L - h_K) \qquad \forall K \in \mathcal{T},\ \forall L \in \mathcal{N}_K \tag{54}$$

  in addition to such a relation as $\sum_{K\in\mathcal{T}} m_K h_K = 0$ (this corresponds to the discrete solution of a homogeneous Neumann problem). One can then prove that, under Hypotheses (H), cons($g_{\mathcal{T}}$) tends to 0 as size($\mathcal{T}$) tends to 0 (see [8] and [18]).
- In the same way, one can compute a mixed finite element approximate for $g_{K,L}$ which also satisfies that cons($g_{\mathcal{T}}$) tends to 0 as size($\mathcal{T}$) tends to 0 (see [4]).

In order to compute a solution of (47)–(50), we shall now describe an algorithm, denoted by Algorithm (A) below.

Algorithm (A).

Initialization: $u_K^{(0)} = 1$ and $p_K^{(0)} = 1$ for all $K \in \mathcal{T}$.

Iterations: Let $n \in \mathbb{N}^\star$. Assume that $u_K^{(n-1)}$ and $p_K^{(n-1)}$ are known for all $K \in \mathcal{T}$.

1. Computation of $\{p_K^{(n)},\ K \in \mathcal{T}\}$:

$$\begin{aligned} &\text{If } \sum_{L\in\mathcal{N}_K} (g_{K,L}^{+} u_L^{(n-1)} - g_{K,L}^{-} u_K^{(n-1)}) + F_K < 0, \text{ then } p_K^{(n)} = 0.\\ &\text{If } \sum_{L\in\mathcal{N}_K} (g_{K,L}^{+} u_L^{(n-1)} - g_{K,L}^{-} u_K^{(n-1)}) + F_K \geq 0, \text{ then } p_K^{(n)} = p_K^{(n-1)}. \end{aligned} \tag{55}$$

2. Computation of $\{u_K^{(n)},\ K \in \mathcal{T}\}$, solution to the following linear system:

$$\begin{aligned} \sum_{L\in\mathcal{N}_K} (g_{K,L}^{+} u_L^{(n)} - g_{K,L}^{-} u_K^{(n)}) &= -F_K \qquad \forall K \in \mathcal{T} \text{ s.t. } p_K^{(n)} = 0,\\ u_K^{(n)} &= 1 \qquad \forall K \in \mathcal{T} \text{ s.t. } p_K^{(n)} = 1. \end{aligned} \tag{56}$$

The following proposition gives a monotonicity property of Algorithm (A).

PROPOSITION 4.2 (a monotonicity property of Algorithm (A)). *Under Hypotheses* (H), *let $\mathcal{T}$ be an admissible mesh of Ω, and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real numbers such that* (47) *and* (48) *are satisfied. Let $n \in \mathbb{N}^\star$ be given such that there exists a family $\{(p_K^{(k)}, u_K^{(k)})$, $K \in \mathcal{T}$, $k = 0, \dots, n-1\}$ such that* (55) *and* (56) *hold in addition to $u_K^{(k)} \geq 0$ for all $K \in \mathcal{T}$, $k = 0, \dots, n-1$. Let $(p_K^{(n)})_{K\in\mathcal{T}}$ be given by* (55).

Then, for all family of reals $(w_K, s_K)_{K\in\mathcal{T}}$ such that $s_K \geq 0$ for all $K \in \mathcal{T}$ and such that

$$(57)\qquad \begin{aligned} \sum_{L\in\mathcal{N}_K} (g_{K,L}^{+} w_L - g_{K,L}^{-} w_K) &= -s_K \qquad \forall K \in \mathcal{T} \text{ s.t. } p_K^{(n)} = 0,\\ w_K &= s_K \qquad \forall K \in \mathcal{T} \text{ s.t. } p_K^{(n)} = 1, \end{aligned}$$

the property $w_K \geq 0$ for all $K \in \mathcal{T}$ holds.

Let us first remark that Proposition 4.2 suffices to prove that the matrix of the linear system (57) is invertible. Since in the case $s_K = 0$ for all $K \in \mathcal{T}$, for any family $(w_K)_{K\in\mathcal{T}}$ satisfying (57), then $(-w_K)_{K\in\mathcal{T}}$ also satisfies (57), which proves that $w_K = 0$ for all $K \in \mathcal{T}$. We therefore state the following corollary.

COROLLARY 4.3. *Under the hypotheses of Proposition* 4.2, *for all families $(s_K)_{K\in\mathcal{T}}$ of reals, there exists one and only one family of reals $(w_K)_{K\in\mathcal{T}}$ such that* (57) *holds.*

Proof of Proposition 4.2. Let us assume the hypotheses of Proposition 4.2, and let $(w_K, s_K)_{K\in\mathcal{T}}$ be a family of reals such that $s_K \geq 0$ for all $K \in \mathcal{T}$ and such that (57) holds. Let us assume that the set $\mathcal{T}_- = \{K \in \mathcal{T};\, w_K < 0\}$ is not empty. Then, if $K \in \mathcal{T}_-$, one has $p_K^{(n)} = 0$, since $w_K = s_K \geq 0$ for $K \in \mathcal{T}$ such that $p_K^{(n)} = 1$. We therefore have

$$(58)\qquad \sum_{L\in\mathcal{N}_K} (g_{K,L}^{+} w_L - g_{K,L}^{-} w_K) + s_K = 0 \qquad \forall K \in \mathcal{T}_-.$$

Summing (58) for $K \in \mathcal{T}_-$ leads to

$$(59)\qquad \sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K\setminus\mathcal{T}_-} (g_{K,L}^{+} w_L - g_{K,L}^{-} w_K) + \sum_{K\in\mathcal{T}_-} s_K = 0.$$

Since $w_K < 0$ for $K \in \mathcal{T}_-$ and $w_L \geq 0$ for $L \notin \mathcal{T}_-$, (59) gives $s_K = 0$ for all $K \in \mathcal{T}_-$ and $g_{K,L} \geq 0$ for all (K, L) such that $K \in \mathcal{T}_-$ and $L \in \mathcal{N}_K \setminus \mathcal{T}_-$. Let $k < n$ be the greatest integer such that there exists $K \in \mathcal{T}_-$ with $p_K^{(k)} = 1$ and $p_K^{(k+1)} = 0$ (such a k exists since $p_K^{(0)} = 1$ for all $K \in \mathcal{T}$). We then have, for all $K \in \mathcal{T}_-$, $p_K^{(k+1)} = 0$ (otherwise this would be in contradiction with the choice of k), and therefore one has $\sum_{L\in\mathcal{N}_K} (g_{K,L}^{+} u_L^{(k)} - g_{K,L}^{-} u_K^{(k)}) + F_K \leq 0$.

For $K \in \mathcal{T}_-$ such that $p_K^{(k)} = 1$ and $p_K^{(k+1)} = 0$, one has $\sum_{L\in\mathcal{N}_K} (g_{K,L}^{+} u_L^{(k)} - g_{K,L}^{-} u_K^{(k)}) + F_K < 0$. We thus get

$$\sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K,\, L\notin\mathcal{T}_-} (g_{K,L}^{+} u_L^{(k)} - g_{K,L}^{-} u_K^{(k)}) + \sum_{K\in\mathcal{T}_-} F_K < 0.$$

On the other hand, since $u_L^{(k)} \geq 0$ and since $g_{K,L} \geq 0$ for all (K, L) such that $K \in \mathcal{T}_-$

and $L \in \mathcal{N}_K \setminus \mathcal{T}_-$ and $F_K \geq 0$, we can write

$$\begin{aligned} 0 &\leq \sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K,\, L\notin\mathcal{T}_-} g^+_{K,L} u^{(k)}_L \\ &\leq \sum_{K\in\mathcal{T}_-} \sum_{L\in\mathcal{N}_K,\, L\notin\mathcal{T}_-} (g^+_{K,L} u^{(k)}_L - g^-_{K,L} u^{(k)}_K) + \sum_{K\in\mathcal{T}_-} F_K < 0, \end{aligned}$$

which is impossible. This contradiction proves that $\mathcal{T}_-$ is empty, which concludes the proof of the proposition. □

We can now prove the following proposition, which states that Algorithm (A) is well defined and leads to a solution of (50) for some $n \leq \operatorname{card}(\mathcal{T})$.

Proposition 4.4 (convergence of an algorithm and existence of a discrete solution). *Under Hypotheses* (H), *let $\mathcal{T}$ be an admissible mesh of Ω, and let $(g_{K,L})_{K\in\mathcal{T},L\in\mathcal{N}_K}$ be a family of real numbers such that* (47) *and* (48) *are satisfied. Then the following hold.*

1. *There exists a unique family $\{(p^{(n)}_K, u^{(n)}_K),\ K \in \mathcal{T},\ n \in \mathbb{N}\}$ solution of Algorithm* (A).
2. *For all $K \in \mathcal{T}$ and all $n \in \mathbb{N}$, one has $u^{(n)}_K \geq 0$.*
3. *For all $K \in \mathcal{T}$, the sequence $(u^{(n)}_K)_{n\in\mathbb{N}}$ is nonincreasing.*
4. *There exists $n \leq$ card($\mathcal{T}$) such that, setting $u_K = u^{(n)}_K$ for all $K \in \mathcal{T}$, the family $\{u_K,\ K \in \mathcal{T}\}$ is such that $u^{(p)}_K = u_K$ for all $K \in \mathcal{T}$ and $p \geq n$. This family is therefore a solution of* (49) *and* (50) *such that*

$$0 \leq u_K \leq 1 \qquad \forall K \in \mathcal{T}. \tag{60}$$

Proof. The family $\{(p^{(0)}_K, u^{(0)}_K),\ K \in \mathcal{T}\}$ is uniquely defined and satisfies $u^{(0)}_K \geq 0$ for all $K \in \mathcal{T}$.

Let us prove the first two items of the above proposition by induction. Let $n \in \mathbb{N}^\star$; we assume that there exists a family $\{(p^{(k)}_K, u^{(k)}_K),\ K \in \mathcal{T},\ k = 0, \ldots, n-1\}$ such that (55) and (56) hold in addition to $u^{(k)}_K \geq 0$ for all $K \in \mathcal{T}$, $k = 0, \ldots, n-1$. Let $(p^{(n)}_K)_{K\in\mathcal{T}}$ be given by (55). We can then apply Proposition 4.2 and Corollary 4.3, setting $s_K = 1$ for all $K \in \mathcal{T}$ such that $p^{(n)}_K = 1$ and $s_K = F_K \geq 0$ for all $K \in \mathcal{T}$ such that $p^{(n)}_K = 0$. We thus immediately get the existence and the uniqueness of $u^{(n)}_K \geq 0$ for all $K \in \mathcal{T}$ such that (56) holds. This suffices to prove the first two items at the level n.

We can now prove that $u^{(n)}_K \leq u^{(n-1)}_K$ for all $K \in \mathcal{T}$. Indeed let us consider $w_K = u^{(n-1)}_K - u^{(n)}_K$ for all $K \in \mathcal{T}$. We have, for all $K \in \mathcal{T}$ such that $p^{(n)}_K = 0$, $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u^{(n)}_L - g^-_{K,L}u^{(n)}_K) + F_K = 0$ and $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u^{(n-1)}_L - g^-_{K,L}u^{(n-1)}_K) + F_K \leq 0$, which gives, by subtraction

$$\sum_{L\in\mathcal{N}_K} (g^+_{K,L} w_L - g^-_{K,L} w_K) := -s_K,$$

with $s_K \geq 0$. For all $K \in \mathcal{T}$ such that $p^{(n)}_K = 1$, we have

$$w_K = s_K := 0.$$

We can then apply Proposition 4.2, and we get that $0 \leq w_K$ for all $K \in \mathcal{T}$, which is the third item of the proposition. Let us prove the last item.

The definition of the algorithm gives $p_K^{(n)} = p_K^{(n-1)}$ or $p_K^{(n)} = 0$ for all K and all $n \in \mathbb{N}^\star$. Then, setting $A_n = \{K \in \mathcal{T};\, p_K^{(n)} = 0\}$, one has $\mathrm{card}(A_n) \geq \mathrm{card}(A_{n-1})$ for all $n \in \mathbb{N}^\star$. Since $\mathrm{card}(A_0)$ =0, there exists $n \leq \mathrm{card}(\mathcal{T}) + 1$ such that $\mathrm{card}(A_n)$ $= \mathrm{card}(A_{n-1})$. For this value of n one has $p_K^{(n)} = p_K^{(n-1)}$ for all $K \in \mathcal{T}$. If $p_K^{(n-1)} = 1$, one has $u_K^{(n-1)} = 1$ and $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L^{(n-1)} - g^-_{K,L}u_K^{(n-1)}) + F_K \geq 0$ (since $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L^{(n-1)} - g^-_{K,L}u_K^{(n-1)}) + F_K < 0$ gives $p_K^{(n)} = 0$).

If $p_K^{(n-1)} = 0$, one has $\sum_{L\in\mathcal{N}_K}(g^+_{K,L}u_L^{(n-1)} - g^-_{K,L}u_K^{(n-1)}) + F_K = 0$ and $u_K^{(n-1)} \leq 1$ thanks to the fact that the sequence $(u_K^{(n)})_{n\in\mathbb{N}}$ is nonincreasing and $u_K^{(0)} = 1$.

Therefore, setting $u_K = u_K^{(n-1)}$ for all $K \in \mathcal{T}$, the family $\{u_K,\ K \in \mathcal{T}\}$ is a solution of (49) and (50). It is also obvious to see that $u_K^{(p)} = u_K$ for all $K \in \mathcal{T}$ and for all $p \geq n - 1$.

This concludes the proof of Proposition 4.4. □

Remark 4.1. Under Hypotheses (H), assuming that Λ is a scalar function and following a method similar to the proof of uniqueness of Proposition 3.5, it is possible to prove that there exists a unique solution to (49) and (50), with the choice (54) for the discrete fluxes.

We then have the following proposition.

Proposition 4.5 (weak bounded variation inequality). *Under Hypotheses* (H), *let $\mathcal{T}$ be an admissible mesh of Ω in the sense of Definition* 4.1, *and let $g_\mathcal{T}$ be a family of reals which satisfies* (47) and (48). *Let $(u_K)_{K\in\mathcal{T}}$ be a solution of scheme* (49) and (50) *such that* (60) *holds. Then there exists $C > 0$, which only depends on d, Ω, g, F and not on $\mathcal{T}$, such that*

$$\sum_{(K,L)\in\mathcal{E}} |g_{K,L}|(u_K - u_L)^2 \leq C. \tag{61}$$

Proof. We multiply (50) by $(1 - u_K)$; we sum on K. We get $T_{17} + T_{18} = 0$ with

$$T_{17} = \sum_{K\in\mathcal{T}} (1 - u_K) \sum_{L\in\mathcal{N}_K} (g^+_{K,L}u_L - g^-_{K,L}u_K)$$

and

$$T_{18} = \sum_{K\in\mathcal{T}} (1 - u_K)F_K.$$

We have $T_{17} = T_{19} + T_{20}$, with

$$T_{19} = \sum_{K\in\mathcal{T}} (1 - u_K) \sum_{L\in\mathcal{N}_K} g^+_{K,L}(u_L - u_K)$$

and, using (48),

$$T_{20} = \sum_{K\in\mathcal{T}} (1 - u_K)u_K G_K.$$

We develop T_{19}: we get

$$\begin{aligned} T_{19} = \ & \frac{1}{2}\sum_{K\in\mathcal{T}} (1 - u_K)^2 \sum_{L\in\mathcal{N}_K} g^+_{K,L} + \frac{1}{2}\sum_{K\in\mathcal{T}}\sum_{L\in\mathcal{N}_K} g^+_{K,L}(u_L - u_K)^2 \\ & -\frac{1}{2}\sum_{K\in\mathcal{T}} (1 - u_L)^2 \sum_{L\in\mathcal{N}_K} g^+_{K,L}. \end{aligned}$$

Since $g^+_{K,L} = g^-_{L,K}$, we get

$$T_{19} = \frac{1}{2} \sum_{K \in \mathcal{T}} (1 - u_K)^2 G_K + \frac{1}{2} \sum_{(K,L) \in \mathcal{E}} |g_{K,L}| (u_L - u_K)^2.$$

Gathering the previous results, we get the conclusion. □

We can now state the convergence of the scheme to a weak process solution. This convergence result is obtained in the sense of the nonlinear weak-$\star$ convergence, defined in [7], which is a convenient way to understand the convergence towards a Young measure. Indeed, a bounded sequence $(u_n)_{n\in\mathbb{N}}$ of $L^\infty(\Omega)$ converges in the nonlinear weak-$\star$ sense to some function $u \in L^\infty(\Omega \times (0,1))$ if, for all $\xi \in C^0(\mathbb{R})$, the sequence $(\xi(u_n))_{n\in\mathbb{N}}$ converges for the weak-$\star$ topology of $L^\infty(\Omega)$ to the function $x \mapsto \int_0^1 \xi(u(x,\alpha))\mathrm{d}\alpha$ (the notation $\mathrm{d}\alpha$ stands here for the Lebesgue measure on $(0,1)$). A main compactness result is that from a bounded sequence of $L^\infty(\Omega)$, it is possible to extract a subsequence converging in the nonlinear weak-$\star$ sense (see [7] or [8] for more details).

PROPOSITION 4.6 (convergence of the scheme to a weak process solution). *Under Hypotheses* (H), *let* $(\mathcal{T}^{(m)}, g_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ *be a sequence such that, for all* $m \in \mathbb{N}$, $\mathcal{T}^{(m)}$ *is an admissible mesh of* Ω *in the sense of Definition* 4.1 *and* $g_{\mathcal{T}^{(m)}}$ *is a family of reals such that* (47) and (48) *are satisfied. We assume that* $\lim_{m\to\infty} \mathrm{size}(\mathcal{T}^{(m)}) = 0$, *that there exists* $R > 0$ *s.t* $\mathrm{regul}(\mathcal{T}^{(m)}) \le R$ *for all* $m \in \mathbb{N}$ *(see Definition* 4.1 *for the definitions of* size *and* regul*), and that* $\lim_{m\to\infty} \mathrm{cons}(g_{\mathcal{T}^{(m)}}) = 0$. *For all* $m \in \mathbb{N}$, *we denote by* $u_{\mathcal{T}^{(m)}}$ *a solution of scheme* (49)–(50) *such that* (60) *holds. Then, from the sequence* $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$, *one can extract a subsequence, again denoted* $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$, *such that the corresponding sequence* $(u_{\mathcal{T}^{(m)}} g)_{m\in\mathbb{N}}$ *converges in the nonlinear weak-*$\star$ *sense (see above for the sense of this convergence) to a weak process solution of problem* (10)–(11) *in the sense of Definition* 1.1.

Proof. Using the property (60) satisfied by $u_{\mathcal{T}^{(m)}}$, we can deduce the existence of a subsequence, again denoted $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$, such that the corresponding sequence $(u_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ converges in the nonlinear weak-$\star$ sense to some function $u \in L^\infty(\Omega \times (0,1))$. We shall now prove that u is the weak process solution of problem (10)–(11) in the sense of Definition 1.1. Let $\varphi \in C^1(\overline{\Omega}, \mathbb{R}_+)$, and let $\xi \in C^1(\mathbb{R})$ be a convex function with $\xi'(1) \ge 0$. Let $m \in \mathbb{N}$, and let $(\mathcal{T}^{(m)})$ be the corresponding admissible mesh of the subsequence. For simplicity, we do not mention the index m until we consider some convergence properties as $m \to \infty$. We get from (50), using $\xi'(u_K) = \xi'(1) + \xi'(u_K) - \xi'(1)$, that

$$\xi'(u_K) \left(\sum_{L \in \mathcal{N}_K} (g^+_{K,L} u_L - g^-_{K,L} u_K) + F_K \right) \ge 0 \qquad \forall K \in \mathcal{T}. \tag{62}$$

We can then multiply (62) by φ_K, where we denote $\varphi_K = \frac{1}{m_K} \int_K \varphi(x)\mathrm{d}x$, and we sum on $K \in \mathcal{T}$. We get $T_{21} + T_{22} \ge 0$, with

$$T_{21} = \sum_{K \in \mathcal{T}} \xi'(u_K)\varphi_K \sum_{L \in \mathcal{N}_K} (g^+_{K,L} u_L - g^-_{K,L} u_K)$$

and

$$T_{22} = \sum_{K \in \mathcal{T}} \xi'(u_K)\varphi_K F_K.$$

We have $T_{21} = T_{23} + T_{24}$ with

$$T_{23} = \sum_{K\in\mathcal{T}} \xi'(u_K)u_K\varphi_K \sum_{L\in\mathcal{N}_K} g_{K,L}$$

and

$$T_{24} = \sum_{K\in\mathcal{T}} \xi'(u_K)\varphi_K \sum_{L\in\mathcal{N}_K} g^+_{K,L}(u_L - u_K).$$

Since $\sum_{L\in\mathcal{N}_K} g_{K,L} = \int_K \mathrm{div} g(x)\mathrm{d}x$, we thus get that

$$\lim_{m\to\infty} T^{(m)}_{23} = \int_\Omega \int_0^1 \xi'(u(x,\alpha))u(x,\alpha)\varphi(x)\mathrm{div} g(x)\mathrm{d}\alpha\mathrm{d}x.$$

On the other hand, we have

$$T_{24} \le T_{25} := \sum_{K\in\mathcal{T}} \varphi_K \sum_{L\in\mathcal{N}_K} g^+_{K,L}(\xi(u_L) - \xi(u_K)).$$

Gathering by edges, we get

$$T_{25} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K))(\varphi_K g^+_{K,L} - \varphi_L g^-_{K,L}).$$

Let us compare T_{25} with T_{26} defined by

$$T_{26} = -\sum_{K\in\mathcal{T}} \xi(u_K) \int_K \mathrm{div}(\varphi(x)g(x))\mathrm{d}x.$$

We have, on one hand, that

$$\lim_{m\to\infty} T^{(m)}_{26} = -\int_\Omega \int_0^1 \xi(u(x,\alpha))\mathrm{div}(\varphi(x)g(x))\mathrm{d}\alpha\mathrm{d}x,$$

and on the other hand, we have

$$T_{26} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K)) \int_{K|L} \varphi(x)g(x)\cdot \mathbf{n}_{K,L}\mathrm{d}s(x).$$

Thus we get that

$$T_{25} - T_{26} = T_{27} + T_{28} + T_{29},$$

with

$$T_{27} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K)) \left(\varphi_K g^+_{K,L} - \varphi_L g^-_{K,L} - \frac{g_{K,L}}{m_{KL}} \int_{K|L} \varphi(x)\mathrm{d}s(x)\right),$$

$$T_{28} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K)) \left(g_{K,L} - \bar{g}_{K,L}\right) \left(\frac{1}{m_{KL}} \int_{K|L} \varphi(x)\mathrm{d}s(x)\right),$$

$$T_{29} = \sum_{(K,L)\in\mathcal{E}} (\xi(u_L) - \xi(u_K)) \left(\int_{K|L} (\frac{\bar{g}_{K,L}}{m_{KL}} - g(x)\cdot \mathbf{n}_{K,L})\varphi(x)\mathrm{d}s(x)\right)$$

(recall that $\bar{g}_{K,L}$ is defined by (53)). In the following, we designate by C_i various real numbers which can depend on $d, \Omega, g, F, \varphi, \xi$ but not on $\mathcal{T}$. Using $|\xi(u_K) - \xi(u_L)| \le$

$C_1\,|u_K - u_L|$ and the Cauchy–Schwarz inequality,

$$\left|\varphi_K - \frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)\right| \le \mathrm{diam}(K)C_2\,,$$

and

$$\left|\varphi_L - \frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)\right| \le \mathrm{diam}(L)C_2\,,$$

we get

$$|T_{27}|^2 \le C_3\left(\sum_{(K,L)\in\mathcal{E}}|g_{K,L}|(u_K-u_L)^2\right)\left(\sum_{(K,L)\in\mathcal{E}}|g_{K,L}|(\mathrm{diam}(K)^2+\mathrm{diam}(L)^2)\right).$$

Using (61) and

$$\sum_{(K,L)\in\mathcal{E}}|g_{K,L}|(\mathrm{diam}(K)^2+\mathrm{diam}(L)^2) \le C_4\,\mathrm{size}(\mathcal{T}),$$

we thus get that

$$\lim_{m\to\infty}|T_{27}^{(m)}| = 0.$$

We now turn to the study of T_{28}. Since we have

$$T_{28} = -\sum_{K\in\mathcal{T}}\xi(u_K)\sum_{L\in\mathcal{N}_K}(g_{K,L}-\bar{g}_{K,L})\left(\frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)\right),$$

we get, using the property (48),

$$T_{28} = -\sum_{K\in\mathcal{T}}\xi(u_K)\sum_{L\in\mathcal{N}_K}(g_{K,L}-\bar{g}_{K,L})\left(\frac{1}{m_{KL}}\int_{K|L}\varphi(x)\mathrm{d}s(x)-\varphi_K\right).$$

Thus, thanks to the Cauchy–Schwarz inequality and using (52), we get

$$T_{28}^2 \le C_5\,\mathrm{cons}(g_{\mathcal{T}}).$$

Thus

$$\lim_{m\to\infty}|T_{28}^{(m)}| = 0.$$

We conclude with the study of T_{29}. Since

$$T_{29} = -\sum_{K\in\mathcal{T}}\xi(u_K)\sum_{L\in\mathcal{N}_K}\left(\int_{K|L}\left(\frac{\bar{g}_{K,L}}{m_{KL}} - g(x)\cdot\mathbf{n}_{K,L}\right)(\varphi(x)-\varphi_K)\mathrm{d}s(x)\right)$$

and since $\int_{K|L}(\frac{\bar{g}_{K,L}}{m_{KL}} - g(x)\cdot\mathbf{n}_{K,L})(\varphi(x)-\varphi_K)\mathrm{d}s(x) \le C_6\,m_{KL}\mathrm{diam}(K)^2$, we easily get

$$\lim_{m\to\infty}|T_{29}^{(m)}| = 0.$$

Gathering these results gives

$$\lim_{m\to\infty} T_{25}^{(m)} = -\int_\Omega \int_0^1 \xi(u(x,\alpha))\mathrm{div}(\varphi(x)g(x))\mathrm{d}\alpha\mathrm{d}x.$$

Finally, we easily get

$$\lim_{m\to\infty} T_{22}^{(m)} = \int_\Omega \int_0^1 \xi'(u(x,\alpha))\varphi(x)F(x)\mathrm{d}\alpha\mathrm{d}x.$$

Gathering the previous results, we get $T_{23} + T_{25} + T_{22} \le 0$. Passing to the limit $m \to \infty$ in this inequality, we get

$$\begin{aligned}&+\int_\Omega \int_0^1 u(x,\alpha)\xi'(u(x,\alpha))\varphi(x)\mathrm{div}g(x)\mathrm{d}\alpha\mathrm{d}x\\ &-\int_\Omega \int_0^1 \xi(u(x,\alpha))\mathrm{div}(\varphi(x)g(x))\mathrm{d}\alpha\mathrm{d}x\\ &+\int_\Omega \int_0^1 \xi'(u(x,\alpha))\varphi(x)F(x)\mathrm{d}\alpha\mathrm{d}x \ge 0,\end{aligned}$$

which is exactly Definition 3.3. □

Thanks to the uniqueness result, we now classically conclude with the following convergence theorem (similar proofs can be found in [8]).

THEOREM 4.7 (strong convergence of the scheme to a weak solution). *Under Hypotheses* (H), *let $\mathcal{T}$ be an admissible mesh of Ω in the sense of Definition* 4.1*, and let $g_\mathcal{T}$ be a family of reals such that* (47) *and* (48) *are satisfied. Then the function $u_\mathcal{T} g$, where $u_\mathcal{T}$ is a solution of scheme* (49)–(50) *such that* (60) *holds, converges in $L^p(\Omega)^d$ for all $p \in [1,\infty)$ to $\tilde{g}$, the unique weak solution to problem* (10)–(11) *in the sense of Definition* 1.1*, as* size$(\mathcal{T})$ *tends to* 0*,* cons$(g_\mathcal{T})$ *tends to* 0*, and* regul$(\mathcal{T})$ *remains bounded (see Definition* 4.1 *for the definitions of* size$(\mathcal{T})$ *and* regul$(\mathcal{T})$*, and see* (52) *for the definition of* cons$(g_\mathcal{T})$*).*

Proof. Under Hypotheses (H), let $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$ be a sequence of admissible meshes of Ω in the sense of Definition 4.1 such that $\lim_{m\to\infty} \mathrm{size}(\mathcal{T}^{(m)}) = 0$. For all $m \in \mathbb{N}$, we denote by $u_{\mathcal{T}^{(m)}}$ a solution of scheme (47)–(50) such that (60) holds. Using Proposition 4.6, from the sequence $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$, one can extract a subsequence, again denoted $(\mathcal{T}^{(m)})_{m\in\mathbb{N}}$, such that the corresponding sequence $(u_{\mathcal{T}^{(m)}})_{m\in\mathbb{N}}$ converges in the nonlinear weak-$\star$ sense to a weak process solution u of problem (10)–(11) in the sense of Definition 1.1. We then get that the limit of $\int_\Omega g(x)^2(u_{\mathcal{T}^{(m)}}(x) - \int_0^1 u(x,\alpha)\mathrm{d}\alpha)^2\mathrm{d}x$ as $m \to \infty$ is equal to $\int_\Omega g(x)^2(\int_0^1 u(x,\alpha)^2\mathrm{d}\alpha - 2(\int_0^1 u(x,\alpha)\mathrm{d}\alpha)^2 + (\int_0^1 u(x,\alpha)\mathrm{d}\alpha)^2)\mathrm{d}x = 0$, using Proposition 3.5 which stands that $\tilde{g}(x) = u(x,\alpha)g(x)$, for a.e. $x \in \Omega$ and $\alpha \in (0,1)$. This proves that $(u_{\mathcal{T}^{(m)}}g)_{m\in\mathbb{N}}$ converges to $\tilde{g}$ in $L^2(\Omega)^d$. The uniqueness of $\tilde{g}$ gives the conclusion of the theorem. □

5. Numerical results.

5.1. One-dimensional example. We again consider the following data, studied in section 2: $\Omega = (-1,1)$, $g : x \mapsto x^3 - x$, and $F : x \mapsto 1/2$. We recall that the weak solution is the function $\tilde{g}$ given by $\tilde{g} = ug$, where the function u is such that $u : x \mapsto 1$ for all $x \in (-1, -\sqrt{1/2}) \cup (\sqrt{1/2}, 1)$ and $u : x \mapsto 1/(2(1-x^2))$ for all $x \in (-\sqrt{1/2}, \sqrt{1/2})$. We use Algorithm (A) to solve the nonlinear system (49)–(50)

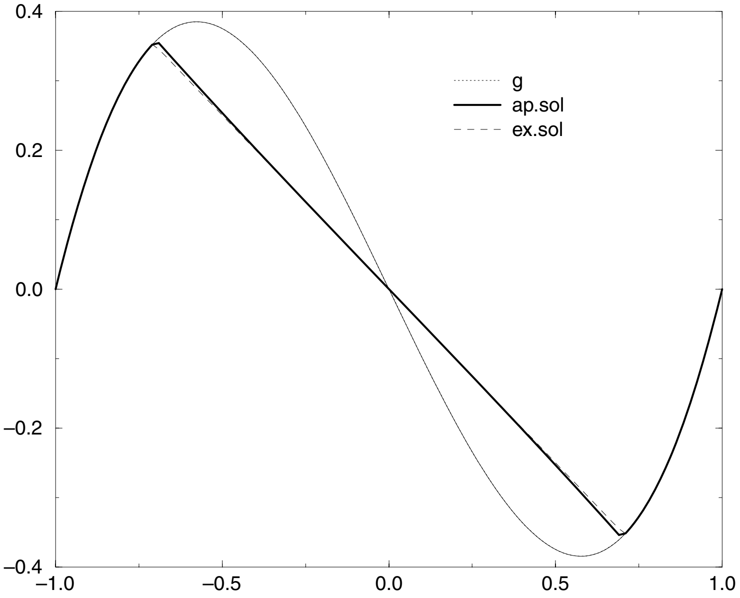


FIG. 1. *Approximate solution (ap.sol) and exact solution (ex.sol) with* 100 *control volumes.*

with $g_{K,L} = \bar{g}_{K,L}$ ($\bar{g}_{K,L}$ is defined in (53)). We get, with 100 uniform control volumes, the results given in Figure 1. The exact solution $\tilde{g}$ is represented by the dashed line (and denoted by "ex.sol." in the legend). The approximate solution of (49)–(50) is $u_{\mathcal{T}}$. Figure 1 gives, with the solid line, the product of $u_{\mathcal{T}}$ with the exact function g (and this product is denoted by "ap.sol." in the legend). The dashed line and the solid line are very close to one another. The last line, namely the grey dotted one, represents the exact function g.

It is interesting to remark that Algorithm (A) converges for a significantly smaller number of iterations than card($\mathcal{T}$). The table below gives, for different numbers of control volumes, the number of iterations until $p_K^{(n)} = p_K^{(n+1)}$ for all $K \in \mathcal{T}$.

Number of control volumes	Number of iterations	$\|\tilde{g} - u_{\mathcal{T}} g\|_{L^1(\Omega)}$
10	3	0.031757
50	9	0.006969
100	17	0.003488
500	76	0.000699
1000	151	0.000348
5000	748	0.000070
10000	1496	0.000035
50000	7473	0.000007

We observe that this number behaves as 1/size($\mathcal{T}$), whereas the error in $L^1(\Omega)$ behaves as size($\mathcal{T}$).

5.2. Two-dimensional examples. We use the coupled finite volume scheme (48)–(54) in order to compute $g_{\mathcal{T}}$. We consider the following data: $\Omega = (0,1)^2$, $\Lambda(x) = \mathrm{I}_d$, and $F(x) = 1/100$ for a.e. $x \in \Omega$, $g = \nabla h$, where h is a solution of

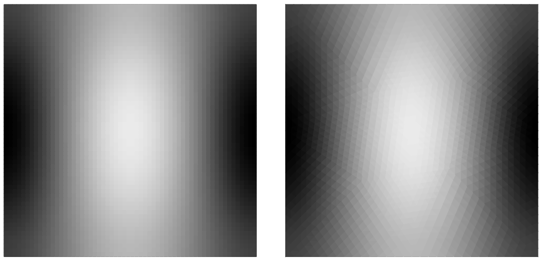


FIG. 2. *Value of h from* 0 *(black) to* 0.00111 *(white): rectangular* 60×60 *mesh (left) and triangular mesh with* 3650 *triangles (right).*

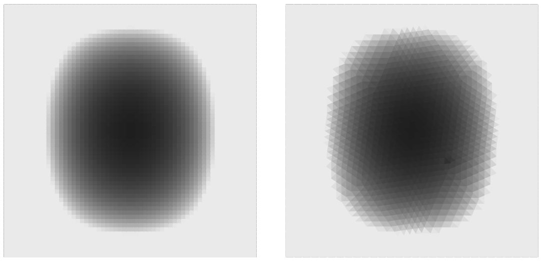


FIG. 3. *Value of u from* 0.48 *(black) to* 1 *(white): rectangular* 60×60 *mesh (left) and triangular mesh with* 3650 *triangles (right).*

the homogeneous Neumann problem

$$-\Delta h(x,y) = y(1-y)(-x^2+x-1/6) \qquad \forall (x,y) \in (0,1)^2,$$

$$\nabla h \cdot \mathbf{n} = 0 \text{ on } \partial\Omega.$$

These data have been chosen since they represent a kind of generalization in two dimensions of the one-dimensional case presented above. Two meshes have been tested. With a rectangular 60×60 mesh, the convergence of Algorithm (A) is obtained after 10 iterations; with a triangular mesh with 3650 triangles, 14 iterations are necessary to converge. The results obtained after the resolution of h by the finite volume method are presented in Figure 2. The corresponding values of the function u such that ug is the weak solution are given in Figure 3, and the values of g_x, g_y, $\tilde{g}_x$, $\tilde{g}_y$ which are the components of g and $\tilde{g}$ are given in Figures 4 and 5 for the rectangular mesh.

These results show the efficiency of the numerical method. In particular, we can remark that the approximate solution obtained with the rectangular mesh is very close to the approximate solution obtained with the triangular mesh.

The following table gives, for rectangular meshes, the number of iterations needed by Algorithm (A) for convergence.

Number of control volumes	Number of iterations
10×10	3
50×50	9
100×100	16
150×150	23
200×200	30

We again observe that this number behaves as $1/\text{size}(\mathcal{T})$.

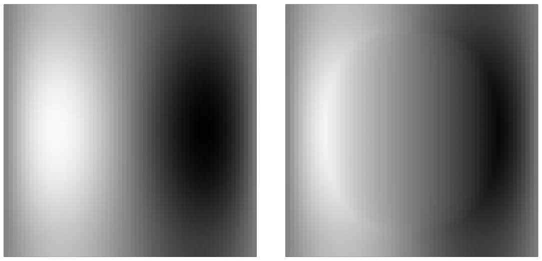


Fig. 4. *Value of g_x (left) and of $\tilde{g}_x$ (right) from* -0.00342 *(black) to* 0.00342 *(white).*

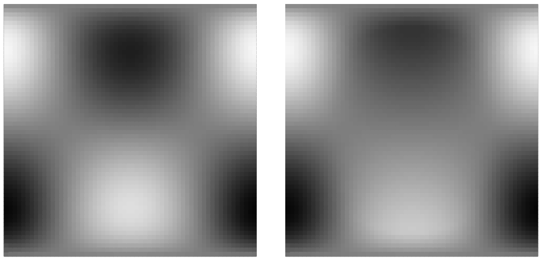


Fig. 5. *Value of g_y (left) and of $\tilde{g}_y$ (right) from* -0.00094 *(black) to* 0.00094 *(white).*

6. Conclusions.

We have been able to prove the existence and the uniqueness of the weak solution to problem (10)–(11) in the sense of Definition 1.1, and we have proved the convergence of a numerical scheme, under Hypotheses (H). At this time, we have not yet derived an error estimate although we can guess that it will be possible to follow the same steps as that of a scalar nonlinear hyperbolic problem, since the basis of proof of the uniqueness theorem is the doubling variable technique of Krushkov. It is, however, probable that the error estimate that we shall obtain will be not sharp. Moreover, the mathematical problem is not directly formulated as a function of h but on g. We have only briefly mentioned in remarks that some of the results of this paper can be obtained without the assumption $g = \Lambda\nabla h$. However, this is not the case for all of them. Finally, much work remains to be done in order to handle the complete problem (2)–(5).

REFERENCES

[1] R. S. Anderson and N. F. Humphrey, *Interaction of weathering and transport processes in the evolution of arid landscapes*, in Quantitative Dynamics Stratigraphy, T. A. Cross, ed., Prentice–Hall, Englewood Cliffs, NJ, 1989, pp. 349–361.

[2] S. N. Antontsev, G. Gagneux, and G. Vallet, *On some stratigraphic control problems*, J. Appl. Mech. Tech. Phys., 44 (2003), pp. 821–828.

[3] R. DiPerna, *Measure-valued solutions to conservation laws*, Arch. Ration. Mech. Anal., 88 (1985), pp. 223–270.

[4] J. Droniou, R. Eymard, D. Hilhorst, and X. D. Zhou, *Convergence of a finite-volume mixed finite-element method for a system of an elliptic-hyperbolic system*, IMA J. Numer. Anal., 23 (2003), pp. 507–538.

[5] R. Eymard, T. Gallouët, V. Gervais, and R. Masson, *Convergence of a numerical scheme for stratigraphic modeling*, SIAM J. Numer. Anal., 43 (2005), pp. 474–501.

[6] R. Eymard, T. Gallouët, D. Granjeon, R. Masson, and Q. H. Tran, *Multi-lithology stratigraphic model under maximum erosion rate constraint*, Internat. J. Numer. Methods Engrg., 60 (2004), pp. 527–548.

[7] R. Eymard, T. Gallouët, and R. Herbin, *Existence and uniqueness of the entropy solution to a nonlinear hyperbolic equation*, Chinese Ann. Math. Ser. B, 16 (1995), pp. 1–14.

[8] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume method*, in Handbook of Numerical Analysis, Vol. VII, North-Holland, Amsterdam, 2000, pp. 715–1022.

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

[10] G. Gagneux and G. Vallet, *Sur des problèmes d'asservissements stratigraphiques*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 715–739.

[11] D. Granjeon, P. Joseph, and B. Doligez, *Using a 3-D stratigraphic model to optimize reservoir description*, Hart's Petroleum Eng. Internat., November (1998), pp. 51–58.

[12] S. N. Krushkov, *First order quasilinear equations with several space variables*, Mat. Sb., 10 (1970), pp. 217–243 (in Russian).

[13] A. Michel, *A finite volume scheme for two-phase immiscible flow in porous media*, SIAM J. Numer. Anal., 41 (2003), pp. 1301–1317.

[14] P. Pedregal, *Optimization, relaxation and Young measures*, Bull. Amer. Math. Soc. (N.S.), 36 (1999), pp. 27–58.

[15] J. C. Rivenaes, *Impact of sediment transport efficiency on large scale sequence architecture: Results from stratigraphic computer simulation*, Basin Res., 4 (1992), pp. 133–146.

[16] T. Roubiček, *Nonlinear Partial Differential Equations with Applications*, Internat. Ser. Numer. Math., Birkhäuser, Basel, Boston, Berlin, 2005.

[17] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier (Grenoble), 15 (1965), pp. 189–258.

[18] M. H. Vignal, *Convergence of a finite volume scheme for a system of an elliptic equation and a hyperbolic equation*, Modél. Math. Anal. Numér., 30 (1996), pp. 841–872.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2371–2401

APPROXIMATIONS FOR VISCOSITY SOLUTIONS OF HAMILTON–JACOBI EQUATIONS WITH LOCALLY VARYING TIME AND SPACE GRIDS*

JIANLIANG QIAN†

Abstract. A new monotone finite difference scheme is introduced that approximates viscosity solutions of first-order nonlinear Hamilton–Jacobi equations. The main feature of the scheme is that it allows for locally varying time and space grids, making it ideal for use with adaptive algorithms. Explicit a priori error estimates are given to establish convergence. Numerical examples, including adaptive mesh refinement examples, demonstrate the effectiveness of the proposed scheme.

Key words. local time-stepping, error estimates, monotone schemes, Hamilton–Jacobi

AMS subject classifications. Primary: 54C40, 14E20; Secondary: 46E25, 20C20

DOI. 10.1137/04060682x

1. Introduction.

We consider Hamilton–Jacobi equations

$$u_t + H(t, x, u, \nabla u) = 0 \qquad \text{in } \mathbb{R}^d \times (0, T),$$

where H is a nondecreasing function of u. These equations arise in many areas of applied mathematics, like optimal control, differential games, seismic wave propagation, terrain navigation of robotic navigation, and financial mathematics [16], among many others. They also appear when modeling evolving interfaces in geometry, fluid mechanics, computer vision [34, 28, 21, 42], and materials science [36]; they are essential when dealing with level set methods [30], numerical methods that have reached widespread popularity. Therefore, there is general interest in designing efficient numerical methods for such equations; see [26, 24, 27, 41, 6, 7] and references therein.

Osher and Sanders [29] designed a class of monotone schemes for scalar conservation laws with locally varying time and space grids. Inspired by this work, we design a new class of monotone schemes for Hamilton–Jacobi equations. In this paper, we shall obtain an upper bound of the L^∞-norm of the difference between the approximate solution v_h and the viscosity solution of the model Hamilton–Jacobi equation

$$u_t + H(\nabla u) = 0 \qquad \text{in } \mathbb{R}^d \times (0, T).$$

The class of monotone schemes we consider here includes the classical Lax–Friedrichs scheme, the monotone schemes considered by Crandall and Lions [15], the intrinsic monotone scheme of Abgrall [1], and the monotone schemes devised by Kossioris, Makridakis, and Souganidis [22]; it consists of numerical schemes that can be characterized in terms of the so-called numerical Hamiltonian $\widehat{H}$ which approximates the exact Hamiltonian H. Therefore, all the schemes proposed in [15, 1, 22] can be formulated under this new framework.

Our error estimate is of the form

$$\| u - v_h \|_{L^\infty(Q_h)} \le C_1 \, h^{\frac{1}{2}},$$

*Received by the editors April 15, 2004; accepted for publication (in revised form) June 14, 2005; published electronically January 6, 2006. This research was supported by ONR Grant N00014-02-1-0720.

http://www.siam.org/journals/sinum/43-6/60682.html

†Department of Mathematics, UCLA, Los Angeles, CA 90095-1555 (qian@math.ucla.edu).

where Q_h denotes the set of space-time points at which v_h is defined, h denotes the so-called mesh-size, and C_1 depends on the $W^{1,\infty}$-norm of u. The order of convergence in h, $\frac{1}{2}$, is the now classical order of convergence already obtained in the previously mentioned papers. However, if the solution is smooth, then we can also prove that

$$\| u - v_h \|_{L^\infty(Q_h)} \leq C_2 \, h,$$

where C_2 depends on the $W^{2,\infty}$-norm of u; this seems to be a new result.

Just as was done in [15], the error estimate we present is obtained by using the *structure* of the numerical scheme and three key properties of its numerical Hamiltonian $\widehat{H}$, namely, consistency, monotonicity, and local smoothness. In this respect, our result is similar to the error estimates for monotone schemes for nonlinear scalar hyperbolic conservation laws obtained by Cockburn and Gremaud [9, 10] and by Cockburn, Gremaud, and Yang [11].

The technique we use to prove our result is a variation of the technique used by Albert et al. [2] to prove a posteriori error estimates for steady-state Hamilton–Jacobi equations. Those estimates are of the form

$$\| u - v_h \|_{L^\infty(\mathbb{R}^d)} \leq \Phi(v_h)$$

and are obtained by using a modification of the elegant technique that Crandall, Evans, and Lions [14] devised to study viscosity solutions of Hamilton–Jacobi equations, where Φ is a nonlinear functional depending only on the computed solution v_h. Our technique is also related to the one used by Cockburn and Qian [12] to prove continuous dependence results for steady-state Hamilton–Jacobi equations, but in that work we dealt only with steady Hamilton–Jacobi equations and no time variable is involved in the error estimates. In this work we apply this doubling variable technique to deal with time-dependent Hamilton–Jacobi equations with locally varying time and space grids.

The devising of such schemes with locally varying time and space grids is fundamental to the development of efficient adaptive methods for Hamilton–Jacobi equations. As is well known, the viscosity solution for Hamilton–Jacobi equations is continuous but may develop kinks in finite time; namely, its gradient might be discontinuous although the initial data is smooth. To resolve such sharp kinks with high resolution, one might use higher-order schemes [30, 31, 26, 20, 41, 6]. On the other hand, one might also use an adaptive strategy to resolve the kinks with lower-order methods.

In general there are two distinct starting points for devising adaptive methods for PDEs. One is getting the most accurate solution for a fixed cost, and the other is attaining a fixed accuracy for a minimum cost.

The so-called adaptive mesh redistribution method is based on the former starting point: getting the most accurate solution with a fixed number of mesh points; see [19, 33, 35, 25, 8, 38, 39, 37] and references therein. In such methods, one dynamically maps the physical domain to a computational domain via an invertible mesh generator. The mesh generator is derived from a variational principle so that the dynamic behavior of the solution can be taken into account. However, the resulting mesh typically concentrates a high number of mesh points around a localized region where the solution is singular or nearly singular [19, 8, 39]; therefore, globally chosen time steps need to be small to be proportional to the smallest mesh size in space, according to the Courant-Friedrichs-Lewy (CFL) stability condition. Nevertheless, for Hamilton–Jacobi equations viscosity solutions are usually Lipschitz continuous and thus are differentiable almost everywhere, and the resulting kinks are expected to develop in quite localized regions. Hence, the small time steps are really needed only in those localized regions, and it is meaningful to develop schemes with locally varying time and space grids. In [37], Tan et al. successfully tested a class of adaptive mesh redistribution methods with locally varying time steps for nonlinear hyperbolic conservation laws. Although an adaptive mesh redistribution method was proposed for Hamilton–Jacobi equations in [39], the method can be made more efficient based on a monotone scheme with locally varying time and space grids. For a nice review on mesh redistribution methods for computational fluid dynamics, see [40].

The so-called adaptive mesh refinement method is based on the following starting point: attaining a prespecified accuracy with a minimum cost; see [3, 5, 23, 17] for such methods for nonlinear hyperbolic conservation laws and see [32, 13] for such methods for Hamilton–Jacobi equations. The local adaptive mesh refinement methods in [3, 5] are based on ad hoc local truncation error estimation procedures and thus are not optimal. The adaptive methods in [23, 17] are based on a posteriori error estimates for hyperbolic conservation laws and thus are optimal to some extent. In particular, the adaptive algorithm proposed by Gosse and Makridakis [17] enforces a strict error control mechanism on each mesh cell by refining and coarsening computational grids according to their local a posteriori error estimates so that a given error tolerance is satisfied for each cell at every time step. Interestingly, a similar error control mechanism was used for solving a class of Hamilton–Jacobi equations in [32], where an adaptive mesh refinement method for eikonal equations is designed based on asymptotic truncation error estimates and some numerical ODE techniques. However, an optimal approach for triggering mesh refinement and coarsening should be based on a posteriori error estimates. The adaptive mesh refinement method proposed by Cockburn and Yenikaya [13] is based on such an a posteriori error estimate [2] and thus has a rigorous error control for steady Hamilton–Jacobi equations; the study of the effectivity index carried out in that paper indicates that the adaptive method has optimal complexity. However, for the time-dependent Hamilton–Jacobi equations the adaptive mesh refinement method must allow time space meshes to vary locally. In this work, we construct a class of monotone finite difference methods for Hamilton–Jacobi equations with locally varying time and space grids and obtain convergence rates. As mentioned above, this class of methods can be used for both adaptive mesh redistribution and adaptive mesh refinement methods.

The results obtained for the model time-dependent Hamilton–Jacobi equation can be extended to the more general case,

$$H(x, t, u, u_t, \nabla u) = 0,$$

without major difficulty, provided that the mapping $u \mapsto H(x, t, u, p_t, p_x)$ is nondecreasing and the mapping $p_t \mapsto H(x, t, u, p_t, p_x)$ is strictly increasing.

We are currently incorporating the proposed scheme with a posteriori error estimates to design optimal adaptive methods for Hamilton–Jacobi equations. This constitutes the subject of an ongoing work. However, to demonstrate the effectiveness of the new scheme we will present some numerical examples for one- and two-dimensional Hamilton–Jacobi equations; in particular, we present adaptive mesh refinement (AMR) examples for two-dimensional Hamilton–Jacobi equations using the monotone scheme presented here as the driving method and a local truncation error estimator in [3] as the indicator for the mesh refinement.

The paper is organized as follows. In section 2, we define the viscosity solution and describe the new monotone schemes. In section 3, we state, discuss, and prove our main results. Section 4 presents some examples of monotone schemes for which our results hold. In section 5 we present some numerical experiments to demonstrate the effectiveness of the scheme.

2. New monotone schemes for Hamilton–Jacobi equations. We consider the following model time-dependent Hamilton–Jacobi equation:

$$u_t + H(\nabla u) = 0 \qquad \text{in } \mathbb{R}^d \times (0,T), \tag{2.2.1}$$

$$u(t=0) = u_0 \qquad \text{in } \mathbb{R}^d, \tag{2.2.2}$$

where u and u_0 are periodic in each coordinate with period 1, and $H \in C(\mathbb{R}^d)$. We begin by defining the notion of viscosity solution and describing the monotone schemes of interest. Then, we state, discuss, and prove the main results.

2.1. Viscosity solutions. We begin by defining the viscosity solutions of the initial value problem (2.2.1) and (2.2.2). To do this, we need the notions of *semidifferentials* of a function on $\mathbb{R}^d \times (0,T)$. The *superdifferential* of a function u at a point $(x,t) \in \mathbb{R}^d \times (0,T)$, $D^+u(x,t)$, is the set of all vectors $p = (p_x, p_t)$ in $\mathbb{R}^d \times \mathbb{R}$ such that

$$\limsup_{(y,s)\in\mathbb{R}^d\times(0,T)\to(x,t)} \left(\frac{u(y,s) - \{u(x,t) + (s-t)\,p_t + (y-x)\cdot p_x\}}{\|(y,s)-(x,t)\|} \right) \le 0,$$

and the *subdifferential* of u at a point (x,t), $D^-u(x,t)$, is the set of all vectors p in $\mathbb{R}^{d+1}$ such that

$$\liminf_{(y,s)\in\mathbb{R}^d\times(0,T)\to(x,t)} \left(\frac{u(y,s) - \{u(x,t) + (s-t)\,p_t + (y-x)\cdot p_x\}}{\|(y,s)-(x,t)\|} \right) \ge 0.$$

We are now ready to define the viscosity solution of (2.2.1).

Definition 2.1 (see [14]). *A viscosity solution u of the initial-value problem for the Hamilton–Jacobi equation* (2.2.1) *is a periodic, continuous function on* $\mathbb{R}^d \times [0,T]$ *satisfying $u(t=0) = u_0$ such that for all (x,t) in $\mathbb{R}^d \times (0,T]$,*

$$\sigma\,(\,p_t + H(p_x)\,) \le 0 \qquad \forall\, p = (p_x, p_t) \in D^\sigma u(x,t), \qquad \sigma \in \{+,-\}.$$

2.2. The monotone schemes. The numerical schemes we consider determine the values of a function v_h on a standard grid $Q_h = G_h \times \{t^n\}_{n=0}^{N_T}$ of $\mathbb{R}^d \times [0,T]$; the spatial grid G_h is periodic with period 1 in each of the canonical directions of $\mathbb{R}^d$. To avoid cluttered notation, we use v instead of v_h to denote the numerical solution in the following derivation. These schemes take the form

$$v(y,0) = u_0(y) \qquad \forall\, y \in G_h, \tag{2.2.3}$$

$$v(y,t^{n+1}) = v(y,t^n) - \Delta t^n \widehat{H}_y(\partial_{\delta_y} v(y,t^n)) \quad \forall\, y \in G_h, \quad n = 0, \dots, N_T - 1, \tag{2.2.4}$$

where $\widehat{H}_y(\partial_{\delta_y} v(y,t^n))$ is an approximation to $H(\nabla v(y,t^n))$,

$$\partial_{\delta_y} v(y,t^n) = (\partial_{\delta_{y,1}} v(y,t^n), \dots, \partial_{\delta_{y,N_y}} v(y,t^n)),$$

and

$$\partial_{\delta_{y,i}} v(y,t^n) = \frac{v(y,t^n) - v(y-\delta_{y,i},t^n)}{|\,\delta_{y,i}\,|}, \text{ where } y - \delta_{y,i} \in G_h, \qquad i = 1, \dots, N_y.$$

Here N_y denotes the number of edges at node y and $\delta_{y,i}$ denotes the position vector of the ith edge at node y. We shall show later that many important numerical Hamiltonians found in current literature have such a structure.

We denote bounded functions on G_h by $l^\infty(G_h)$ equipped with the norm

$$\|v\|_\infty = \sup_{y\in G_h} |v(y)|,$$

where $v \in l^\infty(G_h)$. Such functions can be identified with piecewise linear functions, still denoted by v.

In addition, we assume that the numerical Hamiltonian $\widehat{H}$ has the following properties:

(i) consistency: $\widehat{H}_y(\partial_{\delta_y} v(y)) = H(p)$ if $\nabla v = p \in \mathbb{R}^d$;

(ii) monotonicity: $\widehat{H}_y$ is nondecreasing in each of its arguments;

(iii) smoothness: $\widehat{H}_y$ is locally Lipschitz, that is,

$$|\, \widehat{H}_y(z_1) - \widehat{H}_y(z_2)\,| \le L(M)\, \|\, z_1 - z_2\,\|_\infty, \quad \text{where } \|\, z_i\,\|_\infty \le M, \quad i = 1, 2.$$

As is well known, the first property ensures that $\widehat{H}_y$ is exact for v with constant gradients on the whole domain so that we are approximating the viscosity solution with the correct Hamiltonian. The second property is *precisely* the one on the numerical Hamiltonian that is required to obtain the a priori error estimates. The third property, unlike the previous two, is not really essential and can be relaxed; however, we do not know of any numerical scheme used in practice that does not satisfy it.

It is not difficult to verify that the function

$$\mathbb{G}_y(v(y,t); v(y-\delta_{y,1},t), \ldots, v(y-\delta_{y,N_y},t)) = v(y,t) - \Delta t \widehat{H}_y(\partial_{\delta_y} v(y,t)) \tag{2.2.5}$$

is nondecreasing in each of its arguments for all $y \in G_h$ for a small enough $\Delta t \ge 0$. As is well known, when this is the case, the scheme is said to be a monotone scheme. In what follows, we assume that the above schemes are monotone.

Next, we show that the well-known Lax–Friedrichs scheme satisfies the above conditions.

Example 2.2. *The Lax–Friedrichs scheme.* The Lax–Friedrichs scheme [15, 31] on the uniform Cartesian grid

$$G_h = \{(i,j) = (x_0 + (i-1)\Delta x, y_0 + (j-1)\Delta y)\}$$

reads as follows:

$$\begin{aligned} v_{i,j}^{n+1} = v_{i,j}^n &- \Delta t^n H\left(\frac{v_{i+1,j}^n - v_{i-1,j}^n}{2\Delta x}, \frac{v_{i,j+1}^n - v_{i,j-1}^n}{2\Delta y}\right) \\ &+ \Delta t^n \omega_x \frac{v_{i+1,j}^n - 2v_{i,j}^n + v_{i-1,j}^n}{\Delta x^2} + \Delta t^n \omega_y \frac{v_{i,j+1}^n - 2v_{i,j}^n + v_{i,j-1}^n}{\Delta y^2}, \end{aligned}$$

where

$$\omega_x = \sup_{(x,y)\in\mathbb{R}^d} \frac{1}{2}\, |H_1(\cdot,\cdot)|\, \Delta x,$$

$$\omega_y = \sup_{(x,y)\in\mathbb{R}^d} \frac{1}{2}\, |H_2(\cdot,\cdot)|\, \Delta y,$$

and $H_i(p_1,p_2) = \frac{\partial H}{\partial p_i}(p_1,p_2)$ for $i = 1, 2$.

Since the grid G_h is Cartesian, for each $y = (i, j) \in G_h$ we have $N_y = 4$; the quantities $\partial_{\delta_{y,i}} v(y)$ are thus the following:

$$\partial_{-\Delta x} v_{i,j} = \frac{v_{i,j} - v_{i+1,j}}{\Delta x}, \qquad \partial_{\Delta x} v_{i,j} = \frac{v_{i,j} - v_{i-1,j}}{\Delta x},$$
$$\partial_{-\Delta y} v_{i,j} = \frac{v_{i,j} - v_{i,j+1}}{\Delta y}, \qquad \partial_{\Delta y} v_{i,j} = \frac{v_{i,j} - v_{i,j-1}}{\Delta y};$$

therefore,

$$\begin{aligned}\widehat{H}_y = {} & H\left(\frac{1}{2}(\partial_{\Delta x} v_{i,j} - \partial_{-\Delta x} v_{i,j}), \frac{1}{2}(\partial_{\Delta y} v_{i,j} - \partial_{-\Delta y} v_{i,j})\right) \\ & + \frac{\omega_x}{\Delta x}(\partial_{-\Delta x} v_{i,j} + \partial_{\Delta x} v_{i,j}) + \frac{\omega_y}{\Delta y}(\partial_{-\Delta y} v_{i,j} + \partial_{\Delta y} v_{i,j}).\end{aligned}$$

It is easy to verify that the above $\widehat{H}_y$ satisfies properties (i), (ii), and (iii).

In section 4, we give more examples of schemes satisfying these three properties.

2.3. A new class of monotone schemes. Consider a regular triangulation $\mathcal{T}_h$ of $\mathbb{R}^d$ and define the grid G_h to be the collection of vertices y_j of simplexes in $\mathcal{T}_h$. In the two-dimensional case, these simplexes are triangles. Consider a function v on the standard grid Q_h. Here we enumerate the points y_j of the grid G_h and identify the function v defined on G_h with the point $(v_1, \ldots, v_N)$, where $v_j = v(y_j)$; similar considerations lead to $v_j^n = v(y_j, t^n)$ on Q_h.

Next we denote by Ω_j a control-volume centered at y_j used to define the average of the solution $u(x, t)$ and its numerical gradients at y_j and time t. The control-volume can be taken as Abgrall's intrinsic control-volume [1], the covolume introduced in [22], or the nonconforming dual-volume in [22]. Denote the collection of control-volumes Ω_j as Ω_h.

At each time level t^n, decompose Ω_h into two sets, $\bigcup_{j\in\mathcal{C}^n} \Omega_j$ and $\bigcup_{j\notin\mathcal{C}^n} \Omega_j$, where $\mathcal{C}^n$ is any subset of integers. Let $\Delta t^n = t^{n+1} - t^n$ be the time increment associated to $j \in \mathcal{C}^n$. For $j \notin \mathcal{C}^n$, define the following fractional time increment: $[t^n, t^{n+1}) = \bigcup_{l=0}^{M-1} [t^{n+\eta_l}, t^{n+\eta_{l+1}})$, where $t^{n+\eta_l}$ is defined below. Let $\{\sigma_k\}_{k=1}^M$ satisfy $\sigma_k > 0$ for $k = 1, \ldots, M$ and $\sum_{k=1}^M \sigma_k = 1$. Define η_l to be the partial sum: $\eta_l = \sum_{k=1}^l \sigma_k$ with $\eta_0 = 0$. Then define $t^{n+\eta_{l+1}} = t^{n+\eta_l} + \sigma_{l+1}\Delta t^n$. See Figure 1 for an illustration of such a decomposition in the one-dimensional case.

In the adaptive computation, the set $\mathcal{C}^n$ can be constructed through a posteriori error estimates so that it identifies where the solution is smooth and the complement of the set $\mathcal{C}^n$ identifies where the solution has higher or discontinuous gradients. Therefore, for $j \in \mathcal{C}^n$ we may use large time steps; for $j \notin \mathcal{C}^n$ we may use small time steps according to the local CFL condition shown below; see [13] for such identification mechanisms for the one-dimensional steady case of Hamilton–Jacobi equations.

We shall consider the piecewise linear continuous approximation of the solution on Q_h; namely, the gradient of the numerical solution is piecewise constant. We propose to advance from time level t^n to time level t^{n+1} via a predictor-corrector type scheme. Next we give a formal derivation of the scheme.

Integrating (2.2.1) over Ω_j from $t^{'}$ to $t^{''}$, we have

$$\text{(2.2.6)} \quad \frac{1}{|\Omega_j|}\int_{\Omega_j} u(x, t^{''})dx = \frac{1}{|\Omega_j|}\int_{\Omega_j} u(x, t^{'})dx - \frac{1}{|\Omega_j|}\int_{t'}^{t^{''}}\int_{\Omega_j} H(\nabla u(x, t))dxdt;$$

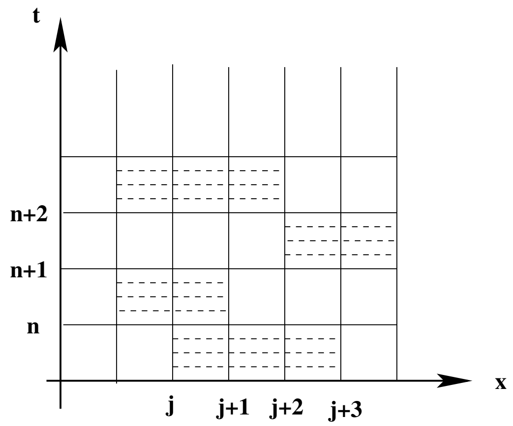


FIG. 1. *An illustration of decomposing nodes into two different sets and substepping in time.*

denote

$$v_j(t) = \frac{1}{|\Omega_j|}\int_{\Omega_j} u(x,t)dx. \tag{2.2.7}$$

For $j \in \mathcal{C}^n$, define

$$v_j(t) = \frac{(t^{n+1}-t)}{\Delta t^n} v_j(t^n) + \frac{(t-t^n)}{\Delta t^n} v_j(t^{n+1}) \quad \text{when } t \in [t^n, t^{n+1});$$

for $j \notin \mathcal{C}^n$, define

$$v_j(t) = \frac{(t^{n+\eta_{k+1}}-t)}{\sigma_{k+1}\Delta t^n} v_j(t^{n+\eta_k}) + \frac{(t-t^{n+\eta_k})}{\sigma_{k+1}\Delta t^n} v_j(t^{n+\eta_{k+1}})$$

when $t \in [t^{n+\eta_k}, t^{n+\eta_{k+1}})$ for $k = 0, \dots, M-1$.

In (2.2.6), for $j \in \mathcal{C}^n$, we take $t^{'} = t^n$ and $t^{''} = t^{n+1}$; for $j \notin \mathcal{C}^n$, we take $t^{'} = t^{n+\eta_k}$ and $t^{''} = t^{n+\eta_{k+1}}$ for $k = 0, \dots, M-1$. Formally substitute the numerical Hamiltonian $\widehat{H}$ for H in (2.2.6). For $j \in \mathcal{C}^n$, we obtain

$$v_j^{n+1} = v_j^n - \Delta t^n \frac{1}{|\Omega_j|}\int_{\Omega_j} \widehat{H}_j(\partial_{\delta_j} v_j^n)dx;$$

for $j \notin \mathcal{C}^n$, we obtain

$$v_j^{n+\eta_{k+1}} = v_j^{n+\eta_k} - \sigma_{k+1}\Delta t^n \frac{1}{|\Omega_j|}\int_{\Omega_j} \widehat{H}_j(\partial_{\delta_j} v_j^{n+\eta_k})dx$$

for $k = 0, \dots, M-1$. Here j corresponds to node y,

$$\partial_{\delta_j} v_j^n = (\partial_{\delta_{j,1}} v(y,t^n), \dots, \partial_{\delta_{j,N_y}} v(y,t^n))$$

and

$$\partial_{\delta_{j,i}} v_j^n = \frac{v(y,t^n) - v(y-\delta_{j,i},t^n)}{|\,\delta_{j,i}\,|}, \text{ where } y - \delta_{j,i} \in G_h, \qquad i = 1, \dots, N_j;$$

N_j denotes the number of edges at node y indexed by j, $\delta_{j,i}$ denotes the position vector of the ith edge at node y, and $|\delta_{j,i}|$ is the length of the ith edge associated to node j.

Because the approximation is piecewise linear on G_h, we can simplify the above formulation to the following: for $j \in \mathcal{C}^n$,

$$v_j^{n+1} = v_j^n - \Delta t^n \widehat{H}_j(\partial_{\delta_j} v_j^n);$$

for $j \notin \mathcal{C}^n$,

$$v_j^{n+\eta_{k+1}} = v_j^n - \sum_{l=0}^{k} \sigma_{l+1} \Delta t^n \widehat{H}_j(\partial_{\delta_j} v_j^{n+\eta_l})$$

for $k = 0, \dots, M-1$.

Therefore we have the predictor scheme

$$v_j^{n+\eta_k} = \begin{cases} v_j^n - \eta_k \Delta t^n \widehat{H}_j(\partial_{\delta_j} v_j^n), & j \in \mathcal{C}^n, \\ v_j^n - \sum_{l=0}^{k-1} \sigma_{l+1} \Delta t^n \widehat{H}_j(\partial_{\delta_j} v_j^{n+\eta_l}), & j \notin \mathcal{C}^n \end{cases} \tag{2.2.8}$$

for $k = 1, \dots, M-1$; the corrector is

$$v_j^{n+1} = \begin{cases} v_j^n - \Delta t^n \widehat{H}_j(\partial_{\delta_j} v_j^n), & j \in \mathcal{C}^n, \\ v_j^n - \sum_{l=0}^{M-1} \sigma_{l+1} \Delta t^n \widehat{H}_j(\partial_{\delta_j} v_j^{n+\eta_l}), & j \notin \mathcal{C}^n. \end{cases} \tag{2.2.9}$$

The above predictor-corrector type scheme is monotone under the following local CFL condition:

$$\sum_{i=1}^{N_j} \frac{\Delta t^n}{|\delta_{j,i}|} \frac{\partial \widehat{H}_j}{\partial p_{j,i}} \le 1 \qquad \text{for } j \in \mathcal{C}^n, \tag{2.2.10}$$

$$\sum_{i=1}^{N_j} \frac{\sigma_{l+1} \Delta t^n}{|\delta_{j,i}|} \frac{\partial \widehat{H}_j}{\partial p_{j,i}} \le 1 \qquad \text{for } j \notin \mathcal{C}^n, \tag{2.2.11}$$

where

$$\widehat{H}_j = \widehat{H}_j(\partial_{\delta_j} v_j) = \widehat{H}_j(p_{j,1}, \dots, p_{j,N_j}).$$

The above CFL condition is obtained by using the definition of monotone schemes defined through the operator (2.2.5). The condition (2.2.10) is the usual CFL restriction to make the scheme stable. The condition (2.2.11) is to enforce smaller time steps when local directional derivatives vary significantly.

Notice that in practice we do not need to compute all the solutions at the intermediate steps for $i \in \mathcal{C}^n$; only at those is involved in the computation for $j \notin \mathcal{C}^n$ is there a need for computation and storage; this, in turn, can be dealt with by using ghost points around those $j \notin \mathcal{C}^n$ as illustrated in the numerical examples.

2.4. Stability properties of the new monotone scheme. Let $\vec{G}_{\Delta t^n}$ be the self-map of $l^\infty(G_h)$ defined by (2.2.9), i.e.,

$$v^{n+1} = \vec{G}_{\Delta t^n}(v^n) \tag{2.2.12}$$

at each time step t^n.

We identify $\lambda \in \mathbb{R}$ with the constant function on G_h. From the form of $\vec{G}_{\Delta t^n}$, it is clear that

$$\vec{G}_{\Delta t^n}(v+\lambda) = \vec{G}_{\Delta t^n}(v) + \lambda$$

for all $v \in l^\infty(G_h)$ and $\lambda \in \mathbb{R}$. Since $\vec{G}_{\Delta t^n}$ is monotone by the CFL condition (2.2.10) and (2.2.11), $\vec{G}_{\Delta t^n}$ is a nonexpansive operator on $l^\infty(G_h)$ [15]. Namely,

$$\|\vec{G}_{\Delta t^n}(w) - \vec{G}_{\Delta t^n}(v)\|_\infty \le \|w - v\|_\infty$$

for all $w, v \in l^\infty(G_h)$.

Next we assume that the Lipschitz constant of v is preserved by the operator $\vec{G}_{\Delta t^n}$ at each time step t^n if v is Lipschitz continuous; this assumption can be verified for individual schemes, for example, the monotone finite difference schemes in [15] and the intrinsic monotone scheme [1]. Therefore, if the given initial data v^0 is Lipschitz continuous with L_0 as a Lipschitz constant, then L_0 will be the Lipschitz constant for each v^n $(n = 1, 2, \dots)$.

Because the numerical Hamiltonian $\hat{H}$ is locally Lipschitz continuous, we have

$$K = \sup\{|\widehat{H}_j(p_{j,1}, \dots, p_{j,N_j})| : |p_{j,i}| \le L_0, 1 \le i \le N_j, j \in G_h, 0 \le n \le N_T\} < \infty. \tag{2.2.13}$$

Furthermore, we can estimate, for any $j \in G_h$,

$$\begin{aligned} |v_j^{n+m} - v_j^n| &\le |v_j^{n+m} - v_j^{n+m-1}| + \cdots + |v_j^{n+1} - v_j^n| \\ &\le K\Delta t^{n+m-1} + \cdots + K\Delta t^n \\ &= K(t^{n+m} - t^n); \end{aligned} \tag{2.2.14}$$

thus

$$\|v^{n+m} - v^n\|_\infty \le K(t^{n+m} - t^n).$$

We can summarize the above properties in the following.

Proposition 2.3. *Let the scheme be monotone and $\vec{G}_{\Delta t^n}$ be the self-map of $l^\infty(G_h)$ as given in* (2.2.12). *Then we have*

(1) $\vec{G}_{\Delta t^n}(u) \le \vec{G}_{\Delta t^n}(v)$ *for* $u, v \in l^\infty(G_h)$, $u \le v$;

(2) $\vec{G}_{\Delta t^n}(v+\lambda) = \vec{G}_{\Delta t^n}(v) + \lambda$ *for* $u \in l^\infty(G_h)$, $\lambda \in \mathbb{R}$;

(3) $\|\vec{G}_{\Delta t^n}(u) - \vec{G}_{\Delta t^n}(v)\|_\infty \le \|u - v\|_\infty$ *for all* $u, v \in l^\infty(G_h)$;

(4) $\|v^{n+m} - v^n\|_\infty \le K(t^{n+m} - t^n)$ *for* $m, n \ge 0$, *where K is defined as* (2.2.13), *for a given initial Lipschitz continuous function v^0 with L_0 as a Lipschitz constant, provided that the Lipschitz constant of v^n is preserved by the operator $\vec{G}_{\Delta t^n}$ at each time step t^n if v^n is Lipschitz continuous.*

3. The error estimate for the scheme.

3.1. The main result. The main result of this section gives an upper bound for the following seminorms:

$$\begin{aligned} |\, w - v \,|_{-,Q_h} &= \sup_{(x,t)\in Q_h} (\, w(x,t) - v(x,t)\,)^+, \\ |\, w - v \,|_{+,Q_h} &= \sup_{(x,t)\in Q_h} (\, v(x,t) - w(x,t)\,)^+, \end{aligned}$$

where $w^+ \equiv \max\{0, w\}$.

In addition, we need to introduce the quantity

$$\omega_\epsilon(w; y, t, p_x) = w(y + \epsilon_x\, p_x, t) - w(y, 0) - \frac{t^2}{2\,\epsilon_t} - \frac{\epsilon_x}{2}\,|\, p_x\,|^2, \tag{3.3.1}$$

which can be bounded in terms of the moduli of continuity of w. For example, if $w(t=0) \in W^{1,\infty}(\mathbb{R}^d)$ and $w_t \in L^\infty(0, T; \mathbb{R}^d)$ we easily get

$$|\,\omega_\epsilon(w; y, t, p_x)\,| \le \frac{|\,w_t\,|^2_{L^\infty(0,T;\mathbb{R}^d)}}{2}\epsilon_t + \frac{|\,w(0)\,|_{W^{1,\infty}(\mathbb{R}^d)}}{2}\epsilon_x. \tag{3.3.2}$$

With the notation introduced above, we have the following result.

THEOREM 3.1 (a priori error estimate). *Assume that H is continuous in $\mathbb{R}^d$ and u_0 is Lipschitz continuous with L_0 as a Lipschitz constant. Let u be the viscosity solution of the equation* (2.2.1), (2.2.2) *and let v be the approximate solution given by the scheme* (2.2.3), (2.2.8), *and* (2.2.9) *under the local CFL condition* (2.2.10) *and* (2.2.11). *In addition, the assumption for Proposition* 2.3(4) *holds. Define $h = \max_{y\in G_h} \max_{i=1,\dots,N_y} |\,\delta_i\,|$ and $\Delta t = \max_{0\le i\le N_T} \Delta t^i$. Let $\frac{h}{\Delta t}$ be fixed. Then we have*

$$\|\,u - v\,\|_{L^\infty(Q_h)} \le C_3\sqrt{\Delta t},$$

where C_3 is a constant depending only on u_0, L_0, T, and $\widehat{H}_y$ as well as the solution u.

Note that the above L$^\infty$-norm is implied by

$$\|\,u - v\,\|_{L^\infty(Q_h)} = \max\{|\,u - v\,|_{+,Q_h}, |\,u - v\,|_{-,Q_h}\}.$$

THEOREM 3.2 (a priori error estimate for smooth solutions). *Assume that H is smooth in $\mathbb{R}^d$ and u_0 is Lipschitz continuous with L_0 as a Lipschitz constant. Let u be the smooth viscosity solution of the equation* (2.2.1), (2.2.2) *and let v be the approximate solution given by the scheme* (2.2.3), (2.2.8), *and* (2.2.9) *with the CFL condition* (2.2.10) *and* (2.2.11). *Define $h = \max_{y\in G_h} \max_{i=1,\dots,N_y} |\,\delta_i\,|$ and $\Delta t = \max_{0\le i\le N_T} \Delta t^i$. Let $\frac{h}{\Delta t}$ be fixed. Then we have*

$$\|\,u - v\,\|_{L^\infty(Q_h)} \le C_4\Delta t,$$

where C_4 is a constant depending only on u_0, L_0, T, and $\widehat{H}_y$ as well as the second-order derivatives of the solution u.

Remark. The technique used to prove Theorem 3.1 and Theorem 3.2 is related to the one used to prove a posteriori error estimates [2, 12]. One may also use such a technique to establish local a posterior error estimates for Hamilton–Jacobi equations. Analogous ideas can be used in the context of hyperbolic conservation laws. For example, borrowing some continuous dependence results from nonlinear conservation laws, Gosse and Makridakis [17] have established such local a posterior error estimates in the context of one-dimensional scalar hyperbolic conservation laws.

3.2. Proof of Theorem 3.1.

We prove the above result for $\sigma = -$; the proof for $\sigma = +$ is entirely analogous.

We attempt to obtain an estimate of the quantity

$$\Delta = |\,u - v\,|_{-,Q_h},$$

which we assume to be strictly positive since otherwise there is nothing to prove. We introduce the auxiliary function

$$\psi(x,y,t,s) = u(x,t) - v(y,s) - (1-\theta)\frac{(t+s)}{2T}\Delta - \frac{(t-s)^2}{2\epsilon_t} - \frac{|x-y|^2}{2\epsilon_x},$$

where the parameter θ belongs to $(0,1)$ and ϵ_x, ϵ_t to $(0,\infty)$; there is a point $(\hat{x},\hat{y},\hat{t},\hat{s})$: $(\hat{x},\hat{t}) \in \mathbb{R}^d \times [0,T]$ and $(\hat{y},\hat{s}) \in Q_h$ such that

$$\psi(\hat{x},\hat{y},\hat{t},\hat{s}) \geq \psi(x,y,t,s) \qquad \forall\,(x,t) \in \mathbb{R}^d \times [0,T] \text{ and } \forall\,(y,s) \in Q_h.$$

The existence of such a point easily follows from the fact that both u and v are continuous and periodic in space with the same period.

To obtain the desired estimate, we proceed in four steps.

Step 1. The case $\hat{s} > 0$ and $\hat{t} > 0$. By the construction of the function ψ, we have that $(\hat{x},\hat{t})$ is a maximum point on $\mathbb{R}^d \times (0,T]$ for

$$(x,t) \to u(x,t) - v(\hat{y},\hat{s}) - (1-\theta)\frac{(t+\hat{s})}{2T}\Delta - \frac{(t-\hat{s})^2}{2\epsilon_t} - \frac{|x-\hat{y}|^2}{2\epsilon_x};$$

thus $(\hat{p_t}, \hat{p_x}) \in D^+u(\hat{x},\hat{t})$, where

$$\hat{p}_t = \frac{\hat{t}-\hat{s}}{\epsilon_t} + \frac{1-\theta}{2T}\Delta \quad \text{ and } \quad \hat{p}_x = \frac{\hat{x}-\hat{y}}{\epsilon_x}.$$

By the assumption, u being the viscosity solution gives us that

$$\hat{p}_t + H(\hat{p}_x) \leq 0. \tag{3.3.3}$$

On the other hand, we note that $(\hat{y},\hat{s})$ is a maximum point on Q_h for

$$(y,s) \to u(\hat{x},\hat{t}) - v(y,s) - (1-\theta)\frac{(\hat{t}+s)}{2T}\Delta - \frac{(\hat{t}-s)^2}{2\epsilon_t} - \frac{|\hat{x}-y|^2}{2\epsilon_x};$$

therefore, it follows that

$$\begin{aligned} v(y,s) &\geq v(\hat{y},\hat{s}) + \left(\frac{(\hat{t}-\hat{s})}{\epsilon_t} - \frac{(1-\theta)\Delta}{2T}\right)(s-\hat{s}) - \frac{(s-\hat{s})^2}{2\,\epsilon_t} \\ &\quad + (y-\hat{y})\cdot\frac{(\hat{x}-\hat{y})}{\epsilon_x} - \frac{|y-\hat{y}|^2}{2\,\epsilon_x} \\ &=: V(y,s) \qquad \forall\,(y,s) \in Q_h; \end{aligned}$$

furthermore, the above inequality can be rewritten as

$$\begin{aligned} v(y,s) &\geq v(\hat{y},\hat{s}) - \frac{(1-\theta)\Delta}{T}(s-\hat{s}) + \hat{p}_t(s-\hat{s}) - \frac{(s-\hat{s})^2}{2\,\epsilon_t} \\ &\quad + (y-\hat{y})\cdot\hat{p}_x - \frac{|y-\hat{y}|^2}{2\,\epsilon_x} \\ &= V(y,s) \qquad \forall\,(y,s) \in Q_h. \end{aligned}$$

In the current case, for some $n_0 \geq 0$ and j_0, we have $\hat{s} = t^{n_0+1}$ and $\hat{y} = y_{j_0}$; thus, taking $s = t^{n_0}$ and $y = y_j$ in the above inequality leads to

$$\begin{aligned} v_j^{n_0} &\geq v_{j_0}^{n_0+1} + \frac{(1-\theta)\Delta}{T}\Delta t^{n_0} - \hat{p}_t \Delta t^{n_0} - \frac{(\Delta t^{n_0})^2}{2\,\epsilon_t} \\ &\quad + (y_j - \hat{y}) \cdot \hat{p}_x - \frac{|y_j - \hat{y}|^2}{2\,\epsilon_x} \\ &= V_j^{n_0} \qquad \forall\, y_j \in G_h. \end{aligned}$$

Next we use the scheme (2.2.8) and (2.2.9) to propagate the above relation from level $s = t^{n_0}$ to level $s = t^{n_0+1}$. To do this, we shall make use of the boundedness of the stencil size of the scheme.

Let $\mathcal{D}_{j_0}^{n_0+\eta_l}$ denote at time level $t^{n_0+\eta_l}$ the set of points $y_j \in Q_h$ involved in computing $v_{j_0}^{n_0+1}$ from time level t^{n_0} to t^{n_0+1} by the predictor-corrector scheme (2.2.8), (2.2.9). Therefore,

$$\mathcal{D}_{j_0}^{n_0+\eta_M} = \{y_{j_0}\} \quad \text{ and } \quad \mathcal{D}_{j_0}^{n_0+\eta_{M-1}} = \{y_{j_0}, y_{j_0,1}, \ldots, y_{j_0,N_{j_0}}\};$$

then $\mathcal{D}_{j_0}^{n_0+\eta_{M-2}}$ consists of all the points needed in evaluating $v_{j_0}^{n_0+\eta_{M-1}}$ and $v_{j_0,i}^{n_0+\eta_{M-1}}$ for $i = 1, \ldots, N_{j_0}$, and so on. Because M is finite, every $\mathcal{D}_{j_0}^{n_0+\eta_l}$ for $l = 0, \ldots, M$ is a finite set and

$$\mathcal{D}_{j_0}^{n_0+\eta_M} \subset \mathcal{D}_{j_0}^{n_0+\eta_{M-1}} \subset \cdots \subset \mathcal{D}_{j_0}^{n_0}.$$

Now define

$$C_{j_0} h = \max\{|y_j - \hat{y}| = |y_j - y_{j_0}| : y_j \in \mathcal{D}_{j_0}^{n_0+\eta_1}\}, \tag{3.3.4}$$

where C_{j_0} is a fixed positive constant and h is the mesh size.

Next we estimate $v_j^{n_0+\eta_1}$ for all $y_j \in Q_h$. By the predictor (2.2.8), it follows for all $y_j \in Q_h$,

$$\begin{aligned} v_j^{n_0+\eta_1} &= v_j^{n_0} - \eta_1 \Delta t^{n_0} \widehat{H}_j(\partial_{\delta_j} v_j^{n_0}) \\ &\geq V_j^{n_0} - \eta_1 \Delta t^{n_0} \widehat{H}_j(\partial_{\delta_j} V_j^{n_0}) \\ &= V_j^{n_0} - \eta_1 \Delta t^{n_0} \widehat{H}_j\left(\frac{V_j^{n_0} - V_{j,1}^{n_0}}{|\,\delta_{j,1}\,|}, \ldots, \frac{V_j^{n_0} - V_{j,N_j}^{n_0}}{|\,\delta_{j,N_j}\,|}\right) \end{aligned}$$

by the monotonicity of the scheme which is guaranteed by the local CFL condition (2.2.10), (2.2.11) and the relation that $v_j^{n_0} \geq V_j^{n_0}$, for all $y_j \in G_h$, shown above. Since for $i = 1, \ldots, N_j$,

$$\frac{V_j^{n_0} - V_{j,i}^{n_0}}{|\,\delta_{j,i}\,|} = \frac{\delta_{j,i}}{|\,\delta_{j,i}\,|} \cdot \hat{p}_x + \frac{\delta_{j,i}}{|\,\delta_{j,i}\,|} \cdot \frac{\delta_{j,i} - 2(y_j - \hat{y})}{2\,\epsilon_x},$$

we obtain by using the Lipschitz continuity of the Hamiltonian $\widehat{H}_j$ that

$$\begin{aligned}
v_j^{n_0+\eta_1} &\geq V_j^{n_0} - \eta_1 \Delta t^{n_0} \widehat{H}_j \left(\frac{\delta_{j,1}}{|\,\delta_{j,1}\,|} \cdot \hat{p}_x, \dots, \frac{\delta_{j,N_j}}{|\,\delta_{j,N_j}\,|} \cdot \hat{p}_x \right) \\
&\quad - \frac{L\eta_1 \Delta t^{n_0}}{2\,\epsilon_x} \max_{1\leq i\leq N_j} \left| \frac{\delta_{j,i}}{|\,\delta_{j,i}\,|} \cdot (\delta_{j,i} - 2(y_j - \hat{y})) \right| \\
&\geq V_j^{n_0} - \eta_1 \Delta t^{n_0} H(\hat{p}_x) - \frac{L\eta_1 \Delta t^{n_0}}{2\,\epsilon_x} \max_{1\leq i\leq N_j} |\delta_{j,i} - 2(y_j - \hat{y})| \\
&\geq V_j^{n_0} - \eta_1 \Delta t^{n_0} H(\hat{p}_x) - \frac{L\eta_1 \Delta t^{n_0}}{2\,\epsilon_x} \max_{1\leq i\leq N_j} |\delta_{j,i}\,| - \frac{L\eta_1 \Delta t^{n_0}}{\epsilon_x} |y_j - \hat{y}| \\
&\geq V_j^{n_0} - \eta_1 \Delta t^{n_0} H(\hat{p}_x) - \frac{L\,h\eta_1 \Delta t^{n_0}}{2\,\epsilon_x} - \frac{L\eta_1 \Delta t^{n_0}}{\epsilon_x} |y_j - \hat{y}|.
\end{aligned} \tag{3.3.5}$$

In the above derivation, we used the consistency of numerical Hamiltonian in the following way: setting $\mathcal{L}u(x,t) = u(y_j, t^{n_0}) + \hat{p}_x \cdot (x - y_j) + p_t(t - t^{n_0})$, we have

$$H(\hat{p}_x) = \widehat{H}_j(\partial_{\delta_j} \mathcal{L}u(y_j, t^{n_0})) \qquad \text{and} \qquad \partial_{\delta_{j,i}} \mathcal{L}u(y_j, t^{n_0})) = \hat{p}_x \cdot \frac{\delta_{j,i}}{|\,\delta_{j,i}\,|}.$$

Hence, for $j \in \mathcal{D}_{j_0}^{n_0+\eta_1}$, invoking (3.3.4) we find from (3.3.5) that

$$\begin{aligned}
v_j^{n_0+\eta_1} &\geq V_j^{n_0} - \eta_1 \Delta t^{n_0} H(\hat{p}_x) - \frac{\eta_1\, L\, h \Delta t^{n_0}}{2\,\epsilon_x} - \frac{\eta_1\, L \Delta t^{n_0}}{\epsilon_x} |y_j - \hat{y}| \\
&\geq V_j^{n_0} - \eta_1 \Delta t^{n_0} H(\hat{p}_x) - \frac{\eta_1\, L\, h \Delta t^{n_0}}{2\,\epsilon_x} - \frac{\eta_1\, L\, C_{j_0}\, h \Delta t^{n_0}}{\epsilon_x} \\
&= V_j^{n_0} - \eta_1 \Delta t^{n_0} H(\hat{p}_x) - \frac{\eta_1\, L\, h \Delta t^{n_0}(1 + 2\,C_{j_0})}{2\,\epsilon_x} \\
&=: V_j^{n_0+\eta_1}.
\end{aligned} \tag{3.3.6}$$

Next we estimate $v_j^{n_0+\eta_2}$. Apparently, for $j \in \mathcal{C}^{n_0}$ and $j \in \mathcal{D}_{j_0}^{n_0+\eta_2}$, by the predictor (2.2.8), we have that

$$\begin{aligned}
v_j^{n_0+\eta_2} &= v_j^{n_0} - \eta_2 \Delta t^{n_0} \widehat{H}_j(\partial_{\delta_j} v_j^{n_0}) \\
&\geq V_j^{n_0} - \eta_2 \Delta t^{n_0} \widehat{H}_j(\partial_{\delta_j} V_j^{n_0}) \\
&\geq V_j^{n_0} - \eta_2 \Delta t^{n_0} H(\hat{p}_x) - \frac{L\,h\eta_2 \Delta t^{n_0}}{2\,\epsilon_x} - \frac{L\eta_2 \Delta t^{n_0}}{\epsilon_x} |y_j - \hat{y}| \\
&\geq V_j^{n_0} - \eta_2 \Delta t^{n_0} H(\hat{p}_x) - \frac{\eta_2\, L\, h \Delta t^{n_0}(1 + 2\,C_{j_0})}{2\,\epsilon_x} \\
&=: V_j^{n_0+\eta_2}.
\end{aligned}$$

For $j \notin \mathcal{C}^{n_0}$ and $j \in \mathcal{D}_{j_0}^{n_0+\eta_2}$, still by the predictor (2.2.8) the inequality (3.3.6), and the monotonicity of the scheme, we get

$$\begin{aligned}
v_j^{n_0+\eta_2} &= v_j^{n_0+\eta_1} - \sigma_2 \Delta t^{n_0} \widehat{H}_j(\partial_{\delta_j} v_j^{n_0+\eta_1}) \\
&\geq V_j^{n_0+\eta_1} - \sigma_2 \Delta t^{n_0} \widehat{H}_j(\partial_{\delta_j} V_j^{n_0+\eta_1}) \\
&= V_j^{n_0+\eta_1} - \sigma_2 \Delta t^{n_0} \widehat{H}_j \left(\frac{V_j^{n_0+\eta_1} - V_{j,1}^{n_0+\eta_1}}{|\,\delta_{j,1}\,|}, \dots, \frac{V_j^{n_0+\eta_1} - V_{j,N_j}^{n_0+\eta_1}}{|\,\delta_{j,N_j}\,|} \right).
\end{aligned}$$

Since for $i = 1, \ldots, N_j$,

$$\frac{V_j^{n_0+\eta_1} - V_{j,i}^{n_0+\eta_1}}{|\,\delta_{j,i}\,|} = \frac{V_j^{n_0} - V_{j,i}^{n_0}}{|\,\delta_{j,i}\,|};$$

moreover, by the smoothness and consistency of the numerical Hamiltonian, we get

$$\begin{aligned} v_j^{n_0+\eta_2} &\geq V_j^{n_0+\eta_1} - \sigma_2 \Delta t^{n_0} \widehat{H}_j \left(\tfrac{V_j^{n_0} - V_{j,1}^{n_0}}{|\,\delta_{j,1}\,|}, \ldots, \tfrac{V_j^{n_0} - V_{j,N_j}^{n_0}}{|\,\delta_{j,N_j}\,|} \right) \\ &\geq V_j^{n_0+\eta_1} - \sigma_2 \Delta t^{n_0} \left(H(\hat{p}_x) + \tfrac{L\,h}{2\,\epsilon_x} + \tfrac{L}{\epsilon_x} |y_j - \hat{y}| \right) \\ &= V_j^{n_0} - \eta_2 \Delta t^{n_0} H\,(\hat{p}_x) - \tfrac{\eta_1\,L\,h \Delta t^{n_0}(1+2\,C_{j_0})}{2\,\epsilon_x} \\ &\quad - \tfrac{\sigma_2\,L\,h\Delta t^{n_0}}{2\,\epsilon_x} - \tfrac{\sigma_2\,L \Delta t^{n_0}}{\epsilon_x} |y_j - \hat{y}| \\ &\geq V_j^{n_0} - \eta_2 \Delta t^{n_0} H\,(\hat{p}_x) - \tfrac{\eta_1\,L\,h \Delta t^{n_0}(1+2\,C_{j_0})}{2\,\epsilon_x} \\ &\quad - \tfrac{\sigma_2\,L\,h\Delta t^{n_0}}{2\,\epsilon_x} - \tfrac{\sigma_2\,L \Delta t^{n_0} C_{j_0}\,h}{\epsilon_x} \\ &= V_j^{n_0} - \eta_2 \Delta t^{n_0} H\,(\hat{p}_x) - \tfrac{\eta_2\,L\,h \Delta t^{n_0}(1+2\,C_{j_0})}{2\,\epsilon_x} \\ &=: V_j^{n_0+\eta_2}, \qquad j \in \mathcal{D}_{j_0}^{n_0+\eta_2}. \end{aligned}$$

Inductively, we get for $j \in \mathcal{D}_{j_0}^{n_0+\eta_k}$, whether $j \in \mathcal{C}^n$ or not,

$$\begin{aligned} v_j^{n_0+\eta_k} &\geq V_j^{n_0} - \eta_k \Delta t^{n_0} H\,(\hat{p}_x) - \frac{\eta_k\,L\,h \Delta t^{n_0}(1+2\,C_{j_0})}{2\,\epsilon_x} \\ &=: V_j^{n_0+\eta_k}, \qquad k = 1, \ldots, M. \end{aligned}$$

Therefore, taking $k = M$ in the above: for $j \in \mathcal{D}_{j_0}^{n_0+1}$,

$$v_j^{n_0+1} \geq V_j^{n_0} - \Delta t^{n_0} H\,(\hat{p}_x) - \frac{L\,h \Delta t^{n_0}(1+2\,C_{j_0})}{2\,\epsilon_x}; \tag{3.3.7}$$

that is, for $j = j_0$ in the inequality (3.3.7), we have

$$\begin{aligned} v_{j_0}^{n_0+1} &\geq V_{j_0}^{n_0} - \Delta t^{n_0} H\,(\hat{p}_x) - \frac{L\,h \Delta t^{n_0}(1+2\,C_{j_0})}{2\,\epsilon_x} \\ &= v_{j_0}^{n_0+1} + \frac{(1-\theta)\Delta}{T} \Delta t^{n_0} - \hat{p}_t \Delta t^{n_0} - \frac{(\Delta t^{n_0})^2}{2\,\epsilon_t} - \Delta t^{n_0} H\,(\hat{p}_x) \\ &\quad - \frac{L\,h \Delta t^{n_0}(1+2\,C_{j_0})}{2\,\epsilon_x}; \end{aligned}$$

hence

$$\begin{aligned} \frac{(1-\theta)\Delta}{T} &\leq \hat{p}_t + H\,(\hat{p}_x) + \frac{\Delta t^{n_0}}{2\,\epsilon_t} + \frac{L\,h(1+2\,C_{j_0})}{2\,\epsilon_x} \\ &\leq \frac{\Delta t^{n_0}}{2\,\epsilon_t} + \frac{L\,h(1+2\,C_{j_0})}{2\,\epsilon_x}, \end{aligned} \tag{3.3.8}$$

where we have used the relation (3.3.3).

Step 2. The case $\hat{s} = 0$ and $\hat{t} \geq 0$. In this case, we no longer can use the definition of the approximate solution, so we rely instead on simple algebraic manipulations. We proceed as follows. Let $(\underline{y}, \underline{t}) \in Q_h$ be such that $|\,u - v\,|_{-,Q_h} = u(\underline{y}, \underline{t}) - v(\underline{y}, \underline{t})$. Then

$$\psi(\underline{y}, \underline{y}, \underline{t}, \underline{t}) = |\,u - v\,|_{-,Q_h} - (1-\theta)\frac{\underline{t}}{T}\Delta.$$

On the other hand,

$$\begin{aligned}\psi(\underline{y},\underline{y},\underline{t},\underline{t}) &\le \psi(\hat{x},\hat{y},\hat{t},\hat{s}) = \psi(\hat{x},\hat{y},\hat{t},0)\\ &= u(\hat{x},\hat{t}) - v(\hat{y},0) - \frac{\hat{t}^2}{2\epsilon_t} - \frac{|\hat{x}-\hat{y}|^2}{2\epsilon_x} - (1-\theta)\frac{\hat{t}}{2T}\Delta\\ &\le u(\hat{x},\hat{t}) - u(\hat{y},0) - \frac{\hat{t}^2}{2\epsilon_t} - \frac{|\hat{x}-\hat{y}|^2}{2\epsilon_x} - (1-\theta)\frac{\hat{t}}{2T}\Delta\\ &= \omega_\epsilon(u;\hat{y},\hat{t},\hat{p}_x) - (1-\theta)\frac{\hat{t}}{2T}\Delta\end{aligned}$$

since $\hat{x} = \hat{y} + \epsilon_x\,\hat{p}_x$; see (3.3.1) for definition of ω_ϵ. This implies that

$$\left(1 - (1-\theta)\frac{2\underline{t}-\hat{t}}{2T}\right)\Delta \le \omega_\epsilon(u;\hat{y},\hat{t},\hat{p}_x);$$

therefore by (3.3.2)

$$\Delta \le \frac{1}{\theta}\left(\frac{|\,u_t\,|^2_{L^\infty(0,T;\mathbb{R}^d)}}{2}\epsilon_t + \frac{|\,u(0)\,|_{W^{1,\infty}(\mathbb{R}^d)}}{2}\epsilon_x\right)$$

since

$$(1-\theta)\frac{2\underline{t}-\hat{t}}{2T} \le (1-\theta).$$

Step 3. The case $\hat{s} > 0$ and $\hat{t} = 0$. In this case, we also rely on simple algebraic manipulations. We proceed as follows. Let $(\underline{y},\underline{t}) \in Q_h$ be such that $|\,u - v\,|_{-,Q_h} = u(\underline{y},\underline{t}) - v(\underline{y},\underline{t})$. Then

$$\psi(\underline{y},\underline{y},\underline{t},\underline{t}) = |\,u - v\,|_{-,Q_h} - (1-\theta)\frac{\underline{t}}{T}\Delta.$$

On the other hand,

$$\begin{aligned}\psi(\underline{y},\underline{y},\underline{t},\underline{t}) &\le \psi(\hat{x},\hat{y},\hat{t},\hat{s}) = \psi(\hat{x},\hat{y},0,\hat{s})\\ &= u(\hat{x},0) - v(\hat{y},\hat{s}) - \frac{\hat{s}^2}{2\epsilon_t} - \frac{|\hat{x}-\hat{y}|^2}{2\epsilon_x} - (1-\theta)\frac{\hat{s}}{2T}\Delta\\ &\le u(\hat{x},0) - u(\hat{y},0) + v(\hat{y},0) - v(\hat{y},\hat{s}) - \frac{\hat{s}^2}{2\epsilon_t} - \frac{|\hat{x}-\hat{y}|^2}{2\epsilon_x} - (1-\theta)\frac{\hat{s}}{2T}\Delta\\ &\le \omega_\epsilon(u;\hat{y},0,\hat{p}_x) + K\hat{s} - \frac{\hat{s}^2}{2\epsilon_t} - (1-\theta)\frac{\hat{s}}{2T}\Delta\end{aligned}$$

by $\hat{x} = \hat{y} + \epsilon_x\,\hat{p}_x$ and Proposition 2.3(4).

Now we have to estimate $\hat{s}$ carefully. Since

$$\psi(\hat{x},\hat{y},\hat{t},\hat{s}) = \psi(\hat{x},\hat{y},0,\hat{s}) \ge \psi(\hat{x},\hat{y},0,0),$$

we get

$$v(\hat{y},0) - v(\hat{y},\hat{s}) \ge (1-\theta)\frac{\hat{s}}{2T}\Delta + \frac{\hat{s}^2}{2\epsilon_t}.$$

By Proposition 2.3(4), we get

$$\hat{s} \leq 2K\epsilon_t.$$

This implies that

$$\left(1-(1-\theta)\frac{t}{T}\right)\Delta \leq \omega_\epsilon(u;\hat{y},0,\hat{p}_x)+2K^2\epsilon_t;$$

therefore by (3.3.2)

$$\Delta \leq \frac{1}{\theta}\left(\frac{|\,u(0)\,|_{W^{1,\infty}(\mathbb{R}^d)}}{2}\epsilon_x+2K^2\epsilon_t\right),$$

since

$$(1-\theta)\frac{t}{T} \leq (1-\theta).$$

Step 4. Conclusion. Putting together the above inequalities and the bound (3.3.8), we get

$$\Delta \leq \max\left\{\frac{A+C}{\theta},\frac{B}{1-\theta}\right\},$$

where

$$A=\frac{|\,u_t\,|^2_{L^\infty(0,T;\mathbb{R}^d)}}{2}\epsilon_t+\frac{|\,u(0)\,|_{W^{1,\infty}(\mathbb{R}^d)}}{2}\epsilon_x,$$

$$B=T\left(\frac{\Delta t^{n_0}}{2\,\epsilon_t}+\frac{L\,h(1+2\,C_{j_0})}{2\,\epsilon_x}\right),$$

$$C=\frac{|\,u(0)\,|_{W^{1,\infty}(\mathbb{R}^d)}}{2}\epsilon_x+2K^2\epsilon_t.$$

Hence, we obtain that

$$\Delta \leq A+C+B$$

by taking the limit when θ tends to $\frac{A+C}{A+B+C}\in[0,1]$.

Note that according to classical results on the viscosity solution, L_0 is also a Lipschitz constant for u; see [15]. The result now follows from the above inequality by minimizing the right-hand side with respect to ϵ. This completes the proof of Theorem 3.1. □

3.3. Proof of Theorem 3.2 when *u* is the smooth solution. Consider $\sigma=-$ only; the proof for $\sigma=+$ is similar.

Let $(\underline{y},\underline{t})\in Q_h$ be such that $|\,u-v\,|_{-,Q_h}=u(\underline{y},\underline{t})-v(\underline{y},\underline{t})$. To estimate $u(\underline{y},\underline{t})-v(\underline{y},\underline{t})$ directly, we compute the local truncation error of the scheme (2.2.3), (2.2.8), (2.2.9). To simplify the presentation, we concentrate on the scheme (2.2.3), (2.2.4) only, since a similar analysis applies to (2.2.3), (2.2.8), (2.2.9).

If u is the smooth solution of (2.2.2), (2.2.1), then

$$\begin{aligned}u(y,t^n+\Delta t^n) &= u(y,t^n)+\Delta t^n u_t+\frac{1}{2}(\Delta t^n)^2 u_{tt}+o((\Delta t)^3)\\ &= u(y,t^n)-\Delta t^n H(\nabla u)+\frac{1}{2}(\Delta t^n)^2(\nabla_p H)^T(\partial^2_{i,j}\, u)(\nabla_p H)+o((\Delta t)^3),\end{aligned}$$

where $\partial^2_{i,j}\, u$ is the Hessian matrix of u.

Substituting the true solution into the scheme (2.2.4) and using the Taylor series expansion of the numerical scheme, we obtain

$$\begin{aligned}\mathbb{G}_y &= \mathbb{G}_y(u(y,t^n);u(y-\delta_{y,1},t^n),\ldots,u(y-\delta_{y,N_y},t^n))\\ &= u(y,t^n)-\Delta t^n\widehat{H}_y(\partial_{\delta_y}u(y,t^n))\\ &= u(y,t^n)-\Delta t^n\widehat{H}_y\left(\nabla u\cdot\frac{\delta_{y,1}}{|\delta_{y,1}|},\ldots,\nabla u\cdot\frac{\delta_{y,N_y}}{|\delta_{y,N_y}|}\right)\\ &\quad-\frac{1}{2}(\Delta t^n)^2\sum_{i=1}^{i=N_y}\frac{|\delta_{y,i}|}{\Delta t^n}\frac{\partial\widehat{H}_y}{\partial p_i}\left(\frac{\delta_{y,i}}{|\delta_{y,i}|}\right)^T(\partial^2_{l,m}\, u)\left(\frac{\delta_{y,i}}{|\delta_{y,i}|}\right)+o((\Delta t)^3)\\ &= u(y,t^n)-\Delta t^n H(\nabla u)\\ &\quad-\frac{1}{2}(\Delta t^n)^2\sum_{i=1}^{i=N_y}\frac{|\delta_{y,i}|}{\Delta t^n}\frac{\partial\widehat{H}_y}{\partial p_i}\left(\frac{\delta_{y,i}}{|\delta_{y,i}|}\right)^T(\partial^2_{l,m}\, u)\left(\frac{\delta_{y,i}}{|\delta_{y,i}|}\right)+o((\Delta t)^3),\end{aligned}$$

where we have used the consistency of the numerical Hamiltonian.

Consequently, the local truncation error is

$$\begin{aligned}&u(y,t^n+\Delta t^n)-\mathbb{G}_y(u(y,t^n);u(y-\delta_{y,1},t^n),\ldots,u(y-\delta_{y,N_y},t^n))\\ &=\ \frac{1}{2}(\Delta t^n)^2(\nabla_p H)^T(\partial^2_{i,j}\, u)(\nabla_p H)\\ &\quad+\frac{1}{2}(\Delta t^n)^2\sum_{i=1}^{i=N_y}\frac{|\delta_{y,i}|}{\Delta t^n}\frac{\partial\widehat{H}_y}{\partial p_i}\left(\frac{\delta_{y,i}}{|\delta_{y,i}|}\right)^T(\partial^2_{l,m}\, u)\left(\frac{\delta_{y,i}}{|\delta_{y,i}|}\right)+o((\Delta t)^3).\end{aligned}$$

Since $\frac{\partial\widehat{H}_y}{\partial p_i}\geq 0$ by the monotonicity of the numerical Hamiltonian, we have that the first two terms on the right-hand side have the same sign, and they cannot be identical to zero unless the solution is a linear function. This essentially shows that the monotone finite-difference schemes for Hamilton–Jacobi equations have only first-order accuracy; see [18] for similar results for hyperbolic conservation laws.

Next we estimate $|\, u-v\,|_{-,Q_h}=u(\underline{y},\underline{t})-v(\underline{y},\underline{t})$.

If $\underline{t}=0$, then we are done. So assume $\underline{y}=y_{j_0}$ for some j_0 and $\underline{t}=t^{n_0+1}$ for some integer $n_0\geq 0$. Define $\mathcal{D}^n_{j_0}$ to be the set of points $y_j\in Q_h$ at the time level $t=t^n$ involved in computing $v(y_{j_0},t^{n_0+1})$ by marching the scheme (2.2.4) from $t=0$ to $t=t^{n_0+1}$. Apparently,

$$\mathcal{D}^{n_0}_{j_0}\subset\mathcal{D}^{n_0-1}_{j_0}\subset\cdots\subset\mathcal{D}^0_{j_0};$$

$\mathcal{D}^0_{j_0}$ is a finite set according to the boundedness of the stencil size of the scheme. Therefore, using the above local truncation error analysis, we have the following

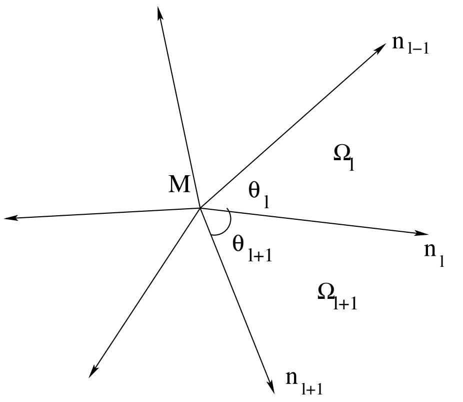


Fig. 2. *Triangulation for the intrinsic monotone scheme.*

estimates for computed solutions:

$$\begin{aligned}
&v(y_j,0)=u(y_j,0) \qquad \text{for } j\in\mathcal{D}^0_{j_0},\\
&v(y_j,t^1)=u(y_j,t^1)+o((\Delta t)^2) \qquad \text{for } j\in\mathcal{D}^1_{j_0},\\
&\dots\\
&v(y_j,t^{n_0})=u(y_j,t^{n_0})+o(n_0(\Delta t)^2) \qquad \text{for } j\in\mathcal{D}^{n_0}_{j_0}.
\end{aligned}$$

Finally, we obtain that

$$|\,u-v\,|_{-,Q_h}=v(y_{j_0},t^{n_0+1})-u(y_{j_0},t^{n_0+1})\le C_4\Delta t,$$

where C_4 depends on the initial condition u_0, T, H, and $\widehat{H}_y$ as well as second-order derivatives of the solution u.

This completes the proof. □

4. Examples of monotone schemes. In this section, we display several examples of monotone schemes for which our results hold.

Example 4.1 (The monotone schemes of [15, 31]). In a way similar to that used to deal with the Lax–Friedrich scheme, the monotone schemes considered in [15, 31] can be recast in our framework and proven to have the desired properties.

Example 4.2 (Abgrall's intrinsic monotone scheme [1]). To describe this scheme, we need to introduce some notation. Let $\mathcal{T}$ be a triangulation of $\mathbb{R}^2$. We denote the triangles by T and their vertices by M_i; the grid G_h is the collection of the vertices M_i. To each vertex M_i we associate a family of angular sectors $\{\Omega^i_l\}_{l=1}^{N_i}$, defined as the inner angles at M_i of all the triangles, T^i_l, having M_i as a vertex. Denote by θ^i_l the angle of Ω^i_l and by n^i_l the unit vector of the half-line $D^i_l=\Omega^i_l\cap\Omega^i_{l+1}$ pointing outward; see Figure 2.

For any set $\{v_i\}$, we denote by v the piecewise-linear function on $\mathcal{T}$ such that $v(M_i)=v_i$; note that $\nabla v|_{T^i_l}=\nabla v_{T^i_l}$ is constant in the triangle T^i_l. The intrinsic monotone scheme for (2.2.1) at a generic point M_i reads

$$v_i^{n+1} = v_i^n - \Delta t^n \left(H\left(\frac{1}{2\pi} \sum_{l=1}^{N_i} \theta_l^i \nabla v_{T_l^i} \right) - \omega \sum_{l=1}^{N_i} \beta_l^i \nabla v_{T_l^i} \cdot n_l^i \right),$$

where

$$\omega = \frac{\eta C(L)}{\pi}, \qquad \beta_l^i = \tan\left(\frac{\theta_l^i}{2}\right) + \tan\left(\frac{\theta_{l+1}^i}{2}\right).$$

Here η is a suitably chosen positive parameter, and $C(L)$ is the Lipschitz constant for H in the ball $B_L = \{p : \|p\| \le L\}$; we denote by $\|\cdot\|$ the usual Euclidean norm.

We have N_i directional derivatives at M_i:

$$p_l^i = -\nabla v_{T_l^i} \cdot n_l^i = -\nabla v_{T_{l+1}^i} \cdot n_l^i, \qquad l = 1, \dots, N_i.$$

Furthermore,

$$\nabla v_{T_l^i} = \frac{1}{\sin^2 \theta_l^i} \left((p_l^i \cos\theta_l^i - p_{l-1}^i) n_{l-1}^i + (p_{l-1}^i \cos\theta_l^i - p_l^i) n_l^i \right);$$

thus the numerical Hamiltonian can be expressed as

$$\widehat{H}_i \left(p_1^i, \dots, p_{N_i}^i\right) = H\left(\frac{1}{2\pi} \sum_{l=1}^{N_i} \theta_l^i \nabla v_{T_l^i} \right) + \omega \sum_{l=1}^{N_i} \beta_l^i p_l^i.$$

The consistency of $\widehat{H}_i$ holds since

$$\sum_{l=1}^{N_i} \beta_l^i p \cdot n_l^i = 0$$

for all $\nabla v = p \in \mathbb{R}^2$ [1].

Differentiating the above $\widehat{H}$ with respect to p_j^i, we have that

$$\begin{aligned} \frac{\partial \widehat{H}_i}{\partial p_j^i} &= \frac{1}{2\pi} \frac{\theta_j^i}{\sin^2 \theta_j^i} \nabla H \cdot (\cos\theta_j^i n_{j-1}^i - n_j^i) \\ &\quad + \frac{1}{2\pi} \frac{\theta_{j+1}^i}{\sin^2 \theta_{j+1}^i} \nabla H \cdot (\cos\theta_{j+1}^i n_{j+1}^i - n_j^i) + \omega\beta_j \ge 0 \end{aligned}$$

if η is chosen as follows:

$$\eta = \frac{1}{4\sin^2 \frac{\alpha}{2}} \max(\alpha, \sin\alpha + (\pi - \alpha)\cos\alpha),$$

where α denotes the smallest angle of the triangles T of the triangulation $\mathcal{T}$. Therefore, $\widehat{H}_i$ is nondecreasing in each of its arguments.

The smoothness of $\widehat{H}_i$ follows by the local smoothness of H.

Example 4.3 (The covolume scheme of [22]). Following [22], we consider a regular triangulation $\mathcal{T}_h$ of $\mathbb{R}^2$ and define the grid G_h to be the collection of vertices M_i of triangles in $\mathcal{T}_h$. Next, we construct a dual mesh by joining the circumcenters of

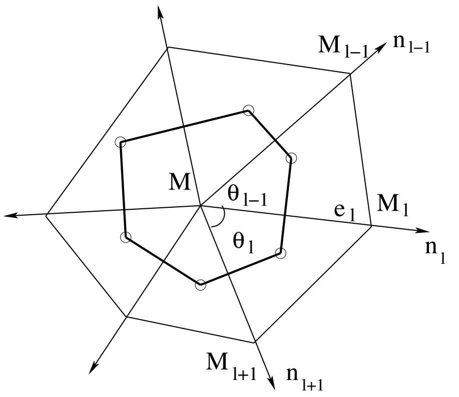


FIG. 3. *Triangulation for the co-volume scheme.*

the triangles in $\mathcal{T}_h$ and denote by V_i the covolume bounded by all the edges of the dual mesh that are perpendicular to edges containing the vertex M_i. Denote by M_l^i, $1 \le l \le N_i$, the vertices linked to M_i by an edge; we enumerate them in the clockwise direction. Denote by e_l^i the line segment joining M_i and M_l^i and by $e_{l,\perp}^i$ the edge of V_i that intersects perpendicularly e_l^i. Moreover, the triangle with vertices M_i, M_l^i, and M_{l+1}^i is denoted by T_l^i, the angle between e_l^i and e_{l+1}^i by θ_l^i, and the unit vector along e_l^i directed towards M_l^i by n_l^i. Finally, let $m(V_i)$ and $m(e_l^i)$ denote the area and length of V_i and e_l^i, respectively; see Figure 3.

Just like in the case of Abgrall's intrinsic monotone scheme, given the set of values $\{v_i\}$, we define a piecewise-linear interpolant v on $\mathcal{T}_h$ by $v(M_i) = v_i$. The covolume scheme for (2.2.1) is

$$v_i^{n+1} = v_i^n - \Delta t^n \left(\widehat{H}_i(\nabla v^n_{T_1^i}, \dots, \nabla v^n_{T^i_{N_i}}) \right),$$

where

$$\widehat{H}_i = \frac{1}{m(V_i)} \sum_{l:T_l^i \cap V_i \neq \emptyset} m(V_i \cap T_l^i) H(\nabla v_{T_l^i}) - \frac{\epsilon_{h,i}}{m(V_i)} \sum_l (\nabla v_{T_l^i}, n_l^i) m(e_{l,\perp}^i).$$

We have N_i directional derivatives at M_i:

$$p_l^i = -\nabla v_{T_{l-1}^i} \cdot n_l^i = -\nabla v_{T_l^i} \cdot n_l^i,$$
$$l = 1, \dots, N_i.$$

Furthermore, we can express $\nabla v_{T_l^i}$ in terms of p_l^i,

$$\nabla v_{T_l^i} = \frac{1}{\sin^2 \theta_l^i} \left((p_{l+1}^i \cos \theta_l^i - p_l^i) n_l^i + (p_l^i \cos \theta_l^i - p_{l+1}^i) n_{l+1}^i \right).$$

$\widehat{H}_i$ is consistent by the fact that

$$\sum_l (p, n_l^i) m(e_{l,\perp}^i) = 0$$

for all $p \in \mathbb{R}^2$ [22].

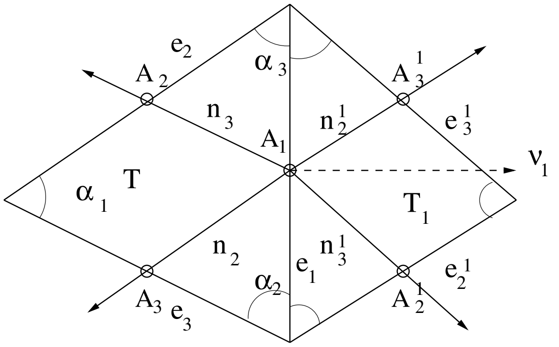


FIG. 4. *Triangulation for the edge-centered scheme.*

Differentiating the above $\widehat{H}_i$ with respect to p^i_j, we have that

$$\begin{aligned}\frac{\partial \widehat{H}_i}{\partial p^i_j} &= \frac{m(V_i \cap T^i_j)}{m(V_i)} \frac{1}{\sin^2 \theta^i_j} \nabla H \cdot (\cos \theta^i_j n^i_{j+1} - n^i_j) \\ &+ \frac{m(V_i \cap T^i_{j-1})}{m(V_i)} \frac{1}{\sin^2 \theta^i_{j-1}} \nabla H \cdot (\cos \theta^i_{j-1} n^i_{j-1} - n^i_j) \\ &+ \frac{\epsilon_{h,i}}{m(V_i)} m(e^i_{l,\perp}) \geq 0\end{aligned}$$

provided that $\epsilon_{h,i}$ is chosen such that

$$\epsilon_{h,i} \geq C^{'} C(L) \max_l (h_{T^i_l}),$$

where $h_{T^i_l}$ is the diameter of the triangle T^i_l and $C^{'}$ is a constant independent of the triangulation. To estimate the constant $C^{'}$, we have used the regularity of the triangulation [22]. The monotonicity of $\widehat{H}_i$ follows.

The smoothness of $\widehat{H}_i$ is implied by the local smoothness of H.

Example 4.4 (The edge-centered schemes of [22]). To describe these schemes, we need to introduce more notation. Consider a regular triangulation $\mathcal{T}_h$ of $\mathbb{R}^2$ with the property that all the inner angles α of the triangles in $\mathcal{T}_h$ satisfy that $\alpha \leq \omega_0 < \frac{\pi}{2}$. Given a triangle T of $\mathcal{T}_h$, we denote by $e_\ell, \ell = 1, 2, 3$ the edges of the triangle and by T_1 the triangle that shares the edge e_1 with T. The midpoints of the edges of T, T_ℓ are denoted by A_ℓ and A^1_ℓ, respectively; $\ell = 1, 2, 3$, are named in the counterclockwise direction. The common midpoints are A_1, A^1_1. The unit normal vector ν_1 to the common edge is directed toward T_1. Denote the inner angles facing e_ℓ by α_ℓ. At the common midpoint A_1, we introduce four outward unit direction vectors n_ℓ, n^1_ℓ such that $n_\ell \parallel e_\ell, n^1_\ell \parallel e^1_\ell, \ell = 2, 3$. See Figure 4.

Define the grid G_h to be the collection of the midpoints of the edges of all the triangles in $\mathcal{T}_h$. The approximating function v_h will lie in the space X_h of nonconforming piecewise linear functions defined on $\mathcal{T}_h$ [22]; namely, v_h is continuous at every $A \in G_h$ and piecewise linear in every triangle $T \in \mathcal{T}_h$.

The edge-centered scheme for (2.2.1) at a generic point A_1 is

$$v^{n+1}_{A_1} = v^n_{A_1} - \Delta t^n \left(\widehat{H}_{A_1}(\nabla v^n_T, \nabla v^n_{T_1}) \right),$$

where

$$\widehat{H}_{A_1} = H\left(\frac{1}{m(T)+m(T_1)}\left(m(T)\nabla v_T + m(T_1)\nabla v_{T_1}\right)\right) - \theta_h^{A_1}(\nabla v_{T_1} - \nabla v_T)\cdot \nu_1.$$

At A_1, we have four possible directions $n_\ell, n_\ell^1, \ell = 2, 3$; thus $N_{A_1} = 4$. Accordingly, we define four quantities:

$$\begin{aligned}
p_1 &= \frac{v_{A_1} - v_{A_2}}{m(\overline{A_1A_2})} = -\nabla v_T \cdot n_3,\\
p_2 &= \frac{v_{A_1} - v_{A_3}}{m(\overline{A_1A_3})} = -\nabla v_T \cdot n_2,\\
p_3 &= \frac{v_{A_1} - v_{A_2^1}}{m(\overline{A_1A_2^1})} = -\nabla v_{T_1} \cdot n_3^1,\\
p_4 &= \frac{v_{A_1} - v_{A_3^1}}{m(\overline{A_1A_3^1})} = -\nabla v_{T_1} \cdot n_2^1.
\end{aligned}$$

Furthermore, we have that

$$\begin{aligned}
\nu_1 &= -\frac{\cos\alpha_3}{\sin\alpha_1}n_3 - \frac{\cos\alpha_2}{\sin\alpha_1}n_2,\\
-\nu_1 = \nu_1^1 &= -\frac{\cos\alpha_3^1}{\sin\alpha_1^1}n_3^1 - \frac{\cos\alpha_2^1}{\sin\alpha_1^1}n_2^1,\\
\nabla v_T &= \frac{1}{\sin^2\alpha_1}\left((p_1\cos\alpha_1 - p_2)n_2 + (p_2\cos\alpha_1 - p_1)n_3\right),\\
\nabla v_{T_1} &= \frac{1}{\sin^2\alpha_1^1}\left((p_3\cos\alpha_1^1 - p_4)n_2^1 + (p_4\cos\alpha_1^1 - p_3)n_3^1\right).
\end{aligned}$$

Therefore, we can express $\widehat{H}_{A_1}$ as

$$\begin{aligned}
\widehat{H}_{A_1}(p_1, p_2, p_3, p_4) = H&\left(\frac{1}{m(T)+m(T_1)}\left(m(T)\nabla v_T + m(T_1)\nabla v_{T_1}\right)\right)\\
&+ \theta_h^{A_1}\left(\frac{\cos\alpha_3}{\sin\alpha_1}p_1 + \frac{\cos\alpha_2}{\sin\alpha_1}p_2 + \frac{\cos\alpha_3^1}{\sin\alpha_1^1}p_3 + \frac{\cos\alpha_2^1}{\sin\alpha_1^1}p_4\right).
\end{aligned}$$

Differentiating the above Hamiltonian with respect to, say, p_1, we have that

$$\frac{\partial\widehat{H}_{A_1}}{\partial p_1} = \frac{m(T)}{m(T)+m(T_1)}\frac{1}{\sin^2\alpha_1}\nabla H\cdot(\cos\alpha_1 n_2 - n_3) + \theta_h\frac{\cos\alpha_3}{\sin\alpha_1} \geq 0$$

if

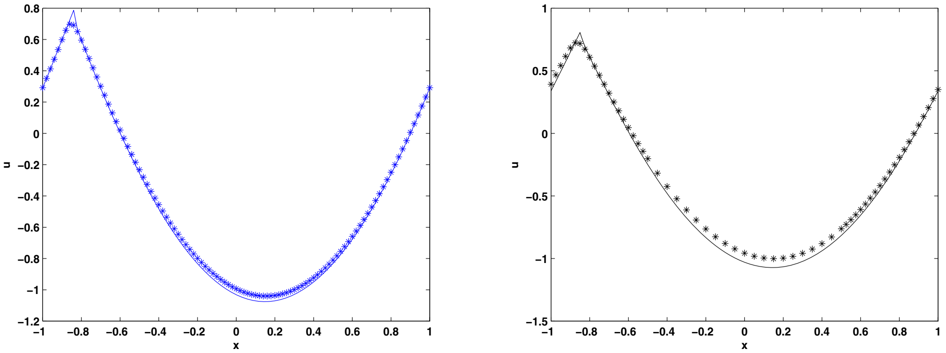
$$\theta_h^{A_1} \geq \frac{m(T)}{m(T)+m(T_1)}\frac{\|\nabla H\|}{\cos\omega_0}.$$

By similar considerations for $p_i, i = 2, 3, 4$, we have that

$$\frac{\partial\widehat{H}_{A_1}}{\partial p_i} \geq 0,\ i = 1, 2, 3, 4,$$

provided that

$$\theta_h^{A_1} = \frac{\max(m(T), m(T_1))}{m(T)+m(T_1)}\frac{\|\nabla H\|}{\cos\omega_0}.$$



(a) (b)

FIG. 5. *Local time stepping for the convex Hamiltonian. Numerical solution (in dots) and exact solution (in solid lines).* (a): *a uniform mesh;* (b): *a nonuniform mesh.*

Hence the monotonicity of $\widehat{H}_{A_1}$ follows.

The consistency of $\widehat{H}_{A_1}$ is obvious and the smoothness follows by the local smoothness of H.

5. Numerical experiments. We show some numerical examples to demonstrate the effectiveness of the new scheme.

Example 5.1 (A one-dimensional Hamilton–Jacobi equation). We solve

$$u_t + H(u_x) = 0, \qquad -1 \le x \le 1, \tag{5.5.1}$$

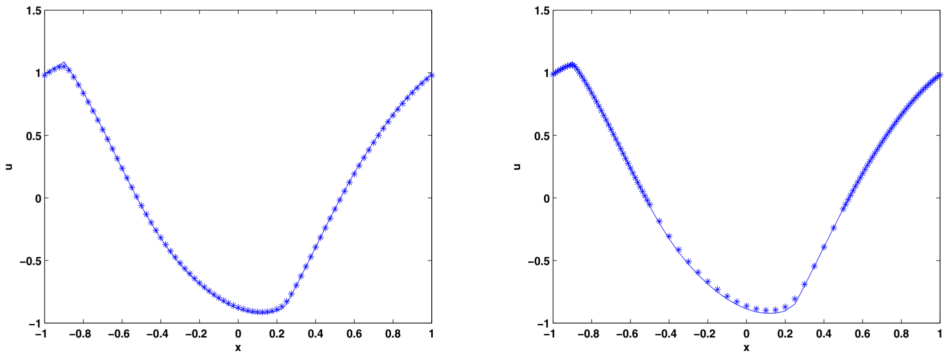
$$u(x,0) = -\cos\ \pi x, \tag{5.5.2}$$

with a convex $H(p) = \frac{(p+\alpha)^2}{2}$ and a nonconvex $H(p) = -\cos(p+\alpha)$.

We take $\alpha = 1$ and use the first order monotone scheme based on the Lax–Friedrichs numerical Hamiltonian to compute the viscosity solution. We use both uniform meshes and locally varying time and space grids to compute the solution up to $t = \frac{1.5}{\pi^2}$ when the solution has a discontinuous derivative.

Figure 5 shows solutions for the case of the convex Hamiltonian. Figure 5(a) shows the exact solution (in solid lines) and the solution computed with a uniform mesh of $dx = 0.02$ (in stars). Figure 5(b) shows the exact solution (in solid lines) and the solution computed with a nonuniform mesh (in stars). In the nonuniform mesh case, we have predetermined to use a mesh size of $dx = 0.025$ on the intervals $[-1,-0.5]$ and $[0.5,1]$ and a mesh size of $dx = 0.05$ on the interval $[-0.5,0.5]$; correspondingly we have used time steps on these different intervals according to the local CFL conditions. Although we did not use an optimal strategy to indicate where the solution is smooth or has kinks, the results show that the scheme based on locally varying space and time grids is convergent.

Figure 6 shows solutions for the case of the nonconvex Hamiltonian. Figure 6(a) shows the exact solution (in solid lines) and the solution computed with a uniform mesh of $dx = 0.025$ (in stars). Figure 6(b) shows the exact solution (in solid lines) and the solution computed with a nonuniform mesh (in stars). In the nonuniform mesh case, we have predetermined to use a mesh size of $dx = 0.0125$ on the intervals $[-1,-0.5]$ and $[0.5,1]$ and a mesh size of $dx = 0.05$ on the interval $[-0.5,0.5]$; correspondingly we have used time steps on these different intervals according to the local



(a) (b)

FIG. 6. *Local time stepping for the nonconvex Hamiltonian. Numerical solution (in dots) and exact solution (in solid lines).* (a): *a uniform mesh;* (b): *a nonuniform mesh.*

CFL conditions. Although we did not use an optimal strategy to indicate where the solution is smooth or has kinks, the results show that the scheme based on locally varying space and time grids is convergent.

Example 5.2 (A two-dimensional non-convex Hamilton–Jacobi equation: a Riemann problem). We use the proposed scheme in the AMR method [3, 5] to solve a two-dimensional nonconvex Riemann problem [31],

$$u_t + \sin(u_x + u_y) = 0, \qquad -2 \leq x, y \leq 2, \tag{5.5.3}$$

$$u(x, y, 0) = \pi(|y| - |x|), \tag{5.5.4}$$

to investigate the behavior of the AMR method and the convergence to the viscosity solution. In the following we assume that the reader is familiar with the version of AMR presented in [3].

We use the first-order Lax–Friedrichs monotone scheme [31] as the driving method in the AMR method to compute the viscosity solution. To simplify the implementation, we use the local truncation error estimator [3, 5] to flag where the computational mesh should be refined or coarsened; then a grid generation procedure dynamically creates or removes rectangular fine patches. On different patches, the time step-size is set according to the local CFL condition (2.2.10) and (2.2.11) so that efficient time stepping can be achieved on different levels of computational grids (or patches). Different from the AMR method for hyperbolic conservation laws where conservation has to be ensured across interfaces between coarse and fine grid patches [3, 5], the AMR method for Hamilton–Jacobi equations only needs to update the solution on a coarse patch by using the solution on a fine patch if available. To further simplify the implementation, we only allow the mesh refinement ratio to be two.

In the computation shown here, we have taken the coarsest mesh to be 40 × 40, the CFL number to be 0.85, the tolerance for the Richardson error estimator [3] to be 0.002528, the maximum mesh refinement to be 2, and the number of ghost points to be 4.

We compute the solution up to $t = 1.0$ by the AMR method. Figure 7 shows the solution in pseudocolor and its computational mesh; the AMR method has refined the mesh in the neighborhood of the two axes where the solution has rapid changes according to the local truncation error estimator; see the solution shown in Figure

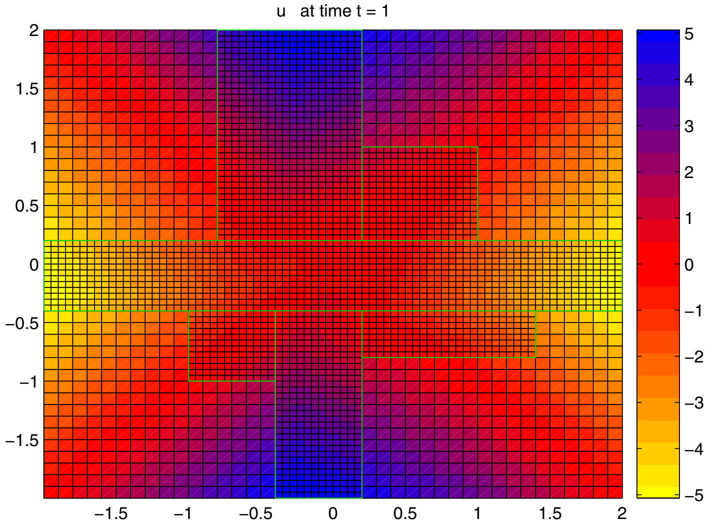
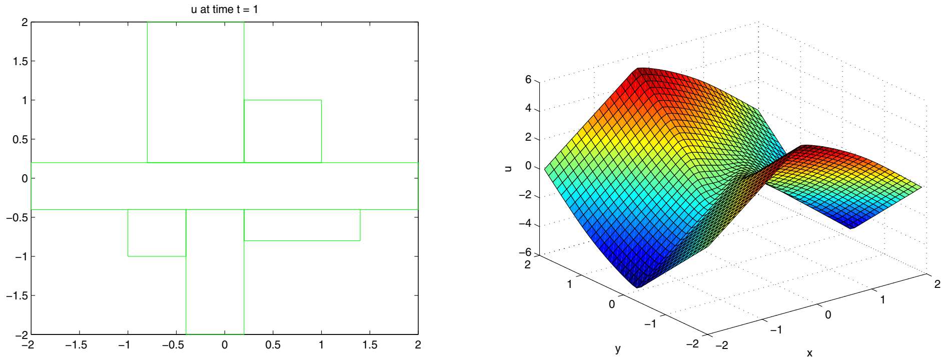


Fig. 7. *The solution at $t = 1$ by the AMR method and its computational mesh.*



(a)

(b)

Fig. 8. *The two-dimensional Hamilton–Jacobi equation by the AMR method.* (a): *$t = 1$ with seven grids, including the underlying coarsest mesh;* (b): *the solution at $t = 1$. Notice that the refined meshes are along the two axes, where the solution changes rapidly.*

8(b). Figure 8(a) shows the computational mesh at $t = 1$ consisting of seven patches (or grids) including the underlying coarsest mesh; Figure 8(b) shows the solution obtained by the AMR method at $t = 1$.

Figure 9 shows the calibration results at $t = 1$, where we have compared three solutions: the one computed by the AMR method, and the two solutions computed on the uniform 40×40 mesh by the first-order monotone Lax–Friedrichs scheme [31] and the third-order weighted essentially nonoscillatory (WENO) Lax–Friedrichs scheme [20]; as we can see, the solution by the AMR method based on the first-order monotone Lax–Friedrichs scheme is much more accurate than the one by the first-order monotone Lax–Friedrichs scheme on the uniform mesh if the solution by the third WENO scheme is accepted as the “true” solution.

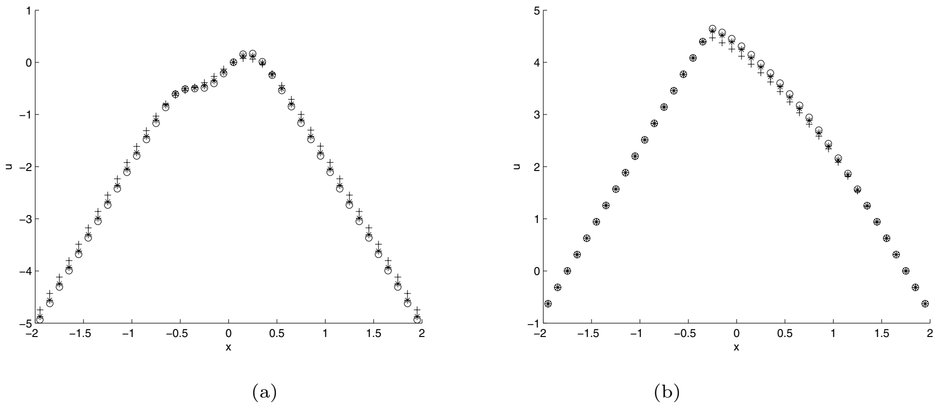


FIG. 9. *Calibration for the solution at $t = 1$ by the AMR method. WENO third-order: o; AMR: *; first-order Lax–Friedrichs scheme: +.* (a): $y = 0.05$*;* (b): $y = -1.75$.

We also point out that since in this example the coarsest mesh size is $h = 0.1$, the finest mesh size in the AMR method is $\frac{h}{2} = 0.05$, and $h^3 = 0.001$, the AMR solution based on the first-order scheme is not as accurate as the one by the third-order WENO scheme on the coarsest mesh. Therefore, the advantage of the AMR method will be more significant if a third-order WENO scheme is incorporated into the AMR method; this is an ongoing work.

Example 5.3 (A two-dimensional convex Hamilton–Jacobi equation). Now we use the proposed scheme in the AMR method [3, 5] to solve a two-dimensional convex periodic problem [31],

$$u_t + \frac{(u_x + u_y + 1)^2}{2} = 0, \qquad -2 \le x, y \le 2, \tag{5.5.5}$$

$$u(x, y, 0) = -\cos \pi \left(\frac{x + y}{2} \right), \tag{5.5.6}$$

to further investigate the behavior of the AMR method and the convergence to the viscosity solution.

We still use the first-order Lax–Friedrichs monotone scheme [31] as the driving method in the AMR method to compute the viscosity solution. Other implementation details are the same as those for the nonconvex Riemann problem.

In the computation shown here, we have taken the coarsest mesh to be 40 × 40, the CFL number to be 0.85, the tolerance for the Richardson error estimator [3] to be 0.0004525, the maximum mesh refinement to be 3, the mesh refinement ratio to be 2, and the number of ghost points to be 4.

We compute the solution up to $t = 0.152$ by the AMR method, when the solution has developed discontinuous gradients. Figure 10 shows the solution in pseudocolor and its computational mesh; the AMR method has refined the mesh where the solution has rapid changes according to the local truncation error estimator; see the solution shown in Figure 11(b). Figure 11(a) shows the computational mesh at $t = 0.152$ consisting of six patches (or grids) including the underlying coarsest mesh, the refined mesh in all the domain at the second level, and four local patches at the third level; Figure 11(b) shows the solution obtained by the AMR method at $t = 0.152$. Because the rapid changes in the solution are not aligned with horizontal or vertical direc-

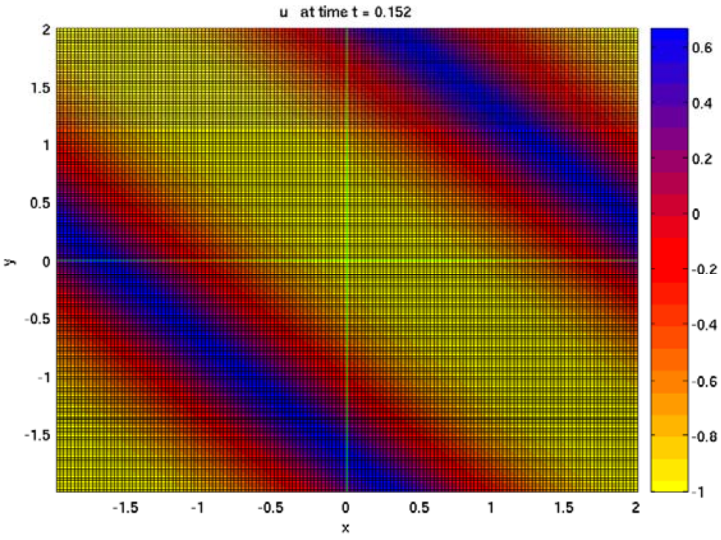
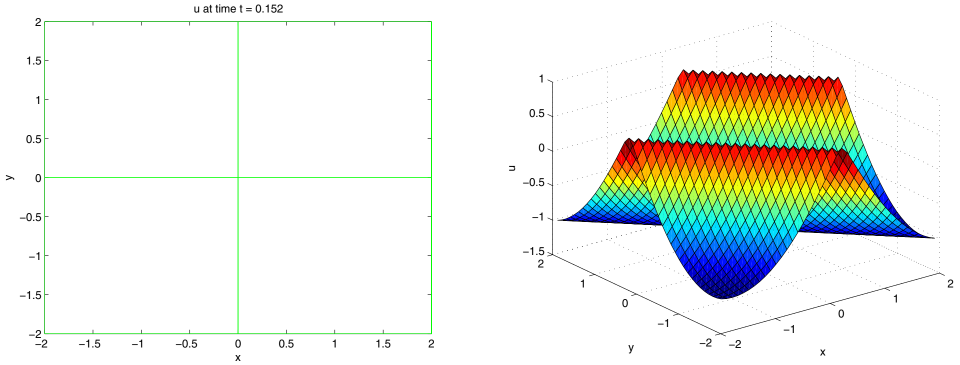


FIG. 10. *The solution at $t = 0.152$ by the AMR method and its computational mesh.*

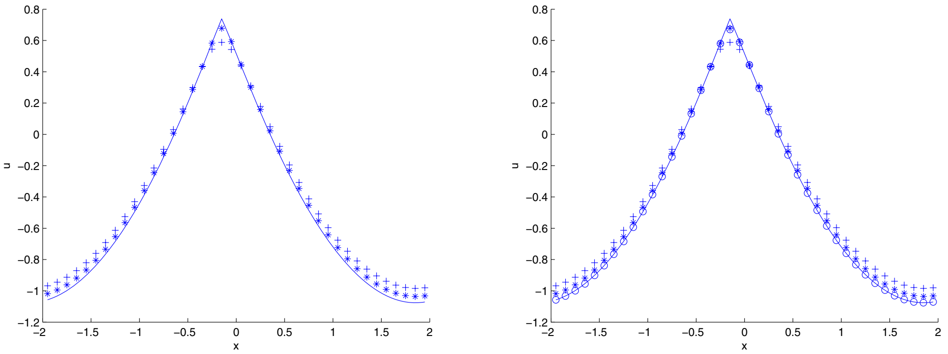


(a) (b)

FIG. 11. *The two-dimensional Hamilton–Jacobi equation by the AMR method.* (a): $t = 0.152$ *with six grids, including the underlying coarsest mesh, the refined mesh in all the domain at the second level and four local patches at the third level;* (b): *the solution at* $t = 0.152$.

tions, the AMR method based on rectangular meshes without rotation has refined the mesh everywhere so as to capture those sharp features; therefore, to have a more efficient AMR method for such features, one may use a version of the AMR method with rotations [4] or use a posteriori error estimators rather than local truncation error estimators to indicate more accurately where the computational mesh should be refined.

Figure 12 shows the calibration results at $t = 0.152$, where we have compared four solutions: the exact solution, the one computed by the AMR method presented here and the two solutions computed on the uniform 40×40 mesh by the first-order monotone Lax–Friedrichs scheme [31] and the third-order WENO Lax–Friedrichs scheme [20]; as we can see, the solution by the AMR method based on the first-order mono-



(a) (b)

FIG. 12. *Calibration for the solution at $t = 0.152$ for $y = -1.75$.* (a) *exact: -; AMR: *; first-order Lax–Friedrichs scheme: +.* (b) *exact: -; AMR: *; WENO third-order: o; first-order Lax–Friedrichs scheme: +.*

tone Lax–Friedrichs scheme is much more accurate than the one by the first-order monotone Lax–Friedrichs scheme on the uniform mesh; moreover, the AMR method provides very sharp resolution at the kink which is as good as the one by the third-order WENO scheme.

Example 5.4 (A two-dimensional Hamilton–Jacobi equation from geometrical optics). Now we use the proposed scheme in the AMR method [3, 5] to solve a two-dimensional periodic problem from geometrical optics [30, 26, 39],

$$u_t + \sqrt{u_x^2 + u_y^2 + 1} = 0, \qquad 0 \le x, y \le 1, \tag{5.5.7}$$

$$u(x, y, 0) = 0.25(\cos(2\pi x) - 1)(\cos(2\pi y) - 1) - 1. \tag{5.5.8}$$

We still use the first-order Lax–Friedrichs monotone scheme [31] as the driving method in the AMR method to compute the viscosity solution. Other implementation details are the same as those for the nonconvex Riemann problem.

In the computation shown here, we have taken the coarsest mesh to be 40 × 40, the CFL number to be 0.85, the tolerance for the Richardson error estimator [3] to be 0.0008250, the maximum mesh refinement to be 3, the mesh refinement ratio to be 2, and the number of ghost points to be 4.

We compute the solution up to $t = 0.60562$ by the AMR method, when the solution has developed discontinuous gradients. Figure 13 shows the solution in pseudocolor and its computational mesh; the AMR method has refined the mesh where the solution has rapid changes according to the local truncation error estimator; see the solution shown in Figure 14(b). Figure 14(a) shows the computational mesh at $t = 0.60562$ consisting of two grids including the underlying coarsest mesh and a refined patch in a neighborhood of (0.5, 0.5) at the second level; Figure 14(b) shows the solution obtained by the AMR method at $t = 0.60562$. Because the rapid changes in the solution are near the point (0.5, 0.5), the AMR method has refined the mesh in a neighborhood of that point so as to capture those sharp features. The results shown here can be compared with those in [39], in which a mesh redistribution method was used to cluster more mesh points in a neighborhood of (0.5, 0.5); see [39] for more details.

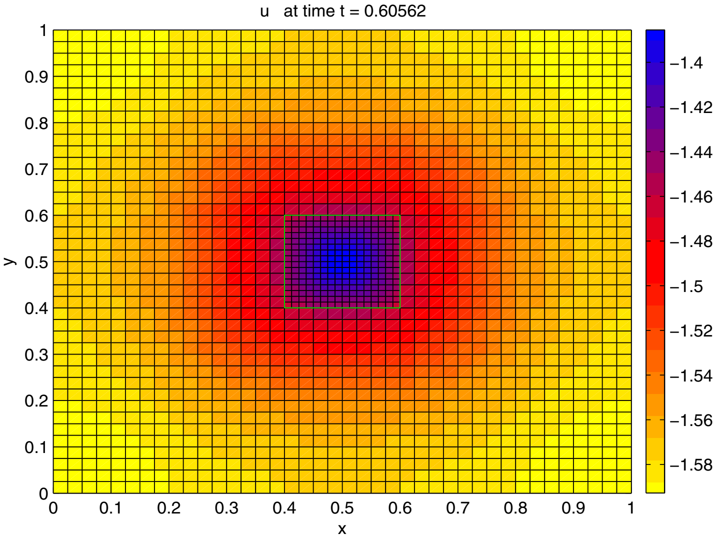


FIG. 13. *The solution at* $t = 0.60562$ *by the AMR method and its computational mesh.*

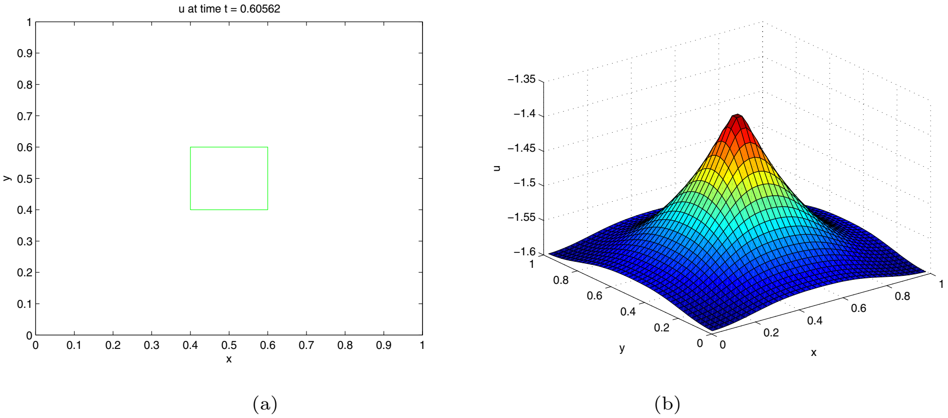


FIG. 14. *The two-dimensional Hamilton–Jacobi equation by the AMR method.* (a): $t = 0.60562$ *with two grids, including the underlying coarsest mesh and a refined patch in a neighborhood of* $(0.5, 0.5)$ *at the second level;* (b): *the solution at* $t = 0.60562$.

The above AMR examples based on the first-order Lax–Friedrichs scheme show that the proposed scheme fares well in comparison with existing methods. Currently we are incorporating a posteriori error estimators and WENO schemes into the AMR method so that we can have optimal mesh refinement and higher-order accuracy during the computation; a fully numerical assessment of such AMR methods and their performance as well as various conditions on the spatial and temporal meshes is an ongoing work.

Acknowledgment. The author thanks Prof. B. Cockburn, S. Osher and E. Tadmor for discussions and their generosity of sharing ideas with him. The author also would like to thank referees for many helpful suggestions. The author also would like to thank Prof. Marsha Berger for the AMR code that facilitates the implementation of examples used in the paper.

REFERENCES

[1] R. Abgrall, *Numerical discretization of the first-order Hamilton-Jacobi equations on triangular meshes*, Comm. Pure Appl. Math., 49 (1996), pp. 1339–1377.

[2] S. Albert, B. Cockburn, D. French, and T. Peterson, *A posteriori error estimates for general numerical methods for Hamilton-Jacobi equations. Part* I*: The steady state case*, Math. Comp., (2002), pp. 222–232.

[3] M. Berger and P. Colella, *Local adaptive mesh refinement for shock hydrodynamics*, J. Comput. Phys., 82 (1989), pp. 64–84.

[4] M. Berger and J. Oliger, *Adaptive mesh refinement for hyperbolic partial differential equations*, J. Comput. Phys., 53 (1984), pp. 484–512.

[5] M. J. Berger and R. J. LeVeque, *Adaptive mesh refinement using wave propagation algorithms for hyperbolic systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2298–2316.

[6] S. Bryson and D. Levy, *High order central WENO schemes for multi-dimensional Hamilton-Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 1339–1369.

[7] T. Cecil, J. Qian, and S. J. Osher, *Numerical methods for high dimensional Hamilton-Jacobi equations using radial basis functions*, J. Comput. Phys., 196 (2004), pp. 327–347.

[8] H. D. Ceniceros and T. Y. Hou, *An efficient dynamically adaptive mesh for potentially singular solutions*, J. Comput. Phys., 172 (2001), pp. 609–639.

[9] B. Cockburn and P.-A. Gremaud, *A priori error estimates for numerical methods for scalar conservation laws. Part* I*: The general approach*, Math. Comp., 65 (1996), pp. 533–573.

[10] B. Cockburn and P.-A. Gremaud, *A priori error estimates for numerical methods for scalar conservation laws. Part* II*: Flux-splitting monotone schemes on irregular Cartesian grids*, Math. Comp., 66 (1997), pp. 547–572.

[11] B. Cockburn, P.-A. Gremaud, and J. X. Yang, *A priori error estimates for hyperbolic conservation laws. Part* III*: Multidimensional flux-splitting monotone schemes in non-Cartesian grids*, SIAM J. Numer. Anal., 35 (1998), pp. 1775–1803.

[12] B. Cockburn and J. Qian, *Continuous dependence results for Hamilton-Jacobi equations*, in Collected Lectures on the Preservation of Stability Under Discretization, D. Estep and S. Tavener, eds., SIAM, Philadelphia, 2002, pp. 67–90.

[13] B. Cockburn and B. Yenikaya, *An adaptive method with rigorous error control for the Hamilton-Jacobi equations, I. The One-Dimensional Steady State Case*, Appl. Numer. Math., 52 (2005), no. 2–3, pp. 175–195.

[14] M.G. Crandall, L. C. Evans, and P. L. Lions, *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc. 282 (1984), 478–502.

[15] M. G. Crandall and P. L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.

[16] M. Falcone and R. Ferretti, *Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations*, Numer. Math., 67 (1994), pp. 315–344.

[17] L. Gosse and C. Makridakis, *Two a posteriori error estimates for one-dimensional scalar conservation laws*, SIAM J. Numer. Anal., 38 (2000), pp. 964–988.

[18] A. Harten, J. M. Hyman, and P. D. Lax, *On finite difference approximations and entropy conditions for shocks*, Comm. Pure Appl. Math, 29 (1976), pp. 297–322.

[19] W. Huang and R. D. Russell, *Moving mesh stragegy based on a gradient flow equation for two-dimensonal problems*, SIAM J. Sci. Comput., 20 (1999), pp. 998–1015.

[20] G. S. Jiang and D. Peng, *Weighted ENO schemes for Hamilton-Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.

[21] C. Y. Kao, S. J. Osher, and J. Qian, *Lax-Friedrichs sweeping schemes for static Hamilton-Jacobi equations*, J. Comput. Phys., 196 (2004), pp. 367–391.

[22] G. Kossioris, Ch. Makridakis, and P. E. Souganidis, *Finite volume schemes for Hamilton-Jacobi equations*, Numer. Math., 83 (1999), pp. 427–442.

[23] D. Kroner and M. Ohlberger, *A posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multidimensions*, Math. Comp., 69 (2000), pp. 25–39.

[24] A. Kurganov and E. Tadmor, *New high-resolution semi-discrete central schemes for Hamilton-Jacobi equations*, J. Comput. Phys., 160 (2000), pp. 720–742.

[25] S. Li and L. Petzold, *Moving mesh methods with upwinding schemes for time-dependent PDEs*, J. Comput. Phys., 131 (1997), pp. 368–377.

[26] C. T. Lin and E. Tadmor, *High-resolution nonoscillatory central schemes for Hamilton-Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2163–2186.

[27] C. T. Lin and E. Tadmor, *L^1-stability and error estimates for approximate Hamilton-Jacobi equations*, Numer. Math. 88 (2001), 2163–2186.

[28] R. Malladi and J. A. Sethian, *An o(nlogn) algorithm for shape modeling*, Proc. Natl. Acad. Sci., 93 (1996), pp. 9389–9392.

[29] S. Osher and R. Sanders, *Numerical approximations for nonlinear conservation laws with locally varying time and space grids*, Math. Comp., 41 (1983), pp. 321–336.

[30] S. J. Osher and J. A. Sethian, *Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[31] S. J. Osher and C. W. Shu, *High-order Essentially Nonoscillatory schemes for Hamilton-Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.

[32] J. Qian and W. W. Symes, *Adaptive finite difference method for traveltime and amplitude*, Geophysics, 67 (2002), pp. 167–176.

[33] W. Ren and X. P. Wang, *An iterative grid redistribution method for singular problems in multiple dimensions*, J. Comput. Phys., 159 (2000), pp. 246–273.

[34] E. Rouy and A. Tourin, *A viscosity solutions approach to shape-from-shading*, SIAM J. Numer. Anal., 29 (1992), pp. 867–884.

[35] K. Salari and S. Steinberg, *Flux-corrected transport in a moving grid*, J. Comput. Phys., 111 (1994), pp. 24–32.

[36] J. A. Sethian, *Level Set Methods*, Cambridge University Press, Cambridge, UK, 1996.

[37] Z. Tan, Z. Zhang, Y. Huang, and T. Tang, *Moving mesh methods with locally varying time steps*, J. Comput. Phys., 200 (2004), pp. 347–367.

[38] H. Z. Tang and T. Tang, *Adaptive mesh methods for one- and two-dimensional hyperbolic conservation laws*, SIAM J. Numer. Anal., 41 (2003), pp. 487–515.

[39] H. Z. Tang, T. Tang, and P. Zhang, *An adaptive mesh redistribution method for nonlinear Hamilton-Jacobi equations in two- and three- dimensions*, J. Comput. Phys., 188 (2003), pp. 543–572.

[40] T. Tang, *Moving Mesh Methods for Computational Fluid Dynamics*, CSCAMM-05-04, University of Maryland, College Park, MD, USA, 2005.

[41] Y.-T. Zhang and C.-W. Shu, *High order WENO schemes for Hamilton-Jacobi equations on triangular meshes*, SIAM J. Sci. Comput., 24 (2003), pp. 1005–1030.

[42] H. K. Zhao, *Fast sweeping method for eikonal equations*, Math. Comp., 74 (2005), pp. 603–627.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2402–2422

NUMERICAL APPROXIMATION OF A TWO-PHASE FLOW PROBLEM IN A POROUS MEDIUM WITH DISCONTINUOUS CAPILLARY FORCES[*]

GUILLAUME ENCHÉRY[†], R. EYMARD[‡], AND A. MICHEL[§]

Abstract. We consider a simplified model of a two-phase flow through a heterogeneous porous medium. Focusing on the capillary forces motion, a nonlinear degenerate parabolic problem is approximated in a domain shared in two homogeneous parts, each of them being characterized by its relative permeability and capillary curves functions of the phase saturations. We first give a weak form of the conservation equations on the whole domain, with a new general expression of the conditions at the interface between the two regions. We then propose a finite volume scheme for the approximation of the solution, which is shown to converge to a weak solution in one-, two-, or three-dimensional domains. We conclude with some numerical tests.

Key words. flows in porous media, capillarity, nonlinear PDE of parabolic type, finite volume methods

AMS subject classifications. 76S05, 76B45, 35K55, 74S10

DOI. 10.1137/040602936

1. Introduction.

Simulations of two-phase flows through heterogeneous porous media are widely used in petroleum engineering. For example, for exploration purposes, the basin modeling aims to reconstruct the geological history of a sedimentary basin and in particular the migration of hydrocarbon components at geological time scale. The reservoir simulation is devoted to the understanding and the prediction of fluid flows occurring during production processes.

One of the most important consequences of the presence of heterogeneities in a porous medium is the phenomenon of capillary entrapment. This phenomenon occurs at the interface between two geological layers where discontinuous capillary thresholds appear. Indeed, if the mean pore radius in one layer is smaller than in the other, the oil phase must reach an access pressure so that the oil phase can enter the least permeable layer. In a sedimentary basin, this mechanism can induce the formation of oilfields. On the other hand, in reservoir engineering, the capillary trapping can reduce the recovery factor since large quantities of oil can remain trapped. Therefore, for this kind of application, one needs a precise understanding of this phenomenon on the physical plane as on the mathematical plane.

The physical principles which govern these flows and the mathematical models can be found in [2], [3], [4], [7]. However, the phenomenon of capillary trapping and its mathematical modelization have been completed only in some simplified cases [5], [9], [14].

[*] Received by the editors January 7, 2004; accepted for publication (in revised form) July 5, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sinum/43-6/60293.html

[†] Weierstrass-Institut für Angewandte Analysis und Stochastik, Mohrenstr. 39, 10117 Berlin, Germany (enchery@wias-berlin.de).
[‡] Université de Marne-La-Vallée, 5 bd Descartes, Champs sur Marne, 77454 Marne-La-Vallée, France (eymard@math.univ-mlv.fr).
[§] Institut Français du Pétrole, 1-4 av. Bois Préau, 92852 Rueil-Malmaison, France (anthony.michel@ifp.fr).

The aim of this paper is to propose a general model for this phenomenon and to give the mathematical study of the convergence of a scheme which can be used in the industrial context.

We thus consider an incompressible and immiscible oil-water flow through a one-, two-, or three-dimensional heterogeneous and isotropic porous medium Ω. Using Darcy's law, the conservation of oil and water phases is given for all $(x,t) \in \Omega \times (0,T)$ by

$$(1.1) \qquad \begin{cases} -\phi(x)\dfrac{\partial u(x,t)}{\partial t} - \operatorname{div}\Big(\mu_w(x,u(x,t))(\nabla p_w(x,t) - \rho_w g)\Big) = 0, \\ \phi(x)\dfrac{\partial u(x,t)}{\partial t} - \operatorname{div}\Big(\mu_o(x,u(x,t))(\nabla p_o(x,t) - \rho_o g)\Big) = 0, \\ p_o(x,t) - p_w(x,t) = \pi(x,u(x,t)), \end{cases}$$

where the function ϕ is the porosity of the medium, $u \in [0,1]$ is the oil saturation (and therefore $1-u$ is the water saturation), $\pi(x,u)$ is the capillary pressure, and g is the gravity acceleration. The indices o and w, respectively, stand for the oil and the water phase. Thus, for $\beta = o, w$, p_β is the pressure of the phase β, $\mu_\beta(x,u)$ is the mobility of the phase β, and ρ_β is the density of the phase β. The unknowns of the problem are the functions u, p_w, and p_o.

Focusing on the modeling of flow at the interface between two different porous materials, we make the following assumptions.

Assumption 1.1.

H1-1. *The domain Ω is such that $\Omega = \Omega_1 \bigcup \Omega_2$. The subdomains Ω_1 and Ω_2 are disjoint open segments (if $d=1$), polygonal (if $d=2$), or polyhedral (if $d=3$) bounded connected subsets of $\mathbb{R}^d$. We assume that the common boundary between Ω_1 and Ω_2, $\Gamma = \partial\overline{\Omega}_1 \bigcap \partial\overline{\Omega}_2$, has a strictly positive and finite $d-1$-measure. The real $T > 0$ is the length of the considered time period.*

H1-2. *The function ϕ takes the strictly positive constant value $0 < \phi_i < 1$ in Ω_i for $i = 1,2$.*

H1-3. *For $\beta \in \{o,w\}$, $i = 1,2$, and for all $x \in \Omega_i$ $\mu_\beta(x,.) = \mu_{\beta,i}$. $\mu_{o,i}$ is a strictly increasing continuous function satisfying $\mu_{o,i}(u) = \mu_{o,i}(0) = 0$ for all $u \le 0$ and $\mu_{o,i}(u) = \mu_{o,i}(1)$ for all $u \ge 1$. $\mu_{w,i}$ is a strictly decreasing continuous function satisfying $\mu_{w,i}(u) = \mu_{w,i}(1) = 0$ for all $u \ge 1$ and $\mu_{w,i}(u) = \mu_{w,i}(0)$ for all $u \le 0$.*

H1-4. *For all $x \in \Omega_i$, $\pi(x,.) = \pi_i \in C^0(\mathbb{R},\mathbb{R})$ and π_i is such that its restriction $\pi_{i|[0,1]}$ to $[0,1]$ is strictly increasing, belongs to $C^1([0,1],\mathbb{R})$, and satisfies $\pi_i(u) = \pi_i(0)$ for all $u \le 0$ and $\pi_i(u) = \pi_i(1)$ for all $u \ge 1$. We assume that $\pi_1(0) \le \pi_2(0) \le \pi_1(1) \le \pi_2(1)$. We denote by $u_1^\star$ the unique real in $[0,1]$ satisfying $\pi_1(u_1^\star) = \pi_2(0)$. Thus, for all $u \in [0,u_1^\star)$, we have $\pi_1(u) < \pi_2(u)$. We denote by $u_2^\star$ the unique real in $[0,1]$ satisfying $\pi_2(u_2^\star) = \pi_1(1)$. Thus, for all $u \in (u_2^\star,1]$, we have $\pi_1(u) < \pi_2(u)$. (See Figure 1.1.)*

H1-5. *The initial condition in saturation $u_{\text{ini}} \in L^\infty(\Omega)$ and $0 \le u_{\text{ini}}(x) \le 1$ for almost everywhere (a.e.) $x \in \Omega$.*

The following conditions must be satisfied on the traces of u_i, $p_{\beta,i}$, and $\nabla p_{\beta,i}$ on $\Gamma \times (0,T)$, respectively, denoted by $u_{i,\Gamma}$, $p_{\beta,i,\Gamma}$, and $(\nabla p)_{\beta,i,\Gamma}$ (see [3]):

1. For any $\beta = o,w$, the flux of the phase β must be continuous:

$$(1.2)\ \mu_{\beta,1}(u_{1,\Gamma})((\nabla p)_{\beta,1,\Gamma} - \rho_\beta g).\overrightarrow{n}_{1,\Gamma} = -\mu_{\beta,2}(u_{2,\Gamma})((\nabla p)_{\beta,2,\Gamma} - \rho_\beta g).\overrightarrow{n}_{2,\Gamma},$$

where $\overrightarrow{n}_{i,\Gamma}$ is the unit normal of Γ outward to Ω_i.

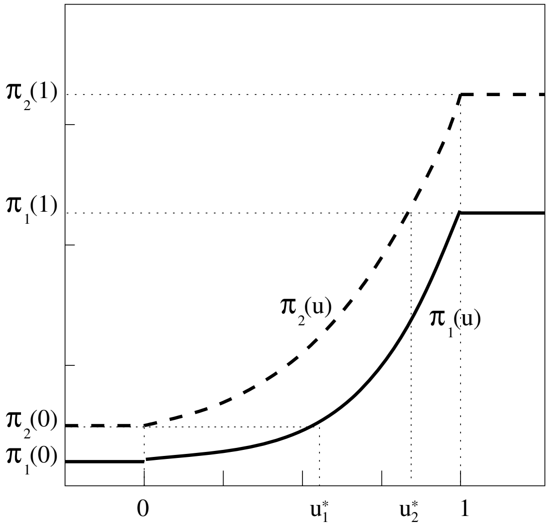


FIG. 1.1. *Functions* π_i, $i = 1, 2$.

2. For any $\beta = o, w$, either (p_β is continuous) or (p_β is discontinuous and $\mu_\beta = 0$); since the saturation is itself discontinuous across Γ, one must express the mobility at the upstream side of the interface. This gives

$$\mu_{\beta,1}(u_{1,\Gamma})(p_{\beta,1,\Gamma} - p_{\beta,2,\Gamma})^+ - \mu_{\beta,2}(u_{2,\Gamma})(p_{\beta,2,\Gamma} - p_{\beta,1,\Gamma})^+ = 0 \tag{1.3}$$

along with $p_{o,i,\Gamma} - p_{w,i,\Gamma} = \pi_i(u_{i,\Gamma})$, for $i = 1, 2$, where we denote, for all $a \in \mathbb{R}$, $a^+ = \max(a, 0)$.

The relations (1.3) can be directly expressed in terms of relations between $u_{i,\Gamma}$ and $p_{\beta,i,\Gamma}$, $\beta = o, w$, $i = 1, 2$:

1. If $0 \le u_{1,\Gamma} < u_1^\star$, then $\mu_{w,1}(u_{1,\Gamma}) > 0$; this implies $p_{w,1,\Gamma} \le p_{w,2,\Gamma}$. Since $\pi_1(u_{1,\Gamma}) < \pi_2(0) \le \pi_2(u_{2,\Gamma})$, we get $p_{o,1,\Gamma} < p_{o,2,\Gamma}$, which in turn implies $\mu_{o,2}(u_{2,\Gamma}) = 0$, and thus $u_{2,\Gamma} = 0$. Therefore $\mu_{w,2}(u_{2,\Gamma}) > 0$ and $p_{w,2,\Gamma} \le p_{w,1,\Gamma}$. Thus $p_{w,2,\Gamma} = p_{w,1,\Gamma}$. In this case, the oil phase is trapped in Ω_1, and the water flows across Γ.
2. If $u_1^\star \le u_{1,\Gamma}$ and $u_{2,\Gamma} \le u_2^\star$, then $\pi_2(0) \le \pi_1(u_{1,\Gamma})$, and $\pi_2(u_{2,\Gamma}) \le \pi_1(1)$. Since $\mu_{o,1}(u_{1,\Gamma}) > 0$, then $p_{o,1,\Gamma} \le p_{o,2,\Gamma}$ and $\mu_{o,2}(u_{2,\Gamma}) = 0$ or $p_{o,1,\Gamma} = p_{o,2,\Gamma}$. Similarly, since $\mu_{w,2}(u_{2,\Gamma}) > 0$, then $p_{w,1,\Gamma} \ge p_{w,2,\Gamma}$ and $\mu_{w,1}(u_{1,\Gamma}) = 0$ or $p_{w,1,\Gamma} = p_{w,2,\Gamma}$. Therefore, we get $p_{o,1,\Gamma} - p_{w,1,\Gamma} \le p_{o,2,\Gamma} - p_{w,2,\Gamma}$, which gives $\pi_1(u_{1,\Gamma}) \le \pi_2(u_{2,\Gamma})$. If we consider the case $\mu_{o,2}(u_{2,\Gamma}) = 0$, we get $u_{2,\Gamma} = 0$ and thus $\pi_2(0) = \pi_1(u_{1,\Gamma})$. Similarly, if we consider the case $\mu_{w,1}(u_{1,\Gamma}) = 0$, we get $\pi_2(u_{2,\Gamma}) = \pi_1(1)$. If we have at the same time $\mu_{o,2}(u_{2,\Gamma}) > 0$ and $\mu_{w,1}(u_{1,\Gamma}) > 0$, then $p_{o,1,\Gamma} = p_{o,2,\Gamma}$ and $p_{w,1,\Gamma} = p_{w,2,\Gamma}$, which implies $\pi_1(u_{1,\Gamma}) = \pi_2(u_{2,\Gamma})$. Therefore, in all cases, we get $\pi_1(u_{1,\Gamma}) = \pi_2(u_{2,\Gamma})$, and consequently $p_{o,1,\Gamma} = p_{o,2,\Gamma}$ and $p_{w,1,\Gamma} = p_{w,2,\Gamma}$. In this case, both phases flow across Γ.
3. If $u_2^\star < u_{2,\Gamma} \le 1$, a similar discussion yields $u_{1,\Gamma} = 1$ and $p_{o,1,\Gamma} = p_{o,2,\Gamma}$. In this case, the water phase is trapped in Ω_1, and the oil flows across Γ.

A consequence of this discussion is that in all cases, the resulting condition on the oil saturations at the boundary Γ is given by $\hat{\pi}_1(u_{1,\Gamma}) = \hat{\pi}_2(u_{2,\Gamma})$, defining the functions $\hat{\pi}_1$ and $\hat{\pi}_2$ by $\hat{\pi}_1 \;:\; u \mapsto \max(\pi_1(u), \pi_2(0))$ and $\hat{\pi}_2 \;:\; u \mapsto \min(\pi_2(u), \pi_1(1))$.

Now let us introduce the global pressure

$$\tilde{p}_i(x,t) = p_{w,i}(x,t) + \int_0^{u_i(x,t)} \frac{\mu_{o,i}(a)}{\mu_{o,i}(a) + \mu_{w,i}(a)} \pi_i'(a) da$$

(first introduced by Chavent; see, for example, [7]) and the functions $\eta_i \;:\; u \mapsto \frac{\mu_{o,i}(u)\mu_{w,i}(u)}{\mu_{o,i}(u)+\mu_{w,i}(u)}$ and $\varphi_i \;:\; u \mapsto \int_0^u \eta_i(a)\pi_i'(a)da$. We denote by L_{φ_i} the Lipschitz constant of φ_i and by C_η an upper bound of $\eta_i(u)$, $u \in \mathbb{R}$, $i = 1$ and 2. Using these notations we have for $(x,t) \in \Omega_i \times (0,T)$, $i = 1, 2$,

$$(1.4) \quad \begin{cases} \phi_i \dfrac{\partial u_i(x,t)}{\partial t} - \operatorname{div}\Big(\mu_{o,i}(u_i(x,t))(\nabla \tilde{p}_i(x,t) - \rho_o g)\Big) - \Delta\varphi_i(u_i(x,t)) = 0, \\ -\operatorname{div}\left(\displaystyle\sum_{\beta=o,w} \mu_{\beta,i}(u_i(x,t))\nabla\tilde{p}_i(x,t) - \sum_{\beta=o,w} \mu_{\beta,i}(u_i(x,t))\rho_\beta g\right) = 0. \end{cases}$$

We neglect in the first equation of (1.4) the term div $[\mu_{o,i}(u_i(x,t))(\nabla\tilde{p}_i(x,t) - \rho_o g)]$ in front of $\Delta\varphi_i(u_i(x,t))$, since this is sufficient to get the mathematical properties which are involved in the oil trapping phenomenon, as shown in the numerical examples at the end of this paper. Equations (1.2), (1.3), and (1.4) then produce within this simplified case the following equations, the solution of which are the functions $u_i(x,t)$, $(x,t) \in \Omega_i \times (0,T)$:

$$(1.5) \qquad \phi_i \frac{\partial u_i}{\partial t} - \Delta\varphi_i(u_i) = 0, \text{in } \Omega_i \times (0,T) \text{ for all } i \in \{1,2\},$$

$$(1.6) \qquad \nabla\varphi_1(u_{1,\Gamma}).\overrightarrow{n}_{1,\Gamma} = -\nabla\varphi_2(u_{2,\Gamma}).\overrightarrow{n}_{2,\Gamma} \text{ on } \Gamma \times (0,T),$$

and

$$(1.7) \qquad \hat{\pi}_1(u_{1,\Gamma}) = \hat{\pi}_2(u_{2,\Gamma}),$$

which summarizes the discussion induced by (1.3). Considering the problem of the migration of oil, we prescribe a homogeneous Neumann condition, which is expressed by

$$(1.8) \qquad \eta(.,u)\nabla\pi(.,u).\overrightarrow{n} = 0 \text{ on } \partial\Omega \times (0,T).$$

For $t = 0$, we have

$$(1.9) \qquad u(x,0) = u_{\text{ini}} \text{ in } \Omega.$$

Before giving the weak formulation of the problem we prove the following lemma.

LEMMA 1.2. *Under Assumption* 1.1*, let* $\Psi \;:\; [\pi_2(0), \pi_1(1)] \to \mathbb{R}$ *be the strictly increasing function defined by* $p \mapsto \Psi(p) = \int_{\pi_2(0)}^p \min(\eta_1(\pi_1^{(-1)}(a)), \eta_2(\pi_2^{(-1)}(a)))da$. *For all* $i \in \{1,2\}$*, the function* $\Psi \circ \hat{\pi}_i \circ \varphi_i^{(-1)}$ *is Lipschitz continuous with a constant lower than* 1.

Proof. For $i = 1$ or 2, let a be real such that $\varphi_1(u_1^\star) < a < \varphi_1(1)$ if $i = 1$, $0 < a < \varphi_2(u_2^\star)$ if $i = 2$. Within such a condition, we have $\hat{\pi}_i(\varphi_i^{(-1)}(a)) = \pi_i(\varphi_i^{(-1)}(a))$.

Let us calculate the derivative of the function $\pi_i \circ \varphi_i^{(-1)}$. Let $b \neq a$ be a real such that $\varphi_1(u_1^\star) < b < \varphi_1(1)$ if $i = 1$, $0 < b < \varphi_2(u_2^\star)$ if $i = 2$; setting $A = \varphi_i^{(-1)}(a)$ and $B = \varphi_i^{(-1)}(b)$, we have

$$\frac{\pi_i(\varphi_i^{(-1)}(b)) - \pi_i(\varphi_i^{(-1)}(a))}{b - a} = \frac{\pi_i(B) - \pi_i(A)}{\varphi_i(B) - \varphi_i(A)}.$$

Let us denote by $I(A, B)$ the interval $[A, B]$ if $B \geq A$, $[B, A]$ otherwise. Using the definition of φ_i, we have

$$\begin{aligned}\left(\min_{C \in I(A,B)} \eta_i(C)\right)(\pi_i(B) - \pi_i(A)) &\leq \varphi_i(B) - \varphi_i(A) \\ \leq \left(\max_{C \in I(A,B)} \eta_i(C)\right)(\pi_i(B) - \pi_i(A)),&\end{aligned}$$

and therefore there exists $C \in I(A, B)$ such that $\varphi_i(B) - \varphi_i(A) = \eta_i(C)(\pi_i(B) - \pi_i(A))$. Thus

$$\frac{\pi_i(\varphi_i^{(-1)}(b)) - \pi_i(\varphi_i^{(-1)}(a))}{b - a} = \frac{1}{\eta_i(C)},$$

which gives, letting $b \to a$, $(\pi_i \circ \varphi_i^{(-1)})'(a) = \frac{1}{\eta_i(\varphi_i^{(-1)}(a))}$. We thus get that the function $\Psi \circ \hat{\pi}_i \circ \varphi_i^{(-1)}$ has a derivative in a which is

$$(\Psi \circ \hat{\pi}_i \circ \varphi_i^{(-1)})'(a) = \Psi'(\pi_i(\varphi_i^{(-1)}(a)))(\pi_i \circ \varphi_i^{(-1)})'(a) = \frac{\Psi'(\pi_i(\varphi_i^{(-1)}(a)))}{\eta_i(\varphi_i^{(-1)}(a))}.$$

Using the definition of Ψ, we get $\Psi'(\pi_i(y)) \leq \eta_i(y)$ for $y = \varphi_i^{(-1)}(a)$. Gathering these results, we get that

$$(\Psi \circ \hat{\pi}_i \circ \varphi_i^{(-1)})'(a) \leq 1.$$

If $i = 1$ and $0 < a < \varphi_1(u_1^\star)$, or if $i = 2$ and $\varphi_2(u_2^\star) < a < 1$, then the function $\Psi \circ \hat{\pi}_i \circ \varphi_i^{(-1)}$ is constant, which implies a zero derivative. This completes the proof of the lemma. □

The system (1.5)–(1.9) is a nonlinear parabolic problem defined on a heterogeneous domain. Since in the general case, such a problem does not have any strong solution, we now give the definition of a weak solution to this problem.

Definition 1.3. *Under Assumption* 1.1*, a weak solution u of the problem* (1.5)–(1.9) *is defined by*

1. *for all $i \in \{1, 2\}$, $u = u_i$ in $\Omega_i \times (0, T)$ with*

$$u_i \in L^\infty(\Omega_i \times (0, T)),\ 0 \leq u_i \leq 1 \text{ a.e. and } \varphi_i(u_i) \in L^2(0, T; H^1(\Omega_i));$$

2. *for all $\psi \in C_{test} = \{h \in H^1(\Omega \times (0, T)),\ \ h(., T) = 0\}$,*

$$\sum_{i=1}^{2} \left[\begin{array}{l} \displaystyle\int_0^T \int_{\Omega_i} [\phi_i u_i(x,t)\psi_t(x,t) - \nabla\varphi_i(u_i(x,t)).\nabla\psi(x,t)]\, dxdt + \\ \displaystyle\int_{\Omega_i} \phi_i u_{\text{ini}}(x,0)\psi(x,0)dx \end{array} \right] = 0,$$

3. *the function $w : \Omega \times (0,T) \to \mathbb{R}$ defined by $(x,t) \mapsto \Psi(\hat{\pi}_i(u_i(x,t)))$ for a.e. $(x,t) \in \Omega_i \times (0,T)$, $i = 1,2$, belongs to $L^2(0,T;H^1(\Omega))$.*

Remark 1.4. This weak formulation is sufficient to impose (1.5), (1.6), (1.8), (1.9) on regular solutions. The last condition given in Definition 1.3 is a functional method to impose the condition (1.7).

In the homogeneous case, i.e., $\phi_1 = \phi_2$, $\pi_1 = \pi_2$, and $\eta_1 = \eta_2$, classical results of existence and uniqueness of a solution are available (see, for instance, [1] and [6] for a uniqueness result in more general cases). A simplified case of (1.5)–(1.9) has been handled in the heterogeneous case in [5], where the authors handle the case $d = 1$, $\Omega_1 = (-\infty,0)$, $\Omega_2 = (0,+\infty)$, and for $i = 1,2$, $\phi_i = 1$, $\eta_i(u) = k_i u$, and $\pi_i(u) = (1+u)/\sqrt{k_i}$, where $0 < k_2 < k_1$. (Note that only the problem of the oil trapping is considered here, since the physical conditions $\eta_i(1) = 0$ is not ensured.) Under additional hypotheses of regularity on the initial data, the authors get the existence and the uniqueness of the solution to the problem (1.5)–(1.9). We focus in this paper on the convergence of a numerical scheme for the approximation of u, in the general framework of Assumption 1.1. Up to a subsequence, we prove (see Theorem 2.15) the convergence of the finite volume scheme given by (2.2)–(2.4) to a weak solution in the sense of Definition 1.3. As an immediate consequence, the convergence of the scheme gives the existence of a solution to the problem (1.5)–(1.9) (see Corollary 2.17). Similar works have already been done, for example, in [12], [13] in the case of a homogeneous domain. Therefore, in the following proofs, we only insist on the new elements which appear in our study, mainly related to the presence of two domains linked by (1.6)–(1.7) (or (2.4) for the discrete problem). We end this study with numerical results (see section 3) and concluding remarks on ongoing works and future prospects (see section 4).

2. Study of a finite volume scheme.

In this section, we study a finite volume scheme discretizing (1.5)–(1.9). First we define an admissible discretization of $\Omega \times (0,T)$.

2.1. Admissible discretization of $\Omega \times (0,T)$.

Definition 2.1 (admissible mesh). *We denote by $\mathcal{M}$ an admissible finite volume discretization on a domain Ω; $\mathcal{M}$ is composed of a triplet $(\mathcal{T},\mathcal{E},\mathcal{P})$ with $\mathcal{T} = \mathcal{T}_1 \bigcup \mathcal{T}_2$, $\mathcal{E} = \mathcal{E}_1 \bigcup \mathcal{E}_2$, and $\mathcal{P} = \mathcal{P}_1 \bigcup \mathcal{P}_2$, which satisfy the following properties:*

- *For $i \in \{1,2\}$, $\mathcal{T}_i$ is a family of control volumes which are nonempty open polygonal convex disjoint subsets of Ω_i. These elements satisfy $\cup_{K\in\mathcal{T}_i}\overline{K} = \overline{\Omega}_i$. We denote by $\partial K = \overline{K} \setminus K$ the boundary of volume K and by $m(K)$ its measure (its length for $d = 1$, its area for $d = 2$, its volume for $d = 3$).*
- *For $i \in \{1,2\}$, $\mathcal{E}_i$ stands for the set of the edges of the control volumes in $\mathcal{T}_i$. For all $\sigma \in \mathcal{E}_i$, there exist a hyperplane E of $\mathbb{R}^d$ and a control volume $K \in \mathcal{T}_i$ such that $\overline{\sigma} = E \bigcap \partial K$ and σ is a nonempty open subset of E. We denote by $\mathcal{E}_K$ the subset of $\mathcal{E}$ composed of the edges of the volume K. Then we have $\partial K = \cup_{\sigma\in\mathcal{E}_K}\overline{\sigma}$. For any $\sigma \in \mathcal{E}_i$, we have*
  - *either $\sigma \in \mathcal{E}_{int,i} = \{\sigma \in \mathcal{E}_i,\ \exists\ (K,L) \in \mathcal{T}_i^2,\ K \neq L,$ such that $\overline{\sigma} = \overline{K} \cap \overline{L} \neq \emptyset\}$ (in that case σ is also denoted by $K|L$),*
  - *or $\sigma \in \mathcal{E}_\Gamma = \{\sigma \in \mathcal{E}_i,\ \exists\,(K,L) \in \mathcal{T}_1 \times \mathcal{T}_2,\ K \neq L,$ such that $\overline{\sigma} = \overline{K} \bigcap \overline{L} \neq \emptyset\}$,*
  - *or $\sigma \in \mathcal{E}_{ext,i} = \{\sigma \in \mathcal{E}_i,\ \exists\ K\ \in \mathcal{T}_i$ such that $\bar{\sigma} = \partial K \bigcap (\partial\Omega_i \setminus \Gamma) \neq \emptyset\}$.*
- *For $i \in \{1,2\}$, $\mathcal{P}_i$ refers to a family of points $(x_K)_{K\in\mathcal{T}}$ satisfying the following properties:*
  - *$x_K \in K$,*

- *for all $L \in \mathcal{T}_j$, $j \in \{1, 2\}$, the straight line (x_K, x_L) going through x_K and x_L is orthogonal to $K|L$.*

We also set

- $\mathcal{T}_\Gamma = \{(K, L),\ K \in \mathcal{T}_1,\ L \in \mathcal{T}_2,\ K|L \in \mathcal{E}_\Gamma\}$,
- $\mathcal{E}_{int} = \mathcal{E}_{int,1} \bigcup \mathcal{E}_{int,2} \bigcup \mathcal{E}_\Gamma$,
- $\mathcal{E}_{ext} = \mathcal{E}_{ext,1} \bigcup \mathcal{E}_{ext,2}$.

For $i = 1, 2$, the set of the neighboring volumes of a volume $K \in \mathcal{T}_i$ within Ω_i is represented by $N(K) = \{L \in \mathcal{T}_i,\ K|L \in \mathcal{E}_K\}$. The unit normal of an edge $K|L \in \mathcal{E}_{int}$ outward to K is denoted by $\overrightarrow{n}_{K,L}$. The area of an edge σ is denoted by $m(\sigma)$. For all $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_K$, $d_{K,\sigma}$ stands for the euclidean distance between x_K and the edge σ and for $K|L \in \mathcal{E}_{int}$, $d_{K|L}$ is the euclidean distance between x_K and x_L. Using these notations the transmissivity $\tau_{K|L}$ through $K|L$ is equal to $m(K|L)/d_{K|L}$ and, for $\sigma \in \mathcal{E}_{ext}$ with $\sigma \in \mathcal{E}_K$, the transmissivity $\tau_{K,\sigma}$ through σ is equal to $m(\sigma)/d_{K,\sigma}$. For $i \in \{1, 2\}$ and $K|L \in \mathcal{E}_{int,i}$, we denote by $D_{K|L}$ the union of the two cones with the respective vertices x_K and x_L and the basis $K|L$. For $\sigma \in \mathcal{E}_{ext}$ such that $\sigma \in \mathcal{E}_K$, D_σ is the cone with vertex x_K and basis σ.

We set $\mathrm{size}(\mathcal{M}) = \sup\{\mathrm{diam(K)}, \mathrm{K} \in \mathcal{T}\}$. *The regularity of the mesh is defined by*

$$\mathrm{regul}(\mathcal{M}) = \frac{\mathrm{size}(\mathcal{M})}{\min_{K \in \mathcal{T}, \sigma \in \mathcal{E}_K} d_{K,\sigma}}. \tag{2.1}$$

In this paper, for the sake of simplicity, we restrict our study to constant time steps. But all results stated in the following can be adjusted to variable time steps.

DEFINITION 2.2 (admissible time discretization of $(0, T)$). *A discretization of $(0, T)$ is given by an integer $M \in \mathbb{N}$ such that $\delta t = \frac{T}{M+1}$. The increasing sequence of times $(t_n)_{n \in \{0 \dots M+1\}}$ which discretizes $(0, T)$ is then given by $t_n = n\delta t$.*

DEFINITION 2.3 (admissible discretization of $\Omega \times (0, T)$). *An admissible discretization $\mathcal{D}$ of $\Omega \times (0, T)$ is composed of a pair $(\mathcal{M}, M)$, where $\mathcal{M}$ is an admissible discretization of Ω and $M \in \mathbb{N}$ (see Definitions* 2.1 *and* 2.2*). We then denote* $\mathrm{size}(\mathcal{D}) = \max(\mathrm{size}(\mathcal{M}), \delta t)$.

2.2. Discrete functional properties. Let $\mathcal{D}$ be an admissible discretization of the domain $\Omega \times (0, T)$ (see Definition 2.3), $K \in \mathcal{T}$, and $n \in \{0 \dots M\}$. For a variable u, we denote by u_K^{n+1} its approximation over the volume K and over the time interval $]n\delta t, (n+1)\delta t]$ and by $(u_K^0)_{K \in \mathcal{T}}$ a piecewise constant approximation of the initial condition. We denote by

- $\mathcal{X}(\mathcal{T})$ the set of piecewise constant functions over the mesh $\mathcal{T}$: $u_\mathcal{T} \in \mathcal{X}(\mathcal{T})$ is defined for all $x \in \Omega$ by $u_\mathcal{T}(x) = u_K$ for $x \in K$,
- $\mathcal{X}(\mathcal{D})$ the set of piecewise constant functions over the discretization $\mathcal{D}$: $u_\mathcal{D} \in \mathcal{X}(\mathcal{D})$ is defined for all $n \in \{0 \dots M\}$ by $u_\mathcal{D}(., t) = u_\mathcal{T}^{n+1} \in \mathcal{X}(\mathcal{T})$ for $t \in]n\delta t, (n+1)\delta t]$.

We introduce the notation $\delta u_{K,L} = u_L - u_K$.

For $i \in \{1, 2\}$, the discrete $L^2(0, T; H^1(\Omega_i))$-seminorm is defined as follows.

DEFINITION 2.4. *Let $\Omega \times (0, T)$ be a domain satisfying* H1-1 *and $\mathcal{D}$ be an admissible discretization of this domain in the sense of Definition* 2.3. *For $i \in \{1, 2\}$, the $L^2(0, T; H^1(\Omega_i))$-seminorm of a function $u_\mathcal{D} \in \mathcal{X}(\mathcal{D})$ is defined by*

$$|u_\mathcal{D}|_{1,\mathcal{D},i}^2 = \sum_{n=0}^{M} \delta t \sum_{K|L \in \mathcal{E}_{int,i}} \tau_{K|L} (\delta u_{K,L}^{n+1})^2.$$

2.3. An implicit scheme. The initial condition u_K^0 is given by

$$u_K^0 = \frac{1}{m(K)} \int_K u_{\mathrm{ini}}(x)\, dx,\ \forall K \in \mathcal{T}. \tag{2.2}$$

For the following time steps, $n \in \{0, \dots, M\}$, we compute a discrete solution in saturation $(u_K^{n+1})_{K\in\mathcal{T}}$ thanks to the scheme

$$\begin{aligned} &m(K)\phi_i \frac{u_K^{n+1} - u_K^n}{\delta t} \sum_{L\in N(K)} \tau_{K|L} \left(\varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1})\right) \\ &+ \sum_{\sigma\in\mathcal{E}_\Gamma \bigcap \mathcal{E}_K} \tau_{K,\sigma} \left(\varphi_i(u_K^{n+1}) - \varphi_i(u_{K,\sigma}^{n+1})\right) = 0,\ K \in \mathcal{T}_i,\ i \in \{1, 2\}, \end{aligned} \tag{2.3}$$

where for all $(K, L) \in \mathcal{T}_\Gamma$ and for given values of u_K^{n+1} and u_L^{n+1}, the values $u_{K,K|L}^{n+1}$, $u_{L,K|L}^{n+1} \in [0, 1]$ are the unique solutions (according to Lemma 2.5 below) of the system

$$\left\{ \begin{array}{rcl} \tau_{K,K|L}(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,\sigma}^{n+1})) & = & \tau_{L,K|L}(\varphi_2(u_{L,\sigma}^{n+1}) - \varphi_2(u_L^{n+1})), \\ \hat{\pi}_1(u_{K,\sigma}^{n+1}) & = & \hat{\pi}_2(u_{L,\sigma}^{n+1}). \end{array} \right. \tag{2.4}$$

Lemma 2.5. *Under Assumption* 1.1, *let* $\alpha_i > 0$ *be given for* $i = 1, 2$. *Let* $(a, b) \in \mathbb{R}^2$. *Then there exists one and only one pair* $(c, d) \in [0, 1]^2$ *such that*

$$\alpha_1(\varphi_1(a) - \varphi_1(c)) = \alpha_2(\varphi_2(d) - \varphi_2(b))$$

and

$$\hat{\pi}_1(c) = \hat{\pi}_2(d).$$

We then denote $c = U_1(a, b, \alpha_1, \alpha_2)$ *and* $d = U_2(a, b, \alpha_1, \alpha_2)$. *Then the functions* U_1 *and* U_2 *are continuous and nondecreasing with respect to* a *and* b. *Moreover, the following inequalities hold:*

$$\begin{aligned} &0 \le (\varphi_1(a) - \varphi_1(c))(\pi_1(a) - \pi_1(c)) \le (\varphi_1(a) - \varphi_1(c))(\pi_1(a) - \pi_2(b)), \\ &0 \le (\varphi_2(d) - \varphi_2(b))(\pi_2(d) - \pi_2(b)) \le (\varphi_2(d) - \varphi_2(c))(\pi_1(a) - \pi_2(b)). \end{aligned} \tag{2.5}$$

Proof. Let us take as unknowns the values $C = \varphi_1(c)$ and $D = \varphi_2(d)$ and let us denote $A = \varphi_1(a)$ and $B = \varphi_2(b)$. Then (C, D) is solution of

$$\alpha_1 C + \alpha_2 D = \alpha_1 A + \alpha_2 B, \tag{2.6}$$

$$\hat{\pi}_1(\varphi_1^{(-1)}(C)) = \hat{\pi}_2(\varphi_2^{(-1)}(D)). \tag{2.7}$$

Let us first consider the case where $\alpha_1 A + \alpha_2 B \le \alpha_1\varphi_1(u_1^\star)$. Since this implies $C \le \varphi_1(u_1^\star)$, we have necessarily $D = 0$ according to (2.7). Thus the solution is obtained, taking $D = 0$ and $C = (\alpha_1 A + \alpha_2 B)/\alpha_1$. In this case, since $D \le B$, we have $C \ge A$, and since $\pi_2(b) \ge \pi_2(0) \ge \pi_1(c) \ge \pi_1(a)$, we get (2.5).

We now consider the case where $\alpha_1\varphi_1(u_1^\star) < \alpha_1 A + \alpha_2 B < \alpha_1\varphi_1(1) + \alpha_2\varphi_2(u_2^\star)$. Since in this case we necessarily have $\varphi_1(u_1^\star) < C$ and $D < \varphi_2(u_2^\star)$ (see (2.7)), the relation $C = \varphi_1(\pi_1^{(-1)}(\pi_2(\varphi_2^{(-1)}(D))))$ holds, and since the function $D \mapsto \alpha_1\varphi_1 (\pi_1^{(-1)}(\pi_2(\varphi_2^{(-1)}(D)))) + \alpha_2 D$ is continuous and strictly increasing, the system has one and only one solution (C, D). We then get in this case that $\pi_1(c) = \pi_2(d)$, and since $\pi_1(a) - \pi_1(c)$ has the same sign as $\pi_2(d) - \pi_2(b)$, we get (2.5).

Finally, the case $\alpha_1\varphi_1(1) + \alpha_2\varphi_2(u_2^\star) \leq \alpha_1 A + \alpha_2 B$ is symmetric with the first case, and we get $C = \varphi_1(1)$ and $D = (\alpha_1(A - \varphi_1(1)) + \alpha_2 B)/\alpha_2$. We then have in this case $C \geq A$ and thus $D \leq B$, and since $\pi_2(b) \geq \pi_2(d) \geq \pi_1(1) \geq \pi_1(a)$, we again get (2.5).

In all these cases, C and D have been expressed as continuous nondecreasing functions of A and B, so the same conclusion holds for c and d as functions of a and b. □

Remark 2.6. It is possible to show that C and D, seen as functions of $A = \varphi_1(a)$ and $B = \varphi_2(b)$, verify, for a.e. $(a, b) \in \mathbb{R}^2$,

$$0 \leq \frac{\partial C}{\partial A} \leq 1,\ 0 \leq \frac{\partial D}{\partial A} \leq \frac{\alpha_1}{\alpha_2},\ 0 \leq \frac{\partial C}{\partial B} \leq \frac{\alpha_2}{\alpha_1},\ \text{and } 0 \leq \frac{\partial D}{\partial B} \leq 1.$$

Now we can state the L^∞-stability of the scheme and then the existence of a solution to (2.2)–(2.4).

2.4. L^∞-stability of the scheme.

If Ω were a homogeneous porous medium we could prove that the discrete solution in saturation satisfies a maximum principle depending on the initial condition [12]. Here, in presence of a heterogeneity, this result no longer holds.

PROPOSITION 2.7. *Under Assumption* 1.1*, let $\mathcal{D}$ be an admissible discretization of the domain $\Omega \times (0, T)$ (see Definition* 2.3*) and $u_\mathcal{T}^{n+1} \in \mathcal{X}(\mathcal{T})$, $n \in \{0 \ldots M\}$, the solution to the system* (2.2)–(2.4)*. (The existence and uniqueness of such a solution is shown in Proposition* 2.8*.) Then $u_\mathcal{T}^{n+1}$ satisfies*

$$\textit{for all } K \in \mathcal{T},\ 0 \leq u_K^{n+1} \leq 1. \tag{2.8}$$

Proof. For all $K \in \mathcal{T}_i$, $i \in \{1, 2\}$, (2.2)–(2.4) imply

$$u_K^{n+1} = H_K(u_K^n, (u_L^{n+1})_{L\in\mathcal{T}})$$

with

$$H_K(a, (a_L)_{L\in\mathcal{T}}) = \frac{1}{1+\lambda_K}\left(a + \lambda_K a_K + \frac{\delta t}{m(K)\phi_i}\left(\begin{array}{l} \displaystyle\sum_{L\in N(K)} \tau_{K|L}\left(\varphi_i(a_L) - \varphi_i(a_K)\right) \\ \displaystyle + \sum_{\sigma\in\mathcal{E}_\Gamma \bigcap \mathcal{E}_K} \tau_{K,\sigma}\left(\varphi_i(a_{K,\sigma}) - \varphi_i(a_K)\right) \end{array} \right)\right)$$

and

$$\lambda_K = \frac{\delta t L_\varphi}{m(K)\phi_i}\left(\sum_{L\in N(K)} \tau_{K|L} + \sum_{\sigma\in\mathcal{E}_\Gamma \bigcap \mathcal{E}_K} \tau_{K,\sigma} \right)$$

and where for all $(K, L) \in \mathcal{T}_\Gamma$, $a_{K,K|L}$ is defined by $a_{K,K|L} = U_1(a_K, a_L, \tau_{K,K|L}, \tau_{L,K|L})$ and $a_{L,K|L} = U_2(a_K, a_L, \tau_{K,K|L}, \tau_{L,K|L})$. (The functions U_1 and U_2 are defined in Lemma 2.5.)

Lemma 2.5 implies that the function $H_K(a,(a_L)_{L\in\mathcal{T}})$ is nondecreasing with respect to a and to a_L for all $L \in \mathcal{T}$ (including the case $L = K$).

Let us prove the above proposition by induction on n. It is true for $n = 0$. We assume that is true for n, and that there is $K_{\max} \in \mathcal{T}$ such that $K_{\max} = \max_{K\in\mathcal{T}}(u_K^{n+1})$ and $u_{K_{\max}}^{n+1} > 1$. Using the monotony of the function $H_{K_{\max}}$, we have

$$1 < u_{K_{\max}}^{n+1} \le H_{K_{\max}}(1,(u_{K_{\max}}^{n+1})_{L\in\mathcal{T}}) = \frac{1+\lambda_{K_{\max}}u_{K_{\max}}^{n+1}}{1+\lambda_{K_{\max}}}.$$

We then get a contradiction with the existence of such a $K_{\max}$. In the same way, we prove that there is no $K_{\min} \in \mathcal{T}_i$ such that $K_{\min} = \min_{K\in\mathcal{T}}(u_K^{n+1})$ and $u_{K_{\min}}^{n+1} < 0$. □

2.5. Existence and uniqueness of a discrete solution.

Proposition 2.8. *Under Assumption* 1.1*, let $\mathcal{D}$ be an admissible discretization of the domain $\Omega \times (0,T)$ (see Definition* 2.3*). Then for all $n \in \{0\dots M\}$, there exists one and only one solution $u_{\mathcal{T}}^{n+1} \in \mathcal{X}(\mathcal{T})$ to the system* (2.2)–(2.4).

Proof. The system composed of (2.2)–(2.4) can be seen as a system with unknowns $(u_K^{n+1})_{K\in\mathcal{T}}$ thanks to Lemma 2.5.

We set $N = \mathrm{card}(\mathcal{T})$ and we consider the application $\psi : \mathbb{R}^N \times [0,1] \to \mathbb{R}^N$ defined by $((u_K)_{K\in\mathcal{T}},\lambda) \mapsto (v_K)_{K\in\mathcal{T}}$ with for all $K \in \mathcal{T}$

$$v_K = m(K)\phi_i\frac{u_K - u_K^n}{\delta t} + \lambda\sum_{L\in N(K)}\tau_{K|L}\left(\varphi_i(u_K) - \varphi_i(u_L)\right) + \lambda\sum_{\sigma\in\mathcal{E}_\Gamma\bigcap\mathcal{E}_K}\tau_{K,\sigma}\left(\varphi_i(u_K) - \varphi_i(u_{K,\sigma})\right),$$

where for all $(K,L) \in \mathcal{T}_\Gamma$ we take $u_{K,K|L} = U_1(u_K,u_L,\tau_{K,K|L},\tau_{L,K|L})$ and $u_{L,K|L} = U_2(u_K,u_L,\tau_{K,K|L},\tau_{L,K|L})$. (The functions U_1 and U_2 are defined in Lemma 2.5.)

The function ψ is continuous with respect to each one of its arguments. Moreover, reproducing the proof of the Proposition 2.7 we can prove that for all $\lambda \in [0,1]$, $\psi((u_K)_{K\in\mathcal{T}},\lambda) = (0)_{K\in\mathcal{T}}$ implies $u_K \in [0,1]$ for all $K \in \mathcal{T}$. Since $\psi((u_K)_{K\in\mathcal{T}},0)$ is linear, an argument based on the topological degree (see [11] and references therein) implies that $\psi((u_K)_{K\in\mathcal{T}},1) = (0)_{K\in\mathcal{T}}$ admits at least one solution.

Turning now to the proof of uniqueness, we assume that for a given $n \in \{0\dots M\}$, $(u_K)_{K\in\mathcal{T}}$ and $(\tilde{u}_K)_{K\in\mathcal{T}}$ are two solutions of (2.2)–(2.4). Using for all $K \in \mathcal{T}$, the functions H_K defined in the proof of Proposition 2.7, we get that

$$\max(u_K,\tilde{u}_K) \le H_K(u_K^n,(\max(u_L,\tilde{u}_L))_{L\in\mathcal{T}})$$

and

$$\min(u_K,\tilde{u}_K) \ge H_K(u_K^n,(\min(u_L,\tilde{u}_L))_{L\in\mathcal{T}}).$$

If we multiply the above inequalities by $(1+\lambda_K)m(K)\phi_i$, if we substract the second inequality from the first one, and if we sum the result over $K \in \mathcal{T}$, the exchange terms between all the pairs of neighboring grid blocks and in particular the terms including λ_K vanish, and we obtain

$$\sum_{i=1,2}\sum_{K\in\mathcal{T}_i} m(K)\phi_i|u_K - \tilde{u}_K| \le 0,$$

which proves the uniqueness of the solution. □

2.6. Convergence. The remaining part of this section is devoted to the convergence proof of the scheme (2.2)–(2.4). The first step consists in obtaining some compactness properties for the sequence of approximated solutions. This will be done thanks to Kolmogorov's theorem. In particular this theorem requires that the space and time translates of the approximated solutions remain bounded.

2.6.1. Upper bound on the space translates.

PROPOSITION 2.9. *Under Assumption* 1.1, *let $\mathcal{D}$ be an admissible discretization of the domain $\Omega \times (0,T)$ in the sense of Definition* 2.3. *Let $u_{\mathcal{D}} \in \mathcal{X}(\mathcal{D})$ be the solution of* (2.2)–(2.4). *Then, there is $C_1 > 0$ only depending on η_j, π_j, Ω_j, $j \in \{1,2\}$ such that*

$$
\begin{aligned}
0 &\le \sum_{n=0}^{M} \& \sum_{(K,L)\in\mathcal{E}_\Gamma} \tau_{K,K|L} \left(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,K|L}^{n+1})\right) \left(\pi_1(u_K^{n+1}) - \pi_2(u_L^{n+1})\right) \\
&= \sum_{n=0}^{M} \& \sum_{(K,L)\in\mathcal{E}_\Gamma} \tau_{L,K|L} \left(\varphi_2(u_{L,K|L}^{n+1}) - \varphi_2(u_L^{n+1})\right) \left(\pi_1(u_K^{n+1}) - \pi_2(u_L^{n+1})\right) \le C_1
\end{aligned}
\tag{2.9}
$$

and for $i \in \{1,2\}$ there exists $C_2 > 0$ depending on C_1 and on C_η such that

$$
|\varphi_i(u_{\mathcal{D}})|^2_{1,\mathcal{D},i} \le C_2\,. \tag{2.10}
$$

Proof. For $n \in \{0\dots M\}$ and $K \in \mathcal{T}_i$, we multiply the equation (2.3) by $\pi_i(u_K^{n+1})$ and we sum over the discretization $\mathcal{D}$. This leads to

$$
\sum_{\substack{i=1\dots 2,\\ n=0\dots M,\\ K\in\mathcal{T}_i}} \left(\begin{array}{l} \Bigg(m(K)\phi_i(u_K^{n+1} - u_K^n) + \& \Bigg(\displaystyle\sum_{L\in N(K)} \tau_{K|L} \left(\varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1})\right) \\ \quad + \displaystyle\sum_{\sigma\in\mathcal{E}_\Gamma \bigcap \mathcal{E}_K} \tau_{K,\sigma} \left(\varphi_i(u_K^{n+1}) - \varphi_i(u_{K,\sigma}^{n+1})\right) \Bigg) \Bigg) \pi_i(u_K^{n+1}) \end{array} \right) = 0.
$$

Accumulation term. Since the function $\pi_i(.)$ is nondecreasing, the function g_i defined by $g_i(u) = \int_0^u \pi_i(a)\,da$ is therefore convex. So we have

$$
(u_K^{n+1} - u_K^n)\pi_i(u_K^{n+1}) \ge g_i(u_K^{n+1}) - g_i(u_K^n).
$$

Thus we get

$$
\sum_{n=0}^{M} \sum_{K\in\mathcal{T}_i} m(K)\phi_i(u_K^{n+1} - u_K^n)\pi_i(u_K^{n+1}) \ge \sum_{K\in\mathcal{T}_i} m(K)\phi_i(g_i(u_K^{M+1}) - g_i(u_K^0)).
$$

Moreover, we notice that

$$
\left| \sum_{K\in\mathcal{T}_i} m(K)\phi_i(g_i(u_K^{M+1}) - g_i(u_K^0)) \right| \le m(\Omega_i)\Big(\int_0^1 |\pi_i(a)|\,da \Big).
$$

Diffusion term. As $\varphi_i(b) - \varphi_i(a) \le C_\eta \int_a^b \pi_i^{'}(u)\,du$, we have

$$
\begin{aligned}
&\sum_{n=0}^{M} \& \sum_{K|L\in\mathcal{E}_{int,i}} \tau_{K|L} \left(\varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1})\right) \left(\pi_i(u_K^{n+1}) - \pi_i(u_L^{n+1})\right) \\
&\ge \frac{1}{C_\eta} \sum_{n=0}^{M} \& \sum_{K|L\in\mathcal{E}_{int,i}} \tau_{K|L} \Big(\varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1})\Big)^2.
\end{aligned}
$$

For $(K,L) \in \mathcal{T}_\Gamma$, we apply (2.5). This leads to

$$\tau_{K,\sigma}(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,\sigma}^{n+1}))(\pi_1(u_K^{n+1}) - \pi_2(u_L^{n+1})) \geq 0.$$

Finally, gathering the lower and upper bounds we obtained, we get

$$\sum_{i=1}^{2} |\varphi_i(u_\mathcal{D})|^2_{1,\mathcal{D},i} \leq C_\eta \sum_{i=1}^{2} m(\Omega_i)\Big(\int_0^1 |\pi_i(a)|\, da\Big) = C_2$$

and

$$\begin{aligned}0 &\leq \sum_{n=0}^{M} \delta t \sum_{\sigma=K|L\in\mathcal{E}_\Gamma} \tau_{K,\sigma}\left(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,\sigma}^{n+1})\right)\left(\pi_1(u_K^{n+1}) - \pi_2(u_L^{n+1})\right) \\ &\leq \sum_{i=1}^{2} m(\Omega_i)\Big(\int_0^1 |\pi_i(a)|\, da\Big) = C_1\,,\end{aligned}$$

which concludes the proof. □

We recall the following result, given in [11].

LEMMA 2.10. *Under Assumption* 1.1*, let $\mathcal{D}$ be an admissible discretization of the domain $\Omega \times (0,T)$ in the sense of Definition* 2.3*. Let $u_\mathcal{D} \in \mathcal{X}(\mathcal{D})$ be given by* (2.2)–(2.4)*. Let $i = 1,2$ and $\xi \in \mathbb{R}^d$. We define the domain $\Omega_{i,\xi}$ by*

$$\Omega_{i,\xi} = \{x \in \Omega_i \ / \ [x, x+\xi] \subset \Omega_i\}.$$

Then the function $\varphi_i(u_\mathcal{D})$ satisfies

$$\int_0^T\!\!\int_{\Omega_{i,\xi}} |\varphi_i(u_\mathcal{D}(x+\xi,t) - \varphi_i(u_\mathcal{D}(x,t)|^2 dxdt \leq |\xi|\Big(|\xi| + 2\text{size}(\mathcal{M})\Big)|\varphi_i(u_\mathcal{D})|^2_{1,\mathcal{D},i}. \tag{2.11}$$

This result produces the following proposition.

PROPOSITION 2.11. *Under Assumption* 1.1*, let $\mathcal{D}$ be an admissible discretization of the domain $\Omega \times (0,T)$ in the sense of Definition* 2.3*. Let $u_\mathcal{D} \in \mathcal{X}(\mathcal{D})$ be given by* (2.2)–(2.4)*. Let $i = 1,2$ and ω_i be an open bounded subset of Ω_i with a regular boundary. We define the function $\varphi_{\mathcal{D},\omega_i}$ by $\varphi_{\mathcal{D},\omega_i}(x,t) = \varphi_i(u_\mathcal{D}(x,t))$ for a.e. $(x,t) \in \omega_i \times (0,T)$, $\varphi_{\mathcal{D},\omega_i}(x,t) = 0$ if $(x,t) \notin \omega_i \times (0,T)$. Then there exists $C_3 > 0$, only depending on T, η_j, π_j, Ω_j, $j \in \{1,2\}$ and of ω_i, such that*

$$\|\varphi_{\mathcal{D},\omega_i}(.+\xi,.) - \varphi_{\mathcal{D},\omega_i}\|^2_{L^2(\mathbb{R}^{d+1})} \leq C_3\,|\xi|\Big(|\xi|+1\Big) \textit{ for all } \xi \in \mathbb{R}^d. \tag{2.12}$$

Proof. This result is a direct consequence of Proposition 2.9 and of Lemma 2.10 and of the fact that the measure of $\{x \in \omega_i,\ [x,x+\xi] \not\subset \omega_i\}$ is bounded by $C_{\omega_i}|\xi|$. □

2.6.2. Upper bound on the time translates.

PROPOSITION 2.12. *Under Assumption* 1.1*, let $\mathcal{D}$ be an admissible discretization of the domain $\Omega \times (0,T)$ in the sense of Definition* 2.3*. Let $u_\mathcal{D} \in \mathcal{X}(\mathcal{D})$ be given by* (2.2)–(2.4)*. Let $i = 1,2$ and ω_i be an open bounded subset of Ω_i with a regular boundary. We define the function $\varphi_{\mathcal{D},\omega_i}$ by $\varphi_{\mathcal{D},\omega_i}(x,t) = \varphi_i(u_\mathcal{D}(x,t))$ for a.e. $(x,t) \in \omega_i \times (0,T)$, $\varphi_{\mathcal{D},\omega_i}(x,t) = 0$ if $(x,t) \notin \omega_i \times (0,T)$. Then there exists $C_4 > 0$, only* *depending on* T*,* η_j*,* π_j*,* ϕ_j*,* Ω_j*,* $j \in \{1,2\}$ *and of* ω_i*, such that, for* $\mathrm{size}(\mathcal{M})$ *small enough,*

$$\text{(2.13)} \qquad \int_{\mathbb{R}} \int_{\Omega} \Big(\varphi_{\mathcal{D},\omega_i}(x,t+\tau) - \varphi_{\mathcal{D},\omega_i}(x,t) \Big)^2 dxdt \leq C_4 \, |\tau| \textit{ for all } \tau \in \mathbb{R}.$$

Proof. We suppose that $\tau \in (0,T)$ (the case $\tau < 0$ is deduced from $\tau > 0$ and the case $\tau > T$ is a consequence of an easy bound of $\int_{\mathbb{R}} \int_{\Omega} (\varphi_{\mathcal{D},\omega_i}(x,t+\tau) - \varphi_{\mathcal{D},\omega_i}(x,t))^2) dxdt$). Let $i = 1,2$ and let $\Theta_i \in C_c^\infty(\Omega_i,[0,1])$ be such that for all $x \in \omega_i$, $\Theta_i(x) = 1$. We suppose that $\mathrm{size}(\mathcal{M})$ is small enough so that Θ_i vanishes on all $K \in \mathcal{T}_i$ having edges on the boundary of Ω_i. For all $K \in \mathcal{T}_i$, we set $\Theta_{i,K} = \frac{1}{m(K)} \int_K \Theta_i(x) \, dx$.

Since the function φ_i is Lipschitz continuous, we have

$$\int_0^{T-\tau} \int_{\Omega} \Theta_i(x) \phi_i \Big(\varphi_i(u_{\mathcal{D}}(x,t+\tau)) - \varphi_i(u_{\mathcal{D}}(x,t)) \Big)^2 dxdt \leq L_\varphi \int_0^{T-\tau} A(t) \, dt$$

with

$$A(t) = \int_{\Omega} \Theta_i(x) \phi_i \Big(\varphi_i(u(x,t+\tau)) - \varphi_i(u(x,t)) \Big) \Big(u(x,t+\tau) - u(x,t) \Big) \, dx.$$

Following the method used in [11], we first write $A(t)$ as

$$A(t) = \sum_{K \in \mathcal{T}_i} \left(m(K) \Theta_{i,K} \phi_i \Big(\varphi_i(u_K^{n_1(t)+1}) - \varphi_i(u_K^{n_0(t)+1}) \Big) \sum_{n=0}^{M} \mathcal{X}_n(t,t+\tau)(u_K^{n+1} - u_K^n) \right),$$

where the indices $n_0(t)$ and $n_1(t)$ satisfy $n_0(t)\delta t < t \leq (n_0(t)+1)\delta t$, $n_1(t)\delta t < t + \tau \leq (n_1(t)+1)\delta t$ and the function $\mathcal{X}_n(a,b)$ is such that $\mathcal{X}_n(a,b) = 1$ if $a < b$ and $n\delta t \in [a,b[$, and $\mathcal{X}_n(a,b) = 0$ otherwise.

Using the definition of the scheme, we get

$$\begin{aligned} A(t) = \quad & \sum_{K \in \mathcal{T}_i} \left(\Theta_{i,K} \Big(\varphi_i(u_K^{n_1(t)+1}) - \varphi_i(u_K^{n_0(t)+1}) \Big) \right. \\ & \left. \qquad \sum_{n=0}^{M} \mathcal{X}_n(t,t+\tau) \sum_{L \in N(K)} \delta t \tau_{K|L} \Big(\varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1}) \Big) \right). \end{aligned}$$

Gathering the terms by edges leads to

$$\begin{aligned} A(t) = \quad & \sum_{n=0}^{M} \delta t \mathcal{X}_n(t,t+\tau) \sum_{K|L \in \mathcal{E}_{int,i}} \tau_{K|L} \left[\begin{array}{c} \Theta_{i,K} \Big(\varphi_i(u_K^{n_1(t)+1}) - \varphi_i(u_K^{n_0(t)+1}) \Big) \\ -\Theta_{i,L} \Big(\varphi_i(u_L^{n_1(t)+1}) - \varphi_i(u_L^{n_0(t)+1}) \Big) \end{array} \right] \\ & \times \Big(\varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1}) \Big). \end{aligned}$$

Applying the equality $2(\Theta_{i,K} a - \Theta_{i,L} b) = (\Theta_{i,K} + \Theta_{i,L})(a-b) + (\Theta_{i,K} - \Theta_{i,L})(a+b)$ we get that

$$A(t) \leq A_0(t) + A_1(t) + A_2(t)$$

with

$$\begin{aligned} A_0(t) = \quad & \sum_{n=0}^{M} \delta t \mathcal{X}_n(t,t+\tau) \sum_{K|L \in \mathcal{E}_{int,i}} \tau_{K|L} \left| \varphi_i(u_K^{n_1(t)+1}) - \varphi_i(u_L^{n_1(t)+1}) \right| \\ & \times \left| \varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1}) \right|, \end{aligned}$$

$$\begin{aligned} A_1(t) = \; & \sum_{n=0}^{M} \delta t \mathcal{X}_n(t, t+\tau) \sum_{K|L \in \mathcal{E}_{int,i}} \tau_{K|L} \left| \varphi_i(u_K^{n_0(t)+1}) - \varphi_i(u_L^{n_0(t)+1}) \right| \\ & \times \left| \varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1}) \right| \end{aligned}$$

and

$$A_2(t) = \sum_{n=0}^{M} \delta t \mathcal{X}_n(t, t+\tau) \sum_{K|L \in \mathcal{E}_{int,i}} \tau_{K|L} L_\varphi \left| \Theta_{i,K} - \Theta_{i,L} \right| \; \left| \varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1}) \right|.$$

We then use Young's inequality, Proposition 2.9, and the regularity of the function Θ to bound $A_0(t)$, $A_1(t)$, and $A_2(t)$ by a sum of terms under the form $\sum_{n=0}^{M} \delta t \mathcal{X}_n(t, t+\tau) a^n$, $\sum_{n=0}^{M} \delta t \mathcal{X}_n(t, t+\tau) a^{n_0(t)}$, and $\sum_{n=0}^{M} \delta t \mathcal{X}_n(t, t+\tau) a^{n_1(t)}$ such that $0 \leq a^n$ for all $n = 0 \ldots, M$ and such that $\delta t \sum_{n=0}^{M} a^n$ is bounded independently on the discretization. We then use the properties

$\int_0^{T-\tau} \sum_{n=0}^{M} \delta t \mathcal{X}_n(t, t+\tau) a^n dt \leq \tau \delta t \sum_{n=0}^{M} a^n$,
$\int_0^{T-\tau} \sum_{n=0}^{M} \delta t \mathcal{X}_n(t, t+\tau) a^{n_0(t)} dt \leq \tau \delta t \sum_{n=0}^{M} a^n$, and
$\int_0^{T-\tau} \sum_{n=0}^{M} \delta t \mathcal{X}_n(t, t+\tau) a^{n_1(t)} dt \leq \tau \delta t \sum_{n=0}^{M} a^n$, proven in [11]. □

2.6.3. Upper bound on the discrete $L^2(0,T;H^1(\Omega))$-seminorm of the function $w_{\mathcal{D}}$.

Let $u_{\mathcal{D}}$ be given by (2.2)–(2.4). We consider $w_{\mathcal{D}}$ defined by $w_K^{n+1} = \Psi(\hat{\pi}_i(u_K^{n+1}))$ for all $i = 1, 2$ and $K \in \mathcal{T}_i$. The following proposition states that the discrete $L^2(0,T;H^1(\Omega))$-seminorm of the function $w_{\mathcal{D}}$ remains bounded. We first recall the definition of this seminorm defined on the whole domain Ω.

Definition 2.13. *Let $\Omega \times (0,T)$ be a domain satisfying* H1-1 *and $\mathcal{D}$ be an admissible discretization of this domain in the sense of Definition* 2.3. *The $L^2(0,T;H^1(\Omega))$-seminorm of a function $u_{\mathcal{D}} \in \mathcal{X}(\mathcal{D})$ is defined by*

$$|u_{\mathcal{D}}|_{1,\mathcal{D}}^2 = \sum_{n=0}^{M} \delta t \sum_{K|L \in \mathcal{E}_{int}} \tau_{K|L} (\delta u_{K,L}^{n+1})^2 = \sum_{i=1,2} |u_{\mathcal{D}}|_{1,\mathcal{D},i}^2 + \sum_{n=0}^{M} \delta t \sum_{(K,L) \in \mathcal{T}_\Gamma} \tau_{K|L} (\delta u_{K,L}^{n+1})^2.$$

Proposition 2.14. *Under Assumption* 1.1, *let $\mathcal{D}$ be an admissible discretization in the sense of Definition* 2.3. *Let $u_{\mathcal{D}} \in \mathcal{X}(\mathcal{D})$ be the solution of* (2.2)–(2.4). *Then, there exists $C_5 > 0$ only depending on η_j, π_j, Ω_j, $j \in \{1, 2\}$ such that*

$$|w_{\mathcal{D}}|_{1,\mathcal{D}}^2 \leq C_5 \,. \tag{2.14}$$

Proof. For $K \in \mathcal{T}_i$ and $L \in N(K)$, using the property of Lipschitz continuity of $\Psi \circ \hat{\pi}_i \circ \varphi_i^{(-1)}$ (see Lemma 1.2), we get

$$(w_K^{n+1} - w_L^{n+1})^2 \leq (\varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1}))^2$$

and therefore we deduce from (2.10)

$$|w_{\mathcal{D}}|_{1,\mathcal{D},i}^2 \leq C_2 \,.$$

We now consider the case $(K, L) \in \mathcal{T}_\Gamma$. We have, since $\hat{\pi}_1(u_{K,K|L}^{n+1}) = \hat{\pi}_2(u_{L,K|L}^{n+1})$,

$$\begin{aligned} \tau_{K|L} (\Psi(\hat{\pi}_1(u_K^{n+1})) - \Psi(\hat{\pi}_2(u_L^{n+1})))^2 \leq \; & \tau_{K,K|L} (\Psi(\hat{\pi}_1(u_K^{n+1})) - \Psi(\hat{\pi}_1(u_{K,K|L}^{n+1})))^2 \\ & + \tau_{L,K|L} (\Psi(\hat{\pi}_2(u_{L,K|L}^{n+1})) - \Psi(\hat{\pi}_2(u_L^{n+1})))^2, \end{aligned}$$

thanks to the convexity of the function $x \mapsto x^2$ and to $1/\tau_{K|L} = 1/\tau_{K,K|L} + 1/\tau_{L,K|L}$. We again use the properties of $\Psi \circ \hat{\pi}_i \circ \varphi_i^{(-1)}$ (see Lemma 1.2):

$$(\Psi(\hat{\pi}_1(u_K^{n+1})) - \Psi(\hat{\pi}_1(u_{K,K|L}^{n+1})))^2 \le (\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,K|L}^{n+1}))^2$$

and

$$(\Psi(\hat{\pi}_2(u_{L,K|L}^{n+1})) - \Psi(\hat{\pi}_2(u_L^{n+1})))^2 \le (\varphi_2(u_L^{n+1}) - \varphi_2(u_{L,K|L}^{n+1}))^2.$$

Now, using (2.5), we have, for all $(K, L) \in \mathcal{T}_\Gamma$,

$$\begin{aligned} &\Big(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,K|L}^{n+1})\Big)^2 \\ &\le \Big(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,K|L}^{n+1})\Big) C_\eta \Big(\pi_1(u_K^{n+1}) - \pi_1(u_{K,K|L}^{n+1})\Big) \\ &\le \Big(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,K|L}^{n+1})\Big) C_\eta \Big(\pi_1(u_K^{n+1}) - \pi_2(u_L^{n+1})\Big). \end{aligned} \tag{2.15}$$

Then, from (2.9) and (2.15), we get

$$\sum_{n=0}^{M} \& \sum_{(K,L)\in\mathcal{T}_\Gamma} \tau_{K,K|L} \Big(\Psi(\hat{\pi}_1(u_K^{n+1})) - \Psi(\hat{\pi}_1(u_{K,K|L}^{n+1}))\Big)^2 \le C_\eta C_1\,,$$

and in the same way

$$\sum_{n=0}^{M} \& \sum_{(K,L)\in\mathcal{T}_\Gamma} \tau_{L,K|L} \Big(\Psi(\hat{\pi}_2(u_{L,K|L}^{n+1})) - \Psi(\hat{\pi}_2(u_L^{n+1}))\Big)^2 \le C_\eta C_1\,.$$

Thus we get

$$\sum_{n=0}^{M} \& \sum_{(K,L)\in\mathcal{T}_\Gamma} \tau_{K|L} (w_K^{n+1} - w_L^{n+1})^2 \le 2C_1\, C_\eta.$$

Gathering the above results proves that there exists $C_6 > 0$, only depending on η_j, π_j, Ω_j, $j \in \{1, 2\}$ such that

$$|w_{\mathcal{D}}|_{1,\mathcal{D}}^2 \le C_6 \qquad \square$$

2.6.4. Convergence of the scheme toward the weak problem. Thanks to the previous propositions, we are now able to prove the following theorem, which states the convergence of the scheme (2.2)–(2.4) toward a solution to the weak problem introduced in Definition 1.3.

THEOREM 2.15. *Under Assumption* 1.1*, let us consider a sequence* $(\mathcal{D}_m)_{m\in\mathbb{N}}$*, of admissible discretizations in the sense of Definition* 2.3*, such that there exists* $\alpha > 0$ *with* $\mathrm{regul}(\mathcal{M}_\mathrm{m}) \le \alpha$ *for all* $m \in \mathbb{N}$ *and such that* $\mathrm{size}(\mathcal{D}_m) \to 0$ *as* $m \to +\infty$*. Let* $u_{\mathcal{D}_m} = u_m \in \mathcal{X}(\mathcal{D}_m)$ *be the solution of* (2.2)–(2.4) *for* $\mathcal{D} = \mathcal{D}_m$*. Then there exists a subsequence of* $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$*, again denoted by* $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$*, and a weak solution* u *of problem* (1.5)–(1.9) *in the sense of Definition* 1.3*, such that* $u_m \to u$ *in* $L^p(\Omega \times (0, T))$ *for all* $p < \infty$*.*

Remark 2.16. A proof that the problem (1.5)–(1.9) admits at most one regular solution can be obtained following the method of [5]. A uniqueness result on the solution of the weak problem given in Definition 1.3 implies that the whole sequence of discrete solutions converges.

Proof.

Step 1: Existence of a convergent subsequence of $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$. For any open subset ω_i of Ω_i, $i = 1, 2$, Propositions 2.7, 2.11, and 2.12 ensure that the hypotheses of Kolmogorov's theorem are satisfied. We thus get the existence of a subsequence of $(\varphi_{\mathcal{D}_m,\omega_i})_{m\in\mathbb{N}}$, converging in $L^2(\omega_i \times (0,T))$ to some function $\varphi_{\omega_i} \in L^2(\omega_i \times (0,T))$. Using an increasing sequence of domains $\omega_{i,k}$ which converges toward Ω_i, we can extract, thanks to a diagonal process, a subsequence again denoted by $(\mathcal{D}_m, u_m)_{m\in\mathbb{N}}$ such that $(\varphi_{\mathcal{D}_m,\omega_{i,m}})_{m\in\mathbb{N}}$ converges in $L^2(\omega_{i,k} \times (0,T))$ for all $k \in \mathbb{N}$, to some bounded function $\tilde{\varphi}_i \in L^2(\omega_{i,k} \times (0,T))$ for all $k \in \mathbb{N}$. We then obtain that $(\varphi_i(u_m))_{m\in\mathbb{N}}$ converges in $L^2(\Omega_i \times (0,T))$ to $\tilde{\varphi}_i$. Since φ_i is continuous and strictly increasing, this implies that, up to a subsequence, $(u_m)_{m\in\mathbb{N}}$ converges toward a function $u_i \in L^2(\Omega_i \times (0,T)) \bigcap L^\infty(\Omega_i \times (0,T))$ for all $i \in \{1, 2\}$.

To prove that $\varphi_i(u_i) \in L^2(0,T; H^1(\Omega_i))$ for all $i \in \{1, 2\}$, it is sufficient to show that $\frac{\partial\varphi_i(u_i)}{\partial x} \in L^2(\Omega_i \times (0,T))$. Let $m \in \{0 \dots M\}$, $\psi_i \in C_c^\infty(\Omega_i \times (0,T))$ and $\epsilon > 0$ be such that $\mathrm{supp}(\psi_i) = \{(x,t) \in \Omega_i \times (0,T) \ / \ \mathrm{dist(x, \mathbb{R}^d \setminus \Omega_i)} \leq \epsilon\}$. Using the Cauchy–Schwarz inequality and Lemma 2.10, we have for all $|\xi| \leq \epsilon$

$$\int_{\Omega_{i,\xi}\times(0,T)} \Big(\varphi_i(u_m(x+\xi,t)) - \varphi_i(u_m(x,t))\Big)\psi_i(x,t)dxdt$$
$$\leq \Big(|\xi|(|\xi| + 2\mathrm{size}(\mathcal{M}_\mathrm{m}))\mathrm{C}_2\Big)^{\frac{1}{2}} \|\psi_\mathrm{i}\|_{\mathrm{L}^2(\Omega_\mathrm{i}\times(0,\mathrm{T}))}.$$

Passing to the limit and after a change of variable we obtain

$$\int_{\Omega_{i,\xi}\times(0,T)} \Big(\psi_i(x-\xi,t) - \psi_i(x,t)\Big)\varphi_i(u_i(x,t))dxdt \tag{2.16}$$
$$\leq |\xi|(C_2\,)^{\frac{1}{2}}||\psi_i||_{L^2(\Omega_i\times(0,T))}.$$

Now if we denote by $\{e_i, i = 1 \dots d\}$ the canonical basis of $\mathbb{R}^d$ and if we take $\xi = \lambda e_i$, $i \in \{1 \dots d\}$ with $|\lambda| < \epsilon$ in (2.16), we then have as $\epsilon \to 0$

$$-\int_{\Omega_{i,\xi}\times(0,T)} \frac{\partial\psi_i(x,t)}{\partial x_i}\varphi_i(u_i(x,t))dxdt \leq (C_2\,)^{\frac{1}{2}}\|\psi_i\|_{L^2(\Omega_i\times(0,T))}$$
$$\text{for all } \psi_i \in C_c^\infty(\Omega_i \times (0,T)),$$

which implies that $\frac{\partial\varphi_i(u_i)}{\partial x} \in L^2(\Omega_i \times (0,T))$.

Step 2: u is a weak solution to the problem (1.5)–(1.9). Let us consider $\tilde{C}_{test} = \{h \in C^2(\overline{\Omega} \times [0,T]) \ / \ h(.,T) = 0\}$ which is dense in C_{test}. Let $\psi \in \tilde{C}_{test}$ and, for $m \in \mathbb{N}$, let u_m be given by (2.2)–(2.4) for $\mathcal{D} = \mathcal{D}_m$. For all $n \in \{0 \dots M\}$ and for all $K \in \mathcal{T}$, we multiply the equation (2.3) by $\psi_K^n = \psi(x_K, n\delta t)$, and we sum these equalities over the volume control set and $n = 0, \dots, M$. We get $\sum_{i=1}^2(E_{i,1,m} +$

$E_{i,2,m}) + E_{1|2,m} = 0$ with

$$E_{i,1,m} = \sum_{n=0}^{M} \sum_{K\in\mathcal{T}_i} m(K)\phi_i(u_K^{n+1} - u_K^n)\psi_K^n,$$
$$E_{i,2,m} = -\sum_{n=0}^{M} \delta t \sum_{K\in\mathcal{T}_i} \sum_{L\in N(K)} \tau_{K|L}\Big(\varphi_i(u_L^{n+1}) - \varphi_i(u_K^{n+1})\Big)\psi_K^n,$$
$$E_{1|2,m} = \sum_{n=0}^{M} \delta t \sum_{(K,L)\in\mathcal{T}_\Gamma} \tau_{K,K|L}\left(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,K|L}^{n+1})\right)\left(\psi_K^n - \psi_L^n\right).$$

Following some classical proofs (see [11]), we get that

$$\lim_{m\to+\infty} E_{i,1,m} = -\int_0^T \int_{\Omega_i} \phi_i u_i(x,t)\psi_t(x,t)dxdt - \int_{\Omega_i} \phi_i u_{\mathrm{ini}}(x)\psi(x,0)dx.$$

Convergence of $\boldsymbol{E_{i,2,m}}$. Gathering the terms by edges in $E_{i,2,m}$ leads to

$$E_{i,2,m} = \sum_{n=0}^{M} \delta t \sum_{\sigma=K|L\in\mathcal{E}_{int,i}} \tau_{K|L}\Big(\varphi_i(u_K^{n+1}) - \varphi_i(u_L^{n+1})\Big)\Big(\psi_K^n - \psi_L^n\Big).$$

We apply the method presented, for example, in [10] (which is a discrete version of a strong-weak convergence) to conclude that

$$\lim_{m\to+\infty} E_{i,2,m} = \int_0^T \int_{\Omega_i} \nabla\varphi_i(u_i)(x,t).\nabla\psi(x,t)\,dx\,dt. \tag{2.17}$$

Convergence of $\boldsymbol{E_{1|2,m}}$. We have

$$E_{1|2,m}^2 \le \left(\sum_{n=0}^{M} \delta t \sum_{(K,L)\in\mathcal{T}_\Gamma} \tau_{K,K|L}\Big(\varphi_1(u_K^{n+1}) - \varphi_1(u_{K,K|L}^{n+1})\Big)^2\right)$$
$$\times \left(\sum_{n=0}^{M} \delta t \sum_{(K,L)\in\mathcal{T}_\Gamma} m(K|L)\frac{(\psi_K^n - \psi_L^n)^2}{d_{K,K|L}}\right).$$

But we notice that, thanks to the regularity of the function ψ, there exists $C_\psi > 0$ such that $|\psi_K^n - \psi_L^n| \le C_\psi d_{K|L}$, which implies with (2.1)

$$\sum_{n=0}^{M} \delta t \sum_{(K,L)\in\mathcal{T}_\Gamma} m(K|L)\frac{(\psi_K^n - \psi_L^n)^2}{d_{K,K|L}} \le 4Tm(\Gamma)C_\psi^2\alpha\mathrm{size}(\mathcal{M}).$$

Thus, using (2.9) and (2.15), we get

$$\sum_{n=0}^{M} \delta t \sum_{(K,L)\in\mathcal{T}_\Gamma} \tau_{K,K|L}\Big(\varphi_i(u_K^{n+1}) - \varphi_i(u_{K,K|L}^{n+1})\Big)^2 \le C_\eta C_1\,.$$

Gathering the above results produces

$$\lim_{m\to+\infty} E_{1|2,m} = 0.$$

Step 3: Let us prove that $\boldsymbol{w \in L^2(0,T;H^1(\Omega))}$. Following the proofs of Lemma 2.10 and of $\varphi(u_i) \in L^2(0,T;H^1(\Omega_i))$ (see Step 1), we obtain that $w \in L^2(0,T;H^1(\Omega))$ using inequality (2.14). $\square$

As an immediate consequence of Theorem 2.15 we get the following corollary.

COROLLARY 2.17. *Under Assumption* 1.1, *problem* (1.5)–(1.9) *admits at least one weak solution in the sense of Definition* 1.3.

As an illustration of the previous results we now give numerical results.

3. Numerical results. Let us consider a domain Ω such that $\Omega_1 = (0,1)$ and $\Omega_2 = (1,2)$. The mobilities are given by

$$\eta_o(u) = \begin{cases} u & \text{if } 0 \le u \le 1, \\ 0 & \text{if } u < 0, \\ 1 & \text{otherwise,} \end{cases} \qquad \eta_w(u) = \begin{cases} 1-u & \text{if } 0 \le u \le 1, \\ 1 & \text{if } u < 0, \\ 0 & \text{otherwise,} \end{cases}$$

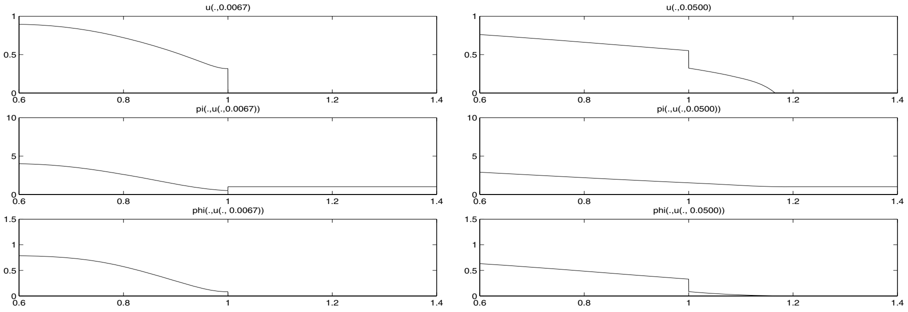
and the capillary pressure is given by

$$\pi_1(u) = \begin{cases} 5u^2 & \text{if } 0 \le u \le 1, \\ 0 & \text{if } u < 0, \\ 5 & \text{otherwise,} \end{cases} \qquad \pi_2(u) = \begin{cases} 5u^2+1 & \text{if } 0 \le u \le 1, \\ 1 & \text{if } u < 0, \\ 6 & \text{otherwise.} \end{cases}$$

In that case, $u_1^\star = \frac{1}{\sqrt{5}}$, $u_2^\star = \frac{2}{\sqrt{5}}$. For the initial condition we take

$$u_{\text{ini}}(x) = \begin{cases} 0.9 & \text{if} \qquad x < 0.9, \\ 0 & \text{otherwise.} \end{cases}$$

To discretize the domains Ω_i, we use a regular mesh such that $\text{dx} = \text{size}(\mathcal{M}) = 10^{-2}$ for all $i \in \{1,2\}$ and we use a constant time step $\delta t = \frac{1}{6}.10^{-3}$. Figure 3.1 represents functions $u(.,t)$, $\pi(.,u(.,t))$, $\varphi(.,u(.,t))$ for $t = 0.007$ and $t = 0.05$. In the first case



(a) (b)

FIG. 3.1. $u(.,t)$, $\pi(.,u(.,t))$, $\varphi(.,u(.,t))$ *for* $t = 0.007$ (a) *and* $t = 0.05$ (b).

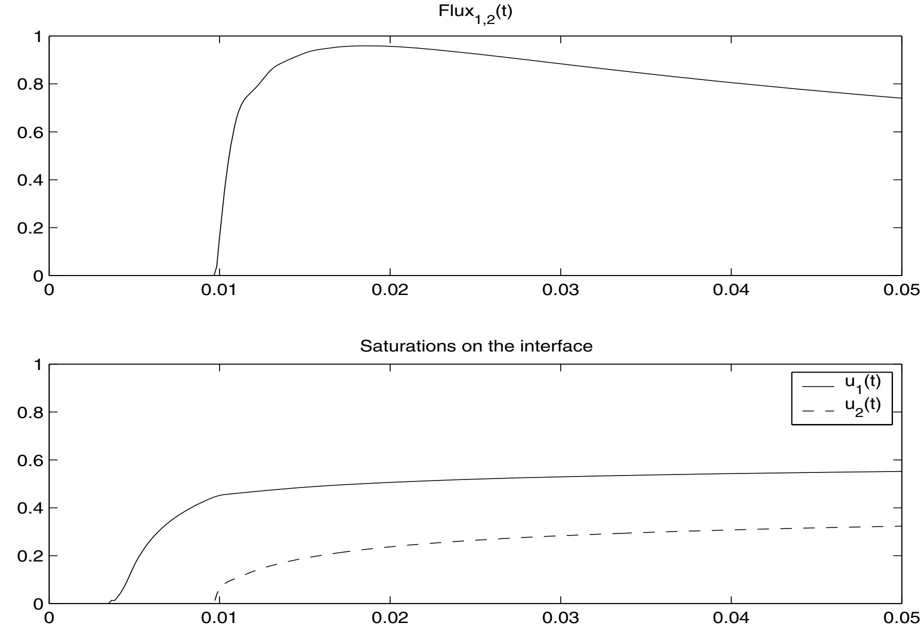
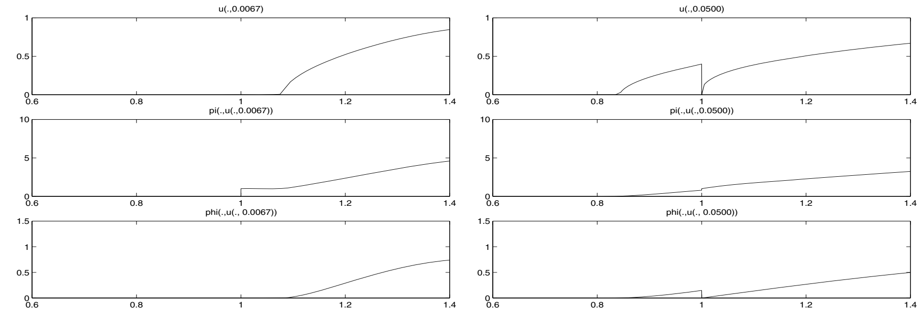


FIG. 3.2. *Evolution of the flux and of the saturations on the interface.*



(a) (b)

FIG. 3.3. $u(.,t)$, $\pi(.,u(.,t))$, $\varphi(.,u(.,t))$ *for* $t = 0.007$ (a) *and* $t = 0.05$ (b).

oil is trapped under the interface Γ located in $x = 1$ and the capillary pressure is discontinuous, whereas in the second case oil can flow through Γ and the continuity of the capillary pressure is ensured. Figure 3.2 represents the evolution of the flux and of the saturations on the interface Γ according to the time variable. We have also done tests with the initial condition

$$u_{\text{ini}}(x) = \begin{cases} 0.9 & \text{if} \qquad\qquad x > 1.2, \\ 0 & \text{otherwise,} \end{cases}$$

where oil already lies in the capillary barrier. Figures 3.3 and 3.4 show the results we obtained. We notice that although the capillary pressure is discontinuous, oil can flow

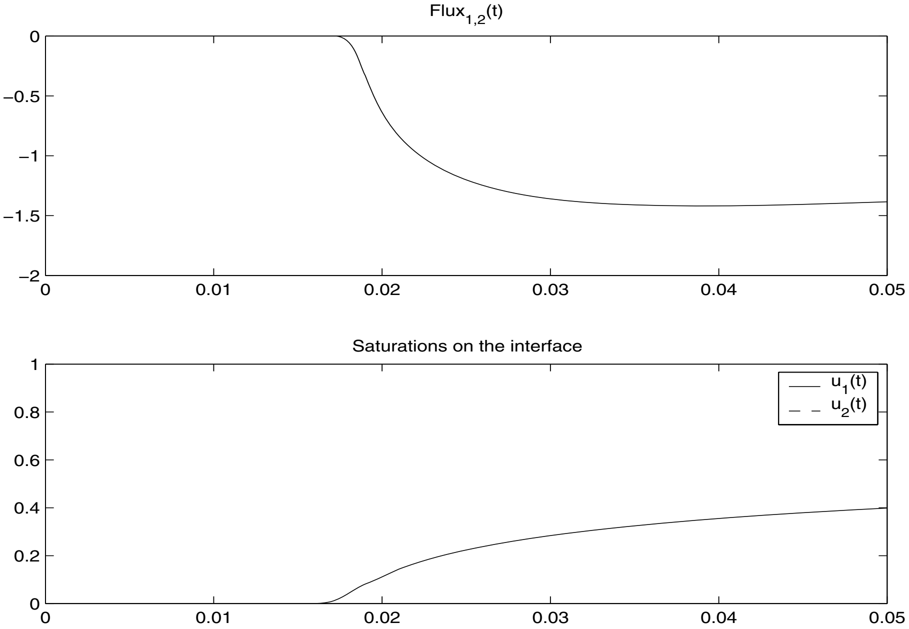


FIG. 3.4. *Evolution of the flux and of the saturations on the interface.*

through Γ from Ω_2 to Ω_1 while satisfying the conditions (2.4) since for all $t \in [0, 0.05]$, $u_2(t) = 0$.

4. Concluding remarks. In this paper we have established a convergence property for the scheme (2.2)–(2.4) toward a weak solution of the problem (1.5)–(1.9) in the sense of Definition 1.3. It remains to prove the uniqueness of such a weak solution. Further work will be done taking into account a total flux and the gravity gradient (see [8]).

REFERENCES

[1] H. W. ALT AND S. LUCKHAUS, *Quasilinear elliptic-parabolic differential equations*, Math. Z., (1983), pp. 311–341.

[2] K. AZIZ AND A. SETTARI, *Petroleum Reservoir Simulation*, Elsevier Applied Science Publishers, London, 1979.

[3] J. BEAR, *Dynamic of Fluids in Porous Media*, Dover, New York, 1972.

[4] J. BEAR, *Modeling Transport Phenomena in Porous Media*, Springer, New York, 1996, pp. 27–63.

[5] M. BERTSCH, R. D. PASSO, AND C. VAN DUIJN, *Analysis of oil trapping in porous media flow*, SIAM J. Math. Anal., 35 (2003), pp. 245–267.

[6] J. CARILLO, *Entropy solutions for nonlinear degenerate problems*, Arch. Ration. Mech. Anal., 147 (1999), pp. 269–361.

[7] G. CHAVENT AND J. JAFFRÉ, *Mathematical Models and Finite Elements for Reservoir Simulation*, Stud. Math. Appl., vol. 17, North–Holland, Amsterdam, 1986.

[8] G. ENCHÉRY, *Modèles et schémas numériques pour la simulation de bassin*, Ph.D. thesis, Université de Marne-La-Vallée, 2004.

[9] B. G. ERSLAND, M. S. ESPEDAL, AND R. NYBO, *Numerical methods for flow in a porous medium with internal boundaries*, Comput. Geosci., 2 (1998), pp. 217–240.

[10] R. EYMARD AND T. GALLOUËT, *H-convergence and numerical schemes for elliptic equations*, SIAM J. Numer. Anal., 41 (2003), pp. 539–562.

[11] R. Eymard, T. Gallouët, and R. Herbin, *The Finite Volume Method*, Ph. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2000.

[12] R. Eymard, T. Gallouët, D. Hilhorst, and Y. N. Slimane, *Finite volumes and nonlinear diffusion equations*, M2AN, 32 (1998), pp. 747–761.

[13] A. Michel, *Convergence de schémas volumes finis pour des problèmes de convection diffusion non linéaires*, Ph.D. thesis, Université de Provence, 2001.

[14] C. J. van Duijn, J. Molenaar, and M. J. de Neef, *The effect of capillary forces on immiscible two-phase flow in heterogeneous porous media*, Transport in Porous Media, 21 (1995), pp. 71–93.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2423–2449

ERROR ESTIMATES FOR FINITE VOLUME APPROXIMATIONS OF CLASSICAL SOLUTIONS FOR NONLINEAR SYSTEMS OF HYPERBOLIC BALANCE LAWS*

VLADIMIR JOVANOVIĆ[†] AND CHRISTIAN ROHDE[‡]

Abstract. We consider a general class of finite volume schemes on unstructured but quasi-uniform meshes for first-order *systems* of hyperbolic balance laws on unstructured meshes. Provided the system is equipped with at least one entropy-entropy flux tuple and the associated Cauchy problem allows for a classical solution u we give conditions such that the finite volume approximation u_h converges to u if the mesh parameter h tends to zero. In fact we prove an error estimate of the form $\|u - u_h\|_{L^2} \leq C\sqrt{h}$, where C is independent of h. The proof relies on a stability result for classical solutions in the class of entropy solutions due to Dafermos [*Arch. Rational Mech. Anal.*, 94 (1979), pp. 373–389] and DiPerna [*Indiana Univ. Math. J.*, 28 (1979), pp. 137–188].

Finally, we present examples such that the conditions to apply the general convergence estimate can be satisfied (at least in part). The examples cover general scalar equations, weakly coupled systems, and the system of elastodynamics in one dimension. Moreover, we generalize the concept of entropy conservative methods due to Tadmor [*Math. Comp.*, 49 (1987), pp. 91–103] and show how this can be used to establish the convergence of finite volume methods for the system's case.

Key words. hyperbolic conservation laws, entropy-entropy flux tuples, classical solutions, finite volume schemes on unstructured meshes, entropy conservative schemes

AMS subject classifications. 65M12, 35L60

DOI. 10.1137/S0036142903438136

1. Introduction.

Let us consider a nonlinear convection process for a vector valued state variable $u = u(x,t) \in \mathcal{U}$ of $m \in \mathbb{N}$ components, where the state space $\mathcal{U}$ is an open subset of $\mathbb{R}^m$, $x = (x_1, \ldots, x_d)^T$ denotes the vector of spatial coordinates and $t \geq 0$ stands for time. Assume that the dynamics for u in the time interval $[0,T)$, $T > 0$, are given by the solution of the Cauchy problem

$$\partial_t u + \sum_{i=1}^{d} \partial_i G_i(u) = B(u) \text{ in } \mathbb{R}^d \times (0,T), \tag{1.1}$$

$$u(.,0) = u_0 \text{ in } \mathbb{R}^d. \tag{1.2}$$

Here $G_i \in [C^2(\mathcal{U})]^m$, $i = 1, \ldots, m$, is the flux in x_i-direction and $B \in [C^1(\mathcal{U})]^m$ some source. We suppose that there exists a vector $\bar{u} \in \mathcal{U}$ with $B(\bar{u}) = 0$ and for the initial function we have $u_0 - \bar{u} \in [L^\infty(\mathbb{R}^d)]^m \cap [H^1(\mathbb{R}^d)]^m \cap [C^1(\mathbb{R}^d)]^m$.

As a structural prerequisite we suppose that (1.1) is equipped with an entropy-entropy flux tuple $(\eta, q_1, \ldots, q_d)^T \in C^3(\mathcal{U}) \times [C^2(\mathcal{U})]^d$, i.e., η is a function that is uniformly convex on compact sets and the consistency relation

$$(\nabla q_i(w))^T = (\nabla \eta(w))^T DG_i(w) \tag{1.3}$$

holds for $i = 1, \ldots, d$ and all $w \in \mathcal{U}$.

[*]Received by the editors November 27, 2003; accepted for publication (in revised form) July 5, 2005; published electronically January 6, 2006.
http://www.siam.org/journals/sinum/43-6/43813.html

[†]Faculty of Natural Sciences and Mathematics, University of Banja Luka, Mladena Stojanovića 2, 78000 Banja Luka, Bosnia and Herzegovina (vladimir@mathematik.uni-freiburg.de).

[‡]Fakultät für Mathematik, Universität Bielefeld, Postfach 100 131, D-33501 Bielefeld, Germany (chris@math.uni-bielefeld.de).

This work is devoted to classical solutions of (1.1). A function $u \in [C^1(\mathbb{R}^d \times [0,T])]^m$ is called a *classical solution* of (1.1), (1.2) if (1.1), (1.2) are satisfied pointwise everywhere. Throughout the paper we suppose that the following assumption is true.

Assumption 1.1. *There is classical solution u of the Cauchy problem* (1.1)*,* (1.2)*. In addition we have $\int_0^T \int_{\mathbb{R}^d} |Du|^2\,dxdt < \infty$, $\sup_{\mathbb{R}^d\times[0,T]} |Du| < \infty$ and there exists a convex compact set $S \subset\subset \mathcal{U}$ such that*

$$u(x,t),\ \bar{u} \in S \quad \textit{for all } (x,t) \in \mathbb{R}^d \times [0,T)]. \tag{1.4}$$

It is well known that in general classical solutions for the Cauchy problem (1.1), (1.2) can exist only for short times [23]. For important exceptions due to the presence and structure of the source term B we refer to the last paragraph of the introduction. As a not-so-strong notion of solutions, entropy solutions are considered. A function $v \in [L^\infty_{loc}(\mathbb{R}^d \times [0,T))]^m$ is called an *entropy solution* of (1.1), (1.2) if it is a weak solution of (1.1), (1.2) and if

(1.5)
$$\int_0^T \int_{\mathbb{R}^d} \eta(u)\partial_t\omega + \sum_{i=1}^d q_i(u)\partial_i\omega + \nabla\eta(u)\cdot B(u)\omega\,dxdt \geq -\int_{\mathbb{R}^d} \eta(u_0)\omega(.,0)\,dxdt$$

holds for all nonnegative functions $\omega \in C_0^{0,1}(\mathbb{R}^d \times [0,T))$. Here $C_0^{0,1}(\mathbb{R}^d \times [0,T))$ is the space of Lipschitz-continuous functions with compact support on $\mathbb{R}^d \times [0,T)$. There is no general multidimensional well-posedness theory for globally-in-time defined entropy solutions. The only result in this direction is the following stability result for classical solutions in the class of entropy solutions. We cite it from the book [5] (where it is proven for the case $B \equiv 0$).

Theorem 1.2 (see [5, Theorem 5.2.1]). *Let u from Assumption* 1.1 *be given. Suppose furthermore that $v \in [L^\infty(\mathbb{R}^d \times [0,T))]^m$ is an entropy solution of* (1.1) *with $v(.,0) = v_0$ for some $v_0 \in [L^\infty(\mathbb{R}^d)]^m$.*

If also v takes values in S there is a constant $\lambda > 0$ and a positive function $\mathcal{C} \in C^0([0,T])$ depending on S such that we have for all $t \in [0,T)$ and $R > 0$

$$\int_{\{|x|\leq R\}} |u(x,t) - v(x,t)|^2\,dx \leq \mathcal{C}(t)\int_{\{|x|\leq R+\lambda t\}} |u_0 - v_0|^2\,dx. \tag{1.6}$$

In particular we observe for $u_0 = v_0$ that classical solutions of (1.1)*,* (1.2) *are unique in the class of entropy solutions.*

We shall use this stability theory for classical solutions due to Dafermos [4] and DiPerna [6] to set up a convergence theory for finite volume schemes applied to the Cauchy problem (1.1), (1.2). We point out that our results hold only if (1.1), (1.2) is classically solvable.

In section 2 we present the precise setting of our problem and introduce an extension of Theorem 1.2: in Theorem 2.2, we estimate the L^2-difference between u and an *arbitrary discontinuous function* v in terms of some error measures. This theorem generalizes Theorem 1.2 to the inhomogeneous case. In section 3 we describe a class of standard finite volume schemes in d space dimensions on an unstructured but quasi-uniform mesh. Denote the mesh parameter by h which is the diameter of the largest volume. We assume that the numerical approximation $u_h : \mathbb{R}^d \times [0,T] \to \mathbb{R}^m$ is uniformly bounded in $[L^\infty(\mathbb{R}^d \times [0,T])]^m$ and satisfies a cell entropy inequality (Assumption 3.1).

These strong but realistic assumptions are then used in section 4 to establish the error bound. With our main result, Theorem 4.4, we prove an estimate of the form

$$\|u - u_h\|_{L^2(\mathbb{R}^d \times [0,T])} \leq C\sqrt{h}, \tag{1.7}$$

where C is some constant independent of h. The proof relies on the generalized stability estimate in Theorem 2.2. We point out that the estimate is given in terms of the L^2-norm (in space) rather than in the L^1-norm, as in the results of Kruzkov [19] and Kuznetsov [20] for scalar equations such that we cannot follow these more usual convergence proofs for finite volume schemes.

In the final part of the paper (section 5) we consider several applications of the convergence theorem. First we consider the example of weakly coupled hyperbolic systems and apply monotone numerical fluxes for each component. Then it is possible to satisfy all assumptions required in Theorem 4.4. The result improves the known estimates which give an estimate of order $h^{1/4}$ in the L^1-norm [3, 8, 34]. These estimates are probably not optimal (see the sharpness results in [32, 26] in the scalar convex case and [31] in the nonconvex case). However, the $h^{1/4}$ results are proven for less regular discontinuous solutions! We stress that our result holds in particular in the scalar case $m = 1$. It gives then a new improved estimate.

As an example for a (strongly coupled) system for which all assumptions on the discrete solution can be verified we consider the one-dimensional system of elastodynamics in section 5.2.

For the final application we return to the general Cauchy problem (1.1), (1.2). Tadmor has introduced in one space dimension a class of semidiscrete second-order entropy conservative schemes [29], (see also [30]). Fully discrete and high-order entropy conservative schemes were established later in [21, 22]. Here we generalize this concept to the multidimensional case on unstructured meshes. To our knowledge, entropy conservative schemes in multiple space dimensions have not been introduced before. Adding artificial viscosity terms, we obtain a new class of finite volume schemes such that the numerical approximation satisfies a cell entropy inequality and an entropy dissipation bound. If the numerical approximation is uniformly bounded we can apply Theorem 4.4 without further restrictions on the system (1.1).

We conclude the introduction with a number of remarks.

The use of Dafermos' approach to derive estimates for approximations of classical solutions goes back to [1]. Another recent application in nonlinear electrodynamics can be found in [2]. Let us note that in one space dimension a result similar to (1.7) has been communicated by Vila [35]. Convergence rates for numerical methods in one space dimension for smooth parts of a weak solution (which might contain singularities!) were established in [28].

In the general case, (1.1), (1.2) has no global smooth solution but there are many physically relevant situations where a classical solution pertains for all times, among these of course problems for linear systems. In the context of hydromechanics we mention the results for special initial data obtained in [10, 27]. If one takes into account a source term B, much more interesting scenarios exist where dissipative effects lead to smooth solutions close to equilibrium states. We mention just a few, like relaxation effects, damping effects through radiation or through gravitational forces, frictional damping, or memory effects [11, 16, 15, 36, 37]. Surveys on general systems with classical solutions and the rôle of the so-called null condition are [17, 33].

Notation. By $C, C_0, C_1, \dots$ we shall denote nonnegative constants depending on a given classical solution but not on the mesh parameter h.

2. Preliminaries and an L^2-stability theorem.

Throughout this section we consider the general problem (1.1), (1.2) and suppose that Assumption 1.1 holds.

We mention that the existence of an entropy-entropy flux tuple implies that the symmetry relation

$$(\nabla^2\eta(w)DG_i(w))^T = \nabla^2\eta(w)DG_i(w) \tag{2.1}$$

holds for $i = 1, \dots, d$ and $w \in \mathcal{U}$.

Without loss of generality, we can assume that we have for $\bar{u}$ with $B(\bar{u}) = 0$ the relations

$$\eta(\bar{u}) = 0, \;\; \nabla\eta(\bar{u}) = 0, \tag{2.2}$$

$$q_i(\bar{u}) = 0, \;\; \nabla q_i(\bar{u}) = 0 \;\; (i = 1, \dots, d). \tag{2.3}$$

We define

$$a := \min_{v \in S} \|\nabla^2\eta(v)\|, \quad b := \max_{v \in S} \|\nabla^2\eta(v)\|. \tag{2.4}$$

Note that $a, b > 0$. Thanks to (2.2), (2.3) we see that we have for all $v \in S$ the inequalities

$$\begin{gathered} a|v-\bar{u}|^2 \le \eta(v) \le b|v-\bar{u}|^2, \;\; |\nabla\eta(v)| \le c|v-\bar{u}|, \\ |q_i(v)| \le c|v-\bar{u}|^2, \;\; |\nabla q_i(v)| \le c|v-\bar{u}| \;\; (i = 1, \dots, d), \end{gathered} \tag{2.5}$$

where the constant c depends only on the set S.

Our aim now is to derive an L^2-stability result similar to but more general than Theorem 1.2. For that purpose, we introduce Dafermos' relative entropy h and entropy fluxes f_i, $i = 1, \dots, d$, by

$$h(v, w) = \eta(w) - \eta(v) - \nabla\eta(v) \cdot (w - v),$$

$$f_i(v, w) = q_i(w) - q_i(v) - \nabla\eta(v) \cdot [G_i(w) - G_i(v)].$$

Note that the function h has locally quadratic growth in $w - v$. This is also true for f_i thanks to (1.3). We conclude that

$$a|v-w|^2 \le h(v, w) \le b|v-w|^2 \qquad (v, w \in S) \tag{2.6}$$

holds and that there is a constant $\lambda = \lambda(S) > 0$, such that

$$\left\{\sum_{i=1}^{d} [f_i(v, w)]^2\right\}^{1/2} \le \lambda h(v, w) \qquad (v, w \in S). \tag{2.7}$$

Similarly, there exists a constant $\alpha = \alpha(S, \|Du\|_{L^\infty}) \geq 0$, such that we have for all $v \in S$

$$\begin{aligned}
&\sum_{i=1}^{d} \big|\nabla^2\eta(u)\partial_i u \cdot \big[G_i(v) - G_i(u) - DG_i(u)(v-u)\big]\big| \leq \frac{\alpha}{3} h(u,v),\\
&\big|B(u) \cdot \big[\nabla\eta(v) - \nabla\eta(u) - \nabla^2\eta(u)(v-u)\big]\big| \leq \frac{\alpha}{3} h(u,v),\\
&\big|[B(u) - B(v)] \cdot [\nabla\eta(u) - \nabla\eta(v)]\big| \leq \frac{\alpha}{3} h(u,v)
\end{aligned} \tag{2.8}$$

on $\mathbb{R}^d \times [0,T]$ for a solution u of (1.1) that satisfies Assumption 1.1. For $R, \lambda > 0$ define the cone

$$\mathcal{C}_R = \{(x,t) \,|\, |x| \leq R + \lambda(T-t),\ x \in \mathbb{R}^d,\ t \in [0,T]\}.$$

We introduce the following measures.

Definition 2.1. *Let $v \in [L^\infty_{loc}(\mathbb{R}^d \times (0,T))]^m$ be an arbitrary function with values in $\mathcal{U}$. The* weak consistency measure *$\mu_v : [C_0^{0,1}(\mathbb{R}^d \times [0,T])]^m \to \mathbb{R}$ and the* dissipation measure *$\nu_v : C_0^{0,1}(\mathbb{R}^d \times [0,T]) \to \mathbb{R}$ are defined by*

$$\langle \mu_v, \pi \rangle = -\int_0^T \int_{\mathbb{R}^d} v \cdot \partial_t \pi + \sum_{i=1}^{d} G_i(v) \cdot \partial_i \pi + B(v) \cdot \pi \; dxdt - \int_{\mathbb{R}^d} u_0 \cdot \pi(.,0)\, dx,$$

$$\begin{aligned}
\langle \nu_v, \omega \rangle = -\int_0^T \int_{\mathbb{R}^d} &\eta(v)\, \partial_t \omega + \sum_{i=1}^{d} q_i(v)\, \partial_i \omega\\
&+ \nabla\eta(v) \cdot B(v)\, \omega \; dxdt - \int_{\mathbb{R}^d} \eta(u_0)\, \omega(.,0)\, dx.
\end{aligned}$$

These measures are important since one can estimate the L^2-distance between the solution u and an *arbitrary function* v in terms of the measures.

Theorem 2.2. *Suppose that Assumption* 1.1 *for the solution u of* (1.1), (1.2) *holds. Let $v \in [L^\infty_{loc}(\mathbb{R}^d \times (0,T))]^m$ be a function with values in the set S (from Assumption* 1.1*). Then, for α given by* (2.8)*, the following estimate holds:*

$$\int_0^T \int_{\mathbb{R}^d} e^{-\alpha t} \xi_R\, h(u,v)\, dxdt \leq \langle \nu_v, \varphi_R \rangle - \langle \mu_v, \psi_R \rangle. \tag{2.9}$$

Here we have $\varphi_R(x,t) = e^{-\alpha t}(T-t)\xi_R(x,t)$, $\psi_R = \varphi_R \nabla\eta(u)$, and

$$\xi_R(x,t) = \gamma(|x| - R - \lambda(T-t)), \quad \gamma(s) = \begin{cases} 1 & : s \leq 0,\\ 1-s & : s \in (0,1),\\ 0 & : s \geq 1. \end{cases}$$

$R > 0$ is an arbitrary number and the parameter λ is given by (2.7).

Proof. Let $\omega \in C_0^{0,1}(\mathbb{R}^d \times [0,T])$ be any function with $\omega(.,T) = 0$ and let $\pi = \omega\nabla\eta(u)$. Note that the classical solution u satisfies (1.5) with equality. Partial integration and application of (1.1) yield

$$\begin{aligned}
&-\int_0^T\int_{\mathbb{R}^d} h(u,v)\,\partial_t\omega+\sum_{i=1}^d f_i(u,v)\,\partial_i\omega\ \ dxdt\\
&=\ \langle\nu_v,\omega\rangle+\int_0^T\int_{\mathbb{R}^d}\omega\,B(v)\cdot\nabla\eta(v)-\omega\,B(u)\cdot\nabla\eta(u)\,dxdt\\
&\qquad+\int_0^T\int_{\mathbb{R}^d}\left[\partial_t\omega\,\nabla\eta(u)\cdot(v-u)+\sum_{i=1}^d\partial_i\omega\,\nabla\eta(u)\cdot[G_i(v)-G_i(u)]\right]dxdt\\
&=\langle\nu_v,\omega\rangle+\int_0^T\int_{\mathbb{R}^d}\omega\,B(v)\cdot\nabla\eta(v)-\omega\,B(u)\cdot\nabla\eta(u)\,dxdt\\
&\qquad+\int_0^T\int_{\mathbb{R}^d}\big[\partial_t[\omega\,\nabla\eta(u)]-\omega\,\nabla^2\eta(u)\partial_t u\big]\cdot\big[v-u\big]\,dxdt\\
&\qquad+\int_0^T\int_{\mathbb{R}^d}\sum_{i=1}^d\big[\partial_i[\omega\,\nabla\eta(u)]-\omega\,\nabla^2\eta(u)\partial_i u\big]\cdot\big[G_i(v)-G_i(u)\big]\,dxdt\\
&=\langle\nu_v,\omega\rangle-\langle\mu_v,\pi\rangle+\int_0^T\int_{\mathbb{R}^d}\omega\,B(v)\cdot\nabla\eta(v)-B(v)\cdot\pi-\omega\,\nabla^2\eta(u)B(u)\cdot(v-u)\,dxdt\\
&\qquad+\int_0^T\int_{\mathbb{R}^d}\sum_{i=1}^d\omega\big[\nabla^2\eta(u)DG_i(u)\partial_i u\cdot(v-u)-\nabla^2\eta(u)\partial_i u\cdot[G_i(v)-G_i(u)]\big]\,dxdt.
\end{aligned}$$

The symmetry of the operators $\nabla^2\eta(u)$ and $\nabla^2\eta(u)DG_i(u)$ (see (2.1)) implies

$$\begin{aligned}
-\int_0^T\int_{\mathbb{R}^d} &h(u,v)\,\partial_t\omega+\sum_{i=1}^d f_i(u,v)\,\partial_i\omega\ dxdt\\
=\quad&\langle\nu_v,\omega\rangle-\langle\mu_v,\pi\rangle+\int_0^T\int_{\mathbb{R}^d}\omega\,B(u)\cdot\big[\nabla\eta(v)-\nabla\eta(u)-\nabla^2\eta(u)(v-u)\big]\,dxdt\\
&+\int_0^T\int_{\mathbb{R}^d}\omega\,[B(u)-B(v)]\cdot[\nabla\eta(u)-\nabla\eta(v)]\,dxdt\\
&-\int_0^T\int_{\mathbb{R}^d}\sum_{i=1}^d\omega\,\nabla^2\eta(u)\partial_i u\cdot\big[G_i(v)-G_i(u)-DG_i(u)(v-u)\big]\,dxdt.
\end{aligned}$$

Using the definition (2.8) of the constant α, one obtains

(2.10)
$$-\int_0^T\int_{\mathbb{R}^d} h(u,v)\,\partial_t\omega+\sum_{i=1}^d f_i(u,v)\,\partial_i\omega\ dxdt\le\langle\nu_v,\omega\rangle-\langle\mu_v,\pi\rangle+\alpha\int_0^T\int_{\mathbb{R}^d}\omega\,h(u,v)\,dxdt.$$

Since

$$\begin{aligned}
&\partial_t\varphi_R(x,t)=-\alpha\varphi_R(x,t)-e^{-\alpha t}\xi_R(x,t)+\lambda e^{-\alpha t}(T-t)\gamma'\big(|x|-R-\lambda(T-t)\big),\\
&\partial_i\varphi_R(x,t)=\gamma'\big(|x|-R-\lambda(T-t)\big)\frac{x_i}{|x|}e^{-\alpha t}(T-t),
\end{aligned}$$

for $i = 1, \dots, d$, we have

$$\begin{aligned}
&-\int_0^T \int_{\mathbb{R}^d} h(u,v)\,\partial_t\varphi_R + \sum_{i=1}^d f_i(u,v)\,\partial_i\varphi_R \;dxdt\\
&= \alpha\int_0^T \int_{\mathbb{R}^d} h(u,v)\,\varphi_R\,dxdt + \int_0^T \int_{\mathbb{R}^d} e^{-\alpha t}\xi_R\,h(u,v)\,dxdt\\
&\quad + \int_0^T \int_{\mathbb{R}^d} e^{-\alpha t}(T-t)\Big[-\gamma'(|x| - R - \lambda(T-t))\Big]\\
&\quad \cdot\Big[\sum_{i=1}^d f_i(u,v)\frac{x_i}{|x|} + \lambda h(u,v)\Big]\,dxdt\\
&\geq \alpha\int_0^T \int_{\mathbb{R}^d} h(u,v)\,\varphi_R\,dxdt + \int_0^T \int_{\mathbb{R}^d} e^{-\alpha t}\xi_R\,h(u,v)\,dxdt.
\end{aligned}$$

The last estimate follows from $\gamma' \leq 0$ almost everywhere (a.e.) on $\mathbb{R}$ and

$$\sum_{i=1}^d f_i(u,v)\frac{x_i}{|x|} + \lambda h(u,v) \geq 0 \quad \text{for } (x,t) \in \mathbb{R}^d \times [0,T],\ x \neq 0,$$

due to (2.7). Finally, if we plug $\omega = \varphi_R$ in (2.10) we get (2.9). □

Theorem 2.2 gives an estimate in terms of Dafermos' relative entropy. Due to the properties of the entropy and our assumptions we can pass to the more convenient L^2-norm.

Corollary 2.3. *Under the conditions stated in Theorem* 2.2 *we have*

$$a\int_0^T \int_{\{|x|\leq R\}} e^{-\alpha t}|u-v|^2 dxdt \leq \langle \nu_v, \varphi_R\rangle - \langle \mu_v, \psi_R\rangle.$$

Here $a > 0$ is the constant from (2.4).

Proof. The statement is a direct consequence of (2.6), Theorem 2.2, and the fact that $\xi_R = 1$ on $\{x \in \mathbb{R}^d \,|\, |x| \leq R\} \times [0,T]$. □

Remark 2.4.

(i) To recover the statement of Theorem 1.2 from Theorem 2.2 let v be an entropy solution as in Theorem 1.2. Then we compute from (2.9)

$$\begin{aligned}
a\int_0^T \int_{\mathbb{R}^d} e^{-\alpha t}\xi_R|u-v|^2\,dxdt &\leq T\int_{\mathbb{R}^d} h(v_0,u_0)\xi_R(x,0)\,dx\\
&\leq b\,T\int_{\{|x|\leq R+\lambda T\}} |u_0 - v_0|^2\,dx.
\end{aligned}$$

The pointwise estimate in time as in Theorem 1.2 can be obtained by a slightly different choice of test functions in Theorem 2.2.

(ii) Due to $B(\bar{u}) = 0$ the function $u = \bar{u}$ is a classical solution of (1.1) and initial condition $u_0 = \bar{u}$. If we choose $v \equiv u$, $v_0 \equiv u_0$ in (i) and let $R \to \infty$, we obtain

$$\int_0^T \int_{\mathbb{R}^d} |u-\bar{u}|^2\,dxdt \leq d_1\int_{\mathbb{R}^d} |u_0 - \bar{u}|^2\,dx \tag{2.11}$$

with a constant d_1 depending on the set S, since $\xi_R \to 1$ and $\alpha = 0$.

(iii) From (1.1), Assumption 1.1, and (ii) we conclude that there is a constant $d_2 > 0$ such that

$$\int_0^T \int_{\mathbb{R}^d} |\partial_t u|^2 \, dxdt \leq d_2. \tag{2.12}$$

3. Discretization. We introduce a class of finite volume schemes to solve (1.1), (1.2) numerically. Beside that we make some assumptions on the corresponding numerical approximation.

3.1. The unstructured mesh. We consider a quasiuniform triangulation $\mathcal{T}_h$ of $\mathbb{R}^d$: this is a set of convex polyhedra, with the property

$$\operatorname{diam}(K) \leq h, \;\; |K| \geq c_0 h^d \quad (K \in \mathcal{T}_h), \tag{3.1}$$

and

$$|e| \geq c_0 h^{d-1} \;\; (e \in \mathcal{E}(K)), \tag{3.2}$$

for some constants $h, \; c_0 > 0$, where $|K|$ is the Lebesgue measure of K on $\mathbb{R}^d$ and $|e|$ the $\mathbb{R}^{d-1}$-Lebesgue measure of the edge e. For a given $K \in \mathcal{T}_h$ the set $\mathcal{E}(K)$ contains all edges of K. Note that for simplices (triangles, tetrahedra) (3.2) is a consequence of (3.1).

We assume that for each $K \in \mathcal{T}_h$ and each edge e of K, there exists exactly one neighboring cell $K_e \in \mathcal{T}_h$. Let furthermore $F = \bigcup_{K \in \mathcal{T}_h} \mathcal{E}(K)$ be the set of all edges.

The mesh with respect to t is uniform: $t^n = n\Delta t$ $(n \in \mathbb{N} \cup \{0\})$. Here $\Delta t > 0$ is such that there is a $N \in \mathbb{N}$ with $N\Delta t = T$. With this number we define the set $\mathcal{N} = \{0, 1, \ldots, N-1\}$.

We will impose on Δt and h the following condition: there exists a constant $c_1 > 0$ independent on $\Delta t, \; h$, such that

$$\frac{\Delta t}{h} \geq c_1. \tag{3.3}$$

Denote by $h_{K,e}$ the height of K that corresponds to $e \in \mathcal{E}(K)$. Then, we have $|e|\,|h_{K,e}| = c(d)|K|$ for a constant $c(d)$ that depends only on the dimension d. For later use we note that the mesh condition (3.3) yields

$$\Delta t|e| = \frac{(\Delta t|e|)^2}{|K|} \frac{|K| h_{K,e}}{\Delta t|e| h_{K,e}} \leq \frac{1}{c(d)} \frac{h}{\Delta t} \frac{(\Delta t|e|)^2}{|K|} \leq \frac{1}{c(d)c_1} \frac{(\Delta t|e|)^2}{|K|} \tag{3.4}$$

and

$$\frac{(\Delta t|e|)^2}{|K|} = \frac{c^2(d)(\Delta t)^2|K|}{h_{K,e}^2} \geq c^2(d) \frac{\Delta t|K|}{h} \frac{\Delta t}{h} \geq c^2(d)c_1 \frac{\Delta t|K|}{h}. \tag{3.5}$$

By Bramble–Hilbert-like techniques one can deduce that for $K \in \mathcal{T}_h, \; e \in \mathcal{E}(K)$, and $z \in [C^1(\overline{K})]^m$

$$\int_K \left| z - \frac{1}{|K|} \int_K z \, dx \right|^2 dx \leq C_0 h^2 \int_K |Dz|^2 dx, \tag{3.6}$$

$$\left| \frac{1}{|K|} \int_K z \, dx - \frac{1}{|e|} \int_e z \, d\sigma \right| \leq C_0 \frac{h^{1/2}}{|e|^{1/2}} \Big(\int_K |Dz|^2 dx \Big)^{1/2} \tag{3.7}$$

holds for some constant $C_0 = C_0(d, c_0)$, under the mesh conditions (3.1), (3.2).

3.2. The finite volume method. To solve the Cauchy problem (1.1), (1.2) numerically we use the finite volume scheme

$$\begin{aligned} u_K^{n+1} &= u_K^n - \frac{\Delta t}{|K|} \sum_{e \in \mathcal{E}(K)} |e| \, g_{K,e}^n(u_K^n, u_{K_e}^n) + \Delta t \, B(u_K^n), \\ u_K^0 &= \frac{1}{|K|} \int_K u_0(x) \, dx. \end{aligned} \tag{3.8}$$

From the iteratives u_K^n we define the piecewise constant approximation $u_h : \mathbb{R}^d \times [0,T] \to \mathbb{R}^m$ of u by

$$u_h(x,t) = u_K^n \quad \text{for } \; x \in K, \; t \in [t^n, t^{n+1}), \tag{3.9}$$

where $K \in \mathcal{T}_h$ and $n \in \mathcal{N}$.

For the numerical flux $g_{K,e}^n$ in (3.8) we suppose the usual consistency and conservation properties.

- For all $n \in \mathcal{N}, \; K \in \mathcal{T}_h, \; e \in \mathcal{E}(K)$ we have

$$g_{K,e}^n(v,v) = \sum_{i=1}^d n_{K,e}^i G_i(v) \quad (v \in \mathcal{U}), \tag{3.10}$$

where $n_{K,e} = (n_{K,e}^1, \dots, n_{K,e}^d)^T$ is the unit outward normal to $e \in \mathcal{E}(K)$.

- For all $n \in \mathcal{N}, \; K \in \mathcal{T}_h, \; e \in \mathcal{E}(K)$, it holds

$$g_{K,e}^n(v,w) = -g_{K_e,e}^n(w,v) \quad (v,w \in \mathcal{U}). \tag{3.11}$$

A special consequence of (3.10) is

$$\sum_{e \in \mathcal{E}(K)} |e| \, g_{K,e}^n(u_K^n, u_K^n) = 0 \quad (n \in \mathcal{N}, \; K \in \mathcal{T}_h). \tag{3.12}$$

We introduce now the numerical entropy flux $q_{K,e}^n$, having the following properties.

- For all $M > 0$ there exists a constant $c(M) > 0$, such that for all $n \in \mathcal{N}$, $K \in \mathcal{T}_h, \; e \in \mathcal{E}(K)$,

$$|q_{K,e}^n(v,w)| \leq c(M)\big(|v - \bar{u}|^2 + |w - \bar{u}|^2\big) \quad (|v - \bar{u}|, |w - \bar{u}| \leq M). \tag{3.13}$$

- For all $n \in \mathcal{N}, \; K \in \mathcal{T}_h, \; e \in \mathcal{E}(K)$, we have

$$q_{K,e}^n(v,v) = \sum_{i=1}^d n_{K,e}^i q_i(v) \quad (v \in \mathcal{U}). \tag{3.14}$$

- For all $n \in \mathcal{N}, \; K \in \mathcal{T}_h, \; e \in \mathcal{E}(K)$, it is

$$q_{K,e}^n(v,w) = -q_{K_e,e}^n(w,v) \quad (v,w \in \mathcal{U}). \tag{3.15}$$

The conditions on the numerical flux functions can be satisfied by standard choices. In particular this is true for the Lax–Friedrichs flux (cf. [18]).

Before we proceed with the convergence proof we make the following strong assumption on u_h.

ASSUMPTION 3.1. *Let u_h be the finite volume approximation defined by* (3.8), (3.9). *Suppose that*

(H.1) u_h *takes values in the set* S *(from Assumption* 1.1*).*

(H.2) *For each* $n \in \mathcal{N}$, $K \in \mathcal{T}_h$, *the following relation holds:*

$$\begin{aligned}\eta(u_K^{n+1}) - \eta(u_K^n) &+ \frac{\Delta t}{|K|} \sum_{e \in \mathcal{E}(K)} |e|\, q_{K,e}^n(u_K^n, u_{K_e}^n) \\ &+ \left(\frac{\Delta t}{|K|}\right)^2 \sum_{e \in \mathcal{E}(K)} |e|^2 \big|g_{K,e}^n(u_K^n, u_{K_e}^n) - g_{K,e}^n(u_K^n, u_K^n)\big|^2 \\ &\qquad \le \Delta t \nabla \eta(u_K^n) \cdot B(u_K^n) + C(\Delta t)^2 |u_K^n - \bar{u}|^2.\end{aligned}$$

The assumption (H.2) expresses that we are supposed to have a cell entropy inequality for the approximate solution u_h. Moreover, the scheme should provide some entropy dissipation to get the weak derivative bound on the fluxes. Together with the L^∞-bound in (H.1) this will restrict the choice of possible flux functions and moreover introduce a CFL-like restriction on the time step. In section 5 we shall consider examples where (H.1) and (H.2) are met.

We conclude this section with several important a priori estimates for u_h, which are essential for our further considerations and are induced by Assumption 3.1. We present them in the next proposition and in the corollary following it. They can be proven in the linear case [13, 34]. Therefore we omit the proofs.

Proposition 3.2. *Let* u_h *be given by* (3.8), (3.9) *and suppose that it obeys Assumption* 3.1. *For the numerical flux* $g_{K,e}^n$ *and the numerical entropy flux* $q_{K,e}^n$ *assume* (3.10), (3.11) *and* (3.13), (3.15), *respectively. Under the mesh condition* (3.1), *we have*

$$\text{(a)} \quad \sum_{K \in \mathcal{T}_h} \sum_{e \in \mathcal{E}(K)} |e|\, q_{K,e}^n(u_K^n, u_{K_e}^n) = 0 \quad (n \in \mathcal{N}),$$

$$\text{(b)} \quad \int_0^T \int_{\mathbb{R}^d} |u_h - \bar{u}|^2 \, dx dt \le C,$$

$$\text{(c)} \quad \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}} \sum_{e \in \mathcal{E}(K)} \frac{(\Delta t |e|)^2}{|K|} \big|g_{K,e}^n(u_K^n, u_{K_e}^n) - g_{K,e}^n(u_K^n, u_K^n)\big|^2 \le C.$$

The a priori estimate (c) in Proposition 3.2 is not optimal with respect to h (and therefore doesn't lead to the optimal error estimate scaling with h instead of $\sqrt{h}$ later). It can be improved a posteriori. For that purpose we introduce the term

$$Q_h = \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}} \sum_{e \in \mathcal{E}(K)} \theta^n \frac{(\Delta t\, |e|)^2}{|K|} \big|g_{K,e}^n(u_K^n, u_{K_e}^n) - g_{K,e}^n(u_K^n, u_K^n)\big|^2 \tag{3.16}$$

with

$$\theta^n = \frac{1}{\Delta t} \int_{t^n}^{t^{n+1}} \theta(t)\, dt$$

and $\theta(t) = e^{-\alpha t}(T - t)$.

Corollary 3.3. *Under the assumptions in the previous proposition, we have*

$$\int_0^T \int_{\mathbb{R}^d} |u - u_h|^2 \, dxdt \leq C, \tag{3.17}$$

$$\sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} |K| \, |u_K^{n+1} - u_K^n|^2 \leq C, \tag{3.18}$$

$$\sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \theta^n |K| \, |u_K^{n+1} - u_K^n|^2 \leq CQ_h + C\Delta t. \tag{3.19}$$

4. The error estimate for classical solutions.

By the L^2-stability result from section 2 we can now tackle the convergence proof. Throughout this section we consider the general problem (1.1), (1.2) and suppose that Assumption 1.1 holds. Moreover we take Assumption 3.1 for the numerical approximation u_h to be valid.

4.1. Identification of the error measure and estimates.

We are now in a position to replace the function v in Corollary 2.3 by u_h. Furthermore, it turns out that one can pass to limits for $R \to \infty$. This is the subject of the next proposition. In addition, since $\varphi_R \to \theta$, the limit function becomes independent of x. This is essential for the derivation of error estimates.

Proposition 4.1. *Suppose that u is the solution of* (1.1), (1.2) *which satisfies Assumption* 1.1. *For the numerical fluxes $g_{K,e}^n$ and $q_{K,e}^n$ we assume the properties* (3.10), (3.11) *and* (3.13), (3.15), *respectively. Assume further that for u_h defined by* (3.9) *Assumption* 3.1 *holds. Then, under the mesh condition* (3.1), *it holds that*

$$a \int_0^T \int_{\mathbb{R}^d} e^{-\alpha t} |u - u_h|^2 dxdt \leq \langle \nu_h, \theta \rangle - \langle \mu_h, \psi \rangle,$$

for α given by (2.8), $\theta(t) = e^{-\alpha t}(T - t)$, $\psi = \theta \nabla \eta(u)$, *and*

$$\begin{aligned}\langle \mu_h, \psi \rangle = &- \int_0^T \int_{\mathbb{R}^d} (u_h - \bar{u}) \cdot \partial_t \psi + \sum_{i=1}^d [G_i(u_h) - G_i(\bar{u})] \cdot \partial_i \psi + B(u_h) \cdot \psi \; dxdt \\ &- \int_{\mathbb{R}^d} (u_0 - \bar{u}) \cdot \psi(.,0) \, dx, \\ \langle \nu_h, \theta \rangle = &- \int_0^T \int_{\mathbb{R}^d} \eta(u_h) \, \theta' + \nabla \eta(u_h) \cdot B(u_h) \, \theta \; dxdt - \int_{\mathbb{R}^d} \eta(u_0) \, \theta(0) \, dx.\end{aligned}$$

Proof. From Definition 2.1 we deduce

$$\begin{aligned}\langle \mu_{u_h}, \psi_R \rangle = &- \int_0^T \int_{\mathbb{R}^d} (u_h - \bar{u}) \cdot \partial_t \psi_R + \sum_{i=1}^d [G_i(u_h) - G_i(\bar{u})] \cdot \partial_i \psi_R + B(u_h) \cdot \psi_R \; dxdt \\ &- \int_{\mathbb{R}^d} (u_0 - \bar{u}) \cdot \psi_R(.,0) \, dx\end{aligned}$$

because ψ_R has a compact support. According to (b) in Proposition 3.2, we have $u_h - \bar{u} \in [L^2(\mathbb{R}^d \times (0,T))]^m$. Due to (2.11), (2.12), and (2.5) one infers ψ, $\partial_t \psi$, $\partial_i \psi$,

$B(u_h)$, $\nabla\eta(u_h) \in [L^2(\mathbb{R}^d \times (0,T))]^m$, and $\eta(u_h), q_i(u_h) \in [L_1(\mathbb{R}^d \times (0,T))]^m$. On the other side $\xi_R \to 1$, $\partial_t \xi_R \to 0$, $\partial_i \xi_R \to 0$ a.e. on $\mathbb{R}^d \times [0,T]$, $i = 1,\dots,d$, for $R \to \infty$, and all these terms remain bounded. Therefore one can pass to limits in Corollary 2.3. ▯

To obtain error estimates, it remains to estimate the expression $\langle \nu_h, \theta\rangle - \langle \mu_h, \psi\rangle$. This is done in the next two lemmas. We need the following inequality:

$$\max_{[t^n,t^{n+1}]} \theta(t) \le 2e^{\alpha\Delta t}\theta^n \quad (n \in \mathcal{N}). \tag{4.1}$$

Lemma 4.2. *Under the assumptions in Proposition* 4.1 *we have*

$$\langle \nu_h, \theta\rangle - \langle \mu_h, \psi\rangle \le L + R + \int_0^T \int_{\mathbb{R}^d} \theta(t) B(u_h) \cdot [\nabla\eta(u) - \nabla\eta(u_h)]\, dxdt + Ch^2,$$

where

$$L = \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \int_K \theta(t^{n+1}) \big[\eta(u_K^{n+1}) - \eta(u_K^n) - \nabla\eta(u(x,t^{n+1}))\cdot(u_K^{n+1} - u_K^n)\big]\, dx,$$

$$R = \frac{1}{2} \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}} \sum_{e\in\mathcal{E}(K)} \left[\sum_{i=1}^d n_{K,e}^i \big(G_i(u_K^n) - G_i(u_{K_e}^n)\big)\right] \cdot \int_{t^n}^{t^{n+1}} \int_e \psi\, d\sigma dt.$$

Proof. For the sake of simplicity assume that $\bar{u} = 0$ and $G_i(\bar{u}) = 0$ for $i = 1,\dots,d$. Partial integration with respect to t yields

$$\begin{aligned} &\int_0^T \int_{\mathbb{R}^d} u_h \cdot \psi_t\, dxdt + \int_{\mathbb{R}^d} u_0(x)\cdot\psi(x,0)\, dx \\ &= \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \int_K \psi(x,t^{n+1})\cdot(u_K^n - u_K^{n+1})\, dx + \sum_{K\in\mathcal{T}_h} \int_K \psi(x,0)\cdot(u_0(x) - u_K^0)\, dx \end{aligned}$$

and

$$\begin{aligned} &-\int_0^T \int_{\mathbb{R}^d} \eta(u_h)\theta'(t)\, dxdt - \int_{\mathbb{R}^d} \eta(u_0(x))\theta(0)\, dx \\ &= \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \int_K \theta(t^{n+1}) \big[\eta(u_K^{n+1}) - \eta(u_K^n)\big]\, dx + \sum_{K\in\mathcal{T}_h} \int_K \varphi(x,0) \big[\eta(u_K^0) - \eta(u_0(x))\big]\, dx. \end{aligned}$$

After applying the Gauss theorem, one obtains

$$\sum_{i=1}^d \int_{t^n}^{t^{n+1}} \int_K G_i(u_K^n)\cdot\partial_i\psi\, dxdt = \sum_{e\in\mathcal{E}(K)} \left[\sum_{i=1}^d n_{K,e}^i G_i(u_K^n)\right] \cdot \int_{t^n}^{t^{n+1}} \int_e \psi\, d\sigma dt.$$

Therefore, taking into account that

$$n_{K,e}^i\big(G_i(u_K^n) - G_i(u_{K_e}^n)\big) = \frac{1}{2}\big[n_{K,e}^i\big(G_i(u_K^n) - G_i(u_{K_e}^n)\big) + n_{K_e,e}^i\big(G_i(u_{K_e}^n) - G_i(u_K^n)\big)\big],$$

we have

$$\begin{aligned}
&\int_0^T \int_{\mathbb{R}^d} \sum_{i=1}^d G_i(u_h)\cdot \partial_i\psi \, dxdt\\
&\qquad = \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}} \sum_{e\in\mathcal{E}(K)} \left[\sum_{i=1}^d n_{K,e}^i G_i(u_K^n)\right]\cdot \int_{t^n}^{t^{n+1}} \int_e \psi \, d\sigma dt\\
&\qquad = \sum_{n\in\mathcal{N}} \sum_{e\in F} \left[\sum_{i=1}^d n_{K,e}^i \big(G_i(u_K^n) - G_i(u_{K_e}^n)\big)\right]\cdot \int_{t^n}^{t^{n+1}} \int_e \psi \, d\sigma dt\\
&\qquad = \frac{1}{2}\sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}} \sum_{e\in\mathcal{E}(K)} \left[\sum_{i=1}^d n_{K,e}^i \big(G_i(u_K^n) - G_i(u_{K_e}^n)\big)\right]\cdot \int_{t^n}^{t^{n+1}} \int_e \psi \, d\sigma dt.
\end{aligned}$$

From the equation

$$\begin{aligned}
\langle \nu_h, \varphi\rangle - \langle \mu_h, \psi\rangle = \int_0^T \int_{\mathbb{R}^d} u_h\cdot\psi_t \ + \sum_{i=1}^d G_i(u_h)\cdot\partial_i\psi \ \ dxdt + \int_{\mathbb{R}^d} u_0(x)\cdot\psi(x,0)\, dx\\
+ \int_0^T \int_{\mathbb{R}^d} \theta B(u_h)\cdot[\nabla\eta(u) - \nabla\eta(u_h)] - \eta(u_h)\theta' \, dxdt - \int_{\mathbb{R}^d} \eta(u_0(x))\theta(0)\, dx,
\end{aligned}$$

the relations derived above, and the fact that $\psi = \theta\nabla\eta(u)$, one concludes that

$$\langle \nu_h, \varphi\rangle - \langle \mu_h, \psi\rangle = L + R + J + \int_0^T \int_{\mathbb{R}^d} \theta(t) B(u_h)\cdot[\nabla\eta(u) - \nabla\eta(u_h)]\, dxdt, \tag{4.2}$$

where

$$J = \sum_{K\in\mathcal{T}_h} \int_K \theta(0)\big[\eta(u_K^0) - \eta(u_0(x)) - \nabla\eta(u_0(x))\cdot(u_K^0 - u_0(x))\big]\, dx.$$

The inequality (3.6) implies $\left\|u_K^0 - u_0\right\|_{L^2(K)} \le Ch\|Du_0\|_{L^2(K)}$, which leads to

$$\begin{aligned}
J &= \sum_{K\in\mathcal{T}_h} \int_K \theta(0) h(u_0(x), u_K^0)\, dx\\
&\le bT \sum_{K\in\mathcal{T}_h} \int_K |u_K^0 - u_0(x)|^2 dx \le Ch^2 \int_{\mathbb{R}^d} |Du_0|^2 dx. \qquad \square
\end{aligned}$$

In the next step we estimate L from Lemma 4.2.

LEMMA 4.3. *Let the assumptions in Proposition* 4.1 *hold. Assume additionally the mesh conditions* (3.2), (3.3). *Then for the term L from Lemma* 4.2, *we have*

$$\begin{aligned}
L \le P - \frac{1}{2} Q_h + \frac{a}{2}\int_0^T \int_{\mathbb{R}^d} e^{-\alpha t} |u - u_h|^2 \, dxdt\\
+ \int_0^T \int_{\mathbb{R}^d} \theta\, B(u_h)\cdot[\nabla\eta(u_h) - \nabla\eta(u)]\, dxdt + C(\Delta t + h),
\end{aligned}$$

where

$$P = \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}} \sum_{e\in\mathcal{E}(K)} \big[g_{K,e}^n(u_K^n, u_{K_e}^n) - g_{K,e}^n(u_K^n, u_K^n)\big]\cdot \int_{t^n}^{t^{n+1}} \int_e \psi \, d\sigma dt.$$

Proof. Assume again that $\bar{u} = 0$. The term L can be rewritten as

$$L = E + I, \tag{4.3}$$

where

$$E = \frac{1}{\Delta t} \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K \theta(t) \big[\eta(u_K^{n+1}) - \eta(u_K^n) - \nabla\eta(u(x,t^{n+1})) \\ \cdot (u_K^{n+1} - u_K^n) \big] \, dxdt,$$

$$I = \frac{1}{\Delta t} \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K \Big[[\theta(t^{n+1}) - \theta(t)] \\ \big[\eta(u_K^{n+1}) - \eta(u_K^n) - \nabla\eta(u(x,t^{n+1})) \cdot (u_K^{n+1} - u_K^n) \big] \Big] \, dxdt.$$

We split up $I = I_1 + I_2 + I_3$ according to

$$I_1 = \frac{1}{\Delta t} \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K \Big[[\theta(t^{n+1}) - \theta(t)] \\ \big[\eta(u_K^{n+1}) - \eta(u_K^n) - \nabla\eta(u_K^n) \cdot \big(u_K^{n+1} - u_K^n\big) \big] \Big] \, dxdt,$$

$$I_2 = \frac{1}{\Delta t} \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K [\theta(t^{n+1}) - \theta(t)] \\ \big[\nabla\eta(u(x,t)) - \nabla\eta(u(x,t^{n+1})) \big] \cdot \big[u_K^{n+1} - u_K^n \big] \, dxdt,$$

$$I_3 = \frac{1}{\Delta t} \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K [\theta(t^{n+1}) - \theta(t)] \\ \big[\nabla\eta(u_K^n) - \nabla\eta(u(x,t)) \big] \cdot \big[u_K^{n+1} - u_K^n \big] \, dxdt.$$

Because of $|\theta(t^{n+1}) - \theta(t)| \le C\Delta t$ for $t \in [t^n, t^{n+1}]$ and

$$|\eta(u_K^{n+1}) - \eta(u_K^n) - \nabla\eta(u_K^n) \cdot (u_K^{n+1} - u_K^n)| \le b|u_K^{n+1} - u_K^n|^2,$$

we obtain by applying the a priori estimate (3.18)

$$I_1 \le C \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}_h} \Delta t \, |K| \, |u_K^{n+1} - u_K^n|^2 \le C\Delta t.$$

For $t \in [t^n, t^{n+1}]$ one obtains

$$\big| \nabla\eta(u(x,t)) - \nabla\eta(u(x,t^{n+1})) \big| = \Big| \int_{t^{n+1}}^{t} \nabla^2\eta(u(x,s)) \partial_t u(x,s) \, ds \Big| \le C \int_{t^n}^{t^{n+1}} |\partial_t u| \, ds.$$

The integral Cauchy–Schwarz inequality implies

$$I_2 \le C\Delta t \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K |\partial_t u| \, |u_K^{n+1} - u_K^n| \, dxdt \\ \le C\Delta t \sum_{n \in \mathcal{N}} \sum_{K \in \mathcal{T}_h} \left(\int_{t^n}^{t^{n+1}} \int_K |\partial_t u|^2 \, dxdt \right)^{1/2} (\Delta t \, |K|)^{1/2} |u_K^{n+1} - u_K^n|.$$

Finally, the Cauchy–Schwarz inequality applied to the last sum together with (2.12), (3.18) gives

$$I_2 \le C\Delta t \left(\int_0^T \int_{\mathbb{R}^d} |\partial_t u|^2 \, dxdt\right)^{1/2} \left(\sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \Delta t \, |K| \, |u_K^{n+1} - u_K^n|^2\right)^{1/2} \le C(\Delta t)^{3/2}.$$

For the term I_3, we have

$$\begin{aligned} I_3 &\le C \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K |u - u_K^n| \, |u_K^{n+1} - u_K^n| \, dxdt \\ &\le C \left(\int_0^T \int_{\mathbb{R}^d} e^{-\alpha t} |u - u_h|^2 \, dxdt\right)^{1/2} \left(\sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \Delta t \, |K| \, |u_K^{n+1} - u_K^n|^2\right)^{1/2} \\ &\le C(\Delta t)^{1/2} \left(\int_0^T \int_{\mathbb{R}^d} e^{-\alpha t} |u - u_h|^2 \, dxdt\right)^{1/2} \\ &\le \frac{a}{2} \int_0^T \int_{\mathbb{R}^d} e^{-\alpha t} |u - u_h|^2 \, dxdt + C\Delta t. \end{aligned}$$

Note that we needed here the fact, that $\int_0^T \int_{\mathbb{R}^d} e^{-\alpha t} |u - u_h|^2 \, dxdt < \infty$ (see (3.17)). Decompose E as $E = E_1 + E_2 + E_3$ with

$$\begin{aligned} E_1 &= \frac{1}{\Delta t} \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K \theta(t) \left[\eta(u_K^{n+1}) - \eta(u_K^n)\right] dxdt, \\ E_2 &= \frac{1}{\Delta t} \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K \theta(t) \left[\nabla\eta(u(x,t)) - \nabla\eta(u(x,t^{n+1}))\right] \cdot \left[u_K^{n+1} - u_K^n\right] dxdt, \\ E_3 &= -\frac{1}{\Delta t} \sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \int_{t^n}^{t^{n+1}} \int_K \theta(t) \nabla\eta(u(x,t)) \cdot (u_K^{n+1} - u_K^n) \, dxdt. \end{aligned}$$

If we apply (H.2) in Assumption 3.1, we infer using (a) in Proposition 3.2 that

$$\begin{aligned} E_1 &\le -Q_h + \int_0^T \int_{\mathbb{R}^d} \theta \nabla\eta(u_h) \cdot B(u_h) \, dxdt + C\Delta t \int_0^T \int_{\mathbb{R}^d} |u_h - \bar{u}|^2 \, dxdt \\ &\le -Q_h + \int_0^T \int_{\mathbb{R}^d} \theta \nabla\eta(u_h) \cdot B(u_h) \, dxdt + C\Delta t \\ &\qquad \text{(here we used (b) in Proposition 3.2).} \end{aligned}$$

Proceeding similarly as for the term I_2, one obtains, with the help of (4.1) and (3.19),

$$\begin{aligned} E_2 &\le C \left(\int_0^T \int_{\mathbb{R}^d} |\partial_t u|^2 \, dxdt\right)^{1/2} \left(\sum_{n\in\mathcal{N}} \sum_{K\in\mathcal{T}_h} \theta^n \, \Delta t \, |K| \, |u_K^{n+1} - u_K^n|^2\right)^{1/2} \\ &\le C(\Delta t)^{1/2} (Q_h)^{1/2} + C\Delta t \\ &\le \frac{Q_h}{4} + C\Delta t. \end{aligned}$$

From (3.8) and (3.12) follows

$$
\begin{aligned}
E_3 = \quad & \sum_{n\in\mathcal{N}}\sum_{K\in\mathcal{T}}\sum_{e\in\mathcal{E}(K)} |e|\left[g^n_{K,e}(u^n_K,u^n_{K_e}) - g^n_{K,e}(u^n_K,u^n_K)\right] \\
& \cdot \int_{t^n}^{t^{n+1}} \theta(t) \int_K \frac{1}{|K|}\nabla\eta(u)\,dxdt \\
& - \int_0^T \int_{\mathbb{R}^d} \theta\nabla\eta(u)\cdot B(u_h)\,dxdt.
\end{aligned}
$$

Define a new term

$$
\begin{aligned}
\overline{E}_3 = & \sum_{n\in\mathcal{N}}\sum_{K\in\mathcal{T}}\sum_{e\in\mathcal{E}(K)} |e|\big[g^n_{K,e}(u^n_K,u^n_{K_e}) - g^n_{K,e}(u^n_K,u^n_K)\big] \\
& \cdot \int_{t^n}^{t^{n+1}} \theta\left\{\frac{1}{|K|}\int_K \nabla\eta(u)\,dx - \frac{1}{|e|}\int_e \nabla\eta(u)\,d\sigma\right\}dt.
\end{aligned}
$$

Then, we have

$$
E_3 = P + \overline{E}_3 - \int_0^T \int_{\mathbb{R}^d} \theta\nabla\eta(u)\cdot B(u_h)\,dxdt. \tag{4.4}
$$

The term $\overline{E}_3$ can be estimated, using (3.7), as follows:

$$
\begin{aligned}
\overline{E}_3 \le\ & Ch^{1/2}\sum_{n\in\mathcal{N}}\sum_{K\in\mathcal{T}}\sum_{e\in\mathcal{E}(K)} |e|^{1/2}\big|g^n_{K,e}(u^n_K,u^n_{K_e}) - g^n_{K,e}(u^n_K,u^n_K)\big| \\
& \int_{t^n}^{t^{n+1}} \theta(t)\left(\int_K |Du|^2\,dx\right)^{1/2} dt \\
\le\ & Ch^{1/2}\sum_{n\in\mathcal{N}}\sum_{K\in\mathcal{T}}\sum_{e\in\mathcal{E}(K)} (\Delta t\,|e|)^{1/2}\big|g^n_{K,e}(u^n_K,u^n_{K_e}) - g^n_{K,e}(u^n_K,u^n_K)\big| \\
& \cdot\left(\max_{[t^n,t^{n+1}]}\theta(t)\right)\left(\int_{t^n}^{t^{n+1}}\int_K |Du|^2\,dxdt\right)^{1/2} \\
\le\ & Ch^{1/2}(Q_h)^{1/2}\left(\int_0^T\int_{\mathbb{R}^d} |Du|^2\,dxdt\right)^{1/2} \quad \text{(here we used (3.4))} \\
\le\ & \frac{Q_h}{4} + Ch.
\end{aligned}
$$

Finally, the conclusion follows from (4.3), the estimates for $I_1, I_2, I_3, E_1, E_2, \overline{E}_3$, and (4.4). □

4.2. ‘. The main theorem The assertions of Proposition 4.1, Lemma 4.2, and Lemma 4.3 are summarized in the error estimate given in the next theorem.

THEOREM 4.4. *Let u be the classical solution of the Cauchy problem* (1.1), (1.2), *satisfying Assumption* 1.1. *Let further the numerical fluxes $g^n_{K,e}$ and $q^n_{K,e}$ have the properties* (3.10), (3.11) *and* (3.13), (3.15), *respectively. Then, for the finite volume approximation u_h defined by* (3.8), (3.9) *that obeys Assumption* 3.1 *and the mesh which satisfies* (3.1), (3.2), (3.3), *the a priori error estimate*

$$
a\int_0^T\int_{\mathbb{R}^d} e^{-\alpha t}|u-u_h|^2\,dxdt + Q_h \le C(\Delta t + h)
$$

holds for a constant C which doesn't depend on the mesh. The constant α is given by (2.8)*, the constant a by* (2.4)*, and the term Q_h by* (3.16)*.*

Proof. From Lemma 4.2 and Lemma 4.3 we get

(4.5)
$$\langle \nu_h, \theta\rangle - \langle \mu_h, \psi\rangle \leq P + R - \frac{1}{2}Q_h + \frac{a}{2}\int_0^T \int_{\mathbb{R}^d} e^{-\alpha t}|u-u_h|^2\,dxdt + C(\Delta t + h).$$

Taking into account the consistency (3.10) and conservation (3.11) properties of the flux $g^n_{K,e}$, one obtains

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{d} n^i_{K,e}\big(G(u^n_K) - G(u^n_{K_e})\big) + g^n_{K,e}(u^n_K, u^n_{K_e}) - g^n_{K,e}(u^n_K, u^n_K)\\
&= \frac{1}{2}\big[g^n_{K,e}(u^n_K, u^n_K) - g^n_{K,e}(u^n_{K_e}, u^n_{K_e})\big] + g^n_{K,e}(u^n_K, u^n_{K_e}) - g^n_{K,e}(u^n_K, u^n_K)\\
&= \frac{1}{2}\big[g^n_{K,e}(u^n_K, u^n_{K_e}) - g^n_{K,e}(u^n_K, u^n_K)\big] - \frac{1}{2}\big[g^n_{K_e,e}(u^n_{K_e}, u^n_K) - g^n_{K_e,e}(u^n_{K_e}, u^n_{K_e})\big].
\end{aligned}$$

After rearranging the sums in P and R, we obtain that $P+R=0$. Now, the conclusion follows from (4.5) and Proposition 4.1. □

As discussed in section 3.2 after Proposition 3.2 one can improve the weak derivative estimate in the cell entropy inequality in (H.2). This is done in the concluding corollary.

COROLLARY 4.5. *Under the assumptions of Theorem* 4.4*, it is*

$$\sum_{t^{n+1}\leq T/2}\ \sum_{K\in\mathcal{T}_h}\ \sum_{e\in\mathcal{E}(K)} \Delta t\,|K| \left|\frac{g^n_{K,e}(u^n_K, u^n_{K_e}) - g^n_{K,e}(u^n_K, u^n_K)}{h}\right|^2 \leq C.$$

Proof. From $Q_h \leq Ch$ and (3.5) we deduce

$$\sum_{n\in\mathcal{N}}\ \sum_{K\in\mathcal{T}}\ \sum_{e\in\mathcal{E}(K)} \theta^n\,\Delta t\,|K| \left|\frac{g^n_{K,e}(u^n_K, u^n_{K_e}) - g^n_{K,e}(u^n_K, u^n_K)}{h}\right|^2 \leq C.$$

Because of

$$\theta^n \geq \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} e^{-\alpha t^{n+1}}T/2\,dt \geq \frac{T}{2}e^{-\alpha T/2},$$

for $t^{n+1}\leq T/2$, we get the statement. □

5. Applications. To apply Theorem 4.4 the Assumptions 1.1, 3.1 have to established. Although the latter assumption for the numerical scheme is reasonable, it is hard to verify it for general systems of type (1.1). In the concluding section we discuss subclasses of (1.1) and corresponding finite-volume discretizations such that both assumptions can be satisfied at least in part.

5.1. Monotone schemes for weakly coupled systems. We shall now apply the theory previously developed in this paper to the class of weakly coupled systems with bounded source terms. The system (1.1) is called weakly coupled if the flux functions $G_i = (G_{i1}(u), \ldots, G_{im}(u))^T$, $i = 1, \ldots, d$, satisfy

(5.1)
$$G_{ik}(u) = G_{ik}(u_k) \quad (k = 1, \ldots, m).$$

The coupling is only due to the source term vector B. For simplicity assume $\mathcal{U} = \mathbb{R}^m$. As before let $\bar{u} \in \mathbb{R}^m$ with $B(\bar{u}) = 0$ and assume that there is a constant $\beta > 0$ such that

$$|\nabla B_k(w)| \le \beta \quad (w \in \mathbb{R}^m,\ k = 1, \dots, m). \tag{5.2}$$

In this case the growth of the solution can be controlled and there is for each $T > 0$ a unique Kruzkov solution $u \in L^\infty(0, T; [L^\infty(\mathbb{R}^d) \cap BV(\mathbb{R}^d)]^m)$ of the Cauchy problem provided that we have $u_0 \in [L^\infty(\mathbb{R}^d) \cap BV(\mathbb{R}^d)]^m$ [24]. Moreover, due to (5.1), entropy tuples can be constructed by

$$\eta(w) = \sum_{k=1}^{m} \eta_k(w_k),\ q_i(w) = \sum_{k=1}^{m} q_{ik}(w_k) \quad (w \in \mathbb{R}^m,\ i = 1, \dots, d) \tag{5.3}$$

such that $\eta_k \in C^3(\mathbb{R})$ are uniformly convex functions and

$$q_{ik} = q_{ik}(w_k) = \int_{\bar{u}}^{w_k} \eta_k'(z) G_{ik}'(z)\, dz \quad (k = 1, \dots, m,\ i = 1, \dots, d).$$

Of course a scalar inhomogeneous conservation law can be written in terms of a weakly coupled system, but moreover there are physically relevant examples of weakly coupled systems with $m > 1$ that admit global-in-time classical solutions for initial data close to a constant state $\bar{u}$ [37].

To discretize the initial value problem for a weakly coupled system we use the finite volume scheme (3.8) such that the numerical flux function $g_{K,e}^n : \mathbb{R}^{2m} \to \mathbb{R}^m$ satisfies (3.10), (3.11) for all $n \in \mathcal{N}$, $K \in \mathcal{T}_h$, $e \in \mathcal{E}(K)$. Moreover we suppose that we have

$$g_{K,e}^n(w, \tilde{w}) = (h_{K,e}^n(w_1, \tilde{w}_1), \dots, h_{K,e}^n(w_m, \tilde{w}_m))^T \quad (w, \tilde{w} \in \mathbb{R}^m), \tag{5.4}$$

where $h_{K,e}^n$ is a scalar locally Lipschitz continuous numerical flux function, i.e., for all $M > 0$, there exists a constant $L_G(M) > 0$, such that we have for all $z_1, z_2, \tilde{z}_1, \tilde{z}_2 \in [-M, M]$

$$|h_{K,e}^n(z_1, z_2) - h_{K,e}^n(\tilde{z}_1, \tilde{z}_2)| \le L_G(M)\big(|z_1 - \tilde{z}_1| + |z_2 - \tilde{z}_2|\big). \tag{5.5}$$

We suppose that $h_{K,e}^n$ is a monotone flux, i.e., we have for all $n \in \mathcal{N}$, $K \in \mathcal{T}_h$, $e \in \mathcal{E}(K)$, $z_1, z_2 \in \mathbb{R}$

$$\frac{\partial}{\partial z_2} h_{K,e}^n(z_1, z_2) \le 0. \tag{5.6}$$

Finally, let there be a numerical entropy flux function $q_{K,e}^n$ for $g_{K,e}^n$ such that

$$\text{(3.13), (3.14), and (3.15) hold.} \tag{5.7}$$

Taking the Lax–Friedrichs flux or the Engquist–Osher flux for $h_{K,e}^n$ provides us with examples of fluxes that can satisfy all above conditions [18].

The property (5.6) is fundamental to ensure the L^∞-stability of a finite volume scheme, and the following statement can be found in any textbook on the subject (e.g., [18]). Let $M > 0$ and $K \in \mathcal{T}_h$. Assume that $z \in \mathbb{R}$ and $z_e \in \mathbb{R}$, $e \in \mathcal{E}(K)$, are from $[\bar{z}-M,\bar{z}+M]$ for $\bar{z}\in\mathbb{R}$. Then there exists a constant c_0 depending only on the mesh such that under the CFL-like time restriction

$$\Delta t \leq \frac{c_0}{L_G(M)}h \tag{5.8}$$

we have

$$z - \frac{\Delta t}{|K|}\sum_{e\in\mathcal{E}(K)} |e| h^n_{K,e}(z,z_e) \in [\bar{z}-M,\bar{z}+M]. \tag{5.9}$$

Monotone finite volume schemes for weakly coupled systems has been analyzed in [14, 25] by variants of Diperna's theory of measure-valued solutions and Kruzkov's L^1-stability estimate for scalar equations. For the Kruzkov solution it is shown that the approximation u_h obtained by the monotone finite volume scheme converges and satisfies the estimate

$$\|u-u_h\|_{L^1(Q\times[0,T])} \leq Ch^{1/4}$$

on each compact subset $Q\subset\mathbb{R}^d$. Theorem 4.4 allows us to prove a better estimate with respect to the order if u is a classical solution.

Theorem 5.1. *Consider the Cauchy problem* (1.1), (1.2) *for fluxes G_i, $i=1,\ldots,d$, and source B such that* (5.1) *and* (5.2) *hold. Assume that this Cauchy problem has a classical solution u such that Assumption* 1.1 *for u (and u_0) is satisfied with some $\bar{u}\in\mathbb{R}^m$.*

Let for $M=\exp(\beta T)\|u_0-\bar{u}\|_{L^\infty(\mathbb{R}^d)}$ the time step restriction (5.8) *hold and assume the mesh conditions* (3.1), (3.2), (3.3).

Let a finite volume approximation u_h obtained by a finite volume scheme (3.8) *with numerical fluxes satisfying* (3.10), (3.11), (5.4), (5.5), (5.6), *and* (5.7) *be given. Then there exists constants $C_1, C_2>0$ such that we have*

$$\|u_h(.,t)-\bar{u}\|_{L^\infty(\mathbb{R}^d)} \leq M \quad (t\in[0,T]), \tag{5.10}$$

$$\|u_h(.,t)-\bar{u}\|_{L^2(\mathbb{R}^d)} \leq \exp(C_1\beta T)\|u_0-\bar{u}\|_{L^2(\mathbb{R}^d)} \quad (t\in[0,T]), \tag{5.11}$$

and

$$\|u-u_h\|_{L^2(\mathbb{R}^d\times[0,T])} \leq C_2 h^{1/2}. \tag{5.12}$$

The constant C_1 depends on B and the constant C_2 depends on T, G_i, β, and u_0 but not on h.

Proof. Assume that we have have proven for some $n\in\mathcal{N}$

$$\|u_h(.,t)-\bar{u}\|_{L^\infty(\mathbb{R}^d)} \leq \exp\left(\beta t^n\right)\|u_0-\bar{u}\|_{L^\infty(\mathbb{R}^d)}. \tag{5.13}$$

Since Δt satisfies the condition (5.8) we deduce from (5.2) and (5.9) for $k=1,\ldots,m$ with $\bar{z}=\bar{u}_k$

$$\begin{aligned}|u^{n+1}_{K,k}-\bar{u}_k| &\leq \left| u^n_{K,k}-\bar{u}_k - \frac{\Delta t}{|K|}\sum_{e\in\mathcal{E}(K)}|e|h^n_{K,e}(u^n_{K,k},u^n_{K_e,k})\right| + \Delta t\beta\|u_h(.,t^n)-\bar{u}\|_{L^\infty(\mathbb{R}^d)}\\ &\leq (1+\Delta t\beta)\|u_h(.,t^n)-\bar{u}\|_{L^\infty(\mathbb{R}^d)}\\ &\leq \exp\left(\beta t^{n+1}\right)\|u_0-\bar{u}\|_{L^\infty(\mathbb{R}^d)}.\end{aligned}$$

The statement (5.10) follows now by induction.

Finite volume schemes for weakly coupled systems are analyzed in [25]. In Lemma 4.4 of [25] it has been proven under the assumptions of the theorem that (H.2) in Assumption 3.1 holds provided u_h is bounded in L^∞ and we choose our entropy tuple $(\eta, q_1, \dots, q_d)$ according to $\eta_k(w_k) = (w_k - \bar{u})^2/2$, $k = 1, \dots, m)$, in (5.3). This implies in particular (5.11) if we proceed iteratively like for the L^∞-bound. Since Assumption 1.1 and 3.1 hold we can apply Theorem 4.4 and get (5.12). □

5.2. System of elastodynamics. The next application is the system of elastodynamics in one dimension. For the unknown vector $u = (w, v)^T$ with strain $w : \mathbb{R} \times [0, T) \to [-1, \infty)$ and velocity $v : \mathbb{R} \times [0, T) \to \mathbb{R}$ the system is given by

$$\partial_t w - \partial_x v = 0, \quad \partial_t v - \partial_x \sigma(w) = 0 \text{ in } \mathbb{R} \times (0, T), \tag{5.14}$$

$$w(x, 0) = w_0(x), \qquad v(x, 0) = v_0(x) \text{ in } \mathbb{R}. \tag{5.15}$$

The stress-strain relation $\sigma \in C^2(\mathbb{R})$ is supposed to satisfy $\sigma'(w) > 0$, $w\,\sigma''(w) > 0$ for all $w \in \mathbb{R}$. Then the system has the strictly convex entropy pair $\eta(w, v) = \frac{1}{2}v^2 + \int^w \sigma(s)\,ds$, $q(w, v) = -v\sigma(w)$. It is known (see [7]) for $N > 0$ that the compact set

$$S_N = \{(w, v) \subset [-1, \infty) \times \mathbb{R} : |y(w, v)| \le N, \ |z(w, v)| \le N\} \tag{5.16}$$

is convex and invariant for the underlying system, where $y(w, v) = -\int_{w_0}^{w} \sqrt{\sigma'(s)}\,ds + v$, $z(w, v) = -\int_{w_0}^{w} \sqrt{\sigma'(s)}\,ds - v$ are the Riemann invariants.

Moreover S_N is invariant for the discrete solution operator defined through the Lax–Friedrichs scheme

$$u_i^{n+1} = u_i^n - \frac{\alpha}{2}\big(f(u_{i+1}^n) - f(u_{i-1}^n)\big) + \frac{1}{2}\big(u_{i-1}^n - 2u_i^n + u_{i+1}^n\big),$$

provided $\alpha \sup_{(w,v)\in S_N} \sqrt{\sigma'(w)} \le 1$. Here we used $f(u) = (-v, -\sigma(w))^T$. This can be shown in the analogous way as it was done in [12] for the appropriate invariant regions of the p-system.

As the numerical scheme for (5.14), (5.15) we consider now

$$u_i^{n+1} = u_i^n - \lambda\big[g(u_i^n, u_{i+1}^n) - g(u_{i-1}^n, u_i^n)\big], \tag{5.17}$$

$$u_i^0 = \frac{1}{\Delta x}\int_{x_i}^{x_{i+1}} u_0(x)\,dx \tag{5.18}$$

with the uniform mesh parameters Δx, $\Delta t > 0, \lambda = \frac{\Delta t}{\Delta x}$. The numerical flux is given by $g(u_1, u_2) = \frac{1}{2}[f(u_1) + f(u_2)] + \frac{1}{2\mu}(u_1 - u_2)$, where μ is a positive parameter. For the mesh we assume only that (3.3) holds. Since (5.17) may be rewritten as

$$\begin{aligned} u_i^{n+1} &= u_i^n - \frac{\lambda}{2}\big[f(u_{i+1}^n) - f(u_{i-1}^n)\big] + \frac{\lambda}{2\mu}\big(u_{i-1}^n - 2u_i^n + u_{i+1}^n\big) \\ &= \Big(1 - \frac{\lambda}{\mu}\Big)u_i^n + \frac{\lambda}{\mu}\Big[u_i^n - \frac{\mu}{2}\big(f(u_{i+1}^n) - f(u_{i-1}^n)\big) + \frac{1}{2}\big(u_{i-1}^n - 2u_i^n + u_{i+1}^n\big)\Big], \end{aligned}$$

one infers that S_N is an invariant region for (5.17) as well, if $\lambda/\mu \le 1$ and $\mu \sup_{(w,v)\in S_N} \sqrt{\sigma'(w)} \le 1$.

LEMMA 5.2. *Let S_N be an arbitrary invariant region of the form* (5.16) *and suppose that the parameters λ, μ satisfy $0 < \mu \leq \mu_0,\ \ \lambda/\mu \leq 1,\ \ \mu \sup_{(w,v)\in S_N} \sqrt{\sigma'(w)} \leq 1$,*

$$\frac{2\lambda(1+\gamma)L^2}{\mu l} \leq 1, \tag{5.19}$$

for some positive constants μ_0, L, l, γ (defined below in the proof) depending only on S_N and σ. Thereby L/μ is the Lipschitz constant for g on S_N. If $u_0 \in L^\infty(\mathbb{R})$ takes values in S_N, then we have for the scheme (5.17), (5.18)

$$\begin{aligned}
&(A1) \quad u_i^n \in S_N,\\
&(A2) \quad \eta(u_i^{n+1}) - \eta(u_i^n) + \lambda[q_h(u_i^n, u_{i+1}^n) - q_h(u_{i-1}^n, u_i^n)]\\
&\qquad\qquad + L^2\Big(\frac{\lambda}{\mu}\Big)^2\big(|u_{i+1}^n - u_i^n|^2 + |u_i^n - u_{i-1}^n|^2\big) \leq 0
\end{aligned}$$

for all $i \in \mathbb{Z},\ n \in \mathcal{N}$.

The numerical entropy flux in this case is $q_h(u_1,u_2) := \frac{1}{2}[q(u_1)+q(u_2)] + \frac{1}{2\mu}[\eta(u_1) - \eta(u_2)]$.

Proof. (A1) has already been proved above. By the Taylor expansion, we get

$$\begin{aligned}
E_{in} &:= \eta(u_i^{n+1}) - \eta(u_i^n) + \lambda\big[q_h(u_i^n, u_{i+1}^n) - q_h(u_{i-1}^n, u_i^n)\big]\\
&= \nabla\eta(u_i^n)\cdot\big[-\lambda\big(g(u_i^n,u_{i+1}^n) - g(u_{i-1}^n,u_i^n)\big)\big] + \lambda\big[q_h(u_i^n, u_{i+1}^n) - q_h(u_{i-1}^n, u_i^n)\big]\\
&\quad + \frac{1}{2}\lambda^2\big[g(u_i^n,u_{i+1}^n) - g(u_{i-1}^n,u_i^n)\big]^T \nabla^2\eta(\xi)\,\big[g(u_i^n,u_{i+1}^n) - g(u_{i-1}^n,u_i^n)\big]\\
&= E_{in}^{(1)} + E_{in}^{(2)} + \frac{1}{2}\lambda^2\big[g(u_i^n,u_{i+1}^n) - g(u_{i-1}^n,u_i^n)\big]^T \nabla^2\eta(\xi)\,\big[g(u_i^n,u_{i+1}^n) - g(u_{i-1}^n,u_i^n)\big],
\end{aligned}$$

where

$$\begin{aligned}
E_{in}^{(1)} &= \lambda\nabla\eta(u_i^n)\cdot\big[g(u_i^n,u_i^n) - g(u_i^n,u_{i+1}^n)\big] + \lambda\big[q_h(u_i^n,u_{i+1}^n) - q_h(u_i^n,u_i^n)\big],\\
E_{in}^{(2)} &= \lambda\nabla\eta(u_i^n)\cdot\big[g(u_{i-1}^n,u_i^n) - g(u_i^n,u_i^n)\big] + \lambda\big[q_h(u_i^n,u_i^n) - q_h(u_{i-1}^n,u_i^n)\big].
\end{aligned}$$

Define $H(a,b) := \lambda\nabla\eta(a)\cdot[g(a,a) - g(a,b)] + \lambda[q_h(a,b) - q_h(a,a)]$. In our particular case, one easily verifies that for $a = (a_1,a_2),\ b = (b_1,b_2)$,

$$\begin{aligned}
H(a,b) = \frac{\lambda}{2}\Big[&\sigma(a_1)(b_2 - a_2) + a_2\big(\sigma(b_1) - \sigma(a_1)\big) + \frac{1}{\mu}\sigma(a_1)(b_1 - a_1) + \frac{1}{\mu}a_2(b_2 - a_2)\\
&+ a_2\sigma(a_1) - b_2\sigma(b_1) + \frac{1}{2\mu}a_2^2 + \frac{1}{\mu}\int^{a_1}\sigma(s)\,ds - \frac{1}{2\mu}b_2^2 - \frac{1}{\mu}\int^{b_1}\sigma(s)\,ds\Big],
\end{aligned}$$

$$\nabla_b^2 H(a,b) = -\frac{\lambda}{\mu}\begin{bmatrix} \frac{1}{2}\sigma'(b_1) - \frac{\mu}{2}(a_2 - b_2)\sigma''(b_1) & \frac{\mu}{2}\sigma'(b_1)\\ \frac{\mu}{2}\sigma'(b_1) & \frac{1}{2}\end{bmatrix} =: -\frac{\lambda}{\mu}A(a,b).$$

Moreover, $H(a,a) = 0,\ \ \nabla_b H(a,a) = 0$. Since $A(a,b)$ is positive definite for $\mu = 0$ and $\inf_{S_N}\sigma' > 0$, we conclude that there exist positive constants $\mu_0,\ l$ with the property $\lambda_2 \geq l$ for all $0 \leq \mu \leq \mu_0$ and $a,b \in S$, where λ_2 is the lower eigenvalue of $A(a,b)$. Summarizing the considerations presented above, one obtains $H(a,b) \leq -\frac{\lambda}{\mu}\cdot\frac{l}{2}|a-b|^2$ for all $a,b \in S_N$. From $E_{in}^{(1)} = H(u_i^n, u_{i+1}^n)$ we conclude $E_{in}^{(1)} \leq -\frac{\lambda}{\mu}\cdot\frac{l}{2}|u_{i+1}^n - u_i^n|^2$.

Considering $\bar{H}(a,b) := \lambda\nabla\eta(a)\cdot[g(b,a)-g(a,a)]+\lambda[q_h(a,a)-q_h(b,a)]$ in the similar way, we have $E_{in}^{(2)} \le -\frac{\lambda}{\mu}\cdot\frac{l}{2}|u_i^n-u_{i-1}^n|^2$. Let $L>0$ be a constant which depends on S_N, such that L/μ is a Lipschitz constant for g on S_N. Finally, by (5.19), with $\gamma = \max\{\sup_{a\in S}\sigma'(a_1),1\}$, it is

$$\begin{aligned} E_{in} &\le \left(-\frac{\lambda l}{2\mu}+\left(\frac{\lambda}{\mu}\right)^2\gamma L^2\right)\left(|u_{i+1}^n-u_i^n|^2+|u_i^n-u_{i-1}^n|^2\right)\\ &\le -L^2\left(\frac{\lambda}{\mu}\right)^2\left(|u_{i+1}^n-u_i^n|^2+|u_i^n-u_{i-1}^n|^2\right). \qquad \square \end{aligned}$$

Taking into account (A1), (A2) we see that Assumption (3.1) can be satisfied for $d=1$. Note that the scheme (5.17), (5.18) matches with the multidimensional scheme (3.8), if one takes $K=(e_-,\,e_+)$, $\mathcal{E}(K)=\{e_-,e_+\}, |e_\pm|=1$, and $n_{K_{e_\pm}}=\pm1$.

Thus Theorem (4.4) applies and we have an error estimate for the Lax–Friedrichs finite volume approximation for (5.14), (5.15) if the latter has a classical solution (which for general smooth initial data can only be guaranteed for T small).

THEOREM 5.3. *Suppose that the Cauchy problem* (5.14), (5.15) *has a classical solution u that satisfies Assumption* 1.1 *for $d=1$ with S having the form* (5.16). *Furthermore let λ, μ satisfy the conditions stated in the Lemma* 5.2.

Then for the numerical solution u_h defined from the scheme (5.17), (5.18) *and the uniform mesh satisfying* (3.3), *the error estimate from Theorem* 4.4 *holds with the constant C depending only on u, c_1, σ, and S.*

5.3. Entropy conservative and entropy dissipative schemes for arbitrary sytems. We consider finally general systems (1.1) of balance laws and suppose that the initial value problem (1.1), (1.2) has a classical solution u that obeys Assumption 1.1.

We shall construct a finite volume scheme that satisfies the cell entropy inequality (H.2) together with the weak H^1-bound. The technique relies on the auxiliary construction of a scheme that satisfies a cell entropy equation, i.e., provides no entropy dissipation. Adding of artificial viscosity then leads to an entropy dissipative scheme. The scheme will be constructed in three steps starting from two semidiscrete versions to exemplify the idea behind the construction and a final step in section 5.3.3. The final finite volume scheme will be given as an implicit-in-time scheme. Throughout the section we assume $d=3$.

5.3.1. Entropy variables and entropy conservative schemes. As the first step to derive the entropy conservative schemes we introduce entropy variables [9]. Since the entropy η is supposed to be uniformly convex function the relation

$$\tilde{u}(u) = \nabla\eta(u) \qquad (u\in\mathcal{U})$$

defines a change of variables on $\mathcal{U}$. The function $\tilde{u}=\tilde{u}(u)$ is called entropy variable. We define furthermore the entropy variable state space $\tilde{\mathcal{U}}=\tilde{u}(\mathcal{U})$ and denote for $i=1,2,3$

$$\tilde{G}_i=\tilde{G}_i(\tilde{u}) := G_i(u), \quad \tilde{q}_i=\tilde{q}_i(\tilde{u}) := q_i(u).$$

Thus we can in particular rewrite the system (1.1) in the equivalent form (neglecting the source term B for simplicity)

$$\partial_t u+\sum_{i=1}^{3}\partial_i\tilde{G}_i(\tilde{u}) = 0 \text{ in } \mathbb{R}^3\times(0,T).$$

It turns out to be convenient to work with the function

$$\psi(\tilde{u}) = \begin{pmatrix} \psi_1(\tilde{u}) \\ \psi_2(\tilde{u}) \\ \psi_3(\tilde{u}) \end{pmatrix} := \begin{pmatrix} \tilde{u} \cdot \tilde{G}_1(\tilde{u}) - \tilde{q}_1(\tilde{u}) \\ \tilde{u} \cdot \tilde{G}_2(\tilde{u}) - \tilde{q}_2(\tilde{u}) \\ \tilde{u} \cdot \tilde{G}_3(\tilde{u}) - \tilde{q}_3(\tilde{u}) \end{pmatrix} \quad (\tilde{u} \in \tilde{\mathcal{U}}). \tag{5.20}$$

We observe that ψ satisfies

$$\tilde{G}_i(\tilde{u}) = \nabla\psi_i(\tilde{u}) \quad (i = 1, 2, 3,\ \tilde{u} \in \tilde{\mathcal{U}}). \tag{5.21}$$

Let us consider for $t \in [0, T]$ the following semidiscrete finite volume scheme

$$u'_K(t) = -\frac{1}{|K|} \sum_{e \in \mathcal{E}(K)} |e| \tilde{g}_{K,e}(\tilde{u}_K(t), \tilde{u}_{K_e}(t)), \qquad u_K(0) = \frac{1}{|K|} \int_K u_0(x)\, dx. \tag{5.22}$$

The numerical flux function $\tilde{g}_{K,e} : \mathbb{R}^{2m} \to \mathbb{R}^m$ is given for $\tilde{u}_1, \tilde{u}_2 \in \tilde{\mathcal{U}}$ by

$$\tilde{g}_{K,e}(\tilde{u}_1, \tilde{u}_2) = \sum_{i=1}^{3} n^i_{K,e} \tilde{g}^*_i(\tilde{u}_1, \tilde{u}_2), \tag{5.23}$$

where $n_{K,e} = (n^1_{K,e}, n^2_{K,e} n^3_{K,e})^T$ is as before the unit outward normal of the edge e of K and $\tilde{g}^*_i$, $i = 1, 2, 3$ is given by the three-dimensionally generalized *Tadmor flux*

$$\tilde{g}^*_i(\tilde{u}_1, \tilde{u}_2) = \int_0^1 \tilde{G}_i(\tilde{u}_1 + s(\tilde{u}_1 - \tilde{u}_2))\, ds. \tag{5.24}$$

The vector $\tilde{u}_K(t)$ in (5.22) is given for $t \in [0, T]$ by

$$\tilde{u}_K(t) = \tilde{u}(u_K(t)).$$

The flux function $\tilde{g}^*_i$ has been introduced in [29] to construct entropy conservative schemes in one space dimension. An overview on this type of scheme can be found in [30].

THEOREM 5.4. *Let there be a solution $u_K : [0, T] \to \mathbb{R}^m$ of* (5.22). *Then for $K \in \mathcal{T}$ and $t \in [0, T]$ the function u_K satisfies the cell entropy equation*

$$\frac{d}{dt}\eta(u_K(t)) + \frac{1}{|K|} \sum_{e \in \mathcal{E}(K)} |e| \tilde{q}_{K,e}(\tilde{u}_K(t), \tilde{u}_{K_e}(t)) = 0. \tag{5.25}$$

The numerical entropy flux $\tilde{q}_{K,e} : \mathbb{R}^{2m} \to \mathbb{R}$ is given for $\tilde{u}_1, \tilde{u}_2 \in \tilde{\mathcal{U}}$ by

$$\tilde{q}_{K,e}(\tilde{u}_1, \tilde{u}_2) = \frac{1}{2}(\tilde{u}_1 + \tilde{u}_2) \cdot \tilde{g}_{K,e}(\tilde{u}_1, \tilde{u}_2) - \frac{1}{2} n_{K,e} \cdot (\psi(\tilde{u}_2) + \psi(\tilde{u}_1)). \tag{5.26}$$

$\tilde{q}_{K,e}$ is consistent with the entropy flux, i.e.,

$$\tilde{q}_{K,e}(\tilde{u}, \tilde{u}) = \sum_{i=1}^{3} n^i_{K,e} \tilde{q}_i(\tilde{u}) \quad (\tilde{u} \in \tilde{\mathcal{U}}).$$

Proof. From (5.21) and the definition (5.24) of the Tadmor flux we observe for $i = 1, 2, 3$ and $t \in [0, T]$

$$\begin{aligned} \psi_i(\tilde{u}_{K_e}(t)) - \psi_i(\tilde{u}_K(t)) &= \int_0^1 \nabla\psi_i(\tilde{u}_K(t) + s(\tilde{u}_{K_e}(t) - \tilde{u}_K(t))) \cdot (\tilde{u}_{K,e}(t) - \tilde{u}_K(t))\, ds \\ &= \left(\int_0^1 \tilde{G}_i(\tilde{u}_K(t) + s(\tilde{u}_{K_e}(t) - \tilde{u}_K(t)))\, ds \right) \cdot (\tilde{u}_{K_e}(t) - \tilde{u}_K(t)) \\ &= \tilde{g}^*_i(\tilde{u}_K(t), \tilde{u}_{K_e}(t)) \cdot (\tilde{u}_{K_e}(t) - \tilde{u}_K(t)). \end{aligned}$$

Thus we compute with the definition of the numerical entropy flux $\tilde{q}_{K,e}$

$$\begin{aligned}
&\sum_{e\in\mathcal{E}(K)} |e|\tilde{q}_{K_e}(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\\
&= \frac{1}{2}\sum_{e\in\mathcal{E}(K)} |e|(\tilde{u}_K(t)+\tilde{u}_{K_e}(t))\cdot\tilde{g}_{K,e}(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\\
&\quad -\frac{1}{2}\sum_{e\in\mathcal{E}(K)} |e|n_{K,e}\cdot(\psi(\tilde{u}_K(t))+\psi(\tilde{u}_{K_e}(t)))\\
&= \frac{1}{2}\sum_{e\in\mathcal{E}(K)} |e|(\tilde{u}_K(t)+\tilde{u}_{K_e}(t))\cdot\tilde{g}_{K,e}(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\\
&\quad -\frac{1}{2}\sum_{e\in\mathcal{E}(K)} |e|n_{K,e}\cdot(\psi(\tilde{u}_{K_e}(t))-\psi(\tilde{u}_K(t)))\\
&= \frac{1}{2}\sum_{e\in\mathcal{E}(K)} |e|(\tilde{u}_K(t)+\tilde{u}_{K_e}(t))\cdot\tilde{g}_{K,e}(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\\
&\quad -\frac{1}{2}\sum_{e\in\mathcal{E}(K)} |e|n_{K,e}\cdot\begin{pmatrix}\tilde{g}_1^*(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\cdot(\tilde{u}_{K_e}(t)-\tilde{u}_K(t))\\ \tilde{g}_2^*(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\cdot(\tilde{u}_{K_e}(t)-\tilde{u}_K(t))\\ \tilde{g}_3^*(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\cdot(\tilde{u}_{K_e}(t)-\tilde{u}_K(t))\end{pmatrix}\\
&= \frac{1}{2}\sum_{e\in\mathcal{E}(K)} |e|(\tilde{u}_K(t)+\tilde{u}_{K_e}(t))\cdot\tilde{g}_{K,e}(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\\
&\quad -\frac{1}{2}\sum_{e\in\mathcal{E}(K)} |e|\tilde{g}_{K,e}(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\cdot(\tilde{u}_{K_e}(t)-\tilde{u}_K(t))\\
&= \sum_{e\in\mathcal{E}(K)} |e|\tilde{g}_{K,e}(\tilde{u}_K(t),\tilde{u}_{K_e}(t))\cdot\tilde{u}_K(t).
\end{aligned}$$

From the last equation we conclude that multiplication of (5.22) with $\tilde{u}_K(t) = \nabla\eta(u_K(t))$ implies (5.25). The consistency of $\tilde{q}_{K,e}$ can be checked in a straightforward manner. □

5.3.2. Entropy dissipative schemes. Based on the results of section 5.3 we shall now propose a class of schemes that satisfies condition (H.2) from Assumption 3.1 written in the semidiscrete case.

For $K \in \mathcal{T}$ and $e \in \mathcal{E}(K)$ let the numbers $\lambda_{K,e} \in \mathbb{R}_{>0}$ be given such that we have for all $K \in \mathcal{T}$ and $e \in \mathcal{E}(K)$ the relation $\lambda_{K_e,e} = \lambda_{K,e}$. Here K_e denotes the volume that shares the edge e with K. We define the entropy dissipative flux $\tilde{g}^{\triangle}_{K,e} : \mathbb{R}^{2m} \to \mathbb{R}^m$ by

$$\tilde{g}^{\triangle}_{K,e}(\tilde{u}_1,\tilde{u}_2) = \tilde{g}_{K,e}(\tilde{u}_1,\tilde{u}_2) + \frac{1}{\lambda_{K,e}}(\tilde{u}_1-\tilde{u}_2) \quad (\tilde{u}_1,\tilde{u}_2 \in \tilde{\mathcal{U}}).$$

Using this (consistent) numerical flux we get for $t \in [0, T]$ the following semidiscrete finite volume scheme to approximate solutions of (1.1), (1.2).

$$u_K'(t) = -\frac{1}{|K|} \sum_{e \in \mathcal{E}(K)} |e| \tilde{g}^{\triangle}_{K,e}(\tilde{u}_K(t), \tilde{u}_{K_e}(t)), \quad u_K(0) = \frac{1}{|K|} \int_K u_0(x)\, dx. \tag{5.27}$$

The construction is based on artificial viscosity and therefore it is possible to get a cell entropy inequality as stated below.

THEOREM 5.5. *Let there be a solution $u_K : [0, T] \to \mathbb{R}^m$ of* (5.27). *Then for $K \in \mathcal{T}$ and $t \in [0, T]$ the function u_K satisfies*

$$\begin{aligned} &\frac{d}{dt}\eta(u_K(t)) + \frac{1}{|K|} \sum_{e \in \mathcal{E}(K)} |e| \tilde{q}^{\triangle}_{K,e}(\tilde{u}_K(t), \tilde{u}_{K_e}(t)) \\ &\qquad = -\frac{1}{2|K|} \sum_{e \in \mathcal{E}(K)} \frac{|e|}{\lambda_{K,e}} |\tilde{u}_{K_e}(t) - \tilde{u}_K(t)|^2 \leq 0. \end{aligned} \tag{5.28}$$

The numerical entropy flux $\tilde{q}^{\triangle}_{K,e} : \mathbb{R}^{2m} \to \mathbb{R}$ is given for $\tilde{u}_1, \tilde{u}_2 \in \mathbb{R}^m$ by

$$\tilde{q}^{\triangle}_{K,e}(\tilde{u}_1, \tilde{u}_2) = \tilde{q}_{K,e}(\tilde{u}_1, \tilde{u}_2) + \frac{1}{2\lambda_{K,e}}(\tilde{u}_1 - \tilde{u}_2) \cdot (\tilde{u}_1 + \tilde{u}_2). \tag{5.29}$$

$\tilde{q}^{\triangle}_{K,e}$ is consistent with the entropy flux, i.e.,

$$\tilde{q}^{\triangle}_{K,e}(\tilde{u}, \tilde{u}) = \sum_{i=1}^{3} n^i_{K,e} \tilde{q}_i(\tilde{u}) \quad (\tilde{u} \in \mathcal{U}).$$

Proof. For $K \in \mathcal{T}$ and $t \in [0, T]$ we consider the term

$$\begin{aligned} &\sum_{e \in \mathcal{E}(K)} \frac{|e|}{\lambda_{K,e}} (\tilde{u}_K(t) - \tilde{u}_{K_e}(t)) \cdot \tilde{u}_K(t) \\ &\qquad = \sum_{e \in \mathcal{E}(K)} \frac{|e|}{2\lambda_{K,e}} |\tilde{u}_K(t) - \tilde{u}_{K_e}(t)|^2 \\ &\qquad + \sum_{e \in \mathcal{E}(K)} \frac{|e|}{2\lambda_{K,e}} (\tilde{u}_K(t) - \tilde{u}_{K_e}(t)) \cdot (\tilde{u}_K(t) + \tilde{u}_{K_e}(t)). \end{aligned}$$

The inequality (5.28) follows as in (the proof of) Theorem 5.4 and using the definition (5.29). □

5.3.3. Fully discrete entropy dissipative schemes. In this final step we propose a class of schemes that satisfies condition (H.2) from Assumption 3.1 in the fully discrete case. Relying on entropy conservative schemes and artificial dissipation these seem to be possible only in an implicit manner. Theorem 4.4 does not apply directly to implicit schemes (5.30) since the theorem is formulated and proven for the explicit case. However, the implicit version can be proven along the same lines.

We rely on the notations from section 5.3.2.

Consider the following fully discrete finite volume scheme for (1.1), (1.2) neglecting the source term for simplicity:

$$\begin{aligned} u_K^{n+1} &= u_K^n - \frac{\Delta t}{|K|} \sum_{e\in\mathcal{E}(K)} |e| \tilde{g}^{\triangle}_{K,e}(\tilde{u}_K^{n,n+1}, \tilde{u}_{K_e}^{n,n+1}), \\ \tilde{u}_K^{n,n+1} &= \int_0^1 \tilde{u}\big(s u_K^{n+1} + (1-s) u_K^n\big)\, ds, \qquad (5.30) \\ u_K^0 &= \frac{1}{|K|} \int_K u_0(x)\, dx. \end{aligned}$$

For this implicit version we get the following theorem.

THEOREM 5.6. *For $\Delta t/h$ sufficiently small there is a solution $u_K^n \in \mathbb{R}^m$ of* (5.30) *for each $n \in \mathcal{N}$ and $K \in \mathcal{T}$.*

The function $u_h : \mathbb{R}^3 \times [0,T] \to \mathbb{R}^m$ defined as in (3.9) *satisfies*

$$\begin{aligned} &\eta(u_K^{n+1}) - \eta(u_K^n) + \frac{\Delta t}{|K|} \sum_{e\in\mathcal{E}(K)} |e| \tilde{q}^{\triangle}_{K,e}(\tilde{u}_K^{n,n+1}, \tilde{u}_{K_e}^{n,n+1}) \qquad (5.31) \\ &\quad + \frac{\Delta t}{2|K|} \sum_{e\in\mathcal{E}(K)} \frac{|e|}{\lambda_{K,e}} \big|\tilde{u}_K^{n,n+1} - \tilde{u}_{K_e}^{n,n+1}\big|^2 \leq 0. \end{aligned}$$

Proof. The proof of existence follows easily with the implicit function theorem.

To prove the cell entropy inequality we multiply (5.30) with $\tilde{u}_K^{n,n+1}$. Then we obtain the numerical entropy flux term and the dissipation term in (5.31) exactly as in the proof of Theorem 5.5 if we interchange $\tilde{u}_K(t)$ with $\tilde{u}_K^{n,n+1}$ (and $\tilde{u}_{K_e}(t)$ with $\tilde{u}_{K_e}^{n,n+1}$). It remains to consider the product

$$\big(u_K^{n+1} - u_K^n\big) \cdot \tilde{u}_K^{n,n+1} = \int_0^1 \nabla\eta\big(s u_K^{n+1} + (1-s) u_K^n\big) \cdot \big(u_K^{n+1} - u_K^n\big)\, ds = \eta(u_K^{n+1}) - \eta(u_K^n).$$

This concludes the proof. □

REFERENCES

[1] C. Arvanitis, C. Makridakis, and A. Tzavaras, *Stability and convergence of a class of finite element schemes for hyperbolic systems of conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 1357–1393.

[2] Y. Brenier, *Hydrodynamic structure of the augmented Born-Infeld equations*, Arch. Rational Mech. Anal., 172 (2004), pp. 65–91.

[3] B. Cockburn, F. Coquel, and P.G. LeFloch, *An error estimate for finite volume methods for multidimensional conservation laws*, Math. Comp., 63 (1994), pp. 77–103.

[4] C.M. Dafermos, *The second law of thermodynamics and stability*, Arch. Rational Mech. Anal., 94 (1979), pp. 373–389.

[5] C.M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Springer-Verlag, New York, 2000.

[6] R. DiPerna, *Uniqueness of solutions to hyperbolic conservation laws*, Indiana Univ. Math. J., 28 (1979), pp. 137–188.

[7] R. DiPerna, *Convergence of approximate solutions to conservation laws*, Arch. Ration. Mech. Anal., 82 (1983), pp. 27–70.

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

[9] K.O. Friedrichs and P.D. Lax, *Systems of conservation laws with a convex extension*, Proc. Natl. Acad. Sci. USA, 68 pp. 1686–1688, 1971.

[10] M. Grassin, *Global smooth solutions to Euler equations for a perfect gas*, Indiana Univ. Math. J., 47 (1998), pp. 1397–1432.

[11] B. Hanouzet and R. Natalini, *Global existence of smooth solutions for partially dissipative hyperbolic systems with a convex entropy*, ARMA, 169 (2003), pp. 89–117.

[12] D. Hoff, *A finite difference scheme for a system of two conservation laws with artificial viscosity*, Math. Comput., 33 (1979), pp. 1171–1193.

[13] V. Jovanović and C. Rohde, *Finite-volume schemes for Friedrichs systems in multiple space dimensions: A-priori and a-posteriori error estimates*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 104–131.

[14] Th. Katsaounis and Ch. Makridakis, *Finite volume relaxation schemes for multidimensional conservation laws*, Math. Comput., 70 (2001), pp. 533–553.

[15] S. Kawashima, Y. Nikkuni, and S. Nishibata, *The initial value problem for hyperbolic-elliptic coupled systems and applications to radiation hydrodynamics*, in Analysis of Systems of Conservation Laws, H. Freistühler, ed., Monogr. Surv. Pure Appl. Math. 99, Chapman and Hall/CRC, Boca Raton, FL, 1998, pp. 87–127.

[16] S. Kawashima, *Global solutions to the equation of viscoelasticity with fading memory*, J. Differential Equations, 101 (1993), pp. 388–420.

[17] S. Klainerman, *The Null Condition and Global Existence to Nonlinear Wave Equations*, Lecture Notes in Appl. Math. 23, 1986.

[18] D. Kröner, *Numerical Schemes for Conservation Laws*, John Wiley, Chichester, 1997.

[19] S.N. Kruzkov, *First order quasilinear equations in several independent variables*, Math. USSR-Sb., 10 (1970), pp. 217–243.

[20] N.N. Kuznetsov, *The accuracy of some approximate methods for computing weak solutions of quasi-linear first order partial differential equation*, Zh. Vychisl. Mat. Mat. Fiz., 16 (1976), pp. 1489–1502.

[21] P.G. LeFloch, J.M. Mercier, and C. Rohde, *Fully discrete, entropy conservative schemes of arbitrary order*, SIAM J. Numer. Anal., 40 (2002), pp. 1968–1992.

[22] P.G. LeFloch and C. Rohde, *High-order schemes, entropy inequalities, and nonclassical shocks*, SIAM J. Numer. Anal., 37 (2000), pp. 2023–2060.

[23] A. Majda, *Compressible Fluid Flow and Systems of Conservation Laws in Several Space Variables*, Springer, New York, 1984.

[24] C. Rohde, *Entropy solutions for weakly coupled hyperbolic systems in several space dimensions*, Z. Angew. Math. Phys., 49 (1998), pp. 470–499.

[25] C. Rohde, *Upwind finite volume schemes for weakly coupled systems of conservation laws*, Numer. Math., 81 (1998), pp. 85–124.

[26] F. Sabac, *The optimal convergence rate of monotone finite difference methods for hyperbolic conservation laws*, SIAM J. Numer. Anal., 34 (1997), pp. 2306–2318.

[27] D. Serre *Solutions classiques globales des équations d'Euler pour un fluide parfait compressible*, Ann. Inst. Fourier, 47 (1997), pp. 139–153.

[28] E. Tadmor and T. Tang, *Pointwise error estimates for scalar conservation laws with piecewise smooth solutions*, SIAM J. Numer. Anal., 36 (1999), pp. 1739–1758.

[29] E. Tadmor, *The numerical viscosity of entropy stable schemes for systems of conservation laws*, Math. Comp., 49 (1987), pp. 91–103.

[30] E. Tadmor, *Entropy stability theory for difference approximations of nonlinear conservation laws and related time dependent problems*, http://www.cscamm.umd.edu/people/faculty/tadmore/pub/TV-and-entropy-stability/Tadmor_Acta03.pdf, Acta Numerica, (2003), pp. 451–512.

[31] T. Tang, Z.-H. Teng, and Z. Xin, *Fractional rate of convergence for viscous approximation to nonconvex conservation laws*, SIAM J. Numer. Anal., 35 (2003), pp. 98–122.

[32] T. Tang and Z.-H. Teng, *The sharpness of kuznetsov's $O(\sqrt{\Delta x})$ L^1-error estimate for monotone difference schemes*, Math. Comp., 64 (1995), pp. 581–589.

[33] L. Ta-Tsien, *Global Classical Solutions for Quasilinear Hyperbolic Systems*, Wiley/Masson, Paris, 1994.

[34] J.P. Vila, *Convergence and error estimates in finite volume schemes for general multidimensional scalar conservation laws*, RAIRO Modél. Math. Anal. Numer., 28 (1994), pp. 267–295.

[35] J.P. Vila, Lecture given at the workshop Finite-Volume Methods, Freiburg, Germany, December, 2000.

[36] L. Ying, T. Yang, and C. Zhu, *Existence of global smooth solutions for Euler equations with symmetry*, Partial Differential Equations, 22 (1998), pp. 1361–1387.

[37] W.-A. Yong, *An entropy condition and global existence result for hyperbolic balance laws*, ARMA, 172 (2004), pp. 247–266.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2450–2477

A WAVENUMBER INDEPENDENT BOUNDARY ELEMENT METHOD FOR AN ACOUSTIC SCATTERING PROBLEM*

S. LANGDON[†] AND S. N. CHANDLER-WILDE[†]

Abstract. In this paper we consider the impedance boundary value problem for the Helmholtz equation in a half-plane with piecewise constant boundary data, a problem which models, for example, outdoor sound propagation over inhomogeneous flat terrain. To achieve good approximation at high frequencies with a relatively low number of degrees of freedom, we propose a novel Galerkin boundary element method, using a graded mesh with smaller elements adjacent to discontinuities in impedance and a special set of basis functions so that, on each element, the approximation space contains polynomials (of degree ν) multiplied by traces of plane waves on the boundary. We prove stability and convergence and show that the error in computing the total acoustic field is $\mathcal{O}(N^{-(\nu+1)}\log^{1/2} N)$, where the number of degrees of freedom is proportional to $N \log N$. This error estimate is independent of the wavenumber, and thus the number of degrees of freedom required to achieve a prescribed level of accuracy does not increase as the wavenumber tends to infinity.

Key words. Galerkin method, high frequency, Helmholtz equation

AMS subject classifications. 35J05, 65R20

DOI. 10.1137/S0036142903431936

1. Introduction.

High-frequency scattering problems are of enormous interest to the mathematics, physics, and engineering communities, with applications to electromagnetic scattering, radar problems, high frequency acoustics, and geophysical waves. Although these problems have a long pedigree, their numerical solution continues to pose considerable difficulties. Many problems of scattering of time-harmonic acoustic or electromagnetic waves can be formulated as the Helmholtz equation

$$\Delta u + k^2 u = 0, \tag{1.1}$$

in $\mathbb{R}^d \backslash \Omega$, $d = 2, 3$, supplemented with appropriate boundary conditions. Here Ω is the scattering object and $k > 0$ (the wavenumber) is an arbitrary positive constant, proportional to the frequency of the incident wave.

Standard schemes for solving (1.1) become prohibitively expensive as $k \to \infty$. For standard boundary element or finite element schemes, where the approximation space typically consists of piecewise polynomials, the number of degrees of freedom per wavelength must remain fixed in order to maintain accuracy, with the rule of thumb in the engineering literature a requirement for 6 to 10 elements per wavelength. Often in applications this results in excessively large systems when the wavelength is small compared to the size of the obstacle. These difficulties have been well documented; see, for example, [44, 45]. For the finite element method the situation is arguably worse in that additional pollution effects are known to be important [5, 33], these being phase errors in wave propagation across the domain, so that the degrees of freedom per wavelength need to increase somewhat to retain accuracy as k increases.

The development of more efficient numerical schemes for high frequency scattering problems has attracted much recent attention in the literature. In the case of

*Received by the editors July 22, 2003; accepted for publication (in revised form) July 14, 2005; published electronically January 6, 2006. This work was supported by the EPSRC via grant GR/M59433/01 and by a Leverhulme Trust Early Career Fellowship awarded to the first author.

http://www.siam.org/journals/sinum/43-6/43193.html

[†]Department of Mathematics, University of Reading, Whiteknights, PO Box 220, Berkshire RG6 6AX, UK (s.langdon@reading.ac.uk, s.n.chandler-wilde@reading.ac.uk).

boundary element methods, a great deal of effort has focused on the fast solution of the large systems which arise, using preconditioned iterative methods (e.g., [22]) combined with fast multipole (e.g., [24, 25]) or fast Fourier transform based methods (e.g., [9, 21]) to carry out the matrix-vector multiplications efficiently. The reduction in the computing cost achieved by the use of these schemes increases the upper limit on the frequency for which accurate results can be obtained in a reasonable time. However, as the size of the system still grows at least linearly with k in two dimensions (2D), quadratically in three dimensions (3D), this upper limit is not removed altogether.

An increasingly popular approach in the literature for higher frequencies is to use either a finite element or a boundary element method in which the approximation space is enriched with plane wave or Bessel function solutions of (1.1), in order to represent efficiently the highly oscillatory solution when k is large. This idea has been applied to both finite element (e.g., [4, 13, 41, 29, 14, 43]) and boundary element schemes [26, 1, 24, 44, 45, 20, 43, 27, 10, 46, 28]. Promising numerical results are reported, but most of the papers are lacking in mathematical analysis, especially with regard to how any error estimates depend on the wavenumber k. As the present paper follows the same general approach of enriching the approximation space, we survey this body of work in a little more detail.

The methods in this category fall approximately into three groups, distinguished by how the enrichment is carried out. In one group the distinguishing feature is that a large number of solutions of the Helmholtz equation are used to form the approximation space. Most commonly the approximation space consists of standard finite element basis functions multiplied by plane waves traveling in a large number of directions, approximately uniformly distributed on the unit circle (in 2D) or sphere (in 3D). This is the approach in the generalized finite element method of Babuška and Melenk [4], the ultra weak variational formulation of Cessenat and Després [13, 14], and the least squares method of Monk and Wang [41]; see also [43, 32, 37]. In the boundary element context this approach is used in the microlocal discretization method of de La Bourdonnaye et al. [26, 27] and in the work of Perrey-Debain et al. [44, 45, 43, 46]. The theoretical analysis carried out (e.g., [4]) and computational results (e.g., [43]) confirm that these methods converge very rapidly as the number of plane wave directions used increases. Moreover, the computational results suggest that to achieve a required accuracy, the number of degrees of freedom needed is reduced by a large factor compared with conventional h-version finite or boundary element methods. However, in the case of boundary element methods, while constants of proportionality are reduced very significantly [44, 45, 43, 46], it is not clear that, asymptotically, the number of degrees of freedom increases any less fast than linearly with k for 2D problems, the same rate of increase as for conventional boundary element methods.

At the other extreme, the second group of papers, using direct integral equation methods, are distinguished by using only one solution of the Helmholtz equation to enrich the approximation space, namely, the known incident field. This approach amounts to applying conventional boundary element methods to the ratio of the total field to the incident field, rather than to the total field directly. This simple idea, employed for the impedance boundary value problem we consider in this paper in [15], seems particularly appropriate in the case of smooth convex obstacles, as physical optics predicts that this ratio is approximately constant on the illuminated side and approximately zero on the shadow side at high frequencies. This approach is used for smooth convex obstacles in 2D in [1], where a standard Galerkin boundary element method with uniform mesh is applied to the ratio of the scattered field to the incident field. In fact, this paper appears to be the first in which the dependence of the error estimates on the wavenumber k is indicated. The error estimate stated in [1] is that the relative error in the best approximation from a boundary element space of piecewise polynomials of degree $\leq \nu$ is $\mathcal{O}(h^{\nu}) + \mathcal{O}((hk^{1/3})^{\nu+1})$. While clearly better than the (at least) linear dependence on k of conventional boundary element methods, the number of degrees of freedom needed to maintain accuracy is still predicted to grow like $k^{1/3}$ as k increases, and moreover the analysis does not guarantee that the Galerkin method solution is close to this best approximation in the limit as $k \to \infty$.

The method of [1] is applied in [24], where results for realistic 3D scatterers are shown. This approach has also recently been applied in [10]. In this latter paper, which focuses on 2D scattering by a sound-soft circular obstacle, the numerical scheme is not completely defined. However, one of the main features of the numerical scheme is that a coordinate stretching is carried out in a k-dependent neighborhood of the shadow boundary (of length $O(k^{-1/3})$). The numerical results in [10, 11] for scattering by a circle suggest that after this transformation, the slowly varying normal derivative of the ratio of scattered to incident field can be approximated using Fourier series basis functions in the L^2 norm with a number of degrees of freedom which remains fixed as the wavenumber k tends to infinity. The authors do not attempt to establish this wavenumber independence theoretically by a rigorous error analysis.

The third group of papers is intermediate in approach between the first and second groups, attempting to identify, by geometrical optics or geometrical theory of diffraction considerations or otherwise, the important wave propagation directions at high frequency. They then incorporate the oscillatory part of this high-frequency asymptotics into the approximation space for the numerical solution. This is the approach in the finite element method of Giladi and Keller [29] and in the boundary integral equation method of Bruno, Geuzaine, and Reitich [12], our own recent work [20], and the present paper. The paper [12], generalizing [10], considers specifically the case of multiple scattering between two 2D convex obstacles and employs a Neumann series approach, solving for each of the multiple scatters in turn, and factoring out a geometric optics estimate of the main oscillatory behavior at each step. We remark that the distinction between the second and third groups of papers is somewhat blurred in that, arguably, for a smooth convex obstacle the only important wave direction to include in the ansatz for the scattered field and its normal derivative on the boundary is the incident wave direction.

The last two groups of papers have in common that, while the number of degrees of freedom may be reduced, very significantly, by incorporating the oscillatory behavior of the solution, the work required to compute a typical matrix entry of the linear system to be solved increases significantly. In particular, in boundary integral equation based methods, a typical entry of the full system matrix corresponds to an integration over a part of the boundary which is large in diameter compared to the wavelength so that the integrand is highly oscillatory. The problem of efficient evaluation of these integrals is tackled by the fast multipole method in [24]. In [10], ideas from the method of stationary phase are used to reduce the support of the integrand, and quadrature rules based on the trapezoidal rule, exponentially accurate for smooth periodic functions, are employed. Numerical results using this approach for 2D scattering by a circle are encouraging and appear to indicate a fixed computational cost as

k tends to infinity. A somewhat similar approach for evaluating the integrals to that of Bruno et al. [10] is employed for scattering by smooth 3D convex scatterers in [28]. We note that there has been considerable recent interest in the efficient evaluation of highly oscillatory integrals for a variety of applications; see [34, 35] and the references therein.

As an instance of the third group of papers, the authors and Ritter recently proposed [20] a new high-frequency boundary element method for the specific problem of 2D acoustic scattering by an inhomogeneous impedance plane. For this new scheme it was shown [20] that the number of degrees of freedom needed to maintain accuracy as $k \to \infty$ grows only logarithmically with k. This appears to be the best theoretical estimate to date for a numerical method for a scattering problem in terms of the dependence on the wavenumber.

In this paper we will be concerned with the numerical solution of the same problem, proposing modifications of the numerical scheme of [20]. For our modified scheme we are able to show, employing somewhat more elaborate arguments than those of [20], that for a fixed number of degrees of freedom the error is bounded independently of the wavenumber k. To our knowledge, this is the first such numerical analysis result for any scattering problem.

The problem we will consider is one of acoustic scattering of an incident wave by a planar surface with spatially varying acoustical surface impedance. This problem has attracted much attention in the literature (see, for example, [17, 30, 31, 16, 21, 6, 48]), both in its own right and also as a model of the scattering of an incident acoustic or electromagnetic wave by an infinite rough surface [8, 47, 36, 7]. In the case in which there is no variation in the acoustical properties of the surface or the incident field in some fixed direction parallel to the surface, the problem is effectively two-dimensional. Adopting Cartesian coordinates $0x_1x_2x_3$, let this direction be that of the x_3-axis and the surface be the plane $x_2 = 0$. Assuming further that the incident wave and scattered fields are time harmonic, the total acoustic field $u^t \in \mathcal{C}(\overline{U}) \cap \mathcal{C}^2(U)$ then satisfies (1.1) in $U := \{(x_1, x_2) \in \mathbb{R}^2 : x_2 > 0\}$, supplemented with the impedance boundary condition

$$\frac{\partial u^t}{\partial x_2} + \mathrm{i}k\beta u^t = f \quad \text{on } \Gamma := \{(x_1, 0) : x_1 \in \mathbb{R}\} \tag{1.2}$$

with $f \equiv 0$, where $k = \omega/c > 0$. Here $\omega = 2\pi\mu$, μ is the frequency of the incident wave and c is the speed of sound in U. The acoustic pressure at time t, position (x_1, x_2, x_3), is then given by $\mathrm{Re}(\mathrm{e}^{-\mathrm{i}\omega t} u^t(x))$ for $x = (x_1, x_2) \in \overline{U}$.

In outdoor sound propagation, the relative surface admittance β depends on the frequency and the ground properties and is often assumed in modeling to be piecewise constant and constant outside some finite interval $[a, b]$ (see, for example, [17, 30, 31]), with β taking a different value for each ground surface type (grassland, forest floor, road pavement, etc. [3]). Thus, for some real numbers $a = t_0 < t_1 < \cdots < t_n = b$, the relative surface admittance at $(x_1, 0)$ on Γ is given by

$$\beta(x_1) = \begin{cases} \beta_j, & x_1 \in (t_{j-1}, t_j], \\ \beta_c, & x_1 \in \mathbb{R} \backslash (t_0, t_n]. \end{cases} \tag{1.3}$$

If the ground surface is to absorb rather than emit energy, the condition $\mathrm{Re}\beta \geq 0$ must be satisfied. We assume throughout that, for some $\epsilon > 0$,

$$\mathrm{Re}\beta_c \geq \epsilon, \quad \mathrm{Re}\beta_j \geq \epsilon, \quad |\beta_c| \leq \epsilon^{-1}, \quad |\beta_j| \leq \epsilon^{-1}, \quad j = 1, \ldots, n. \tag{1.4}$$

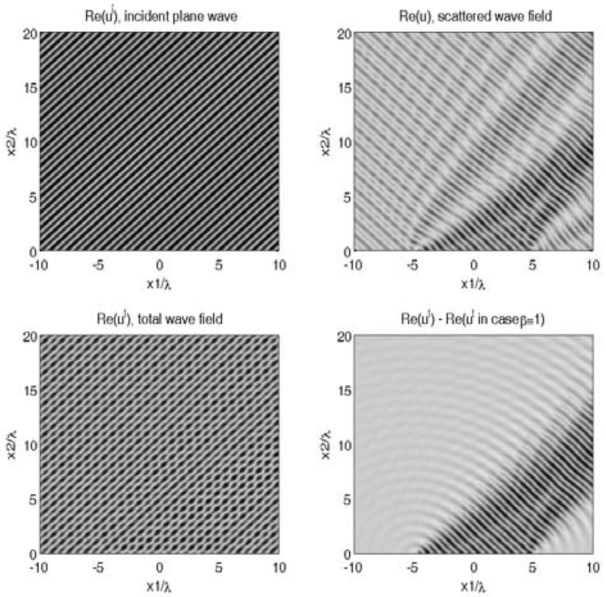


FIG. 1.1. *Acoustic scattering by an impedance boundary.*

For simplicity of exposition, we restrict our attention to the case of plane wave incidence, so that the incident field u^i is given by $u^i(x) = \exp[\mathrm{i}k(x_1 \sin\theta - x_2 \cos\theta)]$, where $\theta \in (-\pi/2, \pi/2)$ is the angle of incidence. The reflected or scattered part of the wave field is $u := u^t - u^i \in C(\overline{U}) \cap C^2(U)$, and this also satisfies (1.1) and (1.2) with

$$f(x_1) := \mathrm{i}k\mathrm{e}^{\mathrm{i}kx_1 \sin\theta}(\cos\theta - \beta(x_1)), \quad x_1 \in \mathbb{R}. \tag{1.5}$$

In Figure 1.1 we show scattering by a typical impedance plane. In this particular example, the surface admittance β is given by

$$\beta(x_1) = \begin{cases} 0.505 - 0.3\mathrm{i}, & x_1 \in (-5\lambda, 5\lambda], \\ 1, & x_1 \in \mathbb{R}\backslash(-5\lambda, 5\lambda], \end{cases}$$

where $\lambda = c/\mu = 2\pi/k$ is the wavelength. There are discontinuities in impedance at $x_1 = -5\lambda$ and at $x_1 = 5\lambda$. The incident plane wave ($\theta = \pi/4$ in this example) can be seen in the top left and the scattered wave in the top right of Figure 1.1. This scattered wave is a combination of reflected and diffracted rays. The diffracted rays, propagating radially from the points $(-5\lambda, 0)$ and $(5\lambda, 0)$, can be seen more clearly in the bottom right of Figure 1.1, where we have subtracted from the total field u^t the (known) total field in the case that $\beta \equiv 1$.

To achieve good approximations with a relatively low number of degrees of freedom, a boundary element method approach was used in [20] with ideas in the spirit of the geometrical theory of diffraction (GTD) being used to identify and subtract off the leading order behavior (namely, the incident and reflected rays) as $k \to \infty$. The remaining scattered wave (consisting of the rays diffracted at impedance discontinuities as visible in the lower right corner of Figure 1.1) can then be expressed (on the boundary Γ) as the product of the known oscillatory functions $\mathrm{e}^{\pm\mathrm{i}kx_1}$ and unknown nonoscillatory functions denoted as $f_j^{\pm}$. Rigorous bounds were established in [20] on the derivatives of the nonoscillatory functions $f_j^\pm$ both adjacent to and away from discontinuities in impedance. Using these bounds a Galerkin method was developed, using a graded mesh with elements very large compared to the wavelength away from discontinuities in β, in order to take advantage of the smooth behavior of $f_j^\pm$ away from these points, and a special set of basis functions so that on each element the approximation space consists of polynomials (of degree ν) multiplied by $\mathrm{e}^{\pm \mathrm{i}kx_1}$, so as to obtain a piecewise polynomial representation of the nonoscillatory functions $f_j^\pm$. Using this approach, it was shown in [20] that the error in computing an approximation to $u^t|_\Gamma$ on $[a,b]$ in the L_2 norm is $\mathcal{O}(\log^{\nu+3/2}(k(b-a))M^{-(\nu+1)})$, where M is the number of degrees of freedom.

In this paper we consider the same problem as in [20] and use a similar approach. We again subtract off the leading order behavior as $k \to \infty$ on each interval and express the scattered wave as a product of oscillatory and nonoscillatory functions. However, here (in section 2) we prove sharper bounds on the nonoscillatory functions $f_j^\pm$ away from impedance discontinuities. Based on these bounds, in section 3 we propose a Galerkin method similar to that in [20], but with a different approximation space. As in [20] this consists of polynomials (of degree ν) multiplied by $\mathrm{e}^{\pm \mathrm{i}kx_1}$, but unlike in [20] the choice of whether to use $\mathrm{e}^{+\mathrm{i}kx_1}$ or $\mathrm{e}^{-\mathrm{i}kx_1}$ on each element is dictated by how close the element is to each impedance discontinuity, and the graded mesh is chosen differently so that when k is large compared to N we do not discretize the entire domain. This is key to achieving a convergence rate independent of the wavenumber.

In section 3 we present an error analysis for this new approach, and we show that the error in computing an approximation to $u^t|_\Gamma$ on $[a,b]$ is $\mathcal{O}(N^{-(\nu+1)}\log^{1/2}\min(N,k(b-a)))$, in the L_2 norm, using a number of degrees of freedom proportional to $N\log\min(N,k(b-a))$. As $\min(N,k(b-a)) \le N$, this error estimate shows that the error is bounded independently of k for a fixed number of degrees of freedom. We believe this to be the first proof for any scattering problem that, for a fixed discretization, the error does not grow as the size (in terms of number of wavelengths) of the scattering object to be discretized tends to infinity. Moreover, for fixed k, as $N \to \infty$ the extra logarithmic dependence on N of the error estimate and the number of degrees of freedom disappears, and we retain the same asymptotic convergence rate as in [20].

Whereas in [20] results were proved regarding only the approximation of $u^t|_\Gamma$, here we also show, in Theorem 3.6, that the total acoustic field at any point $x \in U$ can be computed to a similar order of accuracy. In section 4 we discuss the practical implementation of our approach, and we present some numerical results demonstrating that the theoretically predicted behavior is achieved. Finally in section 5 we present some conclusions and discuss possible future extensions of the ideas presented here.

2. Integral equation formulation and regularity of the solution. In the rest of this paper, ν is the degree of the polynomial approximations used in the Galerkin method described in section 3, and ϵ, in the range $0<\epsilon<1$, is the constant in the bound (1.4). Throughout C_ϵ, C_ν, and $C_{\epsilon,\nu}$ denote constants depending only on ϵ, ν, and both ϵ and ν, respectively, each not necessarily the same at each occurence.

We begin by stating the problem we wish to solve precisely and reformulating it as an integral equation. For $H \ge 0$, let $U_H := \{(x_1,x_2) : x_2 > H\}$ and $\Gamma_H := \{(x_1,H) : x_1 \in \mathbb{R}\}$. To determine the scattered field u uniquely we impose the radiation condition proposed in [16] that, for some $H>0$, u can be written in the half plane U_H as the double layer potential

$$u(x) = \int_{\Gamma_H} \frac{\partial H_0^{(1)}(k|x-y|)}{\partial y_2} \phi(y)\,\mathrm{d}s(y), \quad x \in U_H, \tag{2.1}$$

for some density $\phi \in L_\infty(\Gamma_H)$, where $H_0^{(1)}$ is the Hankel function of the first kind of order zero. The boundary value problem that we wish to solve for u is thus as follows.

Boundary value problem. *Given $k > 0$ (the wavenumber), $\theta \in (-\pi/2, \pi/2)$ (the angle of incidence) and β given by* (1.3)*, find $u \in \mathcal{C}(\overline{U}) \cap \mathcal{C}^2(U)$ such that*

(i) *u is bounded in the horizontal strip $U \backslash U_H$ for every $H > 0$;*
(ii) *u satisfies the Helmholtz equation* (1.1) *in U;*
(iii) *u satisfies the impedance boundary condition* (1.2) *on Γ (in the weak sense explained in* [16]*), with $f \in L_\infty(\Gamma)$ given by* (1.5)*;*
(iv) *u satisfies the radiation condition* (2.1)*, for some $H > 0$ and $\phi \in L_\infty(\Gamma_H)$.*

For $\beta^* \in \mathbb{C}$ with $\mathrm{Re}\beta^* > 0$ let $G_{\beta^*}(x, y)$ denote the Green's function for the above problem in the case of constant relative surface impedance, which satisfies (1.2), with $\beta \equiv \beta^*$ and $f \equiv 0$, and the standard Sommerfeld radiation and boundedness conditions. Explicit representations and efficient calculation methods for G_{β^*} are discussed in [18]. We shall require later the following bounds on G_{β^*} [20, (2.9), (2.10)], which hold provided $\mathrm{Re}\beta^* \geq \epsilon$ and $|\beta^*| \leq \epsilon^{-1}$:

$$|G_{\beta^*}(x, y)| \leq \frac{C_\epsilon(1 + kx_2)}{(k|x-y|)^{3/2}}, \quad x \in \overline{U},\ y \in \Gamma,\ x \neq y, \tag{2.2}$$

$$|G_{\beta^*}(x, y)| \leq C_\epsilon(1 - \log(k|x-y|)), \quad x \in \overline{U},\ y \in \Gamma,\ 0 < k|x-y| \leq 1. \tag{2.3}$$

The following result is shown in [20].

THEOREM 2.1. *If u satisfies the above boundary value problem, then*

$$u(x) = \int_\Gamma G_{\beta^*}(x, y)(\mathrm{i}k(\beta(y) - \beta^*)u(y) - f(y))\,\mathrm{d}s(y), \quad x \in \overline{U}. \tag{2.4}$$

Conversely, if $u|_\Gamma \in BC(\Gamma)$ (the space of bounded and continuous functions on Γ) and u satisfies (2.4)*, for some β^* with $\mathrm{Re}\beta^* > 0$, then u satisfies the above boundary value problem. Moreover,* (2.4) *has exactly one solution with $u|_\Gamma \in BC(\Gamma)$, and hence the boundary value problem has exactly one solution.*

We denote the (known) solution of the above boundary value problem in the special case $\beta \equiv \beta^*$ by u_{β^*} and the corresponding total field by $u^t_{\beta^*} := u^i + u_{\beta^*}$. Then it is easily seen [20] that u_{β^*} is the plane wave $u_{\beta^*}(x) = R_{\beta^*}(\theta)\exp[\mathrm{i}k(x_1 \sin\theta + x_2 \cos\theta)]$, where $R_{\beta^*}(\theta) := (\cos\theta - \beta^*)/(\cos\theta + \beta^*)$ is a reflection coefficient. Moreover, it is shown rigorously in [20] that u^t satisfies

$$u^t(x) = u^t_{\beta^*}(x) + \mathrm{i}k \int_\Gamma G_{\beta^*}(x, y)(\beta(y) - \beta^*)u^t(y)\,\mathrm{d}s(y), \quad x \in \overline{U}. \tag{2.5}$$

We note that the approximate and numerical solution of this integral equation has been extensively studied; see, for example, [42, 30, 17, 21, 20].

To make explicit the dependence on the wavenumber k in the results we obtain, it is useful to introduce new, dimensionless variables. Thus, define $\phi(s) := u^t((s/k, 0))$, $\psi_{\beta^*}(s) := u^t_{\beta^*}((s/k, 0))$, and $\kappa_{\beta^*}(s) := G_{\beta^*}((s/k, 0), (0, 0))$, $s \in \mathbb{R}$. Then (2.5) restricted to Γ is the following second kind boundary integral equation for ϕ:

$$\phi(s) = \psi_{\beta^*}(s) + \mathrm{i}\int_{-\infty}^{\infty} \kappa_{\beta^*}(s - t)(\beta(t/k) - \beta^*)\phi(t)\,\mathrm{d}t, \quad s \in \mathbb{R}. \tag{2.6}$$

It is the main concern in the remainder of the paper to solve this equation numerically in the case when $\beta^* = \beta_c$. Clearly,

$$\psi_{\beta^*}(s) = (1 + R_{\beta^*}(\theta))\mathrm{e}^{\mathrm{i}s\sin\theta}, \tag{2.7}$$

and it is shown in [20], using the representation for G_{β^*} in [18], that

$$\kappa_{\beta^*}(s) = \frac{\mathrm{i}}{2}H_0^{(1)}(|s|) + \frac{\beta^{*2}\mathrm{e}^{\mathrm{i}|s|}}{\pi}\int_0^\infty \frac{t^{-1/2}e^{-|s|t}}{(t-2\mathrm{i})^{1/2}(t^2-2\mathrm{i}t-\beta^{*2})}\,\mathrm{d}t + C_{\beta^*}e^{\mathrm{i}|s|(1-\hat{a}_+)} \tag{2.8}$$

$$= \mathrm{e}^{\mathrm{i}|s|}\breve{\kappa}_{\beta^*}(s), \quad s \in \mathbb{R}\backslash\{0\}, \tag{2.9}$$

where $\hat{a}_\pm := 1 \mp (1-\beta^{*2})^{\frac{1}{2}}$, with $\mathrm{Re}\{(1-\beta^{*2})^{1/2}\} \geq 0$,

$$C_{\beta^*} := \begin{cases} \frac{\beta^*}{(1-\beta^{*2})^{1/2}}, & \mathrm{Im}\beta^* < 0, \mathrm{Re}(\hat{a}_+) < 0, \\ \frac{\beta^*}{2(1-\beta^{*2})^{1/2}}, & \mathrm{Im}\beta^* < 0, \mathrm{Re}(\hat{a}_+) = 0, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$\breve{\kappa}_{\beta^*}(s) := \frac{1}{\pi}\int_0^\infty \frac{r^{\frac{1}{2}}(r-2\mathrm{i})^{\frac{1}{2}}}{r^2-2\mathrm{i}r-\beta^{*2}}\mathrm{e}^{-r|s|}\,\mathrm{d}r + C_{\beta^*}e^{-\mathrm{i}|s|\hat{a}_+}, \quad s \in \mathbb{R}\backslash\{0\}. \tag{2.10}$$

Clearly the only dependence on k in the known terms in (2.6) is in the impedance function $\beta(t/k)$. We shall see shortly that the oscillating part of $\kappa_{\beta^*}(s)$ is contained in the factor $\mathrm{e}^{\mathrm{i}|s|}$ in (2.9), $\breve{\kappa}_{\beta^*}(s)$ becoming increasingly smooth as $s \to \pm\infty$.

In view of (1.3), if we set $\beta^* = \beta_c$ in (2.6), the interval of integration reduces to the finite interval $[\tilde{a}, \tilde{b}]$, where $\tilde{a} := ka = kt_0$, $\tilde{b} := kb = kt_n$. Explicitly, (2.6) becomes

$$\phi(s) = \psi_{\beta_c}(s) + \mathrm{i}\int_{\tilde{a}}^{\tilde{b}} \kappa_{\beta_c}(s-t)(\beta(t/k)-\beta_c)\phi(t)\,\mathrm{d}t, \quad s \in \mathbb{R}, \tag{2.11}$$

with ψ_{β_c} and κ_{β_c} given by (2.7) and (2.8), respectively, with $\beta^* = \beta_c$. This integral equation is studied, in the case $\beta_c = 1$, in [16]. From [16, Theorem 4.17] it follows that

$$\|\phi\|_\infty \leq C_\epsilon\|\psi_1\|_\infty = C_\epsilon|1 + R_1(\theta)| \leq C_\epsilon\cos\theta. \tag{2.12}$$

As in [20], and as discussed in the introduction, our numerical scheme for solving (2.11) is based on a consideration of the contribution of the reflected and diffracted ray paths in the spirit of the GTD. In particular, to leading order as $k \to \infty$, on the interval (t_{j-1}, t_j) it seems reasonable to suppose that the total field $\phi \approx \psi_{\beta_j}$, the total field there would be if the whole boundary had the admittance β_j of the interval (t_{j-1}, t_j), given explicitly by (2.7) with $\beta^* = \beta_j$. In fact, for $s \neq \tilde{t}_j := kt_j$, $j = 0, \ldots, n$, it follows from theorem 2.3 below that $\phi(s) \to \Psi(s)$ as $k \to \infty$, where

$$\Psi(s) := \begin{cases} \psi_{\beta_j}(s), & s \in (\tilde{t}_{j-1}, \tilde{t}_j], \quad j = 1, \ldots, n, \\ \psi_{\beta_c}(s), & s \in \mathbb{R}\backslash(\tilde{t}_0, \tilde{t}_n]. \end{cases} \tag{2.13}$$

In our numerical scheme we compute the difference between ϕ and Ψ, i.e.,

$$\Phi(s) := \phi(s) - \Psi(s), \quad s \in \mathbb{R}, \tag{2.14}$$

which may be thought of as the correction to the leading order field due to scattering from impedance discontinuities. Clearly, from (2.11) we have that

$$\Phi = \Psi_\beta^{\beta_c} + K_\beta^{\beta_c}\Phi, \tag{2.15}$$

where $\Psi_\beta^{\beta_c} \in L_\infty(\mathbb{R})$ is given by $\Psi_\beta^{\beta_c} := \psi_{\beta_c} - \Psi + K_\beta^{\beta_c}\Psi$, and

$$K_\beta^{\beta_c}\chi(s) := \mathrm{i}\int_{\tilde{a}}^{\tilde{b}} \kappa_{\beta_c}(s-t)(\beta(t/k)-\beta_c)\chi(t)\,\mathrm{d}t.$$

Equation (2.15) will be the integral equation that we solve numerically. By setting $\beta^* = \beta_j$ in (2.6) we obtain explicit expressions for Φ on each subinterval, namely,

$$\Phi(s) = \mathrm{e}^{\mathrm{i}s} f_j^+(s-\tilde{t}_{j-1}) + \mathrm{e}^{-\mathrm{i}s} f_j^-(\tilde{t}_j - s), \quad s \in (\tilde{t}_{j-1}, \tilde{t}_j],\ j = 1, \dots, n, \tag{2.16}$$

where for $j = 1, \dots, n$, $f_j^+, f_j^- \in \mathcal{C}[0,\infty)$ are defined by

$$f_j^+(r) := \int_{-\infty}^{\tilde{t}_{j-1}} \check{\kappa}_{\beta_j}(r + \tilde{t}_{j-1} - t)\mathrm{e}^{-\mathrm{i}t}\mathrm{i}(\beta(t/k)-\beta_j)\phi(t)\,\mathrm{d}t, \tag{2.17}$$

$$f_j^-(r) := \int_{\tilde{t}_j}^{\infty} \check{\kappa}_{\beta_j}(t - \tilde{t}_j + r)\mathrm{e}^{\mathrm{i}t}\mathrm{i}(\beta(t/k)-\beta_j)\phi(t)\,\mathrm{d}t \tag{2.18}$$

with $\check{\kappa}_{\beta_j}$ given by (2.10) with $\beta^* = \beta_j$. Similarly, from (2.11),

$$\Phi(s) = \begin{cases} \mathrm{e}^{\mathrm{i}s} f_{n+1}^+(s - \tilde{t}_n), & s > \tilde{t}_n, \\ \mathrm{e}^{-\mathrm{i}s} f_0^-(\tilde{t}_0 - s), & s < \tilde{t}_0, \end{cases} \tag{2.19}$$

where f_{n+1}^+, f_0^- are given by (2.17), (2.18), respectively, with $\beta_0 := \beta_c$ and $\beta_{n+1} := \beta_c$.

The first term in (2.16) can be viewed as an explicit summation of all the diffracted rays scattered at the discontinuity in impedance at t_{j-1} which travel from left to right along (t_{j-1}, t_j). Similarly, the other term in (2.16) is the contribution to the diffracted field diffracted by the discontinuity at t_j. In the remainder of this section, so as to design an efficient discretisation for Φ, we investigate in detail the behavior of the integrals $f_j^{\pm}$. As a first step, we prove the following bounds on $|\check{\kappa}_{\beta^*}^{(m)}(s)|$, for $m = 0, 1, \dots$, $s \in (0, \infty)$, which were stated without proof in [20].

LEMMA 2.2. *Suppose that* $\mathrm{Re}\beta^* \geq \epsilon$, $|\beta^*| \leq \epsilon^{-1}$ *hold for some* $\epsilon > 0$. *Then, for* $m = 0, 1, \dots,$ *there exist constants* c_m, *dependent only on* m *and* ϵ, *such that*

$$|\check{\kappa}_{\beta^*}^{(m)}(s)| \leq \begin{cases} c_m(1 + |\log s|), & m = 0, \\ c_m s^{-m}, & m \geq 1, \end{cases} \qquad \textit{for } 0 < s \leq 1,$$

$$|\check{\kappa}_{\beta^*}^{(m)}(s)| \leq c_m s^{-\frac{3}{2}-m} \qquad \textit{for } s > 1.$$

Proof. Throughout the proof, c_m is a constant dependent only on m and ϵ, not necessarily the same at each occurrence. Let

$$F(z) := \frac{z^{1/2}(z-2\mathrm{i})^{1/2}}{z^2 - 2\mathrm{i}z - \beta^{*2}} = \frac{z^{1/2}(z-2\mathrm{i})^{1/2}}{(z - \mathrm{i}\hat{a}_+)(z - \mathrm{i}\hat{a}_-)}, \quad z \in \mathbb{C}, \tag{2.20}$$

where $\mathrm{Re}\, z^{1/2}, \mathrm{Re}\,(z-2\mathrm{i})^{1/2} \geq 0$, and $\hat{a}_\pm = 1 \mp \sqrt{1-\beta^{*2}}$, as before, with $\mathrm{Re}\sqrt{1-\beta^{*2}} \geq 0$. Then $F(z)$ has simple poles at $z = \mathrm{i}\hat{a}_+$ (which may lie near the real axis if $\mathrm{Re}\hat{a}_+$ is small) and $z = \mathrm{i}\hat{a}_-$ (which cannot lie near the real axis as $\mathrm{Re}\hat{a}_- \geq 1$). Recalling (2.10) we then have, at least provided $\mathrm{Re}\hat{a}_+ \neq 0$ or $\mathrm{Im}\hat{a}_+ > 0$, so that the pole at $\mathrm{i}\hat{a}_+$ does not lie on the positive real axis,

$$|\check{\kappa}^{(m)}_{\beta^*}(s)| \leq \frac{1}{\pi}\left|\int_0^\infty F(r) r^m \mathrm{e}^{-rs}\,\mathrm{d}r\right| + \left|C_{\beta^*}\hat{a}_+^m \mathrm{e}^{\mathrm{Im}\hat{a}_+ s}\right|, \quad s > 0. \tag{2.21}$$

Now, since $\mathrm{Re}\beta^* \geq \epsilon$, it is easy to see that $\mathrm{Im}\hat{a}_+ = 0$ if and only if $\beta^* \in [\epsilon, 1]$, and in this case $\mathrm{Re}\hat{a}_+ \geq 1 - \sqrt{1-\epsilon^2} = \epsilon^2/(1+\sqrt{1-\epsilon^2}) > \epsilon^2/2$. We thus define $S_\epsilon := \{\beta^* : \mathrm{Re}\beta^* \geq \epsilon, |\beta^*| \leq \epsilon^{-1}, \mathrm{Re}\hat{a}_+ \leq \epsilon^2/4\}$. Then S_ϵ is closed and bounded, and $|\mathrm{Im}\hat{a}_+|$ and $|\sqrt{1-\beta^{*2}}|$ are both continuous and nonzero on S_ϵ. Thus, for some $\eta > 0$,

$$|\mathrm{Im}\hat{a}_+| \geq \eta \quad \text{and} \quad \left|\sqrt{1-\beta^{*2}}\right| \geq \eta \tag{2.22}$$

for all $\beta^* \in S_\epsilon$.

Next, we note that if $\mathrm{Re}\hat{a}_+ > 0$, then $C_{\beta^*} = 0$, while if $\mathrm{Re}\hat{a}_+ \leq 0$, then $\beta^* \in S_\epsilon$, so that (2.22) holds. Moreover, if $C_{\beta^*} \neq 0$, then $\mathrm{Im}\beta^* < 0$, and so $\mathrm{Im}\hat{a}_+ < 0$. Since also $|\hat{a}_\pm| \leq 1 + \sqrt{1+\epsilon^{-2}}$, we see that

$$\left|C_{\beta^*}\hat{a}_+^m \mathrm{e}^{\mathrm{Im}\hat{a}_+ s}\right| \leq c_m \mathrm{e}^{-\eta s}, \quad s > 0.$$

We turn to bounding the first term on the right-hand side of (2.21). To do this we consider the two cases $|\mathrm{Re}\hat{a}_+| > \epsilon^2/4$ and $|\mathrm{Re}\hat{a}_+| \leq \epsilon^2/4$ separately.

First, suppose $|\mathrm{Re}\hat{a}_+| > \epsilon^2/4$. Then

$$|F(r)| \leq C_\epsilon r^{1/2}, \quad r > 0, \tag{2.23}$$

and thus

$$\left|\int_0^\infty F(r) r^m \mathrm{e}^{-rs}\,\mathrm{d}r\right| \leq C_\epsilon \int_0^\infty r^{m+1/2}\mathrm{e}^{-rs}\,\mathrm{d}r \leq c_m s^{-m-3/2}, \;\; s > 0. \tag{2.24}$$

This bound suffices when $s > 1$, but for $0 < s \leq 1$ we need a sharper bound.

We proceed by establishing bounds on the mth derivatives of the first two terms on the right-hand side of (2.8) for $0 < s \leq 1$. It can easily be deduced from the power series representations defining the Bessel functions that there exist constants C_j, $j = 0, \ldots$, such that, for $0 < z \leq 1$,

$$|H_0^{(1)}(z)| \leq C_0(1 + |\log z|), \tag{2.25}$$

$$\left|\frac{d^m}{dz^m} H_0^{(1)}(z)\right| \leq C_m z^{-m}, \quad m = 1, 2, \ldots. \tag{2.26}$$

Next note that, for $0 < s \leq 1$, the mth derivative of the second term in (2.8) has absolute value not more than

$$\left|\frac{\beta^{*2}}{\pi}\int_0^\infty \frac{(\mathrm{i}-t)^m \mathrm{e}^{-st} t^{-1/2}\,\mathrm{d}t}{(t-2\mathrm{i})^{1/2}(t^2 - 2\mathrm{i}t - \beta^{*2})}\right| \leq \frac{\epsilon^{-2}}{\pi}\int_0^\infty \frac{(1+t^2)^{m/2}\mathrm{e}^{-st}t^{-1/2}\,\mathrm{d}t}{(t^2+4)^{1/4}\,|(t-\mathrm{i}\hat{a}_+)(t-\mathrm{i}\hat{a}_-)|} \tag{2.27}$$

$$\leq C_\epsilon\left[\int_0^1 t^{-1/2}\,\mathrm{d}t + \int_1^\infty t^{m-1}\mathrm{e}^{-st}\,\mathrm{d}t\right] \tag{2.28}$$

$$\leq \begin{cases} C_\epsilon(1 - \log s), & m = 0, \\ c_m s^{-m}, & m = 1, 2, \ldots. \end{cases} \tag{2.29}$$

Combining (2.25), (2.26), and (2.29) and recalling (2.9) the result follows.

Now we consider the case $0 \leq \mathrm{Re}\hat{a}_+ \leq \epsilon^2/4$. (The proof for the case $-\epsilon^2/4 \leq \mathrm{Re}\hat{a}_+ < 0$ is similar.) As $\beta \in S_\epsilon$, (2.22) holds. If $\mathrm{Im}\hat{a}_+ > 0$, then the bounds (2.23) and (2.28) hold and we proceed as above. If $\mathrm{Im}\hat{a}_+ < 0$, however, $F(z)$ has a pole at $z = \mathrm{i}\hat{a}_+$ with $\mathrm{Re}(\mathrm{i}\hat{a}_+) > \eta$, $0 \leq \mathrm{Im}(\mathrm{i}\hat{a}_+) \leq \epsilon^2/4$. To bound the integrals on the left-hand side of (2.24) and (2.27) in this case, uniformly in β^*, we first deform the path of integration. Define Γ_ϵ to be the semicircle, center $(-\mathrm{Im}\hat{a}_+, 0)$, radius $\tilde{\eta} := \min(1/2, \eta)$, lying in the lower half plane. (Note that by (2.22), $\mathrm{Re}z > \eta/2$ for $z \in \Gamma_\epsilon$.) Let $\gamma_\epsilon = [0, -\mathrm{Im}\hat{a}_+ - \eta/2] \cup [-\mathrm{Im}\hat{a}_+ + \eta/2, \infty)$. Then, by Cauchy's theorem, it follows from (2.21) that, for $\mathrm{Re}\hat{a}_+ > 0$,

$$|\check{\kappa}_{\beta^*}^{(m)}(s)| \leq \frac{1}{\pi}\left|\int_{\gamma_\epsilon} F(r) r^m \mathrm{e}^{-rs}\,\mathrm{d}r + \int_{\Gamma_\epsilon} F(r) r^m \mathrm{e}^{-rs}\,\mathrm{d}r\right|, \quad s > 0. \tag{2.30}$$

By continuity arguments, taking the limit $\mathrm{Re}\hat{a}_+ \to 0^+$ in (2.30), equation (2.30) holds also for $\mathrm{Re}\hat{a}_+ = 0$. For $r \in \gamma_\epsilon$ the bound (2.23) holds, and so the integral over γ_ϵ is bounded by the right-hand side of (2.24). Further,

$$\left|\int_{\Gamma_\epsilon} F(r) r^m \mathrm{e}^{-rs}\,\mathrm{d}r\right| \leq \frac{\pi\eta}{2}\max_{r\in\Gamma_\epsilon}|F(r) r^m \mathrm{e}^{-rs}| \leq c_m \mathrm{e}^{-\eta s/2},$$

so we obtain the required bound for $s \geq 1$. To obtain the desired bound for $0 < s \leq 1$ we proceed as in the case $|\mathrm{Re}\hat{a}_+| > \epsilon^2/4$, but deforming the path of integration as above to bound the left-hand side of (2.27). □

The following result is a slight sharpening of [20, Theorem 2.6], obtained by combining the bounds in Lemma 2.2 and (2.12) with the representations (2.17) and (2.18).

THEOREM 2.3. *Suppose* (1.4) *holds for some* $\epsilon > 0$. *Then, for* $r > 0$, $j = 1, \ldots, n$, $m = 0, 1, \ldots$, *there exist constants* c_m, *dependent only on* m *and* ϵ, *such that*

$$\left|f_j^{\pm(m)}(r)\right| \leq c_m \cos\theta E_m(r),$$

where

$$E_m(r) = \begin{cases} 1, & m = 0, \\ 1 - \log r, & m = 1, \\ r^{1-m}, & m \geq 2, \end{cases} \qquad \text{for } 0 < r \leq 1,$$
$$E_m(r) = r^{-\frac{1}{2}-m} \qquad \text{for } r > 1.$$

Remark. Using the identical argument it can easily be shown that $|{f_{n+1}^{+}}^{(m)}(r)|$, $|{f_0^{-}}^{(m)}(r)| \leq c_m \cos\theta E_m(r)$, $r > 0$, for $m = 0, 1, \ldots$, where c_m is the same constant as in Theorem 2.3.

To prove the main result of this section, a sharper bound on $|f_j^{\pm}(r)|$ when $r > 1$ (Theorem 2.6), we require the bounds in the following two lemmas.

LEMMA 2.4. *Suppose* $p < -1$ *and* $q \leq 0$. *Then there exists a constant* C, *independent of* r, *such that, for* $r \geq 1$,

$$\int_0^\infty (t + r)^p (1 + t)^q\,\mathrm{d}t \leq \begin{cases} C r^{p+q+1}, & q \neq -1, \\ C r^p \log(1 + r), & q = -1. \end{cases}$$

Proof.

$$\int_0^\infty (t+r)^p(1+t)^q\,\mathrm{d}t \le r^p\int_0^r (1+t)^q\,\mathrm{d}t + r^q\int_r^\infty (t+r)^p\,\mathrm{d}t,$$

and the result follows. □

LEMMA 2.5. *Suppose $q \le 0$. Then there exists a constant C, independent of r and D, such that, for $r \ge 1$ and $D > 0$,*

$$\int_0^{2D} (s+r)^{-3/2}(1+2D-s)^q\,\mathrm{d}s \le \begin{cases} Cr^{-3/2}, & q < -1,\\ Cr^{-3/2}\log(1+r), & q = -1,\\ Cr^{-1/2+q}, & -1 < q \le 0. \end{cases}$$

Proof. Splitting the integration range as $[0, 2D] = [0, D] \cup [D, 2D]$, and making the change of variable $t := 2D - s$, we see that

$$\int_0^{2D} (s+r)^{-3/2}(1+2D-s)^q\,\mathrm{d}s \le I_A + I_B,$$

where

$$I_A := (1+D)^q\int_0^D (s+r)^{-3/2}\,\mathrm{d}s, \quad I_B := (D+r)^{-3/2}\int_0^D (1+t)^q\,\mathrm{d}t.$$

Further,

$$\begin{aligned} I_A &= \frac{2D(1+D)^q}{r^{1/2}(D+r)^{1/2}((D+r)^{1/2}+r^{1/2})}\\ &\le \frac{2(1+D)^{q+1}}{r^{1/2}(D+r)} \le \begin{cases} 2r^{-3/2}, & q \le -1,\\ 2r^{-1/2+q}, & -1 < q \le 0. \end{cases} \end{aligned} \tag{2.31}$$

For $q \ne -1$, $I_B = (1+q)^{-1}(D+r)^{-3/2}((1+D)^{q+1} - 1)$. Thus

$$|1+q|I_B \le \begin{cases} r^{-3/2}, & q < -1,\\ r^{-1/2+q}, & -1 < q \le 0. \end{cases}$$

To bound I_B in the case that $q = -1$ we need to consider the cases $r \ge D$ and $r < D$ separately. For $r \ge D$,

$$I_B = (D+r)^{-3/2}\log(1+D) \le r^{-3/2}\log(1+r).$$

For $r < D$ we split the range of integration as $[0, D] = [0, r] \cup [r, D]$ and note that

$$(D+r)^{-3/2}\int_0^r (1+t)^{-1}\,\mathrm{d}t \le r^{-3/2}\log(1+r),$$

$$(D+r)^{-3/2}\int_r^D (1+t)^{-1}\,\mathrm{d}t \le \frac{(D-r)}{(D+r)^{3/2}(1+r)} \le r^{-3/2}.$$

This completes the proof. □

We are now ready to prove the main result of this section, the following sharper bound on $|f_j^\pm(r)|$ when $r > 1$, on which the design of our numerical scheme is based.

THEOREM 2.6. *Suppose* (1.4) *holds for some $\epsilon > 0$. Then for $r > 1$, $j = 0, \dots, n$,*

$$\left|f_{j+1}^+(r)\right|, \left|f_j^-(r)\right| \le C_\epsilon \frac{r^{-3/2}n^3}{\cos\theta}.$$

Proof. First we consider $f_j^-(r)$. Recalling (2.14), for $j = 0,\dots,n$, $f_j^-(r) = I_1(r) + I_2(r)$, where

$$I_1(r) := \int_{\tilde{t}_j}^{\infty} \breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)\mathrm{e}^{\mathrm{i}t}\mathrm{i}(\beta(t/k) - \beta_j)\Psi(t)\,\mathrm{d}t,$$

$$I_2(r) := \int_{\tilde{t}_j}^{\infty} \breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)\mathrm{e}^{\mathrm{i}t}\mathrm{i}(\beta(t/k) - \beta_j)\Phi(t)\,\mathrm{d}t.$$

We begin by establishing a bound on I_1. Recalling (2.13) and (2.7),

$$I_1(r) = \sum_{m=j+1}^{n} \mathrm{i}(\beta_m - \beta_j) \int_{\tilde{t}_{m-1}}^{\tilde{t}_m} \breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)(1 + R_{\beta_m}(\theta))\mathrm{e}^{\mathrm{i}t(\sin\theta+1)}\,\mathrm{d}t$$
$$+ \mathrm{i}(\beta_c - \beta_j) \int_{\tilde{t}_n}^{\infty} \breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)(1 + R_{\beta_c}(\theta))\mathrm{e}^{\mathrm{i}t(\sin\theta+1)}\,\mathrm{d}t.$$

Integrating by parts,

$$I_1(r) = \sum_{m=j+1}^{n} \frac{(\beta_m - \beta_j)(1 + R_{\beta_m}(\theta))}{\sin\theta + 1} \left(\left[\breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)\mathrm{e}^{\mathrm{i}t(\sin\theta+1)}\right]_{\tilde{t}_{m-1}}^{\tilde{t}_m} \right.$$
$$\left. - \int_{\tilde{t}_{m-1}}^{\tilde{t}_m} \breve{\kappa}'_{\beta_j}(t - \tilde{t}_j + r)\mathrm{e}^{\mathrm{i}t(\sin\theta+1)}\,\mathrm{d}t \right)$$
$$+ \frac{(\beta_c - \beta_j)(1 + R_{\beta_c}(\theta))}{\sin\theta + 1} \left(\left[\breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)\mathrm{e}^{\mathrm{i}t(\sin\theta+1)}\right]_{\tilde{t}_n}^{\infty} \right.$$
$$\left. - \int_{\tilde{t}_n}^{\infty} \breve{\kappa}'_{\beta_j}(t - \tilde{t}_j + r)\mathrm{e}^{\mathrm{i}t(\sin\theta+1)}\,\mathrm{d}t \right).$$

Now from Lemma 2.2, for $r > 1$,

$$\left|\breve{\kappa}_{\beta_j}(\tilde{t}_m - \tilde{t}_j + r)\right| \le C_\epsilon(\tilde{t}_m - \tilde{t}_j + r)^{-3/2} \le C_\epsilon r^{-3/2}, \quad m = j,\dots,n.$$

Thus, noting that $|1 + R_{\beta_m}(\theta)| = |2\cos\theta/(\cos\theta + \beta_m)| \le 2\cos\theta/\epsilon$ and $|\beta_m - \beta_j| \le 2/\epsilon$, and using Lemma 2.2 again to bound $\breve{\kappa}'_{\beta_j}$, we have, for $r > 1$,

$$|I_1(r)| \le C_\epsilon \frac{(n+1-j)\cos\theta}{\sin\theta + 1} \left[r^{-3/2} + \int_{\tilde{t}_j}^{\infty} \left|\breve{\kappa}'_{\beta_j}(t - \tilde{t}_j + r)\right| \mathrm{d}t \right]$$

$$\le C_\epsilon \frac{r^{-3/2} n \cos\theta}{\sin\theta + 1}. \tag{2.32}$$

We next bound I_2. Recalling (2.16) and (2.19),

$$|I_2(r)| \le \frac{2}{\epsilon} \left(J_\infty^+ + \sum_{m=j+1}^{n} (J_m^+ + J_m^-) \right), \tag{2.33}$$

where

$$J_m^+ := \int_{\tilde{t}_{m-1}}^{\tilde{t}_m} |\breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)||f_m^+(t - \tilde{t}_{m-1})|\,\mathrm{d}t, \tag{2.34}$$

$$J_m^- := \int_{\tilde{t}_{m-1}}^{\tilde{t}_m} |\breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)||f_m^-(\tilde{t}_m - t)|\,\mathrm{d}t, \tag{2.35}$$

$$J_\infty^+ := \int_{\tilde{t}_n}^{\infty} |\breve{\kappa}_{\beta_j}(t - \tilde{t}_j + r)||f_{n+1}^+(t - \tilde{t}_n)|\,\mathrm{d}t. \tag{2.36}$$

First we bound J_∞^+. Applying Lemma 2.2 and Theorem 2.3, and noting the remark after Theorem 2.3, for $r > 1$,

$$J_\infty^+ \leq C_\epsilon \cos\theta \int_{\tilde{t}_n}^{\infty} (t - \tilde{t}_n + r)^{-3/2}(1 + t - \tilde{t}_n)^{-1/2}\, \mathrm{d}t \leq C_\epsilon r^{-1}\cos\theta,$$

making the change of variables $s = t - \tilde{t}_n$ and using Lemma 2.4 with $p = -3/2$ and $q = -1/2$. Arguing similarly, $J_m^+ \leq C_\epsilon r^{-1}\cos\theta$. To bound J_m^-, again using Lemma 2.2 and Theorem 2.3 we have

$$J_m^- \leq C_\epsilon \cos\theta \int_{\tilde{t}_{m-1}}^{\tilde{t}_m} (t - \tilde{t}_{m-1} + r)^{-3/2}(1 + \tilde{t}_m - t)^{-1/2}\, \mathrm{d}t \leq C_\epsilon r^{-1}\cos\theta,$$

making the change of variables $s = t - \tilde{t}_{m-1}$ and using Lemma 2.5 with $D = (\tilde{t}_m - \tilde{t}_{m-1})/2$ and $q = -1/2$. Thus, recalling (2.33), $|I_2(r)| \leq C_\epsilon r^{-1} n\cos\theta$.

So far in the argument we have shown that, for $r > 1$, $j = 0, \ldots, n$,

$$|f_j^-(r)| \leq |I_1(r)| + |I_2(r)| \leq C_\epsilon r^{-1} n\cos\theta\left(\frac{r^{-1/2}}{1+\sin\theta} + 1\right). \tag{2.37}$$

Proceeding in a similar way, we can show that, for $r > 1$, $j = 1, \ldots, n+1$,

$$|f_j^+(r)| \leq C_\epsilon r^{-1} n\cos\theta\left(\frac{r^{-1/2}}{1-\sin\theta} + 1\right). \tag{2.38}$$

Next, starting from (2.33)–(2.36), we can use (2.37) and (2.38) to establish sharper bounds on I_2 and hence a sharper bound on f_j^-. Using (2.38) in (2.36), we have for $r > 1$ that

$$\begin{aligned} J_\infty^+ &\leq C_\epsilon n\cos\theta \int_{\tilde{t}_n}^{\infty} (t - \tilde{t}_n + r)^{-3/2}(1 + t - \tilde{t}_n)^{-1}\left(1 + \frac{(1 + t - \tilde{t}_n)^{-1/2}}{1 - \sin\theta}\right) \mathrm{d}t \\ &\leq C_\epsilon r^{-3/2} n\cos\theta(\log(1+r) + (1-\sin\theta)^{-1}), \end{aligned} \tag{2.39}$$

making the change of variable $s = t - \tilde{t}_n$ and using Lemma 2.4 with $p = -3/2$ and $q = -1, -3/2$. Arguing similarly, we can show that

$$J_m^+ \leq C_\epsilon r^{-3/2} n\cos\theta(\log(1+r) + (1-\sin\theta)^{-1}), \quad m = j+1, \ldots, n, \tag{2.40}$$

and, using (2.37) and Lemma 2.2,

$$\begin{aligned} J_m^- &\leq C_\epsilon n\cos\theta \int_{\tilde{t}_{m-1}}^{\tilde{t}_m} (t - \tilde{t}_{m-1} + r)^{-3/2}(1 + \tilde{t}_m - t)^{-1}\left(1 + \frac{(1 + \tilde{t}_m - t)^{-1/2}}{1 + \sin\theta}\right) \mathrm{d}t \\ &\leq C_\epsilon r^{-3/2} n\cos\theta(\log(1+r) + (1+\sin\theta)^{-1}), \end{aligned}$$

where again we make the change of variable $s = t - \tilde{t}_{m-1}$ and use Lemma 2.5 with $D := (\tilde{t}_m - \tilde{t}_{m-1})/2$ and $q = -1, -3/2$. Combining this with (2.39) and (2.40),

$$|I_2(r)| \leq C_\epsilon r^{-3/2} n^2 \cos\theta(\log(1+r) + (1-\sin^2\theta)^{-1}) \leq C_\epsilon \frac{r^{-3/2}\log(1+r)n^2}{\cos\theta}.$$

Thus

$$|f_j^-(r)| \leq |I_1| + |I_2| \leq C_\epsilon \frac{r^{-3/2}\log(1+r)n^2}{\cos\theta}. \tag{2.41}$$

In a similar way it can be proved that

$$|f_j^+(r)| \leq C_\epsilon \frac{r^{-3/2}\log(1+r)n^2}{\cos\theta}. \tag{2.42}$$

To obtain sharper bounds still on $f_j^\pm$, removing the dependence on $\log r$ in (2.41), (2.42), we note that it follows from (2.41) and (2.42) that

$$\int_0^\infty |f_j^\pm(r)|\,\mathrm{d}r \leq C_\epsilon \frac{n^2}{\cos\theta}.$$

Using this bound and the bounds in Lemma 2.2 in (2.34)–(2.36), we see that

$$J_m^\pm \leq C_\epsilon r^{-3/2}\int_0^\infty |f_m^\pm(s)|\,\mathrm{d}s \leq C_\epsilon \frac{r^{-3/2}n^2}{\cos\theta},$$

and an identical bound holds on J_∞^+. Hence, recalling (2.33),

$$|I_2(r)| \leq C_\epsilon \frac{r^{-3/2}n^3}{\cos\theta},$$

and combining this with (2.32) the desired bound on $f_j^-(r)$ follows. The desired bound on $f_j^+(r)$ follows similarly. □

3. Galerkin method and error analysis. Our aim now is to design a numerical method for the solution of (2.15), supported by a full error analysis, for which the error bounds are independent of the parameter $k(b-a)$. To achieve this we will work in $L_2(\mathbb{R})$, and to that end we introduce the operator $Q : L_\infty(\mathbb{R}) \to L_2(\mathbb{R})$ defined by

$$Q\chi(s) := \begin{cases} \chi(s), & s \in [\tilde{a},\tilde{b}] = [\tilde{t}_0,\tilde{t}_n], \\ 0, & s \in \mathbb{R}\backslash[\tilde{a},\tilde{b}]. \end{cases}$$

Writing $\Phi^* := Q\Phi$, and noting that $K_\beta^{\beta_c}\Phi = K_\beta^{\beta_c}\Phi^*$, it follows from (2.15) that

$$\Phi^* - QK_\beta^{\beta_c}\Phi^* = Q\Psi_\beta^{\beta_c}, \tag{3.1}$$

where Φ^* and $Q\Psi_\beta^{\beta_c}$ are both in $L_2(\mathbb{R})$.

Existence and boundedness of $(I - QK_\beta^{\beta_c})^{-1} : L_2(\mathbb{R}) \to L_2(\mathbb{R})$ are shown in [20], where it is also shown that the unique solution $\Phi^* = (I - QK_\beta^{\beta_c})^{-1}Q\Psi_\beta^{\beta_c}$ of (3.1) satisfies $\|\Phi^*\|_2 \leq C_1\|Q\Psi_\beta^{\beta_c}\|_2$ with $C_1 = \mathrm{Re}\beta_c/(\mathrm{Re}\beta_c - \|\beta-\beta_c\|_\infty)$ if

$$|\beta_j - \beta_c| < \mathrm{Re}\beta_c, \quad j = 1,\ldots,n, \tag{3.2}$$

and C_1 unspecified but dependent only on ϵ and β_c if (3.2) does not hold.

To approximate the solution $\Phi^* = Q\Phi$ of (3.1) we use a Galerkin method, similar to that in [20], but with the approximation space chosen in a different way so as to take advantage of our stronger bound on Φ (Theorem 2.6), in order to remove the dependence of the error estimates on $k(b-a)$. As in [20], on each interval $(\tilde{t}_{j-1},\tilde{t}_j)$, we approximate $f_j^+(s-\tilde{t}_{j-1})$ and $f_j^-(\tilde{t}_j - s)$ in (2.16) by conventional piecewise polynomial approximations, rather than approximating Φ itself. This makes sense since, as quantified by Theorems 2.3 and 2.6, the functions $f_j^+(s-\tilde{t}_{j-1})$ and $f_j^-(\tilde{t}_j - s)$ are smooth (their higher order derivatives are small) away from $\tilde{t}_{j-1}$ and $\tilde{t}_j$, respectively. To approximate $f_j^+(s-\tilde{t}_{j-1})$ and $f_j^-(\tilde{t}_j - s)$ we use piecewise polynomials of a fixed degree $\nu \geq 0$ on a graded mesh, the mesh grading adapted in an optimal way to the bounds on $f_j^{\pm(m)}$ in Theorems 2.3 and 2.6.

To begin, we define a graded mesh on a general interval $[0, A]$, for $A > 1$, with more mesh points near 0 and less near A. This mesh is identical to that defined in [20, Definition 3.1]; the difference here is in how we choose the value of A when we apply this mesh to the discretization of each interval $[\tilde{t}_{j-1}, \tilde{t}_j]$. Whereas in [20], A was chosen as a function of $\tilde{t}_j - \tilde{t}_{j-1}$ and the functions $f_j^\pm$ were approximated over the whole interval $[\tilde{t}_{j-1}, \tilde{t}_j]$, here we choose A as a function of N, a positive integer, where the size of N also determines the density of the mesh on $[0, A]$. A judicious choice of $A = A(N)$, as described below, allows us to discretize only a subsection of the interval $[\tilde{t}_{j-1}, \tilde{t}_j]$, near to $\tilde{t}_{j-1}$ and $\tilde{t}_j$, and to approximate $f_j^\pm$ by zero away from these points without harming the overall accuracy of our scheme. This is the key to achieving error estimates independent of $k(b-a)$.

The mesh we use also has similarities to that used in [40] for solving (1.1) in the case $k = \mathrm{i}\tau$, $\tau > 0$, τ large, where a similar idea of only discretizing a subsection of the boundary as $k \to \infty$ was used to establish error bounds independent of τ.

Definition 3.1. *For $A > 1$ and $N = 2, 3, \ldots$, the mesh $\Lambda_{N,A} = \{y_0, \ldots, y_{N+N_A}\}$ consists of the points $y_i = (i/N)^q$, $i = 0, \ldots, N$, where $q = 1 + 2\nu/3$, together with the points $y_{N+j} = A^{j/N_A}$, $j = 1, \ldots, N_A$, where $N_A = \lceil N^* \rceil$, the smallest integer $\geq N^*$, and $N^* := -\log A/[q \log(1 - 1/N)]$.*

The mesh $\Lambda_{N,A}$ is a composite mesh with a polynomial grading on $[0, 1]$ and a geometric grading on $[1, A]$. The definition of N_A ensures a smooth transition between the two parts of the mesh. Precisely, the definition of N^* is such that, in the case $N_A = N^*$, it holds that $y_{N+1}/y_N = y_N/y_{N-1}$, so that y_{N-1} and y_N are points in both the polynomial and the geometric parts of the mesh. It is shown in [20] that the total number of subintervals $N + N_A$ of the mesh on $[0, A]$ satisfies

$$N + N_A < \left(\frac{3}{2} + \frac{\log A}{q}\right) N. \tag{3.3}$$

Let $\Pi_{A,N,\nu} := \{\sigma : \sigma|_{[y_{j-1}, y_j]}$ is a polynomial of degree $\leq \nu$, $j = 1, \ldots, N + N_A\}$, and let P_N^* be the orthogonal projection operator from $L_2(0, A)$ to $\Pi_{A,N,\nu}$, so that setting $p = P_N^* f$ minimizes $\|f - p\|_{2,(0,A)} = \{\int_0^A |f(t) - p(t)|^2 \,\mathrm{d}t\}^{1/2}$ over all $p \in \Pi_{A,N,\nu}$. The mesh $\Lambda_{N,A}$ is designed to approximately minimize $\|f - P_N^* f\|_{2,(0,A)}$, over all meshes with the same number of points, when $f \in \mathcal{C}^\infty(0, \infty)$ with $|f^{(\nu+1)}(s)| = E_{\nu+1}(s)$, $s > 0$, where $E_{\nu+1}$ is defined as in Theorem 2.3. It achieves this by ensuring that $\|f - P_N^* f\|_{2,(y_{j-1}, y_j)}$ is approximately constant for $j = 1, \ldots, N + N_A$, i.e., by equidistributing the approximation error over the intervals of the mesh, as shown in the proof of the following result in [20].

Theorem 3.2. *Suppose that $f \in \mathcal{C}^\infty(0, \infty)$ and $|f'(s)| \leq E_1(s)$, $|f^{(\nu+1)}(s)| \leq E_{\nu+1}(s)$, $s > 0$. Then*

$$\|f - P_N^* f\|_{2,(0,A)} \leq C_\nu \frac{1 + \log^{1/2} A}{N^{\nu+1}}.$$

To form our approximation space on $[\tilde{a}, \tilde{b}] = [\tilde{t}_0, \tilde{t}_n]$, we begin by defining

$$A_j := \min \left\{ \alpha \frac{n^3 N^{\nu+1}}{\cos\theta}, \tilde{t}_j - \tilde{t}_{j-1} \right\}, \tag{3.4}$$

where $\alpha \geq 1$ is an absolute constant which will be determined experimentally and whose value will not effect the asymptotic convergence rates. The reason for our choice of A_j will become apparent shortly, in the proof of Theorem 3.3. Clearly A_j is bounded independently of $k(b-a)$. As we are primarily concerned with the high-frequency problem, we assume for simplicity that $A_j \geq 1$, $j = 1, \ldots, n$, but remark that in the case $A_j < 1$ for any value of j then we can define Λ_{N,A_j} to be an appropriate subset of the points y_i, and this will give similar approximation properties to those achieved using Λ_{N,A_j} when $A_j \geq 1$. For $j = 1, \ldots, n$ we define the two meshes $\Omega_j^+ := \tilde{t}_{j-1} + \Lambda_{N,A_j}$, $\Omega_j^- := \tilde{t}_j - \Lambda_{N,A_j}$. Letting $e_\pm(s) := \mathrm{e}^{\pm \mathrm{i}s}$, $s \in \mathbb{R}$, we then define $V_{\Omega_j^+,\nu} := \{\sigma e_+ : \sigma \in \Pi_{\Omega_j^+,\nu}\}$, $V_{\Omega_j^-,\nu} := \{\sigma e_- : \sigma \in \Pi_{\Omega_j^-,\nu}\}$, for $j = 1, \ldots, n$, where

$$\begin{aligned}
\Pi_{\Omega_j^+,\nu} := \{\sigma \in L_2(\mathbb{R}) : \sigma|_{(\tilde{t}_{j-1}+y_{m-1}, \tilde{t}_{j-1}+y_m)} &\text{ is a polynomial of degree} \leq \nu, \text{ for} \\
m = 1, \ldots, N + N_{A_j}, \text{ and } \sigma|_{\mathbb{R}\setminus[\tilde{t}_{j-1}, \tilde{t}_{j-1}+A_j]} &= 0\}, \\
\Pi_{\Omega_j^-,\nu} := \{\sigma \in L_2(\mathbb{R}) : \sigma|_{(\tilde{t}_j - y_m, \tilde{t}_j - y_{m-1})} &\text{ is a polynomial of degree} \leq \nu, \text{ for} \\
m = 1, \ldots, N + N_{A_j}, \text{ and } \sigma|_{\mathbb{R}\setminus[\tilde{t}_j - A_j, \tilde{t}_j]} &= 0\},
\end{aligned}$$

and $y_0, \ldots, y_{N_{A_j}}$ are the points of the mesh Λ_{N,A_j}. Our approximation space is then $V_{\Omega,\nu}$, the linear span of $\bigcup_{j=1,\ldots,n}\{V_{\Omega_j^+,\nu} \cup V_{\Omega_j^-,\nu}\}$.

Let $(\cdot,\cdot)$ denote the usual inner product on $L_2(\mathbb{R})$, $(\chi_1, \chi_2) := \int_{-\infty}^{\infty} \chi_1(s)\overline{\chi_2}(s)\,\mathrm{d}s$, $\chi_1, \chi_2 \in L_2(\mathbb{R})$. Then our Galerkin method approximation, $\Phi_N \in V_{\Omega,\nu}$, is defined by

$$(\Phi_N, \rho) = (\Psi_\beta^{\beta_c}, \rho) + (K_\beta^{\beta_c}\Phi_N, \rho) \quad \text{for all } \rho \in V_{\Omega,\nu}; \tag{3.5}$$

equivalently,

$$\Phi_N - P_N K_\beta^{\beta_c}\Phi_N = P_N Q \Psi_\beta^{\beta_c}, \tag{3.6}$$

where $P_N : L_2(\mathbb{R}) \to V_{\Omega,\nu}$ is the operator of orthogonal projection onto $V_{\Omega,\nu}$. Equation (3.5) can be written explicitly as a system of M_N linear algebraic equations, where M_N, the dimension of $V_{\Omega,\nu}$, i.e., the number of degrees of freedom, is given by

$$M_N = 2(\nu+1)\sum_{j=1}^{n}(N + N_{A_j}). \tag{3.7}$$

By (3.3) and (3.4), where $\bar{A} := (A_1 \ldots A_n)^{1/n} \leq (A_1 + \cdots + A_n)/n$,

$$M_N < (\nu+1)Nn\left[3 + \frac{2\log\bar{A}}{q}\right] \leq (\nu+1)Nn\left[3 + \frac{2}{q}\log\min\left(\frac{\alpha n^3 N^{\nu+1}}{\cos\theta}, \frac{k(b-a)}{n}\right)\right].$$

Using an argument similar to that for the Galerkin method in [20], it can be shown that, provided (3.2) holds, (3.6) is uniquely solvable and

$$\|(I - P_N K_\beta^{\beta_c})^{-1}\| \leq \frac{\mathrm{Re}\beta_c}{\mathrm{Re}\beta_c - \|\beta - \beta_c\|_\infty}, \tag{3.8}$$

and thus

$$\|\Phi^* - \Phi_N\|_2 \leq \frac{\mathrm{Re}\beta_c}{\mathrm{Re}\beta_c - \|\beta - \beta_c\|_\infty}\|\Phi^* - P_N\Phi^*\|_2. \tag{3.9}$$

There is also a description in [20] of how one can perturb the original problem in such a way that the condition (3.2) on β is forced to hold, and the solution of the perturbed problem is arbitrarily close in an arbitrarily large bounded region to the solution of the original problem. In any case, numerical results in [38] suggest that the Galerkin scheme we propose is stable and convergent even when (3.2) does not hold. In this case the bound (3.9) does not apply, however.

It remains to bound $\|\Phi^* - P_N\Phi^*\|_2$, showing that our approximation space is well adapted to approximate Φ^*. We introduce P_N^+ and P_N^-, the orthogonal projection operators from $L_2(\mathbb{R})$ onto $\Pi_{\Omega^+,\nu}$ and $\Pi_{\Omega^-,\nu}$, respectively, where $\Pi_{\Omega^\pm,\nu}$ denotes the linear span of $\bigcup_{j=1,\dots,n}\Pi_{\Omega_j^\pm,\nu}$. We also define

$$f_+(s) := \begin{cases} f_j^+(s-\tilde{t}_{j-1}), & s\in(\tilde{t}_{j-1},\tilde{t}_j],\ j=1,\dots,n,\\ 0, & s\in\mathbb{R}\backslash(\tilde{t}_0,\tilde{t}_n],\end{cases}$$
$$f_-(s) := \begin{cases} f_j^-(\tilde{t}_j - s), & s\in(\tilde{t}_{j-1},\tilde{t}_j],\ j=1,\dots,n,\\ 0, & s\in\mathbb{R}\backslash(\tilde{t}_0,\tilde{t}_n].\end{cases}$$

Then we have the following error estimate.

THEOREM 3.3. *If* (1.4) *holds for some* $\epsilon > 0$, *then*

$$\|f_+ - P_N^+ f_+\|_2 \le C_{\epsilon,\nu}\frac{n^{1/2}}{N^{\nu+1}}\left(1+\log^{1/2}\left(\min\left(\alpha\frac{n^3N^{\nu+1}}{\cos\theta}, k(b-a)\right)\right)\right),$$

where α *is the constant in* (3.4), *and the identical bound holds on* $\|f_- - P_N^- f_-\|_2$.

Proof. We prove the result for $\|f_+ - P_N^+ f_+\|_2$, the bound on $\|f_- - P_N^- f_-\|_2$ can be proved in a similar way. Recalling (3.4),

$$\begin{aligned}\|f_+ - P_N^+ f_+\|_2^2 &= \|f_+ - P_N^+ f_+\|_{2,(\tilde{a},\tilde{b})}^2\\ &= \sum_{j=1}^n\left[\|f_+ - P_N^+ f_+\|_{2,(\tilde{t}_{j-1},\tilde{t}_{j-1}+A_j)}^2 + \|f_+ - P_N^+ f_+\|_{2,(\tilde{t}_{j-1}+A_j,\tilde{t}_j)}^2\right].\end{aligned}$$

Now, by Theorems 2.3 and 3.2,

$$\|f_+ - P_N^+ f_+\|_{2,(\tilde{t}_{j-1},\tilde{t}_{j-1}+A_j)} \le C_{\epsilon,\nu}\cos\theta\frac{1+\log^{1/2}A_j}{N^{\nu+1}}.$$

If $\alpha n^3N^{\nu+1}/\cos\theta \ge \tilde{t}_j - \tilde{t}_{j-1}$, then $A_j = \tilde{t}_j - \tilde{t}_{j-1}$, in which case

$$\|f_+ - P_N^+ f_+\|_{2,(\tilde{t}_{j-1}+A_j,\tilde{t}_j)} = 0.$$

If $\alpha n^3N^{\nu+1}/\cos\theta < \tilde{t}_j - \tilde{t}_{j-1}$, then $A_j = \alpha n^3N^{\nu+1}/\cos\theta$, and then, recalling the definition of $\Pi_{\Omega^+,\nu}$ and Theorem 2.6,

$$\begin{aligned}\|f_+ - P_N^+ f_+\|_{2,(\tilde{t}_{j-1}+A_j,\tilde{t}_j)}^2 = \|f_+\|_{2,(\tilde{t}_{j-1}+A_j,\tilde{t}_j)}^2 &\le C_\epsilon\frac{n^6}{\cos^2\theta}\int_{A_j}^\infty s^{-3}\,\mathrm{d}s\\ &= C_\epsilon\frac{n^6}{2\cos^2\theta}A_j^{-2} = \frac{C_\epsilon}{2\alpha^2}N^{-2(\nu+1)},\end{aligned}$$

and recalling that $\alpha \ge 1$ the result follows. □

To use the above error estimate, note from (2.16) that $\Phi^* = e_+ f_+ + e_- f_-$. But $e_+ P_N^+ f_+ + e_- P_N^- f_- \in V_{\Omega,\nu}$, and $P_N \Phi^*$ is the best approximation to Φ^* in $V_{\Omega,\nu}$. So

$$\begin{aligned}\|\Phi^* - P_N\Phi^*\|_2 &\le \|\Phi^* - (e_+ P_N^+ f_+ + e_- P_N^- f_-)\|_2 \\ &= \|e_+(f_+ - P_N^+ f_+) + e_-(f_- - P_N^- f_-)\|_2 \\ &\le \|e_+\|_\infty \|f_+ - P_N^+ f_+\|_2 + \|e_-\|_\infty \|f_- - P_N^- f_-\|_2.\end{aligned}$$

Applying Theorem 3.3 we obtain the following result.

Theorem 3.4. *If* (1.4) *holds for some* $\epsilon > 0$*, then*

$$\|\Phi^* - P_N\Phi^*\|_2 \le C_{\epsilon,\nu} \frac{n^{1/2}}{N^{\nu+1}} \left(1 + \log^{1/2}\left(\min\left(\alpha \frac{n^3 N^{\nu+1}}{\cos\theta}, k(b-a)\right)\right)\right),$$

where α *is the constant in* (3.4).

Combining this result with the stability bound (3.9) we obtain our final error estimate for the approximation of Φ by Φ_N.

Theorem 3.5. *If* (1.4) *holds for some* $\epsilon > 0$*, and* (3.2) *is satisfied, then*

$$\|\Phi - \Phi_N\|_{2,(\tilde{a},\tilde{b})} = \|\Phi^* - \Phi_N\|_2 \le \frac{C_{\epsilon,\nu} n^{1/2}(1 + \log^{1/2}(\min(\alpha n^3 N^{\nu+1}/\cos\theta, k(b-a))))}{(\mathrm{Re}\beta_c - \|\beta - \beta_c\|_\infty) N^{\nu+1}},$$

where α *is the constant in* (3.4)*. Further, the number of degrees of freedom* M_N *satisfies*

$$M_N \le C_\nu N n \left[1 + \log\min\left(\frac{\alpha n^3 N^{\nu+1}}{\cos\theta}, \frac{k(b-a)}{n}\right)\right].$$

We finish by considering the computation of an approximation to u^t throughout the upper half plane U, once the Galerkin solution Φ_N has been computed. Recalling (2.13) and (2.14) we define $\phi_N \in L_2(\tilde{a}, \tilde{b})$, an approximation to ϕ on $(\tilde{a}, \tilde{b})$, by

$$\phi_N(s) := \Phi_N(s) + \psi_{\beta_j}(s), \quad s \in (\tilde{t}_{j-1}, \tilde{t}_j],\ j = 1, \dots, n,$$

where ψ_{β_j} is given explicitly by (2.7). Then, recalling that $u^t((y_1, 0)) = \phi(ky_1)$, we define an approximation to u^t by replacing $u^t(y)$ by its approximation $\phi_N(ky_1)$ in (2.5), to give the approximation u_N^t defined by

$$u_N^t(x) := u_{\beta_c}^t(x) + \mathrm{i}k \int_a^b G_{\beta_c}(x, (y_1, 0))(\beta(y_1) - \beta_c)\phi_N(ky_1)\,\mathrm{d}y_1. \tag{3.10}$$

From (2.2) and (2.3), and using properties of standard single-layer potentials [23], it follows that $u_N^t \in C^2(U) \cap C(\overline{U})$ and satisfies the Helmholtz equation (1.1) in U. Further, from Theorem 3.5 we deduce the following error estimate.

Theorem 3.6. *If* (1.4) *holds for some* $\epsilon > 0$*, and* (3.2) *is satisfied, then*

$$|u^t(x) - u_N^t(x)| \le \frac{C_{\epsilon,\nu} n^{1/2}(1 + \log^{1/2}(\min(\alpha n^3 N^{\nu+1}/\cos\theta, k(b-a))))}{(\mathrm{Re}\beta_c - \|\beta - \beta_c\|_\infty) N^{\nu+1}}$$

for $x \in \overline{U}$*, where* α *is the constant in* (3.4).

Proof. Subtracting (3.10) from (2.5) and using the Cauchy–Schwarz inequality and the definitions of Φ^* and ϕ_N, we see that

$$\begin{aligned}|u^t(x) - u^t_N(x)| &= \left| \int_{\tilde{a}}^{\tilde{b}} G_{\beta_c}(x, (t/k, 0))(\beta(t/k) - \beta_c)(\Phi(t) - \Phi_N(t))\,\mathrm{d}t \right| \\ &\le \|\beta - \beta_c\|_\infty \left\{ \int_{-\infty}^{\infty} |G_{\beta_c}(x, (t/k, 0))|^2 \,\mathrm{d}t \right\}^{1/2} \|\Phi - \Phi_N\|_{2,(\tilde{a},\tilde{b})}.\end{aligned}$$

Now, defining $H = kx_2$ and using (2.2) we see that for $H \ge 1/2$ it holds that

$$\begin{aligned}\int_{-\infty}^{\infty} |G_{\beta_c}(x, (t/k, 0))|^2 \,\mathrm{d}t &\le C_\epsilon (1 + H)^2 \int_{-\infty}^{\infty} \frac{\mathrm{d}t}{(t^2 + H^2)^{3/2}} \\ &= 2C_\epsilon \frac{(1 + H)^2}{H^2} \int_0^\infty \frac{\mathrm{d}s}{(1 + s^2)^{3/2}} \le C_\epsilon \int_0^\infty \frac{\mathrm{d}s}{(1 + s^2)^{3/2}}.\end{aligned}$$

Using (2.2) and (2.3) we see that, for $0 \le H < 1/2$,

$$\begin{aligned}\int_{-\infty}^{\infty} |G_{\beta_c}(x, (t/k, 0))|^2 \,\mathrm{d}t &\le C_\epsilon \left(\int_{\sqrt{1-H^2}}^{\infty} \frac{(1 + H)^2 \mathrm{d}t}{(t^2 + H^2)^{3/2}} + \int_0^{\sqrt{1-H^2}} \left(1 - \frac{1}{2}\log(t^2 + H^2)\right) \mathrm{d}t \right) \\ &\le C_\epsilon \left(\frac{9}{4} \int_{\sqrt{3}/2}^{\infty} \frac{\mathrm{d}t}{t^3} + \int_0^1 (1 - \log t)\,\mathrm{d}t \right).\end{aligned}$$

Thus $|u^t(x) - u^t_N(x)| \le C_\epsilon \|\Phi - \Phi_N\|_{2,(\tilde{a},\tilde{b})}$, and the result follows from Theorem 3.5. □

4. Implementation and numerical results.

We restrict our attention in this section to the case $\nu = 0$. The implementation of the scheme is similar for higher values of ν. Recalling (3.5), the equation we wish to solve is

$$(\Phi_N, \rho) - (K^{\beta_c}_\beta \Phi_N, \rho) = (\Psi^{\beta_c}_\beta, \rho) \quad \text{for all } \rho \in V_{\Omega,0}. \tag{4.1}$$

Writing Φ_N as a linear combination of basis functions of $V_{\Omega,0}$, we have $\Phi_N(s) = \sum_{j=1}^{M_N} v_j \rho_j(s)$, where M_N is given by (3.7) and ρ_j is the jth basis function, defined by

$$\begin{aligned}\rho_j(s) &:= \frac{\mathrm{e}^{\mathrm{i}s}\chi_{[s^+_{\tilde{j}}, s^+_{\tilde{j}-1})}(s)}{(s^+_{\tilde{j}} - s^+_{\tilde{j}-1})^{1/2}}, \quad j = \tilde{j} + 2\sum_{m=1}^{p-1}(N + N_{A_m}),\ \tilde{j} = 1, \dots, N + N_{A_p}, \\ \rho_j(s) &:= \frac{\mathrm{e}^{-\mathrm{i}s}\chi_{[s^-_{\tilde{j}}, s^-_{\tilde{j}-1})}(s)}{(s^-_{\tilde{j}} - s^-_{\tilde{j}-1})^{1/2}}, \quad j = \tilde{j} + N + N_{A_p} + 2\sum_{m=1}^{p-1}(N + N_{A_m}),\ \tilde{j} = 1, \dots, N + N_{A_p},\end{aligned}$$

for $p = 1, \dots, n$, where $s^+_l \in \Omega^+_p$, $s^-_l \in \Omega^-_p$ for $l = 0, \dots, N + N_{A_p}$, and $\chi_{[s_1,s_2)}$ denotes the characteristic function of the interval $[s_1, s_2)$. Equation (4.1) then becomes the linear system

$$\sum_{j=1}^{M_N} v_j((\rho_j, \rho_m) - (K^{\beta_c}_\beta \rho_j, \rho_m)) = (\Psi^{\beta_c}_\beta, \rho_m), \quad m = 1, \dots, M_N. \tag{4.2}$$

If k is large compared to N, then, from the definition of A_j in (3.4), it is clear that the two meshes Ω^+_j and Ω^-_j will not overlap. In this case the basis functions ρ_j,

$j = 1, \dots, M_N$, form an orthonormal basis for $V_{\Omega,\nu}$ (this is not true for the Galerkin method described in [20]), and hence the condition number of our linear system (4.2) will be bounded by (see, e.g., [2, section 3.6.3])

$$\begin{aligned}\|(I - P_N K_\beta^{\beta_c})\|_2 \|(I - P_N K_\beta^{\beta_c})^{-1}\|_2 &\leq (1 + \|K_\beta^{\beta_c}\|_2) \left(\frac{\mathrm{Re}\beta_c}{\mathrm{Re}\beta_c - \|\beta - \beta_c\|_\infty} \right) \\ &\leq \left(1 + \frac{\|\beta - \beta_c\|_\infty}{\mathrm{Re}\beta_c}\right) \left(\frac{\mathrm{Re}\beta_c}{\mathrm{Re}\beta_c - \|\beta - \beta_c\|_\infty} \right) \\ &= \frac{\mathrm{Re}\beta_c + \|\beta - \beta_c\|_\infty}{\mathrm{Re}\beta_c - \|\beta - \beta_c\|_\infty}, \qquad (4.3)\end{aligned}$$

where we have used (3.8) (under the assumption that (3.2) holds) and the facts that $\|K_\beta^{\beta_c}\|_2 \leq \|\beta - \beta_c\|_\infty / \mathrm{Re}\beta_c$ (see, e.g., [20, (3.2)]) and $\|P_N\|_2 = 1$. The fact that we can establish such a bound on the condition number of our linear system is in direct contrast to some other schemes in the literature where the approximation space consists of plane wave basis functions, e.g., [41, 44, 45], where serious difficulties due to ill-conditioning have been reported.

To evaluate the coefficients $(K_\beta^{\beta_c} \rho_j, \rho_m)$ and $(\Psi_\beta^{\beta_c}, \rho_m)$ of (4.2) we must compute some integrals numerically. The exact formulas are given in [38], but note that after some integrations are carried out analytically, the most difficult of these take the forms

$$\int_0^\infty \frac{(\mathrm{i} - r)F(r)}{r(r - 2\mathrm{i})}\,\mathrm{d}r, \int_0^\infty \frac{(1 - \mathrm{e}^{rs})F(r)}{r^2}\,\mathrm{d}r, \int_0^\infty \frac{(1 - \mathrm{e}^{rs})F(r)}{r(r - 2\mathrm{i})}\,\mathrm{d}r,$$

where $s < 0$ and $F(r)$ is given by (2.20). These integrals are similar in difficulty to integral representations for the Green's function G_{β^*}, for which very efficient numerical schemes are proposed in [18]. The integrands are not oscillatory and the coefficients do not become more difficult to evaluate as $k \to \infty$.

As a numerical example, we take $\theta = \pi/4$, $n = 1$, and

$$\beta(s) = \begin{cases} 0.505 - 0.3\mathrm{i}, & s \in [-m\lambda, m\lambda], \\ 1, & s \notin [-m\lambda, m\lambda], \end{cases}$$

for $m = 5$, 10, 20, 40, 80, 160, 320, 640, 1280, 2560, and 5120, where $k = 1$ and $\lambda = 2\pi$ is the wavelength. This experiment is equivalent to fixing the interval $[a, b] = [t_0, t_1]$ and decreasing the wavelength. The assumption (3.2) is satisfied, so that Theorem 3.5 holds. For each value of m, we compute Φ_N with $\nu = 0$, $\alpha = 25\sqrt{2}$ (so that $\alpha n^3/\cos\theta = \sqrt{2}\alpha = 50$, this value chosen experimentally) and $N = 2$, 4, 8, 16, 32, 64. For the purpose of computing errors, we take the "exact" solution (Φ^*) to be the solution computed with $\sqrt{2}\alpha = 1000$ and $N = 128$. Whereas for the scheme of [20] the number of degrees of freedom needed to maintain accuracy increases logarithmically with respect to $k(b - a)$ as $k(b - a) \to \infty$, here the number needed to maintain accuracy remains bounded as $k(b - a) \to \infty$, as we shall see below.

In Figure 4.1 we plot $|\Phi^*|$ and $|\Phi_2|$ for $m = 10$. Noting the logarithmic scales on the plots, it is clear that $|\Phi^*|$ is highly peaked near the discontinuities in impedance. Recalling that Φ is a correction term, namely, the difference between the true solution and the solution that there would be if the impedance were constant everywhere, the reason for this is clear. On the plot of $|\Phi_2|$ we also show the two grids Ω_1^+ and Ω_1^-. For

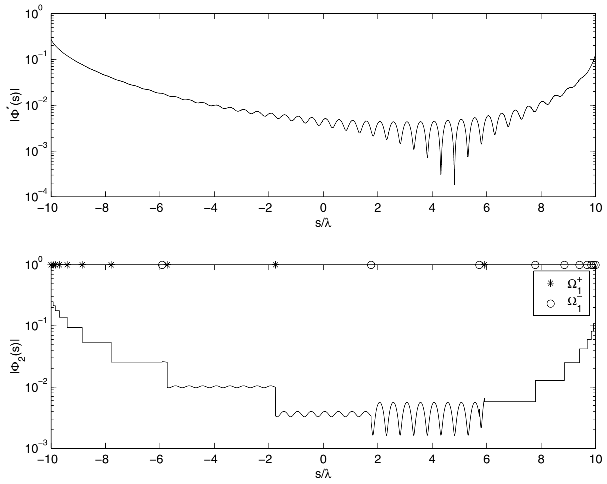


FIG. 4.1. *Plot of* $|\Phi^*|$ *and* $|\Phi_2|$*,* $m = 10$*, so that* $b - a = 20\lambda$*.*

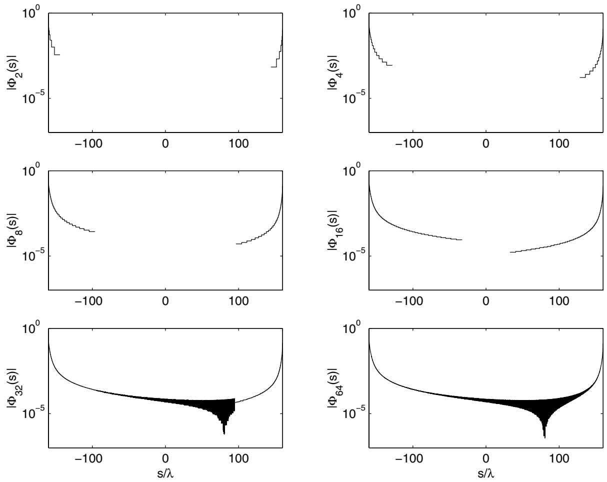


FIG. 4.2. *Plot of* $|\Phi_N|$*,* $N = 2$*, 4, 8, 16, 32, 64 for* $m = 160$*, so that* $b - a = 320\lambda$*.*

s/λ less than about -6 and for s/λ greater than about 6 the grids do not overlap, and on these regions $\Phi_2(s) = e^{\mathrm{i}s}\times$ (piecewise constants) and $\Phi_2(s) = e^{-\mathrm{i}s}\times$ (piecewise constants), respectively. Thus $|\Phi_2(s)|$ is piecewise constant where the grids do not overlap, and this can be clearly seen in Figure 4.1. Where the grids overlap, (roughly between $s/\lambda = -6$ and $s/\lambda = 6$) the oscillatory nature of $\Phi_2(s)$ is more apparent.

In Figure 4.2 we plot $|\Phi_N|$ for $m = 160$ and for $N = 2, 4, 8, 16, 32, 64$. Again noting the logarithmic scales on each plot, $|\Phi_N|$ is highly peaked near the impedance discontinuities, much more so than for $m = 10$. As N increases so we discretize a larger part of the domain $[-m\lambda, m\lambda]$, as well as having a finer mesh near the discontinuities in impedance at $-m\lambda$, $m\lambda$. For $N = 2, 4, 8, 16$ the piecewise constant approximation can be clearly seen, as the grids Ω_1^+ and Ω_1^- do not overlap. For $N = 32$ the grids overlap between about $s\lambda = -100$ and $s\lambda = 100$. For $N = 64$, each grid covers the whole domain $[-m\lambda, m\lambda]$.

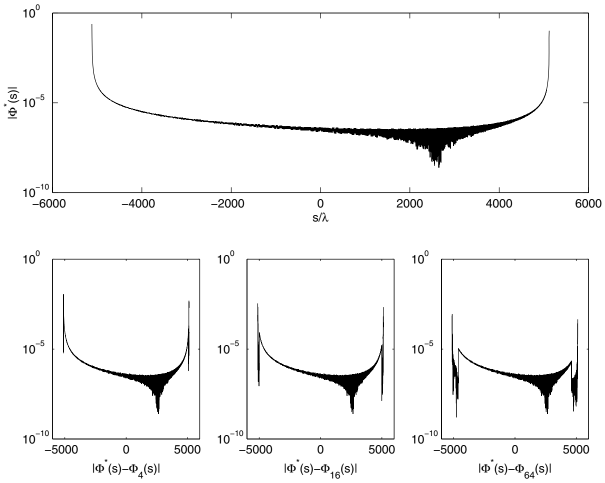


FIG. 4.3. *Plot of* $|\Phi^*|$ *and* $|\Phi^* - \Phi_N|$, $N = 4$, 16, 64 *for* $m = 5120$, *so that* $b - a = 10240\lambda$.

TABLE 4.1
$\|\Phi^* - \Phi_N\|_2/\|\Phi^*\|_2$ *for* $m = 10$, 160, *and* 5120, *and increasing* N.

$(b-a)/\lambda$	N	M_N	$\|\Phi^* - \Phi_N\|_2/\|\Phi^*\|_2$	EOC	COND
20	2	18	1.635×10^{-1}	1.1	1.8
	4	42	7.393×10^{-2}	1.1	2.6
	8	90	3.525×10^{-2}	1.0	8.1
	16	182	1.773×10^{-2}	1.0	94.0
	32	370	8.875×10^{-3}	1.0	625.5
	64	742	4.557×10^{-3}		2551.6
320	2	18	1.647×10^{-1}	1.2	1.8
	4	46	7.399×10^{-2}	1.0	2.0
	8	106	3.622×10^{-2}	1.0	2.0
	16	240	1.790×10^{-2}	1.0	2.1
	32	530	8.662×10^{-3}	0.9	2.1
	64	1094	4.537×10^{-3}		92.7
10240	2	18	1.639×10^{-1}	1.2	1.8
	4	46	6.918×10^{-2}	0.8	2.0
	8	106	3.881×10^{-2}	1.2	2.0
	16	240	1.751×10^{-2}	1.1	2.1
	32	530	8.076×10^{-3}	0.8	2.1
	64	1154	4.579×10^{-3}		2.1

In Figure 4.3 we plot $|\Phi^*|$ and $|\Phi^* - \Phi_N|$ for $m = 5120$ and for $N = 4$, 16 and 64. In this case the interval $[-m\lambda, m\lambda]$ is over 10,000 wavelengths long, and so even for $N = 64$ the grids Ω_1^+ and Ω_1^- do not overlap. As m increases, so $|\Phi^*|$ becomes even more peaked, and the benefit of clustering the grid points around the impedance discontinuities becomes even more apparent.

For $m = 10$, 160, and 5120 the relative L_2 errors $\|\Phi^* - \Phi_N\|_2/\|\Phi^*\|_2$ are shown in Table 4.1. (All L_2 norms are computed by approximating by discrete L_2 norms, sampling at 100,000 evenly spaced points in the relevant interval for the function whose norm is to be evaluated.) The estimated order of convergence is given by

$$\mathrm{EOC} := \log_2\left(\frac{\|\Phi^* - \Phi_N\|_2}{\|\Phi^* - \Phi_{2N}\|_2}\right).$$

Table 4.2
$\|\Phi^* - \Phi_{16}\|_2/\|\Phi^*\|_2$ *for increasing interval length.*

$(b-a)/\lambda$	M_N	$\|\Phi^* - \Phi_{16}\|_2/\|\Phi^*\|_2$	$\|\Phi^* - \Phi_{16}\|_2$	COND
10	162	1.746×10^{-2}	7.936×10^{-3}	181.5
20	182	1.773×10^{-2}	8.059×10^{-3}	94.0
40	204	1.775×10^{-2}	8.068×10^{-3}	24.7
80	226	1.766×10^{-2}	8.027×10^{-3}	8.2
160	240	1.761×10^{-2}	8.000×10^{-3}	2.1
320	240	1.790×10^{-2}	8.122×10^{-3}	2.1
640	240	1.749×10^{-2}	7.916×10^{-3}	2.1
1280	240	1.650×10^{-2}	7.435×10^{-3}	2.1
2560	240	1.616×10^{-2}	7.216×10^{-3}	2.1
5120	240	1.556×10^{-2}	6.831×10^{-3}	2.1
10240	240	1.751×10^{-2}	7.433×10^{-3}	2.1

For this example, Theorem 3.5 predicts that

$$\|\Phi^* - \Phi_N\|_2 \le \frac{C}{N}(1 + \log^{1/2}(\min(\sqrt{2}\alpha N, 2m\lambda))),$$

so that we expect EOC ≈ 1, and this is what we see. For each value of m, the number of degrees of freedom M_N increases approximately in proportion to $N \log N$ as N increases until the two grids Ω_1^+ and Ω_1^- each cover the whole domain $[-m\lambda, m\lambda]$ (i.e., until $\sqrt{2}\alpha N \ge 2m\lambda$), after which M_N increases only proportionally to N as N increases further. For $m = 10$, the whole domain is covered by the grids for $N = 4$; for $m = 160$ this occurs for $N = 64$ but for $m = 5120$ the two grids do not overlap even for $N = 64$. The condition numbers for the matrix of the linear system (4.2) (denoted by COND) satisfy the bound (4.3), which predicts that COND ≤ 3.75 if the grids do not overlap, i.e., if $N \le 16$ for $m = 160$, for all values of N when $m = 5120$. For $N \le 32$ the number of degrees of freedom is the same for $m = 160$ and $m = 5120$, and yet the relative L_2 error is almost the same for the two cases $b - a = 320\lambda$ and $b - a = 10240\lambda$.

In Table 4.2 we fix $N = 16$ and show $\|\Phi^* - \Phi_{16}\|_2/\|\Phi^*\|_2$ and also $\|\Phi^* - \Phi_{16}\|_2$ for increasing values of $m = (b - a)/2\lambda$. As m increases, the number of degrees of freedom increases logarithmically for those values of m for which $\sqrt{2}\alpha N \ge 2m\lambda$, i.e., for $m \le 40$, but as m increases further for $m \ge 80$ the number of degrees of freedom remains constant, and yet both the relative and the actual L_2 error also remain roughly constant as m grows. For $m = 5120$ the interval is of length greater than 10,000 wavelengths, and yet we achieve almost 1% relative error with only 240 degrees of freedom. As in Table 4.1, the condition number of the linear system (4.2) is bounded by (4.3), so that COND ≤ 3.75, when m is sufficiently large that the grids Ω_1^+ and Ω_1^- do not overlap, i.e., for $m \ge 160$.

In the last figure and table we show numerical computations of the total field above the boundary, i.e., $u_N^t(x)$ given by (3.10). We note that computing $u_N^t(x)$ requires, for each point x, the computation of the highly oscillatory integral (3.10), which is evaluated here using accurate but slow "black box" techniques. In the future it is hoped that more efficient quadrature schemes can be developed, taking advantage of the fact that the oscillatory parts of both G_{β_c} and ϕ_N are known explicitly. We note that Iserles [34, 35] has recently proposed and analyzed Filon-type quadrature methods appropriate for the efficient evaluation of highly oscillatory integrals, which we expect may be appropriate.

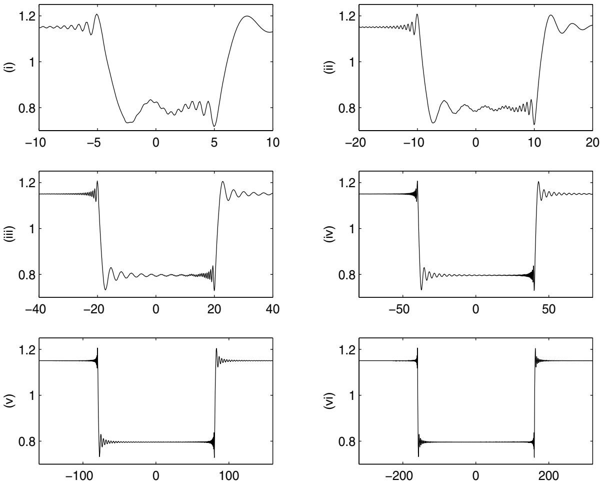


FIG. 4.4. $|u^t(x)|$ *(on the y-axis) against* x_1/λ *(on the x-axis) for* $x = (x_1, \lambda)$, $x_1 \in [-2m\lambda, 2m\lambda]$, *plotted for* $m = 5$ *(plot* (i)*),* $m = 10$ *(plot* (ii)*),* $m = 20$ *(plot* (iii)*),* $m = 40$ *(plot* (iv)*),* $m = 80$ *(plot* (v)*), and* $m = 160$ *(plot* (vi)*).*

TABLE 4.3
$|u^t(x) - u^t_N(x)|$ *for* $m = 10$ *and* $m = 160$*, and increasing* N*.*

		$x = (m\lambda/2, \lambda)$		$x = (m\lambda, \lambda)$	
m	N	$\|u^t(x) - u^t_N(x)\|$	EOC	$\|u^t(x) - u^t_N(x)\|$	EOC
10	2	3.894×10^{-4}	1.5	1.108×10^{-4}	1.3
	4	1.421×10^{-4}	2.5	4.514×10^{-5}	3.5
	8	2.432×10^{-5}	1.0	4.068×10^{-6}	0.2
	16	1.183×10^{-5}	2.7	3.448×10^{-6}	0.8
	32	1.841×10^{-6}	1.0	2.014×10^{-6}	1.1
	64	9.350×10^{-7}		9.108×10^{-7}	
160	2	1.059×10^{-4}	2.0	5.278×10^{-4}	2.6
	4	2.572×10^{-5}	0.4	8.790×10^{-5}	2.8
	8	1.978×10^{-5}	0.0	1.283×10^{-5}	0.3
	16	1.981×10^{-5}	2.1	1.060×10^{-5}	0.8
	32	4.474×10^{-6}	0.9	6.029×10^{-6}	3.4
	64	2.431×10^{-6}		5.634×10^{-7}	

In Figure 4.4 we plot $|u^t_{128}(x)|$ for $x = (x_1, \lambda)$, $x_1 \in [-2m\lambda, 2m\lambda]$, i.e., the absolute value of the total acoustic field one wavelength above the plane, as computed with $\sqrt{2}\alpha = 1000$ and $N = 128$, for $m = 5$ (plot (i)), $m = 10$ (plot (ii)), $m = 20$ (plot (iii)), $m = 40$ (plot (iv)), $m = 80$ (plot (v)), and $m = 160$ (plot (vi)). In each plot the x-axis represents x_1/λ and the y-axis represents $|u^t(x)|$. One can clearly see that the wave diffracted from the impedance discontinuities at $x = (-m\lambda, 0)$ and $x = (m\lambda, 0)$ is a significant component of the total field only within a small number of wavelengths of the impedance discontinuities. Figure 1.1 shows a surface plot of the incident, scattered and total wave fields up to 10 wavelengths above the plane for this same example with $m = 5$.

We also computed $u^t_N(x)$ for $x = (m\lambda/2, \lambda)$ and $x = (m\lambda, \lambda)$ for $m = 10$ and $m = 160$ and for $\sqrt{2}\alpha = 50$, $N = 2$, 4, 8, 16, 32, and 64. Taking the values for $\alpha = 500\sqrt{2}$, $N = 128$ to be the "exact" values, the errors are shown in Table 4.3. The estimated order of convergence is calculated as

$$\mathrm{EOC} := \log_2 \left(\frac{|u^t(x) - u^t_N|}{|u^t(x) - u^t_{2N}|} \right),$$

and from Theorem 3.6 we would expect EOC ≈ 1. The convergence rate is rather irregular, but broadly speaking it is at least as good as expected, and the actual and relative errors are both very small. At every point x it holds that $0.7 < |u^t(x)| < 0.9$.

Further numerical results for $\theta \approx \pi/2$, i.e., grazing incidence, can be found in [39].

5. Conclusions and discussion. In this paper we have presented a Galerkin boundary element method for an acoustic scattering problem, and we have demonstrated, via both an a priori error analysis and numerical examples, that the number of degrees of freedom required for an accurate solution is bounded independently of the wavenumber. Our numerical method and analysis are for a specific scattering problem, namely the 2D problem of scattering by an unbounded flat surface with piecewise constant surface impedance, this problem being important in the theory of outdoor noise propagation and in an electromagnetic context.

As we discussed in our review of the literature, our method is an instance of the general idea of expressing the solution of the scattering problem as a finite sum of known oscillatory terms (given by the leading order behavior of the solution as $k \to \infty$) multiplied by unknown more slowly oscillating terms, these smoother components much more suitable for approximation by standard finite element functions than the original solution. Our results add to the evidence of the theory and numerical experiments of other authors [1, 29, 24, 10, 12] that this general methodology has promise for a range of scattering problems.

Specifically, we anticipate that many of the details of our numerical scheme and analysis will be applicable to other interesting scattering problems. This is clearest in the case of 2D acoustic scattering by a convex polygon, in the case that a homogeneous Dirichlet condition or an impedance boundary condition with constant impedance holds on each side. For this problem we expect that the behavior of the total field on each side of the polygon (after subtraction of the leading order high frequency asymptotics given by physical optics) will be very similar to the behavior quantified in Theorems 2.3 and 2.6. Thus the same mesh may be applicable and much of the same analysis. For more discussion of scattering by a 2D polygon see [20, section 6], [19].

Moreover, we expect that our mesh design will be relevant more generally, at least for representing certain components of the total field. In the case of three-dimensional scattering by convex polyhedra it seems to us likely that the mesh we propose will be useful in representing the variation of edge scattered waves in the direction perpendicular to the edge. In the case of 2D convex curvilinear polygons something close to the mesh we use on each interval $[t_{j-1}, t_j]$ may be appropriate on each side of the polygon, especially at higher frequencies when our mesh becomes more localized near the ends of the intervals just as the waves diffracted by the corners become more localized near the corners.

Acknowledgment. Some of this work was carried out while the authors participated in the 2003 Programme Computational Challenges in Partial Differential Equations at the Isaac Newton Institute, Cambridge, UK.

REFERENCES

[1] T. Abboud, J. C. Nédélec, and B. Zhou, *Méthodes des équations intégrales pour les hautes fréquences*, C.R. Acad. Sci. I Math., 318 (1994), pp. 165–170.

[2] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.

[3] K. Attenborough, *Acoustical impedance models for outdoor ground surfaces*, J. Sound Vib., 99 (1985), pp. 521–544.

[4] I. Babuška and J. Melenk, *The partition of unity method*, Internat. J. Numer. Meth. Engrg., 40 (1997), pp. 727–758.

[5] I. Babuška and S. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers*, SIAM J. Numer. Anal., 34 (1997), pp. 2392–2423.

[6] F. X. Becot, P. J. Thorsson, and W. Kropp, *An efficient application of equivalent sources to noise propagation over inhomogeneous ground*, Acta Acoust. United Ac., 88 (2002), pp. 853–860.

[7] P. Boulanger, K. Attenborough, and Q. Qin, *Effective impedance of surfaces with porous roughness: Models and data*, J. Acoust. Soc. Am., 117 (2005), pp. 1146–1156.

[8] P. Boulanger, K. Attenborough, S. Taherzadeh, T. WatersFuller, and K. M. Li, *Ground effect over hard rough surfaces*, J. Acoust. Soc. Am., 104 (1998), pp. 1474–1482.

[9] O. Bruno and L. Kunyansky, *A fast, high-order algorithm for the solution of surface scattering problems: Basic implementation, tests and applications*, J. Comput. Phys., 169 (2001), pp. 80–110.

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

[11] O. P. Bruno, C. A. Geuzaine, and F. Reitich, *A new high-order high-frequency integral equation method for the solution of scattering problems* i: *Single-scattering configurations*, in Proceedings of the 20th Annual Review of Progress in Applied Computational Electromagnetics, Syracuse, NY, 2004.

[12] O. P. Bruno, C. A. Geuzaine, and F. Reitich, *A new high-order high-frequency integral equation method for the solution of scattering problems* ii: *Multiple-scattering configurations*, Proceedings of the 20th Annual Review of Progress in Applied Computational Electromagnetics, Syracuse, NY, 2004.

[13] O. Cessenat and B. Després, *Application of an ultra weak variational formulation of elliptic PDEs to the two-dimensional Helmholtz problem*, SIAM J. Numer. Anal., 35 (1998), pp. 255–299.

[14] O. Cessenat and B. Després, *Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation*, J. Comp. Acoust., 11 (2003), pp. 227–238.

[15] S. N. Chandler-Wilde, *Ground Effects in Environmental Sound Propagation*, Ph.D. thesis, University of Bradford, UK, 1988.

[16] S. N. Chandler-Wilde, *The impedance boundary value problem for the Helmholtz equation in a half-plane*, Math. Methods Appl. Sci., 20 (1997), pp. 813–840.

[17] S. N. Chandler-Wilde and D. C. Hothersall, *Sound propagation above an inhomogeneous impedance plane*, J. Sound Vib., 98 (1985), pp. 475–491.

[18] S. N. Chandler-Wilde and D. C. Hothersall, *Efficient calculation of the Green's function for acoustic propagation above a homogeneous impedance plane*, J. Sound Vib., 180 (1995), pp. 705–724.

[19] S. N. Chandler-Wilde and S. Langdon, *A Galerkin boundary element method for high frequency scattering by convex polygons* (in preparation).

[20] S. N. Chandler-Wilde, S. Langdon, and L. Ritter, *A high-wavenumber boundary-element method for an acoustic scattering problem*, Philos. Trans. R. Soc. London A, 362 (2004), pp. 647–671.

[21] S. N. Chandler-Wilde, M. Rahman, and C. R. Ross, *A fast two-grid and finite section method for a class of integral equations on the real line with application to an acoustic scattering problem in the half-plane*, Numer. Math., 93 (2002), pp. 1–51.

[22] S. H. Christiansen and J. C. Nédélec, *Preconditioners for the numerical solution of boundary integral equations from electromagnetism*, C.R. Acad. Sci. I Math., 331 (2000), pp. 733–738.

[23] D. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, John Wiley, New York, 1983.

[24] E. Darrigrand, *Coupling of fast multipole method and microlocal discretization for the 3-D Helmholtz equation*, J. Comput. Phys., 181 (2002), pp. 126–154.

[25] E. Darve and P. Havé, *A fast multipole method for maxwell equations stable at all frequencies*, Philos. Trans. R. Soc. London A, 362 (2004), pp. 603–628.

[26] A. de La Bourdonnaye, *A microlocal discretization method and its utilization for a scattering problem*, C.R. Acad. Sci. I Math., 318 (1994), pp. 385–388.

[27] A. de La Bourdonnaye and M. Tolentino, *Reducing the condition number for microlocal discretization problems*, Philos. Trans. R. Soc. London A, 362 (2004), pp. 541–559.

[28] M. Ganesh, S. Langdon, and I. H. Sloan, *Efficient evaluation of highly oscillatory acoustic scattering surface integrals*, Reading University Numerical Analysis Report 6/05, submitted for publication to J. Comp. Appl. Math.

[29] E. Giladi and J. Keller, *A hybrid numerical asymptotic method for scattering problems*, J. Comput. Phys., 174 (2001), pp. 226–247.

[30] D. Habault, *Sound propagation above an inhomogeneous plane*, J. Sound Vib., 100 (1985), pp. 55–67.

[31] D. C. Hothersall and J. N. B. Harriott, *Approximate models for sound propagation above multi-impedance plane boundaries*, J. Acoust. Soc. Am., 97 (1995), pp. 918–926.

[32] T. Huttunen, P. Monk, F. Collino, and J. P. Kaipio, *The ultra-weak variational formulation for elastic wave problems*, SIAM J. Sci. Comput., 25 (2004), pp. 1717–1742.

[33] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Springer-Verlag, New York, 1998.

[34] A. Iserles, *On the numerical quadrature of highly-oscillating integrals* I*: Fourier transforms*, IMA J. Numer. Anal., 24 (2004), pp. 365–391.

[35] A. Iserles, *On the numerical quadrature of highly-oscillating integrals* II*: Irregular oscillations*, IMA J. Numer. Anal., 25 (2005), pp. 25–44.

[36] I. M. Kaganova, *The impedance boundary conditions and effective surface impedance of inhomogeneous metals*, Phys. B Cond. Matter, 338 (2003), pp. 38–43.

[37] O. Laghrouche, P. Bettess, E. Perrey-Debain, and J. Trevelyan, *Wave interpolation finite elements for helmholtz problems with jumps in the wave speed*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 367–381.

[38] S. Langdon and S. N. Chandler-Wilde, *A Galerkin boundary element method for an acoustic scattering problem, with convergence rate independent of frequency*, in Proceedings of the 4th UK Conference on Boundary Integral Methods, S. Amini, ed., Salford University Press, 2003, pp. 67–76.

[39] S. Langdon and S. N. Chandler-Wilde, *A GTD-based boundary element method for a surface scattering problem*, Proc. Inst. Acoustics, 25 (2003), pp. 224–233.

[40] S. Langdon and I. G. Graham, *Boundary integral methods for singularly perturbed boundary value problems*, IMA J. Numer. Anal., 21 (2001), pp. 217–237.

[41] P. Monk and D. Q. Wang, *A least squares method for the Helmholtz equation*, Comput. Methods Appl. Math., 175 (1999), pp. 121–136.

[42] P. M. Morse and K. U. Ingard, *Theoretical Acoustics*, McGraw–Hill, New York, 1968.

[43] E. Perrey-Debain, O. Lagrouche, P. Bettess, and J. Trevelyan, *Plane-wave basis finite elements and boundary elements for three-dimensional wave scattering*, Philos. Trans. R. Soc. London A, 362 (2004), pp. 561–577.

[44] E. Perrey-Debain, J. Trevelyan, and P. Bettess, *Plane wave interpolation in direct collocation boundary element method for radiation and wave scattering: Numerical aspects and applications*, J. Sound Vib., 261 (2003), pp. 839–858.

[45] E. Perrey-Debain, J. Trevelyan, and P. Bettess, *Use of wave boundary elements for acoustic computations*, J. Comput. Acoust., 11 (2003), pp. 305–321.

[46] E. Perrey-Debain, J. Trevelyan, and P. Bettess, *On wave boundary elements for radiation and scattering problems with piecewise constant impedance*, IEEE Trans. Ant. Prop., 53 (2005), pp. 876–879.

[47] J. R. Poirier, A. Bendali, and P. Borderies, *Impedance boundary condition for rapidly oscillating surface scatterers*, in *Mathematical and Numerical Aspects of Wave Propagation*, A. Bermudez, D. Gomez, P. Joly, and J. E. Roberts, eds., SIAM, Philadelphia, 2000, pp. 528–532.

[48] M. Shimoda, R. Iwaki, and M. Miyoshi, *Scattering of an electromagnetic plane wave by a plane with local change of surface impedance*, IEICE Trans. Electronics, E87C (2004), pp. 44–51.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2478–2499

ERROR ESTIMATES FOR DISCONTINUOUS GALERKIN APPROXIMATIONS OF IMPLICIT PARABOLIC EQUATIONS*

K. CHRYSAFINOS[†] AND NOEL J. WALKINGTON[‡]

Abstract. We analyze the classical discontinuous Galerkin method for "implicit" parabolic equations. Symmetric error estimates for schemes of arbitrary order are presented. The ideas developed allow certain assumptions frequently required in previous work to be relaxed. For example, different discrete spaces are allowed at each time step, and the spatial operator is not required to be self-adjoint or independent of time. Error estimates are posed in terms of projections of the exact solution onto the discrete spaces and are valid under the minimal regularity guaranteed by the natural energy estimate. These projections are local and enjoy optimal approximation properties when the solution is sufficiently regular.

Key words. error estimates, discontinuous Galerkin, implicit parabolic equations

AMS subject classification. 65M60

DOI. 10.1137/040602948

1. Introduction.

We consider implicit parabolic partial differential equations of the form

$$(M(t)u)_t + A(t)u = F(t), \qquad u(0) = u_0. \tag{1.1}$$

The operators act on Hilbert spaces related through the standard pivot construction, $U \hookrightarrow H \simeq H' \hookrightarrow U'$, where each embedding is continuous and dense. Then, $A(.) : U \to U'$ is a linear map and $F(.) \in U'$. It is assumed that $M(.) : H \to H$ is a self-adjoint positive definite operator.

Conservation laws for systems undergoing diffusion may take the form of (1.1) when the capacity changes with time; for example, in a porous medium the porosity could change as the medium collapses due to oil being extracted from the reservoir. Classical parabolic equations (i.e., equations with M the identity) also take the form of (1.1) under a time-dependent change of coordinates, common examples being diffusion on surfaces (more generally manifolds) which are in motion and the Lagrange (or characteristic) Galerkin formulation of the convection diffusion equation [10, 19].

Here the classical discontinuous Galerkin (DG) scheme for approximating solutions of (1.1) is analyzed, and fully discrete error estimates are derived under minimal regularity assumptions. The class of DG schemes considered is classical in the sense that the discrete solutions may be discontinuous in time but are conforming in space, i.e., are in (a subspace of) U at each time. Our analysis extends the ideas introduced in [6] and addresses the following issues which have not yet been adequately considered in the literature.

[*] Received by the editors January 7, 2004; accepted for publication (in revised form) May 13, 2005; published electronically January 12, 2006. This research was supported in part by National Science Foundation grants DMS-0208586 and ITR-0086093. This work was also supported by the NSF through the Center for Nonlinear Analysis.

http://www.siam.org/journals/sinum/43-6/60294.html

[†] Instutüt für Numerische Simulation, Wegelstr 6, University of Bonn, Bonn 53115, Germany (chrysafinos@ins.uni-bonn.de).

[‡] Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213 (noelw@cmu.edu).

- A systematic treatment of DG approximations of implicit parabolic equations of the form (1.1) has not been considered in the past.
- The natural setting for (1.1) involves time dependent norms (Hilbert scales); this gives rise to technical problems not encountered in the analysis of classical parabolic equations.
- The operator $A(.)$ may depend upon time and is not required to be self-adjoint.
- The subspaces of U used for the DG approximations may be different on each time interval $(t^{n-1}, t^n]$. This adds a significant complication to the analysis which is present even when $A = 0$. Indeed, the first step in our analysis is to consider the DG scheme for an auxiliary equation which reduces to an implicit ODE when the coercivity constant vanishes. This limiting case plays an important role in the error analysis.

  Different subspaces are an essential ingredient of adaptive strategies used in conjunction with a posteriori error estimates to give guaranteed error bounds. Retriangulation is also necessary for many algorithms based upon a Lagrangian coordinate system; below we present an example.
- The operator $A(.)$ is not required to be strictly coercive; semicoercivity of the form $\langle A(.)u, u\rangle \geq c|u|^2_{U(.)} - C\|u\|^2_{H(.)}$ is assumed. Here $|.|_{U(.)}$ is a seminorm such that $\|.\|^2_{U(.)} = |.|^2_{U(.)} + \|.\|^2_{H(.)}$. This causes significant problems in the analysis of DG schemes since the classical Gronwall argument, used for the continuous problem, fails in the discrete setting. This failure is due to the elementary observation that functions of the form $\chi_{[0,\hat{t})}u_h$ are not polynomial in time unless $\hat{t}$ is a partition point, so these functions are not available as test functions in the discrete setting.[1] Below these issues are circumvented by constructing polynomial approximations to the characteristic functions $\chi_{[0,\hat{t})}$.

As stated above, our analysis does not require any regularity above and beyond the natural bounds that follow from the usual energy estimate. This is essential for control problems where solutions of the dual problem typically will not exhibit any additional regularity. Care is taken to keep track of how the various constants depend upon the coercivity constant of $A(.)$. This is important for the analysis of problems like the convection diffusion equation where the coercivity constant is small.

We present an example which can be analyzed within the general framework developed here but falls outside of the theory developed, for example, in Thomée's text [28].

Example: Diffusion on manifolds. As an illustrative example, consider diffusion on a cell membrane, $\mathcal{S}(t) \subset \mathbb{R}^3$, which is being transported in an ambient fluid with velocity $\mathbf{V} = \mathbf{V}(t, x)$. In order to avoid triangulating the manifold at each step, a numerical scheme may compute the triangulation of a reference configuration, $\mathcal{S}_r$, and at each time t construct a mapping $x(t, .) : \mathcal{S}_r \to \mathcal{S}(t) \subset \mathbb{R}^3$. The reference configuration is typically $\mathcal{S}(0)$ or possibly the unit sphere S^2. If $\mathcal{S}_r$ is locally parameterized by coordinates $X \in U \subset \mathbb{R}^2$, the diffusion equation with diffusion constant $\sigma > 0$ takes the form

$$u_t - (1/J)\mathrm{div}_X \left(\sigma J (F^T F)^{-1} \nabla_X u\right) = 0, \tag{1.2}$$

where F is the 3×2 matrix with components $F_{i\alpha} = \partial x_i / \partial X_\alpha$ and $J = \sqrt{\det(F^T F)}$

[1] Here $\chi_{[0,\hat{t})}$ is the characteristic function equal to 1 on $[0, \hat{t})$ and equal to zero otherwise.

is the determinant of the first fundamental form. The determinant J satisfies

$$J_t = J\,(I - \mathbf{n}\otimes\mathbf{n})\cdot(\nabla_x \mathbf{V}) = J\sum_{ij}(\delta_{ij} - n_i n_j)(\partial V_i/\partial x_j),$$

where $\mathbf{n} = \mathbf{n}(t, X)$ is the normal to $\mathcal{S}(.)$. It follows that the diffusion equation is of the form (1.1) with $M(.)u = Ju$ and

$$A(.)u = -(I - \mathbf{n}\otimes\mathbf{n})\cdot(\nabla_x \mathbf{V})uJ - \operatorname{div}_X\left(\sigma J(F^T F)^{-1}\nabla_X u\right).$$

As $\mathcal{S}(.)$ and the solution $u(.)$ evolve, adaptive error control may modify the triangulation of $\mathcal{S}_r$ giving rise to different discrete subspaces on different intervals $(t^{n-1}, t^n]$ of $[0, T]$. It is also possible that $\mathcal{S}(t)$ will undergo large shears in which case matrix $F^T F$ will become ill conditioned; in this situation the geometry of $\mathcal{S}_r$ no longer resembles that of $\mathcal{S}(t)$, so the reference configuration needs to be updated and retriangulated. This example is considered in more detail in section 6 below.

1.1. Related results. The DG method was first introduced by Lasaint and Raviart [17] to simulate neutron transport. There is an abundant literature concerning applications of the DG scheme in hyperbolic problems; see, e.g., [5, 15, 29] and references within. The DG method for ordinary differential equations was considered by Delfour, Hager, and Trochu in [7]. They showed that the DG scheme was super convergent at the partition points (order $2k+2$ for polynomials of degree k).

In the context of parabolic equations DG schemes were first analyzed for linear parabolic problems by Jamet in [14] where $\mathcal{O}(\tau^k)$ results were proved and then by Eriksson, Johnson, and Thomée [13] where $\mathcal{O}(\tau^{2k-1})$ estimates are established at the partition points for smooth solutions. An excellent exposition of their results and, more generally, the DG method for parabolic equations can be found in Thomée's book [28]. In [28] nodal and interior estimates are presented in various norms. One may also consult [20] for the analysis of a related formulation based on the backward Euler scheme. The relation between the DG scheme and adaptive techniques was studied in [11] and [12]. Finally, some results concerning the analysis of parabolic integro-differential equations by discontinuous Galerkin method are presented in [18] (see also references therein).

In [10] DuPont and Liu introduce the concept of "symmetric error estimates" for parabolic problems. They define such an error estimate to be one of the form

$$|||u - u_h||| \le C \inf_{w_h\in\mathcal{U}_h} |||u - w_h|||,$$

where u and u_h are the exact and approximate solutions respectively, $|||.|||$ is an appropriate norm, and $\mathcal{U}_h$ is the discrete subspace in which approximation solutions are sought. While estimates of this form are standard for elliptic problems, this is not the case for evolution problems. For example, error estimates for evolution problems approximated by the implicit Euler scheme frequently involve terms of the form $\|u_{tt}\|_{L^2(\Omega)}$. Symmetric error estimates are useful for problems where the solution u may not be very regular, such as control problems, and are used to develop a posteriori error estimates for adaptive schemes. Symmetric error estimates for moving mesh finite element methods were studied in [10, 19] (see also references therein). Mesh modification techniques for finite elements have also been introduced in [21] and [22]. For some earlier work on convection-dominated problems based on the methods of characteristics and mesh modification, one may consult [8] and [9], respectively.

An alternative to the symmetric error estimates are estimates of the form

$$|||u - u_h||| \leq C|||u - \mathbb{P}_h u|||, \tag{1.3}$$

where $\mathbb{P}_h : \mathcal{U} \to \mathcal{U}_h$ is a projection which exhibits optimal interpolation properties if u is sufficiently smooth. Estimates of this form enjoy the same advantages of those proposed by DuPont and Liu. Theorem 5.1 below provides an estimate of the form (1.3) for implicit parabolic equations of the form (1.1), where the projection $\mathbb{P}_h u$ is the numerical approximation of an auxiliary equation using the DG scheme and is not local. However,

$$|||u - \mathbb{P}_h u||| \leq |||u - \mathbb{P}_h^{loc} u||| + |||\mathbb{P}_h u - \mathbb{P}_h^{loc} u|||,$$

where $\mathbb{P}_h^{loc}$ is a local projection, so the first term can be estimated using classical interpolation theory. The second term, $|||\mathbb{P}_h u - \mathbb{P}_h^{loc} u|||$, vanishes if the same subspace of U is used in each partition (t^{n-1}, t^n); otherwise, it depends solely upon the jump in the interpolant of the exact solution at the partition points $\{t^n\}_{n=0}^N$. The size of the constant C in (1.3) and its dependence on various constants play an important role; below we are careful to state the dependence of the constant upon the various coercivity constants and bounds assumed for the operator A.

Error estimates for Lagrange–Galerkin approximations of convection dominated problems for divergence-free velocity fields vanishing on the boundary are presented in [4]. Issues related to the stability of Lagrange–Galerkin approximations are also discussed in [23]. Recently there has been a lot of work on the development and analysis of discontinuous (in space) Galerkin methods for elliptic problems. A comprehensive survey and comparison of this work can be found in [3], which contains many references related to this approach.

1.2. Outline. In section 2 we introduce spaces and structural assumptions on the operators which guarantee that (1.1) is a well-posed implicit parabolic equation. Discrete spaces used for the DG approximation of (1.1) are introduced in section 3, and "discrete characteristic functions" are constructed on these spaces. These were introduced in [6] and are modified here to accommodate the time-dependent spaces.

We formulate and analyze the DG scheme for an auxiliary equation in section 4. This section focuses on the difficulties that arise in the presence of time-dependent norms and when different subspaces of U are used at every time step. Finally in section 5 error estimates are developed for the DG approximation of (1.1). The approximate solution of (1.1) is first compared with that of the auxiliary equation in Theorem 5.1. The results of section 4 are then used to obtain error estimates which take the form of the sum of (i) the "local truncation error," (ii) projection errors between different subspaces, and (iii) errors in the initial data.

One technical distinction between the error estimate developed for the classical parabolic problem in [6, Theorem 3.1] and Theorem 5.1 of section 5 is that the latter assumes the existence of an inverse hypothesis of the form $\|u_h\|_{U(t)} \leq C_{inv}(h)\|u_h\|_{H(t)}$ for u_h in the discrete subspaces of U. In Theorem 5.1 the product $\tau C_{inv}(h)$ enters into the error estimate where τ is the time step size. For classical second order parabolic problems, this term will be of order $\mathcal{O}(1)$ if $\tau \sim h$ and quasi-uniform finite element meshes with no small angles are used.

1.3. Notation. Spaces $H(t) = (H, \|.\|_{H(t)})$ and $U(t) = (U, \|.\|_{U(t)})$ with time dependent norms are used. The pivot spaces $H(t)$ have inner product $(u, v)_{H(t)} = (M(t)u, v)_H$, and the norms on the these spaces are often denoted by $|.|_{H(t)} \equiv \|.\|_{H(t)}$.

Each norm and inner product will be explicitly subscripted; while this is rather cumbersome, it helps to minimize confusion due the plethora of spaces and projections. Notation of the form $L^2[0,T;U(.)]$, $H^1[0,T;U'(.)]$, etc., is used to indicate the temporal regularity of functions with values in $U(.)$, $U'(.)$, etc.

Approximations of (1.1) will be constructed on a partition $0 = t^0 < t^1 < \cdots < t^N = T$ of $[0,T]$ with time step size denoted by $\tau^n = t^n - t^{n-1}$. On each interval of the form $(t^{n-1}, t^n]$ a subspace U_h^n of U is specified, and the approximate solutions will lie in the space

$$\mathcal{U}_h = \{u_h \in L^2[0,T;U(.)] \mid u_h|_{(t^{n-1},t^n]} \in \mathcal{P}_k(t^{n-1}, t^n; U_h^n)\}.$$

Here $\mathcal{P}_k(t^{n-1}, t^n; U_h^n)$ is the space of polynomials of degree k or less having values in U_h^n. Notice that, by convention, functions in $\mathcal{U}_h$ have been chosen to be left continuous with right limits. We will write u^n for $u_h(t^n) = u_h(t_-^n)$ and let u_+^n denote $u(t_+^n)$. This notation is also used with functions like the error $e = u - u_h$. The exact solution, u, is assumed to be in $C[0,T;H(.)]$ so that the jump in the error at t^n, denoted by $[e^n]$, is equal to $[u^n] = u_+^n - u^n$.

2. Implicit parabolic equations.

In this section structural assumptions required for the analysis of the implicit parabolic problem

$$(M(t)u)_t + A(t)u = F(t), \qquad u(0) = u_0, \tag{2.1}$$

are introduced. To characterize the time dependence of $A(.)$, equivalent norms on U of the form $\|u\|_{U(t)}^2 = |u|_{H(t)}^2 + |u|_{U(t)}^2$ are considered where $|.|_{U(t)}$ is a seminorm on U (the principle part) and $|.|_{H(t)} = (M(t).,.)_H$ is a norm on H with Riesz map (the symmetric positive operator) $M(t)$. Let $a(.;u,v)$ denote the natural bilinear form associated with $A(.)$, and assume that the spaces $(U(t), H(t))$ satisfy $U(t) \hookrightarrow H(t) \hookrightarrow U'(t)$, where each embedding is dense and continuous and the embedding constant is independent of time.

2.1. Structural assumptions.

The existence theory for implicit evolution equations of the form (2.1) almost always requires the operators $M(.)$ to be Riesz maps for a Hilbert space [26, 27], and this assumption is used in the analysis below.

Assumption 1. The operators $M(\cdot)$ are nonnegative and self-adjoint, and there exist constants $c(t) > 0$ such that

$$(M(t)u, u)_H \geq c(t)|u|_H^2.$$

It follows for each $t \geq 0$ that $(M(t)u, v)$ is an inner product on H, which is denoted by $(.,.)_{H(t)}$.

Definition 2.1. *$H(t)$ is the Hilbert space with underlying set H and inner product $(u,v)_{H(t)} = (M(t)u, v)_H$.*

With this notation it is possible to state the structural hypotheses which guarantee that (2.1) is parabolic in nature and facilitate the development of error estimates.

Assumption 2.

1. Smoothness of $M(t)$: For each $t > 0$ there exists a symmetric bilinear form, $\mu(t,.,.)$, satisfying

$$\frac{d}{dt}(u,v)_{H(t)} = (u_t, v)_{H(t)} + (u, v_t)_{H(t)} + \mu(t; u, v)$$

for u, $v \in H^1[0,T;H]$, and there exists $C_\mu > 0$ independent of time such that

$$|\mu(t,u,v)| \le C_\mu |u|_{H(t)} |v|_{H(t)}.$$

2. Equivalence of norms on $U(t)$: For each $0 < \tau \le T$ there exists $C_u > 0$ such that for all s, $t \ge 0$ with $|t-s| < \tau$

$$1/C_u \le \|v\|_{U(t)}/\|v\|_{U(s)} \le C_u \qquad \forall\, v \in U.$$

3. Continuity of the bilinear form and data: There exist nonnegative constants $0 \le c_a \le C_a$ such that

$$|a(t;u,v)| \le \Big(c_a|u|^2_{U(t)} + C_a|u|^2_{H(t)}\Big)^{1/2}\Big(c_a|v|^2_{U(t)} + C_a|v|^2_{H(t)}\Big)^{1/2},$$

and there exists a (weighted dual) norm $\|.\|_*$ equivalent to $\|.\|_{U'}$ such that

$$|\langle F(t),u\rangle| \le \|F(t)\|_* \Big(c_a|u|^2_{U(t)} + C_a|u|^2_{H(t)}\Big)^{1/2}.$$

4. Coercivity of the bilinear form: There exist constants $C_\alpha \in \mathbb{R}$ and $c_\alpha > 0$ such that

$$a(t;u,u) \ge c_\alpha |u|^2_{U(t)} - C_\alpha |u|^2_{H(t)}.$$

In this context the natural weak statement of (2.1) is to find $u \in \mathcal{U} \equiv L^2[0,T;U(.)] \cap H^1[0,T;U'(.)]$ such that

$$\begin{aligned}(u(T),v(T))_{H(T)} &+ \int_0^T \Big(-(u,v_t)_{H(t)} + a(.;u,v)\Big)\\ &= (u_0,v(0))_{H(0)} + \int_0^T \langle F,v\rangle \qquad \forall\, v \in \mathcal{U}.\end{aligned} \tag{2.2}$$

2.2. Properties of $H(t)$. The smoothness assumption 2.1 guarantees that the norms on the pivot spaces $H(t)$ vary continuously with t. The following lemma quantifies this and will be used ubiquitously below.

LEMMA 2.2. *Let w, $z \in H$ and $s \le t$. Then* $\mathrm{e}^{C_\mu(s-t)} \le |z|^2_{H(t)}/|z|^2_{H(s)} \le \mathrm{e}^{C_\mu(t-s)}$ *and*

$$|(w,z)_{H(t)} - (w,z)_{H(s)}| \le (t-s)C_\mu \mathrm{e}^{C_\mu(t-s)} |w|_{H(\xi_1)} |z|_{H(\xi_2)}, \qquad \xi_1,\ \xi_2 \in [s,t].$$

Proof. The differentiability of $(.,.)_{H(.)}$ implies

$$(w,z)_{H(t)} - (w,z)_{H(s)} = \int_s^t \frac{d}{d\xi}(w,z)_{H(\xi)}\, d\xi = \int_s^t \mu(\xi,w,z).$$

Putting $w = z$ gives

$$|z|^2_{H(t)} = |z|^2_{H(s)} + \int_s^t \mu(\xi,z,z)\, d\xi \le |z|^2_{H(s)} + C_\mu \int_s^t |z|^2_{H(\xi)}\, d\xi.$$

Gronwall's inequality then shows $|z|^2_{H(t)} \le |z|^2_{H(s)} \mathrm{e}^{C_\mu(t-s)}$. Since this argument is symmetric in s and t, the first inequality follows.

The second inequality now follows from the intermediate value theorem:

$$|(w,z)_{H(t)} - (w,z)_{H(s)}| \le C_\mu \int_s^t |w|_{H(\xi)} |z|_{H(\xi)}\, d\xi = C_\mu(t-s)|w|_{H(\hat\xi)}|z|_{H(\hat\xi)}$$

for some $\hat\xi \in [s,t]$. Upon introducing a factor of $\mathrm{e}^{C_\mu(t-s)}$ each instance of $\hat\xi$ on the right may be replaced by any $\xi \in [s,t]$. □

3. Discrete spaces.

Discrete subspaces $\mathcal{U}_h$ of $L^2[0,T;U(.)]$ are constructed from a partition $0=t^0<t^1<\cdots<t^N=T$ and a sequence of subspaces $\{U_h^n\}_{n=1}^N$ of U as

$$\mathcal{U}_h=\{u_h\in L^2[0,T;U(.)]\ \mid\ u_h|_{(t^{n-1},t^n]}\in\mathcal{P}_k(t^{n-1},t^n;U_h^n)\}.$$

Projections of H onto the subspaces U_h^n with respect to the norms $H(t)$ appear in the analysis below; the following notation is used to denote these.

Notation 1. *$P_n(t)$ is the projection $P_n(t):H(t)\to U_h^n$ characterized by $P_n(t)u\in U_h^n$, $(P_n(t)u,v_h)_{H(t)}=(u,v_h)_{H(t)}$ for all $v_h\in U_h^n$.*

3.1. Discrete characteristic functions.

Estimates for the solution $u(t)$ of an evolution equation are frequently obtained by multiplying the equation by $\chi_{[0,t)}u$. This choice of test functions is not available in the discrete context unless the terminal time is one of the partition points. To estimate the solution at times $t\in[t^{n-1},t^n)$ we first recall the discrete characteristic functions introduced in [6, section 2.3].

The discrete characteristic functions on each interval are invariant under translation, so it is convenient to work on the interval $[0,\tau)$ with $\tau=t^n-t^{n-1}$. The first step is to consider polynomials $p\in\mathcal{P}_k(0,\tau)$. A discrete approximation of $\chi_{[0,t)}p$ is the polynomial $\hat p\in\{\hat p\in\mathcal{P}_k(0,\tau)|\hat p(0)=p(0)\}$ satisfying

$$\int_0^\tau\hat pq=\int_0^t pq\qquad\forall\,q\in\mathcal{P}_{k-1}(0,\tau).$$

The above construction is motivated by the fact that it is possible to select $q=p'$ to obtain $\int_0^\tau\hat pp'=\int_0^t pp'=(1/2)(p^2(t)-p^2(0))$.

This elementary construction extends to approximate functions of the form $\chi_{[0,t)}v$ for $v_h\in\mathcal{P}_k(0,\tau;V)$, where V is any semi-inner product space. If $v\in\mathcal{P}_k(0,\tau;V)$, it may be written as $v=\sum_{i=0}^k p_i(t)v_i$, where $\{p_i\}\subset\mathcal{P}_k(0,\tau)$ and $\{v_i\}\subset V$. Defining $\hat v=\sum_{i=0}^k\hat p_i(t)v_i$ it is clear that $\hat v\in\mathcal{P}_k(0,\tau;V)$ satisfies

$$\hat v(0)=v(0)\text{ and }\int_0^\tau(\hat v,w)_V=\int_0^t(v,w)_V\qquad\forall\,w\in\mathcal{P}_{k-1}(0,\tau;V).\tag{3.1}$$

We recall the following elementary lemma from [6, Lemma 2.7] which shows that the mapping $v\mapsto\hat v$ is continuous on $\mathcal{P}_k(0,\tau,V)$.

Lemma 3.1. *Let V be a semi-inner product space. Then the mapping*

$$v=\sum_{i=0}^k p_i(t)v_i\mapsto\hat v=\sum_{i=0}^k\hat p_i(t)v_i$$

on $\mathcal{P}_k(0,\tau;V)$ is continuous in $\|\cdot\|_{L^2[0,\tau;V]}$. In particular, there exists $C_k>0$ depending upon only k such that

$$\|\hat v\|_{L^2[0,\tau;V]}\le C_k\|v\|_{L^2[0,\tau;V]}\quad\textit{and}\quad\|\hat v-\chi_{[0,t)}v\|_{L^2[0,\tau;V]}\le C_k\|v\|_{L^2[0,\tau;V]}.$$

Notice that the above construction is purely algebraic in the sense that (3.1) holds for any choice of inner product on V. A major complication encountered with the time-dependent spaces is that the analogous construction with V replaced by $H(t)$ is no longer algebraic; the time dependence of $H(t)$ enters into the definition. In this situation estimates in other spaces, such as $U(t)$, are no longer automatic.

For the implicit evolution problem it is necessary to modify the above construction. If $u \in \mathcal{P}_k(0,\tau;U_h)$, define a discrete approximation $\tilde{u}$ of $\chi_{[0,\tau)}u$ by the following: $\tilde{u} \in \mathcal{P}_k(0,\tau;U_h)$ is the function satisfying

$$\tilde{u}(0) = u(0) \quad \text{and} \quad \int_0^\tau (\tilde{u}, w)_{H(.)} = \int_0^t (u, w)_{H(.)} \qquad \forall\, w \in \mathcal{P}_{k-1}(0,\tau;U_h). \tag{3.2}$$

A slight modification of Lemma 2.4 from [6] can be used to establish bounds for $\tilde{u}$ in $L^2[0,\tau;H(.)]$.

Lemma 3.2. *The mapping $u \mapsto \tilde{u}$ in $\mathcal{P}_k[0,\tau;H(.)]$ is linear and continuous, and there exists a constant C_k depending only upon k such that*

$$\|\tilde{u} - u\|_{L^2[0,\tau;H(.)]} \le C_k \mathrm{e}^{C_\mu \tau} \|u\|_{L^2[t,\tau;H(.)]}.$$

Moreover

$$\|\tilde{u} - \chi_{[0,t)}u\|_{L^2[0,\tau;H(.)]} \le (1 + C_k \mathrm{e}^{C_\mu \tau}) \|u\|_{L^2[t,\tau;H(.)]},$$

and $\|\tilde{u}\|_{L^2[0,\tau;H(.)]} \le (1 + C_k \mathrm{e}^{C_\mu \tau}) \|u\|_{L^2[0,\tau;H(.)]}$

The proof is essentially the same as in [6, Lemma 2.4]; the only difference concerns the scaling argument required to show that C_k can be chosen to be independent of time. To do this Lemma 2.2 is used to bound $\|u - \tilde{u}\|_{L^2[0,\tau,H(.)]}$ by $\|u - \tilde{u}\|_{L^2[0,\tau,H(\tau/2)]}$ to remove the implicit time dependence through $H(t)$.

To bound $\tilde{u}$ in $L^2[0,\tau;U(.)]$, the difference, $\tilde{u} - \hat{u}$, with the algebraic projection $\hat{u}$ is first estimated in the weaker norm $L^2[0,\tau;H(.)]$.

Lemma 3.3. *Let $u \in \mathcal{P}_k(0,\tau,U_h^n)$ and $\tilde{u}$ be the projections defined in* (3.2). *If $\hat{u}$ is the* algebraic *projection characterized by* (3.1), *then*

$$\|\hat{u} - \tilde{u}\|_{L^2[0,\tau;H(.)]} \le C_\mu^{1/2} C(k,\mu) \tau \|u\|_{L^2[0,\tau;H(.)]},$$

where $C(\mu,k)$ is a constant depending on k and μ through C_μ and C_k, the constant in Lemma 3.2.

Proof. In this proof $C(k,\mu)$ denotes a constant depending only on C_k and C_μ which may change from step to step. Recall that $\hat{u} \in \mathcal{P}_k(0,\tau;U_h)$ satisfies $\hat{u}(0) = u(0) = \tilde{u}(0)$,

$$\int_0^\tau (\hat{u}, w)_{H(0)} = \int_0^t (u, w)_{H(0)}, \qquad w \in \mathcal{P}_{k-1}(0,\tau;U_h),$$

and $\|\hat{u}\|_{L^2[0,\tau,H(0)]} \le C_k \|u\|_{L^2[0,\tau;H(0)]}$. If $w \in \mathcal{P}_{k-1}(0,\tau;U_h)$, then

$$\begin{aligned} \int_0^\tau (\tilde{u} - \hat{u}, w)_{H(.)} &= \int_0^t (u, w)_{H(.)} - \int_0^\tau (\hat{u}, w)_{H(.)} \\ &= \int_0^t \Big((u, w)_{H(.)} - (u, w)_{H(0)} \Big) - \int_0^\tau \Big((\hat{u}, w)_{H(.)} - (\hat{u}, w)_{H(0)} \Big). \end{aligned}$$

Since $(\hat{u} - \tilde{u})(0) = 0$ it follows that $(\hat{u} - \tilde{u})(s) = s\bar{u}(s)$, where $\bar{u} \in \mathcal{P}_{k-1}(0,\tau;U_h)$. Putting $w = \bar{u}$ and using Lemma 2.2 to estimate the right-hand side gives

$$\int_0^\tau s|\bar{u}|^2_{H(s)}\, ds \le \int_0^\tau s C_\mu e^{C_\mu s} \big(|u(s)|_{H(0)} + |\hat{u}(s)|_{H(0)} \big) |\bar{u}(s)|_{H(s)}\, ds.$$

An application of the Cauchy–Schwarz inequality then shows

$$\int_0^\tau s|\bar{u}(s)|^2_{H(s)}\,ds \le C_\mu e^{C_\mu \tau}\int_0^\tau s\big(|u(s)|^2_{H(0)} + |\hat{u}(s)|^2_{H(0)}\big)\,ds \le C_\mu C(k,\mu)\tau \int_0^\tau |u|^2_{H(0)}.$$

Lemma 2.2 is used to compare $|.|_{H(s)}$ with the fixed norm $|.|_{H(0)}$ to obtain

$$\int_0^\tau s|\bar{u}|^2_{H(0)}\,ds \le C_\mu C(k,\mu)\tau \int_0^\tau |u|^2_{H(0)} \le C_\mu C(k,\mu)\tau \int_0^\tau |u|^2_{H(.)}.$$

Using the equivalence of norms on $\mathcal{P}_{k-1}(0,\tau)$ and Lemma 2.2 once again to compare $|.|_{H(0)}$ with $|.|_{H(.)}$ gives

$$\int_0^\tau |\tilde{u}-\hat{u}|^2_{H(.)} = \int_0^\tau s^2|\bar{u}|^2_{H(s)}\,ds \le C_\mu C(k,\mu)\tau^2 \int_0^\tau |u|^2_{H(.)}. \qquad \square$$

To estimate $\|\tilde{u}\|_{L^2[0,\tau;U(.)]}$ an inverse hypothesis is used to bound $\|u_h\|_{U(.)}$ by the weaker norm $|u_h|_{H(.)}$ for $u_h \in U_h$. Recall that in the usual finite element context, the constant depends upon the minimum angle in the mesh.

Corollary 3.4. *Define the "inverse hypothesis constant" $C_{inv}(h)$ by*

$$C_{inv}(h) = \max_{0\le n\le N}\ \sup_{u_h\in U_h^n}\ \sup_{t\in(t^{n-1},t^n]} \frac{|u_h|_{U(t)}}{|u_h|_{H(t)}}.$$

Then there exists a constant $C(k,\mu)$ depending only on C_k, and C_μ such that

$$|\tilde{u}|_{L^2[0,\tau;U(.)]} \le C(k,\mu)\Big(C_u^2|u|_{L^2[0,\tau;U(.)]} + C_\mu^{1/2}\tau C_{inv}(h)\|u\|_{L^2[0,\tau;H(.)]}\Big),$$

where C_u and c_u are the constants in Assumption 2 *and $\tau = \max_{1\le n\le N}(t^n - t^{n-1})$.*

Remark 1. When $U \subset H^1(\Omega)$ and $H = L^2(\Omega)$ the inverse inequality states $C_{inv}(h) \le C/h$ for the classical finite element subspaces constructed over quasi-uniform meshes.

Proof. As in the proof of the lemma, let $\hat{u}$ be the algebraic projection. Then

$$\begin{aligned}
|\tilde{u}|_{L^2[0,\tau;U(.)]} &\le |\hat{u}|_{L^2[0,\tau;U(.)]} + |\hat{u}-\tilde{u}|_{L^2[0,\tau;U(.)]}\\
&\le C_u|\hat{u}|_{L^2[0,\tau;U(0)]} + C_{inv}(h)\|\hat{u}-\tilde{u}\|_{L^2[0,\tau;H(.)]}\\
&\le C_kC_u|u|_{L^2[0,\tau;U(0)]} + C_\mu^{1/2}C(k,\mu)\tau C_{inv}(h)\|u\|_{L^2[0,\tau;H(.)]}\\
&\le C(\mu,k)\Big(C_u^2|u|_{L^2[0,\tau;U(.)]} + C_\mu^{1/2}\tau C_{inv}(h)\|u\|_{L^2[0,\tau;H(.)]}\Big). \qquad \square
\end{aligned}$$

4. DG scheme for an auxiliary PDE. We consider approximating implicit parabolic equations of the form

$$(Mu)_t + Bu = f, \qquad u(0) = u_0, \tag{4.1}$$

where $B(.) : U(.) \to U'(.)$ is the operator corresponding to the bilinear form $b(.) : U(.) \times U(.) \to \mathbb{R}$ defined by

$$b(t;u,v) = \eta(u,v)_{U(t)} + C_\mu(u,v)_{H(t)}$$

and the coefficient $\eta \ge 0$ is constant. Since $b(.)$ is symmetric and positive definite it induces a norm on $U(.)$ which is denoted as $\|u\|^2_{B(.)} = b(.;u,u)$. The situation where

η is small is important, and the estimates obtained below are valid when $\eta = 0$, which corresponds to the situation where (4.1) is an ODE in H.

It is assumed that there exists a unique solution $u \in L^2[0,T;U(.)] \cap H^1[0,T;U'(.)] \hookrightarrow C[0,T;H(.)]$ satisfying the associated weak problem

$$\begin{aligned}(4.2)\quad &(u(T),v(T))_{H(T)} + \int_0^T -\langle u, v_t\rangle_{H(.)} + b(.;u,v) \\ &\quad = (u_0, v(0))_{H(0)} + \int_0^T \langle f, v\rangle \qquad \forall\, v \in L^2[0,T;U(.)] \cap H^1[0,T;U'(.)].\end{aligned}$$

Given a partition $0 = t^0 < t^1 < \cdots < t^N = T$ of $[0,T]$ and a collection $\{U_h^n\}_{n=0}^N$ of subspaces of U, the DG method constructs an approximate solution $u_h|_{(t_{n-1},t_n]} \in \mathcal{P}_k(t^{n-1},t^n;U_h^n)$ satisfying

$$\begin{aligned}&(u^n, v^n)_{H(t^n)} + \int_{t^{n-1}}^{t^n} \big(-(u_h, v_{ht})_{H(.)} + b(.;u_h,v_h)\big) - (u^{n-1}, v_+^{n-1})_{H(t^{n-1})} \\ (4.3)\qquad &\qquad = \int_{t^{n-1}}^{t^n} \langle f, v_h\rangle \quad \forall\, v_h \in \mathcal{P}_k(t^{n-1},t^n;U_h^n).\end{aligned}$$

The next definition characterizes the local truncation error in the present context.

Definition 4.1. (1) *The projection* $\mathbb{P}_n^{loc} : C[t^{n-1},t^n;H(.)] \to \mathcal{P}_k(t^{n-1},t^n;U_h^n)$ *satisfies* $(\mathbb{P}_n^{loc}u)^n = P_n(t^n)u(t^n)$, *and*

$$\int_{t^{n-1}}^{t^n} (u - \mathbb{P}_n^{loc}u, v_h)_{H(.)} = 0 \qquad \forall\, v_h \in \mathcal{P}_{k-1}(t^{n-1},t^n;U_h^n).$$

Here we have used the convention $(\mathbb{P}_n^{loc}u)^n \equiv (\mathbb{P}_n^{loc}u)(t^n)$.

(2) *The projection* $\mathbb{P}_h^{loc} : C[0,T;H(.)] \to \mathcal{U}_h$ *satisfies*

$$\mathbb{P}_h^{loc}u \in \mathcal{U}_h \qquad \text{and} \qquad (\mathbb{P}_h^{loc}u)|_{(t^{n-1},t^n]} = \mathbb{P}_n^{loc}(u|_{[t^{n-1},t^n]}).$$

(3) $\mathbb{P}_h : \{u \in C[0,T;H(.)] \mid (Mu) \in H^1[0,T;U'(.)]\} \to \mathcal{U}_h$ *is the discontinuous Galerkin solution of* (4.1) *with* $f = u_t + Bu$ *and initial data* u_0 *specified.*

Remark 2. Notice that $\mathbb{P}_n^{loc}u$ is the solution of the DG approximation of $(M(.)u)' = f$ on $(t^{n-1},t^n]$ with $u(t^{n-1})$ specified as the initial data. It follows that $\sup_{t^{n-1}\le t\le t^n} |u - \mathbb{P}_n^{loc}u|_{H(.)}$ (or related norms) measures the local truncation error of the scheme with $\eta = 0$.

The following theorem estimates the error at the partition points and is the analogue of [6, Theorem 2.2]. We remind the reader that $P_n(t) : H(t) \to U_h^n$ is the projection from $H(t)$ onto the discrete space U_h^n.

Theorem 4.2. *Let Assumptions* 1 *and* 2 *hold, and let* u *and* u_h *satisfy* (4.1) *and* (4.3), *respectively. Then the error* $\hat{e}^n = P_n(t^n)u(t^n) - u^n$ *at the partition points satisfies*

$$\begin{aligned}&(1/2)|\hat{e}^n|_{H(t^n)}^2 + (1/4)\int_0^{t^n} \|\hat{e}\|_{B(.)}^2 + (1/4)\sum_{i=0}^{n-1} |\hat{e}^i - \hat{e}_+^i|_{H(t^{n-1})}^2 \le (1/2)|\hat{e}^0|_{H(0)}^2 \\ &+ \sum_{i=0}^{n-1} \min\Big((C(k,C_u)/\tau^{i+1}\eta)\|P_{i+1}\big(I-P_i\big)u(t^i)\|_{U'(t^i)}^2, |(I-P_i)u(t^i)|_{H(t^i)}^2\Big) \\ &+ \int_0^{t^n} \|(I - \mathbb{P}_h^{loc})u\|_{B(.)}^2,\end{aligned}$$

where the projections in the sum are evaluated at t^i *(*$P_i = P_i(t^i)$ *and* $P_{i+1} = P_{i+1}(t^i)$*).*

Proof. Let $e = u - u_h$ be the total error and note that the Galerkin orthogonality gives

$$(e^n, v^n)_{H(t^n)} + \int_{t^{n-1}}^{t^n} \big(-(e, v_{ht})_{H(.)} + b(.; e, v_h)\big) - (e^{n-1}, v_+^{n-1})_{H(t^{n-1})} = 0.$$

Letting $\hat{e} = \mathbb{P}_n^{loc} u - u_h = e - (I - \mathbb{P}_n^{loc})u$ and using the properties of $\mathbb{P}_n^{loc}$ gives

$$\begin{aligned}(\hat{e}^n, v^n)_{H(t^n)} &+ \int_{t^{n-1}}^{t^n} \Big(-(\hat{e}, v_{ht})_{H(.)} + b(.; \hat{e}, v_h)\Big) - (\hat{e}^{n-1}, v_+^{n-1})_{H(t^{n-1})} \\ &= ((I - P_{n-1})u(t^{n-1}), v_+^{n-1})_{H(t^{n-1})} - \int_{t^{n-1}}^{t^n} b(.; (I - \mathbb{P}_n^{loc})u, v_h).\end{aligned} \tag{4.4}$$

Setting $v_h = \hat{e}$ and using Assumption 2, (4.4) becomes

$$\begin{aligned}(1/2)|\hat{e}^n|_{H(t^n)}^2 &+ \int_{t^{n-1}}^{t^n} \|\hat{e}\|_{B(.)}^2 + (1/2)|\hat{e}^{n-1} - \hat{e}_+^{n-1}|_{H(t^{n-1})}^2 \\ &= (1/2)|\hat{e}^{n-1}|_{H(t^{n-1})}^2 + ((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1})_{H(t^{n-1})} \\ &\quad - \int_{t^{n-1}}^{t^n} (1/2)\mu(.; \hat{e}, \hat{e}) + b(.; (I - \mathbb{P}_n^{loc})u, \hat{e}).\end{aligned} \tag{4.5}$$

The last two terms on the right are bounded using Assumption 2 and the Cauchy–Schwarz inequality; specifically,

$$\begin{aligned}\int_{t^{n-1}}^{t^n} (1/2)\mu(.; \hat{e}, \hat{e}) + b(.; (I - \mathbb{P}_n^{loc})u, \hat{e}) &\le \int_{t^{n-1}}^{t^n} (C_\mu/2)\|\hat{e}\|_{H(.)}^2 + \|(I - \mathbb{P}_n^{loc})u\|_{B(.)}\|\hat{e}\|_{B(.)} \\ &\le \int_{t^{n-1}}^{t^n} (C_\mu/2)\|\hat{e}\|_{H(.)}^2 + \|(I - \mathbb{P}_n^{loc})u\|_{B(.)}^2 \\ &\quad + (1/4)\|\hat{e}\|_{B(.)}^2.\end{aligned}$$

The jump term on the right or (4.5) is bounded two different ways. Since $\hat{e}^{n-1} \in U^{n-1}$ an estimate independent of η is computed as

$$\begin{aligned}((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1})_{H(t^{n-1})} &= ((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1} - \hat{e}^{n-1})_{H(t^{n-1})} \\ &\le |(I - P_{n-1})u(t^{n-1})|_{H(t^{n-1})}^2 \\ &\quad + (1/4)|\hat{e}_+^{n-1} - \hat{e}^{n-1}|_{H(t^{n-1})}^2\end{aligned}$$

(we write $P_{n-1} = P_{n-1}(t^{n-1})$ and similarly $P_n = P_n(t^{n-1})$ below). An alternative estimate is obtained upon writing

$$\begin{aligned}((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1})_{H(t^{n-1})} &= (P_n(I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1})_{H(t^{n-1})} \\ &\le \|P_n(I - P_{n-1})u(t^{n-1})\|_{U'(t^{n-1})}\|\hat{e}_+^{n-1}\|_{U(t^{n-1})}.\end{aligned}$$

The following "inverse" estimate for functions in $\mathcal{P}_k(t^{n-1}, t^n, U^n)$ is used to bound the $\|\hat{e}_+^{n-1}\|_{U(t^{n-1})}$:

$$\|\hat{e}_+^{n-1}\|_{U(t^{n-1})}^2 \le (C_k/\tau^n) \int_{t^{n-1}}^{t^n} \|\hat{e}\|_{U(t^{n-1})}^2 \le (C_k C_u^2/\tau^n) \int_{t^{n-1}}^{t^n} \|\hat{e}\|_{U(.)}^2.$$

The finite dimensionality of $\mathcal{P}_k(t^{n-1}, t^n)$ and a scaling argument shows that the constant C_k appearing in the first inequality depends only upon k. It follows that

$$((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1})_{H(t^{n-1})} \leq (C(C_k, C_u)/\tau^n \eta) \|P_n(I - P_{n-1})u(t^{n-1})\|^2_{U'(t^{n-1})} + (\eta/2) \int_{t^{n-1}}^{t^n} \|\hat{e}\|^2_{U(.)}.$$

Substituting these estimates into (4.5), recalling the definition of $b(.;.,.)$, and summing completes the proof. □

The following theorem compares the (global) solution, $\mathbb{P}_h u$, of the DG scheme with the local projections, $\mathbb{P}_n^{loc} u$, at arbitrary times.

Theorem 4.3. *Let Assumptions* 1 *and* 2 *hold, and let u and u_h satisfy* (4.1) *and* (4.3), *respectively. Let $\hat{e} = \mathbb{P}_h^{loc} u - u_h$. Then there exists a constant*

$$C = C(C_k, C_\mu, C_u, \sqrt{\eta} \tau C_{inv}(h))$$

such that

$$\begin{aligned} &|\hat{e}(t)|^2_{H(t)} + \int_0^{t^n} \Big(\eta \|\hat{e}\|^2_{U(.)} + C_\mu |\hat{e}|^2_{H(.)} \Big) + \sum_{i=0}^{n-1} |[\hat{e}^i]|^2_{H(t^i)} \\ &\leq C \Big(|\hat{e}^0|^2_{H(0)} + \int_0^{t^n} \Big(\eta \|(I - \mathbb{P}_h^{loc})u\|^2_{U(.)} + C_\mu |(I - \mathbb{P}_h^{loc})u|^2_{H(.)} \Big) \\ &\quad + \sum_{i=0}^{n-1} \min \Big(1/(\tau^{i+1}\eta) \|P_{i+1}(I - P_i)u(t^i)\|^2_{U'(t^i)}, |(I - P_i)u(t^i)|^2_{H(t^i)} \Big) \end{aligned}$$

for any time $t \in (t^{n-1}, t^n]$. Here $[\hat{e}^i] = \hat{e}_+^i - \hat{e}^i$ is the jump in $\hat{e}$ at t^i, $\tau = \max_{1<i\leq n}(t^i - t^{i-1})$, and the projections in the sum are evaluated at t^i.

Proof. Given Theorem 4.2 it suffices to bound $|\hat{e}(t)|_{H(t)}$ for $t \in (t^{n-1}, t^n]$. Recall inequality (4.4)

$$\begin{aligned} &(\hat{e}^n, v^n)_{H(t^n)} + \int_{t^{n-1}}^{t^n} \Big(-(\hat{e}, v_{ht})_{H(.)} + b(.; \hat{e}, v_h) \Big) - (\hat{e}^{n-1}, v_+^{n-1})_{H(t^{n-1})} \\ &\quad = ((I - P_{n-1})u(t^{n-1}), v_+^{n-1})_{H(t^{n-1})} - \int_{t^{n-1}}^{t^n} b(.; (I - \mathbb{P}_n^{loc})u, v_h), \end{aligned}$$

and rewrite it as

$$\begin{aligned} &\int_{t^{n-1}}^{t^n} \Big((\hat{e}_{ht}, v_h)_{H(.)} + \mu(.; \hat{e}, v_h) + b(., \hat{e}, v_h) \Big) + (\hat{e}_+^{n-1} - \hat{e}^{n-1}, v_+^{n-1})_{H(t^{n-1})} \\ &\quad = ((I - P_{n-1})u(t^{n-1}), v_+^{n-1})_{H(t^{n-1})} - \int_{t^{n-1}}^{t^n} b(.; (I - \mathbb{P}_n^{loc})u, v_h). \end{aligned}$$

Next let $v_h = \tilde{e}$ be the discrete approximation of $\chi_{[t^{n-1},t)}\hat{e}$ characterized by equation (3.2) to get

$$\begin{aligned} &\int_{t^{n-1}}^{t} (\hat{e}_t, \hat{e})_{H(.)} + \int_{t^{n-1}}^{t^n} \mu(.; \hat{e}, \tilde{e}) + (\hat{e}_+^{n-1} - \hat{e}^{n-1}, \hat{e}_+^{n-1})_{H(t^{n-1})} \\ &= ((I - P_{n-1})u(t^{n-1}), \hat{e}_+^{n-1})_{H(t^{n-1})} - \int_{t^{n-1}}^{t^n} b(.; (I - \mathbb{P}_n^{loc})u, \tilde{e}) + b(.; \hat{e}, \tilde{e}). \end{aligned}$$

The differentiability properties of $(.,.)_{H(.)}$ allow the first term integrated so that

$$\begin{aligned}(1/2)|\hat{e}(t)|^2_{H(t)} &- (1/2)\int_{t^{n-1}}^{t} \mu(.;\hat{e},\hat{e}) + \int_{t^{n-1}}^{t^n} \mu(.;\hat{e},\tilde{e})\\ &+ (1/2)|\hat{e}^{n-1}_{+} - \hat{e}^{n-1}|^2_{H(t^{n-1})} - (1/2)|\hat{e}^{n-1}|^2_{H(t^{n-1})}\\ &= ((I-P_{n-1})u(t^{n-1}), \hat{e}^{n-1}_{+})_{H(t^{n-1})}\\ &\quad - \int_{t^{n-1}}^{t^n} b(.;(I-\mathbb{P}^{loc}_n)u,\tilde{e}) + b(.;\hat{e},\tilde{e}).\end{aligned}$$

Use Lemma 3.2 and Corollary 3.4 to bound $\tilde{e}$ in terms of $\hat{e}$; specifically,

- Using Lemma 3.2 shows

$$\begin{aligned}\int_{t^{n-1}}^{t^n} \mu(.;\hat{e},\tilde{e}) &\le C_\mu \|\hat{e}\|_{L^2[t^{n-1},t^n;H(.)]} \|\tilde{e}\|_{L^2[t^{n-1},t^n;H(.)]}\\ &\le C_\mu(1 + C_k e^{C_\mu \tau^n}) \|\hat{e}\|^2_{L^2[t^{n-1},t^n;H(.)]}\end{aligned}$$

so that

$$(1/2)\int_{t^{n-1}}^{t} \mu(.;\hat{e},\hat{e}) - \int_{t^{n-1}}^{t^n} \mu(.;\hat{e},\tilde{e}) \le C_\mu C(k,\mu) \int_{t^{n-1}}^{t^n} |\hat{e}|^2_{H(.)}.$$

- Expanding the definition of $b(.;.,.)$ and recalling Corollary 3.4 shows

$$\begin{aligned}&\int_{t^{n-1}}^{t^n} b(.;\hat{e},\tilde{e}) \le (1/2)\int_{t^{n-1}}^{t^n} \|\hat{e}\|^2_{B(.)} + \Big(\eta\|\tilde{e}\|^2_{U(.)} + C_\mu|\tilde{e}|^2_{H(.)}\Big)\\ &\le (1/2)\int_{t^{n-1}}^{t^n} \|\hat{e}\|^2_{B(.)}\\ &\quad + C(k,\mu)\Big(\eta\big(C_u^4\|\tilde{e}\|^2_{U(.)} + C_\mu \tau^2 C_{inv}(h)^2 \|\hat{e}\|^2_{H(.)}\big) + C_\mu\|\hat{e}\|^2_{H(.)}\Big)\\ &\le (1/2)C(k,\mu,u,\sqrt{\eta}\tau C_{inv}(h)) \int_{t^{n-1}}^{t^n} \|\hat{e}\|^2_{B(.)}.\end{aligned}$$

- Similarly

$$\begin{aligned}\int_{t^{n-1}}^{t^n} b(.;(I-\mathbb{P}^{loc}_n)u,\tilde{e}) \le (1/2)\int_{t^{n-1}}^{t^n} &\|(I-\mathbb{P}^{loc}_n)u\|^2_{B(.)}\\ &+ C(k,\mu,u,\sqrt{\eta}\tau C_{inv}(h))\|\hat{e}\|^2_{B(.)}.\end{aligned}$$

It follows that

$$\begin{aligned}(1/2)|\hat{e}(t)|^2_{H(t)} &+ (1/2)|\hat{e}^{n-1}_{+} - \hat{e}^{n-1}|^2_{H(t^{n-1})} - (1/2)|\hat{e}^{n-1}|^2_{H(t^{n-1})}\\ &\le ((I-P_{n-1})u(t^{n-1}), \hat{e}^{n-1}_{+})_{H(t^{n-1})}\\ &\quad + \int_{t^{n-1}}^{t^n} \|(I-\mathbb{P}^{loc}_n)u\|^2_{B(.)} + C(k,\mu,u,\sqrt{\eta}\tau C_{inv}(h))\|\hat{e}\|^2_{B(.)}.\end{aligned} \tag{4.6}$$

As in the proof of Theorem 4.2 the first term on the right can be bounded by

$$|(I-P_{n-1})u(t^{n-1})|^2_{H(t^{n-1})} + (1/4)|\hat{e}^{n-1}_{+} - \hat{e}^{n-1}|^2_{H(t^{n-1})}$$

or

$$C(k,u)/(\tau^n\eta)\|P_n(I-P_{n-1})u(t^{n-1})\|^2_{U'(t^{n-1})}+(\eta/2)\int_{t^{n-1}}^{t^n}\|\hat{e}\|^2_{U(.)},$$

so

$$|\hat{e}(t)|^2_{H(t)}\leq|\hat{e}^{n-1}|^2_{H(t^{n-1})}+\int_{t^{n-1}}^{t^n}C(k,\mu,u,\sqrt{\eta}\tau C_{inv}(h))\|\hat{e}\|^2_{B(.)}+\|(I-\mathbb{P}_n^{loc})u\|^2_{B(.)}$$
$$+2\max\Big(C(k,u)/(\tau^n\eta)\|P_n(I-P_{n-1})u(t^{n-1})\|^2_{U'(t^{n-1})},\,|(I-P_{n-1})u(t^{n-1})|^2_{H(t^{n-1})}\Big).$$

The theorem then follows upon using Theorem 4.2 to bound the first two terms on the right. ☐

5. DG scheme for implicit parabolic equations.

In this section DG approximations of (2.2) are considered. It is shown that the error for the parabolic PDE can be bounded by the error of the DG approximation of the auxiliary equation introduced in the previous section.

5.1. DG scheme.

To approximate the solution of the weak formulation (2.2), let $0=t^0<t^1<\cdots<t^N=T$ be a partition of $[0,T]$, and on each partition construct a closed subspace $U_h^n\subset U$. The discontinuous Galerkin approximates the solution of (2.2) on $(t^{n-1},t^n]$ by $u_h\in\mathcal{P}_k(t^{n-1},t^n;U_h^n)$ satisfying

$$(u^n,v^n)_{H(t^n)}+\int_{t^{n-1}}^{t^n}\Big(-(u_h,v_{ht})_{H(t)}+a(.;u_h,v_h)\Big)$$
$$-(u^{n-1},v_+^{n-1})_{H(t^{n-1})}=\int_{t^{n-1}}^{t^n}\langle F,v_h\rangle\quad\forall v_h\in\mathcal{P}_k(t^{n-1},t^n;U_h^n).\tag{5.1}$$

The stability and error estimates are established using very similar arguments; for this reason we will just focus on the error estimate. The Galerkin orthogonality condition shows that the error $e=u-u_h$ satisfies

$$(e^n,v^n)_{H(t^n)}+\int_{t^{n-1}}^{t^n}\Big(-(e,v_{ht})_{H(t)}+a(.;e,v_h)\Big)-(e^{n-1},v_+^{n-1})_{H(t^{n-1})}=0\tag{5.2}$$

for all $v_h\in\mathcal{P}_k(t^{n-1},t^n;U_h^n)$. Decompose the error as $e=e_p+e_h\equiv(u-\mathbb{P}_hu)+(\mathbb{P}_hu-u_h)$, where $\mathbb{P}_h:\{u\in C[0,T;H(.)]\mid M(.)u\in H^1[0,T;U'(.)]\}\to\mathcal{U}_h$ is the projection introduced in Definition 4.1. The orthogonality condition (5.2) becomes

$$(e_h^n,v^n)_{H(t^n)}+\int_{t^{n-1}}^{t^n}\Big(-(e_h,v_{ht})_{H(.)}+a(.;e_h,v_h)\Big)-(e_h^{n-1},v_+^{n-1})_{H(t^{n-1})}$$
$$=-(e_p^n,v^n)_{H(t^n)}+\int_{t^{n-1}}^{t^n}(e_p,v_{ht})_{H(.)}+(e_p^{n-1},v_+^{n-1})_{H(t^{n-1})}$$
$$-\int_{t^{n-1}}^{t^n}a(.;e_p,v_h).$$

By construction, $\mathbb{P}_hu$ is the discontinuous Galerkin approximation of (4.1), so e_p satisfies the orthogonality condition (4.4). It follows that the first three terms of the right-hand side simplify to $-\int_{t^{n-1}}^{t^n} \eta(e_p, v_h)_{U(.)} + C_\mu(e_p, v_h)_{H(.)}$ so that (with $\eta = c_a$)

$$(e_h^n, v^n)_{H(t^n)} + \int_{t^{n-1}}^{t^n} (-(e_h, v_{ht})_{H(.)} + a(.; e_h, v_h)) - (e_h^{n-1}, v_+^{n-1})_{H(t^{n-1})}$$

$$= -\int_{t^{n-1}}^{t^n} \big(a(\cdot; e_p, v_h) + c_a(e_p, v_h)_{U(.)} + C_\mu(e_p, v_h)_{H(.)}\big). \tag{5.3}$$

A preliminary estimate satisfied by the error at the partition points is obtained by setting $v_h = e_h$ and using the coercivity of $a(.; ., .)$:

$$\begin{aligned}\frac{1}{2}|e_h^n|^2_{H(t^n)} &+ c_\alpha \int_{t^{n-1}}^{t^n} |e_h|^2_{U(.)} + \frac{1}{2}|e_{h+}^{n-1} - e_h^{n-1}|^2_{H(t^{n-1})} \\ &\le \frac{1}{2}|e_h^{n-1}|^2_{H(t^{n-1})} - \int_{t^{n-1}}^{t^n} \Big((1/2)\mu(\cdot; e_h, e_h) + a(.; e_p, e_h) + c_a(e_p, e_h)_{U(.)} \\ &\quad + C_\mu(e_p, e_h)_{H(.)} - C_\alpha |e_h|^2_{H(.)}\Big).\end{aligned} \tag{5.4}$$

Using Assumption 2 on the continuity of $a(.; ., .)$ and $\mu(.; ., .)$ shows

$$\begin{aligned}&|e_h^n|^2_{H(t^n)} + c_\alpha \int_{t^{n-1}}^{t^n} |e_h|^2_{U(.)} + |e_h^{n-1} - e_{h+}^{n-1}|^2_{H(t^{n-1})} \le |e_h^{n-1}|^2_{H(t^{n-1})} \\ &+ \int_{t^{n-1}}^{t^n} \Big((1 + 2c_a/c_\alpha)\big(2c_a|e_p|^2_{U(.)} + (2C_a + C_\mu)|e_p|^2_{H(.)}\big) + 2(C_\alpha + C_\mu + C_a)|e_h|^2_{H(.)}\Big).\end{aligned} \tag{5.5}$$

The inequalities $c_\alpha \le c_a \le C_a$ were used to derive expressions on the right that are independent of the coercivity constant.

Notice that the last term on the right-hand side involves $|e_h(s)|_{H(s)}$ at times $s \in (t^{n-1}, t^n)$, so the discrete Gronwall inequality is not applicable. Below the discrete characteristic functions developed in section 3.1 are used to obtain an expression similar to the above with $\sup_{t^{n-1} \le s \le t^n} |e_h(s)|_{H(s)}$ on the left-hand side.

Remark 3. One way to circumvent this problem is to bound temporal derivatives of the solution [7, 28] so that $e_h(s)$, $s \in (t^{n-1}, t^n)$, can be controlled by the values at the partition points. Bounds on the temporal derivatives of the discrete solution are frequently obtained by assuming A to be self-adjoint. Clearly this line of argument fails for solutions having minimal regularity.

An alternative approach developed in [16] and [2] is to construct a discrete approximation of $(e_h(t) - e_h^{n-1})/(t - t^{n-1})$ in $\mathcal{P}_k(t^{n-1}, t^n; U_h)$. A formal calculation with $v(t) = (e_h(t) - e_h^{n-1})/(t - t^{n-1} + \epsilon)$ shows

$$\begin{aligned}\int_{t^{n-1}}^{t^n} (e_h', v)_H + (e_{h+}^{n-1} - e_h^{n-1}, v_+^{n-1})_H &= \int_{t^{n-1}}^{t^n} (e_h', v)_H + \epsilon|v_+^{n-1}|^2_H \\ &= (1/2)\int_{t^{n-1}}^{t^n} (e_h - e_h^{n-1})^2/(t - t^{n-1} + \epsilon)^2 \\ &\quad + (1/2)|e_h^n - e_h^{n-1}|^2_H/(t^n - t^{n-1} + \epsilon) + (1/2)\epsilon|v_+^{n-1}|^2_H,\end{aligned}$$

which was used to bound $\|e_h\|_{L^2(t^{n-1}, t^n)}$.

5.2. Error estimate. We are now ready to state and prove our main error estimate for the DG approximation of (1.1).

THEOREM 5.1. *Let $U(.) \hookrightarrow H \hookrightarrow U'(.)$ be a dense embedding of Hilbert spaces satisfying Assumption* 1. *Assume each norm $\|.\|_{U(.)}$ is equivalent to $\|.\|_U$, and let $\mathcal{U}_h$ be the subspace of $L^2[0,T;U]$ defined in section* 3. *Let the bilinear form $a : U(.) \times U(.) \to \mathbb{R}$ and the linear form $F : U(.) \to \mathbb{R}$ satisfy Assumption* 2. *Let $u \in \{u \in C[0,T;H(.)] \mid M(.)u \in H^1[0,T;U'(.)]\}$ be the solution of* (2.1) *and $u_h \in \mathcal{U}_h$ be the approximate solution computed using the discontinuous Galerkin scheme* (5.1) *on the partition $0 = t^0 < t^1 < \cdots < t^N = T$, and set $\tau \equiv \max_n t^n - t^{n-1}$.*

Then there exist constants $C > 0$ and $0 < \lambda < 1$ depending only on k (through the constant C_k of Lemma 3.2*), the constants C_a, C_α, C_μ, C_u, c_u, the ratio c_a/c_α, and the product $\sqrt{c_a}\tau C_{inv}(h)$ (defined in Corollary* 3.4*) such that*

$$\begin{aligned}(1-\lambda)|e_h^n|^2_{H(t^n)} + \lambda \sup_{0\le s\le t^n} |e_h(s)|^2_{H(s)} + \sum_{i=0}^{n-1} \mathrm{e}^{C(t^{n-1}-t^i)} |e_h^i - e_{h+}^i|^2_{H(t^i)} \\ + (1-\lambda)\frac{c_\alpha}{2}\int_0^{t^n} \mathrm{e}^{C(t-s)} |e_h(s)|^2_{U(s)}\, ds \\ \le \big(1 + T\mathcal{O}(\tau)\big)\Big(\mathrm{e}^{Ct^n}|e_h^0|^2_{H(0)} \\ + C\lambda \int_0^{t^n} \mathrm{e}^{C(t-s)}\big(c_a |e_p(s)|^2_{U(s)} + (C_a + C_\mu)|e_p(s)|^2_{H(s)}\big)\, ds\Big),\end{aligned}$$

provided $C\tau < 1$. Here $e_h = \mathbb{P}_h u - u_h$ and $e_p = u - \mathbb{P}_h u$, where $\mathbb{P}_h$ is the projection defined in Definition 4.1.

Proof. Rewrite (5.3) as

$$\begin{aligned}\int_{t^{n-1}}^{t^n} \Big((e_{ht}, v_h)_{H(.)} + \mu(.;e_h,v_h) + a(.;e_h,v_h)\Big) + \big(e_{h+}^{n-1} - e_h^{n-1}, v_+^{n-1}\big)_{H(t^{n-1})} \\ = -\int_{t^{n-1}}^{t^n} a(.;e_p,v_h) + b(.;e_p,v_h),\end{aligned}$$

where $b(t;u,v) = c_a(u,v)_{U(t)} + C_\mu(u,v)_{H(t)}$. Set $v_h = \tilde{e}_h$, where $\tilde{e}_h$ is the discrete approximation of $\chi_{[t^{n-1},t)} e_h$ constructed in section 3.1, to obtain

$$\begin{aligned}\int_{t^{n-1}}^{t} (e_{ht}, e_h)_{H(.)} + \big(e_{h+}^{n-1} - e_h^{n-1}, e_{h+}^{n-1}\big)_{H(t^{n-1})} \\ = -\int_{t^{n-1}}^{t^n} \Big(\mu(.;e_h,\tilde{e}_h) + a(.;e_h,\tilde{e}_h) + a(.;e_p,\tilde{e}_h) + b(.;e_p,\tilde{e}_h)\Big).\end{aligned}$$

Recalling that $(e_{ht}, e_h)_{H(.)} = (1/2)(d/dt)|e_h|^2_{H(.)} - (1/2)\mu(.;e_h,e_h)$ shows

$$\begin{aligned}\frac{1}{2}|e_h(t)|^2_{H(t)} + \frac{1}{2}|e_h^{n-1} - e_{h+}^{n-1}|^2_{H(t^{n-1})} - \frac{1}{2}|e_h^{n-1}|^2_{H(t^{n-1})} = \int_{t^{n-1}}^{t} \frac{1}{2}\mu(.;e_h,e_h) \\ - \int_{t^{n-1}}^{t^n} \Big(a(.;e_p,\tilde{e}_h) + a(.;e_h,\tilde{e}_h) + b(.;e_p,\tilde{e}_h)_{U(.)} + \mu(.,e_h,\tilde{e}_h)\Big).\end{aligned}$$

Estimating the right-hand side using Lemma 3.2 and Corollary 3.4, as in the derivation of (4.6), gives

$$\begin{aligned} |e_h(t)|^2_{H(t^n)} + |e_h^{n-1} - e_{h+}^{n-1}|^2_{H(t^{n-1})} &\leq |e_h^{n-1}|^2_{H(t^{n-1})} \\ &+ \int_{t^{n-1}}^{t^n} \Big(c_a |e_p|^2_{U(.)} + (C_a + C_\mu)|e_p|^2_{H(.)} \\ &+ c_\alpha C\big(C_k, C_\mu, C_u, c_a/c_\alpha\big)|e_h|^2_{U(.)} + C(...)|e_h|^2_{H(.)}\Big). \end{aligned} \tag{5.6}$$

Here $C(...)$ is a constant depending upon $C_a, C_k, C_\mu, \sqrt{c_a}\tau C_{inv(h)}$, and c_a/c_α. The remainder of the proof parallels the proof of [6, Theorem 3.1]. Specifically, construct the convex combination of $(1-\lambda)$ from (5.5) and λ from (5.6), and choose the coefficient, λ, so that the term involving $|e_h(t)|^2_{U(.)}$ on the right-hand side of (5.6) is dominated by the corresponding term on the left-hand side of (5.6). Setting

$$\lambda C\big(C_k, C_\mu, C_u, c_a/c_\alpha\big) = (1/2)(1-\lambda) \qquad \text{or} \qquad \lambda = \frac{1}{1 + 2C\big(C_k, C_\mu, C_u, c_a/c_\alpha\big)}$$

leads to an estimate of the form

$$\begin{aligned} &(1-\lambda)|e_h^n|^2_{H(t^n)} + \lambda|e_h(t)|^2_{H(.)} + (1-\lambda)\frac{c_\alpha}{2}\int_{t^{n-1}}^{t^n} |e_h|^2_{U(.)} + |e_h^{n-1} - e_{h+}^{n-1}|^2_{H(t^{n-1})} \\ &\leq |e_h^{n-1}|^2_{H(t^{n-1})} + C(...)\lambda \int_{t^{n-1}}^{t^n} \Big(c_a|e_p|^2_{U(.)} + (C_a + C_\mu)|e_p|^2_{H(.)} + |e_h|^2_{H(.)}\Big), \end{aligned}$$

where $C(...)$ may now depend additionally upon C_α. Bound the first and last terms on the right by

$$|e_h^{n-1}|^2_{H(t^n)} \leq (1-\lambda)|e_h^{n-1}|^2_{H(t^{n-1})} + \lambda \sup_{t^{n-2} < s \leq t^{n-1}} |e_h(s)|^2_{H(s)}$$

and

$$\int_{t^{n-1}}^{t^n} |e_h|^2_{H(.)} \leq \tau^n \sup_{t^{n-1} < s \leq t^n} |e_h(s)|^2_{H(s)}, \qquad \tau^n \equiv t^n - t^{n-1},$$

respectively, and select the time t on the left so that $|e_h(t)|_{H(s)} = \sup_{t^{n-1} < s \leq t^n} |e_h(s)|_{H(s)}$ to get

$$\begin{aligned} (1-\lambda)|e_h^n|^2_{H(t^n)} &+ \lambda(1 - C(...)\tau^n) \sup_{t^{n-1} < s \leq t^n} |e_h(t)|^2_{H(s)} \\ &\qquad + (1-\lambda)\frac{c_\alpha}{2}\int_{t^{n-1}}^{t^n} |e_h|^2_{U(.)} + |e_h^{n-1} - e_{h+}^{n-1}|^2_{H(t^{n-1})} \\ &\leq (1-\lambda)|e_h^{n-1}|^2_{H(t^{n-1})} + \lambda \sup_{t^{n-2} < s \leq t^{n-1}} |e_h(s)|^2_{H(s)} \\ &\qquad + C(...)\lambda \int_{t^{n-1}}^{t^n} (c_a|e_p|^2_{U(.)} + (C_a + C_\mu)|e_p|^2_{H(.)}). \end{aligned}$$

Upon introducing a factor $(1 - C(...)\tau^n)$ in front of the first term, this inequality takes the form

$$(1 - C(...)\tau^n)\alpha^n + \beta^n \leq \alpha^{n-1} + f^n,$$

and the theorem follows from the discrete Gronwall inequality. □

This theorem enables error estimates to be established in various norms. Defining

$$|||e|||_\infty^2 = \sup_{0\le s\le T} |e(s)|_{H(s)}^2 + c_\alpha \int_0^T \mathrm{e}^{C(T-s)} |e(s)|_{U(s)}^2 \, ds$$

and

$$|||e|||_2^2 = \int_0^T \mathrm{e}^{C(T-s)} |e(s)|_{H(s)}^2 \, ds + c_a \int_0^T \mathrm{e}^{C(T-s)} |e(s)|_{U(s)}^2 \, ds,$$

Theorem 4.3 shows that

$$\begin{aligned} |||\mathbb{P}_h u - \mathbb{P}_h^{loc} u|||_\infty^2 \le C \Bigg(& |||u - \mathbb{P}_h^{loc}|||_2^2 \\ & + \sum_{i=0}^{n-1} \min\left(|(I-P_i)u(t^i)|_{H(t^i)}^2, 1/(\tau^{i+1} c_a) \|P_{i+1}(I-P_i)u(t^i)\|_{U'(t^i)}^2 \right) \Bigg). \end{aligned}$$

Here the initial data and constant $\eta \ge 0$ have been selected as $\mathbb{P}_h(0) = P_0(u_0)$ and $\eta = c_a$. Similarly, Theorem 5.1 states

$$\begin{aligned} |||\mathbb{P}_h u - u_h|||_\infty^2 &\le C\Big(|P_0(u_0) - u_h^0|_{H(0)} + |||u - \mathbb{P}_h u|||_2^2 \Big) \\ &\le C\Big(|P_0(u_0) - u_h^0|_{H(0)} + |||u - \mathbb{P}_h^{loc} u|||_2^2 + |||\mathbb{P}_h u - \mathbb{P}_h^{loc}|||_2^2 \Big). \end{aligned}$$

Combining these estimates gives an estimate for the (total) error of the solution.

Theorem 5.2. *Under the assumptions of Theorem* 5.1, *there exists a positive constant $C > 0$ depending only on T, k (through the constant C_k of Lemma* 3.2*), the constants C_a, C_α, C_μ, C_u, c_u, the ratio c_a/c_α, and the product $\sqrt{c_a}\tau C_{inv}(h)$ (defined in Corollary* 3.4*) such that the following estimate holds:*

$$\begin{aligned} |||u - u_h|||_\infty^2 \le C \Bigg(& |P_0(u_0) - u_h^0|_{H(0)} + |||u - \mathbb{P}_h^{loc} u|||_\infty^2 \\ & + \sum_{i=0}^{N-1} \min\left(|(I-P_i)u(t^i)|_{H(t^i)}^2, 1/(\tau^{i+1} c_a) \|P_{i+1}(I-P_i)u(t^i)\|_{U'(t^i)}^2 \right) \Bigg), \end{aligned}$$

where $\mathbb{P}_h^{loc} u$ is the local projection defined in Definition 4.1 *and $P_n(t) : H(t) \to U_h^n$ is the orthogonal projection. A similar estimate also holds with $|||.|||_2$ in place of $|||.|||_\infty$.*

Proof. Using the triangle inequality compute

$$|||u - u_h|||_\infty \le |||u - \mathbb{P}_h^{loc} u|||_\infty + |||\mathbb{P}_h u - \mathbb{P}_h^{loc} u|||_\infty + |||\mathbb{P}_h u - u_h|||_\infty.$$

The estimate now follows from the estimates stated prior to the theorem and the inequality $|||.|||_2 \le C(T) |||.|||_\infty$. □

6. Example: Diffusion on a manifold. This section illustrates how our results apply to the example presented in the introduction. The bilinear forms associated with this problem are

$$(u, v)_{H(.)} = (M(.)u, v)_{L^2(\mathcal{S}_r)} = \int_{\mathcal{S}_r} uv \, J$$

and

$$a(.;u,v) = \int_{\mathcal{S}_r} \Big(\sigma(\nabla v)^T (F^T F)^{-1} \nabla u - (I - \mathbf{n} \otimes \mathbf{n}) \cdot (\nabla_x \mathbf{V}) uv \Big) J.$$

Here $\mathcal{S}_r \subset \mathbb{R}^3$ is homeomorphic to the sphere, and $\mathcal{S}(t) = x(t, \mathcal{S}_r)$, where $x : [0,T] \times \mathcal{S}_r \to \mathbb{R}^3$, determines the present position of a membrane. If $X = (X_1, X_2)$ locally parameterizes the $\mathcal{S}_r$, then $F_{i\alpha} = \partial x_i / \partial X_\alpha$, $J = \sqrt{\det(F^T F)}$, and we write ∇u for $\nabla_X u$. The matrix $F = F(t,X)$ is related to the velocity by $F_t = (\nabla_x \mathbf{V}) F$.

Notice that $(.,.)_{H(.)}$ and the principle part of $a(.;.,.)$ are intrinsic in the sense that

$$\int_{\mathcal{S}_r} uv\, J = \int_{\mathcal{S}(t)} \tilde{u}\tilde{v}, \qquad \text{and} \qquad \int_{\mathcal{S}_r} \sigma(\nabla v)^T . (F^T F)^{-1} \nabla u\, J = \int_{\mathcal{S}(t)} \sigma \operatorname{grad}(\tilde{u}).\operatorname{grad}(\tilde{v}),$$

where $\tilde{u} : \mathcal{S}(t) \to \mathbb{R}$ is related to $u : \mathcal{S}_r \to \mathbb{R}$ by the change of variables $u(t,X) = \tilde{u}(t, x(t,X))$. As indicated above, geometric properties of $\mathcal{S}(t)$, such as the first fundamental form $F^T F$, are determined by the ambient velocity field $\mathbf{V}$; for example, the unit normal $\mathbf{n}$ to $\mathcal{S}(t)$ satisfies

$$\mathbf{n}_t = \left(\mathbf{n}^T (\nabla_x \mathbf{V}) \mathbf{n}\, I - (\nabla_x \mathbf{V})^T\right) \mathbf{n}.$$

The columns of F form a basis for the tangent space of $\mathcal{S}(t)$, and if they become parallel at time $\tilde{t}$, the first fundamental form becomes singular and the hypotheses in Assumption 2 fail. In this situation it is necessary to redefine and triangulate a new reference configuration. The natural choice for the new reference configuration is $\tilde{\mathcal{S}}_r = \mathcal{S}(\tilde{t})$. Since the bilinear forms are intrinsic to the manifold, such a redefinition is covered by our theory in the sense that if $\mathcal{T}_r$ is the triangulation of $\mathcal{S}_r$ at time $\tilde{t}$, then $\{x(\tilde{t}, K) \mid K \in \mathcal{T}_r\}$ becomes the triangulation of $\tilde{\mathcal{S}}_r$ at $\tilde{t}_-$ and the (new) triangulation $\tilde{\mathcal{T}}_r$ of $\tilde{\mathcal{S}}_r$ is the triangulation at $\tilde{t}_+$. The coefficients J, etc., are continuous in the sense that

$$(u,v)_{H(\tilde{t}_-)} = \int_{\mathcal{S}_r} uv\, J = \int_{\tilde{\mathcal{S}}_r} \tilde{u}\tilde{v}\, \tilde{J} = (\tilde{u}, \tilde{v})_{H(\tilde{t}_+)},$$

where $\tilde{J}(\tilde{t}, .) = 1$ and $u(\tilde{t}, X) = \tilde{u}(\tilde{t}, x(\tilde{t}, X))$. Of course the mesh $\tilde{\mathcal{T}}_r$ should be adapted so that the interpolant $\tilde{u}(\tilde{t}_+, .)$ of $\tilde{u}(\tilde{t}_-, .)$ gives a good approximation.

To verify that $M(.)$ satisfies the hypothesis of Assumption 1, recall that

$$J_t = J(I - \mathbf{n} \otimes \mathbf{n}) \cdot (\nabla_x \mathbf{V}) \qquad \text{or} \qquad \ln(J)_t = (I - \mathbf{n} \otimes \mathbf{n}) \cdot (\nabla_x \mathbf{V}).$$

It follows that $\ln(J)_t \leq 2\|\nabla_x \mathbf{V}\|$, where $\|.\|$ is the Frobenius norm, so if $0 < c_0 \leq J(0,.) \leq C_0$, then

$$c_0 e^{-Ct} \leq J(t,.) \leq C_0 e^{Ct}, \qquad \text{where} \qquad C = 2\|\nabla_x \mathbf{V}\|_{L^\infty}. \tag{6.1}$$

We verify that this problem satisfies the hypotheses of Assumption 2 when the seminorm $|.|_{U(t)}$ is defined by

$$|u|^2_{U(.)} = \int_{\mathcal{S}_r} \sigma(\nabla u)^T (F^T F)^{-1} \nabla u\, J.$$

1. Smoothness of $M(t)$: For this example
$$\mu(.;u,v) = \int_{\mathcal{S}_r} uvJ_t = \int_{\mathcal{S}_r} uv(I - \mathbf{n}\otimes\mathbf{n})\cdot(\nabla_x\mathbf{V})$$
so $C_\mu = 2\|\nabla_x\mathbf{V}\|_{L^\infty}$.
2. Equivalence of norms on $U(t)$: If F_1 and F_2 are the columns of F, $F = [F_1, F_2]$, then
$$|u|_{U(.)} = \int_{\mathcal{S}_r} \sigma(\nabla u)^T (F^T F)^{-1} \nabla u\, J = \int_{\mathcal{S}_r} \sigma(\nabla u)^T \frac{1}{J}\begin{bmatrix} |F_2|^2 & -F_1.F_2 \\ -F_1.F_2 & |F_1|^2 \end{bmatrix} \nabla u.$$
Letting $\cos(\theta) = F_1.F_2/|F_1||F_2|$ and $x \in \mathbb{R}^2$, then
$$\begin{aligned}(1-\cos(\theta))\Big((x_1|F_2|)^2 + (x_2|F_2|)^2\Big) &\le x^T \begin{bmatrix} |F_2|^2 & -F_1.F_2 \\ -F_1.F_2 & |F_1|^2 \end{bmatrix} x \\ &\le (1+\cos(\theta))\Big((x_1|F_2|)^2 + (x_2|F_2|)^2\Big).\end{aligned}$$
Since $F_{it} = (\nabla_x\mathbf{V})F_i$ it follows that $\ln(|F_i|)_t \le \|\nabla_x\mathbf{V}\|_{\ell^2}$, so if $0 \le c_0 \le |F_i(0,.)| \le C_0$, a calculation shows that for $s < t$
$$\frac{1-\|\cos(\theta)\|_{L^\infty}}{1+\|\cos(\theta)\|_{L^\infty}} c_0 e^{-C(t-s)} \le \frac{|u|_{U(t)}}{|u|_{U(s)}} \le \frac{1+\|\cos(\theta)\|_{L^\infty}}{1-\|\cos(\theta)\|_{L^\infty}} C_0 e^{C(t-s)}, \tag{6.2}$$
where C is a small multiple of $\|\nabla_x\mathbf{V}\|$ and it is assumed that $\|\cos(\theta)\|_{L^\infty} < 1$. Notice that $\cos(\theta)$ measures the amount of shear the membrane $\mathcal{S}(t)$ experiences when deformed from the reference configuration $\mathcal{S}_r$ and that $|(\cos(\theta))_t| \le C\|\nabla_x\mathbf{V}\|_{L^\infty}$. If the membrane is elastic, it will resist shear, so $|\cos(\theta)|$ will typically be bounded away from one. Examples of flow computations with elastic membranes can be found in [25].
3. Continuity of the bilinear form and data:
$$\begin{aligned}|a(t;u,v)| &\le \sigma\|u\|_{U(t)}\|v\|_{U(t)} + 2\|\nabla_x\mathbf{V}\|_{L^\infty}\|u\|_{H(t)}\|v\|_{H(t)} \\ &\le \Big(c_a|u|^2_{U(t)} + C_a|u|^2_{H(t)}\Big)^{1/2}\Big(c_a|v|^2_{U(t)} + C_a|v|^2_{H(t)}\Big)^{1/2},\end{aligned}$$
where $c_a = \sigma$ and $C_a = 2\|\nabla_x\mathbf{V}\|_{L^\infty}$.
Since the equation is homogeneous ($f = 0$), the continuity hypothesis on f is trivially satisfied.
4. Coercivity of the bilinear form: If $c_\alpha = c_a = \sigma$ and $C_\alpha = C_a = 2\|\nabla_x\mathbf{V}\|_{L^\infty}$, then the bilinear form satisfies
$$a(t;u,u) \ge c_\alpha|u|^2_{U(t)} - C_\alpha|u|^2_{H(t)}.$$

Assuming the existence of constants $0 < c_0 < C_0$ introduced above and assuming that $\|\cos(\theta)\|_{L^\infty} < C_1 < 1$, the approximate solutions of (1.2) computed using the discontinuous Galerkin scheme (4.3) will satisfy the error estimates stated in Theorems 5.1 and 5.2. The estimates in (6.1) and (6.2) show that the norms $\|.\|_{H(.)}$ and $|.|_{U(.)}$ are equivalent to unweighted L^2 and H^1 norms scaled appropriately. If classical Lagrange finite elements with polynomials of degree $\ell > 0$ are used to construct subspaces of $U \sim H^1(\mathcal{S}_r)$ over a quasi-regular triangulation of $\mathcal{S}_r$, then classical interpolation theory shows that the initial error is bounded by
$$\|e_h^0\|_{L^2(\Omega)} \le C\|D_x^\ell u_0\|_{L^2(\mathcal{S}_r)} h^\ell.$$

Similarly, for $1 \leq p \leq \infty$,

$$\|u - \mathbb{P}_h^{loc} u\|_{L^p[0,T;L^2(S_r)]} \leq C \left(\|D_x^{\ell+1} u\|_{L^p[0,T;L^2(S_r)]} h^{\ell+1} + \|D_t^{k+1} u\|_{L^p[0,T;L^2(S_r)]} \tau^{k+1} \right),$$

where $h > 0$ is the usual mesh parameter. The inverse inequality shows

$$\|u - \mathbb{P}_h^{loc} u\|_{L^2[0,T;H^1(S_r)]} \leq C \left(\|D_x^{\ell+1} u\|_{L^2[0,T;L^2(S_r)]} h^{\ell} + \|D_t^{k+1} u\|_{L^2[0,T;L^2(S_r)]} (\tau^{k+1}/h) \right),$$

so

$$|||u - \mathbb{P}_h^{loc} u|||_\infty \leq C \left(\sqrt{\sigma} \left(h^\ell + \tau^{k+1}/h \right) + h^\ell + \tau^k \right),$$

where the constant C depends upon T, k, ℓ, and the norms

$$\|D_x^{\ell+1} u\|_{L^2[0,T;L^2(S_r)]}, \quad \|D_x^{\ell} u\|_{L^\infty[0,T;L^2(S_r)]},$$
$$\|D_t^{k+1} u\|_{L^2[0,T;L^2(S_r)]}, \quad \|D_t^{k} u\|_{L^\infty[0,T;L^2(S_r)]}.$$

When $\sigma << h$ the minimum of the expressions in the jump term is $\|(I - P_i)u(t^i)\|_{L^2(\Omega)}$ and

$$\sum_{i=0}^{N-1} \|(I - P_i)u(t^i)\|^2_{L^2(\Omega)} \leq C \|D_x^\ell u\|^2_{L^\infty[0,T;L^2(S_r)]} N h^{2\ell} \leq CT \|D_x^\ell u\|^2_{L^\infty[0,T;L^2(S_r)]} h^{2\ell}/\tau.$$

This gives an error of size $h^\ell/\sqrt{\tau}$, which is typical of the discontinuous Galerkin method for hyperbolic equations [24]. If σ is $O(1)$, the second term in the minimum can be bounded as $\|P_{i+1}(I - P_i)u(t^i)\|^2_{U'(t^i)} \leq C \|D_x^\ell u\|_{L^\infty[0,T;L^2(S_r)]} h^{\ell+1}$, and

$$\sum_{i=0}^{N-1} 1/(\sigma\tau) \|P_{i+1}(I - P_i)u(t^i)\|^2_{U'(t^i)} \leq (CT/\sigma) \|D_x^\ell u\|^2_{L^\infty[0,T;L^2(S_r)]} h^{2\ell+2}/\tau^2,$$

which gives the optimal $O(h^{\ell+1}/\tau)$ bound (assuming $\tau \sim h$).

Remark 4. (1) The above estimates assume that $\mathcal{S}_r$ is been triangulated exactly using a curvilinear coordinate system.

(2) The geometry of $\mathcal{S}(t)$ doesn't explicitly appear in the error estimates. However, if $\mathcal{S}_r$ has narrow regions of high curvature, the quasi-uniform assumption on the mesh will force a fine triangulation in such regions.

(3) If the exact solution is more regular in time, $u \in L^\infty[0,T;H^{\ell+1}(\Omega)]$, then the jump terms will be one order of h smaller. In this situation the jump terms will be asymptotically negligible compared with $|||u - \mathbb{P}_h^{loc} u|||_\infty$ when $\sigma = O(1)$.

REFERENCES

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

[2] G. Akrivis and C. Makridakis, *Galerkin time-stepping methods for nonlinear parabolic equations*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 261–289.

[3] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[4] M. Bause and P. Knabner, *Uniform error analysis for Lagrange–Galerkin approximations of convection-dominated problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1954–1984.

[5] B. Cockburn, G. E. Karnadiakis, and C. W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods, Newport, RI, 1999, Springer, Berlin, 2000, pp. 3–50.

[6] K. Chrysafinos and N. J. Walkington, *Error estimates for the discontinuous Galerkin methods for parabolic equations*, SIAM J. Numer. Anal., to appear; available online from http://www.math.cmu.edu/cna/publications.html.

[7] M. Delfour, W. Hager, and F. Trochu, *Discontinuous Galerkin methods for ordinary differential equations*, Math. Comp., 36 (1981), pp. 455–473.

[8] J. Douglas, Jr., and T. F. Russell, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.

[9] T. Dupont, *Mesh modification for evolutionary equations*, Math. Comp., 39 (1982), pp. 85–107.

[10] T. F. Dupont and Y. Liu, *Symmetric error estimates for moving mesh Galerkin methods for advection-diffusion equations*, SIAM J. Numer. Anal., 40 (2002), pp. 914–927.

[11] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* I. *A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

[12] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems.* II. *Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.

[13] K. Eriksson, C. Johnson, and V. Thomée, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Modél. Math. Anal. Numér., 29 (1985), pp. 611–643.

[14] P. Jamet, *Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain*, SIAM J. Numer. Anal., 15 (1978), pp. 912–928.

[15] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge, University Press, New York, 1987.

[16] O. Karakashian and C. Makridakis, *A space-time finite element method for nonlinear Schroedinger equation: The discontinuous Galerkin method*, Math. Comp., 67 (1998), pp. 479–499.

[17] P. Lasaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–123.

[18] S. Larsson, V. Thomée, and L. B. Wahlbin, *Numerical solution of parabolic integro-differential equations by the discontinuous Galerkin method*, Math. Comp., 67 (1998), pp. 45–71.

[19] Y. Liu, R. E. Bank, T. F Dupont, S. Garcia, and R. P. Santos, *Symmetric error estimates for moving mesh mixed methods for advection-diffusion equations*, SIAM J. Numer. Anal., 40 (2003), pp. 2270–2291.

[20] M. Luskin and R. Rannacher, *On the smoothing property of the Galerkin method for parabolic equations*, SIAM J. Numer. Anal., 19 (1982), pp. 93–113.

[21] K. Miller, *Moving finite elements* II, SIAM J. Numer. Anal., 18 (1981), pp. 1033–1057.

[22] K. Miller and R. N. Miller, *Moving finite elements* I, SIAM J. Numer. Anal., 18 (1981), pp. 1019–1032.

[23] K. W. Morton, A. Priestley, and E. Süli, *Stability of the Lagrange-Galerkin method with nonexact integration*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 625–653.

[24] T. E. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal., 28 (1991), pp. 133–140.

[25] C. Pozrikidis, *Numerical simulation of the flow-induced deformation of red blood cells*, Ann. Biomed. Eng., 31 (2003), pp. 1194–1205.

[26] R. E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, Boston, 1979; also available online from http://ejde.math.txstate.edu.

[27] R. E. Showalter, *Monotone Operators in Banach Space and Nonlinear Partial Differential Equations*, AMS, Providence, RI, 1997.

[28] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer, Berlin, 1997

[29] N. J. Walkington, *Convergence of the discontinuous Galerkin method for discontinuous solutions*, SIAM J. Numer. Anal., 42 (2005), pp. 1801–1817.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2500–2516

LOCALIZATION OF THE GENERALIZED SAMPLING SERIES AND ITS NUMERICAL APPLICATION*

LIWEN QIAN† AND DENNIS B. CREAMER‡

Abstract. We study the localization of the sampling series with general kernel and obtain its error estimates. To achieve exponentially decaying accuracy for the series to approximate a band-limited function and its derivatives, a sufficient condition is given and a rigorous criterion is provided for the sampling to obtain any desired accuracy. The result includes several known results as special cases and gives new results about wavelet sampling series.

Key words. localization, generalized sampling series, error estimate, Gaussian, Meyer wavelets

AMS subject classifications. Primary, 94A20, 41A30; Secondary, 41A25, 65D25, 65D15

DOI. 10.1137/04061845X

1. Introduction.

Let $\mathbb{R}$ be the set of all real numbers, $\mathbb{Z}$ the set of all integers, $\mathbb{Z}_+$ the set of all nonnegative integers, and $\mathbb{N}$ the set of all positive integers. For any interval $\mathbb{I} \subseteq \mathbb{R}$ and $1 \leq p \leq \infty$, we let $L^p(\mathbb{I})$ be the space of complex-valued Lebesgue measurable functions f on $\mathbb{I}$ for which the norm

$$\|f\|_{p,\mathbb{I}} := \begin{cases} \left\{\int_{\mathbb{I}} |f(x)|^p \mathrm{d}x\right\}^{1/p}, & 1 \leq p < \infty, \\ \operatorname{ess\,sup}\{|f(x)| : x \in \mathbb{I}\}, & p = \infty, \end{cases}$$

is finite. When $\mathbb{I} = \mathbb{R}$ we denote this norm as $\|f\|_p$.

Every function $f \in L^2(\mathbb{R})$ has a *Fourier transform* in $L^2(\mathbb{R})$ which we denote by $\widehat{f}$:

$$\widehat{f}(\omega) := \int_{\mathbb{R}} f(x) \exp(\mathrm{i}x\omega)\mathrm{d}x, \quad w \in \mathbb{R}.$$

A *signal*, that is, a function f in $L^2(\mathbb{R})$, is said to be band-limited with bandwidth σ provided we have for $|w| > \sigma$ that $\widehat{f}(w) = 0$. We denote the totality of such functions by

$$B_\sigma := \{f \in L^2(\mathbb{R}) : \widehat{f}(w) = 0, |w| > \sigma\}.$$

We let $C(\mathbb{R})$ be the space of complex-valued continuous functions on $\mathbb{R}$ with the maximum norm $\|\cdot\|_\infty$. For $h > 0$ we define $\mathcal{S}_h : C(\mathbb{R}) \to C(\mathbb{R})$ for $f \in C(\mathbb{R})$ by the equation

$$(\mathcal{S}_h f)(x) := \sum_{k\in\mathbb{Z}} f(kh)\phi(h^{-1}x - k), \quad x \in \mathbb{R}. \tag{1}$$

We refer to $\mathcal{S}_h$ as a *sampling operator* because it uses the values f on $h\mathbb{Z}$. By $C_c(\mathbb{R})$ we denote the subspace of $C(\mathbb{R})$ consisting of all functions which vanish outside of

*Received by the editors November 8, 2004; accepted for publication (in revised form) May 25, 2005; published electronically January 12, 2006.
http://www.siam.org/journals/sinum/43-6/61845.html

†Department of Computational Science, National University of Singapore, 117543, Singapore (liwen_qian@yahoo.com).

‡P.O. Box 660537, Arcadia, CA 91066. Current address: Code 5580, Naval Research Laboratory, 4555 Overlook Avenue, SW, Washington, DC 20375 (db_creamer@yahoo.com).

some finite subinterval of $\mathbb{R}$. When $\phi \in C_c(\mathbb{R})$, we call (1) the *Schoenberg operator* because of the work of I. J. Schoenberg [1] on spline functions. This hypothesis on ϕ naturally arises in his work on spline functions. When $\phi = \text{sinc}$, we call (1) the *cardinal operator* and denote it by $\mathcal{C}_h$:

$$(\mathcal{C}_h f)(x) := \sum_{k\in\mathbb{Z}} f(kh)\text{sinc}(h^{-1}x-k), \quad x \in \mathbb{R}. \tag{2}$$

A fundamental result in information theory is the Whittaker–Kotel'nikov–Shannon (WKS) sampling theorem [2, 3, 4]. It states that any $f \in B_\sigma$ can be reconstructed from its sampled values $f(x_k)$, where $x_k := k\pi/\sigma$ and $k \in \mathbb{Z}$, by the formula

$$f(x) = \sum_{k\in\mathbb{Z}} f(x_k)\text{sinc}(\sigma x/\pi - k), \quad x \in \mathbb{R},$$

where

$$\text{sinc}(x) := \begin{cases} \frac{\sin \pi x}{\pi x}, & x \in \mathbb{R}\setminus\{0\}, \\ 1, & x = 0, \end{cases}$$

and the series converges absolutely and uniformly on any finite closed interval of $\mathbb{R}$.

The applications of the WKS the sampling theorem have been widely studied [5, 6, 7, 8]. However, the sinc function in the WKS sampling series is not very convenient for practical applications, due largely to the fact that it has a very slow rate of decay at infinity.

An effort has been made to find replacements for the sinc kernel that have better decay properties and are more convenient for numerical computation. It is expected that a modification of the sinc function will yield a better convergence rate of the WKS sampling formula. Many particular convergence factors have been considered; see, for example, [8, 9, 10, 11, 12, 13, 14, 15, 16, 17] and references therein. Another approach is to replace the sinc function with a sampling atom with rapid decay; see, for example, [18, 19] and references therein.

In [20], the *localization operator* $\mathcal{G}_h : B_\sigma \to L^2(\mathbb{R})$, where $h \in (0, \pi/\sigma]$, defined by

$$(\mathcal{G}_h f)(x) := \sum_{k\in\mathbb{Z}} f(kh)\text{sinc}(h^{-1}x-k)\phi_r(h^{-1}x-k), \quad x \in \mathbb{R}, \tag{3}$$

is considered where $\phi_r := \phi(r^{-1}\cdot)$. The function ϕ is even on $\mathbb{R}$ satisfying $\phi(0) = 1$ and that ϕ is continuous at 0. We require $\phi \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R})$, so that the sampling series in (3) is uniformly convergent on $\mathbb{R}$ when $f \in B_\sigma$ for any $\sigma > 0$. Such a function ϕ can be a band-limited function or a duration-limited function.

For practical application, we consider its truncation version defined by

$$(\mathcal{T}_h f)(x) := \sum_{k\in\mathbb{Z}_m(x)} f(kh)\text{sinc}(h^{-1}x-k)\phi_r(h^{-1}x-k), \quad x \in \mathbb{R}, \tag{4}$$

where $\mathbb{Z}_m(x) := \{k \in \mathbb{Z} : |[h^{-1}x]-k| \le m\}$ for $m \in \mathbb{N}$ and $[x]$ denotes the integer part of $x \in \mathbb{R}$. Note that the truncation terms chosen here coincide with what is applied in [17], and it is different from what is considered in, for example, [10, 11, 12, 21], where the truncation is made from $-m$ to m for a given m.

For $h > 0$ and $s \in \mathbb{Z}_+$, we define the sth order derivative of the localization operator by $\mathcal{G}_h^{(s)} f := \mathcal{D}^{(s)}\mathcal{G}_h$, and correspondingly for the truncated form $\mathcal{T}_h^{(s)} f := \mathcal{D}^{(s)}\mathcal{T}_h$, where $\mathcal{D}^{(s)}$ denotes the sth order derivative operator.

We will study how well the operator $\mathcal{G}_h^{(s)}$ or $\mathcal{T}_h^{(s)}$ approximates $\mathcal{D}^{(s)}$. This problem is valuable for numerical application. To this end, we recall for $p \in [1, \infty]$ and $s \in \mathbb{Z}^+$ that the *Sobolev space* is defined as

$$W^{s,p}(\mathbb{R}) := \{f \in L^1(\mathbb{R}), f^{(k)} \in L^p(\mathbb{R}), 1 \leq k \leq s\},$$

where for $f \in W^{s,p}(\mathbb{R})$ and an interval $\mathbb{I} \subseteq \mathbb{R}$, we let

$$\|f\|_{s,p,\mathbb{I}} := \sum_{k=1}^{s} \|f^{(s)}\|_{p,\mathbb{I}}.$$

Typically, for given $m \in \mathbb{N}$ and $\mathbb{I} := \mathbb{R} \setminus [-m, m]$, we denote $\|f\|_{s,p,\mathbb{I}}$ by $\|f\|_{s,p,m}$.

For $h > 0$ and $s \in \mathbb{Z}_+$, error estimations for $\mathcal{T}_h^{(s)} f$ to approximate the sth order derivative of a band-limited function f have been obtained in [20]. When ϕ is the Gaussian function $G(x) := \exp(-x^2/2)$ for $x \in \mathbb{R}$, which is proposed in [16], the explicit error bound is given in [22]. If the sinc kernel in (4) is replaced by $r_1 \mathrm{sinc}(r_1 x)$ for $x \in \mathbb{R}$ and certain $r_1 > 0$, the error estimates are given in [23] and its numerical application for solving Burgers's equation is discussed in [24]. When the sinc kernel in (4) is replaced by a modified version, which is defined for $x \in \mathbb{R}$ and $\lambda \in [0, \pi)$ by the formula $S(x) := \mathrm{sinc}(x) \cos(\lambda x)$, the L^∞ error estimates are given in [25], which shows that the approximation accuracy is comparable with that achieved by using the sinc kernel.

Since different kernels have been efficiently applied and analyzed, we were motivated to replace the sinc kernel in (4) by a general kernel K, that is, to consider

$$(\mathcal{G}_h f)(x) := \sum_{k \in \mathbb{Z}} f(kh) K(h^{-1}x - k) G_r(h^{-1}x - k), \quad x \in \mathbb{R}. \tag{5}$$

Furthermore, we consider its truncated version defined by

$$(\mathcal{T}_h f)(x) := \sum_{k \in \mathbb{Z}_m(x)} f(kh) K(h^{-1}x - k) G_r(h^{-1}x - k), \quad x \in \mathbb{R}. \tag{6}$$

We will give a sufficient condition for the kernel function K so that (6) can approximate a band-limited function f with exponentially decaying accuracy.

In the next section, we review some known results for later comparison and introduce some preliminaries. Error estimates are derived in section 3. Rigorous and practical means for obtaining any desired accuracy for the sampling will be provided in section 4, where we also give a new result of sampling series which is based on Meyer wavelets. Section 5 provides a few numerical experiments to demonstrate the high efficiency of the approximation. The conclusion will be given in the last section.

2. Preliminaries.

We first review a basic result about approximation by the Schoenberg operator (1). For this purpose, we let π_n be the space of polynomials of degree not exceeding n, and we denote for any $h > 0$ the *modulus of continuity* of a function f defined on $\mathbb{R}$ by $w(f; h) := \sup\{|f(x) - f(y)| : |x - y| \leq h\}$. The following result is standard; see, for example, [20, pp. 13–14].

THEOREM 2.1. *If $\phi \in C_c(\mathbb{R})$ and $n \in \mathbb{Z}_+$, then there exists a positive constant c such that for all f with $f^{(n)} \in C(\mathbb{R})$ and $h > 0$ there holds the inequality*

$$\|\mathcal{S}_h f - f\|_\infty \leq c h^n w(f^{(n)}, h)$$

if and only if for every $p \in \pi_n$,

$$p(x) = \sum_{j \in \mathbb{Z}} p(j)\phi(x-j), \quad x \in \mathbb{R}.$$

For later comparison, we introduce the following result [22] for (6) with a sinc kernel to approximate a band-limited function f.

THEOREM 2.2. *If* $f \in B_\sigma$, $h \in (0, \pi/\sigma]$, $K := \operatorname{sinc}$ *in* (6), $r > 0$, $m \in \mathbb{N}$, $m \geq sr/\sqrt{2}$, $\alpha_0 := \min\{m/r, r(\pi - h\sigma)\}$, *and* $s \in \mathbb{Z}_+$, *then*

$$\|f^{(s)} - \mathcal{T}_h^{(s)} f\|_\infty \leq \beta_0 \exp(-\alpha_0^2/2),$$

where $\beta_0 := \frac{e^\pi r(s+1)!}{h^s \pi \alpha_0}(\sqrt{2\sigma}\|f\|_2 + 2r\|f\|_\infty)$.

The following inequality about the *Mills' ratio*, which is defined for $x \in \mathbb{R}$ as

$$M(x) := \exp\left(\frac{x^2}{2}\right)\int_x^{+\infty} \exp\left(-\frac{t^2}{2}\right)\mathrm{d}t,$$

will be essential for our later use.

LEMMA 2.1. *If* $x \geq 0$, *then*

$$\frac{\pi}{\sqrt{x^2 + 2\pi} + (\pi - 1)x} \leq M(x) \leq \frac{\pi}{\sqrt{(\pi-2)^2 x^2 + 2\pi} + 2x}.$$

Both bounds tend to $\sqrt{\pi/2}$ when $x \to 0$. See, for example, [28, pp. 177–181]. Based on Lemma 2.1, we have the following result [25].

LEMMA 2.2. *If* $x_0 \geq 0$, $s \in \mathbb{Z}_+$, *and* $x \geq \max\{1, \sqrt{s}\}$, *then*

$$\int_x^\infty (t + x_0)^s \exp(-t^2/2)\mathrm{d}t \leq (1+s)(x+x_0)^s \frac{\exp(-x^2/2)}{x}.$$

3. Error estimates. For a kernel function K used in (5), we suppose that K satisfies the following:

(1) for certain $0 \leq a \leq b \leq 2\pi$ that

$$\widehat{K}(w) = 1, \quad |w| \leq a; \widehat{K}(w) = 0, \quad |w| \geq b; \widehat{K}(w) \in [0,1], \quad \text{otherwise}, \tag{7}$$

(2) for given $s \in \mathbb{Z}_+$ and $m \in \mathbb{N}$ that

$$\|K\|_{s,\infty,m} < \infty. \tag{8}$$

For simplicity of notation, we denote

$$c := \min\{2\pi - b, a\},$$

which will be used throughout the rest of the paper. Note that from (7) we have $c \leq \pi$. We will derive the error bounds for (6) to approximate a band-limited function f. We carry out the estimates in three steps.

3.1. Without truncation. The first step is to consider the operator $\mathcal{G}_h$ defined in (5) and to derive for any $s \in \mathbb{Z}_+$ the estimate of the *localization error* $f^{(s)} - \mathcal{G}_h^{(s)} f$. To this end, for a given function $K \in L^1(\mathbb{R})$ and $h, r > 0$, we associate functions μ

and ν defined for $\omega \in \mathbb{R}$ by setting

$$\mu(\omega) := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} \widehat{K}(h\omega - t/r) \exp(-t^2/2) \mathrm{d}t \tag{9}$$

and

$$\nu(\omega) := 1 - \mu(\omega) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} [1 - \widehat{K}(h\omega - t/r)] \exp(-t^2/2) \mathrm{d}t.$$

We need the following result.

Lemma 3.1. *If K satisfies* (7), $h \in (0, c/\sigma]$, $r(c - h\sigma) \geq \max\{1, \sqrt{s}\}$, *and* $\omega \in [-\sigma, \sigma]$, *then*

$$\begin{aligned} &\sqrt{|\omega^s \nu(\omega)|^2 + \sum_{k \in \mathbb{Z} \setminus \{0\}} |(\omega + 2k\pi/h)^s \mu(\omega + 2k\pi/h)|^2} \\ &\quad \leq (1 + 2^{s+1})(1 + s) \sqrt{\frac{2}{\pi}} \Big(\frac{\pi}{h}\Big)^s \frac{\mathrm{e}^{-r^2(c-h\sigma)^2/2}}{r(c - h\sigma)} \end{aligned}$$

and

$$|\omega^s \nu(\omega)| + \sum_{k \in \mathbb{Z} \setminus \{0\}} |(\omega + 2k\pi/h)^s \mu(\omega + 2k\pi/h)| \leq (1 + 2^{s+1})(1 + s) \sqrt{\frac{2}{\pi}} \Big(\frac{\pi}{h}\Big)^s \frac{\mathrm{e}^{-r^2(c-h\sigma)^2/2}}{r(c - h\sigma)}.$$

Proof. From (7) and (9) we observe for $\omega \in [-\sigma, \sigma]$ that

$$|\mu(\omega)| \leq \frac{1}{\sqrt{2\pi}} \int_{|h\omega - t/r| \leq b} \exp(-t^2/2) \mathrm{d}t,$$

which leads to

$$\begin{aligned} S_1(\omega) : &= \sum_{k \in \mathbb{N}} |(\omega + 2k\pi/h)^s \mu(\omega + 2k\pi/h)| \\ &= \frac{1}{\sqrt{2\pi}} \sum_{k \in \mathbb{N}} (\omega + 2k\pi/h)^s \int_{r(2k\pi + h\omega - b)}^{r(2k\pi + h\omega + b)} \exp(-t^2/2) \mathrm{d}t \\ &\leq \frac{1}{\sqrt{2\pi}(rh)^s} \sum_{k \in \mathbb{N}} \int_{r(h\omega + 2k\pi - b)}^{r(h\omega + 2k\pi + b)} (t + rb)^s \exp(-t^2/2) \mathrm{d}t. \end{aligned}$$

Since $b \leq 2\pi$, we have $h\omega + 2(k+2)\pi - b \geq h\omega + 2k\pi + b$ for $k \in \mathbb{N}$ and that

$$\begin{aligned} S_1(\omega) &\leq \frac{1}{\sqrt{2\pi}(rh)^s} \sum_{k \in 2\mathbb{N} - 1} \int_{r(h\omega + 2k\pi - b)}^{r(h\omega + 2k\pi + b)} (t + rb)^s \exp(-t^2/2) \mathrm{d}t \\ &\quad + \frac{1}{\sqrt{2\pi}(rh)^s} \sum_{k \in 2\mathbb{N}} \int_{r(h\omega + 2k\pi - b)}^{r(h\omega + 2k\pi + b)} (t + rb)^s \exp(-t^2/2) \mathrm{d}t \\ &\leq \frac{2}{\sqrt{2\pi}(rh)^s} \int_{r(h\omega + 2\pi - b)}^{\infty} (t + rb)^s \exp(-t^2/2) \mathrm{d}t \\ &\leq \frac{2}{\sqrt{2\pi}(rh)^s} \int_{r(c - h\sigma)}^{\infty} (t + rb)^s \exp(-t^2/2) \mathrm{d}t \\ &\leq \frac{2(1 + s)}{\sqrt{2\pi}} \Big(\frac{2\pi}{h} - \sigma\Big)^s \frac{\exp(-r^2(c - h\sigma)^2/2)}{r(c - h\sigma)}, \end{aligned} \tag{10}$$

where the last step follows from Lemma 2.2. Furthermore, we have

$$\begin{aligned}S_2(\omega) :&= \sum_{-k\in\mathbb{N}} |(\omega+2k\pi/h)^s\mu(\omega+2k\pi/h)|\\ &= \frac{1}{\sqrt{2\pi}}\sum_{-k\in\mathbb{N}}|\omega+2k\pi/h|^s\int_{r(h\omega+2k\pi-b)}^{r(h\omega+2k\pi+b)}\exp(-t^2/2)\mathrm{d}t\\ &= \frac{1}{\sqrt{2\pi}}\sum_{k\in\mathbb{N}}(2k\pi/h-\omega)^s\int_{r(2k\pi-b-h\omega)}^{r(2k\pi+b-h\omega)}\exp(-t^2/2)\mathrm{d}t = S_1(-\omega).\end{aligned}\tag{11}$$

On the other hand, since

$$\nu(\omega)\le 1-\frac{1}{\sqrt{2\pi}}\int_{|h\omega-t/r|\le a}\exp(-t^2/2)\mathrm{d}t = \frac{1}{\sqrt{2\pi}}\int_{|h\omega-t/r|\ge a}\exp(-t^2/2)\mathrm{d}t,$$

from Lemma 2.1 we have

$$\begin{aligned}|\omega^s\nu(\omega)| &\le \frac{|\omega|^s}{\sqrt{2\pi}}\int_{r(h\omega+a)}^{\infty}\mathrm{e}^{-t^2/2}\mathrm{d}t+\frac{|\omega|^s}{\sqrt{2\pi}}\int_{-\infty}^{r(h\omega-a)}\mathrm{e}^{-t^2/2}\mathrm{d}t\\ &\le \sqrt{\frac{2}{\pi}}\sigma^s\frac{\exp(-r^2(c-h\sigma)^2/2)}{r(c-h\sigma)} \le \sqrt{\frac{2}{\pi}}\Big(\frac{\pi}{h}\Big)^s\frac{\exp(-r^2(c-h\sigma)^2/2)}{r(c-h\sigma)},\end{aligned}\tag{12}$$

where the last step follows from $\sigma\le c/h$ and $c\le\pi$. Thus, combining (10), (11), and (12) gives

$$\begin{aligned}&|\omega^s\nu(\omega)|+\sum_{k\in\mathbb{Z}\setminus\{0\}}|(\omega+2k\pi/h)^s\mu(\omega+2k\pi/h)|\\ &\quad\le(1+2^{s+1})(1+s)\sqrt{\frac{2}{\pi}}\Big(\frac{\pi}{h}\Big)^s\frac{\exp(-r^2(c-h\sigma)^2/2)}{r(c-h\sigma)}.\end{aligned}$$

From

$$\begin{aligned}&\sqrt{|\omega^{2s}\nu(\omega)^2|+\sum_{k\in\mathbb{Z}\setminus\{0\}}|(\omega+2k\pi/h)^{2s}\mu(\omega+2k\pi/h)^2|}\\ &\le|\omega^s\nu(\omega)|+\sum_{k\in\mathbb{Z}\setminus\{0\}}|(\omega+2k\pi/h)^s\mu(\omega+2k\pi/h)|,\end{aligned}$$

we finish the proof. □

Now, we are ready to obtain for given $f\in B_\sigma$ the error estimates of $\mathcal{G}_hf-f$ and its derivatives.

THEOREM 3.1. *If $f\in B_\sigma$, K satisfies (7), $h\in(0,c/\sigma]$, $r>0$, $s\in\mathbb{Z}_+$, and $r(c-h\sigma)\ge\max\{1,\sqrt{s}\}$, then*

$$\|f^{(s)}-\mathcal{G}_h^{(s)}f\|_2\le(1+2^{s+1})(1+s)\sqrt{\frac{2}{\pi}}\Big(\frac{\pi}{h}\Big)^s\frac{\mathrm{e}^{-r^2(c-h\sigma)^2/2}}{r(c-h\sigma)}\|f\|_2$$

and

$$\|f^{(s)}-\mathcal{G}_h^{(s)}f\|_\infty\le(1+2^{s+1})(1+s)\frac{\sqrt{2\sigma}}{\pi}\Big(\frac{\pi}{h}\Big)^s\frac{\mathrm{e}^{-r^2(c-h\sigma)^2/2}}{r(c-h\sigma)}\|f\|_2.$$

Proof. Since for $\omega \in \mathbb{R}$ we have

$$K(h^{-1}\cdot -k)\widehat{(\omega)} = h\exp(\mathrm{i}kh\omega)\widehat{K}(h\omega)$$

and

$$G_r(h^{-1}\cdot -k)\widehat{(\omega)} = hr\exp(\mathrm{i}kh\omega)\sqrt{2\pi}\exp(-r^2h^2\omega^2/2),$$

we obtain

$$\begin{aligned} &K(h^{-1}\cdot -k)\widehat{(\omega)} * G_r(h^{-1}\cdot -k)\widehat{(\omega)} \\ &\quad = \sqrt{2\pi}h^2 r\exp(\mathrm{i}kh\omega)\int_{\mathbb{R}}\widehat{K}(h(\omega-\theta))\exp(-r^2h^2\theta^2/2)\mathrm{d}\theta, \end{aligned}$$

which provides for

$$\mathcal{G}_h f := \sum_{k\in\mathbb{Z}} f(kh)K(h^{-1}\cdot -k)G_r(h^{-1}\cdot -k)$$

that

$$\begin{aligned} (\mathcal{G}_h f)\widehat{(\omega)} &= \frac{1}{2\pi}\sum_{k\in\mathbb{Z}} f(kh)K(h^{-1}\cdot -k)\widehat{(\omega)} * G_r(h^{-1}\cdot -k)\widehat{(\omega)} \\ &= \sum_{k\in\mathbb{Z}} f(kh)h\exp(\mathrm{i}kh\omega)\mu(\omega), \end{aligned} \tag{13}$$

where μ is defined in (9). On the other hand, since $c \le \pi$, function f satisfies

$$\widehat{f} \in L^2[-\sigma,\sigma] \subseteq L^2[-c/h, c/h] \subseteq L^2[-\pi/h,\pi/h],$$

and from its Fourier series expansion, we have

$$\widehat{f}(\omega) = \widehat{f}(\omega)\chi_{[-\sigma,\sigma]}(\omega) = \sum_{k\in\mathbb{Z}} hf(kh)\exp(\mathrm{i}kh\omega)\chi_{[-\sigma,\sigma]}(\omega), \quad \omega\in\mathbb{R}. \tag{14}$$

Applying (13) and (14) to $E_1 := f^{(s)} - \mathcal{G}_h^{(s)}f$ gives for all $\omega \in \mathbb{R}$ that

$$\widehat{E_1}(\omega) = (\mathrm{i}\omega)^s\sum_{k\in\mathbb{Z}} hf(kh)\exp(\mathrm{i}kh\omega)(\chi_{[-\sigma,\sigma]}(\omega) - \mu(\omega)).$$

Since f is band-limited to σ, we restrict $\widehat{f}$ to the interval $[-\pi/h, \pi/h]$, extend this function to a $2\pi/h$-periodic function and denote the resulting function as g. Then

$$\widehat{E_1}(\omega) = \begin{cases} (\mathrm{i}\omega)^s\widehat{f}(\omega)\nu(\omega), & |\omega| \le \sigma, \\ -(\mathrm{i}\omega)^s g(\omega)\mu(\omega), & |\omega - 2k\pi/h| \le \sigma, k\in\mathbb{Z}\setminus\{0\}, \\ 0 & \text{otherwise.} \end{cases}$$

We observe that if $\omega \in \mathbb{R}$ such that $|\omega - 2k\pi/h| \le \sigma$, where $k \in \mathbb{Z}\setminus\{0\}$, then

$$g(\omega) = g(\omega - 2k\pi/h) = \widehat{f}(\omega - 2k\pi/h),$$

which provides that

$$\widehat{E_1}(\omega) = \begin{cases} (\mathrm{i}\omega)^s\widehat{f}(\omega)\nu(\omega), & |\omega| \le \sigma, \\ -(\mathrm{i}\omega)^s\widehat{f}(\omega - 2k\pi/h)\mu(\omega), & |\omega - 2k\pi/h| \le \sigma, k\in\mathbb{Z}\setminus\{0\}, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, we have that

$$\|\widehat{E}_1\|_2^2 = \int_{-\sigma}^{\sigma} |\omega^s \widehat{f}(\omega)\nu(\omega)|^2 \mathrm{d}\omega + \sum_{k\in\mathbb{Z}\setminus\{0\}} \int_{-\sigma}^{\sigma} |(\omega + 2k\pi/h)^s \widehat{f}(\omega)\mu(\omega + 2k\pi/h)|^2 \mathrm{d}\omega$$
$$= \int_{-\sigma}^{\sigma} |\widehat{f}(\omega)|^2 \left(|\omega^s \nu(\omega)|^2 + \sum_{k\in\mathbb{Z}\setminus\{0\}} |(\omega + 2k\pi/h)^s \mu(\omega + 2k\pi/h)|^2 \right) \mathrm{d}\omega.$$

Lemma 3.1 implies the L^2 estimates. Since

$$\int_{\mathbb{R}} |\widehat{E_1}(\omega)| \mathrm{d}\omega = \int_{-\sigma}^{\sigma} |\omega^s \widehat{f}(\omega)\nu(\omega)| \mathrm{d}\omega + \sum_{k\in\mathbb{Z}\setminus\{0\}} \int_{-\sigma}^{\sigma} |(\omega + 2k\pi/h)^s \widehat{f}(\omega)\mu(\omega + 2k\pi/h)| \mathrm{d}\omega$$
$$= \int_{-\sigma}^{\sigma} |\widehat{f}(\omega)| \left(|\omega^s \nu(\omega)| + \sum_{k\in\mathbb{Z}\setminus\{0\}} |(\omega + 2k\pi/h)^s \mu(\omega + 2k\pi/h)| \right) \mathrm{d}\omega,$$

applying Lemma 3.1 and the Cauchy–Schwarz inequality to the above gives the L^∞ estimates. □

Note that if $r \to +\infty$, then $\exp(-r^2(-h\sigma + c)^2/2) \to 0$. Thus, from Theorem 3.1 we have the following result.

COROLLARY 3.1. *If $f \in B_\sigma$, K satisfies* (7), *and $h \in (0, c/\sigma]$, then*

$$f(x) = \sum_{k\in\mathbb{Z}} f(kh)K(h^{-1}x - k), \quad x \in \mathbb{R},$$

and the series converges in the L^2 norm and in the L^∞ norm on $\mathbb{R}$. It also converges uniformly on any finite closed interval of $\mathbb{R}$.

If $K :=$ sinc, then the above result reduces to the WKS sampling theorem. Therefore, both Theorem 3.1 and Corollary 3.1 can be viewed as generalizations of the WKS sampling theorem.

3.2. Truncated series. The next step is to estimate for $h > 0$ and $s \in \mathbb{Z}_+$ the *truncation error* $\mathcal{G}_h^{(s)} f - \mathcal{T}_h^{(s)} f$, where $\mathcal{G}_h$ and $\mathcal{T}_h$ are defined in (5) and (6), respectively.

THEOREM 3.2. *If $f \in L^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$, $h, r > 0$, $s \in \mathbb{Z}_+$, $m \in \mathbb{N}$, $m \geq sr/\sqrt{2}$, and K satisfies* (8), *then*

$$\|\mathcal{G}_h^{(s)} f - \mathcal{T}_h^{(s)} f\|_\infty \leq \frac{2\sqrt{\sigma/\pi}\, r^2 s! \|K\|_{s,\infty,m}}{(m-2)h^s} \exp(-(m-2)^2/2r^2)\|f\|_2.$$

Proof. We denote $E_2 := \mathcal{G}_h^{(s)} f - \mathcal{T}_h^{(s)} f$. For $x \in \mathbb{R}$ and $s \in \mathbb{Z}_+$ we have

$$E_2(x) := \sum_{|[x/h]-k|>m} f(kh) \frac{\mathrm{d}^s}{\mathrm{d}x^s} \psi(x/h - k),$$

where

$$\psi(x) := K(x) \exp(-x^2/2r^2).$$

We denote the fractional part of $x \in \mathbb{R}$ by $\{x\}$ and let $l := [x/h] - k$; then we have

$$E_2(x) = \sum_{|l|>m} f(x - lh - \{x/h\}h) \frac{1}{h^s} \psi^{(s)}(l + \{x/h\}),$$

which leads to

$$|E_2(x)| \le \frac{\|f\|_\infty}{h^s} \sum_{|l|>m} |\psi^{(s)}(l + \{x/h\})|. \tag{15}$$

On the other hand, since for $k \in \mathbb{N}$ we have

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp(-x^2/2r^2) = \frac{(-1)^k}{(\sqrt{2}r)^k} H_k(x/\sqrt{2}r) \exp(-x^2/2r^2),$$

where $H_k(x)$ is the kth order Hermite polynomial

$$\exp(-x^2) H_k(x) = (-1)^k \frac{\mathrm{d}^k}{\mathrm{d}x^k} \exp(-x^2),$$

we obtain

$$\psi^{(s)}(x) = \sum_{j_1+j_2=s} \frac{s!}{j_1! j_2!} K^{(j_1)}(x) \frac{(-1)^{j_2}}{(\sqrt{2}r)^{j_2}} H_{j_2}(x/\sqrt{2}r) \exp(-x^2/2r^2).$$

From [29, p. 187] we have

$$H_k(x) = \sum_{j=0}^{[k/2]} \frac{(-1)^j k! (2x)^{k-2j}}{j!(k-2j)!}.$$

For $x \in \mathbb{R}$, $k \in \mathbb{N}$, and $j \in \mathbb{Z}_+$ we denote $a_j := \frac{k!(2x)^{k-2j}}{j!(k-2j)!}$. Then we have $H_k(x) = \sum_{j=0}^{[k/2]} (-1)^j a_j$. It is easily shown that $\{a_j : j \in \mathbb{Z}_+\}$ decrease for $|x| \ge k/2$. This leads to $|H_k(x)| \le |a_0|$. Therefore, for $|x| \ge sr/\sqrt{2}$, we have

$$|H_k(x/\sqrt{2}r)| \le (\sqrt{2}|x|/r)^k,$$

which provides the estimation

$$|\psi^{(s)}(x)| \le \sum_{j_1+j_2=s} \frac{s!}{j_1! j_2!} K^{(j_1)}(x) \frac{|x|^{j_2}}{r^{2j_2}} \exp(-x^2/2r^2).$$

For $x \ge sr\sqrt{2}$ we have $\sum_{n=0}^{s} |x|^n/n! < e^{|x|}$. Thus, for $|x| > m$ we have

$$\begin{aligned} |\psi^{(s)}(x)| &\le s! \|K\|_{s,\infty,m} \exp(|x|/r^2) \exp(-x^2/2r^2) \\ &\le s! \|K\|_{s,\infty,m} \exp(-(|x|-1)^2/2r^2). \end{aligned} \tag{16}$$

Since $m \ge sr/\sqrt{2}$, applying (16) to (15) gives

$$\begin{aligned} |E_2(x)| \le \frac{s! \|K\|_{s,\infty,m} \|f\|_\infty}{h^s} \Bigg\{ &\sum_{l \ge m} \exp(-(l + \{x/h\})^2/2r^2) \\ &+ \sum_{l \ge m} \exp(-(l - \{x/h\})^2/2r^2) \Bigg\}. \end{aligned}$$

From Lemma 2.1,

$$\sum_{l\geq m} \exp(-(l+\{x/h\})^2/2r^2) < \int_{m-1}^{\infty} \exp(-x^2/2r^2)\mathrm{d}x$$
$$\leq \frac{r^2}{m-1}\exp(-(m-1)^2/2r^2).$$

Likewise,

$$\sum_{l\geq m} \exp(-(l-\{x/h\})^2/2r^2) < \int_{m-2}^{\infty} \exp(-x^2/2r^2)\mathrm{d}x$$
$$\leq \frac{r^2}{m-2}\exp(-(m-2)^2/2r^2).$$

Therefore, we obtain

$$|E_2(x)| \leq \frac{2r^2 s!\|K\|_{s,\infty,m}\|f\|_\infty}{(m-2)h^s}\exp(-(m-2)^2/2r^2).$$

By applying the Cauchy–Schwarz inequality and the Parseval identity, we have

$$\|f\|_\infty \leq \frac{1}{2\pi}\int_{-\sigma}^{\sigma}|\widehat{f}(\omega)|\mathrm{d}\omega \leq \frac{\sqrt{2\sigma}}{2\pi}\|\widehat{f}\|_2 = \sqrt{\sigma/\pi}\|f\|_2,$$

from which we obtain the estimate. □

3.3. Main result. For practical applications, only the *truncated* sampling series can be used. Therefore, we will now turn our attention to error estimates for $\mathcal{T}_h$ defined in (6) to approximate a band-limited function and its derivatives.

THEOREM 3.3. *If $f \in B_\sigma$, $h \in (0, c/\sigma]$, $r > 0$, $s \in \mathbb{Z}_+$, $r(c-h\sigma) \geq \max\{1, \sqrt{s}\}$, $m \in \mathbb{N}$, $m \geq sr/\sqrt{2}$, K satisfies* (7) *and* (8), *and* $\alpha := \min\{r(c-h\sigma), (m-2)/r\}$, *then*

$$\|f^{(s)} - \mathcal{T}_h^{(s)}f\|_\infty \leq \beta\exp(-\alpha^2/2)\|f\|_2,$$

where

$$\beta := ((1+2^{s+1})(1+s)\pi^{s-1} + \sqrt{2/\pi}rs!\|K\|_{s,\infty,m})\frac{\sqrt{2\sigma}}{\alpha h^s}.$$

Proof. Since

$$\|f^{(s)} - \mathcal{T}_h^{(s)}f\|_\infty \leq \|f^{(s)} - \mathcal{G}_h^{(s)}f\|_\infty + \|\mathcal{G}_h^{(s)}f - \mathcal{T}_h^{(s)}f\|_\infty,$$

combining the error bounds in Theorems 3.1 and 3.2 gives the estimate with

$$\alpha := \min\{r(c-h\sigma), (m-2)/r\}$$

and

$$\beta := (1+2^{s+1})(1+s)\frac{\sqrt{2\sigma}}{\pi}\left(\frac{\pi}{h}\right)^s\frac{1}{r(c-h\sigma)} + \frac{2\sqrt{\sigma/\pi}r^2 s!\|K\|_{s,\infty,m}}{(m-2)h^s}$$
$$\leq ((1+2^{s+1})(1+s)\pi^{s-1} + \sqrt{2/\pi}rs!\|K\|_{s,\infty,m})\frac{\sqrt{2\sigma}}{\alpha h^s}. \tag{17}$$

Thus, we finish the proof. □

Theorem 3.3 tells us that the approximation can achieve *exponentially decaying* accuracy provided $\alpha > 0$. Note that here the convergence is measured by a balance of various parameters, which include the variance r of the Gaussian, the bandwidth σ of f, the sampling rate h, and the truncation level m. This point of view is different from that found in the literature, where convergence is measured by the sample rate for functions.

In fact, if we specialize error bounds available in the literature to the class of functions we considered, our error estimates are far superior. For example, Theorem 2.1 tells us that if the kernel function is B-splines, the error estimate is of polynomial order. The convergence in Theorem 3.3 could be much faster for given h.

Theorem 3.3 provides great flexibility in approximation by involving many parameters that can be chosen appropriately. For application, simplified error bounds depending only on the truncation level m for a given grid spacing h are desirable. Since the maximal α takes when $r(c - h\sigma) = (m - 2)/r$, we can choose the optimal scaling factor $r = \sqrt{\frac{m-2}{c-h\sigma}}$ in the error bounds of Theorem 3.3. For this and the effects of different parameters on the error estimate, see, for example, [25]. Since the approaches and results there are basically similar to our general cases, we omit the details for simplicity of presentation.

As the proof of the above result mainly requires an application of the Fourier transform, it is possible to use the distributional theory of the Fourier transform to prove analogous results for a function that is bounded by a polynomial at infinity. This case is valuable for the study of radial basis approximation; see, for example, [26]. We will not require such improvements here, although they are important.

We remark that extensive numerical experiments have shown that high accuracy of approximation still holds for not necessarily band-limited functions and their derivatives. However, the explicit error estimates for the aliasing error are not available yet, though some preliminary results have been obtained.

4. Applications. For applications in computation, we deal with finite domain. Let $\mathbb{I}$ be a finite interval on $\mathbb{R}$ and $|\mathbb{I}|$ be its Lebesgue measure. For $p \in [1, \infty]$, the following result gives the estimates in L^p norms for approximation on the interval $\mathbb{I}$.

THEOREM 4.1. *If the hypotheses of Theorem* 3.3 *hold,* α, β *are given in Theorem* 3.3*,* $\mathbb{I} \subset \mathbb{R}$*, and* $p \in [1, \infty]$*, then*

$$\|f^{(s)} - \mathcal{T}_h^{(s)} f\|_{p,\mathbb{I}} \leq \beta_p \exp(-\alpha^2/2)\|f\|_2,$$

where

$$\beta_p := \begin{cases} |\mathbb{I}|^{1/p}\beta, & p \in [1, \infty), \\ \beta, & p = \infty. \end{cases}$$

Proof. It is obvious for $p \in [1, \infty)$ that

$$\left\{ \int_{\mathbb{I}} |f^{(s)}(x) - (\mathcal{T}_h^{(s)} f)(x)|^p \mathrm{d}x \right\}^{1/p} \leq |\mathbb{I}|^{1/p} \|f^{(s)} - \mathcal{T}_h^{(s)} f\|_\infty,$$

and for $p = \infty$ that

$$\|f^{(s)} - \mathcal{T}_h^{(s)} f\|_{\infty,\mathbb{I}} \leq \|f^{(s)} - \mathcal{T}_h^{(s)} f\|_\infty.$$

Thus, from Theorem 3.3 we obtain the conclusion. □

We remark that the L^2 estimate in Theorem 3.2, and thereafter in Theorem 3.3, has not been obtained yet. It is not clear whether or not the L^2 estimate is similar to the L^∞ estimate.

The error estimates in Theorem 4.1 are a useful guide for use in numerical computations. For this purpose, we obtain the following fact, which provides rigorous means for obtaining the desired accuracy of sampling.

THEOREM 4.2. *Given any $\rho > 0$, if the hypotheses of Theorem* 4.1 *hold, $\gamma_p := \ln \beta_p$, and the parameters m, r, and h are chosen such that*

$$\min\{r(c - h\sigma), (m-2)/r\} \geq \sqrt{2(\gamma_p + \rho \ln 10)}, \tag{18}$$

then

$$\|f^{(s)} - \mathcal{T}_h^{(s)} f\|_{p,\mathbb{I}} \leq 10^{-\rho}.$$

Proof. From Theorem 4.1 the result follows directly. □

By using (18), we can choose m, r, and h appropriately to attain any desired accuracy. Roughly speaking, suppose $c = \pi$, $\beta_p = 1$, and h and σ are small enough; then we can choose $r \in [3, 4]$ and $m \sim 30$ to ensure the highest accuracy in a double precision computation, that is, $\rho = 16$.

There are a few examples of kernel function K that satisfy the hypotheses (7) and (8).

Example 1:

$$K(x) := \operatorname{sinc}(x), \quad x \in \mathbb{R}.$$

This is the Shannon kernel discussed in [22]. Here $a = b = c = \pi$.

Example 2:

$$K(x) := r_1 \operatorname{sinc}(r_1 x), \quad x \in \mathbb{R},$$

for $r_1 > 0$. This is the oversampled Shannon kernel discussed in [24]. Here $a = b = r_1\pi$ and $c = \min\{2\pi - r_1\pi, r_1\pi\}$. When $r_1 := 1$ it reduces to the case of Example 1.

Example 3:

$$K(x) := \operatorname{sinc}(x)\cos(\lambda x), \quad x \in \mathbb{R},$$

for $\lambda \in [0, \pi)$. This is the modified sinc kernel studied in [25]. Here $a = \pi - \lambda$, $b = \pi + \lambda$, and $c = \pi - \lambda$. When $\lambda := 0$ it reduces to the case of Example 1.

Example 4:

$$K(x) := \operatorname{sinc}(x)\operatorname{sinc}(x/3), \quad x \in \mathbb{R}.$$

This is the de la Vallée Poussin kernel discussed in [27]. Here $a = c = 2\pi/3$ and $b = 4\pi/3$.

Example 5: The scaling function ϕ of the Meyer wavelet [30] is given by its Fourier transform

$$\widehat{\phi}(w) = \begin{cases} 1, & |w| \leq 2\pi/3, \\ \cos[\frac{\pi}{2}\theta(3|w|/2\pi - 1)], & 2\pi/2 < |w| < 4\pi/3, \\ 0, & |w| \geq 4\pi/3, \end{cases} \tag{19}$$

where θ is a real-valued function satisfying $\theta + \theta(1-\cdot) = 1$. Let $d \in (1/2, 1]$ and

$$\theta(w) = \begin{cases} \frac{2}{\pi}\operatorname{arccot}\left(\frac{w-d}{1-w-d}\right), & 1/2 \le w \le d, \\ 1, & w > d. \end{cases} \tag{20}$$

Using $\theta + \theta(1-\cdot) = 1$, we extend θ on $[0,1]$. Then the sampling function for the wavelet subspace [31] is

$$W(x) = \frac{3\sin(\pi x)\sin(2\pi(d-0.5)x/3)}{2\pi^2(d-0.5)x^2}, \quad x \in \mathbb{R}. \tag{21}$$

See [32, 21]. We have the following estimates, which provide much faster convergence than that given in [32, 21].

COROLLARY 4.1. *If $f \in B_\sigma$, $K := W$ defined in (21), $d \in (1/2,1]$, $h \in (0, (4-2d)\pi/3\sigma]$, $r > 0$, $s \in \mathbb{Z}_+$, $r[(4-2d)\pi/3 - h\sigma] \ge \max\{1, \sqrt{s}\}$, $m \in \mathbb{N}$, $m \ge sr/\sqrt{2}$, and $\alpha_1 := \min\{r[(4-2d)\pi/3 - h\sigma], (m-2)/r\}$, then*

$$\|f^{(s)} - \mathcal{T}_h^{(s)} f\|_\infty \le \beta_1 \exp(-\alpha_1^2/2)\|f\|_2,$$

where $\beta_1 := \left((1+2^{s+1})(1+s)\pi^{s-1} + \sqrt{2/\pi}rs!\frac{s(1+\lambda)^s}{\lambda m^2}\right)\frac{\sqrt{2\sigma}}{\alpha h^s}$.

Proof. Direct computing shows that W satisfies hypotheses (7) with $a = c = (4-2d)\pi/3$. So we need only to estimate for $s \in \mathbb{Z}_+$ and $m \in \mathbb{N}$ the quantity $\|W\|_{s,\infty,m}$. We observe for $\lambda := 2(d-0.5)/3 \in (0, 1/3]$ and $x \in \mathbb{R}$ that

$$W(x) = \frac{\cos(1-\lambda)x - \cos(1+\lambda)x}{2\lambda x^2}$$

and

$$W^{(s)}(x) = \sum_{j+k=s} \frac{s!}{j!k!}\{(1-\lambda)^j \cos[(1-\lambda)x + j\pi/2] \\ -(1+\lambda)^j\cos[(1+\lambda)x + j\pi/2]\}\frac{(-1)^k(k+1)!}{2\lambda x^{k+2}},$$

which provides

$$|W^{(s)}(x)| \le \frac{s!}{2\lambda}\sum_{j+k=s}\frac{(1-\lambda)^j + (1+\lambda)^j}{j!}\frac{k+1}{x^{k+2}}.$$

Since for $j \in \mathbb{Z}_+$ the function $x^j/j!$ increases on $x > 0$, we have

$$|W^{(s)}(x)| \le \frac{(1-\lambda)^s + (1+\lambda)^s}{2\lambda}\sum_{k=0}^{s}\frac{k+1}{x^{k+2}} \le \frac{(1+\lambda)^s}{\lambda}\sum_{k=0}^{s}\frac{k+1}{x^{k+2}} \le \frac{s(1+\lambda)^s}{\lambda x^2}.$$

So, we conclude that

$$\|W\|_{s,\infty,m} \le \frac{s(1+\lambda)^s}{\lambda m^2},$$

and W satisfies (8). Thus, from Theorem 3.3 we finish our proof. □

There are many examples of wavelet sampling series (see, for example, [33, pp. 160–162]) that Theorem 3.3 can be applied to obtain much faster convergence. We omit the details. For the relation between the localized sampling and the Meyer wavelets, we will investigate it further at another occasion.

5. Numerical experiments. We may illustrate the approximation with numerical experiments. We consider the function $f := \operatorname{sinc}^7 \in B_{7\pi}$ and denote the computation domain by $\mathbb{I} \subset \mathbb{R}$. First, we describe the numerical scheme based on applying the operator (6). We start by specifying the spatial grids. Let $h := |\mathbb{I}|/n$ and $x_j := jh$ for $0 \leq j \leq n-1$. Here n is the number of the grid points and we choose $n > 2m$, where m is the truncation level. From (6) we have

$$(\mathcal{T}_h f)(x_j) = \sum_{k \in \mathbb{Z}_m(x_j)} f(x_k) K(x_j - x_k) G_r(x_j - x_k). \tag{22}$$

For $s \in \mathbb{Z}^+$, we use $\mathcal{T}_h^{(s)} f$ to approximate $f^{(s)}$.

Note that the values $\{f(x_k) : -m \leq k \leq n+m-1\}$ are required to implement (22). So, we need to extend the values $\{f(x_k) : 0 \leq k \leq n-1\}$ to $\{f(x_k) : -m \leq k \leq n+m-1\}$. Since the function f decays fast, we will make an antisymmetric extension of the function f and therefore set $f(x_k) := -f(x_{-k})$ for $-m \leq k \leq -1$ and $f(x_k) := -f(x_{2n-k-2})$ for $n \leq k \leq n+m-1$.

All the computations were done on a UNIX workstation with a Fortran 90 compiler. It is simple to measure the errors of the solutions in the discrete p-norms for $p \in [1, \infty]$.

$$|f|_{p,\mathbb{I}} := \begin{cases} \left\{ h \sum_{j=0}^{n-1} |f^{(s)}(x_j) - (\mathcal{T}_h^{(s)} f)(x_j)|^p \right\}^{1/p}, & 1 \leq p < \infty, \\ \max\{|f^{(s)}(x_j) - (\mathcal{T}_h^{(s)} f)(x_j)| : 0 \leq j \leq n-1\}, & p = \infty. \end{cases}$$

Similar to Theorem 4.1, we have the following estimates for the above errors.

COROLLARY 5.1. *If the hypotheses of Theorem* 3.3 *hold,* α *and* β_p *are given in Theorem* 3.3*,* $\mathbb{I} \subset \mathbb{R}$*, and* $p \in [1, \infty]$*, then*

$$|f^{(s)} - \mathcal{T}_h^{(s)} f|_{p,\mathbb{I}} \leq \beta_p \exp(-\alpha^2/2) \|f\|_2.$$

Proof. It is obvious for $p \in [1, \infty)$ that

$$\left\{ h \sum_{j=0}^{n-1} |f^{(s)}(x_j) - (\mathcal{T}_h^{(s)} f)(x_j)|^p \right\}^{1/p} \leq (nh)^{1/p} \|f^{(s)} - \mathcal{T}_h^{(s)} f\|_\infty = |\mathbb{I}|^{1/p} \|f^{(s)} - \mathcal{T}_h^{(s)} f\|_\infty,$$

and for $p = \infty$ that

$$|f^{(s)} - \mathcal{T}_h^{(s)} f|_{\infty,\mathbb{I}} \leq \|f^{(s)} - \mathcal{T}_h^{(s)} f\|_\infty.$$

Therefore, the conclusion follows from Theorem 3.3. □

First, we compare the localized sampling with the WKS sampling series. The accuracy of approximation achieved by the localized sampling in Example 1 is denoted by LS, and the corresponding approximation accuracy obtained by using the truncated version of the cardinal series (2),

$$(\mathcal{C}_{h,m} f)(x) := \sum_{k \in \mathbb{Z}_m(x)} f(kh) \operatorname{sinc}(h^{-1}x - k), \quad x \in \mathbb{R},$$

is denoted by CS. The numerical results of LS and CS approximating f' are reported and compared in Table 1, from which the advantage of the localization with Gaussian multiplier is obvious.

Table 1
Approximation for $f'(x)$ on $x \in [-5,5]$, Example 1.

			LS		CS	
Nodes	m	r	L^2 error	L^∞ error	L^2 error	L^∞ error
100	30	3.0	2.07(−7)	2.34(−7)	1.41(−1)	1.67(−1)
		3.3	3.20(−8)	3.65(−8)		
		3.6	4.92(−9)	5.50(−9)		
		3.9	7.47(−10)	8.02(−10)		
200	35	3.0	1.77(−13)	2.69(−13)	2.40(−1)	2.82(−1)
		3.3	2.98(−14)	5.84(−14)		
		3.6	2.86(−14)	5.41(−14)		
		3.9	2.74(−14)	5.11(−14)		

Table 2
Approximation for $f(x) := \operatorname{sinc}^7(x)$ on $x \in [-5,5]$, Example 2.

				$f'(x)$		$f''(x)$	
Nodes	m	r	r_1	L^2 error	L^∞ error	L^2 error	L^∞ error
100	30	2	1	2.15(−6)	3.57(−6)	1.72(−4)	3.09(−4)
			1.1	1.50(−5)	2.55(−5)	3.73(−5)	6.06(−5)
			1.2	1.10(−4)	1.90(−4)	3.31(−4)	5.56(−4)
		3	1	9.39(−12)	1.60(−11)	3.30(−10)	4.19(−10)
			1.1	4.07(−10)	6.09(−10)	2.31(−9)	2.99(−9)
			1.2	1.96(−8)	3.018(−8)	1.12(−7)	1.52(−7)
200	30	2	1	1.42(−7)	2.39(−7)	2.86(−5)	6.00(−5)
			1.1	1.66(−6)	2.74(−6)	7.86(−7)	1.25(−6)
			1.2	1.81(−5)	2.93(−5)	1.55(−5)	2.93(−5)
		3	1	4.43(−14)	8.90(−14)	3.53(−12)	1.54(−11)
			1.1	1.20(−13)	2.12(−13)	3.26(−12)	1.24(−11)
			1.2	1.72(−11)	3.00(−11)	3.25(−11)	5.89(−11)

Table 3
Approximation for $f'(x)$ on $x \in [-3,3]$, Example 4.

			LMS		MS	
Nodes	m	r	L^2 error	L^∞ error	L^2 error	L^∞ error
100	30	3	2.82(−9)	4.32(−9)	4.05(−3)	4.75(−3)
		3.3	1.63(−10)	2.37(−10)		
		3.6	1.37(−11)	2.17(−11)		
		3.9	9.87(−12)	2.00(−11)		
200	32	3	1.89(−12)	3.30(−12)	2.57(−2)	3.01(−2)
		3.3	6.04(−14)	1.49(−13)		
		3.6	5.71(−14)	1.52(−13)		
		3.9	5.60(−14)	1.55(−13)		

Table 2 reports the numerical results obtained by using the kernel in Example 2. It is observed for $K := r_1 \operatorname{sinc}(r_1 \cdot)$ that if $r_1 \sim 1$ and h decreases, then the error bounds decrease to the double precision accuracy. The computational results confirm the theoretical conclusions of Theorem 4.2.

Finally, we choose $d := 3/4$ in Example 4. The accuracy of approximation achieved by using the localized sampling series with kernel (21) is denoted by LMS, and the corresponding approximation accuracy obtained by using the sampling without localization is denoted by MS. The numerical results of LMS and MS are reported and compared in Table 3. It demonstrates the superior efficiency of the localized sampling in approximation.

For any fixed parameters in Tables 1, 2, and 3, the corresponding L^2 error and L^∞ error are basically of the same order. This is in good agreement with the theoretical conclusions given in Corollary 5.1.

6. Conclusion.

We give a sufficient condition for the kernel K used in (6), so that the localized sampling series can achieve exponentially decaying accuracy for approximating band-limited functions and their derivatives.

For application, our result provides for $p \in [1, \infty]$ the L^p estimates for approximation on finite domain. Numerical experiments demonstrate the high accuracy achieved by using the localized sampling series.

The general kernel obtained here includes some known kernel functions, such as the Shannon kernel, the oversampled Shannon kernel, the modified sinc kernel, and the de la Vallée Poussin kernel, as special cases.

Furthermore, the general kernel sampling gives new results on the interpolating Meyer wavelets. Note that by the localization with Gaussian function, the sampling series achieves much faster convergence for approximation.

Acknowledgments. The first author is very thankful to Professor Say Song Goh, Professor Charles A Micchelli, and Professor Hidemitsu Ogawa for their helpful discussion, which was invaluable to the investigation and presentation of the results.

REFERENCES

[1] I. J. Schoenberg, *Cardinal Spline Interpolation*, CBMS-NSF Regional Conf. Ser. in Appl. Math., 12, SIAM, Philadelphia, 1973.

[2] E. T. Whittaker, *On the functions which are represented by the expansion of the interpolation theory*, Proc. Roy. Soc. Edinburgh Sect. A, 35 (1915), pp. 181–194.

[3] V. Kotel'nikov, *On the carrying capacity of the "ether" and wire in telecommunications, material for the first All-Union Conference on Questions of Communications*, Izd. Red. Upr. Svyazi RKKA, Moscow, 1933 (in Russian).

[4] C. E. Shannon, *Communications in the presence of noise,* Proc. I.R.E., 37 (1949), pp. 10–21.

[5] R. J. Marks II, *Introduction to Shannon Sampling and Interpolation Theory*, Springer-Verlag, New York, 1991.

[6] A. I. Zayed, *Advances in Shannon's Sampling Theory*, CRC Press, Boca Raton, FL, 1993.

[7] J. R. Higgins, *Sampling theory in Fourier and signal analysis: Foundations*, Oxford University Press, New York, 1996.

[8] J. J. Benedetto, *Harmonic Analysis and Applications*, CRC Press, Boca Raton, FL, 1997.

[9] M. Theis, *Über eine Interpolationsformel von de la Vallée Poussin*, Math. Z., 3 (1919), pp. 93–113.

[10] H. D. Helms and J. B. Thomas, *Truncation error of sampling-theorem expansions*, Proc. I.R.E., 50 (1962), pp. 179–184.

[11] D. Jagerman, *Bounds for truncation error of the sampling expansion*, SIAM J. Appl. Math., 14 (1966), pp. 714–723.

[12] L. L. Campbell, *Sampling theorem for the Fourier transform of a distribution with bounded support*, SIAM J. Appl. Math., 16 (1968), pp. 626–636.

[13] A. J. Lee, *Approximate interpolation and the sampling theorem*, SIAM J. Appl. Math., 32 (1977), pp. 731–744.

[14] R. Gervais, Q. I. Rahman, and G. Schmeisser, *A bandlimited function simulating a duration-limited one*, in Anniversary Volume on Approximation Theory and Functional Analysis, Birkhäuser, Basel, 1984, pp. 355–362.

[15] P. L. Butzer and R. L. Stens, *A modification of the Whittaker-Kotel'nikov-Shannon sampling series,* Aequationes Math., 28 (1985), pp. 305–311.

[16] D. K. Hoffman, G. W. Wei, D. S. Zhang, and D. J. Kouri, *Shannon-Gabor wavelet distributed approximating functional*, Chem. Phys. Letters, 287 (1998), pp. 119–124.

[17] K. M. Flornes, Y. Lyubarskii, and K. Seip, *A direct interpolation method for irregular sampling*, Appl. Comput. Harmon. Anal., 7 (1999), pp. 305–314.

[18] H. G. Feichtinger and K. Gröchenig, *Irregular sampling theorems and series expansions of band-limited functions*, J. Math. Anal. Appl., 167 (1992), pp. 530–556.

[19] T. Werther, *Reconstruction from Irregular Samples with Improved Locality*, Master's thesis, University of Vienna, Vienna, Austria, 2000.

[20] L. W. Qian, *The Regularized WKS Sampling Theorem and Its Application to the Numerical Solutions of Partial Differential Equations*, Ph.D. thesis, National University of Singapore, Singapore, 2004.

[21] A. G. García and A. Portal, *Hypercircle inequalities and sampling theory*, Appl. Anal., 82 (2003), pp. 1111–1125.

[22] L. W. Qian, *On the regularized WKS sampling formula*, Proc. Amer. Math. Soc., 131 (2003), pp. 1169–1176.

[23] L. W. Qian and D. B. Creamer, *A modification of the sampling series with a Gaussian multiplier*, Samp. Theory Signal Image Process, to appear.

[24] L. W. Qian and D. B. Creamer, *Localized over-sampling series and its numerical application*, in Proceedings of the International Conference on Scientific and Engineering Computation, Singapore, 2004.

[25] L. W. Qian and H. Ogawa, *Modified sinc kernels for the localized sampling series*, Samp. Theory Signal Image Process., 4 (2005), pp. 121–139.

[26] I. R. H. Jackson, *An order of convergence for some radial functions*, IMA J. Numer. Anal., 9 (1989), pp. 567–587.

[27] L. W. Qian and D. B. Creamer, *Generalized de la Vallée Poussin kernel and its numerical application*, Appl. Numer. Math., submitted.

[28] D. S. Mitrinovic, *Analytic Inequalities*, Springer-Verlag, Berlin, 1970.

[29] E. D. Rainville, *Special Functions*, Macmillan, New York, 1960.

[30] I. Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 61, SIAM, Philadelphia, 1992.

[31] G. G. Walter, *A sampling theorem for wavelet subspaces*, IEEE Trans. Inform. Theory, 38 (1992), pp. 881–884.

[32] N. Atreas and C. Karanikas, *Truncation error on wavelet sampling expansions*, J. Comput. Anal. Appl., 2 (2000), pp. 89–102.

[33] X. P. Shen, *A quadrature formula based on sampling in Meyer wavelet subspaces*, J. Comput. Anal. Appl., 3 (2001), pp. 147–163.

SIAM J. Numer. Anal.
Vol. 43, No. 6, pp. 2517–2543

LEAST-SQUARES FINITE ELEMENT METHODS FOR OPTIMALITY SYSTEMS ARISING IN OPTIMIZATION AND CONTROL PROBLEMS[*]

PAVEL BOCHEV[†] AND MAX D. GUNZBURGER[‡]

Abstract. The approximate solution of optimization and optimal control problems for systems governed by linear, elliptic partial differential equations is considered. Such problems are most often solved using methods based on applying the Lagrange multiplier rule to obtain an optimality system consisting of the state system, an adjoint-state system, and optimality conditions. Galerkin methods applied to this system result in indefinite matrix problems. Here, we consider using modern least-squares finite element methods for the solution of the optimality systems. The matrix equations resulting from this approach are symmetric and positive definite and are readily amenable to uncoupling strategies. This is an important advantage of least-squares principles as they allow for a more efficient computational solution of the optimization problem. We develop an abstract theory that includes optimal error estimates for least-squares finite element methods applied to optimality systems. We then provide an application of the theory to optimization problems for the Stokes equations.

Key words. optimal control, optimization, least-squares finite element methods, optimality systems, Lagrange multipliers

AMS subject classifications. 65N30, 65N22, 49J20, 49K20

DOI. 10.1137/040607848

1. Introduction.

Optimization and control problems for systems governed by partial differential equations arise in many applications. Experimental studies of such problems go back 100 years [22], and computational approaches have been applied since the advent of the computer age. Most of the efforts in the latter direction have employed elementary optimization strategies, but more recently, there has been considerable practical and theoretical interest in the application of sophisticated local and global optimization strategies, e.g., Lagrange multiplier methods, sensitivity or adjoint-based gradient methods, quasi-Newton methods, evolutionary algorithms, etc.

The optimal control or optimization problems we consider consist of

- *state variables*, i.e., variables that describe the system being modeled;
- *control variables* or *design parameters*, i.e., variables at our disposal that can be used to affect the state variables;
- a *state system*, i.e., partial differential equations relating the state and control variables; and
- a *functional* of the state and control variables whose minimization is the goal.

Then, the problems we consider consist of finding state and control variables that minimize the given functional subject to the state system being satisfied. Here, we restrict attention to linear, elliptic state systems and to quadratic functionals.

[*] Received by the editors May 7, 2004; accepted for publication (in revised form) June 13, 2005; published electronically January 12, 2006.
http://www.siam.org/journals/sinum/43-6/60784.html

[†] Computational Mathematics and Algorithms Department, Sandia National Laboratories, Albuquerque, NM 87185-1110 (pbboche@sandia.gov). Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy under contract DE-AC04-94AL85000.

[‡] School of Computational Science, Florida State University, Tallahassee FL 32306-4120 (gunzburg@csit.fsu.edu). The research of this author was supported in part by CSRI, Sandia National Laboratories under contract 18407.

The Lagrange multiplier rule is a standard approach for solving finite-dimensional, constrained optimization problems. It is not surprising then that several popular approaches to solving optimization and control problems constrained by partial differential equations are also based on solving optimality systems deduced from the application of the Lagrange multiplier rule. The optimality system consists of

- the *state system*, i.e., the given partial differential equations that relate the unknown state and control variables;
- an *adjoint* or *costate system* which is also partial differential equations involving the adjoint operator of the state system; and
- an *optimality condition* that reflects the fact that the gradient of the functional vanishes for optimal values of the state and control variables.

The three components of the optimality system are coupled. In the linear constraints/quadratic functional context we consider in this paper, the optimality system, viewed as a coupled system, is a symmetric and weakly coercive linear system in the state, adjoint-state, and control variables.

In the context of finite element methods, optimality systems are usually discretized using Galerkin methods, resulting in typical saddle-point-type matrix problems that are symmetric and indefinite. In many if not most practical situations, the coupled optimality system is a formidable system to solve; compared to solving direct problems involving the state system alone, discrete optimality systems typically involve at least double the number of unknowns. For this reason, many approaches have been proposed for decoupling, through iterative processes, the different components of the optimality system. An extensive discussion of several such strategies in both an abstract setting and for fluid flow problems can be found in [18].

In this paper, we discuss the use of modern least-squares finite element methods for finding approximate solutions of the optimality system. The resulting matrix problems are symmetric and *positive definite*. Moreover, their diagonal blocks are also symmetric and positive definite, thus opening up better possibilities for devising efficient uncoupling methods than is the case for Galerkin discretizations. In order to develop a basic theory for least-squares finite element methods for optimization and control problems, we focus on treating the optimality system as a fully coupled system and only briefly discuss the application of decoupling strategies. Nevertheless, the reader should keep in mind that amenability to efficient uncoupling strategies is, perhaps, the chief reason to consider the application of least-squares principles to optimization problems. The application of least-squares principles to optimality systems was previously discussed, in a concrete setting, in [19].

The approach we have described for finding approximate solutions of optimal control and optimization problems for partial differential equations is of the *differentiate-then-discretize* or *optimize-then-discretize* type. One first applies, at the continuous partial differential equations level, the first-order necessary conditions for finding saddle points of a Lagrangian functional, and then one uses a finite element method, be it of Galerkin or of least-square type, to discretize the resulting optimality system. For the alternative *discretize-then-differentiate* or *discretize-then-optimize* type approach, one reverses the steps: One first discretizes the optimization or control problem by some means and then applies the Lagrange multiplier rule to the resulting discrete optimization problem. The two steps do not, in general, commute so that the discrete systems determined by the two approaches are not the same. A discussion of the relative merits of the two approaches can be found, e.g., in [18]. Here, we focus on the differentiate-then-discretize approach.

Instead of using the Lagrange multiplier rule for solving constrained optimization problems, one may use a penalty method. Penalty/least-squares finite element methods are the subject of the companion paper [9]; see also [20]. Other applications of least-squares finite element methods to optimization problems may be found in [2,3,5,8].

The paper is organized as follows. In section 3, we study, in an abstract setting, Lagrange multiplier methods for quadratic optimization and control problems constrained by linear, elliptic partial differential equations. In section 4, we study least-squares finite element methods for the approximate solution of the optimality system resulting from the application of the Lagrange multiplier rule. In section 5, we provide concrete examples that illustrate the theory of sections 3 and 4. In passing, we briefly remark on several related topics, including decoupling strategies for the solution of the discretized optimality system. Before we embark, however, in section 2, we present mostly well-known results about general constrained optimization problems and their solution via Lagrange multiplier methods. These results serve as the foundation for the considerations of sections 3 and 4. We remark that in several inequalities appearing in the paper, C denotes a positive constant whose value changes with context but that is independent of any of the data or solution functions appearing in the inequalities.

2. Linearly constrained quadratic minimization problems in Hilbert spaces.

In this section, we review the now classical theory (see [12] and also [13,16]) for finite element methods for constrained quadratic minimization problems. The optimization and control problems that are the subject of this paper can be profitably viewed as special cases of the types of problems treated by the classical theory.

Given Hilbert spaces V and S along with their dual spaces V^* and S^*, respectively, the symmetric bilinear form $a(\cdot,\cdot)$ on $V \times V$, the bilinear form $b(\cdot,\cdot)$ on $V \times S$, the functions $f \in V^*$ and $g \in S^*$, and the real number t, we define the *quadratic functional*[1]

$$\mathcal{J}(u) = \frac{1}{2}a(u,u) - \langle f,u\rangle_{V^*,V} + t \qquad \forall\, u \in V, \tag{2.1}$$

the *linear constraint equation*

$$b(u,q) = \langle g,q\rangle_{S^*,S} \qquad \forall\, q \in S, \tag{2.2}$$

where $\langle\cdot,\cdot\rangle$ denotes an appropriate duality pairing, and the *constrained minimization problem*[2]

$$\min_{u\in V} \mathcal{J}(u) \quad \text{subject to (2.2)}. \tag{2.3}$$

[1] The value of t does not affect the minimizer of $\mathcal{J}(\cdot)$. We include it in the definition of $\mathcal{J}(u)$ only to facilitate, in later sections, the identification of concrete functionals with the abstract functional (2.1).

[2] Such problems arise in many applications. A classical example is provided by the functional $\mathcal{J}(\mathbf{v}) = \frac{1}{2}\int_\Omega |\mathbf{v}|^2\, d\Omega$, the linear constraint $\nabla\cdot\mathbf{v} = f$, and the minimization problem $\min J(\mathbf{v})$ subject to $\nabla\cdot\mathbf{v} = f$ in Ω, where the minimization is effected over a suitable function space. For example, in fluid mechanics, this problem is known as the *Kelvin principle* and, in structural mechanics (where $\mathbf{v}$ is a tensor), as the *complimentary energy principle.* For the Kelvin principle, S is the space $L^2(\Omega)$ of all square integrable functions, and V is the space $H(\mathrm{div},\Omega)$ of all square integrable vector fields whose divergencies are also square integrable, $a(\mathbf{u},\mathbf{v}) = \int_\Omega \mathbf{u}\cdot\mathbf{u}\, d\Omega$, $b(\mathbf{v},q) = \int_\Omega q\nabla\cdot\mathbf{v}\, d\Omega$. Also, the operators A and B defined below are the identity and divergence operators, respectively.

The bilinear forms serve to define the associated operators

$$A : V \to V^*, \qquad B : V \to S^*, \qquad \text{and} \qquad B^* : S \to V^* \tag{2.4}$$

through the relations

$$\begin{cases} a(u,v) = \langle Au, v\rangle_{V^*,V} & \forall\, u,v \in V, \\ b(v,q) = \langle Bv, q\rangle_{S^*,S} = \langle B^*q, v\rangle_{V^*,V} & \forall\, v \in V,\, q \in S. \end{cases}$$

The minimization problem (2.3) can then be given the form

$$\min_{u\in V} \mathcal{J}(u) \quad \text{subject to} \quad Bu = g,$$

where the constraint equation $Bu = g$ holds in S^*. We define the subspace

$$Z = \{v \in V \ : \ b(v,q) = 0 \ \forall\, q \in S\} \tag{2.5}$$

and make the following assumptions about the bilinear forms:

$$\begin{cases} a(u,v) \le C_a \|u\|_V \|v\|_V & \forall\, u,v \in V, \\ b(u,q) \le C_b \|u\|_V \|q\|_S & \forall\, u \in V,\ q \in S, \\ a(u,u) \ge 0 & \forall\, u \in V, \\ a(u,u) \ge K_a \|u\|_V^2 & \forall\, u \in Z, \\ \displaystyle\sup_{v\in V, v\neq 0} \frac{b(v,q)}{\|v\|_V} \ge K_b \|q\|_S & \forall\, q \in S, \end{cases} \tag{2.6}$$

where C_a, C_b, K_a, and K_b are all positive constants.

2.1. Existence of solutions. The following result is well known; see, e.g. [21].

Proposition 2.1. *Let the assumptions* (2.6) *hold. Then, the constrained minimization problem* (2.3) *has a unique solution* $u \in V$. □

2.2. Solution via Lagrange multipliers. For all $v \in V$ and $q \in S$, we introduce the Lagrangian functional

$$\mathcal{L}(v,q) = \mathcal{J}(v) + b(v,q) - \langle g, q\rangle_{S^*,S} = \frac{1}{2} a(v,v) + b(v,q) - \langle f, v\rangle_{V^*,V} - \langle g, q\rangle_{S^*,S} + t.$$

Then, the constrained minimization problem (2.3) is equivalent to the unconstrained optimization problem of finding saddle points $(u,p) \in V \times S$ of the Lagrangian functional. These saddle points may be found by solving the optimality system

$$\begin{cases} a(u,v) + b(v,p) &= \langle f, v\rangle_{V^*,V} & \forall\, v \in V, \\ b(u,q) &= \langle g, q\rangle_{S^*,S} & \forall\, q \in S. \end{cases} \tag{2.7}$$

The following result is also well known; see, e.g., [12].

Proposition 2.2. *Let the assumptions* (2.6) *hold. Then, the system* (2.7) *has a unique solution* $(u,p) \in V \times S$. *Moreover,*

$$\|u\|_V + \|p\|_S \le C\big(\|f\|_{V^*} + \|g\|_{S^*}\big), \tag{2.8}$$

and $u \in V$ *is the unique solution of the constrained minimization problem* (2.3). □

In terms of the operators introduced in (2.4), the system (2.7) takes the form

$$\begin{cases} Au + B^*p &= f \quad \text{in } V^*, \\ Bu &= g \quad \text{in } S^*. \end{cases}$$

Remark 2.3. The unique solvability of (2.7) and the estimate (2.8) do not require that the bilinear form $a(\cdot,\cdot)$ be symmetric or that it satisfy the third condition in (2.6). Also, the fourth condition in (2.6) may be weakened to a weak coercivity condition. However, these conditions are required to make the connection between (2.7) and the constrained minimization problem (2.3). So, throughout, we will assume that all the conditions in (2.6) hold.

2.3. Galerkin approximations of the optimality system. We choose (conforming) finite-dimensional subspaces $V^h \subset V$ and $S^h \subset S$, and we then restrict (2.7) to these subspaces, i.e., we seek $u^h \in V^h$ and $p^h \in S^h$ that satisfy

$$\begin{cases} a(u^h,v^h) + b(v^h,p^h) &= \langle f, v^h\rangle_{V^*,V} \qquad \forall\, v^h \in V^h, \\ b(u^h,q^h) &= \langle g, q^h\rangle_{S^*,S} \qquad \forall\, q^h \in S^h. \end{cases} \tag{2.9}$$

This is also the optimality system for the minimization of the functional $\mathcal{J}(\cdot)$ over V^h subject to $b(u^h,q^h) = \langle g,q^h\rangle_{S^*,S}$ for all $q^h \in S^h$. Let

$$Z^h = \{v^h \in V^h \;:\; b(v^h,q^h) = 0 \;\forall\, q^h \in S^h\}.$$

In general, $Z^h \not\subset Z$ even though $V^h \subset V$ and $S^h \subset S$, and so the last two assumptions in (2.6) may not be satisfied with respect to the subspaces. If V^h and S^h are such that the last two assumptions hold, then one obtains the following well-known result; see, e.g., [12].

PROPOSITION 2.4. *Let the hypotheses of Proposition* 2.1 *hold and assume that*

$$a(u^h,u^h) \geq K_a^h \|u^h\|_V^2 \quad \forall\, u^h \in Z^h \tag{2.10}$$

and

$$\sup_{v^h \in V^h, v^h \neq 0} \frac{b(v^h,q^h)}{\|v^h\|_V} \geq K_b^h \|q^h\|_S \quad \forall\, q^h \in S^h, \tag{2.11}$$

where K_a^h and K_b^h are positive constants independent of h. Then, the discrete system (2.9) *has a unique solution $(u^h,p^h) \in V^h \times S^h$, and moreover*

$$\|u^h\|_V + \|p^h\|_S \leq C\big(\|f\|_{V^*} + \|g\|_{S^*}\big).$$

Furthermore, if $(u,p) \in V \times S$ denotes the unique solution of (2.7), *then*

$$\|u-u^h\|_V + \|p-p^h\|_S \leq C\Big(\inf_{v^h \in V^h} \|u-v^h\|_V + \inf_{q^h \in S^h} \|p-q^h\|_S\Big). \tag{2.12}$$

The discrete problem (2.9) is equivalent to a linear system. Indeed, let $\{U_i\}_{i=1}^n$ and $\{P_i\}_{i=1}^m$, where $n = \dim(V^h)$ and $m = \dim(S^h)$, denote bases for V^h and S^h, respectively, and let $\vec{\mathbf{u}} = (u_1,\dots,u_n)^T$ and $\vec{\mathbf{p}} = (p_1,\dots,p_m)^T$ denote the vectors of coefficients in the expansions of u^h and p^h in terms of the respective bases. Furthermore, let $f_i = \langle f, U_i\rangle_{V^*,V}$ for $i = 1,\dots,n$, $g_i = \langle g, P_i\rangle_{S^*,S}$ for $i = 1,\dots,m$,

$\vec{\mathbf{f}} = (f_1, \dots, f_n)^T$, and $\vec{\mathbf{g}} = (g_1, \dots, g_m)^T$, and define the elements of the $n \times n$ matrix $\mathbb{A}$ and the $m \times n$ matrix $\mathbb{B}$ by $\mathbb{A}_{ij} = a(U_i, U_j)$ for $i, j = 1, \dots, n$ and $\mathbb{B}_{ij} = b(U_j, P_i)$ for $i = 1, \dots, m$, $j = 1, \dots, n$, respectively. Then, (2.9) is equivalent to the linear system

$$\begin{pmatrix} \mathbb{A} & \mathbb{B}^T \\ \mathbb{B} & 0 \end{pmatrix} \begin{pmatrix} \vec{\mathbf{u}} \\ \vec{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{g}} \end{pmatrix}. \tag{2.13}$$

Remark 2.5. The coefficient matrix in (2.13) is symmetric and indefinite. This is universal for discretizations of saddle-point problems arising from the use of the Lagrange multiplier rule for constrained optimization problems.

Remark 2.6. The assumptions (2.10) and (2.11) guarantee that the $(m+n) \times (m+n)$ coefficient matrix in (2.13) is invertible and that the norms of its inverse are bounded from above independently of m and n, i.e., independently of the grid size h.

Remark 2.7. The observations made in Remark 2.3 about the bilinear form $a(\cdot, \cdot)$ and (2.7) also apply to (2.9).

3. Quadratic optimization and control problems in Hilbert spaces with linear constraints.

In this section, we specialize the results of section 2 to the type of optimization and control problems described in section 1. We identify the variable u of section 2 with the pair (ϕ, θ), where ϕ and θ are the state and control variables, respectively, of the control problem.

We begin with four given Hilbert spaces Θ, Φ, $\widehat{\Phi}$, and $\widetilde{\Phi}$ along with their dual spaces denoted by $(\cdot)^*$. We assume that $\Phi \subseteq \widehat{\Phi} \subseteq \widetilde{\Phi}$ with continuous embeddings and that $\widetilde{\Phi}$ acts as the pivot space for both the pair $\{\Phi^*, \Phi\}$ and the pair $\{\widehat{\Phi}^*, \widehat{\Phi}\}$ so that we have not only that $\Phi \subseteq \widehat{\Phi} \subseteq \widetilde{\Phi} \subseteq \widehat{\Phi}^* \subseteq \Phi^*$ but also that

$$\langle \psi, \phi \rangle_{\Phi^*, \Phi} = \langle \psi, \phi \rangle_{\widehat{\Phi}^*, \widehat{\Phi}} = (\psi, \phi)_{\widetilde{\Phi}} \quad \forall\, \psi \in \widehat{\Phi}^* \subseteq \Phi^* \quad \text{and} \quad \forall\, \phi \in \Phi \subseteq \widehat{\Phi}, \tag{3.1}$$

where $(\cdot, \cdot)_{\widetilde{\Phi}}$ denotes the inner product on $\widetilde{\Phi}$.

Next, we define the *quadratic functional*

$$\mathcal{J}(\phi, \theta) = \frac{1}{2} a_1(\phi - \widehat{\phi}, \phi - \widehat{\phi}) + \frac{1}{2} a_2(\theta, \theta) \qquad \forall\, \phi \in \Phi,\ \theta \in \Theta, \tag{3.2}$$

where $a_1(\cdot, \cdot)$ and $a_2(\cdot, \cdot)$ are symmetric bilinear forms on $\widehat{\Phi} \times \widehat{\Phi}$ and $\Theta \times \Theta$, respectively, and $\widehat{\phi} \in \widehat{\Phi}$ is a given function. In the language of control theory, Φ is called the *state space,* ϕ the *state variable,* Θ the *control space,* and θ the *control variable.* In many applications, the control space is finite dimensional in which case θ is often referred to as the vector of *design variables.* We note that often Θ is chosen to be a bounded set in a Hilbert space, but for our purposes, we consider the less general situation of Θ itself being a Hilbert space. We make the following assumptions about the bilinear forms $a_1(\cdot, \cdot)$ and $a_2(\cdot, \cdot)$:

$$\begin{cases} a_1(\phi, \mu) \le C_1 \|\phi\|_{\widehat{\Phi}} \|\mu\|_{\widehat{\Phi}} & \forall\, \phi, \mu \in \widehat{\Phi}, \\ a_2(\theta, \nu) \le C_2 \|\theta\|_{\Theta} \|\nu\|_{\Theta} & \forall\, \theta, \nu \in \Theta, \\ a_1(\phi, \phi) \ge 0 & \forall\, \phi \in \widehat{\Phi}, \\ a_2(\theta, \theta) \ge K_2 \|\theta\|_{\Theta}^2 & \forall\, \theta \in \Theta, \end{cases} \tag{3.3}$$

where C_1, C_2, and K_2 are all positive constants. The second term in the functional (3.2) can be interpreted as a penalty term which limits the size of the control θ.

Given another Hilbert space Λ, the additional bilinear forms $b_1(\cdot,\cdot)$ on $\Phi\times\Lambda$ and $b_2(\cdot,\cdot)$ on $\Theta\times\Lambda$, and the function $g\in\Lambda^*$, we define the *linear constraint equation*

$$b_1(\phi,\psi)+b_2(\theta,\psi)=\langle g,\psi\rangle_{\Lambda^*,\Lambda}\quad\forall\,\psi\in\Lambda. \tag{3.4}$$

We make the following assumptions about the bilinear forms $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$:

$$\begin{cases} b_1(\phi,\psi)\le c_1\|\phi\|_\Phi\|\psi\|_\Lambda & \forall\,\phi\in\Phi,\ \ \psi\in\Lambda,\\[2mm] b_2(\theta,\psi)\le c_2\|\psi\|_\Phi\|\theta\|_\Theta & \forall\,\theta\in\Theta,\ \ \psi\in\Lambda,\\[2mm] \displaystyle\sup_{\psi\in\Lambda,\psi\neq0}\frac{b_1(\phi,\psi)}{\|\psi\|_\Lambda}\ge k_1\|\phi\|_\Phi & \forall\,\phi\in\Phi,\\[2mm] \displaystyle\sup_{\phi\in\Phi,\phi\neq0}\frac{b_1(\phi,\psi)}{\|\phi\|_\Phi}>0 & \forall\,\psi\in\Lambda, \end{cases} \tag{3.5}$$

where c_1, c_2, and k_1 are all positive constants. These assumptions suffice to guarantee that, given any $\theta\in\Theta$, the constraint equation (3.4) is uniquely solvable for $\phi\in\Phi$; this observation easily follows from [1] and (3.5).

We consider the *optimal control problem*[3]

$$\min_{(\phi,\theta)\in\Phi\times\Theta}\mathcal{J}(\phi,\theta)\quad\text{subject to}\quad b_1(\phi,\psi)+b_2(\theta,\psi)=\langle g,\psi\rangle_{\Lambda^*,\Lambda}\quad\forall\,\psi\in\Lambda. \tag{3.6}$$

It is easy to verify that the problem (3.6) falls into the framework of section 2. To this end, we let $V\equiv\Phi\times\Theta$, $S\equiv\Lambda$, $\|\{\phi,\theta\}\|_V=\sqrt{\|\phi\|_\Phi^2+\|\theta\|_\Theta^2}$ for all $\{\phi,\theta\}\in V$,

$$\begin{cases} a\big(\{\phi,\theta\},\{\mu,\nu\}\big)\equiv a_1(\phi,\mu)+a_2(\theta,\nu) & \forall\,\phi,\mu\in\Phi,\ \ \theta,\nu\in\Theta,\\ b\big(\{\phi,\theta\},\{\psi\}\big)\equiv b_1(\phi,\psi)+b_2(\theta,\psi) & \forall\,\phi\in\Phi,\ \ \theta\in\Theta,\ \ \psi\in\Lambda,\\ \big\langle f,\{\mu,\nu\}\big\rangle_{V^*,V}\equiv a_1(\mu,\widehat{\phi}) & \forall\,\mu\in\Phi,\ \ \nu\in\Theta,\\ t=\dfrac{1}{2}a_1(\widehat{\phi},\widehat{\phi}). \end{cases} \tag{3.7}$$

From (3.3), it follows that $t\le(C_1/2)\|\widehat{\phi}\|_{\widehat{\Phi}}^2$ and, also using the continuous embedding $\Phi\subseteq\widehat{\Phi}$,

$$\frac{\langle f,\{\mu,\nu\}\rangle_{V^*,V}}{\|\{\mu,\nu\}\|_V}\le\frac{\langle f,\{\mu,\nu\}\rangle_{V^*,V}}{\|\mu\|_\Phi}=\frac{a_1(\mu,\widehat{\phi})}{\|\mu\|_\Phi}\le C_1\|\widehat{\phi}\|_{\widehat{\Phi}}\quad\forall\,\{\mu,\nu\}\in\Phi\times\Theta=V$$

so that $\|f\|_{V^*}\le C_1\|\widehat{\phi}\|_{\widehat{\Phi}}$, i.e., f does indeed belong to V^*. Then, with the obvious identifications $u=\{\phi,\theta\}$, $v=\{\mu,\nu\}$, and $q=\{\psi\}$, the functionals (2.1) and (3.2) are equivalent as are the constraint equations (2.2) and (3.4). The constrained optimization problem (2.3) and the optimal control problem (3.6) are also equivalent.

We will use the framework and results established in section 2 to study the optimal control problem (3.6). Many of the results we discuss are well known, but we repeat them here to establish a context for later discussions.

[3] In section 5 we will consider an example where the linear constraint will be the weak form of the Stokes equations of incompressible viscous flows. We draw attention to the fact that these equations themselves are another example of a problem that fits the abstract setting of section 2 with $\mathcal{J}(\mathbf{v};\mathbf{f})=\frac{1}{2}\int_\Omega|\nabla\mathbf{v}|^2\,d\Omega-\int_\Omega\mathbf{f}\cdot\mathbf{v}\,d\Omega$, $b(\mathbf{v},q)=\int_\Omega q\nabla\cdot\mathbf{v}\,d\Omega$, and $g=0$.

3.1. Existence of optimal states and controls. We begin with the following preliminary result.

LEMMA 3.1. *Let the assumptions* (3.3) *and* (3.5) *hold. Then the spaces* $V \equiv \Phi \times \Theta$ *and* $S \equiv \Lambda$ *and the bilinear forms* $a(\cdot,\cdot)$ *and* $b(\cdot,\cdot)$ *defined in* (3.7) *satisfy the assumptions* (2.6).

Proof. Using (3.3) and the continuous embedding $\Phi \subset \widehat{\Phi}$, we have that

$$\begin{aligned} a(\{\phi,\theta\},\{\mu,\nu\}) &= a_1(\phi,\mu) + a_2(\theta,\nu) \le C_1\|\phi\|_{\widehat{\Phi}}\|\mu\|_{\widehat{\Phi}} + C_2\|\theta\|_\Theta\|\nu\|_\Theta \\ &\le C_1\|\phi\|_\Phi\|\mu\|_\Phi + C_2\|\theta\|_\Theta\|\nu\|_\Theta \le \max\{C_1,C_2\}\sqrt{\|\phi\|_\Phi^2+\|\theta\|_\Theta^2}\sqrt{\|\mu\|_\Phi^2+\|\nu\|_\Theta^2}\end{aligned}$$

for all $\phi,\mu \in \Phi$, $\theta,\nu \in \Theta$ so that $a(u,v) \le C_a\|u\|_V\|v\|_V$ for all $u,v \in V$ with $C_a = \max\{C_1,C_2\}$. Similarly, we have using (3.5) that $b(u,q) \le C_b\|u\|_V\|q\|_S$ for all $u \in V$, $q \in S$ with $C_b = \max\{c_1,c_2\}$. Next, from (3.3) we have that

$$a(\{\phi,\theta\},\{\phi,\theta\}) = a_1(\phi,\phi) + a_2(\theta,\theta) \ge a_2(\theta,\theta) \ge K_2\|\theta\|_\Theta^2 \qquad \forall \phi \in \Phi,\ \ \theta \in \Theta$$

so that $a(u,u) \ge 0$ for all $u \in V$. We next define the subspace $Z \subset \Phi \times \Theta$ by

$$Z = \Big\{\{\phi,\theta\} \in \Phi \times \Theta \ :\ b_1(\phi,\psi) + b_2(\theta,\psi) = 0 \quad \forall\,\psi \in \Lambda\Big\}. \tag{3.8}$$

The assumptions (3.5) imply that, given any $\theta \in \Theta$, the problem

$$b_1(\phi,\psi) = -b_2(\theta,\psi) \quad \forall\,\psi \in \Lambda \tag{3.9}$$

has a unique solution ϕ_θ and, moreover,

$$\|\phi_\theta\|_\Phi \le \frac{c_2}{k_1}\|\theta\|_\Theta; \tag{3.10}$$

see, e.g., [1]. Thus, Z can be completely characterized by $(\phi_\theta,\theta) \in \Phi \times \Theta$, where for arbitrary $\theta \in \Theta$, ϕ_θ is the solution of (3.9). Then, (3.10) and (3.3) imply that

$$\begin{aligned} a(\{\phi_\theta,\theta\},\{\phi_\theta,\theta\}) &= a_1(\phi_\theta,\phi_\theta) + a_2(\theta,\theta) \ge a_2(\theta,\theta) \ge K_2\|\theta\|^2 \\ &\ge \frac{K_2}{2}\min\left\{1,\frac{k_1^2}{c_2^2}\right\}\|\{\phi_\theta,\theta\}\|_V \qquad \forall\,\{\phi_\theta,\theta\} \in Z.\end{aligned}$$

As a result, $a(u,u) \ge K_a\|u\|_V^2$ for all $u \in Z$ with $K_a = \frac{1}{2}K_2\min\{1,\frac{k_1^2}{c_2^2}\}$ so that the third assumption in (2.6) is also satisfied.

To verify the last assumption in (2.6), note that

$$\sup_{\phi\in\Phi,\phi\neq 0}\frac{b_1(\phi,\psi)}{\|\phi\|_\Phi} \ge k_1\|\psi\|_\Lambda \qquad \forall\,\psi \in \Lambda. \tag{3.11}$$

Indeed, assumptions (3.5) imply that (see [1]), for any $\psi \in \Lambda$, the problem

$$b_1(\phi,\mu) = (\psi,\mu)_\Lambda \qquad \forall\,\mu \in \Lambda \tag{3.12}$$

has a unique solution ϕ_ψ and, moreover,

$$\|\phi_\psi\|_\Phi \le \frac{1}{k_1}\|\psi\|_\Lambda. \tag{3.13}$$

Using (3.12) and (3.13), it is easy to see that

$$\frac{b_1(\phi_\psi, \psi)}{\|\phi_\psi\|_\Phi} = \frac{\|\psi\|_\Lambda^2}{\|\phi_\psi\|_\Phi} \geq k_1 \|\psi\|_\Lambda \qquad \forall\, \psi \in \Lambda$$

which immediately implies (3.11). Finally, using (3.11),

$$\sup_{(\phi,\theta)\in\Phi\times\Theta,\, (\phi,\theta)\neq(0,0)} \frac{b(\{\phi,\theta\},\{\psi\})}{\sqrt{\|\phi\|_\Phi^2 + \|\theta\|_\Theta^2}} \geq \sup_{\phi\in\Phi,\, \phi\neq 0} \frac{b_1(\phi,\psi)}{\|\phi\|_\Phi} \geq k_1 \|\psi\|_\Lambda \qquad \forall\, \psi \in \Lambda$$

so that

$$\sup_{u\in V,\, u\neq 0} \frac{b(u,q)}{\|u\|_V} \geq K_b \|q\|_S \quad \forall\, q \in S$$

with $K_b = k_1$. □

Having verified the assumptions (2.6) for the optimal control problem (3.6), we immediately have the following result.

THEOREM 3.2. *Let the assumptions* (3.3) *and* (3.5) *hold. Then, the optimal control problem* (3.6) *has a unique solution* $(\phi, \theta) \in \Phi \times \Theta$.

Proof. The result immediately follows from Proposition 2.1 and Lemma 3.1. □

It is instructive to rewrite the functional (3.2), the constraint (3.4), and the optimal control problem (3.6) in operator notation. To this end, we note that the bilinear forms serve to define operators

$$\begin{array}{lll} A_1 : \widehat{\Phi} \to \widehat{\Phi}^*, & A_2 : \Theta \to \Theta^*, & B_1 : \Phi \to \Lambda^*, \\ B_1^* : \Lambda \to \Phi^*, & B_2 : \Theta \to \Lambda^*, & B_2^* : \Lambda \to \Theta^* \end{array}$$

through the following relations:

$$(3.14) \qquad \begin{cases} a_1(\phi,\mu) = \langle A_1\phi, \mu\rangle_{\widehat{\Phi}^*,\widehat{\Phi}} & \forall\, \phi,\mu \in \widehat{\Phi}, \\ a_2(\theta,\nu) = \langle A_2\theta, \nu\rangle_{\Theta^*,\Theta} & \forall\, \theta,\nu \in \Theta, \\ b_1(\phi,\psi) = \langle B_1\phi, \psi\rangle_{\Lambda^*,\Lambda} = \langle B_1^*\psi, \phi\rangle_{\Phi^*,\Phi} & \forall\, \phi \in \Phi,\;\; \psi \in \Lambda, \\ b_2(\psi,\theta) = \langle B_2\theta, \psi\rangle_{\Lambda^*,\Lambda} = \langle B_2^*\psi, \theta\rangle_{\Theta^*,\Theta} & \forall\, \theta \in \Theta,\;\; \psi \in \Lambda. \end{cases}$$

Then, the functional (3.2) and the constraint (3.4) take the forms

$$(3.15) \qquad \mathcal{J}(\phi,\theta) = \frac{1}{2}\big\langle A_1(\phi-\widehat{\phi}), (\phi-\widehat{\phi})\big\rangle_{\widehat{\Phi}^*,\widehat{\Phi}} + \frac{1}{2}\langle A_2\theta, \theta\rangle_{\Theta^*,\Theta} \qquad \forall\, \phi\in\Phi,\, \theta\in\Theta$$

and

$$(3.16) \qquad B_1\phi + B_2\theta = g \qquad \text{in } \Lambda^*,$$

respectively, and the optimal control problem (3.6) takes the form

$$(3.17) \qquad \min_{(\phi,\theta)\in\Phi\times\Theta} \mathcal{J}(\phi,\theta) \quad \text{subject to (3.16).}$$

Assumptions (3.3) and (3.5) imply that A_1, A_2, B_1, B_2, B_1^*, and B_2^* are bounded with

$$\begin{array}{lll} \|A_1\|_{\widehat{\Phi}\to\widehat{\Phi}^*} \leq C_1, & \|A_2\|_{\Theta\to\Theta^*} \leq C_2, & \|B_1\|_{\Phi\to\Phi^*} \leq c_1, \\ \|B_1^*\|_{\Phi\to\Phi^*} \leq c_1, & \|B_2\|_{\Phi\to\Theta^*} \leq c_2, & \|B_2^*\|_{\Theta\to\Phi^*} \leq c_2 \end{array}$$

and that the operator B_1 is invertible with $\|B_1^{-1}\|_{\Lambda^*\to\Phi} \leq 1/k_1$. Note also that the subspace $Z \subset V = \Phi \times \Theta$ can be defined by

$$Z = \Big\{ \{\phi,\theta\} \in \Phi\times\Theta \;:\; \phi = -B_1^{-1}B_2\theta \quad \forall\, \theta \in \Theta \Big\}.$$

3.2. Solution via Lagrange multipliers and the optimality system. For all $\{\mu, \nu\} \in V = \Phi \times \Theta$ and $\psi \in S = \Lambda$, we introduce the Lagrangian functional

$$\begin{aligned}\mathcal{L}(\{\mu,\nu\},\{\psi\}) &= \mathcal{J}(\{\mu,\nu\}) + b(\{\mu,\nu\},\{\psi\}) - \langle g,\psi\rangle_{\Lambda^*,\Lambda}\\ &= \frac{1}{2}a_1(\mu-\widehat{\phi},\mu-\widehat{\phi}) + \frac{1}{2}a_2(\nu,\nu) + b_1(\mu,\psi) + b_2(\nu,\psi) - \langle g,\psi\rangle_{\Lambda^*,\Lambda}.\end{aligned}$$

Then, (3.6) is equivalent to the unconstrained optimization problem of finding saddle points $(\{\phi,\theta\},\{\lambda\})$ in $V \times S$ of the Lagrangian functional. These saddle points may be found by solving the *optimality system*

$$\left\{\begin{array}{lllllll} a_1(\phi,\mu) & & & + & b_1(\mu,\lambda) & = & a_1(\widehat{\phi},\mu) \quad \forall\,\mu\in\Phi,\\ & & a_2(\theta,\nu) & + & b_2(\nu,\lambda) & = & 0 \quad \forall\,\nu\in\Theta,\\ b_1(\phi,\psi) & + & b_2(\theta,\psi) & & & = & \langle g,\psi\rangle_{\Lambda^*,\Lambda} \quad \forall\,\psi\in\Lambda.\end{array}\right. \tag{3.18}$$

The third equation in the optimality system (3.18) is simply the constraint equation. The first equation is commonly referred to as the *adjoint* or *costate equation* and the Lagrange multiplier λ is referred as the *adjoint* or *costate* variable. The second equation in (3.18) is referred to as the *optimality condition* since it is merely a statement that the gradient of the functional $\mathcal{J}(\cdot,\cdot)$ defined in (3.2) vanishes at the optimum.

Using the framework of section 2.2, the following result is immediate.

THEOREM 3.3. *Let the assumptions* (3.3) *and* (3.5) *hold. Then, the optimality system* (3.18) *has a unique solution* $(\phi,\theta,\lambda) \in \Phi\times\Theta\times\Lambda$. *Moreover*

$$\|\phi\|_\Phi + \|\theta\|_\Theta + \|\lambda\|_\Lambda \le C\big(\|g\|_{\Lambda^*} + \|\widehat{\phi}\|_{\widehat{\Phi}}\big),$$

and $(\phi,\theta)\in\Phi\times\Theta$ *is the unique solution of the optimal control problem* (3.6).

Proof. With the associations $V=\Phi\times\Theta$, $S=\Lambda$, $u=\{\phi,\theta\}$, and $p=\{\lambda\}$, the results immediately follow from Lemma 3.1 and Proposition 2.2. ∎

Using the operators introduced in (3.14) and (3.1), the optimality system (3.18) takes the form

$$\left\{\begin{array}{lllllll} A_1\phi & & & + & B_1^*\lambda & = & A_1\widehat{\phi} \quad \text{in } \Phi^*,\\ & & A_2\theta & + & B_2^*\lambda & = & 0 \quad \text{in } \Theta^*,\\ B_1\phi & + & B_2\theta & & & = & g \quad \text{in } \Lambda^*.\end{array}\right. \tag{3.19}$$

3.3. Galerkin approximation of the optimality system. We choose (conforming) finite dimensional subspaces $\Phi^h\subset\Phi$, $\Theta^h\subset\Theta$, and $\Lambda^h\subset\Lambda$ and then restrict (3.18) to the subspaces, i.e., we seek $(\phi^h,\theta^h,\lambda^h)\in\Phi^h\times\Theta^h\times\Lambda^h$ that satisfies

$$\left\{\begin{array}{llll} a_1(\phi^h,\mu^h) & & +b_1(\mu^h,\lambda^h) & = a_1(\widehat{\phi},\mu^h) \quad \forall\,\mu^h\in\Phi^h,\\ & a_2(\theta^h,\nu^h) & +b_2(\nu^h,\lambda^h) & = 0 \quad \forall\,\nu^h\in\Theta^h,\\ b_1(\phi^h,\psi^h) & +b_2(\theta^h,\psi^h) & & = \langle g,\psi^h\rangle_{\Lambda^*,\Lambda} \quad \forall\,\psi^h\in\Lambda^h.\end{array}\right. \tag{3.20}$$

This is also the optimality system for the minimization of (3.2) over $\Phi^h\times\Theta^h$ subject to the constraint $b_1(\phi^h,\psi^h)+b_2(\psi^h,\theta^h)=\langle g,\psi^h\rangle_{\Lambda^*,\Lambda}$ for all $\psi^h\in\Lambda^h$.

We next define the subspace $Z^h\subset\Phi^h\times\Theta^h$ by

$$Z^h = \Big\{\{\phi^h,\theta^h\}\in\Phi^h\times\Theta^h \ :\ b_1(\phi^h,\psi^h)+b_2(\theta^h,\psi^h)=0 \quad \forall\,\psi^h\in\Lambda^h\Big\}. \tag{3.21}$$

Note that, in general, $Z^h \not\subset Z$ even though $\Phi^h \subset \Phi$, $\Theta^h \subset \Theta$, and $\Lambda^h \subset \Lambda$. Thus, we make the following additional assumptions about $b_1(\cdot,\cdot)$ and Φ^h:

$$(3.22) \qquad \begin{cases} \displaystyle\sup_{\psi^h \in \Lambda^h, \psi^h \neq 0} \frac{b_1(\phi^h, \psi^h)}{\|\psi^h\|_\Lambda} \geq k_1^h \|\phi^h\|_\Phi & \forall\, \phi^h \in \Phi^h, \\ \displaystyle\sup_{\phi^h \in \Phi^h, \phi^h \neq 0} \frac{b_1(\phi^h, \psi^h)}{\|\phi^h\|_V} > 0 & \forall\, \psi^h \in \Lambda^h, \end{cases}$$

where k_1^h is a positive constant whose value is independent of h. Analogous to Lemma 3.1, we have the following result.

LEMMA 3.4. *Let the assumptions* (3.3), (3.5), *and* (3.22) *hold. Then, the spaces* $V^h = \Phi^h \times \Theta^h$ *and* $S^h = \Phi^h$ *and the bilinear forms* $a(\cdot,\cdot)$ *and* $b(\cdot,\cdot)$ *defined in* (3.7) *satisfy the assumptions* (2.10) *and* (2.11).

Proof. The proof proceeds exactly as that for Lemma 3.1; the constants in (2.10) are given by $K_a^h = \frac{1}{2} K_2 \min\{1, \frac{(k_1^h)^2}{c_2^2}\}$ and $K_b^h = k_1^h$. □

We then easily obtain the following results.

THEOREM 3.5. *Let the assumptions* (3.3), (3.5), *and* (3.22) *hold. Then, the discrete optimality system* (3.20) *has a unique solution* $(\phi^h, \theta^h, \lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h$, *and moreover*

$$\|\phi^h\|_\Phi + \|\theta^h\|_\Theta + \|\lambda^h\|_\Lambda \leq C\big(\|g\|_{\Lambda^*} + \|\widehat{\phi}\|_{\widehat{\Phi}}\big).$$

Furthermore, let $(\phi, \theta, \lambda) \in \Phi \times \Theta \times \Lambda$ *denote the unique solution of the optimality system* (3.18) *or, equivalently, of the optimal control problem* (3.6). *Then,*

$$(3.23) \qquad \begin{aligned} &\|\phi - \phi^h\|_\Phi + \|\theta - \theta^h\|_\Theta + \|\lambda - \lambda^h\|_\Lambda \\ &\qquad \leq C\Big(\inf_{\mu^h \in \Phi^h} \|\phi - \mu^h\|_\Phi + \inf_{\xi^h \in \Theta^h} \|\theta - \xi^h\|_\Theta + \inf_{\psi^h \in \Lambda^h} \|\lambda - \psi^h\|_\Lambda\Big). \end{aligned}$$

Proof. The results immediately follow from Proposition 2.4 and Lemma 3.4. □

The discrete optimality system (3.20) is equivalent to the linear system

$$(3.24) \qquad \begin{pmatrix} \mathbb{A}_1 & 0 & \mathbb{B}_1^T \\ 0 & \mathbb{A}_2 & \mathbb{B}_2^T \\ \mathbb{B}_1 & \mathbb{B}_2 & 0 \end{pmatrix} \begin{pmatrix} \vec{\boldsymbol{\phi}} \\ \vec{\boldsymbol{\theta}} \\ \vec{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{0}} \\ \vec{\mathbf{g}} \end{pmatrix},$$

where $\vec{\mathbf{f}}$ and $\vec{\mathbf{g}}$ are defined using $a_1(\widehat{\phi}, \mu^h)$ and $\langle g, \psi^h \rangle_{\Lambda^*,\Lambda}$, respectively, and $\mathbb{A}_k$ and $\mathbb{B}_k$ are defined in the standard manner from the bilinear forms $a_k(\cdot,\cdot)$ and $b_k(\cdot,\cdot)$, $k = 1, 2$, respectively.

Remark 3.6. There are two sets of inf-sup conditions associated with the problems (3.18) and (3.20). First, we have the "inner" conditions (3.5) and (3.22) that involve only the state variable and that guarantee the unique solvability of the state equation and the discrete state equation, respectively, i.e., of the third equations in (3.18) and (3.20). Second, we have the "outer" conditions (2.6) and (2.11) involving the bilinear form $b(\cdot,\cdot)$ defined in (3.7) and that involve both the state and control variables. These latter conditions help guarantee the unique solvability of the optimality system (3.18) and the discrete optimality system (3.20), respectively. Note that the outer conditions and the related saddle-point nature of the optimality systems occur regardless of the nature of the inner problem, i.e., the state equations. For example, even if the state equations involve a strongly coercive bilinear form $b_1(\cdot,\cdot)$ so that the last two inequalities in (3.5) can be replaced by $b_1(\phi,\phi) \geq k_1\|\phi\|_\Phi^2$ for all $\phi \in \Phi$, we would still have the inf-sup condition in the form of the last equation in (2.6).

Remark 3.7. As mentioned in Remark 3.6, the assumptions in (3.22) guarantee the unique solvability of the discrete state equation (the third equation in (3.20)) for the discrete state variable $\phi^h \in \Phi^h$. Thus, if the constraint equation (3.4) is a partial differential equation problem, then the assumptions in (3.22) are the general assumptions on the associated bilinear form and the approximating space that are made to guarantee the stability and convergence of Galerkin finite element discretizations; see, e.g., [1]. Furthermore, because of the nature of the assumptions (3.3) and (3.5), the inf-sup condition on the bilinear form $b(\cdot,\cdot)$ is satisfied merely by assuming that (3.22) holds. Thus, by merely guaranteeing that the discrete constraint equations within the discretized optimal control problem are uniquely solvable for any given discrete control, i.e., assuming that the "inner" inf-sup conditions hold, we have that the "outer" inf-sup condition on the bilinear form $b(\cdot,\cdot)$ holds. The latter, of course, is crucial to the stability and convergence of finite element approximations to any saddle-point problem, including the optimality systems we consider here.

Remark 3.8. The discrete optimality system (3.20) or its matrix equivalent (3.24) has the typical saddle-point structure, and thus, the stability and convergence of the approximations they define depend on the bilinear form $b(\cdot,\cdot) = b_1(\cdot,\cdot) + b_2(\cdot,\cdot)$ satisfying the discrete inf-sup condition (2.11) with respect to $V^h = \Phi^h \times \Theta^h$ and $S^h = \Lambda^h$. In the current context, this assumption is satisfied (see Remark 3.7) merely by assuming that (3.22) holds for the bilinear form $b_1(\cdot,\cdot)$ and for the spaces Φ^h and Λ_h. Thus, as discussed in Remark 3.7, the stability and convergence of solutions of (3.20) or (3.24) depends solely on the ability to stably solve, given any discrete control variable, the discrete state equation for a discrete state variable. On the other hand, if (3.22) does not hold, then there exists a $\phi_0^h \in \Phi^h$ such that $\phi_0^h \neq 0$ and $b_1(\phi_0^h,\psi^h) = 0$ for all $\psi^h \in \Lambda^h$. Then, $b(\{\phi_0^h,0\},\{\psi^h\}) = b_1(\phi_0^h,\psi^h) = 0$ for all $\psi^h \in \Lambda^h$ so that

$$\sup_{\psi^h\in\Lambda^h,\,\psi^h\neq 0} \frac{b(\{\phi_0^h,0\},\{\psi^h\})}{\|\psi^h\|_\Lambda} = 0.$$

It can be shown that this implies that the discrete inf-sup condition (2.11) does not hold so that (3.20) or its matrix equivalent (3.24) may not be solvable, i.e., the coefficient matrix in (3.24) may not be invertible. In fact, the assumptions (3.22) imply that $\mathbb{B}_1$ is uniformly invertible. This and the facts (which follow from (3.3)) that the symmetric matrices $\mathbb{A}_1$ and $\mathbb{A}_2$ are positive semidefinite and positive definite, respectively, are enough to guarantee that the coefficient matrix in (3.24) is invertible. On the other hand, if (3.22) does not hold so that the matrix $\mathbb{B}_1$ has a nontrivial null space, then under the other assumptions that have been made, one cannot guarantee the invertibility of the coefficient matrix in (3.24).

Remark 3.9. Solving the discrete optimality system (3.20) or, equivalently, the linear system (3.24) is often a formidable task. If the constraint equations (3.4) are a system of partial differential equations, then the last (block) row of (3.24) represents a Galerkin finite element discretization of that system. The discrete adjoint equations, i.e., the first row in (3.24), are also a discretization of a system of partial differential equations. Moreover, the dimension of the discrete adjoint vector $\vec{\boldsymbol{\lambda}}$ is essentially the same as that of discrete state vector $\vec{\boldsymbol{\phi}}$. Thus, (3.24) is at least twice the size (we have yet to account for the discrete control variables in $\vec{\boldsymbol{\theta}}$) of the discrete system corresponding to the discretization of the partial differential equation constraints. Thus, if these equations are difficult to approximate, the discrete optimality system will be even more difficult to deal with. For this reason, there have been many approaches suggested for uncoupling the three components of discrete optimality systems such as (3.20) or, equivalently, (3.24). See, e.g., [18] for a discussion of several of these approaches. We note that these approaches rely on the invertibility of the matrices $\mathbb{B}_1$ and $\mathbb{A}_2$, properties that follow from (3.22) and (3.3), respectively.

4. Least-squares finite element methods for the optimality system.

Even if the state equation (3.4) (or (3.16)) involves a symmetric, positive definite operator B_1, i.e., even if the bilinear form $b_1(\cdot,\cdot)$ is symmetric and strongly coercive, the discrete optimality system (3.20) (or (3.24)) obtained through a Galerkin discretization is indefinite. For example, if $B_1 = -\Delta$ with zero boundary conditions, then $\mathbb{B}_1$ is a symmetric, positive definite matrix, but the coefficient matrix in (3.24) is indefinite. In order to obtain a discrete optimality system that is symmetric and positive definite, we will apply a least-squares finite element discretization. In fact, these desirable properties for the discrete system will remain in place even if the state system bilinear form $b_1(\cdot,\cdot)$ is only weakly coercive, i.e., even if the operator B_1 is merely invertible and not necessarily positive definite.

Given a system of partial differential equations, there are many ways to define least-squares finite element methods for determining approximate solutions. Practicality issues can be used to select the “best” methods from among the many choices available. See, e.g., [6] for a discussion of what factors enter into the choice of a particular least-squares finite element method for a given problem. Here, we will consider the most straightforward means for defining a least-squares finite element method. When, in section 5, we consider a specific example, we will return to a discussion of practicality issues in the choice of a least-squares finite element formulation.

4.1. A least-squares finite element method for a generalization of the optimality system.

We start with the generalized form of the optimality system (3.19) written in operator form, i.e.,

$$\left\{\begin{array}{lllllll} A_1\phi & & & + & B_1^*\lambda & = f & \text{in } \Phi^*, \\ & & A_2\theta & + & B_2^*\lambda & = s & \text{in } \Theta^*, \\ B_1\phi & + & B_2\theta & & & = g & \text{in } \Lambda^*, \end{array}\right. \tag{4.1}$$

where $(f,s,g) \in \Phi^* \times \Theta^* \times \Lambda^*$ is a general data triple and $(\phi,\theta,\lambda) \in \Phi \times \Theta \times \Lambda$ is the corresponding solution triple. In the same way that Theorem 3.3 was proved, we have the following result.

PROPOSITION 4.1. *Let the assumptions* (3.3) *and* (3.5) *hold. Then, for any* $(f,s,g) \in \Phi^* \times \Theta^* \times \Lambda^*$*, the generalized optimality system* (4.1) *has a unique solution* $(\phi,\theta,\lambda) \in \Phi \times \Theta \times \Lambda$. *Moreover,*

$$\|\phi\|_\Phi + \|\theta\|_\Theta + \|\lambda\|_\Lambda \le C\big(\|f\|_{\Phi^*} + \|s\|_{\Theta^*} + \|g\|_{\Lambda^*}\big). \tag{4.2}$$

A least-squares functional can be defined by summing the squares of the norms of the residuals of the three equations in (4.1) to obtain

$$\mathcal{K}(\phi,\theta,\lambda;f,s,g) = \|A_1\phi + B_1^*\lambda - f\|^2_{\Phi^*} + \|A_2\theta + B_2^*\lambda - s\|^2_{\Theta^*} + \|B_1\phi + B_2\theta - g\|^2_{\Lambda^*}. \tag{4.3}$$

Clearly, the unique solution of (4.1) is also the solution of the problem

$$\min_{(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda} \mathcal{K}(\phi,\theta,\lambda;f,s,g). \tag{4.4}$$

The first-order necessary conditions corresponding to (4.4) are easily found to be

$$B\big((\phi,\theta,\lambda),(\mu,\nu,\psi)\big) = F\big((\mu,\nu,\psi);(f,s,g)\big) \quad \forall\,(\mu,\nu,\psi)\in\Phi\times\Theta\times\Lambda, \tag{4.5}$$

where

$$\begin{aligned} B\big((\phi,\theta,\lambda),(\mu,\nu,\psi)\big) &= (A_1\mu + B_1^*\psi, A_1\phi + B_1^*\lambda)_{\Phi^*} \\ &+(A_2\nu + B_2^*\psi, A_2\theta + B_2^*\lambda)_{\Theta^*} + (B_1\mu + B_2\nu, B_1\phi + B_2\theta)_{\Lambda^*} \\ &\forall\,(\phi,\theta,\lambda),\ (\mu,\nu,\psi)\in\Phi\times\Theta\times\Lambda \end{aligned} \tag{4.6}$$

and

$$\begin{aligned} F\big((\mu,\nu,\psi);(f,s,g)\big) &= (A_1\mu + B_1^*\psi, f)_{\Phi^*} + (A_2\nu + B_2^*\psi, s)_{\Theta^*} \\ &+(B_1\mu + B_2\nu, g)_{\Lambda^*} \quad \forall\,(\mu,\nu,\psi)\in\Phi\times\Theta\times\Lambda. \end{aligned} \tag{4.7}$$

LEMMA 4.2. *Let the assumptions* (3.3) *and* (3.5) *hold. Then, the bilinear form* $B(\cdot,\cdot)$ *is symmetric and continuous on* $(\Phi\times\Theta\times\Lambda)\times(\Phi\times\Theta\times\Lambda)$*, and the linear functional* $F(\cdot)$ *is continuous on* $(\Phi\times\Theta\times\Lambda)$*. Moreover, the bilinear form* $B(\cdot,\cdot)$ *is coercive on* $(\Phi\times\Theta\times\Lambda)$*, i.e.,*

$$B\big((\phi,\theta,\lambda),(\phi,\theta,\lambda)\big) \ge C(\|\phi\|_\Phi^2 + \|\theta\|_\Theta^2 + \|\lambda\|_\Lambda^2) \quad \forall\,(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda. \tag{4.8}$$

Proof. The symmetry and continuity of the form $B(\cdot,\cdot)$ and the continuity of the form $F(\cdot)$ are clear. From (4.6), we have that

$$B\big((\phi,\theta,\lambda),(\phi,\theta,\lambda)\big) = \|A_1\phi + B_1^*\lambda\|_{\Phi^*}^2 + \|A_2\theta + B_2^*\lambda\|_{\Theta^*}^2 + \|B_1\phi + B_2\theta\|_{\Lambda^*}^2. \tag{4.9}$$

Clearly, for any $(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda$, there exists $(f,s,g)\in\Phi^*\times\Theta^*\times\Lambda^*$ such that (ϕ,θ,λ) is a solution of (4.1). This observation and Proposition 4.1 then yield that

$$\begin{aligned} \|\phi\|_\Phi^2 + \|\theta\|_\Theta^2 + \|\lambda\|_\Lambda^2 &\le C\big(\|f\|_{\Phi^*}^2 + \|s\|_{\Theta^*}^2 + \|g\|_{\Lambda^*}^2\big) \\ &= C\big(\|A_1\phi + B_1^*\lambda\|_{\Phi^*}^2 + \|A_2\theta + B_2^*\lambda\|_{\Theta^*}^2 + \|B_1\phi + B_2\theta\|_{\Lambda^*}^2\big) \\ &\forall\,(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda. \end{aligned} \tag{4.10}$$

Combining (4.9) and (4.10) then easily yields (4.8). □

Remark 4.3. Since

$$\begin{aligned} \mathcal{K}(\phi,\theta,\lambda;0,0,0) &= \|A_1\phi + B_1^*\lambda\|_{\Phi^*}^2 + \|A_2\theta + B_2^*\lambda\|_{\Theta^*}^2 + \|B_1\phi + B_2\theta\|_{\Lambda^*}^2 \\ &= B\big((\phi,\theta,\lambda),(\phi,\theta,\lambda)\big), \end{aligned}$$

the coercivity and continuity of the bilinear form $B(\cdot,\cdot)$ are equivalent to stating that the functional $\mathcal{K}(\phi,\theta,\lambda;0,0,0)$ is norm-equivalent, i.e., that there exist constants $\gamma_1 > 0$ and $\gamma_2 > 0$ such that

$$\gamma_1(\|\phi\|_\Phi^2 + \|\theta\|_\Theta^2 + \|\lambda\|_\Lambda^2) \le \mathcal{K}(\phi,\theta,\lambda;0,0,0) \le \gamma_2(\|\phi\|_\Phi^2 + \|\theta\|_\Theta^2 + \|\lambda\|_\Lambda^2) \tag{4.11}$$

for all $(\phi,\theta,\lambda) \in \Phi \times \Theta \times \Lambda$.

PROPOSITION 4.4. *Let the assumptions* (3.3) *and* (3.5) *hold. Then, for any* $(f,s,g) \in \Phi^* \times \Theta^* \times \Lambda^*$*, the problem* (4.5) *has a unique solution* $(\phi,\theta,\lambda) \in \Phi \times \Theta \times \Lambda$*. Moreover, this solution coincides with the solution of the problems* (4.1) *and* (4.4) *and satisfies the estimate* (4.2).

Proof. The results follow from Lemma 4.2 and the Lax–Milgram lemma. □

We define a finite element discretization of (4.1) or, equivalently, of (4.5) by choosing conforming finite element subspaces $\Phi^h \subset \Phi$, $\Theta^h \subset \Theta$, and $\Lambda^h \subset \Lambda$ and then requiring that $(\phi^h,\theta^h,\lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h$ satisfy

$$B\big((\phi^h,\theta^h,\lambda^h),(\mu^h,\nu^h,\psi^h)\big) = F\big((\mu^h,\nu^h,\psi^h);(f,s,g)\big) \qquad \forall\,(\mu^h,\nu^h,\psi^h) \in \Phi^h \times \Theta^h \times \Lambda^h. \tag{4.12}$$

Note that $(\phi^h,\theta^h,\lambda^h)$ can also be characterized as the solution of the problem

$$\min_{(\phi^h,\theta^h,\lambda^h)\in\Phi^h\times\Theta^h\times\Lambda^h} \mathcal{K}(\phi^h,\theta^h,\lambda^h;f,s,g).$$

PROPOSITION 4.5. *Let the assumptions* (3.3) *and* (3.5) *hold. Then, for any* $(f,h,g) \in \Phi^* \times \Theta^* \times \Lambda^*$*, the problem* (4.12) *has a unique solution* $(\phi^h,\theta^h,\lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h$*. Moreover, we have the optimal error estimate*

$$\begin{aligned}&\|\phi-\phi^h\|_\Phi + \|\theta-\theta^h\|_\Theta + \|\lambda-\lambda^h\|_\Lambda \\ &\qquad \le C\Big(\inf_{\widetilde{\phi}^h\in\Phi^h}\|\phi-\widetilde{\phi}^h\|_\Phi + \inf_{\widetilde{\theta}^h\in\Theta^h}\|\theta-\widetilde{\theta}^h\|_\Theta + \inf_{\widetilde{\lambda}^h\in\Lambda^h}\|\lambda-\widetilde{\lambda}^h\|_\Lambda\Big),\end{aligned} \tag{4.13}$$

where $(\phi,\theta,\lambda) \in \Phi \times \Theta \times \Lambda$ *is the unique solution of the problem* (4.5) *or, equivalently, of the problems* (4.1) *or* (4.4).

Proof. The results follow from Lemma 4.2 and from standard finite element analyses. □

4.2. A least-squares finite element method for the optimality system. The results of section 4.1 easily specialize to the optimality system (3.19). Indeed, letting $f = A_1\widehat{\phi} \in \widehat{\Phi}^* \subset \Phi^*$ and $s = 0$, we have that (4.1) reduces to (3.19). We now have the least-squares functional

$$\mathcal{K}(\phi,\theta,\lambda;\widehat{\phi},g) = \|A_1\phi + B_1^*\lambda - A_1\widehat{\phi}\|^2_{\Phi^*} + \|A_2\theta + B_2^*\lambda\|^2_{\Theta^*} + \|B_1\phi + B_2\theta - g\|^2_{\Lambda^*}, \tag{4.14}$$

the minimization problem

$$\min_{(\phi,\theta,\lambda)\in\Phi\times\Theta\times\Lambda} \mathcal{K}(\phi,\theta,\lambda;\widehat{\phi},g), \tag{4.15}$$

and the first-order necessary conditions

$$B\big((\phi,\theta,\lambda),(\mu,\nu,\psi)\big) = F\big((\mu,\nu,\psi);(A_1\widehat{\phi},0,g)\big) \quad \forall\,(\mu,\nu,\psi) \in \Phi \times \Theta \times \Lambda, \tag{4.16}$$

where $B(\cdot,\cdot)$ and $F(\cdot)$ are defined as in (4.6) and (4.7), respectively.

We define a finite element discretization of (4.16) by again choosing conforming finite element subspaces $\Phi^h \subset \Phi$, $\Theta^h \subset \Theta$, and $\Lambda^h \subset \Lambda$ and then requiring that $(\phi^h,\theta^h,\lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h$ satisfy

$$B\big((\phi^h,\theta^h,\lambda^h),(\mu^h,\nu^h,\psi^h)\big) = F\big((\mu^h,\nu^h,\psi^h);(A_1\widehat{\phi},0,g)\big) \qquad \forall\,(\mu^h,\nu^h,\psi^h) \in \Phi^h \times \Theta^h \times \Lambda^h. \tag{4.17}$$

Then, Proposition 4.5 takes the following form.

THEOREM 4.6. *Let the assumptions* (3.3) *and* (3.5) *hold. Then, for any* $(\widehat{\phi}, g) \in \widehat{\Phi}^* \times \Lambda^*$, *the problem* (4.17) *has a unique solution* $(\phi^h, \theta^h, \lambda^h) \in \Phi^h \times \Theta^h \times \Lambda^h$. *Moreover, we have the optimal error estimate: There exists a constant* $C > 0$ *whose value is independent of* h, *such that*

$$\begin{aligned} &\|\phi - \phi^h\|_\Phi + \|\theta - \theta^h\|_\Theta + \|\lambda - \lambda^h\|_\Lambda \\ &\qquad \le C\Big(\inf_{\widetilde{\phi}^h \in \Phi^h} \|\phi - \widetilde{\phi}^h\|_\Phi + \inf_{\widetilde{\theta}^h \in \Theta^h} \|\theta - \widetilde{\theta}^h\|_\Theta + \inf_{\widetilde{\lambda}^h \in \Lambda^h} \|\lambda - \widetilde{\lambda}^h\|_\Lambda\Big), \end{aligned} \tag{4.18}$$

where $(\phi, \theta, \lambda) \in \Phi \times \Theta \times \Lambda$ *is the unique solution of the problem* (4.16) *or, equivalently, of the problems* (3.19) *or* (3.18). *Note also that* $(\phi, \theta) \in \Phi \times \Theta$ *is the unique solution of the problem* (3.6).

Remark 4.7. The discrete problem (4.17) is equivalent to the linear algebraic system

$$\begin{pmatrix} \mathbb{K}_1 & \mathbb{C}_1^T & \mathbb{C}_2^T \\ \mathbb{C}_1 & \mathbb{K}_2 & \mathbb{C}_3^T \\ \mathbb{C}_2 & \mathbb{C}_3 & \mathbb{K}_3 \end{pmatrix} \begin{pmatrix} \vec{\boldsymbol{\phi}} \\ \vec{\boldsymbol{\theta}} \\ \vec{\boldsymbol{\lambda}} \end{pmatrix} = \begin{pmatrix} \vec{\mathbf{f}} \\ \vec{\mathbf{h}} \\ \vec{\mathbf{g}} \end{pmatrix}. \tag{4.19}$$

Indeed, if one chooses bases $\{\mu_j^h(\mathbf{x})\}_{j=1}^J$, $\{\nu_k^h(\mathbf{x})\}_{k=1}^K$, and $\{\psi_\ell^h(\mathbf{x})\}_{\ell=1}^L$ for Φ_h, Θ_h, and Λ_h, respectively, we then have $\phi^h = \sum_{j=1}^J \phi_j \mu_j^h$, $\theta^h = \sum_{k=1}^K \theta_k \mu_k^h$, and $\lambda^h = \sum_{\ell=1}^L \lambda_\ell \psi_\ell^h$ for some sets of coefficients $\{\phi_j\}_{j=1}^J$, $\{\theta_k\}_{k=1}^K$, and $\{\lambda_\ell\}_{\ell=1}^L$ that are determined by solving (4.19). In (4.19), we have that $\vec{\boldsymbol{\phi}} = (\phi_1, \dots, \phi_J)^T$, $\vec{\boldsymbol{\theta}} = (\theta_1, \dots, \theta_K)^T$, $\vec{\boldsymbol{\lambda}} = (\lambda_1, \dots, \lambda_L)^T$,

$$\begin{aligned}
&\big(\mathbb{K}_1\big)_{ij} = (A_1\mu_i, A_1\mu_j)_{\Phi^*} + (B_1\mu_i, B_1\mu_j)_{\Lambda^*} && \text{for } i,j = 1,\dots,J, \\
&\big(\mathbb{K}_2\big)_{ik} = (A_2\nu_i, A_1\nu_k)_{\Theta^*} + (B_2\nu_i, B_2\nu_k)_{\Lambda^*} && \text{for } i,k = 1,\dots,K, \\
&\big(\mathbb{K}_3\big)_{i\ell} = (B_1^*\psi_i, B_1^*\psi_\ell)_{\Phi^*} + (B_2\psi_i, B_2\psi_\ell)_{\Theta^*} && \text{for } i,\ell = 1,\dots,L, \\
&\big(\mathbb{C}_1\big)_{ij} = (B_2\nu_i, B_1\mu_j)_{\Lambda^*} && \text{for } i = 1,\dots,K,\ j = 1,\dots,J, \\
&\big(\mathbb{C}_2\big)_{ij} = (B_1^*\psi_i, A_1\nu_j)_{\Phi^*} && \text{for } i = 1,\dots,L,\ j = 1,\dots,J, \\
&\big(\mathbb{C}_3\big)_{ik} = (B_2^*\psi_i, A_2\nu_k)_{\Theta^*} && \text{for } i = 1,\dots,L,\ k = 1,\dots,K, \\
&\big(\vec{\mathbf{f}}\big)_i = (A_1\mu_i, A_1\widehat{\phi})_{\Phi^*} + (B_1\mu_i, g)_{\Lambda^*} && \text{for } i = 1,\dots,J, \\
&\big(\vec{\mathbf{h}}\big)_i = (B_2\nu_i, g)_{\Lambda^*} && \text{for } i = 1,\dots,K, \\
&\big(\vec{\mathbf{g}}\big)_i = (B_1^*\psi_i, A_1\widehat{\phi})_{\Phi^*} && \text{for } i = 1,\dots,L.
\end{aligned}$$

Remark 4.8. It easily follows from Lemma 4.2 that the coefficient matrix of (4.19) is *symmetric and positive definite.* This should be compared to the linear system (3.24) that results from a Galerkin finite element discretization of the optimality system (3.18) for which the coefficient matrix is symmetric and *indefinite.*

Remark 4.9. The stability of the discrete problem (4.17), the convergence and optimal accuracy of the approximate solution $(\phi^h, \theta^h, \lambda^h)$, and the symmetry and positive definiteness of the discrete system (4.19) obtained by the least-squares finite element method follow from the assumptions (3.3) and (3.5) that guarantee the well posedness of the infinite-dimensional optimization problem (3.6) and its corresponding optimality system (3.18). It is important to note that all of these desirable properties of the least-squares finite element method do not require that the bilinear form $b_1(\cdot,\cdot)$ and that the finite element spaces Φ^h and Λ^h satisfy the inner (see Remark 3.6) inf-sup conditions (3.22) that are necessary for the well posedness of the Galerkin finite element discretization (3.20) of the optimality system (3.18). In fact, this is why least-squares finite element methods are often an attractive alternative to Galerkin discretizations; see, e.g., [6].

Remark 4.10. The observations made in Remark 3.9 about the possible need to uncouple the equations in (3.24) hold as well for the linear system (4.19). Uncoupling approaches for (3.24) rely on the invertibility of the matrices $\mathbb{B}_1$ and $\mathbb{A}_2$; the first of these is, in general, nonsymmetric and indefinite, even when the necessary discrete inf-sup conditions in (3.22) are satisfied. For (4.19), uncoupling strategies would rely on the invertibility of the matrices $\mathbb{K}_1$, $\mathbb{K}_2$, and $\mathbb{K}_3$; all three of these matrices are symmetric and positive definite even when (3.22) is not satisfied. An example of a simple uncoupling strategy is to apply a block-Gauss–Seidel method to (4.19), which would proceed as follows.

> Start with initial guesses $\vec{\boldsymbol{\phi}}^{(0)}$ and $\vec{\boldsymbol{\theta}}^{(0)}$ for the discretized state and control; then, for $k = 1, 2, \ldots$, successively solve the linear systems
>
> $$\begin{aligned} \mathbb{K}_3\vec{\boldsymbol{\lambda}}^{(k+1)} &= \vec{\mathbf{g}} - \mathbb{C}_2\vec{\boldsymbol{\phi}}^{(k)} - \mathbb{C}_3\vec{\boldsymbol{\theta}}^{(k)}, \\ \mathbb{K}_1\vec{\boldsymbol{\phi}}^{(k+1)} &= \vec{\mathbf{f}} - \mathbb{C}_1^T\vec{\boldsymbol{\theta}}^{(k)} - \mathbb{C}_2^T\vec{\boldsymbol{\lambda}}^{(k+1)}, \\ \mathbb{K}_2\vec{\boldsymbol{\theta}}^{(k+1)} &= \vec{\mathbf{h}} - \mathbb{C}_1\vec{\boldsymbol{\phi}}^{(k+1)} - \mathbb{C}_3^T\vec{\boldsymbol{\lambda}}^{(k+1)} \end{aligned} \tag{4.20}$$
>
> until satisfactory convergence is achieved, e.g., until some norm of the difference between successive iterates is less than some prescribed tolerance.

Since the coefficient matrix in (4.19) is symmetric and positive definite, this iteration will converge. Moreover, all three coefficient matrices $\mathbb{K}_3$, $\mathbb{K}_1$, and $\mathbb{K}_2$ of the linear systems in (4.20) are themselves symmetric and positive definite so that very efficient solution methodologies, including parallel ones, can be applied for their solution. We also note that, in order to obtain faster convergence rates, better uncoupling iterative methods, e.g., over-relaxation schemes or a conjugate gradient method, can be applied instead of the Gauss–Seidel iteration of (4.20).

Remark 4.11. The discrete problem (4.17) (or equivalently, (4.19)) resulting from the least-squares method for the optimality system (3.19) can be viewed as a Galerkin discretization of the system

$$\begin{aligned} &(A_1^*A_1 + B_1^*B_1)\phi + (B_1^*B_2)\theta + (A_1^*B_1^*)\lambda = (A_1^*A_1)\widehat{\phi} + (B_1^*)g && \text{in } \Phi, \\ &(A_2^*A_2 + B_2^*B_2)\theta + (A_2^*B_2^*)\lambda + (B_2^*B_1)\phi = (B_2^*)g && \text{in } \Theta, \\ &(B_1B_1^* + B_2B_2^*)\lambda + (B_1A_1)\phi + (B_2A_2)\theta = (B_1A_1)\widehat{\phi} && \text{in } \Lambda. \end{aligned} \tag{4.21}$$

The first equation of this system is the sum of A_1^* applied to the first equation of the optimality system (3.19) and B_1^* applied to the third equation of that system. The other equations of (4.21) are related to the equations of (3.19) in a similar manner.

The system (4.21) shows that the discrete system (4.19) essentially involves the discretization of "squares" of operators, e.g., $A_1^*A_1$, $B_1^*B_1$, etc. This observation has a profound effect in how one chooses the form of the constraint equation in (3.6), i.e., the form of (3.16). We will return to this point in the next section when we consider a concrete example.

5. Example: Optimization problems for the Stokes system. Let Ω denote an open, bounded domain in $\mathcal{R}^s$, $s = 2$ or 3, with boundary Γ. Let $\mathbf{u}$ and p denote the velocity and pressure fields, respectively, and let $\boldsymbol{\theta}$ denote a distributed control. Then, consider the Stokes system

$$(5.1)\qquad \left\{\begin{array}{lcl} -\Delta\mathbf{u} + \nabla p + \boldsymbol{\theta} & = & \mathbf{g} \\ \nabla\cdot\mathbf{u} & = & 0 \end{array}\right. \quad \text{in } \Omega, \qquad \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma, \qquad \int_\Omega p\, d\Omega = 0$$

and the functionals

$$(5.2)\qquad \text{Case I:}\qquad \mathcal{J}_1(\mathbf{u},\boldsymbol{\theta}) = \frac{1}{2}\int_\Omega |\nabla\times\mathbf{u}|^2\, d\Omega + \frac{\delta}{2}\int_\Omega |\boldsymbol{\theta}|^2\, d\Omega,$$

$$(5.3)\qquad \text{Case II:}\qquad \mathcal{J}_2(\mathbf{u},\boldsymbol{\theta};\widehat{\mathbf{u}}) = \frac{1}{2}\int_\Omega |\mathbf{u}-\widehat{\mathbf{u}}|^2\, d\Omega + \frac{\delta}{2}\int_\Omega |\boldsymbol{\theta}|^2\, d\Omega,$$

where $\mathbf{g}$ and $\widehat{\mathbf{u}}$ are given functions. We study the two problems of finding $(\mathbf{u}, p, \boldsymbol{\theta})$ that minimizes the functional in either (5.2) or (5.3), subject to the Stokes system (5.1) being satisfied. In the first case, i.e., for the functional (5.2), the problem we study is to find a distributed control function $\boldsymbol{\theta}$ that minimizes, in the $\mathbf{L}^2(\Omega)$ sense, the vorticity over the flow domain Ω. In the second case, i.e., for the functional (5.3), the problem we study is to find a distributed control function $\boldsymbol{\theta}$ such that flow velocity $\mathbf{u}$ matches as well as possible, in the $\mathbf{L}^2(\Omega)$ sense, a given velocity field $\widehat{\mathbf{u}}$.

In Remark 4.11, it was noted that least-squares finite element methods for optimization problems result in the "squaring" of the constraint operator, in this case, of the Stokes system (5.1). This results in biharmonic-type terms appearing in the system corresponding to (4.21) or, equivalently, in (4.9). A conforming finite element discretization would then require the use of continuously differentiable approximation spaces. In order to overcome this impracticality, it has become a standard procedure in least-squares finite element methods to write the state system in an equivalent first-order formulation; see, e.g., [6] for a detailed discussion of this issue.

There are many ways to rewrite the Stokes system (5.1) as a first-order system of partial differential equations. Here, we choose the *velocity-vorticity-pressure formulation* that is the most commonly used system for this purpose. Let $\boldsymbol{\omega} = \nabla\times\mathbf{u}$ denote the vorticity. Then, using the well-known vector identity $-\triangle\mathbf{u} = \nabla\times\nabla\times\mathbf{u} - \nabla(\nabla\cdot\mathbf{u}) = \nabla\times\boldsymbol{\omega} - \nabla(\nabla\cdot\mathbf{u})$, the Stokes system (5.1) can be expressed as

$$(5.4)\qquad \left\{\begin{array}{lcl} \nabla\times\boldsymbol{\omega} + \nabla p + \boldsymbol{\theta} & = & \mathbf{g} \\ \nabla\cdot\mathbf{u} & = & 0 \\ \nabla\times\mathbf{u} - \boldsymbol{\omega} & = & \mathbf{0} \end{array}\right. \quad \text{in } \Omega, \qquad \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma, \qquad \int_\Omega p\, d\Omega = 0.$$

Note that the functional (5.2) can now be written as

$$(5.5)\qquad \text{Case I:}\qquad \mathcal{J}_1(\boldsymbol{\omega},\boldsymbol{\theta}) = \frac{1}{2}\int_\Omega |\boldsymbol{\omega}|^2\, d\Omega + \frac{\delta}{2}\int_\Omega |\boldsymbol{\theta}|^2\, d\Omega.$$

Thus, the optimization problems we study are to find $(\mathbf{u}, \boldsymbol{\omega}, p, \boldsymbol{\theta})$ that minimizes the functional in either (5.3) or (5.5), subject to the Stokes system in the form (5.4) being satisfied.

5.1. Precise statement of optimization problems. We recall the space $L^2(\Omega)$ of all square integrable functions with norm $\|\cdot\|_0$ and inner product $(\cdot,\cdot)$, the space $L^2_0(\Omega) \equiv \{q \in L^2(\Omega) : \int_\Omega p d\Omega = 0\}$, the space $H^1(\Omega) \equiv \{v \in L^2(\Omega) : \nabla v \in [L^2(\Omega)]^s\}$, and the space $H^1_0(\Omega) \equiv \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma\}$. A norm for functions $v \in H^1(\Omega)$ is given by $\|v\|_1 \equiv (\|\nabla v\|^2 + \|v\|_0^2)^{1/2}$. The dual space of $H^1_0(\Omega)$ is denoted by $H^{-1}(\Omega)$. The corresponding spaces of vector-valued functions are denoted in bold face, e.g., $\mathbf{H}^1(\Omega) = [H^1(\Omega)]^s$ is the space of vector-valued functions each of whose components belongs to $H^1(\Omega)$. We note the following equivalence of norms [16]:

$$\widetilde{C}_1\|\mathbf{v}\|_1^2 \le \|\nabla\times\mathbf{v}\|_0^2 + \|\nabla\cdot\mathbf{v}\|_0^2 \le \widetilde{C}_2\|\mathbf{v}\|_1^2 \qquad \forall\,\mathbf{v}\in\mathbf{H}^1_0(\Omega) \tag{5.6}$$

for some constants $\widetilde{C}_1 > 0$ and $\widetilde{C}_2 > 0$.

Let $\Phi = \Lambda = \mathbf{H}^1_0(\Omega) \times \mathbf{L}^2(\Omega) \times L^2_0(\Omega)$ and $\Theta = \mathbf{L}^2(\Omega)$ so that $\Phi^* = \Lambda^* = \mathbf{H}^{-1}(\Omega) \times \mathbf{L}^2(\Omega) \times L^2_0(\Omega)$ and $\Theta^* = \mathbf{L}^2(\Omega)$. Let $\widehat{\Phi} = \widetilde{\Phi} = \mathbf{L}^2(\Omega) \times \mathbf{L}^2(\Omega) \times L^2_0(\Omega)$. Then, $\Phi \subset \widehat{\Phi} = \widetilde{\Phi} = \widehat{\Phi}^* \subset \Phi^*$. For $\phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \Phi$, we define the norm

$$\|\phi\|_\Phi = \left(\|\mathbf{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2\right)^{1/2}$$

and likewise for the other product spaces.

We make the associations of

trial functions:	$\phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \Phi$,	$\theta = \{\boldsymbol{\theta}\} \in \Theta$,	$\lambda = \{\mathbf{v}, \boldsymbol{\sigma}, q\} \in \Lambda$,
test functions:	$\mu = \{\widetilde{\mathbf{u}}, \widetilde{\boldsymbol{\omega}}, \widetilde{p}\} \in \Phi$,	$\nu = \{\widetilde{\boldsymbol{\theta}}\} \in \Theta$,	$\psi = \{\widetilde{\mathbf{v}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{r}\} \in \Lambda$,
data:	$g = \{\mathbf{g}, \mathbf{0}, 0\} \in \Lambda^*$,	$\widehat{\phi} = \{\widehat{\mathbf{u}}, \mathbf{0}, 0\} \in \widehat{\Phi}$.	

We next define the bilinear forms

$$a_1(\phi,\mu) = \begin{cases} (\widetilde{\boldsymbol{\omega}}, \boldsymbol{\omega}) & \text{for Case I} \\ (\widetilde{\mathbf{u}}, \mathbf{u}) & \text{for Case II} \end{cases} \qquad \forall\,\phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \widehat{\Phi},\;\; \mu = \{\widetilde{\mathbf{u}}, \widetilde{\boldsymbol{\omega}}, \widetilde{p}\} \in \widehat{\Phi},$$

$$a_2(\theta,\nu) = \delta(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) \qquad \forall\,\theta = \{\boldsymbol{\theta}\} \in \Theta,\;\; \nu = \{\widetilde{\boldsymbol{\theta}}\} \in \Theta,$$

$$b_1(\phi,\psi) = (\boldsymbol{\omega}, \nabla\times\widetilde{\mathbf{v}}) - (p, \nabla\cdot\widetilde{\mathbf{v}}) + (\nabla\times\mathbf{u} - \boldsymbol{\omega}, \widetilde{\boldsymbol{\sigma}}) - (\nabla\cdot\mathbf{u}, \widetilde{r})$$
$$\forall\,\phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \Phi,\;\; \psi = \{\widetilde{\mathbf{v}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{r}\} \in \Lambda,$$

$$b_2(\theta,\psi) = (\boldsymbol{\theta}, \widetilde{\mathbf{v}}) \qquad \forall\,\theta = \{\boldsymbol{\theta}\} \in \Theta,\;\; \psi = \{\widetilde{\mathbf{v}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{r}\} \in \Lambda.$$

For $\mathbf{g} \in \mathbf{H}^{-1}(\Omega)$, we also define the linear functional

$$\langle g, \psi\rangle_{\Lambda^*,\Lambda} = \langle \mathbf{g}, \widetilde{\mathbf{v}}\rangle_{\mathbf{H}^{-1}(\Omega),\mathbf{H}^1_0(\Omega)} \qquad \forall\,\psi = \{\widetilde{\mathbf{v}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{r}\} \in \Lambda.$$

The operators associated with the bilinear forms are then

$$\begin{aligned} &A_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & I & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{for Case I,} \qquad A_1 = \begin{pmatrix} I & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{for Case II,} \\ &A_2 = \delta I, \qquad B_1 = \begin{pmatrix} 0 & \nabla\times & \nabla \\ \nabla\times & -I & 0 \\ -\nabla\cdot & 0 & 0 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} I \\ 0 \\ 0 \end{pmatrix}. \end{aligned} \tag{5.7}$$

It is now easily seen that the functionals $\mathcal{J}_1(\cdot,\cdot)$ and $\mathcal{J}_2(\cdot,\cdot;\cdot)$ defined in (5.5) and (5.3), respectively, can be written in the form (3.2). Likewise, the Stokes system (5.4) can be written in the form (3.4). Thus, the two optimization problems for the Stokes system can both be written in the form (3.6), with $\mathcal{J}(\cdot,\cdot)$ being either $\mathcal{J}_1(\cdot,\cdot)$ or $\mathcal{J}_2(\cdot,\cdot)$ as appropriate. Thus, if the assumptions (3.3) and (3.5) can be verified in the context of the two optimization problems for the Stokes system, then all the results of section 4 will apply to those systems.

Proposition 5.1. *Let the spaces Φ, $\widehat{\Phi}$, Θ, and Λ and the bilinear forms $a_1(\cdot,\cdot)$, $a_2(\cdot,\cdot)$, $b_1(\cdot,\cdot)$, and $b_2(\cdot,\cdot)$ be defined as in this section. Then, the assumptions* (3.3) *and* (3.5) *are satisfied.*

Proof. The four inequalities in (3.3) and the first two inequalities in (3.5) are easily verified with $C_1 = 1$, $C_2 = \delta$, $K_2 = \delta$, $c_1 = 3$, and $c_2 = 1$. The third inequality in (3.5) is verified if, for any $\phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \Phi$, one can find a $\widetilde{\psi} = \{\widetilde{\mathbf{v}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{r}\} \in \Lambda$ such that

$$\begin{aligned} b_1(\phi, \widetilde{\psi}) &= (\boldsymbol{\omega}, \nabla\times\widetilde{\mathbf{v}}) - (p, \nabla\cdot\widetilde{\mathbf{v}}) + (\nabla\times\mathbf{u} - \boldsymbol{\omega}, \widetilde{\boldsymbol{\sigma}}) - (\nabla\cdot\mathbf{u}, \widetilde{r}) \\ &\geq k_1\|\phi\|_\Phi\|\widetilde{\psi}\|_\Lambda = k_1\big(\|\mathbf{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2\big)^{1/2}\big(\|\widetilde{\mathbf{v}}\|_1^2 + \|\widetilde{\boldsymbol{\sigma}}\|_0^2 + \|\widetilde{r}\|_0^2\big)^{1/2} \end{aligned}$$

for some constant $k_1 > 0$. To this end, for any $\phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \Phi$, let $\widetilde{\psi} = \{\widetilde{\mathbf{v}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{r}\} \in \Lambda$ satisfy the system

$$\nabla\times\widetilde{\mathbf{v}} = \boldsymbol{\omega}, \qquad \nabla\cdot\widetilde{\mathbf{v}} = -p, \qquad \widetilde{\boldsymbol{\sigma}} = \nabla\times\mathbf{u}, \qquad \text{and} \qquad \widetilde{r} = -\nabla\cdot\mathbf{u}$$

in Ω. Clearly, from the last two equations, we have that

$$\|\widetilde{\boldsymbol{\sigma}}\|_0^2 + \|\widetilde{r}\|_0^2 = \|\nabla\times\mathbf{u}\|_0^2 + \|\nabla\cdot\mathbf{u}\|_0^2 \leq \widetilde{C}_2\|\mathbf{u}\|_1^2.$$

Also, since $\widetilde{\mathbf{v}} \in \mathbf{H}_0^1(\Omega)$, we have from the first two equations and (5.6) that

$$\widetilde{C}_1\|\widetilde{\mathbf{v}}\|_1^2 \leq (\|\nabla\times\widetilde{\mathbf{v}}\|_0^2 + \|\nabla\cdot\widetilde{\mathbf{v}}\|_0^2) = (\|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2).$$

Combining the last two results yields that

$$\Big(\|\widetilde{\mathbf{v}}\|_1^2 + \|\widetilde{\boldsymbol{\sigma}}\|_0^2 + \|\widetilde{r}\|_0^2\Big) \leq \max\left\{\frac{1}{\widetilde{C}_1}, \widetilde{C}_2\right\}\Big(\|\mathbf{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2\Big).$$

Then, with $\phi = \{\mathbf{u}, \boldsymbol{\omega}, p\} \in \Phi$ and $\widetilde{\psi} = \{\widetilde{\mathbf{v}}, \widetilde{\boldsymbol{\sigma}}, \widetilde{r}\} \in \Lambda$, we have that

$$\begin{aligned} b_1(\phi, \widetilde{\psi}) &= \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2 + \|\nabla\times\mathbf{u}\|_0^2 + \|\nabla\cdot\mathbf{u}\|_0^2 - (\boldsymbol{\omega}, \nabla\times\mathbf{u}) \\ &\geq \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2 + \|\nabla\times\mathbf{u}\|_0^2 + \|\nabla\cdot\mathbf{u}\|_0^2 - \|\boldsymbol{\omega}\|_0\|\nabla\times\mathbf{u}\|_0 \\ &\geq \tfrac{1}{2}\|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2 + \tfrac{1}{2}\|\nabla\times\mathbf{u}\|_0^2 + \|\nabla\cdot\mathbf{u}\|_0^2 \\ &\geq \tfrac{1}{2}\min\{1, \widetilde{C}_1\}\Big(\|\mathbf{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2\Big) \\ &\geq \frac{\min\{1,\widetilde{C}_1\}}{2\sqrt{\max\left\{\frac{1}{\widetilde{C}_1}, \widetilde{C}_2\right\}}}\Big(\|\mathbf{u}\|_1^2 + \|\boldsymbol{\omega}\|_0^2 + \|p\|_0^2\Big)^{1/2}\Big(\|\widetilde{\mathbf{v}}\|_1^2 + \|\widetilde{\boldsymbol{\sigma}}\|_0^2 + \|\widetilde{r}\|_0^2\Big)^{1/2} \\ &= \frac{\min\{1,\widetilde{C}_1\}}{2\sqrt{\max\left\{\frac{1}{\widetilde{C}_1}, \widetilde{C}_2\right\}}}\|\phi\|_\Phi\|\widetilde{\psi}\|_\Lambda. \end{aligned}$$

Thus, with $k_1 = \min\{1, \widetilde{C}_1\}/(2\sqrt{\max\{\frac{1}{\widetilde{C}_1}, \widetilde{C}_2\}})$, the third inequality in (3.5) is verified. Note that k_1 depends only on the comparability constants in (5.6). □

Remark 5.2. We have now verified the assumptions (3.3) and (3.5) for the two optimization problems of finding $(\mathbf{u}, \boldsymbol{\omega}, p, \boldsymbol{\theta})$ that minimize either the functional in (5.3) or (5.5), subject to the Stokes system in the form (5.4) being satisfied. Thus, all the results of sections 3.1 and 3.2 hold. In particular, with the associations already defined between spaces, operators, etc., we could use the Lagrange multiplier rule to characterize the solutions of the optimization problems as solutions of the optimality system (3.19).

Remark 5.3. We could apply, as in section 3.3, a Galerkin finite element method for determining approximate solutions of the optimality system (3.19). Such an approach, unlike least-squares finite element discretizations, does not involve the "squaring" of operators so that there is no need to transform the Stokes system (5.1) into an equivalent first-order form as in (5.4); one would then also use the form (5.2) for the functional $\mathcal{J}_1$ instead of the from (5.5). We then would have $\phi = \{\mathbf{u}, p\}$, $\theta = \{\boldsymbol{\theta}\}$, etc., and use, instead of the operators defined in (5.7), the operators

$$\begin{aligned} &A_1 = \begin{pmatrix} \nabla\times & 0 \\ 0 & 0 \end{pmatrix} \quad \text{for Case I,} \qquad A_1 = \begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix} \quad \text{for Case II,} \\ &A_2 = \delta I, \qquad B_1 = \begin{pmatrix} -\Delta & \nabla \\ -\nabla\cdot & 0 \end{pmatrix}, \qquad B_2 = \begin{pmatrix} I \\ 0 \end{pmatrix}. \end{aligned} \tag{5.8}$$

The assumptions (3.3) and (3.5) can also be verified for the bilinear forms associated with these operators.

Remark 5.4. As noted in section 3.3, a Galerkin discretization of the optimality system (3.19) using either of the forms (5.7) or (5.8) for the operators requires that the assumptions in (3.22) hold. If one uses (5.8), one can easily show that the finite element spaces for the velocity and pressure approximations have to satisfy the inf-sup condition [12, 13, 16, 17]

$$\inf_{q^h \in S^h, q^h \neq 0} \sup_{\mathbf{v}^h \in \mathbf{V}^h, \mathbf{v}^h \neq \mathbf{0}} \frac{\displaystyle\int_\Omega q^h \nabla \cdot \mathbf{v}^h \, d\Omega}{\|q^h\|_0 \|\mathbf{v}^h\|_1} \geq \gamma \tag{5.9}$$

for some constant $\gamma > 0$. This condition guarantees the unique solvability of the discrete Stokes system and restricts the choice of finite element spaces used for the velocity and pressure approximations; see [12, 13, 17] for details. In particular, one cannot use piecewise polynomial spaces of the same order and defined with respect to the same grid for the velocity and pressure approximations. If one instead uses (5.7), an even more onerous inf-sup condition is required of the finite element spaces for the velocity, vorticity, and pressure.

Remark 5.5. Note that (5.9) is a *third* level of inf-sup conditions that we have encountered in our deliberations: (5.9) is necessary and sufficient to guarantee that the inf-sup condition (3.22) holds; the latter is necessary and sufficient to guarantee that the inf-sup condition (2.11) holds.

5.2. Least-squares finite element methods for the two optimization problems. Using the associations of spaces and variables defined in section 5.1 and the operators defined in (5.7), it is easy to see that the least-squares functional (4.14) is given by, for the example problems we are considering,

$$
\begin{aligned}
&\mathcal{K}\big(\{\mathbf{u},\boldsymbol{\omega},p\},\boldsymbol{\theta},\{\mathbf{v},\boldsymbol{\sigma},q\};\widehat{\mathbf{u}},\mathbf{g}\big)\\
&\quad = \|\nabla\times\boldsymbol{\sigma}+\nabla q+\delta_2(\mathbf{u}-\widehat{\mathbf{u}})\|_{-1}^2+\|\nabla\times\mathbf{v}-\boldsymbol{\sigma}+\delta_1\boldsymbol{\omega}\|_0^2+\|\nabla\cdot\mathbf{v}\|_0^2\\
&\qquad +\|\delta\boldsymbol{\theta}+\mathbf{v}\|_0^2\\
&\qquad +\|\nabla\times\boldsymbol{\omega}+\nabla p+\boldsymbol{\theta}-\mathbf{g}\|_{-1}^2+\|\nabla\times\mathbf{u}-\boldsymbol{\omega}\|_0^2+\|\nabla\cdot\mathbf{u}\|_0^2,
\end{aligned}
\tag{5.10}
$$

where

$$
\delta_1=\begin{cases}1 & \text{for Case I,}\\ 0 & \text{for Case II,}\end{cases}\qquad\text{and}\qquad \delta_2=\begin{cases}0 & \text{for Case I,}\\ 1 & \text{for Case II.}\end{cases}
$$

We also have the bilinear form

$$
\begin{aligned}
&B\big(\{\mathbf{u},\boldsymbol{\omega},p\},\boldsymbol{\theta},\{\mathbf{v},\boldsymbol{\sigma},q\};\{\widetilde{\mathbf{u}},\widetilde{\boldsymbol{\omega}},\widetilde{p}\},\widetilde{\boldsymbol{\theta}},\{\widetilde{\mathbf{v}},\widetilde{\boldsymbol{\sigma}},\widetilde{q}\}\big)\\
&\quad = \Big(\nabla\times\boldsymbol{\sigma}+\nabla q+\delta_2\mathbf{u},\,\nabla\times\widetilde{\boldsymbol{\sigma}}+\nabla\widetilde{q}+\delta_2\widetilde{\mathbf{u}}\Big)_{-1}\\
&\qquad +\Big(\nabla\times\mathbf{v}-\boldsymbol{\sigma}+\delta_1\boldsymbol{\omega},\,\nabla\times\widetilde{\mathbf{v}}-\widetilde{\boldsymbol{\sigma}}+\delta_1\widetilde{\boldsymbol{\omega}}\Big)+\Big(\nabla\cdot\mathbf{v},\,\nabla\cdot\widetilde{\mathbf{v}}\Big)\\
&\qquad +\Big(\delta\boldsymbol{\theta}+\mathbf{v},\,\delta\widetilde{\boldsymbol{\theta}}+\widetilde{\mathbf{v}}\Big)\\
&\qquad +\Big(\nabla\times\boldsymbol{\omega}+\nabla p+\boldsymbol{\theta},\,\nabla\times\widetilde{\boldsymbol{\omega}}+\nabla\widetilde{p}+\widetilde{\boldsymbol{\theta}}\Big)_{-1}\\
&\qquad +\Big(\nabla\times\mathbf{u}-\boldsymbol{\omega},\,\nabla\times\widetilde{\mathbf{u}}-\widetilde{\boldsymbol{\omega}}\Big)+\Big(\nabla\cdot\mathbf{u},\,\nabla\cdot\widetilde{\mathbf{u}}\Big),
\end{aligned}
\tag{5.11}
$$

where $(\cdot,\cdot)_{-1}$ denotes the inner product in $\mathbf{H}^{-1}(\Omega)$, and the linear functional

$$
\begin{aligned}
&F\big(\{\widetilde{\mathbf{u}},\widetilde{\boldsymbol{\omega}},\widetilde{p}\},\widetilde{\boldsymbol{\theta}},\{\widetilde{\mathbf{v}},\widetilde{\boldsymbol{\sigma}},\widetilde{q}\};\widehat{\mathbf{u}},\mathbf{g}\big)\\
&\quad = \Big(\delta_2\widehat{\mathbf{u}},\,\nabla\times\widetilde{\boldsymbol{\sigma}}+\nabla\widetilde{q}+\delta_2\widetilde{\mathbf{u}}\Big)_{-1}+\Big(\mathbf{g},\,\nabla\times\widetilde{\boldsymbol{\omega}}+\nabla\widetilde{p}+\widetilde{\boldsymbol{\theta}}\Big)_{-1}.
\end{aligned}
\tag{5.12}
$$

Then, as in (4.16), we have that the unique minimizer of the least-squares functional (5.10) can be characterized as being the solution of the problem: Find $\{\mathbf{u},\boldsymbol{\omega},p\}\in\mathbf{H}_0^1(\Omega)\times\mathbf{L}^2(\Omega)\times L_0^2(\Omega)$, $\boldsymbol{\theta}\in\mathbf{L}^2(\Omega)$, and $\{\mathbf{v},\boldsymbol{\sigma},q\}\in\mathbf{H}_0^1(\Omega)\times\mathbf{L}^2(\Omega)\times L_0^2(\Omega)$ such that

$$
\begin{aligned}
&B\big(\{\mathbf{u},\boldsymbol{\omega},p\},\boldsymbol{\theta},\{\mathbf{v},\boldsymbol{\sigma},q\};\{\widetilde{\mathbf{u}},\widetilde{\boldsymbol{\omega}},\widetilde{p}\},\widetilde{\boldsymbol{\theta}},\{\widetilde{\mathbf{v}},\widetilde{\boldsymbol{\sigma}},\widetilde{q}\}\big)=F\big(\{\widetilde{\mathbf{u}},\widetilde{\boldsymbol{\omega}},\widetilde{p}\},\widetilde{\boldsymbol{\theta}},\{\widetilde{\mathbf{v}},\widetilde{\boldsymbol{\sigma}},\widetilde{q}\};\widehat{\mathbf{u}},\mathbf{g}\big)\\
&\qquad \forall\,\{\widetilde{\mathbf{u}},\widetilde{\boldsymbol{\omega}},\widetilde{p}\}\in\mathbf{H}_0^1(\Omega)\times\mathbf{L}^2(\Omega)\times L_0^2(\Omega),\ \ \widetilde{\boldsymbol{\theta}}\in\mathbf{L}^2(\Omega),\\
&\qquad\qquad \{\widetilde{\mathbf{v}},\widetilde{\boldsymbol{\sigma}},\widetilde{q}\}\in\mathbf{H}_0^1(\Omega)\times\mathbf{L}^2(\Omega)\times L_0^2(\Omega).
\end{aligned}
\tag{5.13}
$$

To define least-squares finite element approximations of the optimization problems, we first choose conforming finite element subspaces $\mathbf{V}^h\subset\mathbf{H}_0^1(\Omega)$, $\mathbf{W}^h\subset\mathbf{L}^2(\Omega)$, $S^h\subset L_0^2(\Omega)$, and $\mathbf{T}^h\subset\mathbf{L}^2(\Omega)$. We then minimize the functional in (5.10) over the subspaces or, equivalently, solve the problem: Find $\{\mathbf{u}^h,\boldsymbol{\omega}^h,p^h\}\in\mathbf{V}^h\times\mathbf{W}^h\times S^h$,

$\boldsymbol{\theta}^h \in \mathbf{T}^h$, and $\{\mathbf{v}^h, \boldsymbol{\sigma}^h, q^h\} \in \mathbf{V}^h \times \mathbf{W}^h \times S^h$ such that

$$\begin{aligned}
&B\big(\{\mathbf{u}^h, \boldsymbol{\omega}^h, p^h\}, \boldsymbol{\theta}^h, \{\mathbf{v}^h, \boldsymbol{\sigma}^h, q^h\}; \{\widetilde{\mathbf{u}}^h, \widetilde{\boldsymbol{\omega}}^h, \widetilde{p}^h\}, \widetilde{\boldsymbol{\theta}}^h, \{\widetilde{\mathbf{v}}^h, \widetilde{\boldsymbol{\sigma}}^h, \widetilde{q}^h\}\big) \\
&\qquad = F\big(\{\widetilde{\mathbf{u}}^h, \widetilde{\boldsymbol{\omega}}^h, \widetilde{p}^h\}, \widetilde{\boldsymbol{\theta}}^h, \{\widetilde{\mathbf{v}}^h, \widetilde{\boldsymbol{\sigma}}^h, \widetilde{q}^h\}; \widehat{\mathbf{u}}, \mathbf{g}\big) \\
&\qquad\qquad \forall\, \{\widetilde{\mathbf{u}}^h, \widetilde{\boldsymbol{\omega}}^h, \widetilde{p}\}^h \in \mathbf{V}^h \times \mathbf{W}^h \times S^h,\ \ \widetilde{\boldsymbol{\theta}}^h \in \mathbf{T}^h, \\
&\qquad\qquad\quad \{\widetilde{\mathbf{v}}^h, \widetilde{\boldsymbol{\sigma}}^h, \widetilde{q}^h\} \in \mathbf{V}^h \times \mathbf{W}^h \times S^h.
\end{aligned} \tag{5.14}$$

Proposition 5.1 and the results of section 4 allow us to prove the following results.

THEOREM 5.6. *Let $\Phi = \Lambda = \mathbf{H}_0^1(\Omega) \times \mathbf{L}^2(\Omega) \times L_0^2(\Omega)$, and let $\Theta = \mathbf{L}^2(\Omega)$. Then,*

(i) *the bilinear form $B(\cdot;\cdot)$ defined in* (5.11) *is symmetric, continuous, and coercive on $\{\Phi \times \Theta \times \Lambda\} \times \{\Phi \times \Theta \times \Lambda\}$.*

Let $\widehat{\mathbf{u}} \in \mathbf{L}^2(\Omega)$ and $\mathbf{g} \in \mathbf{H}^{-1}(\Omega)$ be given. Then

(ii) *the linear functional $F(\cdot)$ defined in* (5.12) *is continuous on $\{\Phi \times \Theta \times \Lambda\}$;*

(iii) *the problem* (5.13) *has a unique solution $(\{\mathbf{u}, \boldsymbol{\omega}, p\}, \boldsymbol{\theta}, \{\mathbf{v}, \boldsymbol{\sigma}, q\}) \in \Phi \times \Theta \times \Lambda$.*

Let $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$, $\mathbf{W}^h \subset \mathbf{L}^2(\Omega)$, $S^h \subset L_0^2(\Omega)$, and $\mathbf{T}^h \subset \mathbf{L}^2(\Omega)$, and let $\Phi^h = \Lambda^h = \mathbf{V}^h \times \mathbf{W}^h \times S^h$ and $\Theta^h = \mathbf{T}^h$. Then,

(iv) *the discrete problem* (5.13) *has a unique solution $(\{\mathbf{u}^h, \boldsymbol{\omega}^h, p^h\}, \boldsymbol{\theta}^h, \{\mathbf{v}^h, \boldsymbol{\sigma}^h, q^h\}) \in \Phi^h \times \Theta^h \times \Lambda^h$;*

(v) *we have the error estimate*

$$\begin{aligned}
&\|\mathbf{u} - \mathbf{u}^h\|_1 + \|\boldsymbol{\omega} - \boldsymbol{\omega}^h\|_0 + \|p - p^h\|_0 + \|\boldsymbol{\theta} - \boldsymbol{\theta}^h\|_0 \\
&\qquad + \|\mathbf{v} - \mathbf{v}^h\|_1 + \|\boldsymbol{\sigma} - \boldsymbol{\sigma}^h\|_0 + \|q - q^h\|_0 \\
&\le C\Big(\inf_{\widetilde{\mathbf{u}}^h \in \mathbf{V}^h} \|\mathbf{u} - \widetilde{\mathbf{u}}^h\|_1 + \inf_{\widetilde{\boldsymbol{\omega}}^h \in \mathbf{W}^h} \|\boldsymbol{\omega} - \widetilde{\boldsymbol{\omega}}^h\|_0 + \inf_{\widetilde{p}^h \in S^h} \|p - \widetilde{p}^h\|_0 \\
&\qquad + \inf_{\widetilde{\boldsymbol{\theta}}^h \in \mathbf{T}^h} \|\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}^h\|_0 + \inf_{\widetilde{\mathbf{v}}^h \in \mathbf{V}^h} \|\mathbf{v} - \widetilde{\mathbf{v}}^h\|_1 \\
&\qquad + \inf_{\widetilde{\boldsymbol{\sigma}}^h \in \mathbf{W}^h} \|\boldsymbol{\sigma} - \widetilde{\boldsymbol{\sigma}}^h\|_0 + \inf_{\widetilde{q}^h \in S^h} \|q - \widetilde{q}^h\|_0 \Big).
\end{aligned} \tag{5.15}$$

Proof. The results follow in a straightforward manner from Proposition 5.1 along with Lemma 4.2, Proposition 4.4, and Theorem 4.6. □

Remark 5.7. Following Remark 4.8, the discrete problem (5.13) is equivalent to a linear algebraic system having a symmetric, positive definite coefficient matrix. In the case of a Galerkin discretization of the optimality system, the coefficient matrix is indefinite.

Remark 5.8. Following Remark 4.9, the results in Theorem 5.6 about the solution of the discrete problem (5.13) follow merely from the conformity of the finite element subspaces, i.e., merely from the inclusions $\mathbf{V}^h \subset \mathbf{H}_0^1(\Omega)$, $\mathbf{W}^h \subset \mathbf{L}^2(\Omega)$, $S^h \subset L_0^2(\Omega)$, and $\mathbf{T}^h \subset \mathbf{L}^2(\Omega)$. In particular, unlike the case of Galerkin finite element discretizations of the optimality system, they do not require that the finite element spaces satisfy additional conditions such as (5.9); see Remark 5.4. In particular, in (5.13), one can choose the same degree piecewise polynomials defined with respect to the same grid for all variables.

Remark 5.9. The discrete problem (5.13) is a rather formidable one in that it involves many unknowns, i.e., 10 scalar fields in two dimensions and 17 scalar fields in three dimensions. However, following Remark 4.10, the discrete problem (5.13) can be efficiently uncoupled, more so than is the case for Galerkin finite element discretizations of optimality systems.

Remark 5.10. A practical Galerkin finite element discretization of the optimality system can use a formulation in terms of the operators defined in (5.8) while the least-squares based discretization employs a formulation in terms of the operators defined in (5.7). Thus, the latter approach involves more unknowns compared to the former that involves 8 scalar fields in two dimensions and 11 scalar fields in three dimensions. This apparent disadvantage of the least-squares approach should be balanced against the advantages discussed in Remarks 5.7, 5.8, and 5.9.

Remark 5.11. Suppose one chooses continuous, piecewise polynomial finite element spaces of degree r for the approximation of all variables; this is permissible for least-squares finite element methods; see Remark 5.8. Suppose also that the solution of the optimality system satisfies $\mathbf{u} \in \mathbf{H}^{r+1}(\Omega) \cap \mathbf{H}_0^1(\Omega)$, $\boldsymbol{\omega} \in \mathbf{H}^r(\Omega)$, $p \in H^r(\Omega) \cap L_0^2(\Omega)$, $\boldsymbol{\theta} \in \mathbf{H}^r(\Omega)$, $\mathbf{v} \in \mathbf{H}^{r+1}(\Omega) \cap \mathbf{H}_0^1(\Omega)$, $\boldsymbol{\sigma} \in \mathbf{H}^r(\Omega)$, and $q \in H^r(\Omega) \cap L_0^2(\Omega)$. Then, the error estimate (5.15) implies that

$$\begin{aligned} &\|\mathbf{u}-\mathbf{u}^h\|_1 + \|\boldsymbol{\omega}-\boldsymbol{\omega}^h\|_0 + \|p-p^h\|_0 + \|\boldsymbol{\theta}-\boldsymbol{\theta}^h\|_0 \\ &\qquad\qquad + \|\mathbf{v}-\mathbf{v}^h\|_1 + \|\boldsymbol{\sigma}-\boldsymbol{\sigma}^h\|_0 + \|q-q^h\|_0 = O(h^r), \end{aligned} \tag{5.16}$$

where h is a measure of the grid size.

5.2.1. Circumventing the use of negative norms. The least-squares functional (5.10) makes use of the $\mathbf{H}^{-1}(\Omega)$ norm. As a result, both the bilinear from $B(\cdot;\cdot)$ and the linear functional $F(\cdot)$ appearing in least-squares finite element discretization (5.14) of the optimality system involve the $H^{-1}(\Omega)$ inner product $(\cdot,\cdot)_{-1}$. Computing the $H^{-1}(\Omega)$ inner product of two functions essentially requires the solution of a Poisson problem, i.e., for two functions $\omega, \sigma \in H^{-1}(\Omega)$, we can write

$$(\omega,\sigma)_{-1} = \int_\Omega \omega v \, d\Omega, \qquad \text{where} \qquad -\Delta v = \sigma \quad \text{in } \Omega \quad \text{and} \quad v = 0 \quad \text{on } \Gamma.$$

Having to solve a Poisson problem every time one has to evaluate the $H^{-1}(\Omega)$ inner product of two functions renders impractical the implementation of (5.14).

There is substantial temptation to avoid the appearance of negative norms in the least-squares finite element formulation by simply replacing the negative norm in (5.10) with the $\mathbf{L}^2(\Omega)$ norm, i.e., to base a least-squares finite element method on the functional

$$\begin{aligned} &\widetilde{\mathcal{K}}(\{\mathbf{u},\boldsymbol{\omega},p\},\boldsymbol{\theta},\{\mathbf{v},\boldsymbol{\sigma},q\};\widehat{\mathbf{u}},\mathbf{g}) \\ &\quad = \|\nabla\times\boldsymbol{\sigma} + \nabla q + \delta_2(\mathbf{u}-\widehat{\mathbf{u}})\|_0^2 + \|\nabla\times\mathbf{v} - \boldsymbol{\sigma} + \delta_1\boldsymbol{\omega}\|_0^2 + \|\nabla\cdot\mathbf{v}\|_0^2 \\ &\qquad + \|\delta\boldsymbol{\theta} + \mathbf{v}\|_0^2 + \|\nabla\times\boldsymbol{\omega} + \nabla p + \boldsymbol{\theta} - \mathbf{g}\|_0^2 + \|\nabla\times\mathbf{u} - \boldsymbol{\omega}\|_0^2 + \|\nabla\cdot\mathbf{u}\|_0^2 \end{aligned}$$

instead of the functional (5.10); note that now we would have to choose $\Phi = \Lambda = \mathbf{H}_0^1(\Omega) \times \mathbf{H}^1(\Omega) \times L_0^2(\Omega)$. Doing this would indeed lead to a discrete problem involving only easily implementable $L^2(\Omega)$ inner products. However, in this case, the norm-equivalence relation does not hold (see [6, 7]) so that the resulting bilinear form associated with the minimization of the functional $\widetilde{\mathcal{K}}$ is not coercive. As a result, the discrete problem will not have a (uniformly, as $h \to 0$) positive definite coefficient matrix, and the least-squares finite element approximations may not be stable and will certainly not be optimally accurate.

Another approach for avoiding the use of $H^{-1}(\Omega)$ inner products is to replace the velocity-vorticity-pressure formulation (5.4) of the Stokes problem with another first-order formulation whose residuals, when measured in $L^2(\Omega)$ norms, do result in a norm-equivalent functional. Such formulations, involving additional unknowns and redundant equations, were developed in [14, 15]. For example, if we use the velocity-velocity gradient-pressure formulation due to [14], we would employ the least-squares functional

$$\begin{aligned}
&\widehat{\mathcal{K}}\big(\{\mathbf{u},\mathbf{U},p\},\boldsymbol{\theta},\{\mathbf{v},\mathbf{V},q\};\widehat{\mathbf{u}},\mathbf{g}\big)\\
&\quad= \|\nabla\cdot\mathbf{V}+\nabla q+\delta_2(\mathbf{u}-\widehat{\mathbf{u}})\|_0^2+\|(\nabla\mathbf{v})^T-\mathbf{V}+\delta_1\boldsymbol{\omega}\|_0^2+\|\nabla\cdot\mathbf{v}\|_0^2\\
&\qquad\quad+\|\delta\boldsymbol{\theta}+\mathbf{v}\|_0^2+\|-\nabla\cdot\mathbf{U}+\nabla p+\boldsymbol{\theta}-\mathbf{g}\|_0^2+\|(\nabla\mathbf{u})^T-\mathbf{U}\|_0^2+\|\nabla\cdot\mathbf{u}\|_0^2\\
&\qquad\quad+\|\nabla(\mathrm{Tr}\mathbf{V})\|_0^2+\|\nabla\times\mathbf{V}\|_0^2+\|\nabla(\mathrm{Tr}\mathbf{U})\|_0^2+\|\nabla\times\mathbf{U}\|_0^2,
\end{aligned}$$

where $(\cdot)^T$ and $\mathrm{Tr}(\cdot)$ denote the transpose and the trace of a tensor and where the components of $\boldsymbol{\omega}$ can be easily expressed as linear combinations of the off-diagonal elements of the tensor $\mathbf{U}$. Instead of the vorticity $\boldsymbol{\omega}=\nabla\times\mathbf{u}$ and adjoint vorticity $\boldsymbol{\sigma}=\nabla\times\mathbf{v}$, the new variables introduced to effect the first-order formulation are $\mathbf{U}=(\nabla\mathbf{u})^T$ and $\mathbf{V}=(\nabla\mathbf{v})^T$. Also, now we have that $\Phi=\Lambda=\mathbf{H}_0^1(\Omega)\times Q\times L_0^2(\Omega)$, where $Q=\{\mathbf{V}\in[H^1(\Omega)]^9 \mid \mathbf{V}\times\mathbf{n}=0\}$. The equations whose residuals appear in the last line of the definition of $\widehat{\mathcal{K}}$ are all redundant in the sense that they are all already implied by the other equations. Note that now we have even more unknowns than that for the velocity-vorticity-pressure; e.g., in three dimensions, the least-squares discrete problem resulting from minimizing the functional $\widehat{\mathcal{K}}$ would now involve 27 scalar fields. Furthermore, the addition of redundant equations requires more regular data and solutions and precludes the use of the least-squares methodology in, e.g., nonconvex polygonal domains.

A third and more practical approach to avoiding the use of $H^{-1}(\Omega)$ inner products is to replace the functional (5.10) by the mesh-weighted functional

$$\begin{aligned}
&\widetilde{\mathcal{K}}_h\big(\{\mathbf{u},\boldsymbol{\omega},p\},\boldsymbol{\theta},\{\mathbf{v},\boldsymbol{\sigma},q\};\widehat{\mathbf{u}},\mathbf{g}\big)\\
&\quad= h^2\|\nabla\times\boldsymbol{\sigma}+\nabla q+\delta_2(\mathbf{u}-\widehat{\mathbf{u}})\|_0^2+\|\nabla\times\mathbf{v}-\boldsymbol{\sigma}+\delta_1\boldsymbol{\omega}\|_0^2+\|\nabla\cdot\mathbf{v}\|_0^2\\
&\qquad\quad+\|\delta\boldsymbol{\theta}+\mathbf{v}\|_0^2+h^2\|\nabla\times\boldsymbol{\omega}+\nabla p+\boldsymbol{\theta}-\mathbf{g}\|_0^2+\|\nabla\times\mathbf{u}-\boldsymbol{\omega}\|_0^2+\|\nabla\cdot\mathbf{u}\|_0^2.
\end{aligned}$$

This approach is motivated by the finite element inverse inequality $C\|\omega^h\|_0\le h^{-1}\|\omega^h\|_{-1}$ which leads to the norm "equivalence" $Ch\|\omega^h\|_0\le\|\omega^h\|_{-1}\le\|\omega^h\|_0$ between the $H^{-1}(\Omega)$ and $L^2(\Omega)$ norms of finite element functions. One can then show, using the analyses developed in [6], that one obtains an optimal convergence for the functional $\widetilde{\mathcal{K}}_h$, even though this functional is not norm-equivalent. One possible drawback of this approach is that the condition number of the resulting matrix may be too large for the practical use of some iterative solution techniques.

Perhaps the most practical approach to avoiding the use of $H^{-1}(\Omega)$ inner products is to replace the $H^{-1}(\Omega)$ norm terms in the functional (5.10) by more sophisticated "equivalent" discrete norms that involve only $L^2(\Omega)$ norms. Such ideas have been widely used in the least-squares finite element literature; see, e.g., [4, 10, 11]. As noted above, the computation of negative norms requires inversion of a Laplacian operator (with zero boundary conditions). It was shown in [10] that for finite element functions, it is equivalent to use the discrete minus one inner product

$$(\omega^h, \sigma^h)_h = \big((L^h + h^2 I)\omega^h, \sigma^h\big)_0\,,$$

where L^h is a discrete inverse Laplacian operator (with zero boundary conditions) that is spectrally equivalent to the inverse Laplacian operator itself. In practice, the computation of $L^h\omega^h$ for any finite element function ω^h is often implemented by using a few multigrid cycles, which makes its computation very efficient. The application of this approach in our context results in the minimization of the functional

$$\begin{aligned}&\overline{\mathcal{K}}_h\big(\{\mathbf{u}^h, \boldsymbol{\omega}^h, p^h\}, \boldsymbol{\theta}^h, \{\mathbf{v}^h, \boldsymbol{\sigma}^h, q^h\}; \widehat{\mathbf{u}}, \mathbf{g}\big)\\ &\quad = \|\nabla\times\boldsymbol{\sigma}^h + \nabla q^h + \delta_2(\mathbf{u}^h - \widehat{\mathbf{u}})\|_h^2 + \|\nabla\times\mathbf{v}^h - \boldsymbol{\sigma}^h + \delta_1\boldsymbol{\omega}^h\|_0^2 + \|\nabla\cdot\mathbf{v}^h\|_0^2\\ &\qquad + \|\delta\boldsymbol{\theta}^h + \mathbf{v}^h\|_0^2 + \|\nabla\times\boldsymbol{\omega}^h + \nabla p^h + \boldsymbol{\theta}^h - \mathbf{g}\|_h^2 + \|\nabla\times\mathbf{u}^h - \boldsymbol{\omega}^h\|_0^2 + \|\nabla\cdot\mathbf{u}^h\|_0^2,\end{aligned}$$

where $\|\omega^h\|_h^2 = ((L^h + h^2 I)\omega^h, \omega^h)_0$. Using the techniques of [4], it can be shown that this functional leads to a practical least-squares finite element method yielding positive definite coefficient matrices and an error estimate such as (5.16).

6. Concluding remarks.

Optimization and control problems governed by partial differential equations are most often solved by Lagrange multiplier techniques that lead to variational equations consisting of the state system, an adjoint-state system, and optimality conditions. Galerkin discretizations of such systems result in discrete problems that are not only formidable in size but are indefinite so that their iterative solution is difficult.

In this paper, we formulated a new approach for the finite element discretization of optimality systems that is based on the application of least-squares principles. The main advantage of this formulation is seen in the better possibilities that it affords for the uncoupling of the discrete optimality equations and their efficient iterative solution. Least-squares principles result in symmetric and positive definite algebraic systems. Moreover, for the optimization and control problems considered in this paper, these linear systems have a 3 × 3 block structure where the diagonal blocks themselves are symmetric and positive definite. As an example of a simple but convergent uncoupling strategy, we considered a block-Gauss–Seidel method. To illustrate the issues involved in the formulation of effective and practical least-squares methods, we considered two optimization problems for the Stokes equations.

REFERENCES

[1] I. Babuska and A. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, Academic, New York, 1972, pp. 1–359.

[2] D. Bedivan and G. Fix, *Least-squares methods for optimal shape design problems*, Comput. Math. Appl., 30 (1995), pp. 17–25.

[3] P. Bochev, *Least-squares methods for optimal control*, Nonlinear Anal., 30 (1997), pp. 1875–1885.

[4] P. Bochev, *Negative norm least-squares methods for the velocity-vorticity-pressure Navier–Stokes equations*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 237–256.

[5] P. Bochev and D. Bedivan, *Least-squares methods for Navier-Stokes boundary control problems*, Int. J. Comput. Fluid Dyn., 9 (1997), pp. 43–58.

[6] P. Bochev and M. Gunzburger, *Least-squares finite element methods for elliptic equations*, SIAM Rev., 40 (1998), pp. 789–837.

[7] P. Bochev and M. Gunzburger, *Analysis of least-squares finite element methods for the Stokes equations*, Math. Comp., 63 (1994), pp. 479–506.

[8] P. Bochev and M. Gunzburger, *Least-squares finite element methods for optimization and control problems for the Stokes equations*, Comput. Math. Appl., 48 (2004), pp. 1035–1057.

[9] P. Bochev and M. Gunzburger, *On least-squares variational principles for the discretization of optimization and control problems*, to appear in Math. Anal. Appl.

[10] J. Bramble, R. Lazarov, and J. Pasciak, *A Least Squares Approach Based on a Discrete Minus One Inner Product for First Order Systems*, Technical report 94-32, Mathematical Science Institute, Cornell University, Ithaca, NY, 1994.

[11] J. Bramble and J. Pasciak, *Least-squares methods for Stokes equations based on a discrete minus one inner product*, J. Comput. Appl. Math., 74 (1996), pp. 155–173.

[12] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers*, RAIRO Anal. Numer. 8 (1974), pp. 129–151.

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

[14] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for the Stokes equations, with application to linear elasticity,* SIAM J. Numer. Anal., 34 (1997), pp. 1727–1741.

[15] C.-L. Chang, *A mixed finite element method for the Stokes problem: An acceleration pressure formulation*, Appl. Math. Comput., 36 (1990), pp. 135–146.

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

[17] M. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic, Boston, 1989.

[18] M. Gunzburger, *Perspectives in Flow Control and Optimization*, SIAM, Philadelphia, 2002.

[19] M. Gunzburger and H. C. Lee, *Analysis and approximation of optimal control problems for first-order elliptic systems in three dimensions*, Appl. Math. Comput., 100 (1999), pp. 49–70.

[20] M. Gunzburger and H. C. Lee, *A penalty/least-squares method for optimal control problems for first-order elliptic systems*, Appl. Math. Comput., 107 (2000), pp. 57–75.

[21] J. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, New York, 1971.

[22] H. Schlichting and K. Gersten, *Boundary Layer Theory*, Springer, Berlin, 2000.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2544–2566

LOCAL PROJECTION STABILIZATION FOR THE OSEEN PROBLEM AND ITS INTERPRETATION AS A VARIATIONAL MULTISCALE METHOD*

M. BRAACK† AND E. BURMAN‡

Abstract. We propose to apply the recently introduced local projection stabilization to the numerical computation of the Oseen equation at high Reynolds number. The discretization is done by nested finite element spaces. Using a priori error estimation techniques, we prove the convergence of the method. The a priori estimates are independent of the local Peclet number and give a sufficient condition for the size of the stabilization parameters in order to ensure optimality of the approximation when the exact solution is smooth. Moreover, we show how this method may be cast in the framework of variational multiscale methods. We indicate what modeling assumptions must be made to use the method for large eddy simulations.

Key words. stabilized finite elements, Galerkin methods, multiscale

AMS subject classifications. 35Q30, 65N12, 65N30, 76D05, 76M10, 76F65

DOI. 10.1137/050631227

1. Introduction.

In this paper, we advocate the use of the two-level stabilization scheme (see Becker and Braack [3]) for the computation of solutions of the Navier–Stokes equations at high Reynolds number. This is one in a group of more recently developed stabilized methods, such as, for instance, Guermond [14], Becker and Braack [2], Burman and Hansbo [5, 6], and Rebollo and Delgado [18].

A main advantage of this approach is that it shares similar conservation properties with a standard Galerkin finite element method. Moreover, one does not need to resort to space-time finite elements for time stepping in order to stay consistent but can apply any higher-order finite difference scheme for the discretization in time. We prove optimal order a priori error estimates for the method. The method remains stable independent of the local Reynolds number. It should be noted that the method is not residual based in the way the streamline upwind Petrov–Galerkin (SUPG) method is, but is flexible with respect to the subgrid model, thus leaving open the possibility of subgrid models that are more complex than the linear one considered here. Of course, the possibility of using Galerkin least-squares (GLS) or residual-free bubbles as a subgrid model remains and will be discussed.

An attractive feature of this method is that it can be cast in the framework of the variational multiscale (VMS) method of [15] as we shall show. The stabilization is acting only on the smallest resolved scales of the flow. Hence, contrary to the method of [15], our large scales do not need any additional stabilization. In fact, the fine scale

*Received by the editors May 11, 2005; accepted for publication (in revised form) June 22, 2005; published electronically January 12, 2006. This work was initiated while the first author was a guest of the Bernoulli Center program on Mathematical Modelling for Cardiovascular Flows; it was completed while the second author was invited by the “Sonderforschungsbereich 359: Reactive Flow, Diffusion and Transport” at the University of Heidelberg. The support of both institutions is gratefully acknowledged.

http://www.siam.org/journals/sinum/43-6/63122.html

†Institute of Applied Mathematics, University of Heidelberg, INF 294, 69120 Heidelberg, Germany (malte.braack@iwr.uni-heidelberg.de).

‡École Polytechnique Fédérale de Lausanne, Institute of Analysis Modelling and Scientific Computing, CH-1015 Lausanne, Switzerland (erik.burman@epfl.ch).

fluctuations allow for both the satisfaction of the inf-sup condition and stabilization of the convective terms.

We begin in section 2 with the variational formulation of the Oseen equations and their discretization by finite elements in space. The local projection stabilization for the Oseen system is formulated in section 3. Afterwards, an a priori error analysis is presented in section 4. There we consider the case of smooth solutions (velocities and pressure are both in the Sobolev space $H^2(\Omega)$) and discuss the behavior of the method for less regular solutions. We present some variants of the stabilization operator and discuss the relation to more standard stabilization techniques in section 5. An interpretation of the stabilization in terms of a VMS method is given in section 6. In a numerical test case, discussed in section 7, we investigate the convergence order for a given exact Navier–Stokes solution and compare the kinetic energy of a nonstationary driven cavity flow with the numerical dissipation. We finish with a short conclusion in section 8. In forthcoming work we will give numerical evidence of the performance of the numerical scheme for turbulent flow in three space dimensions.

2. Variational formulation of the Oseen system.

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a polygonal domain with boundary $\partial\Omega$. The velocities will be denoted by $v = v(x, t)$ and the pressure by $p = p(x, t)$. The gradient in space is denoted by ∇, and the divergence with respect to space is denoted by div. The Navier–Stokes equations read

$$\left.\begin{array}{rcl} \partial_t v + \operatorname{div}(v \otimes v) - \mu\Delta v + \nabla p &=& f, \\ \operatorname{div} v &=& 0 \end{array}\right\} \quad \text{in } \Omega, \tag{2.1}$$

subject to some initial condition $v(\cdot, 0) = v_0$ and suitable boundary conditions for v.

In the analysis, we will consider the Oseen system as an important linearization of (2.1). For ease of presentation, we suppose homogeneous Dirichlet boundary conditions:

$$\left.\begin{array}{rcll} \sigma\, v + \operatorname{div}(\beta \otimes v) - \mu\Delta v + \nabla p &=& f & \text{in } \Omega, \\ \operatorname{div} v &=& 0 & \text{in } \Omega, \\ v &=& 0 & \text{on } \partial\Omega, \end{array}\right\} \tag{2.2}$$

with some given solenoidal vector field β and $\sigma > 0$.

For the variational formulation we use the notation $(\cdot, \cdot)$ for the L^2-scalar product over Ω. The velocity is sought in the Sobolev space $V := [H_0^1(\Omega)]^d$, and the pressure is sought in the space of square-integrable functions with zero mean, $Q := L_0^2(\Omega)$. The product space for the vector $u = \{v, p\}$ is denoted by $X := V \times Q$.

We introduce the bilinear form $A(u, \varphi)$ defined by

$$A(u, \varphi) := (\sigma v, \psi) - (\beta \otimes v, \nabla\psi) - (p, \operatorname{div}\psi) + (\operatorname{div} v, \xi) + (\mu\nabla v, \nabla\psi),$$

and consider f as an element of X' defined by $\langle f, \varphi\rangle := (f, \psi)$ for a test function $\varphi = \{\psi, \xi\} \in X$. The variational formulation of the Oseen problem (2.2) reads

$$u \in X : \quad A(u, \varphi) = \langle f, \varphi\rangle \quad \forall \varphi \in X. \tag{2.3}$$

In order to solve this problem numerically we choose in the following section a finite dimensional subspace, $X_h \subset X$.

2.1. Discrete Galerkin formulation.

We consider shape regular meshes $\mathcal{T}_h = \{K\}$ of hexahedral elements K with the minimum mesh size $h = \min\{h_K : K \in \mathcal{T}_h\}$

(quadrilateral elements for the academical case $d = 2$). The finite element spaces Q_h^r result from isoparametric transformations of polynomials on a reference cell $\hat{K}$:

$$Q_h^r(\Omega, \mathbb{R}) := \{\varphi \in C(\Omega, \mathbb{R}) : \varphi|_K = \hat{\varphi} \circ T_K^{-1}\},$$

where $\hat{\varphi}$ denotes an arbitrary polynomial of maximal degree r on the reference cell $\hat{K}$, and $T_K : \hat{K} \to K$ denotes a polynomial transformation of the same type and degree r. We will treat (bi-/tri-)linear elements ($r = 1$) and (bi-/tri-)quadratic elements ($r = 2$) simultaneously in the analysis. These finite element spaces will simply be called Q_1 in the case of $r = 1$ and Q_2 elements in the case $r = 2$. The discrete pressure space Q_h is the subspace of Q_h^r with zero mean, and the velocity space V_h is the subspace with vanishing trace:

$$Q_h := Q \cap Q_h^r(\Omega, \mathbb{R}), \quad V_h := V \cap [Q_h^r(\Omega, \mathbb{R}^d)]^d.$$

The product space is denoted by X_h:

$$X_h := V_h \times Q_h.$$

In the Galerkin formulation of (2.3) for the space X_h, a discrete solution $u_h \in X_h$ is sought such that

$$A(u_h, \varphi) = \langle f, \varphi \rangle \quad \forall \varphi \in X_h.$$

This formulation is not stable due to the following reasons: (i) violation of the discrete inf-sup (or Babuska–Brezzi) condition for velocity and pressure approximation and (ii) dominating advection (and reaction). Therefore, in the following we present and analyze a stabilization technique based on local projection.

3. Definition of the local projection stabilization. The two-level finite element formulation is as follows: find $u_h \in X_h$ such that

$$A(u_h, \varphi) + S_h(u_h, \varphi) = \langle f, \varphi \rangle \quad \forall \varphi \in X_h. \tag{3.1}$$

In order to specify the stabilization term $S_h(\cdot, \cdot)$, we have to introduce further notations. The discontinuous analogue of Q_h^r is denoted by $Q_{h,disc}^r$:

$$Q_{h,disc}^r(\Omega, \mathbb{R}) := \{\varphi \in L^2(\Omega, \mathbb{R}) : \varphi|_K = \hat{\varphi} \circ T_K^{-1}\}.$$

Furthermore, let $\mathcal{T}_{2h}$ be the coarser mesh obtained by a "global coarsening" of $\mathcal{T}_h$. Obviously, the finer mesh $\mathcal{T}_h$ contains 2^d times more elements than $\mathcal{T}_{2h}$. The corresponding finite element spaces are denoted by $Q_{2h} \subset Q_h$ and $V_{2h} \subset V_h$. Let D_h^v and D_h^p be the following space for pressure and velocities, respectively, of functions allowing discontinuities across elements of $\mathcal{T}_{2h}$:

$$\begin{aligned} D_h^v &:= [Q_{2h,disc}^{r-1}(\Omega, \mathbb{R})]^d, \\ D_h^p &:= Q_{2h,disc}^{r-1}(\Omega, \mathbb{R}). \end{aligned}$$

In the case $r = 1$, these spaces contain patchwise ($K \in V_{2h}$) constants; for $r = 2$, they contain patchwise linear elements. We will make use of the L^2-projection operator

$$\bar{\pi}_h : L^2(\Omega) \to D_h^p,$$

characterized by the property

$$(v - \bar{\pi}_h v, \phi) = 0 \quad \forall \phi \in D_h^p.$$

The operator giving the space fluctuations is denoted by

$$\bar{\varkappa}_h := i - \bar{\pi}_h,$$

with the identity mapping i. We use the same notation $\bar{\pi}_h$, $\bar{\varkappa}_h$ for the mappings on vector-valued functions, for instance, $\bar{\pi}_h : L^2(\Omega)^d \to D_h^v$.

The subgrid model is given by

$$S_h(u, \varphi) := (\delta \bar{\varkappa}_h \nabla v, \bar{\varkappa}_h \nabla \psi) + (\alpha \bar{\varkappa}_h \nabla p, \bar{\varkappa}_h \nabla \xi). \tag{3.2}$$

The parameters α and δ are taken patchwise constant and depend on the local mesh size. The optimal choice of these stabilization parameters will be a result of the following analysis.

We like to end this section with a brief discussion on the numerical costs of the scheme compared to more standard stabilized schemes. Compared to the standard Galerkin formulation, the subgrid models (3.2) lead to a larger stencil in the stiffness matrix due to the projection $\bar{\varkappa}_h$. However, no couplings between pressure and velocities are introduced. Furthermore, a cheaper preconditioner may be used with a smaller stencil as proposed and analyzed for the Stokes system in [2].

4. A priori error analysis.

To tune the stabilization parameters α and δ we use a priori error estimation. Assuming sufficient regularity of the underlying solution, the parameters are chosen in such a way that the method has optimal convergence properties independent of the viscosity. We will prove under the assumption of sufficiently regular pressure and velocity $v \in [H_0^2(\Omega)]^3$, $p \in H^2(\Omega) \cap L_0^2(\Omega)$, that a certain scaling of α and δ gives optimal convergence of the velocities independent of the Reynolds number. A similar result is then proved for the L^2-norm of the pressure. We consider only the interesting case of high Reynolds number, hence assuming that $\mu \leq |\beta| h$. First we prove an estimate for a mesh-dependent norm including the H^1-norm of the velocities and a "subgrid model" error:

$$|\!|\!| u |\!|\!| = |\!|\!| \{v, p\} |\!|\!| := \Big(\|\sigma^{1/2} v\|^2 + \|\mu^{1/2} \nabla v\|^2 + S_h(u, u) \Big)^{1/2},$$

where $\|\cdot\|$ stands for the L^2-norm in Ω. We then use this estimate to recover control of the pressure and show that the L^2-norm error of the pressure is bounded by the mesh-dependent norm of the error of the state vector u_h.

4.1. Properties of the subgrid model.

By the following coercivity result we deduce existence and uniqueness of the discrete velocities.

Lemma 4.1. *We have the coercivity property*

$$|\!|\!| u |\!|\!|^2 = A(u, u) + S_h(u, u) \quad \forall u \in X. \tag{4.1}$$

Proof. The proof follows immediately by integration by parts. □

We have the following approximate Galerkin orthogonality.

Lemma 4.2. *Let $u \in X$ be the solution of the weak formulation of* (2.2) *and let $u_h \in X_h$ be the solution of its discrete version* (3.1). *Then it holds that*

$$A(u - u_h, \varphi) = S_h(u_h, \varphi), \quad \varphi \in X_h.$$

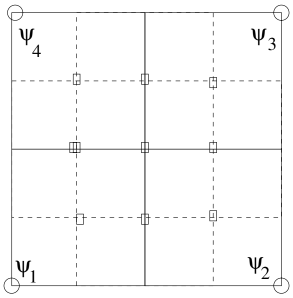


FIG. 4.1. *A patch K of four Q_2 cells in two dimensions with the linears ψ_i used in the proof of Lemmas* 4.4 *and* 4.5.

Proof. The proof is obtained by subtracting (3.1) from the weak formulation of (2.2). □

Since the method is not strongly consistent in the sense that we do not have full Galerkin orthogonality, we must analyze the asymptotic behavior of the subgrid model, i.e., the dependence with respect to the mesh size h. We first prove a result for a modified Clément interpolation operator introduced in [2], here with a generalization to P_2 elements, and use this approximation result to show the asymptotic behavior of the stabilization term.

In the next lemma, we consider for a cell $K \in \mathcal{T}_{2h}$, the space of functions in Q_h with support in K. This space will be denoted by $Q_h(K)$ and has the dimension $(2r-1)^d$. Analogously, the subspace of D_h^p consisting of functions with support in K will be denoted by $D_h^p(K)$. This space has the dimension r^d. For $r = 1$, this space consists only of the patchwise constants. For $r = 2$, a possible basis of these subspaces is indicated in Figure 4.1.

LEMMA 4.3. *The local L^2-orthogonal projection $\pi_K : D_h^p(K) \to Q_h(K)$, characterized for $\psi \in D_h^p(K)$ by the property*

$$(\pi_K\psi, \phi) = (\psi, \phi) \quad \forall \phi \in Q_h(K),$$

is injective.

Proof. We assume $\pi_K\psi = 0$ for $\psi \in D_h^p(K)$. Then $(\psi, \phi) = 0$ for all $\phi \in Q_h(K)$ due to the orthogonality property. In the case $r = 1$, $D_h^p(K)$ consists of constant functions, so that either $\psi = 0$ or $\int_K \phi = 0$ for all $\phi \in Q_h(K)$. Since the latter is not valid (for instance, taking the Lagrange nodal basis function associated to the interior node of K), it follows $\psi = 0$. For $r = 2$, we take as $\phi_i \in Q_h(K)$, $i = 1, \dots, 2^d$, the Lagrange nodal functions associated to the center nodes of the child cells K_i. Since these ϕ_i have a sign, ψ must have zeros in the interior of all child cells K_i, $i = 1, \dots, 2^d$. For a d-linear ψ this is possible only if $\psi = 0$. □

LEMMA 4.4. *Let $\{\psi_1, \dots, \psi_{r^d}\}$ be an arbitrary basis of $D_h^p(K)$. Then the matrix $M = (m_{ij})$, $i, j \in \{1, \dots, r^d\}$, with entries*

$$m_{ij} = (\psi_i, \pi_K\psi_j)_K,$$

is symmetric and positive definite.

Proof. Since $\pi_K\psi_j \in Q_h(K)$, it follows due to the orthogonality property of the L^2-projection that

$$m_{ij} = (\psi_i, \pi_K\psi_j) = (\pi_K\psi_i, \pi_K\psi_j).$$

It follows that M is symmetric. Furthermore, for $\alpha \in \mathbb{R}^{r^d}$ and $\psi_\alpha = \sum \alpha_i \psi_i$ it holds that

$$\begin{aligned} \alpha^T M \alpha &= (\psi_\alpha, \pi_K \psi_\alpha)_K \\ &= (\pi_K \psi_\alpha, \pi_K \psi_\alpha)_K \\ &= \|\pi_K \psi_\alpha\|_K^2. \end{aligned}$$

Since π_K is injective, M is positive definite. □

Remark. The analogous results are valid for vector-valued projections $\pi_K : D_h^v(K) \to V_h(K)$, where $D_h^v(K) \subset D_h^v$ and $V_h(K) \subset V_h$ are defined analogously as $D_h^p(K)$ and $Q_h(K)$, respectively.

In the following, the norm in $H^s(\Omega)$ will be denoted by $\|\cdot\|_s$. The corresponding norms in subsets $K \subset \Omega$ will be denoted by $\|\cdot\|_{s,K}$. Furthermore, we use the notation $\lesssim$ to indicate that there may arise mesh-independent constants in the estimates.

LEMMA 4.5. *There is an interpolation operator*

$$j_h : V \to V_h$$

with the orthogonality property

$$(v - j_h v, \psi) = 0 \quad \forall \psi \in D_h^v, \forall v \in V, \tag{4.2}$$

that has optimal approximation properties in the L^2-norm and H^1-seminorm

$$\|v - j_h v\| \lesssim h^l \|v\|_l \quad \forall v \in [H^l(\Omega)]^d,\ 0 \le l \le r+1, \tag{4.3}$$

$$\|\nabla(v - j_h v)\| \lesssim h^{l-1} \|v\|_l \quad \forall v \in [H^l(\Omega)]^d,\ 0 \le l \le r+1, \tag{4.4}$$

with $r \in \{1, 2\}$, and is H^1-stable:

$$\|j_h v\|_1 \lesssim \|v\|_1 \quad \forall v \in [H^1(\Omega)]^d. \tag{4.5}$$

Proof. The construction uses the Scott and Zhang variant of the Clément interpolation operator $j_h^{Cl} : V \to V_h$ (see [19] and Clément [8]), which already fulfills the approximation properties (4.3) and (4.4), maintains homogeneous Dirichlet values, and has the stability property (4.5). In order to ensure (4.2), we define j_h in the form

$$j_h = j_h^{Cl} + m_h,$$

with a local projection $m_h : V \to \tilde{V}_h$ onto the subspace

$$\tilde{V}_h := \bigoplus_{K \in \mathcal{T}_{2h}} V_h(K) \subset V_h.$$

In order to fulfill (4.2) this mapping must satisfy

$$(m_h v, \psi) = (v - j_h^{Cl} v, \psi), \quad \forall \psi \in D_h^v(K), \forall K \in \mathcal{T}_{2h}. \tag{4.6}$$

If we take a basis $\psi_{K,i}$ of $D_h^v(K)$, $m_h v$ can be expressed on each patch $K \in \mathcal{T}_{2h}$ as a linear combination of the $\pi_K \psi_{K,i}$. Hence property (4.6) is equivalent to solving for each K the linear system $M\alpha = \beta$, with the regular matrix M of Lemma 4.4 and the right-hand side β with coefficients

$$\beta_i = (v - j_h^{Cl} v, \psi_{K,i})_K.$$

Hence, j_h is well defined. In order to prove the approximation property in the L^2-norm (4.3) we will show that

$$\|m_h v\|_K \lesssim \|v - j_h^{Cl} v\|_K, \tag{4.7}$$

because due to (4.7) it follows:

$$\begin{aligned} \|v - j_h v\| &= \|v - j_h^{Cl} v + m_h v\| \\ &\le \|v - j_h^{Cl} v\| + \|m_h v\| \\ &\lesssim \|v - j_h^{Cl} v\|. \end{aligned}$$

Let us verify (4.7): The (v-dependent) solution vector α contains the coefficients of $m_h v$ in the basis $\pi_K \psi_{K,i}$. Similarly, the projection of $m_h v$ onto the vectors $\psi_{K,i}$ can be expressed as a linear combination of the $\psi_{K,i}$ with coefficients $\bar\alpha_{K,i}$:

$$\bar\pi_h m_h v = \sum_{i=1}^{r^d} \bar\alpha_{K,i} \psi_{K,i}.$$

Now, the equality

$$(\bar\pi_h m_h v, \psi_{K,i})_K = (m_h v, \psi_{K,i})_K, \quad i = 1, \dots, r^d, \tag{4.8}$$

can be written with the help of the matrix $N = (n_{ij})$ with coefficients $n_{ij} = (\psi_{K,i}, \psi_{K,j})$:

$$N\bar\alpha = M\alpha.$$

The local mass matrix N is symmetric and positive definite as well. It then follows that

$$\begin{aligned} \lambda_{min}(M)\, \|m_h v\|_K^2 &\le (M\alpha)^T \cdot M\alpha \\ &= (N\bar\alpha)^T N\bar\alpha \\ &\le \lambda_{max}(N)\, \|\bar\pi_h m_h v\|_K^2. \end{aligned}$$

We conclude that there holds on each patch $K \in \mathcal{T}_{2h}$

$$\|m_h v\|_K^2 \lesssim \|\bar\pi_h m_h v\|_K^2 = (m_h v, \bar\pi_h(m_h v))_K, \tag{4.9}$$

where the constant is independent of the mesh size. We estimate on each patch $K \in \mathcal{T}_{2h}$ due to (4.9):

$$\begin{aligned} \|m_h v\|_K^2 &\lesssim \int_K (v - j_h^{Cl} v) \bar\pi_h m_h v \, dx \\ &\le \|v - j_h^{Cl} v\|_K \, \|\bar\pi_h m_h v\|_K \\ &\lesssim \|v - j_h^{Cl} v\|_K \, \|m_h v\|_K, \end{aligned}$$

where we used the L^2-stability of $\bar\pi_h$ in the last inequality. Thus (4.7) follows.

For proving (4.4) we proceed in a similar fashion by applying an inverse estimate:

$$\begin{aligned} \|\nabla m_h v\|^2 &\lesssim \sum_{K \in \mathcal{T}_h} h_K^{-2} \|m_h v\|_K^2 \\ &= \sum_{K \in \mathcal{T}_h} h_K^{-2} \|v - j_h^{Cl} v\|_K^2 \\ &\lesssim \sum_{K \in \mathcal{T}_h} h_K^{2(l-1)} \|v\|_{l,\widetilde{K}}^2 \\ &\lesssim h^{2(l-1)} \|v\|_l^2. \end{aligned}$$

The stability (4.5) follows also due to this last estimate, the stability of j_h^{CL}, and (4.7). □

Remark. The interpolation operator j_h^{Cl} maintains *homogeneous* Dirichlet conditions on (parts of) $\partial\Omega$. For polynomial Dirichlet conditions, the interpolation introduced by Melenk and Wohlmuth [16] can be used. The interpolation operator j_h acts on the velocity space, but the result holds true of course for the scalar space $L^2(\Omega)$. We will therefore use the notation j_h also for the interpolation operator acting on the state variable $u = \{v, p\}$.

In the following analysis, we make use of the interpolation and stability properties of $\bar{\varkappa}_h$.

LEMMA 4.6. *The fluctuation operator $\bar{\varkappa}_h$ has the following interpolation and stability properties:*

$$\|\bar{\varkappa}_h \nabla v\| \lesssim h^r \|v\|_{r+1} \quad \forall v \in H^{r+1}(\Omega), \tag{4.10}$$

$$\|\bar{\varkappa}_h v\| \lesssim \|v\| \qquad \forall v \in L^2(\Omega). \tag{4.11}$$

Proof. The interpolation property (4.10) is an immediate consequence of the patch-wise interpolation of $\bar{\pi}_h$ for the H^1 function $w := \nabla v$:

$$\|\bar{\varkappa}_h \nabla v\|_K = \|w - \bar{\pi}_h w\|_K \lesssim h_K^r \|w\|_{r,K} \le h_K^r \|v\|_{r+1,K} \quad \forall K \in \mathcal{T}_{2h}.$$

Stability of $\bar{\varkappa}_h$ is due to the L^2-stability of $\bar{\pi}_h$:

$$\|\bar{\varkappa}_h v\| \le \|v\| + \|\bar{\pi}_h v\| \lesssim \|v\|. \qquad \square$$

LEMMA 4.7. *For the interpolation operator j_h of Lemma* 4.5 *we have for all $u \in X \cap [H^{r+1}(\Omega)]^{d+1}$*

$$S_h(j_h u, j_h u)^{1/2} \lesssim (\delta^{1/2} + \alpha^{1/2}) h^r (\|v\|_{r+1} + \|p\|_{r+1}).$$

Proof. We start with adding and subtracting u:

$$\begin{aligned} S_h(j_h u, j_h u) &= S_h(u + j_h u - u, u + j_h u - u) \\ &\le S_h(u,u) + S_h(j_h u - u, j_h u - u) + 2 S_h(j_h u - u, u) \\ &\le 2(S_h(u,u) + S_h(j_h u - u, j_h u - u)). \end{aligned}$$

For the first term the result follows immediately by the interpolation property (4.10):

$$\begin{aligned} S_h(u,u) &\le \delta \|\bar{\varkappa}_h \nabla v\|^2 + \alpha \|\bar{\varkappa}_h \nabla p\|^2 \\ &\lesssim \delta h^{2r} \|v\|_{r+1}^2 + \alpha h^{2r} \|p\|_{r+1}^2. \end{aligned}$$

For the second term $S_h(j_h u - u, j_h u - u)$ we have

$$\begin{aligned} \delta \|\bar{\varkappa}_h \nabla (j_h v - v)\|^2 &\lesssim \delta \|\nabla (j_h v - v)\|^2 \\ &\lesssim \delta h^{2r} \|v\|_{r+1}^2, \end{aligned}$$

using the L^2-stability (4.11) of the local projector $\bar{\varkappa}_h$ and the interpolation property (4.4) of j_h. For the pressure contribution of course the same holds. □

4.2. A priori estimate for smooth velocities and pressure. In this subsection, we prove the following a priori estimate for the discrete solution of (3.1).

THEOREM 4.8. *If the solution $u = \{v, p\}$ of* (2.2) *satisfies $u \in [H^{r+1}(\Omega)]^{d+1}$, then we have the a priori estimate*

$$\|\!|u - u_h|\!\| \lesssim a h^{r+\frac{1}{2}} \left(\|v\|_{r+1} + \|p\|_{r+1}\right) \tag{4.12}$$

with

$$a = h^{-1/2}(\mu^{1/2} + \delta^{1/2} + \alpha^{1/2}) + h^{1/2}(\sigma^{1/2} + \delta^{-1/2} + \alpha^{-1/2}). \tag{4.13}$$

Before proceeding with the proof of this theorem, let us briefly comment on its interpretation. An immediate consequence of the inequality (4.12) is that for convection dominated flow, $\delta \sim h$ and $\alpha \sim h$ are the optimal choice of the parameters, yielding an h-independent constant a and the (optimal) convergence order of $h^{r+1/2}$. The positive powers of δ and α in (4.13) represent the dissipative character of the stabilization terms. It follows that too much dissipation will have a negative effect on the precision. The presence of $\delta^{-1/2}$ and $\alpha^{-1/2}$ in (4.13) is due to the stabilizing effect of the subgrid model: The dissipation of the small-scale energy into the unresolved scales avoids artificial energy concentrations on the small scales due to the conservation properties of the Galerkin method. As expected, precision deteriorates for small values of δ and α due to spurious oscillations.

Proof. In the standard fashion we decompose the error in $u - u_h = \eta + \xi$ into an interpolation part $\eta = u - j_h u$ and a projection part $\xi = j_h u - u_h$. Clearly, $\|\!|\eta|\!\| \leq C a h^{r+1/2}$ using the interpolation Lemma 4.5 and the asymptotic bound for the stabilization term of Lemma 4.7. Consider now the discrete error ξ. By coercivity (Lemma 4.1) and the Galerkin orthogonality property (Lemma 4.2) we have

$$\begin{aligned} \|\!|\xi|\!\|^2 &= A(\xi, \xi) + S_h(\xi, \xi) \\ &= A(\eta, \xi) + S_h(j_h u, \xi). \end{aligned}$$

The second term on the right-hand side is bounded by applying the Cauchy–Schwarz inequality followed by Lemma 4.7:

$$\begin{aligned} S_h(j_h u, \xi) &\leq S_h(j_h u, j_h u)^{1/2} S_h(\xi, \xi)^{1/2} \\ &\lesssim (\mu^{1/2} + \alpha^{1/2}) h^r (\|v\|_{r+1} + \|p\|_{r+1}) \|\!|\xi|\!\|. \end{aligned}$$

For the first term on the right-hand side we use the Cauchy–Schwarz inequality and integration by parts, writing ξ^p and ξ^v for the discrete pressure and the velocity error, respectively:

$$A(\eta, \xi) \leq \|\!|\eta|\!\| \, \|\!|\xi|\!\| - (p - j_h p, \operatorname{div} \xi^v) - (v - j_h v, \nabla \xi^p) - (\beta \otimes (v - j_h v), \nabla \xi^v). \tag{4.14}$$

We now use the orthogonality property of the quasi-interpolation operator to obtain upper bounds:

$$\begin{aligned} |(p - j_h p, \operatorname{div} \xi^v)| &= |(p - j_h p, \operatorname{div} \xi^v - \overline{\pi} \operatorname{div} \xi^v)| \\ &\leq \|\delta^{-1/2}(p - j_h p)\| \, \|\delta^{1/2}(\operatorname{div} \xi^v - \overline{\pi} \operatorname{div} \xi^v)\| \\ &\leq \|\delta^{-1/2}(p - j_h p)\| \, S_h(\xi, \xi)^{1/2}, \end{aligned}$$

$$\begin{aligned} |(v - j_h v, \nabla \xi^p)| &= (v - j_h v, \nabla \xi^p - \overline{\pi} \nabla \xi^p) \\ &\leq \|\alpha^{-1/2}(v - j_h v)\| \, \|\alpha^{1/2}(\nabla \xi^p - \overline{\pi} \nabla \xi^p)\| \\ &\leq \|\alpha^{-1/2}(v - j_h v)\| \, S_h(\xi, \xi)^{1/2}, \end{aligned}$$

$$\begin{aligned} |(\beta \otimes (v - j_h v), \nabla \xi^v)| &= |(v - j_h v, (\beta \cdot \nabla) \xi^v - \overline{\pi} (\beta \cdot \nabla) \xi^v)| \\ &\leq \|\delta^{-1/2}(v - j_h v)\| \, S_h(\xi, \xi)^{1/2}. \end{aligned}$$

In summary, we get

$$\begin{aligned}A(\eta,\xi) &\le |\!|\!|\eta|\!|\!|\,|\!|\!|\xi|\!|\!| + (\|\delta^{-1/2}(p-j_hp)\| + \|(\alpha^{-1/2}+\delta^{-1/2})(v-j_hv)\|)S_h(\xi,\xi)^{1/2}\\ &\le \big(|\!|\!|\eta|\!|\!| + \|\delta^{-1/2}(p-j_hp)\| + \|(\alpha^{-1/2}+\delta^{-1/2})(v-j_hv)\|\big)\,|\!|\!|\xi|\!|\!|.\end{aligned}$$

The assertion follows using interpolation properties (4.3) and (4.4) of the quasi interpolant j_h. □

We proceed and prove that the pressure also has optimal convergence properties in the L^2-norm.

LEMMA 4.9. *Let $u = \{v, p\}$ be the solution of* (2.2) *and let $u_h = \{v_h, p_h\}$ be the solution of* (3.1)*. Then there holds*

$$\|p - p_h\| \lesssim a|\!|\!|u - u_h|\!|\!|,$$

where $a = \sigma^{1/2} + |\beta|\sigma^{-1/2} + \mu^{1/2} + \delta^{1/2} + \alpha^{-1/2}h$.

Proof. Following [12], by the surjectivity of the divergence operator there exists $v_p \in [H_0^1(\Omega)]^d$ such that $p - p_h = \operatorname{div} v_p$ and $\|v_p\|_{1,\Omega} \lesssim \|p - p_h\|$. By the H^1-stability property of the quasi interpolant j_h we then have

$$\|j_hv_p\|_{1,\Omega} \lesssim \|p - p_h\|. \tag{4.15}$$

Consider now the equality $p - p_h = \operatorname{div} v_p$. This gives

$$\|p - p_h\|^2 = (p - p_h, \operatorname{div} v_p).$$

We now subtract j_hv_p from v_p in the right-hand side and use the Galerkin orthogonality property in Lemma 4.2 for the test function $\{j_hv_p, 0\}$:

$$\begin{aligned}\|p-p_h\|^2 = (p-p_h, \operatorname{div}(v_p - j_hv_p)) - (\mu\nabla(v-v_h), \nabla j_hv_p)\\ +(\sigma(v-v_h), j_hv_p) + (\beta\otimes(v-v_h), \nabla(j_hv_p)) - S_h(u-u_h, \{j_hv_p, 0\}).\end{aligned}$$

We estimate the resulting parts separately. For the first term we integrate by parts and use the orthogonality property (4.2) of the quasi-interpolation operator j_h to obtain

$$\begin{aligned}(p-p_h, \operatorname{div}(v_p - j_hv_p)) &= (\nabla(p-p_h), v_p - j_hv_p)\\ &= (\bar{\varkappa}_h\nabla(p-p_h), v_p - j_hv_p)\\ &\le S_h(\{0, p-p_h\}, \{0, p-p_h\})^{1/2}\|\alpha^{-1/2}(v_p - j_hv_p)\|\\ &\lesssim \alpha^{-1/2}h|\!|\!|u-u_h|\!|\!|\|v_p\|_1\\ &\lesssim \alpha^{-1/2}h|\!|\!|u-u_h|\!|\!|\,\|p-p_h\|,\end{aligned}$$

where we used the stability property of v_p in the last inequality. Furthermore, we have

$$\begin{aligned}&(\sigma(v-v_h), j_hv_p) + (\beta\otimes(v-v_h), \nabla(j_hv_p))\\ &= (\sigma(v-v_h), j_hv_p) - (v-v_h, (\beta\cdot\nabla)j_hv_p)\\ &\le (\sigma^{1/2} + |\beta|\sigma^{-1/2})|\!|\!|u-u_h|\!|\!|\,\|j_hv_p\|_1.\end{aligned}$$

Similarly we obtain, after application of the Cauchy–Schwarz inequality and (4.11),

$$\begin{aligned}&(\mu\nabla(v-v_h), \nabla j_hv_p) - S_h(u-u_h, \{j_hv_p, 0\})\\ &\le \|\mu^{1/2}\nabla(v-v_h)\|\|\mu^{1/2}\nabla j_hv_p\| + S_h(u-u_h, u-u_h)^{1/2}S_h(\{j_hv_p, 0\}, \{j_hv_p, 0\})^{1/2}\\ &\le |\!|\!|u-u_h|\!|\!|\,(\mu^{1/2}+\delta^{1/2})\|j_hv_p\|_1.\end{aligned}$$

Collecting terms and using (4.15) gives the assertion. □

COROLLARY 4.10. *For the solution of* (3.1) *there holds*

$$\|p_h\|^2 \lesssim \vert\kern-0.25ex\vert\kern-0.25ex\vert u_h \vert\kern-0.25ex\vert\kern-0.25ex\vert^2 + \|f\|^2 = (f, v_h) + \|f\|^2.$$

Hence the pressure is unique.

Proof. Modifying the proof of Lemma 4.9 considering not $p - p_h$ but simply p_h and introducing the right-hand side instead of using Galerkin orthogonality, we have

$$\|p_h\| \lesssim a \vert\kern-0.25ex\vert\kern-0.25ex\vert u_h \vert\kern-0.25ex\vert\kern-0.25ex\vert + \|f\|$$

and conclude by applying Lemma 4.1. □

4.3. Lower regularities. The aim of the smoothness assumptions above is to show that the discretization allows for the quasi-optimal a priori error estimates that are characteristic for stabilized methods. However, for the case of high Reynolds number flows this may seem overly optimistic, and we will therefore discuss what we may prove rigorously in the case where the pressure is only in $H^1(\Omega)$ and the velocities are in $[H^2(\Omega)]^d$. In order to recover optimality of the estimate, the pressure stabilization has to be reduced to $\alpha \sim h^2$. Otherwise, the error would be dominated by the term $\alpha^{1/2}\|p\|_1$, which would lead to a convergence order of only $h^{1/2}$.

LEMMA 4.11. *In the case of less regular velocities and pressure, $v \in [H^2(\Omega)]^d$ and $p \in H^1(\Omega)$, and the use of the stabilization*

$$S_h(u, \varphi) = (\delta \nabla \varkappa_h v, \nabla \varkappa_h \psi) + (\alpha \nabla \varkappa_h p, \nabla \varkappa_h \xi) + (\bar{\varkappa}_h \, div \, v_h, \bar{\varkappa}_h \, div \, \psi),$$

with $\delta \sim h$ and $\alpha \sim h^2$, it holds for $v \in [H^2(\Omega)]^d$, $p \in H^1(\Omega)$, and high local Peclet number that

$$\vert\kern-0.25ex\vert\kern-0.25ex\vert u - u_h \vert\kern-0.25ex\vert\kern-0.25ex\vert \lesssim h^{3/2}\|v\|_2 + h\|p\|_1.$$

Furthermore, for lower regularity $v \in [H^{1+\epsilon}(\Omega)]^d$, $p \in L^2(\Omega)$ with $\epsilon > 0$ and under the assumption

$$\|v - j_h v\| \lesssim h^{1+\epsilon}\|v\|_{1+\epsilon}, \tag{4.16}$$

we have at least convergence $\vert\kern-0.25ex\vert\kern-0.25ex\vert u - u_h \vert\kern-0.25ex\vert\kern-0.25ex\vert \to 0$ for $h \to 0$.

Proof. We begin with the case $v \in [H^2(\Omega)]^d$, $p \in H^1(\Omega)$. The regularity of the pressure is necessary only for the upper bound of the stabilizing term of Lemma 4.7. The lower regularity gives the modified upper bound

$$S_h(j_h u, j_h u)^{1/2} \lesssim \delta^{1/2} h \|v\|_2 + \alpha^{1/2}\|p\|_1.$$

Due to the decrease of the pressure stabilization to $\alpha \sim h^2$ it follows from the proof of Theorem 4.8 that the control of the incompressibility condition has to be increased. This is warranted due to the additional stabilization term

$$\|\bar{\varkappa}_h \operatorname{div} v_h\| \le S_h(u_h, u_h). \tag{4.17}$$

Hence, we deduce

$$\begin{aligned} |(p - j_h p, \operatorname{div} \xi^v)| &= |(p - j_h p, \operatorname{div} \xi^v - \overline{\pi} \operatorname{div} \xi^v)| \\ &\le \|p - j_h p\| \, S_h(u_h, u_h). \end{aligned}$$

For the case $v \in H^{1+\epsilon}(\Omega)$ and $p \in L^2(\Omega)$, of course we cannot expect to get any convergence order. On the other hand, the choice $\alpha \sim h^2$ and (4.17) allow us to prove convergence by a density argument, provided that the interpolants converge. The Scott–Zhang operator is no longer well defined on the space of L^2-functions, and we therefore replace it by the L^2-projection. Assuming quasi-uniform meshes, the same estimate holds. First we check the stabilization operator $S(j_h u, j_h u)$. Using the stability of the projection $\bar{\varkappa}_h$, the L^2-projection onto a piecewise constant on element K, $\pi_{0,K}$, and an inverse inequality, we deduce that

$$\begin{aligned} S(j_h u, j_h u) &\lesssim \sum_K \left(\delta_K \|\nabla j_h v\|_K^2 + \alpha_K \|\nabla j_h p\|_K^2\right) \\ &\lesssim \sum_K \left(h_K \|\nabla v\|_K^2 + \|j_h p - \pi_{0,K} p\|_K^2\right) \\ &\lesssim \sum_K \left(h_K \|\nabla v\|_K^2 + \|p - j_h p\|_K^2 + \|p - \pi_{0,K} p\|_K^2\right) \\ &\to 0 \quad (\text{for } h \to 0). \end{aligned}$$

Convergence of the other terms are achieved in a similar fashion assuming (4.16). □

In the case of low local Reynolds number, i.e., $|\beta| h < \mu$, and if $\{v, p\} \in [H^2(\Omega)]^d \times H^1(\Omega)$, one easily shows that the choice $\delta = 0$ and $\alpha \sim h^2$ leads to optimal a priori error estimates in the energy norm by Theorem 4.8. An error estimate in the L^2-norm for the velocities may then be recovered using a standard Nitsche duality argument.

5. Variants of local projection stabilization.

5.1. Local projection in streamline direction. It should also be noted that from the practical viewpoint it may be more advantageous to use the streamline derivative in the part of the subgrid model acting on the velocity in order to minimize crosswind diffusion; for instance,

$$\begin{aligned} (5.1) \quad S_h^\beta(u, \varphi) := {} & (\delta \bar{\varkappa}_h (\beta \cdot \nabla) v, \bar{\varkappa}_h (\beta \cdot \nabla) \psi) + (\delta \bar{\varkappa}_h \operatorname{div} v, \bar{\varkappa}_h \operatorname{div} \psi) \\ & + (\alpha \bar{\varkappa}_h \nabla p, \bar{\varkappa}_h \nabla \xi). \end{aligned}$$

The following lemma states the fact that the proposed stabilization term (5.1) involving only diffusion in streamline direction can be bounded by the triple norm. As a consequence, taking (5.1) does not affect the order of the numerical scheme.

LEMMA 5.1. *If $\beta \in [W^{1,\infty}(\Omega)]^d$, then it holds for all $v \in V_h$ that*

$$(5.2) \qquad \|\delta^{1/2} \bar{\varkappa}_h (\beta \cdot \nabla) v\| \le C_\beta |||\{v, 0\}|||,$$

where $C_\beta \sim \delta^{1/2} \|\beta\|_{W^{1,\infty}(\Omega)} \sigma^{-1/2} + \|\beta\|_\infty$.

Proof. The proof follows by adding and subtracting $\bar{\pi}_h \beta$, where $\bar{\pi}_h$ denotes the projection on D_h^v (here denoting the space of piecewise constants on the macropatches, regardless of the approximation). We apply the triangle inequality and the H^1 stability of $\bar{\varkappa}_h$:

$$\begin{aligned} \|\delta^{1/2} \bar{\varkappa}_h (\beta \cdot \nabla) v\| &\le \|\delta^{1/2} \bar{\varkappa}_h ((\beta - \bar{\pi}_h \beta) \cdot \nabla) v\| + \|\delta^{1/2} \bar{\varkappa}_h ((\bar{\pi}_h \beta) \nabla) v\| \\ &\lesssim \|\delta^{1/2} ((\beta - \bar{\pi}_h \beta) \cdot \nabla) v\| + \|\delta^{1/2} \bar{\varkappa}_h ((\bar{\pi}_h \beta) \nabla) v\|. \end{aligned}$$

The second term on the right-hand side is simply bounded by

$$\begin{aligned} \|\delta^{1/2} \bar{\varkappa}_h ((\bar{\pi}_h \beta) \nabla) v\| &\le \|\beta\|_\infty \|\delta^{1/2} \bar{\varkappa}_h \nabla v\| \\ &\le \|\beta\|_\infty S_h(u, u)^{1/2}. \end{aligned}$$

The first term can be estimated by the approximation property of $\bar{\pi}_h$ and a local inverse inequality:

$$\begin{aligned}\|((\beta - \bar{\pi}_h\beta)\cdot\nabla)v\| &\lesssim \sum_{K\in\mathcal{T}_{2h}} \|\bar{\varkappa}\beta\|_{K,\infty} h_K^{-1}\|v\|_K \\ &\leq \|\beta\|_{W^{1,\infty}(\Omega)}\|v\|.\end{aligned}$$

This gives

$$\begin{aligned}\|\delta^{1/2}\bar{\varkappa}_h(\beta\cdot\nabla)v\| &\lesssim \|\beta\|_\infty S_h(u,u)^{1/2} + \delta^{1/2}\|\beta\|_{W^{1,\infty}(\Omega)}\|v\| \\ &\leq C_\beta |\!|\!|\{v,0\}|\!|\!| \qquad \square\end{aligned}$$

Note that this result is valid immediately (without any assumptions on β) if the form (5.1) is used in the definition of the triple norm.

5.2. Projection onto a coarser mesh. As a further alternative we may use the nodal interpolant $\pi_h : Q_h \to Q_{2h}$ and take as fluctuation filter $\varkappa_h := i - \pi_h$. The stabilization term for the Oseen system can now be taken as

$$S_h(u,\varphi) := (\delta\nabla\varkappa_h v, \nabla\varkappa_h\psi) + (\alpha\nabla\varkappa_h p, \nabla\varkappa_h\xi). \tag{5.3}$$

When the triple norm $|\!|\!|\cdot|\!|\!|$ is designed with the term $S_h(\cdot,\cdot)$, the coercivity property of Lemma 4.1 and the perturbed Galerkin orthogonality of Lemma 4.2 still hold for this variant. Also the estimate in Lemma 4.7 is still valid if we assume that $u \in H^2(\Omega)^{d+1}$ holds in order to be able to apply the nodal interpolant on u.

This variant can be considered as a generalization of the concept for advection equations of Guermond [11, 14].

5.3. Relation to classical stabilized methods. In this section, we will show the relation between the local projection method analyzed in this paper and the GLS method or the residual-free bubble method. A key feature of the proposed method is the weak consistency: The fact that the stabilization enjoys the right asymptotic behavior without strong consistency allows us to decouple the stabilization of the pressure and the velocities and, even more importantly, allows us to decouple the stabilization from time-stepping terms and source terms. However, to show the relation to the GLS we will reintroduce the strong consistency. Our aim is to show that by using the local projection stabilization we may in fact use GLS on the fine scales only, whereas the coarse scales are stable thanks to the interaction between coarse and fine scales. To this end we consider the full residual

$$\rho(u) := \sigma v + \operatorname{div}(\beta\otimes v) - \mu\Delta v + \nabla p$$

in the stabilization

$$S_{gls}(u_h,\varphi) := (\delta\bar{\varkappa}_h\rho(u_h), \bar{\varkappa}_h\rho(\varphi))_h + (\delta\bar{\varkappa}_h \operatorname{div} v_h, \bar{\varkappa}_h \operatorname{div}\psi), \tag{5.4}$$

where $(\cdot,\cdot)_h := \sum_K (\cdot,\cdot)_K$. To make the formulation strongly consistent we perturb the right-hand side and obtain

$$A(u_h,\varphi) + S_{gls}(u_h,\varphi) = (f, \psi + \delta\bar{\varkappa}_h\rho(\varphi))_h \quad \forall\varphi\in X_h. \tag{5.5}$$

The consistency follows, because for the exact solution u we have

$$S_{gls}(u,\varphi) - (f,\delta\bar{\varkappa}_h\rho(\varphi))_h = (\delta\bar{\varkappa}_h(\rho(u)-f), \bar{\varkappa}_h\rho(\varphi))_h = 0.$$

We have thus reformulated the local projection method as a GLS formulation with the stabilization acting only as a filter on the small scales. An important difference, however, is that the local projection approach using (5.4) does not impose any artificial boundary conditions on the solution in contrast to the case of residual-free bubbles. This should be a definite advantage for nonlinear problems. We take as triple norm

$$|\!|\!|u|\!|\!|_{gls} := \left(\|\sigma^{1/2}v\|^2 + \|\mu^{1/2}\nabla v\|^2 + \|\delta^{\frac{1}{2}}\bar{\varkappa}_h(\beta\cdot\nabla v + \nabla p)\|^2\right)^{1/2}.$$

Note that we still have coercivity. In fact, after minor modifications, Theorem 4.8 remains true for (5.4).

Lemma 5.2. *It holds that*

$$|\!|\!|u - u_h|\!|\!|_{gls} \lesssim h^{r+\frac{1}{2}}\left(\|v\|_{r+1} + \|p\|_{r+1}\right)$$

if $\delta \lesssim \min(\frac{h}{\mu}, \frac{1}{\sigma})$ *with a constant depending on the constant in the* L^2*-stability of the projection* $\bar{\varkappa}_h$ *and the constant in the inverse inequality.*

Proof. We first note that

$$\begin{aligned} S_{gls}(u - j_h u, u - j_h u) = {} & (\delta\bar{\varkappa}_h\rho(u - j_h u_h), \bar{\varkappa}_h\rho(u - j_h u))_h \\ & + (\delta\bar{\varkappa}_h \operatorname{div}(v - j_h v_h), \bar{\varkappa}_h \operatorname{div}(v - j_h v)) \end{aligned}$$

has the right asymptotic, which is immediate, assuming optimal approximation for the second derivatives. One may then show, using the stability of the local projection and standard inverse inequalities, that provided δ satisfies the upper bound in the supposition, there holds

$$\frac{1}{2}\|\delta^{\frac{1}{2}}\bar{\varkappa}_h(\beta\cdot\nabla v_h + \nabla p_h)\|^2 - \frac{1}{2}\|\sigma^{1/2}v_h\|^2 - \frac{1}{2}\|\mu\nabla v_h\|^2 \lesssim S_{gls}(u_h, u_h).$$

The proof now follows from (4.14) in the following fashion (considering here only the modified terms):

$$\begin{aligned} A(\eta,\xi) &\le |\!|\!|\eta|\!|\!|_{gls}\,|\!|\!|\xi|\!|\!|_{gls} - (v - j_h v, \beta\cdot\nabla\xi^v + \nabla\xi^p) \\ &\le |\!|\!|\eta|\!|\!|_{gls}\,|\!|\!|\xi|\!|\!|_{gls} - (\delta^{-1/2}(v - j_h v), \delta^{1/2}\bar{\varkappa}_h(\beta\cdot\nabla\xi^v + \nabla\xi^p)) \\ &\le \left(|\!|\!|\eta|\!|\!|_{gls} + \|\delta^{-1/2}(v - j_h v)\|\right)|\!|\!|\xi|\!|\!|_{gls}. \qquad \square \end{aligned}$$

5.4. Extension to triangular meshes. Until now, we have considered meshes $\mathcal{T}_h$ with quadrilateral (or hexahedral) elements. The corresponding finite elements are d-linear ($r = 1$) or d-quadratic ($r = 2$). This raises the question of whether the described method is applicable also on elements with triangles ($d = 2$) or tetrahedrons ($d = 3$). Of course, the definition of the method carries over to those triangulations without any modification if the patches are defined properly. It has to be assured that for a patch $K \in \mathcal{T}_h$, test functions with support inside K do exist.

Bisection on triangles creates four smaller triangles out of one triangle; see Figure 5.1(a). Since no inner points are created, this strategy would not work in the case $r = 1$: The spaces $V_h(K)$ used in Lemma 4.3 would be empty in this particular case. However, possible patches are sketched in Figure 5.1(b) and (c). In the case $r = 2$ and $d = 2$, bisection leads to spaces $V_h(K)$ of dimension three corresponding to the three internal edges inside the patch; see Figure 5.1(a). The space $D_h(K)$ has the same dimension and can be represented by test functions corresponding to the three nodes of K. Hence, we have $\dim D_h(K) \le \dim V_h(K)$, and the mapping π_K can be defined to be injective. A similar situation occurs for $d = 3$ and $r = 2$: $\dim V_h(K) = 6$ and $\dim D_h(K) = 4$.

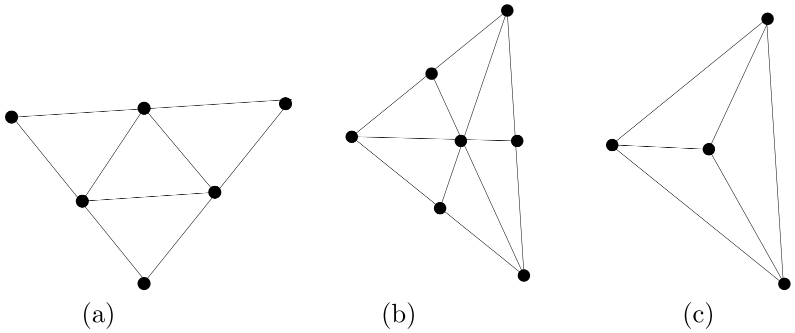


FIG. 5.1. *Possible patches of triangles:* (a) *bisection does not create inner nodes and is therefore not suitable for local projection stabilization in the case $r = 1$;* (b) *and* (c) *create inner nodes.*

6. Link to the variational multiscale method.

Today one of the major challenges in computational fluid dynamics is the accurate computation of different quantities in turbulent flow. Recently, several new approaches have been proposed such as the dynamic multilevel methodology (DML) of Dubois, Jauberteau, and Temam [10] or the VMS method of Hughes, Mazzei, and Jansen [15]. In the latter work, reference is made to residual-free bubble techniques (see Brezzi and Russo [4]) and subgrid viscosity as introduced by Guermond [14] to motivate an approach to large eddy simulation (LES), where the turbulence model acts only on the fine scales. In the following section we will show how the local projection method may be cast in the VMS framework, leading to a stabilized finite element method suitable for high Reynolds number flows.

6.1. Variational formulation of the Navier–Stokes equations.

Let $I := [0, T]$ be the time interval. The velocities are sought in the Bochner space $\mathcal{V}^v := H^1(I, V)$, and the pressure in $\mathcal{V}^p := L^2(I, Q)$. The product space will be denoted by $\mathcal{V} := \mathcal{V}^v \times \mathcal{V}^p$. The test functions are in the space $\mathcal{W} := L^2(I, X)$. The L^2-scalar product over the space-time slab $\Omega_T := \Omega \times I$ will be denoted by $(\cdot, \cdot)_{\Omega_T}$, and its norm by $\|\cdot\|_{\Omega_T}$. Introducing now the state vector $u = \{v, p\} \in \mathcal{V}$, we may write the standard variational formulation of the Navier–Stokes equations (2.1): Find $u \in \mathcal{V}$ such that $v(\cdot, 0) = v_0$ and

$$B(u, \varphi) = \langle f, \varphi \rangle \quad \forall \varphi \in \mathcal{W}, \tag{6.1}$$

where $B(u, \varphi)$ is defined for $\varphi = \{\psi, \xi\}$ by

$$B(u, \varphi) := (\partial_t v, \psi)_{\Omega_T} - (v \otimes v, \nabla \psi)_{\Omega_T} + (\mu \nabla v, \nabla \psi)_{\Omega_T} - (p, \operatorname{div} \psi)_{\Omega_T} + (\operatorname{div} v, \xi)_{\Omega_T}.$$

6.2. Separation of scales on the continuous level.

In the VMS formulation as introduced in [15], a scale separation is performed and the turbulence model acts only on the finer scales. However, as always in turbulence modeling certain model assumptions on the interaction between the scales are made.

To clarify our model assumptions, we use the three-level partition proposed in Collis [9]. Hence we consider a scale separation in large resolved scales denoted by $\bar{v}$, small resolved scales denoted by $\tilde{v}$, and unresolved scales denoted by $\hat{v}$. The solution space is partitioned in a corresponding manner:

$$\mathcal{V} = \bar{\mathcal{V}} \oplus \tilde{\mathcal{V}} \oplus \hat{\mathcal{V}}.$$

The function space $\mathcal{W}$ is partitioned similarly, $\mathcal{W} = \bar{\mathcal{W}} \oplus \tilde{\mathcal{W}} \oplus \hat{\mathcal{W}}$, with corresponding test functions, for instance, $\bar{\varphi} = \{\bar{\psi}, \bar{\xi}\} \in \bar{\mathcal{W}}$. We now write the exact equations of motions for each scale:

$$B(u, \bar{\varphi}) = \langle f, \bar{\varphi}\rangle \quad \forall \bar{\varphi} \in \bar{\mathcal{W}}, \tag{6.2}$$

$$B(u, \tilde{\varphi}) = \langle f, \tilde{\varphi}\rangle \quad \forall \tilde{\varphi} \in \tilde{\mathcal{W}}, \tag{6.3}$$

$$B(u, \hat{\varphi}) = \langle f, \hat{\varphi}\rangle \quad \forall \hat{\varphi} \in \hat{\mathcal{W}}. \tag{6.4}$$

Introducing the linearized Navier–Stokes operator

$$\begin{aligned} B'(u, u', \varphi) := {} & (\partial_t v', \hat{\psi})_{\Omega_T} - (v' \otimes v + v \otimes v', \nabla\psi)_{\Omega_T} \\ & -(p', \nabla \cdot \psi)_{\Omega_T} + (\mu\nabla v', \nabla\psi)_{\Omega_T} + (\nabla \cdot v', \xi)_{\Omega_T}, \end{aligned}$$

the Reynolds stress projection

$$R(v, \psi) := (v \otimes v, \nabla\psi)_{\Omega_T},$$

and the cross-stress projection operator

$$C(v, \hat{v}, \psi) := (v \otimes \hat{v} + \hat{v} \otimes v, \nabla\psi)_{\Omega_T},$$

we may reformulate the exact equations for each scale in a fashion that makes evident the coupling between the scales. Following Collis [9], the exact solution $\bar{v} \in \bar{\mathcal{V}}$ for the resolved large scales fulfills for all $\bar{\varphi} \in \bar{\mathcal{W}}$ the equation

$$\begin{aligned} B(\bar{u}, \bar{\varphi}) + B'(\bar{u}, \tilde{u}, \bar{\varphi}) - R(\tilde{v}, \bar{\psi}) = {} & \langle f, \bar{\varphi}\rangle \\ & -B'(\bar{u}, \hat{u}, \bar{\varphi}) - R(\hat{v}, \bar{\psi}) + C(\tilde{v}, \hat{v}, \bar{\psi}). \end{aligned} \tag{6.5}$$

The first line in (6.5) includes the influence of the resolved scales on the large scales, whereas the second line includes the influence of the unresolved scales on the large scales. In the same fashion, the small resolved scales $\tilde{v} \in \tilde{\mathcal{V}}$ fulfill for all $\tilde{\varphi} \in \tilde{\mathcal{W}}$

$$\begin{aligned} B'(\bar{u}, \tilde{u}, \tilde{\varphi}) - R(\tilde{v}, \tilde{\psi}) = {} & \langle f, \tilde{\varphi}\rangle - B(\bar{u}, \tilde{\varphi}) \\ & -B'(\bar{u}, \hat{u}, \tilde{\varphi}) - R(\hat{v}, \tilde{\psi}) + C(\tilde{v}, \hat{v}, \tilde{\psi}). \end{aligned} \tag{6.6}$$

The unresolved scales $\hat{v} \in \hat{\mathcal{V}}$ finally satisfy the following equation for all $\hat{\varphi} \in \hat{\mathcal{W}}$

$$B'(\bar{u} + \tilde{u}, \hat{u}, \hat{\varphi}) + R(\hat{v}, \hat{\psi}) = \langle f, \hat{\varphi}\rangle - B(\bar{u} + \tilde{u}, \hat{\varphi}).$$

It follows that the equation for the unresolved scales is driven by the residual of the resolved scales. With the equations written in this form it is easy to state the modeling assumptions as follows:

(M1) The unresolved scales $\hat{v}$ have no "direct" influence on the large scales. This means that the second line of (6.5) is set to zero:

$$-B'(\bar{u}, \hat{u}, \bar{\varphi}) - R(\hat{v}, \bar{\psi}) + C(\tilde{v}, \hat{v}, \bar{\psi}) = 0 \quad \forall \bar{\varphi} \in \bar{\mathcal{W}}. \tag{6.7}$$

(M2) The influence of the unresolved scales on the small scales is modeled by an artificial viscosity term

$$S : \mathcal{X} \times \mathcal{X} \to \mathbb{R},$$

with $\mathcal{X} := (\bar{\mathcal{V}} \oplus \tilde{\mathcal{V}}) \cup (\bar{\mathcal{W}} \oplus \tilde{\mathcal{W}})$, acting only on the small resolved scales. Hence we assume in (6.6) that for $\tilde{\varphi} \in \tilde{\mathcal{W}}$

$$S(\tilde{u}, \tilde{\varphi}) \approx B'(\bar{u}, \hat{u}, \tilde{\varphi}) + R(\hat{v}, \tilde{\psi}) - C(\tilde{v}, \hat{v}, \tilde{\psi}). \tag{6.8}$$

The first modeling assumption (M1) can be expected to hold true when the main features of the flow are resolved. This is the large eddy assumption. The second modeling assumption (M2) implies that the unresolved scales only have the effect of dissipating energy from the small resolved scales. Heuristically one may argue that if assumption (M1) is satisfied, then the exact form or size of the subgrid model is of less importance *as long as it allows for a sufficient rate of dissipation of energy from the resolved small scales to the unresolved scales.* Insufficient dissipation may cause buildup of energy in high frequency modes (by the conservation properties of the Galerkin method) leading to spurious oscillations. Excessive dissipation will cause too much damping of the resolved small scales leading to poorer resolution of the large scales through the Reynolds stress coupling.

Using these modeling assumptions and the L^2-projection Πv_0 of the initial conditions onto the resolved scales $\bar{\mathcal{V}}^v \oplus \tilde{\mathcal{V}}^v$ we arrive at the formulation $(\bar{v}+\tilde{v})(\cdot,0) = \Pi v_0$ and

$$\begin{aligned} B(\bar{u}+\tilde{u},\bar{\varphi}) &= \langle f,\bar{\varphi}\rangle \quad \forall\bar{\varphi}\in\bar{\mathcal{W}}, \\ B(\bar{u}+\tilde{u},\tilde{\varphi}) + S(\tilde{v},\tilde{\varphi}) &= \langle f,\tilde{\varphi}\rangle \quad \forall\tilde{\varphi}\in\tilde{\mathcal{W}}. \end{aligned} \tag{6.9}$$

We choose the subgrid viscosity term to be coercive on the small resolved scales $\tilde{u}$, i.e., $S(\tilde{u},\tilde{u}) \geq c\|\nabla\tilde{u}\|^2$ for all $\tilde{u}\in\tilde{\mathcal{W}}$, symmetric $S(u,\varphi)=S(\varphi,u)$ for all $u,\varphi\in\mathcal{X}$, and such that it vanishes on the large resolved scales

$$S(\cdot,\bar{\varphi}) = 0 \quad \forall\bar{\varphi}\in\bar{\mathcal{W}}\cup\bar{\mathcal{V}}. \tag{6.10}$$

6.3. Separation of scales on the discrete level. We introduce some finite element approximation $\mathcal{V}_h$ of $\mathcal{V}$ that will represent the resolved scales $\mathcal{V}_h = \bar{\mathcal{V}}\oplus\tilde{\mathcal{V}}$. This space is then decomposed into large and small resolved scales by choosing $\bar{\mathcal{V}} = \mathcal{V}_H$, where $\mathcal{V}_H \subset \mathcal{V}_h$. To indicate its dependence on h, we equip the subgrid viscosity with a subscript, $S_h(\cdot,\cdot)$. The same discrete space is used for the test space $\mathcal{W}_h = \bar{\mathcal{W}}\oplus\tilde{\mathcal{W}}$. The discrete version of (6.9) becomes the following: Find $u_h\in\mathcal{V}_h$ such that $v_h(\cdot,0) = \pi v_0$ and

$$B(u_h,\varphi) + S_h(\tilde{u}_h,\tilde{\varphi}) = \langle f,\varphi\rangle \quad \forall\varphi\in\mathcal{W}_h, \tag{6.11}$$

or, using the scale separation property (6.10) of $S_h(\cdot,\cdot)$,

$$B(v_h,\varphi) + S_h(u_h,\varphi) = \langle f,\varphi\rangle \quad \forall\varphi\in\mathcal{W}_h. \tag{6.12}$$

Note also that by the properties of $S_h(\cdot,\cdot)$ we have Galerkin orthogonality for the discretization error $u-u_h$ on the large resolved scales:

$$B(u-u_h,\bar{\varphi}) = 0 \quad \forall\bar{\varphi}\in\mathcal{W}_H. \tag{6.13}$$

Let us partition the time interval I into subintervals $I_n = (t_{n-1},t_n]$, $n=1,\ldots,N$, with $0=t_0<t_1<\cdots<t_N=T$ and $\tau_n := t_n - t_{n-1}$. We also introduce the space time slabs $Q_n := I_n\times\Omega$. As the time integration scheme, we use the Crank–Nicholson scheme. It means that we choose piecewise d-linears for the ansatz functions and as test spaces piecewise constants (discontinuous), precisely,

$$\mathcal{V}_h = P_\tau^1(I,X_h), \quad \mathcal{W}_h = P_\tau^0(I,X_h).$$

The spaces $\mathcal{V}_H$ and $\mathcal{W}_H$ are defined analogously by using X_H. With these finite element spaces we now propose the following finite element method: Find $u_h \in u_0 + \mathcal{V}_h$, so that in the nth time step it holds for the restriction $u^n = \{v^n, p^n\} := u_h|_{I_n}$:

$$A_n(u^n, \varphi) + S_h(u^n, \varphi) = g_n(u^{n-1}, \varphi) \quad \forall \varphi \in \mathcal{W}_h, \tag{6.14}$$

with

$$\begin{aligned} A_n(u,\varphi) &:= (\tau_n^{-1} v, \psi) - (v \otimes v, \nabla \psi) - (p, \operatorname{div} \psi) + (\operatorname{div} v, \xi) + (\mu \nabla v, \nabla \psi),\\ g_n(u,\varphi) &:= \langle f, \varphi \rangle + (\tau_n^{-1} v, \psi) - (\mu \nabla v, \nabla \psi) + (v \otimes v, \nabla \psi) - S_h(u, \varphi). \end{aligned}$$

As mentioned before, a widely used linearization of (6.14) is the Oseen linearization (2.3) with $\sigma := \tau_n^{-1}$ and β a suitable approximation on v^n (for instance, the last iterate in the nonlinear iteration).

With these notations, we may take as subgrid model (3.2) or (5.3) with parameters δ and α depending on h. If triangular or tetrahedral meshes are used, both subgrid operators satisfy (6.10) exactly and the analysis stated before shows that they are equivalent. On quadrilateral and hexahedral meshes, the stabilization (5.3) also satisfies (6.10) exactly. If version (3.2) is used on quadrilateral meshes, a small residual may remain due to the cross-term of Q_1 and Q_2 elements. Consequently, we do not have exact scale separation for (3.2) on quadrilaterals and hexahedrons.

7. Numerical example.

Finally, we show numerical examples of this stabilization strategy. As the first step, the convergence rates of $v, \nabla v$ and of p in L^2 are checked numerically on tensor grids and on locally refined meshes. In the next step, we investigate the difference of the kinetic energy of a nonstationary driven cavity flow with the numerical dissipation.

7.1. Convergence order for an exact Navier–Stokes solution.

In order to check that the theoretical proven convergence order is also obtained numerically, we consider a stationary Navier–Stokes problem with known exact smooth solution in the unit square $\Omega := (0,1)^2$:

$$\begin{aligned} v_x(x,y) &:= -\cos(x)\,\sin(y)\,(2\pi^2 + 1),\\ v_y(x,y) &:= \sin(x)\,\cos(y)\,(2\pi^2 - 1),\\ p(x,y) &:= 2(\cos(x) + \cos(y)). \end{aligned}$$

The right-hand side f is obtained by applying the Navier–Stokes operator to this solution. The solution is independent of the viscosity ν since the Laplacian applied to v vanishes; $\Delta v = 0$. However, this is not the case for the discrete solutions. The viscosity is set to $\nu = 10^{-6}$ so that this is smaller than the mesh size.

We investigate the convergence order of v in L^2 and in the H^1 seminorm, and for the pressure we check the L^2-error. Theorem 4.8 assures at least

$$\|\nu \nabla (v - v_h)\| \le |\!|\!| u - u_h |\!|\!| \;=\; \mathcal{O}(h^{r+1/2}).$$

Since $\nu < h$, we expect at least $\mathcal{O}(h^{r-1/2})$ for $\|\nabla(v - v_h)\|$. Due to Theorem 4.9, we get the same order of convergence for $\|p - p_h\|$. With a standard duality argument we obtain one order more for the L^2-error in the velocities. This expectation is summed up in Table 7.1 (second column with label "theoretically") for the case $r = 2$.

Let us first consider the case of equidistant tensor grids; see Figure 7.1 and the third column of Table 7.1. We clearly observe for this example the superconvergence behavior.

Table 7.1

Theoretical and practical convergence rates for different quantities. The practical convergence rates summarize the results of the numerical example.

	Theoretically	Practice Uniform meshes	Practice Nonuniform meshes
$\|\nabla(v-v_h)\|$	$h^{1.5}$	h^2	$h^{1.8}$
$\|v-v_h\|$	$h^{2.5}$	h^3	$h^{2.6}$
$\|p-p_h\|$	$h^{1.5}$	h^2	$h^{1.9}$

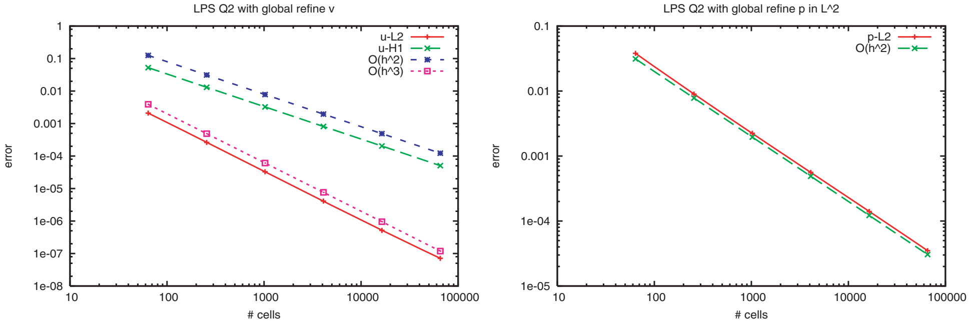


Fig. 7.1. *Convergence for stabilized biquadratic elements (Q_2) on equidistant tensor grids in dependence of the number of cells. Left: $\|v-v_h\| = \mathcal{O}(h^3)$ and $\|\nabla(v-v_h)\| = \mathcal{O}(h^2)$. Right: $\|p-p_h\| = \mathcal{O}(h^2)$.*

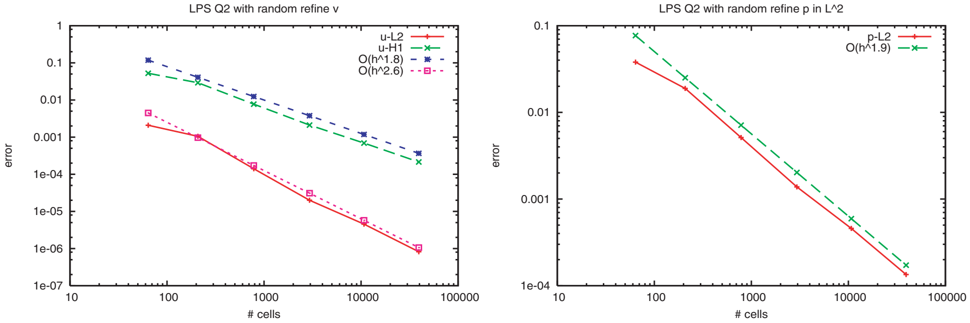


Fig. 7.2. *Convergence for stabilized biquadratic elements (Q_2) on randomly locally refined grids in dependence of the number of cells. Left: $\|v-v_h\| = \mathcal{O}(h^{2.6})$ and $\|\nabla(v-v_h)\| = \mathcal{O}(h^{1.8})$. Right: $\|p-p_h\| = \mathcal{O}(h^{1.9})$.*

On locally refined meshes superconvergence cannot be expected to such an extent. Therefore, we perform the same computations on a sequence of meshes which are obtained by refining approximately 50% of the cells by random. The meshes will be kept quasi-uniform with a ratio between the largest and the smallest mesh size bounded by 3. In order to compare with tensor grids we consider the "global mesh size" $h := n^{-1/2}$. In Figure 7.2 the convergence behavior is plotted on such locally refined meshes. The observed convergence rates are listed in the last column of Table 7.1. As expected, the convergence is reduced compared to tensor grids but is

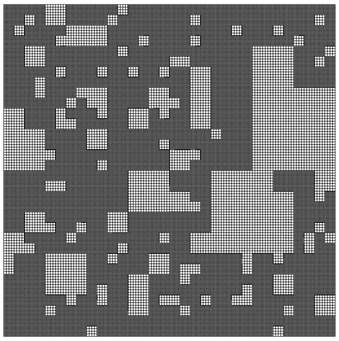


FIG. 7.3. *One of the locally refined meshes used in the numerical example. The refinement is performed by random.*

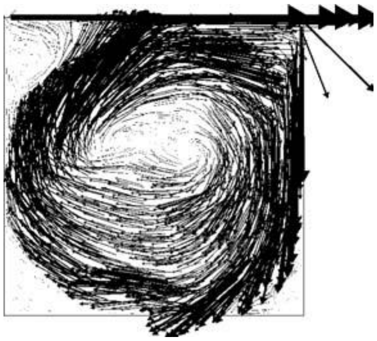


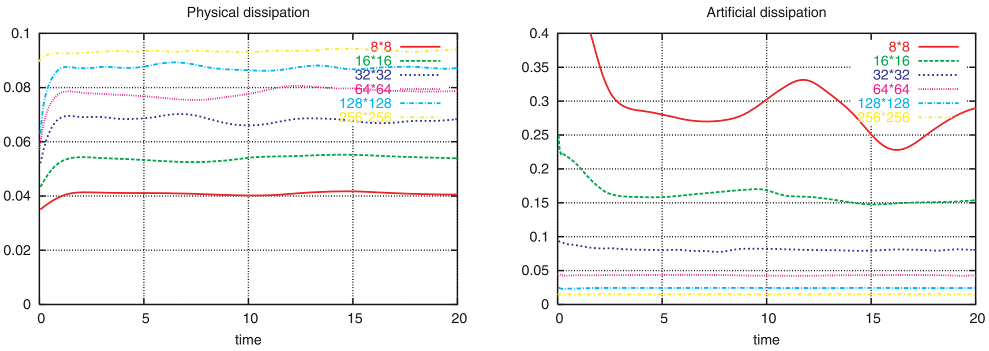
FIG. 7.4. *Flow field of the driven cavity problem at $Re = 10^4$ and $t = 20\,s$.*

still better than the theoretical results. Note that this is not a fault of the proof but is simply due to superconvergence on parts of the domain. We show one of the used meshes in Figure 7.3.

7.2. Driven cavity flow at $Re = 10^5$. The considered problem is a standard nonstationary driven cavity flow in the unit square $\Omega = (0,1)^2 \subset \mathbb{R}^2$. As boundary conditions, we have for the vertical velocity component homogeneous Dirichlet conditions $v_y = 0$ on $\partial\Omega$; for the horizontal velocity component we have $v_x = 1$ on $\partial\Omega \cap \{y = 1\}$ and $v_x = 0$ elsewhere.

Although the configuration has been investigated for many years, it is still not clear when the transition to nonsteady flow exactly occurs. Auteri, Parolini, and Quartapelle [1] found the first Hopf bifurcation at about $Re = 8018$ with a second-order spectral projection method on a mesh with 160^2 nodes. The computation of Peng, Shiau, and Hwang [17] shows that the transition to a periodic solution occurs at $Re = 7\,402 \pm 4$. They state that the flow becomes "chaotic" for $Re > 11\,000$. Further investigations can be found in [7, 13, 20], each with their own critical Reynolds number for the first Hopf bifurcation, but all in the range between 7 400 and 8 375. Hence, in order to guarantee a nonstationary flow we choose for our computations the Reynolds number $Re = 10^4$.

The initial solution $u|_{t=0}$ is chosen as the stationary solution at lower Reynolds number ($Re = 10^3$). The time step in the Crank–Nicholson scheme is chosen constant $\Delta t = 0.05\,s$. In Figure 7.4 the flow is shown at time instant $t = 20\,s$.


Physical dissipation

Artificial dissipation

Fig. 7.5. *Physical dissipation (left) and artificial dissipation (right) for the driven cavity problem at $Re = 10^4$.*

Table 7.2
Mean values of physical and artificial dissipation and their ratio for the sequence of meshes.

# Q_2 cells	8^2	16^2	32^2	64^2	128^2	256^2	512^2
$\overline{e}^h_{art}$	0.299	0.162	0.0810	0.04278	0.0243	0.0149	0.00937
$\overline{e}^h_{phy}$	0.0407	0.0536	0.0677	0.0782	0.0868	0.0934	0.0967
Ratio	7.34	3.02	1.20	0.55	0.28	0.16	0.097

We compare the physical dissipation $e^h_{phy}(t)$ and artificial dissipation $e^h_{art}(t)$, given by

$$\begin{aligned} e^h_{phy}(t) &:= \nu^{1/2}\|\nabla v_h(t)\|, \\ e^h_{art}(t) &:= S_h(u_h(t), u_h(t))^{1/2} \end{aligned}$$

on a sequence of equidistant tensor grids with 8^2 to 256^2 Q_2 cells (which corresponds to 3 151 875 DOFs). In Figure 7.5, these two quantities are plotted in time. On the coarsest mesh considered, the physical dissipation stabilizes slightly above 0.04. This quantity increases under mesh refinement, reaching nearly 0.1 on the mesh with 256^2 Q_2 cells. At the same time, the artificial dissipation part e^h_{art} decreases from about 0.3 to below of 0.01.

For such time-dependent flows, physical meaningful quantities are time averages, denoted by overbars. For example, the mean of physical dissipation will be denoted by

$$\overline{e}^h_{phy} := \frac{1}{T}\int_0^T e^h_{phy}(t)\, dt.$$

In Table 7.2, the averaged quantities and the ratio of artificial to physical dissipation are listed. While the artificial dissipation dominates the physical dissipation on the coarsest mesh, it becomes less than 10% on the finest mesh.

Although it cannot be expected that a solution u_H on a coarse mesh $\mathcal{T}_H$ shows quantitatively the same information as the solution u_h on the finest mesh $\mathcal{T}_h$, time averages should be comparable if the mesh is reasonably fine. We may hope that $\overline{u}_H$ coincides well with the L^2-projection of $\overline{u}_h$ onto X_H, denoted by $\Pi_H\overline{u}_h$. This is illustrated in Figures 7.6 and 7.7, where the isolines of the time-averaged velocity components are shown for the grid with 32^2 Q_2 cells. Although the mesh size differs by a factor of 2^4, the two time averages coincide quite well.

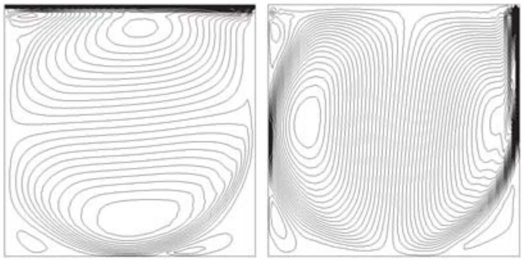


FIG. 7.6. *Time-averaged velocities on a grid with* 32×32 Q_2 *cells; isolines of horizontal velocity component (left) and vertical velocity component (right).*

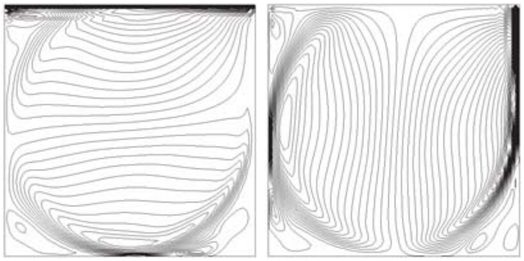


FIG. 7.7. *Time-averaged solution on a grid with* 512×512 Q_2 *cells* L^2*-projected onto the finite element space with* 32×32 Q_2 *cells,* $\Pi_H \overline{u}_h$*; isolines of horizontal velocity component (left) and vertical velocity component (right).*

8. Concluding remarks. We have proposed and analyzed a stabilized finite element method for the Oseen system based on local projections. To assure stability for the equal-order interpolation of velocity and pressure and for the case of high Reynolds number, a sufficient condition on the characteristic length scale of the subgrid model is established. This condition coincides with the condition for optimal-order convergence for the stabilized method when the underlying exact solution is smooth. We have discussed how the choices of stabilization parameters may influence the precision of the computation. Moreover, we have shown that the method can be formulated in a multiscale setting, hence rigorously establishing a link between stabilized methods and the VMS method for Navier–Stokes equations. We hope that this contribution will give additional insight into the close relationship between VMS and stabilized finite element methods. More extensive numerical simulations will be reported in a forthcoming paper.

REFERENCES

[1] F. AUTERI, N. PAROLINI, AND L. QUARTAPELLE, *Numerical investigation on the stability of singular driven cavity flow*, J. Comput. Phys., 183 (2002), pp. 1–25.

[2] R. BECKER AND M. BRAACK, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, Calcolo, 38 (2001), pp. 173–199.

[3] R. BECKER AND M. BRAACK, *A two-level stabilization scheme for the Navier-Stokes equations*, in Numerical Mathematics and Advanced Applications, M. Feistauer et al., eds., ENUMATH 2003, Springer-Verlag, Berlin, 2004, pp. 123–130.

[4] F. BREZZI AND A. RUSSO, *Choosing bubbles for advection-diffusion problems*, Math. Models Methods Appl. Sci., 4 (1994), pp. 571–587.

[5] E. BURMAN AND P. HANSBO, *The edge stabilization method for finite elements in CFD*, in Numerical Mathematics and Advanced Applications, M. Feistauer et al., eds., ENUMATH 2003, Springer-Verlag, Berlin, 2004.

[6] E. Burman and P. Hansbo, *A stabilized non-conforming finite element method for incompressible flow*, Comp. Meth. Mech. Eng., in press, 2005.

[7] W. Cazemir, R. Verstappen, and A. Veldman, *Proper orthogonal decomposition and low-dimensional models for the driven cavity flows*, Phys. Fluids, 10 (1998), pp. 1685–1699.

[8] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO, Anal. Numer., 9 (1975), pp. 77–84.

[9] S. Collis, *Monitoring unresolved scales in multiscale turbulence modeling*, Phys. Fluids, 13 (2001), pp. 1800–1806.

[10] T. Dubois, F. Jauberteau, and R. Temam, *Incremental unknowns, multilevel methods and the numerical simulation of turbulence*, Comput. Methods Appl. Mech. Engrg., 159 (1998), pp. 123–189.

[11] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.

[12] V. Girault and P.-A. Raviart, *Finite Elements for the Navier Stokes Equations*, Springer-Verlag, Berlin, 1986.

[13] O. Goyon, *High-Reynolds number solutions of Navier-Stokes equations using incremental unknowns*, Comput. Methods Appl. Mech. Engrg., 130 (1996), pp. 319–335.

[14] J.-L. Guermond, *Stabilization of Galerkin approximations of transport equations by subgrid modeling*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1293–1316.

[15] J. R. T. Hughes, L. Mazzei, and A. A. Oberai, *The multiscale formulation of large eddy simulation: Decay of homogeneous isotropic turbulence*, Phys. Fluids, 13 (2001), pp. 505–511.

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

[17] Y.-H. Peng, Y.-H. Shiau, and R. Hwang, *Transition in a 2d lid-driven cavity flow*, Comput. & Fluids, 32 (2003), pp. 337–352.

[18] T. C. Rebollo and A. D. Delgado, *A unified analysis of mixed and stabilized finite element solutions of Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 182 (2000), pp. 301–331.

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

[20] G. Tiesinga, F. Wubs, and A. Veldman, *Bifurcation analysis of incompressible flow in a cavity by the Newton-Picard method*, J. Comput. Appl. Math., 140 (2002), pp. 751–772.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2567–2589

GENERALIZED LAGUERRE INTERPOLATION AND PSEUDOSPECTRAL METHOD FOR UNBOUNDED DOMAINS*

GUO BEN-YU[†], WANG LI-LIAN[‡], AND WANG ZHONG-QING[†]

Abstract. In this paper, error estimates for generalized Laguerre–Gauss-type interpolations are derived in nonuniformly weighted Sobolev spaces weighted with $\omega_{\alpha,\beta}(x) = x^\alpha e^{-\beta x},\ \alpha > -1, \beta > 0$. Generalized Laguerre pseudospectral methods are analyzed and implemented. Two model problems are considered. The proposed schemes keep spectral accuracy and, with suitable choice of basis functions, lead to sparse and symmetric linear systems.

Key words. generalized Laguerre–Gauss-type interpolations, pseudospectral method, unbounded domains, exterior problems

AMS subject classifications. 33C45, 41A05, 65N35

DOI. 10.1137/04061324X

1. Introduction.

With the extensive applications of Legendre- and Chebyshev-spectral approximations to PDEs in bounded domains (cf. [2, 3, 4, 6, 7, 8]), considerable progress has been made recently in using spectral methods for solving PDEs in unbounded domains. Among the existing methods, the direct and commonly used approach is based on orthogonal systems in infinite intervals, i.e., the Hermite and Laguerre spectral methods (see, e.g., [5, 6, 9, 10, 17, 19]). In earlier studies, one usually considers Laguerre approximations in spaces weighted with e^{-x}, which are not the most appropriate in some cases. For instance, the approximations of some differential equations in financial mathematics, fluid dynamics, quantum mechanics, and astronomical physics involve different weight functions for derivatives of different orders. In such cases, we have to consider the generalized Laguerre approximation with weight function $\omega_\alpha(x) = x^\alpha e^{-x},\ \alpha > -1$, which was used recently for two-dimensional exterior problems; see [11]. Indeed, from both theoretical and computational points of view, it is more interesting to consider an orthogonal system with a more general weight function: $\omega_{\alpha,\beta}(x) = x^\alpha e^{-\beta x},\ \alpha > -1,\ \beta > 0$. One obvious advantage is that it can provide us a variety of choices of polynomial bases to fit exact solutions of underlying differential equations with various asymptotic behaviors at infinity. Moreover, as we will see later, some other good by-products can be obtained using this new family of orthogonal polynomials.

In actual computations, it is more preferable to use the Laguerre interpolation. As we know, there have been many results on the Laguerre polynomial approximation (e.g., see, [2, 5, 6, 8, 10, 11, 12, 13, 14, 17]), but only a few papers dealing with the error analysis of Laguerre interpolation. Recently, some authors developed the

[*]Received by the editors August 11, 2004; accepted for publication (in revised form) April 6, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/sinum/43-6/61324.html

[†]Department of Mathematics, Division of Computational Science, E-Institute of Shanghai Universities, Shanghai Normal University, Shanghai 200234, People's Republic of China (byguo@shnu.edu.cn, wang72@purdue.edu). The work of the first and second authors was supported in part by the NSF of China (grant 10471095), the SF of Shanghai (grant 04JC14062), the fund of Chinese Education Ministry (grant 20040270002), the funds 04DB15, E-institutes E03004, The Shanghai Leading Academic Discipline Project N.T0401, and major specialties of Shanghai Education Commission.

[‡]Department of Mathematics, Shanghai Normal University, Shanghai 200234, People's Republic of China, and Department of Mathematics, Purdue University, West Lafayette, IN 47907 (lwang@math.purdue.edu).

Laguerre interpolation—for example, the Laguerre interpolation ($\alpha=0,\beta=1$) with its applications to approximation of differential equations (see [19]) and the standard generalized Laguerre interpolation ($\alpha>-1,\beta=1$), which are very useful for approximation of integral equations (see [15, 16]). The objective of this paper is to analyze the generalized Laguerre–Gauss-type interpolation errors with a more general weight $\omega_{\alpha,\beta}(x),\alpha>-1,\beta>0$. In the special case of $\alpha=0,\beta=1$, our new results are better than the previous ones. Moreover, we derive the approximation results in nonuniformly weighted Sobolev spaces, which enables us to develop and analyze efficient generalized Laguerre pseudospectral approximations of a large class of problems in unbounded domains.

This paper is organized as follows. In section 2, we present some basic results on this new generalized Laguerre–Gauss-type interpolation. In section 3, we establish the main approximation results on the generalized Laguerre–Gauss and Laguerre–Gauss–Radau interpolations, which provide us useful tools for numerical analysis of generalized Laguerre pseudospectral methods for unbounded domains. Section 4 is devoted to the generalized Laguerre pseudospectral method for unbounded domains as an important application of the generalized Laguerre–Gauss interpolation. In section 5, we develop a pseudospectral method for exterior problems as an application of the generalized Laguerre–Gauss–Radau interpolation. In section 6, we present some numerical results, which demonstrate the spectral accuracy of proposed schemes. The final section is for some concluding remarks.

2. Generalized Laguerre–Gauss-type interpolations.

In this section, we shall introduce the new generalized Laguerre–Gauss-type interpolations, and study the asymptotic behaviors of the interpolation nodes and weights.

2.1. Notation and preliminaries.

Let $\Lambda=(0,\infty)$ and $\chi(x)$ be a certain weight function on Λ in the usual sense. We define the weighted space $L^2_\chi(\Lambda)$ as usual with the inner product $(u,v)_\chi$ and the norm $\|v\|_\chi$. For simplicity, we denote $\partial_x^k v(x)=\frac{d^k}{dx^k}v(x), k\ge 1$. For any integer $m\ge 0$, $H^m_\chi(\Lambda)=\{v\mid \partial_x^k v\in L^2_\chi(\Lambda), 0\le k\le m\}$ with the seminorm $|v|_{m,\chi}$ and the norm $\|v\|_{m,\chi}$. For any real $r>0$, we define the space $H^r_\chi(\Lambda)$ and its norm $\|v\|_{r,\chi}$ by space interpolation as in [1]. For $\chi(x)\equiv 1$, we drop the subscript χ in the previous notations as usual.

Let $\omega_{\alpha,\beta}(x)=x^\alpha e^{-\beta x}$, $\alpha>-1$, $\beta>0$. In particular, we denote $\omega_\alpha(x)=\omega_{\alpha,1}(x)=x^\alpha e^{-x}$. The new generalized Laguerre polynomial of degree l is defined by

$$\mathcal{L}_l^{(\alpha,\beta)}(x)=\frac{1}{l!}x^{-\alpha}e^{\beta x}\partial_x^l(x^{l+\alpha}e^{-\beta x}),\quad l=0,1,\dots .$$

Let $\mathcal{L}_l^{(\alpha)}(x)$ be the usual generalized Laguerre polynomials that are mutually orthogonal with the weight function $\omega_\alpha(x)$. It is noted that $\mathcal{L}_l^{(\alpha)}(x)=\mathcal{L}_l^{(\alpha,1)}(x)$, and

$$\mathcal{L}_l^{(\alpha,\beta)}(x)=\mathcal{L}_l^{(\alpha)}(y)=\mathcal{L}_l^{(\alpha)}(\beta x),\quad y=\beta x. \tag{2.1}$$

Therefore, it is straightforward to derive the following properties (cf. [18]):

$$\mathcal{L}_l^{(\alpha,\beta)}(0)=\mathcal{L}_l^{(\alpha)}(0)=\frac{\Gamma(l+\alpha+1)}{\Gamma(\alpha+1)\Gamma(l+1)},\quad l\ge 0, \tag{2.2}$$

$$\partial_x\mathcal{L}_l^{(\alpha,\beta)}(x)=-\beta\mathcal{L}_{l-1}^{(\alpha+1,\beta)}(x),\quad l\ge 1, \tag{2.3}$$

$$(l+1)\mathcal{L}_{l+1}^{(\alpha,\beta)}(x)=(2l+\alpha+1-\beta x)\mathcal{L}_l^{(\alpha,\beta)}(x)-(l+\alpha)\mathcal{L}_{l-1}^{(\alpha,\beta)}(x),\quad l\ge 1, \tag{2.4}$$

(2.5)
$$\mathcal{L}_l^{(\alpha,\beta)}(x) = \mathcal{L}_l^{(\alpha+1,\beta)}(x) - \mathcal{L}_{l-1}^{(\alpha+1,\beta)}(x) = \beta^{-1}\Big(\partial_x \mathcal{L}_l^{(\alpha,\beta)}(x) - \partial_x \mathcal{L}_{l+1}^{(\alpha,\beta)}(x)\Big), \quad l \geq 1.$$

The generalized Laguerre polynomials form a complete $L^2_{\omega_{\alpha,\beta}}(\Lambda)$-orthogonal system,

$$\Big(\mathcal{L}_l^{(\alpha,\beta)}, \mathcal{L}_m^{(\alpha,\beta)}\Big)_{\omega_{\alpha,\beta}} = \gamma_l^{(\alpha,\beta)}\delta_{l,m}, \quad \gamma_l^{(\alpha,\beta)} = \frac{\Gamma(l+\alpha+1)}{\beta^{\alpha+1}\Gamma(l+1)}, \tag{2.6}$$

where $\delta_{l,m}$ is the Kronecker symbol. Hence, for any $v \in L^2_{\omega_{\alpha,\beta}}(\Lambda)$, we can write

$$v(x) = \sum_{l=0}^{\infty} \hat{v}_l^{(\alpha,\beta)} \mathcal{L}_l^{(\alpha,\beta)}(x), \qquad \hat{v}_l^{(\alpha,\beta)} = \frac{1}{\gamma_l^{(\alpha,\beta)}}(v, \mathcal{L}_l^{(\alpha,\beta)})_{\omega_{\alpha,\beta}}. \tag{2.7}$$

For integer $N > 0$, $\mathbb{P}_N$ stands for the set of algebraic polynomials of degree $\leq N$. We denote by c a generic positive constant independent of N, β, and any function.

2.2. Generalized Laguerre–Gauss and Laguerre–Gauss–Radau interpolations. Let $\xi_{G,N,j}^{(\alpha,\beta)}$ and $\xi_{R,N,j}^{(\alpha,\beta)}$, $0 \leq j \leq N$, be the zeros of $\mathcal{L}_{N+1}^{(\alpha,\beta)}(x)$ and $x\partial_x \mathcal{L}_{N+1}^{(\alpha,\beta)}(x)$, respectively. They are arranged in ascending order. Denote $\omega_{Z,N,j}^{(\alpha,\beta)}$, $0 \leq j \leq N$, $Z = G, R$, the corresponding Christoffel numbers such that

$$\int_\Lambda \phi(x)\omega_{\alpha,\beta}(x)\,dx = \sum_{j=0}^{N} \phi\Big(\xi_{Z,N,j}^{(\alpha,\beta)}\Big)\omega_{Z,N,j}^{(\alpha,\beta)} \quad \forall \phi \in \mathbb{P}_{2N+\lambda_Z}, \tag{2.8}$$

where $\lambda_z = 1$ and 0 for $Z = G$ and R, respectively. In particular, the usual generalized Laguerre–Gauss-type quadrature nodes and weights are denoted by $\xi_{Z,N,j}^{(\alpha)} := \xi_{Z,N,j}^{(\alpha,1)}$ and $\omega_{Z,N,j}^{(\alpha)} := \omega_{Z,N,j}^{(\alpha,1)}$, $Z = G, R$, respectively. Thanks to (2.1), we have $\xi_{Z,N,j}^{(\alpha,\beta)} = \frac{1}{\beta}\xi_{Z,N,j}^{(\alpha)}$. We next derive the expressions of the weights. Indeed,

$$\omega_{G,N,j}^{(\alpha,\beta)} = \frac{1}{\partial_x \mathcal{L}_{N+1}^{(\alpha,\beta)}\big(\xi_{G,N,j}^{(\alpha,\beta)}\big)} \int_\Lambda \frac{\mathcal{L}_{N+1}^{(\alpha,\beta)}(x)}{x - \xi_{G,N,j}^{(\alpha,\beta)}}\omega_{\alpha,\beta}(x)dx, \quad 0 \leq j \leq N, \tag{2.9}$$

which, along with formula (15.3.5) of [18], leads to

(2.10)
$$\omega_{G,N,j}^{(\alpha,\beta)} = \frac{1}{\beta^{\alpha+1}}\omega_{G,N,j}^{(\alpha)} = \frac{\Gamma(N+\alpha+2)}{\beta^\alpha \Gamma(N+2)} \frac{1}{\xi_{G,N,j}^{(\alpha,\beta)}\Big[\partial_x \mathcal{L}_{N+1}^{(\alpha,\beta)}\big(\xi_{G,N,j}^{(\alpha,\beta)}\big)\Big]^2}, \quad 0 \leq j \leq N.$$

Similarly, for the Gauss–Radau weights, we have

(2.11)
$$\omega_{R,N,j}^{(\alpha,\beta)} = \frac{1}{\partial_x\Big[x\partial_x \mathcal{L}_{N+1}^{(\alpha,\beta)}(x)\Big]\big|_{x=\xi_{R,N,j}^{(\alpha,\beta)}}} \int_\Lambda \frac{x\partial_x \mathcal{L}_{N+1}^{(\alpha,\beta)}(x)}{x - \xi_{R,N,j}^{(\alpha,\beta)}}\omega_{\alpha,\beta}(x)dx, \quad 0 \leq j \leq N,$$

which, together with formula (3.6.2) of [6], yields

(2.12)
$$\omega_{R,N,j}^{(\alpha,\beta)} = \frac{1}{\beta^{\alpha+1}}\omega_{R,N,j}^{(\alpha)} = \begin{cases} \dfrac{(\alpha+1)\Gamma^2(\alpha+1)\Gamma(N+1)}{\beta^{\alpha+1}\Gamma(N+\alpha+2)}, & j = 0, \\[2ex] \dfrac{\Gamma(N+\alpha+1)}{\beta^\alpha \Gamma(N+2)} \dfrac{1}{\mathcal{L}_{N+1}^{(\alpha,\beta)}(\xi_{R,N,j}^{(\alpha,\beta)})\partial_x \mathcal{L}_N^{(\alpha,\beta)}(\xi_{R,N,j}^{(\alpha,\beta)})}, & 1 \leq j \leq N. \end{cases}$$

Note that the earlier two types of quadratures have close relations:

$$\xi^{(\alpha,\beta)}_{R,N,j} = \xi^{(\alpha+1,\beta)}_{G,N-1,j-1}, \quad \omega^{(\alpha,\beta)}_{R,N,j} = \left(\xi^{(\alpha,\beta)}_{R,N,j}\right)^{-1} \omega^{(\alpha+1,\beta)}_{G,N-1,j-1}, \quad 1 \le j \le N. \tag{2.13}$$

Indeed, the first identity follows from (2.3). Moreover, using (2.3), (2.9), (2.13), and the definition of $\xi^{(\alpha+1,\beta)}_{G,N-1,j-1}$, we obtain from (2.11) that for $1 \le j \le N$,

$$\begin{aligned} \omega^{(\alpha,\beta)}_{R,N,j} &= \frac{1}{\partial_x \left(x\mathcal{L}^{(\alpha+1,\beta)}_N(x)\right)\big|_{x=\xi^{(\alpha,\beta)}_{R,N,j}}} \int_\Lambda \frac{x\mathcal{L}^{(\alpha+1,\beta)}_N(x)}{x-\xi^{(\alpha,\beta)}_{R,N,j}} \omega_{\alpha,\beta}(x)\,dx \\ &= \frac{1}{\xi^{(\alpha,\beta)}_{R,N,j}\partial_x \mathcal{L}^{(\alpha+1,\beta)}_N\left(\xi^{(\alpha+1,\beta)}_{G,N-1,j-1}\right)} \int_\Lambda \frac{\mathcal{L}^{(\alpha+1,\beta)}_N(x)}{x-\xi^{(\alpha+1,\beta)}_{G,N-1,j-1}} \omega_{\alpha+1,\beta}(x)\,dx \\ &= \left(\xi^{(\alpha,\beta)}_{R,N,j}\right)^{-1} \omega^{(\alpha+1,\beta)}_{G,N-1,j-1}. \end{aligned} \tag{2.14}$$

To obtain the interpolation error estimates, it is necessary to study the asymptotic behaviors of generalized Laguerre–Gauss interpolation nodes and weights.

- Using Theorem 8.9.2 of [18], we can verify that for a certain fixed number $\eta > 0$,

$$2\beta^{\frac{1}{2}}\left(\left(\xi^{(\alpha,\beta)}_{G,N,j}\right)\right)^{\frac{1}{2}} = \frac{1}{\sqrt{N+1}}\big(j\pi + \mathcal{O}(1)\big) \quad \text{if } 0 < \left(\xi^{(\alpha,\beta)}_{G,N,j}\right) \le \frac{\eta}{\beta}. \tag{2.15}$$

- Theorem 6.31.3 of [18] reveals that for large j,

$$\frac{c_1 j^2}{\beta(N+\frac{\alpha}{2}+\frac{3}{2})} < \left(\xi^{(\alpha,\beta)}_{G,N,j}\right) < \frac{c_2 j^2}{\beta(N+\frac{\alpha}{2}+\frac{3}{2})}, \quad c_1 \cong \frac{\pi^2}{4}, \quad c_2 \cong 4. \tag{2.16}$$

- Let $\tilde{N} = 2(N+1)+\alpha+1$. By Theorem 6.31.2 of [18], the largest node satisfies

$$\xi^{(\alpha,\beta)}_{G,N,N} < \beta^{-1}\left(\tilde{N} + \left(\tilde{N}^2 + \frac{1}{4} - \alpha^2\right)^{1/2}\right) \cong 4\beta^{-1}(N+1). \tag{2.17}$$

- We can verify from formula (15.3.15) of [18] that for a certain fixed number $\eta > 0$,

$$\omega^{(\alpha,\beta)}_{G,N,j} \cong \frac{\pi}{\sqrt{\beta N}} e^{-\beta \xi^{(\alpha,\beta)}_{G,N,j}} \left(\xi^{(\alpha,\beta)}_{G,N,j}\right)^{\alpha+\frac{1}{2}} \quad \text{if } 0 < \left(\xi^{(\alpha,\beta)}_{G,N,j}\right) \le \frac{\eta}{\beta}. \tag{2.18}$$

- Let $\xi^{(\alpha,\beta)}_{G,N,-1} := 0$. By the formulae (2.4), (2.5), and (2.7) of [15],

$$\begin{aligned} \omega^{(\alpha,\beta)}_{G,N,j} &= \frac{1}{\beta^{\alpha+1}} \omega^{(\alpha)}_{G,N,j} \sim \frac{1}{\beta^{\alpha+1}} \omega_\alpha\left(\xi^{(\alpha)}_{G,N,j}\right)\left(\xi^{(\alpha)}_{G,N,j+1} - \xi^{(\alpha)}_{G,N,j}\right) \\ &= \omega_{\alpha,\beta}(\xi^{(\alpha,\beta)}_{G,N,j})\left(\xi^{(\alpha,\beta)}_{G,N,j} - \xi^{(\alpha,\beta)}_{G,N,j-1}\right), \quad 0 \le j \le N. \end{aligned} \tag{2.19}$$

- Thanks to the relation (2.13), we deduce from (2.18) and (2.19) that

$$\begin{aligned} \omega^{(\alpha,\beta)}_{R,N,j} &= \left(\xi^{(\alpha+1,\beta)}_{G,N-1,j-1}\right)^{-1} \omega^{(\alpha+1,\beta)}_{G,N-1,j-1} \\ &\cong \frac{\pi}{\sqrt{\beta(N-1)}} e^{-\beta \xi^{(\alpha,\beta)}_{R,N,j}} \left(\xi^{(\alpha,\beta)}_{R,N,j}\right)^{\alpha+\frac{1}{2}} \quad \text{if } 0 < \xi^{(\alpha,\beta)}_{R,N,j} \le \frac{\eta}{\beta}, \quad 1 \le j \le N, \end{aligned} \tag{2.20}$$

and

$$\begin{aligned}
\omega_{R,N,j}^{(\alpha,\beta)} &= \left(\xi_{R,N,j}^{(\alpha,\beta)}\right)^{-1} \omega_{G,N-1,j-1}^{(\alpha+1,\beta)} \\
&\sim \left(\xi_{R,N,j}^{(\alpha,\beta)}\right)^{-1} \omega_{\alpha+1,\beta}\left(\xi_{G,N-1,j-1}^{(\alpha+1,\beta)}\right)\left(\xi_{G,N-1,j-1}^{(\alpha+1,\beta)} - \xi_{G,N-1,j-2}^{(\alpha+1,\beta)}\right) \\
&= \omega_{\alpha,\beta}\left(\xi_{R,N,j}^{(\alpha,\beta)}\right)\left(\xi_{R,N,j}^{(\alpha,\beta)} - \xi_{R,N,j-1}^{(\alpha,\beta)}\right), \quad 1 \le j \le N.
\end{aligned} \tag{2.21}$$

For notational convenience, we now introduce the discrete inner product and norm,

$$(u,v)_{\omega_{\alpha,\beta},Z,N} = \sum_{j=0}^{N} u\left(\xi_{Z,N,j}^{(\alpha,\beta)}\right) v\left(\xi_{Z,N,j}^{(\alpha,\beta)}\right) \omega_{Z,N,j}^{(\alpha,\beta)},$$

$$\|v\|_{\omega_{\alpha,\beta},Z,N} = (v,v)_{\omega_{\alpha,\beta},Z,N}^{\frac{1}{2}}, \quad Z = G, R.$$

By the exactness of (2.8),

$$(\phi,\psi)_{\omega_{\alpha,\beta},Z,N} = (\phi,\psi)_{\omega_{\alpha,\beta}} \quad \forall \phi\psi \in \mathbb{P}_{2N+\delta_Z}, \tag{2.22}$$

where $\delta_Z = 1, 0$ for $Z = G, R$, respectively. In particular,

$$\|\phi\|_{\omega_{\alpha,\beta},Z,N} = \|\phi\|_{\omega_{\alpha,\beta}} \quad \forall \phi \in \mathbb{P}_N, \quad Z = G, R. \tag{2.23}$$

The generalized Laguerre–Gauss interpolant $\mathcal{I}_{Z,N,\alpha,\beta} v \in \mathbb{P}_N$ is defined by

$$\mathcal{I}_{Z,N,\alpha,\beta} v\left(\xi_{Z,N,j}^{(\alpha,\beta)}\right) = v\left(\xi_{Z,N,j}^{(\alpha,\beta)}\right), \quad Z = G, R, \quad 0 \le j \le N. \tag{2.24}$$

3. Generalized Laguerre interpolation error estimates.

In this section, we estimate the interpolation errors in weighted Sobolev spaces, which provide useful tools for the analysis of generalized Laguerre pseudospectral methods.

3.1. $L^2_{\omega_{\alpha,\beta}}(\Lambda)$-orthogonal projection.

We first recall the $L^2_{\omega_{\alpha,\beta}}(\Lambda)$-orthogonal projection $P_{N,\alpha,\beta} : L^2_{\omega_{\alpha,\beta}}(\Lambda) \to \mathbb{P}_N$, defined by

$$(P_{N,\alpha,\beta} v - v, \phi)_{\omega_{\alpha,\beta}} = 0 \quad \forall \phi \in \mathbb{P}_N.$$

In order to describe approximation errors precisely, we introduce the nonuniformly weighted Sobolev space $A^r_{\alpha,\beta}(\Lambda)$. For any integer $r \ge 0$, its seminorm and norm are given by

$$|v|_{A^r_{\alpha,\beta}} = \|\partial_x^r v\|_{\omega_{\alpha+r,\beta}}, \quad \|v\|_{A^r_{\alpha,\beta}} = \left(\sum_{k=0}^{r} |v|^2_{A^k_{\alpha,\beta}}\right)^{\frac{1}{2}}.$$

For any real $r > 0$, we define the space $A^r_{\alpha,\beta}(\Lambda)$ by space interpolation as in [1].

We have the following basic result; see Theorem 2.1 of [12].

Lemma 3.1. *For any $v \in A^r_{\alpha,\beta}(\Lambda)$, an integer r, and $0 \le \mu \le r$,*

$$\|P_{N,\alpha,\beta} v - v\|_{A^\mu_{\alpha,\beta}} \le c(\beta N)^{\frac{\mu - r}{2}} |v|_{A^r_{\alpha,\beta}}. \tag{3.1}$$

In the analysis of generalized Laguerre–Gauss–Radau interpolation approximation (cf. the proof of Theorem 3.7), we need to estimate $|P_{N,\alpha,\beta} v(0) - v(0)|$.

LEMMA 3.2. *For any $v \in A^r_{\alpha,\beta}(\Lambda)$ and an integer $r > \alpha + 1$,*

$$|P_{N,\alpha,\beta}v(0) - v(0)| \le c(\beta N)^{\frac{\alpha-r+1}{2}} |v|_{A^r_{\alpha,\beta}}. \tag{3.2}$$

Proof. Let $\lambda_l^{(\beta)} = \beta l$. By virtue of (2.3) and (2.6), we find that for $l \ge r$,

$$|v|^2_{A^r_{\alpha,\beta}} = \sum_{l=r}^{\infty} \beta^{2r} \gamma_{l-r}^{(\alpha+r,\beta)} \big(\hat{v}_l^{(\alpha,\beta)}\big)^2, \quad d_{l,r}^{\alpha,\beta} := \frac{(\lambda_l^{(\beta)})^r \gamma_l^{(\alpha,\beta)}}{\gamma_{l-r}^{(\alpha+r,\beta)}} \le c\beta^{2r}.$$

Therefore,

$$\sum_{l=N+1}^{\infty} \big(\lambda_l^{(\beta)}\big)^r \gamma_l^{(\alpha,\beta)} \big(\hat{v}_l^{(\alpha,\beta)}\big)^2 = \sum_{l=N+1}^{\infty} d_{l,r}^{\alpha,\beta} \gamma_{l-r}^{(\alpha+r,\beta)} \big(\hat{v}_l^{(\alpha,\beta)}\big)^2 \le c|v|^2_{A^r_{\alpha,\beta}}. \tag{3.3}$$

Consequently, using (2.2), (2.6), (3.3), and the Cauchy–Schwarz inequality leads to

$$\begin{aligned}
|P_{N,\alpha,\beta}v(0) - v(0)| &= \Big| \sum_{l=N+1}^{\infty} \hat{v}_l^{(\alpha,\beta)} \mathcal{L}_l^{(\alpha,\beta)}(0) \Big| \\
&\le \left(\sum_{l=N+1}^{\infty} (\lambda_l^{(\beta)})^{-r} (\mathcal{L}_l^{(\alpha,\beta)}(0))^2 (\gamma_l^{(\alpha,\beta)})^{-1} \right)^{\frac{1}{2}} \left(\sum_{l=N+1}^{\infty} (\lambda_l^{(\beta)})^{r} \gamma_l^{(\alpha,\beta)} (\hat{v}_l^{(\alpha,\beta)})^2 \right)^{\frac{1}{2}} \\
&\le c\beta^{\frac{\alpha-r+1}{2}} \left(\sum_{l=N+1}^{\infty} \frac{\Gamma(l+\alpha+1)}{l^r \Gamma(l+1)} \right)^{\frac{1}{2}} |v|_{A^r_{\alpha,\beta}}.
\end{aligned}$$

By the Stirling formula, $\Gamma(s+1) = \sqrt{2\pi s} s^s e^{-s}(1 + \mathcal{O}(s^{-\frac{1}{5}}))$. Thus, for $r > \alpha + 1$,

$$\sum_{l=N+1}^{\infty} \frac{\Gamma(l+\alpha+1)}{l^r \Gamma(l+1)} \le c \sum_{l=N+1}^{\infty} l^{\alpha-r} \le cN^{\alpha-r+1}.$$

This completes the proof. □

The approximation errors stated in Lemma 3.1 are measured in the space $A^{\mu}_{\alpha,\beta}(\Lambda)$. However, when we apply the generalized Laguerre approximation to numerical solutions of differential and integral equations, we oftentimes need to estimate them in the standard weighted Sobolev space $H^r_{\omega_{\alpha,\beta}}(\Lambda)$, stated later.

LEMMA 3.3. *If $v \in H^{\mu}_{\omega_{\alpha,\beta}}(\Lambda) \cap A^r_{\alpha-1,\beta}(\Lambda) \cap A^r_{\alpha-\mu,\beta}(\Lambda)$, then for integers $1 \le \mu \le r$,*

$$|P_{N,\alpha,\beta}v - v|_{\mu,\omega_{\alpha,\beta}} \le c\beta^{-\frac{1}{2}} (\beta N)^{\mu - \frac{r}{2}} \big(|v|_{A^r_{\alpha-1,\beta}} + |v|_{A^r_{\alpha-\mu,\beta}} \big). \tag{3.4}$$

Proof. We have

$$|P_{N,\alpha,\beta}v - v|_{1,\omega_{\alpha,\beta}} \le \|P_{N,\alpha,\beta}\partial_x v - \partial_x v\|_{\omega_{\alpha,\beta}} + \|P_{N,\alpha,\beta}\partial_x v - \partial_x P_{N,\alpha,\beta} v\|_{\omega_{\alpha,\beta}}. \tag{3.5}$$

By (3.1) with $\mu = 0$, the first term at the right side of the previous inequality is bounded above by $c(\beta N)^{\frac{1-r}{2}} |v|_{A^r_{\alpha-1,\beta}}$. Hence, it remains to estimate the second term. To do this, let $\partial_x v(x) = \sum_{l=0}^{\infty} \hat{\hat{v}}_l^{(\alpha,\beta)} \mathcal{L}_l^{(\alpha,\beta)}(x)$. By virtue of (2.5) and (2.7), we can derive that $\hat{\hat{v}}_l^{(\alpha,\beta)} = -\beta \sum_{p=l+1}^{\infty} \hat{v}_p^{(\alpha,\beta)}$. Thus, we follow the same lines as in [2, 8] to deduce that

$$\begin{aligned} P_{N,\alpha,\beta}\partial_x v(x) - \partial_x P_{N,\alpha,\beta} v(x) &= -\beta \sum_{l=0}^{N} \mathcal{L}_l^{(\alpha,\beta)}(x) \left(\sum_{p=l+1}^{\infty} \hat{v}_p^{(\alpha,\beta)} \right) \\ &\quad + \beta \sum_{l=0}^{N-1} \mathcal{L}_l^{(\alpha,\beta)}(x) \left(\sum_{p=l+1}^{N} \hat{v}_p^{(\alpha,\beta)} \right) = \hat{\hat{v}}_N^{(\alpha,\beta)} \sum_{l=0}^{N} \mathcal{L}_l^{(\alpha,\beta)}(x). \end{aligned} \tag{3.6}$$

Accordingly, we use (2.6) and (3.1) with $\mu = 0$ to obtain that

$$\begin{aligned} \|P_{N,\alpha,\beta}\partial_x v - \partial_x P_{N,\alpha,\beta} v\|_{\omega_{\alpha,\beta}}^2 &= \big(\hat{\hat{v}}_N^{(\alpha,\beta)}\big)^2 \gamma_N^{(\alpha,\beta)} \sum_{l=0}^{N} \gamma_l^{(\alpha,\beta)} \big(\gamma_N^{(\alpha,\beta)}\big)^{-1} \\ &\le \|P_{N-1,\alpha,\beta}\partial_x v - \partial_x v\|_{\omega_{\alpha,\beta}}^2 \sum_{l=0}^{N} \gamma_l^{(\alpha,\beta)} \big(\gamma_N^{(\alpha,\beta)}\big)^{-1} \\ &\le c(\beta N)^{1-r} |v|_{A_{\alpha-1,\beta}^r}^2 \sum_{l=0}^{N} \gamma_l^{(\alpha,\beta)} \big(\gamma_N^{(\alpha,\beta)}\big)^{-1}. \end{aligned} \tag{3.7}$$

If $\alpha \ge 0$, then $\gamma_l^{(\alpha,\beta)}$ increases as l increases. In this case,

$$\sum_{l=0}^{N} \gamma_l^{(\alpha,\beta)} \big(\gamma_N^{(\alpha,\beta)}\big)^{-1} \le N + 1. \tag{3.8}$$

For $-1 < \alpha < 0$, we use the Stirling formula to deduce that for a suitably large integer $M < N$ and $l \ge M$,

$$\gamma_l^{(\alpha,\beta)} \sim \beta^{-\alpha-1} \Big(1 + \frac{\alpha}{l}\Big)^{l+\frac{1}{2}} (l+\alpha)^{\alpha} \sim \beta^{-\alpha-1} l^{\alpha}. \tag{3.9}$$

Hence, for certain $c_1 > 0$,

$$\sum_{l=0}^{N} \gamma_l^{(\alpha,\beta)} \big(\gamma_N^{(\alpha,\beta)}\big)^{-1} \le cN^{-\alpha} \left(c_1 + c \sum_{l=M}^{N} l^{\alpha} \right) \le cN^{-\alpha}(c_1 + cN^{1+\alpha}) \le cN. \tag{3.10}$$

Inserting (3.8) and (3.10) into (3.7), we obtain the desired result with $\mu = 1$.

Now, we use induction to derive the desired result with $\mu \ge 2$. We shall use the following inverse inequality:

$$\|\phi\|_{r,\omega_{\alpha,\beta}} \le c(\beta N)^r \|\phi\|_{\omega_{\alpha,\beta}} \quad \forall \phi \in \mathbb{P}_N, \quad r > 0.$$

Assume that (3.4) holds for $\mu - 1$. Then we obtain that

$$\begin{aligned} &|P_{N,\alpha,\beta} v - v|_{\mu,\omega_{\alpha,\beta}} \le |P_{N,\alpha,\beta}\partial_x v - \partial_x v|_{\mu-1,\omega_{\alpha,\beta}} + |P_{N,\alpha,\beta}\partial_x v - \partial_x P_{N,\alpha,\beta} v|_{\mu-1,\omega_{\alpha,\beta}} \\ &\quad \le c\beta^{-\frac{1}{2}} (\beta N)^{\mu-1-\frac{r-1}{2}} (|v|_{A_{\alpha-2,\beta}^r} + |v|_{A_{\alpha-\mu,\beta}^r}) + c(\beta N)^{\mu-1} \|P_{N,\alpha,\beta}\partial_x v \\ &\qquad - \partial_x P_{N,\alpha,\beta} v\|_{\omega_{\alpha,\beta}} \le c\beta^{-\frac{1}{2}} (\beta N)^{\mu-\frac{r}{2}-\frac{1}{2}} (|v|_{A_{\alpha-2,\beta}^r} + |v|_{A_{\alpha-\mu,\beta}^r}) \\ &\qquad + c\beta^{-\frac{1}{2}} (\beta N)^{\mu-\frac{r}{2}} |v|_{A_{\alpha-1,\beta}^r}. \end{aligned} \tag{3.11}$$

By the definition of $|\cdot|_{A^r_{\alpha,\beta}}$, we have that

$$|v|_{A^r_{\alpha-2,\beta}} = \|\partial_x^r v\|_{\omega_{\alpha+r-2,\beta}} \le c(|v|_{A^r_{\alpha-\mu,\beta}} + |v|_{A^r_{\alpha-1,\beta}}).$$

This fact with (3.11) implies the desired result. □

3.2. Generalized Laguerre interpolation approximations. We first study the stability of generalized Laguerre–Gauss interpolation.

THEOREM 3.4. *For any* $v \in H^1_{\omega_{\alpha,\beta}}(\Lambda) \cap A^1_{\alpha,\beta}(\Lambda)$,

$$\|\mathcal{I}_{G,N,\alpha,\beta} v\|_{\omega_{\alpha,\beta}} \le c\Big(\beta^{-1}N^{-\frac{1}{2}}|v|_{1,\omega_{\alpha,\beta}} + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}\|v\|_{A^1_{\alpha,\beta}}\Big). \tag{3.12}$$

Proof. By (2.23) and (2.24),

$$\|\mathcal{I}_{G,N,\alpha,\beta} v\|^2_{\omega_{\alpha,\beta}} = \|\mathcal{I}_{G,N,\alpha,\beta} v\|^2_{\omega_{\alpha,\beta},G,N} = \sum_{j=0}^{N} v^2\big(\xi^{(\alpha,\beta)}_{G,N,j}\big)\omega^{(\alpha,\beta)}_{G,N,j} := A_N + B_N, \tag{3.13}$$

where

$$A_N = \sum_{\xi^{(\alpha,\beta)}_{G,N,j} \le \frac{\eta}{\beta}} v^2\big(\xi^{(\alpha,\beta)}_{G,N,j}\big)\omega^{(\alpha,\beta)}_{G,N,j}, \quad B_N = \sum_{\xi^{(\alpha,\beta)}_{G,N,j} > \frac{\eta}{\beta}} v^2\big(\xi^{(\alpha,\beta)}_{G,N,j}\big)\omega^{(\alpha,\beta)}_{G,N,j}.$$

We first estimate A_N. For simplicity of statements, let

$$\Delta^{(\alpha,\beta)}_j = \Big[\xi^{(\alpha,\beta)}_{G,N,j-1},\ \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)\Big], \quad |\Delta^{(\alpha,\beta)}_j| = \xi^{(\alpha,\beta)}_{G,N,j} - \xi^{(\alpha,\beta)}_{G,N,j-1},$$
$$\delta^{(\alpha,\beta)}_{j,+} = \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}} + \big(\xi^{(\alpha,\beta)}_{G,N,j-1}\big)^{\frac{1}{2}}, \quad \delta^{(\alpha,\beta)}_{j,-} = \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}} - \big(\xi^{(\alpha,\beta)}_{G,N,j-1}\big)^{\frac{1}{2}}.$$

By (13.7) of [2], we know that for any $u \in H^1(a,b)$,

$$\sup_{x\in[a,b]} |u(x)| \le c\bigg(\frac{1}{\sqrt{b-a}}\|u\|_{L^2(a,b)} + \sqrt{b-a}\|\partial_x u\|_{L^2(a,b)}\bigg). \tag{3.14}$$

Thus, by (2.18) and (3.14),

$$\begin{aligned} A_N &\le \frac{c}{\sqrt{\beta N}} \sum_{\xi^{(\alpha,\beta)}_{G,N,j} \le \frac{\eta}{\beta}} \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}} \sup_{x\in\Delta^{(\alpha,\beta)}_j} |x^{\alpha} v^2(x)| \\ &\le \frac{c}{\sqrt{\beta N}} \sum_{\xi^{(\alpha,\beta)}_{G,N,j} \le \frac{\eta}{\beta}} \bigg(\big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}} \big(\delta^{(\alpha,\beta)}_{j,+}\big)^{-1}\big(\delta^{(\alpha,\beta)}_{j,-}\big)^{-1} \|x^{\frac{\alpha}{2}} v\|^2_{L^2\left(\Delta^{(\alpha,\beta)}_j\right)} \\ &\quad + \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}} \delta^{(\alpha,\beta)}_{j,+}\delta^{(\alpha,\beta)}_{j,-}\Big(\|x^{\frac{\alpha}{2}}\partial_x v\|^2_{L^2\left(\Delta^{(\alpha,\beta)}_j\right)} + \|x^{\frac{\alpha}{2}-1} v\|^2_{L^2\left(\Delta^{(\alpha,\beta)}_j\right)}\Big)\bigg). \end{aligned} \tag{3.15}$$

We now bound the terms in the previous summation. Using (2.15) yields

$$\begin{aligned} \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}} \delta^{(\alpha,\beta)}_{j,+}\delta^{(\alpha,\beta)}_{j,-} \|x^{\frac{\alpha}{2}-1} v\|^2_{L^2\left(\Delta^{(\alpha,\beta)}_j\right)} &\le \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}} \big(\delta^{(\alpha,\beta)}_{j,+}\big)^2 \int_{\Delta^{(\alpha,\beta)}_j} x^{\alpha-2} v^2(x)dx \\ &\le c\big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{3}{2}} \big(\xi^{(\alpha,\beta)}_{G,N,j-1}\big)^{-2} \int_{\Delta^{(\alpha,\beta)}_j} x^{\alpha} v^2(x)dx \\ &\le c\sqrt{\beta N}\|x^{\frac{\alpha}{2}} v\|^2_{L^2\left(\Delta^{(\alpha,\beta)}_j\right)}. \end{aligned} \tag{3.16}$$

The expression (2.15) implies that for $0 < \xi^{(\alpha,\beta)}_{G,N,j} \le \frac{\eta}{\beta}$,

$$\delta^{(\alpha,\beta)}_{j,-} \sim \frac{1}{\sqrt{\beta N}}, \quad \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}}(\delta^{(\alpha,\beta)}_{j,+})^{-1} \le c, \quad \big(\xi^{(\alpha,\beta)}_{G,N,j}\big)^{\frac{1}{2}}\delta^{(\alpha,\beta)}_{j,+} \le \frac{c}{\beta}. \tag{3.17}$$

Hence, plugging (3.16) and (3.17) into (3.15) gives

$$\begin{aligned} A_N &\le c \sum_{\Delta^{(\alpha,\beta)}_j} \Big(\|x^{\frac{\alpha}{2}} v\|^2_{L^2\left(\Delta^{(\alpha,\beta)}_j\right)} + \beta^{-2}N^{-1}\|x^{\frac{\alpha}{2}}\partial_x v\|^2_{L^2\left(\Delta^{(\alpha,\beta)}_j\right)}\Big) \\ &\le c(\|v\|^2_{\omega_{\alpha,\beta}} + \beta^{-2}N^{-1}|v|^2_{1,\omega_{\alpha,\beta}}). \end{aligned} \tag{3.18}$$

We next estimate B_N in (3.13). By (2.19) and (2.17),

$$\begin{aligned} B_N &\le c \sum_{\xi^{(\alpha,\beta)}_{G,N,j} > \frac{\eta}{\beta}} v^2\big(\xi^{(\alpha,\beta)}_{G,N,j}\big)\omega_{\alpha,\beta}\big(\xi^{(\alpha,\beta)}_{G,N,j}\big)\Big(\xi^{(\alpha,\beta)}_{G,N,j} - \xi^{(\alpha,\beta)}_{G,N,j-1}\Big) \\ &\le c \sup_{x > \frac{\eta}{\beta}} |v^2(x)\omega_{\alpha+1,\beta}(x)| \sum_{\xi^{(\alpha,\beta)}_{G,N,j} > \frac{\eta}{\beta}} \frac{1}{\xi^{(\alpha,\beta)}_{G,N,j}}\Big(\xi^{(\alpha,\beta)}_{G,N,j} - \xi^{(\alpha,\beta)}_{G,N,j-1}\Big) \\ &\le c \sup_{x > \frac{\eta}{\beta}} |v^2(x)\omega_{\alpha+1,\beta}(x)| \int_{\frac{\eta}{\beta}}^{4\beta^{-1}(N+1)} \frac{1}{x}dx. \end{aligned}$$

By a similar argument as in the derivation of Lemma 2.2 of [11], we deduce that

$$\sup_{x\in\Lambda} |v^2(x)\omega_{\alpha+1,\beta}(x)| \le \max(\alpha+1, 2/\beta)\|v\|^2_{A^1_{\alpha,\beta}}. \tag{3.19}$$

Consequently,

$$B_N \le c(1+1/\beta)\ln N \|v\|^2_{A^1_{\alpha,\beta}}. \tag{3.20}$$

The combination of (3.13), (3.18), (3.20), and the fact $\|v\|_{\omega_{\alpha,\beta}} \le \|v\|_{A^1_{\alpha,\beta}}$ leads to the desired result. □

With the aid of the previous theorem, we are able to estimate the interpolation error.

Theorem 3.5. *If $v \in A^r_{\alpha-1,\beta}(\Lambda) \cap A^r_{\alpha,\beta}(\Lambda)$, then for integer $r \ge 1$,*

$$\|\mathcal{I}_{G,N,\alpha,\beta}v - v\|_{\omega_{\alpha,\beta}} \le c(\beta N)^{\frac{1}{2}-\frac{r}{2}}\Big(\beta^{-1}|v|_{A^r_{\alpha-1,\beta}} + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}|v|_{A^r_{\alpha,\beta}}\Big). \tag{3.21}$$

If, in addition, $v \in A^r_{\alpha-\mu,\beta}(\Lambda)$, then for integers $1 \le \mu \le r$,

$$\begin{aligned} |\mathcal{I}_{G,N,\alpha,\beta}v - v|_{\mu,\omega_{\alpha,\beta}} \le c(\beta N)^{\mu+\frac{1}{2}-\frac{r}{2}}\Big(&\beta^{-1}(|v|_{A^r_{\alpha-1,\beta}} + N^{-\frac{1}{2}}|v|_{A^r_{\alpha-\mu,\beta}}) \\ &+ (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}|v|_{A^r_{\alpha,\beta}}\Big). \end{aligned} \tag{3.22}$$

Proof. The use of (3.12), (3.1), and (3.4) with $\mu = 1$ leads to

$$\begin{aligned} &\|\mathcal{I}_{G,N,\alpha,\beta}v - P_{N,\alpha,\beta}v\|_{\omega_{\alpha,\beta}} = \|\mathcal{I}_{G,N,\alpha,\beta}(P_{N,\alpha,\beta}v - v)\|_{\omega_{\alpha,\beta}} \\ &\quad \le c\beta^{-1}N^{-\frac{1}{2}}|P_{N,\alpha,\beta}v - v|_{1,\omega_{\alpha,\beta}} + c(1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}\|P_{N,\alpha,\beta}v - v\|_{A^1_{\alpha,\beta}} \\ &\quad \le c(\beta N)^{\frac{1}{2}-\frac{r}{2}}\Big(\beta^{-1}|v|_{A^r_{\alpha-1,\beta}} + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}|v|_{A^r_{\alpha,\beta}}\Big). \end{aligned} \tag{3.23}$$

Thus, using the previous formula and (3.1) with $\mu = 0$ yields

$$\begin{aligned}\|\mathcal{I}_{G,N,\alpha,\beta}v - v\|_{\omega_{\alpha,\beta}} &\le \|\mathcal{I}_{G,N,\alpha,\beta}v - P_{N,\alpha,\beta}v\|_{\omega_{\alpha,\beta}} + \|P_{N,\alpha,\beta}v - v\|_{\omega_{\alpha,\beta}} \\ &\le c(\beta N)^{\frac{1}{2}-\frac{r}{2}}\Big(\beta^{-1}|v|_{A^r_{\alpha-1,\beta}} + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}|v|_{A^r_{\alpha,\beta}}\Big).\end{aligned} \tag{3.24}$$

This implies (3.21). Next, by (3.4), (3.23), and the inverse inequality as before, we deduce that

$$\begin{aligned}&|\mathcal{I}_{G,N,\alpha,\beta}v - v|_{\mu,\omega_{\alpha,\beta}} \le |\mathcal{I}_{G,N,\alpha,\beta}v - P_{N,\alpha,\beta}v|_{\mu,\omega_{\alpha,\beta}} + |P_{N,\alpha,\beta}v - v|_{\mu,\omega_{\alpha,\beta}} \\ &\quad\le c(\beta N)^{\mu}\|\mathcal{I}_{G,N,\alpha,\beta}v - P_{N,\alpha,\beta}v\|_{\omega_{\alpha,\beta}} + c\beta^{-\frac{1}{2}}(\beta N)^{\mu-\frac{r}{2}}(|v|_{A^r_{\alpha-1,\beta}} + |v|_{A^r_{\alpha-\mu,\beta}}) \\ &\quad\le c(\beta N)^{\mu+\frac{1}{2}-\frac{r}{2}}\Big(\beta^{-1}\big(|v|_{A^r_{\alpha-1,\beta}} + N^{-\frac{1}{2}}|v|_{A^r_{\alpha-\mu,\beta}}\big) + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}|v|_{A^r_{\alpha,\beta}}\Big).\end{aligned} \tag{3.25}$$

This completes the proof. □

We now turn to the generalized Laguerre–Gauss–Radau interpolation. We first study the stability of interpolation, stated later.

Theorem 3.6. *For any $v \in H^1_{\omega_{\alpha,\beta}}(\Lambda) \cap A^1_{\alpha,\beta}(\Lambda)$,*

$$\begin{aligned}\|\mathcal{I}_{R,N,\alpha,\beta}v\|_{\omega_{\alpha,\beta}} \le c\Big(&(\beta N)^{-\frac{\alpha+1}{2}}|v(0)| + \beta^{-1}N^{-\frac{1}{2}}|v|_{1,\omega_{\alpha,\beta}} \\ &+ (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}\|v\|_{A^1_{\alpha,\beta}}\Big).\end{aligned} \tag{3.26}$$

In particular, for $|\alpha| < 1$,

$$\|\mathcal{I}_{R,N,\alpha,\beta}v\|_{\omega_{\alpha,\beta}} \le c\Big(\beta^{-1}N^{-\frac{1}{2}}|v|_{1,\omega_{\alpha,\beta}} + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}}\|v\|_{A^1_{\alpha,\beta}}\Big). \tag{3.27}$$

Proof. Let η be the positive constant in (2.15). By the exactness (2.23),

$$\|\mathcal{I}_{R,N,\alpha,\beta}v\|^2_{\omega_{\alpha,\beta}} = \|\mathcal{I}_{R,N,\alpha,\beta}v\|^2_{\omega_{\alpha,\beta},R,N} = v^2(0)\omega^{(\alpha,\beta)}_{R,N,0} + \widetilde{A}_N + \widetilde{B}_N,$$

where

$$\widetilde{A}_N = \sum_{0<\xi^{(\alpha,\beta)}_{R,N,j}\le\frac{\eta}{\beta}} v^2(\xi^{(\alpha,\beta)}_{R,N,j})\omega^{(\alpha,\beta)}_{R,N,j}, \quad \widetilde{B}_N = \sum_{\xi^{(\alpha,\beta)}_{R,N,j}>\frac{\eta}{\beta}} v^2(\xi^{(\alpha,\beta)}_{R,N,j})\omega^{(\alpha,\beta)}_{R,N,j}.$$

Using the Stirling formula, we have $\omega^{(\alpha,\beta)}_{R,N,0} \le c(\beta N)^{-\alpha-1}$. On the other hand, we observe from (2.13) that the interior nodes $\xi^{(\alpha,\beta)}_{R,N,j}$, $1 \le j \le N$, satisfy asymptotic properties (2.15) and (2.16), while the corresponding weights $\omega^{(\alpha,\beta)}_{R,N,j}$, $1 \le j \le N$, fulfill (2.20) and (2.21). Thus, we can follow the same lines as in the proof of Theorem 3.4 to derive that

$$\widetilde{A}_N \le c(\|v\|^2_{\omega_{\alpha,\beta}} + \beta^{-2}N^{-1}|v|^2_{1,\omega_{\alpha,\beta}}), \quad \widetilde{B}_N \le c(1+\beta^{-1})\ln N\|v\|^2_{A^1_{\alpha,\beta}}.$$

Then the result (3.26) follows from the previous statements.

We next prove (3.27). For any $x \in [0, \frac{1}{\beta}]$ and $|\alpha| < 1$,

$$|v(x) - v(0)| \le \left(\int_0^{\frac{1}{\beta}} x^{-\alpha}e^{\beta x}\,dx\right)^{\frac{1}{2}}\|\partial_x v\|_{L^2_{\omega_{\alpha,\beta}}(0,\frac{1}{\beta})} \le c\beta^{\frac{\alpha-1}{2}}\|\partial_x v\|_{L^2_{\omega_{\alpha,\beta}}(0,\frac{1}{\beta})}. \tag{3.28}$$

Now, let $|v(x_*)| = \min_{x\in[0,1/\beta]} |v(x)|$. Clearly, for $|\alpha| < 1$,

$$|v(x_*)| \le \beta \int_0^{\frac{1}{\beta}} |v(x)|\, dx \le c\beta^{\frac{\alpha+1}{2}} \|v\|_{L^2_{\omega_{\alpha,\beta}}(0,\frac{1}{\beta})}. \tag{3.29}$$

The previous formula with (3.28) gives

$$|v(0)| \le |v(x_*)| + |v(x_*) - v(0)| \le c\Big(\beta^{\frac{\alpha+1}{2}} \|v\|_{\omega_{\alpha,\beta}} + \beta^{\frac{\alpha-1}{2}} |v|_{1,\omega_{\alpha,\beta}}\Big). \tag{3.30}$$

If $0 \le \alpha < 1$, then by (3.30) and the fact $\|v\|_{\omega_{\alpha,\beta}} \le \|v\|_{A^1_{\alpha,\beta}}$, we derive that

$$(\beta N)^{-\frac{\alpha+1}{2}} |v(0)| \le c\Big(\beta^{-1} N^{-\frac{1}{2}} |v|_{1,\omega_{\alpha,\beta}} + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}} \|v\|_{A^1_{\alpha,\beta}}\Big),$$

which, along with (3.26), leads to (3.27) with $0 \le \alpha < 1$. For $-1 < \alpha < 0$, we change slightly the derivation of (3.28) to obtain that for any $x \in [0, \frac{1}{\beta}]$,

$$|v(x) - v(0)| \le \int_0^{\frac{1}{\beta}} |\partial_x v(x)|\, dx \le c\beta^{\frac{\alpha}{2}} \|\partial_x v\|_{L^2_{\omega_{\alpha+1,\beta}}(0,\frac{1}{\beta})}.$$

Correspondingly, (3.30) becomes

$$|v(0)| \le c\Big(\beta^{\frac{\alpha+1}{2}} \|v\|_{\omega_{\alpha,\beta}} + \beta^{\frac{\alpha}{2}} \|\partial_x v\|_{\omega_{\alpha+1,\beta}}\Big) \le c\beta^{\frac{\alpha+1}{2}} (1+\beta^{-\frac{1}{2}}) \|v\|_{A^1_{\alpha,\beta}}.$$

Then the result (3.27) with $-1 < \alpha < 0$ follows from formula (3.26) and the fact $N^{-\frac{\alpha+1}{2}} \le c$. □

The following two theorems describe the error of interpolation $\mathcal{I}_{R,N,\alpha,\beta} v$.

THEOREM 3.7. *If $v \in A^r_{\alpha,\beta}(\Lambda) \cap A^r_{\alpha-1,\beta}(\Lambda)$, then for an integer $r \ge 1$ and $r > \alpha + 1$,*

$$\|\mathcal{I}_{R,N,\alpha,\beta} v - v\|_{\omega_{\alpha,\beta}} \le c(\beta N)^{\frac{1}{2}-\frac{r}{2}} \Big(\beta^{-1} |v|_{A^r_{\alpha-1,\beta}} + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}} |v|_{A^r_{\alpha,\beta}}\Big). \tag{3.31}$$

In particular, if $|\alpha| < 1$, then the previous result holds for all integers $r \ge 1$.

Proof. As a consequence of (3.26),

$$\begin{aligned} \|\mathcal{I}_{R,N,\alpha,\beta} v - P_{N,\alpha,\beta} v\|_{\omega_{\alpha,\beta}} &= \|\mathcal{I}_{R,N,\alpha,\beta}(P_{N,\alpha,\beta} v - v)\|_{\omega_{\alpha,\beta}} \\ &\le c(\beta N)^{-\frac{\alpha+1}{2}} |P_{N,\alpha,\beta} v(0) - v(0)| + c\beta^{-1} N^{-\frac{1}{2}} |P_{N,\alpha,\beta} v - v|_{1,\omega_{\alpha,\beta}} \\ &\quad + c(1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}} \|P_{N,\alpha,\beta} v - v\|_{A^1_{\alpha,\beta}}. \end{aligned} \tag{3.32}$$

According to Lemma 3.2, the first term on the right-hand side of (3.32) is bounded above by $c(\beta N)^{-\frac{r}{2}} |v|_{A^r_{\alpha,\beta}}$ for an integer $r > \alpha + 1$. The other two terms can be estimated by using Lemmas 3.1 and 3.3 with $\mu = 1$ (cf. the proof of (3.21)).

If $|\alpha| < 1$, we use (3.27) to derive (3.32), which does not contain the term $|P_{N,\alpha,\beta} v(0) - v(0)|$, and consequently does not require $r > \alpha + 1$. □

We can follow the same approach as for the proof of (3.22) to derive the following result.

THEOREM 3.8. *If $v \in A^r_{\alpha,\beta}(\Lambda) \cap A^r_{\alpha-1,\beta}(\Lambda) \cap A^{r-1}_{\alpha-\mu,\beta}(\Lambda)$, then for integers $1 \le \mu \le r$ and $r > \alpha + 1$,*

$$\begin{aligned} |\mathcal{I}_{R,N,\alpha,\beta} v - v|_{\mu,\omega_{\alpha,\beta}} \le & c(\beta N)^{\mu+\frac{1}{2}-\frac{r}{2}} \Big(\beta^{-1}\Big(|v|_{A^r_{\alpha-1,\beta}} + N^{-\frac{1}{2}} |v|_{A^r_{\alpha-\mu,\beta}}\Big) \\ & + (1+\beta^{-\frac{1}{2}})(\ln N)^{\frac{1}{2}} |v|_{A^r_{\alpha,\beta}}\Big). \end{aligned} \tag{3.33}$$

In particular, for $|\alpha| < 1$, the previous result holds for all integers $r \ge 1$.

4. Generalized Laguerre pseudospectral method for unbounded domains. This section is devoted to the generalized pseudospectral method based on the generalized Laguerre–Gauss interpolation. Throughout this section, let $\Omega = \Lambda \times S$ and S be the unit spherical surface, $S = \{(\lambda, \theta) | \ 0 \le \lambda < 2\pi, \ \ -\frac{\pi}{2} \le \theta < \frac{\pi}{2}\}$. The Laplacian operator on Ω is given by

$$\Delta v(\rho,\lambda,\theta) = \frac{1}{\rho^2}\partial_\rho(\rho^2\partial_\rho v(\rho,\lambda,\theta)) + \frac{1}{\rho^2\cos\theta}\partial_\theta(\cos\theta\partial_\theta v(\rho,\lambda,\theta)) + \frac{1}{\rho^2\cos^2\theta}\partial_\lambda^2 v(\rho,\lambda,\theta).$$

We consider the following problem:

$$\begin{cases} -\Delta W(\rho,\lambda,\theta) + \mu W(\rho,\lambda,\theta) = F(\rho,\lambda,\theta), \quad \mu > 0, \quad \text{in } \Omega, \\ W(\rho,\lambda+2\pi,\theta) = W(\rho,\lambda,\theta). \end{cases} \tag{4.1}$$

Here, we look for the solution of (4.1) such that $\rho^{\frac{1}{2}} W(\rho,\lambda,\theta) \to 0$ as $\rho \to 0$ and $\rho^{\frac{3}{2}} W(\rho,\lambda,\theta) \to 0$ as $\rho \to \infty$. In addition, the solution $W(\rho,\lambda,\theta)$ satisfies the pole condition, namely, $\partial_\lambda W(\rho,\lambda,\theta) = 0$ for $\theta = \pm\frac{\pi}{2}$.

It is noted that the usual weighted (with the weight $e^{-\beta\rho}$) Galerkin variational formulation of (4.1), on which the generalized Laguerre approximations are often based, is not well posed. One possible way to remedy this deficiency is to find a suitable variable transform such that the weighted variational formulation of the transformed equation becomes well posed. Motivated by [10], we make the variable transform

$$W(\rho,\lambda,\theta) = e^{-\frac{\beta}{2}\rho} U(\rho,\lambda,\theta), \quad F(\rho,\lambda,\theta) = e^{-\frac{\beta}{2}\rho} f(\rho,\lambda,\theta), \tag{4.2}$$

which converts (4.1) into

$$\begin{aligned} &-\partial_\rho^2 U(\rho,\lambda,\theta) - \frac{1}{\rho}(2-\beta\rho)\partial_\rho U(\rho,\lambda,\theta) - \frac{1}{\rho^2\cos\theta}\partial_\theta(\cos\theta\partial_\theta U(\rho,\lambda,\theta)) \\ &\quad - \frac{1}{\rho^2\cos^2\theta}\partial_\lambda^2 U(\rho,\lambda,\theta) + \frac{1}{\rho}\Big(\mu\rho + \beta - \frac{\beta^2}{4}\rho\Big) U(\rho,\lambda,\theta) = f(\rho,\lambda,\theta). \end{aligned} \tag{4.3}$$

To focus on our main idea, we consider only the spherically symmetric case, in which U and f are independent of λ and θ, denoted by $U(\rho)$ and $f(\rho)$, respectively. Accordingly,

$$-\partial_\rho^2 U(\rho) - \frac{1}{\rho}(2-\beta\rho)\partial_\rho U(\rho) + \frac{1}{\rho}\Big(\mu\rho + \beta - \frac{\beta^2}{4}\rho\Big) U(\rho) = f(\rho). \tag{4.4}$$

In addition, $\rho^{\frac{1}{2}} U(\rho) \to 0$ as $\rho \to 0$ and $\rho^{\frac{3}{2}} e^{-\frac{\beta}{2}\rho} U(\rho)$ as $\rho \to \infty$.

With the previous general setup, we now derive a weak formulation of (4.4). First, we observe that for any $v \in H^1_{\omega_{2,\beta}}(\Lambda)$, we have $\partial_\rho v(\rho) = o(\rho^{-\frac{3}{2}})$ and $v(\rho) = o(\rho^{-\frac{1}{2}})$ as $\rho \to 0$, and $\partial_\rho v(\rho) \sim v(\rho) = o(\rho^{-\frac{3}{2}} e^{\frac{\beta\rho}{2}})$ as $\rho \to \infty$. Consequently, if $v \in H^1_{\omega_{2,\beta}}(\Lambda) \cap L^2_{\omega_{1,\beta}}(\Lambda)$, then $\rho^2 v(\rho)\partial_\rho v(\rho) e^{-\beta\rho} \to 0$ as $\rho \to 0, \infty$. Hence, we obtain a weak formulation of (4.4). It is to find $U \in H^1_{\omega_{2,\beta}}(\Lambda) \cap L^2_{\omega_{1,\beta}}(\Lambda)$ such that

$$a_{\mu,\beta}(U,v) = (f,v)_{\omega_{2,\beta}} \quad \forall v \in H^1_{\omega_{2,\beta}}(\Lambda) \cap L^2_{\omega_{1,\beta}}(\Lambda), \tag{4.5}$$

where the bilinear form is defined by

$$a_{\mu,\beta}(u,v) = (\partial_\rho u, \partial_\rho v)_{\omega_{2,\beta}} + \Big(\mu - \frac{\beta^2}{4}\Big)(u,v)_{\omega_{2,\beta}} + \beta(u,v)_{\omega_{1,\beta}}.$$

One can verify that $a_{\mu,\beta}(\cdot,\cdot)$ is continuous and elliptic in $\big(H^1_{\omega_{2,\beta}}(\Lambda)\cap L^2_{\omega_{1,\beta}}(\Lambda)\big)^2$. Indeed,

$$|a_{\mu,\beta}(u,v)|\le c\big((1+\beta)\|u\|_{1,\omega_{2,\beta}}+\beta^{\frac12}\|u\|_{\omega_{1,\beta}}\big)\big((1+\beta)\|v\|_{1,\omega_{2,\beta}}+\beta^{\frac12}\|v\|_{\omega_{1,\beta}}\big),\tag{4.6}$$

and for $\mu>\frac{\beta^2}{4}$, we have

$$a_{\mu,\beta}(v,v)\ge c\big(\|v\|^2_{1,\omega_{2,\beta}}+\beta\|v\|^2_{\omega_{1,\beta}}\big).\tag{4.7}$$

Therefore, if $f\in(H^1_{\omega_{2,\beta}}(\Lambda)\cap L^2_{\omega_{1,\beta}}(\Lambda))'$, then (4.5) admits a unique solution.

The corresponding pseudospectral scheme for (4.5) is to seek $u_N(\rho)\in\mathbb{P}_N$ such that

$$a_{\mu,\beta,N}(u_N,\phi)=(f,\phi)_{\omega_{2,\beta},G,N}\quad\forall\phi\in\mathbb{P}_N,\tag{4.8}$$

where

$$a_{\mu,\beta,N}(u,v)=(\partial_\rho u,\partial_\rho v)_{\omega_{2,\beta},G,N}+\Big(\mu-\frac{\beta^2}{4}\Big)(u,v)_{\omega_{2,\beta},G,N}+\beta(u,v)_{\omega_{1,\beta},G,N}.$$

According to (2.22), (4.8) is equivalent to

$$a_{\mu,\beta}(u_N,\phi)=(\mathcal{I}_{G,N,2,\beta}f,\phi)_{\omega_{2,\beta}}\quad\forall\phi\in\mathbb{P}_N.\tag{4.9}$$

Before analyzing the convergence of (4.8), we first consider a special orthogonal projection $P^1_{N,\beta}:H^1_{\omega_{2,\beta}}(\Lambda)\cap L^2_{\omega_{1,\beta}}(\Lambda)\to\mathbb{P}_N$, defined by

$$\big(\partial_\rho(P^1_{N,\beta}v-v),\partial_\rho\phi\big)_{\omega_{2,\beta}}+\big(P^1_{N,\beta}v-v,\phi\big)_{\omega_{2,\beta}}+\big(P^1_{N,\beta}v-v,\phi\big)_{\omega_{1,\beta}}=0\quad\forall\phi\in\mathbb{P}_N.\tag{4.10}$$

To analyze its approximation error, we need the following two imbedding inequalities which are the special cases of Lemmas 2.1 and 2.2 of [12].

- If $v\in L^2_{\omega_{0,\beta}}(\Lambda)$, $\partial_\rho v\in L^2_{\omega_{2,\beta}}(\Lambda)$, and $v(\frac{1}{\beta})=0$, then

$$\|v\|_{\omega_{0,\beta}}\le c\|\partial_\rho v\|_{\omega_{2,\beta}}.\tag{4.11}$$

- If $v\in H^1_{\omega_{2,\beta}}(\Lambda)\cap L^2_{\omega_{0,\beta}}(\Lambda)$, then

$$\|v\|^2_{\omega_{2,\beta}}\le 8\beta^{-2}\big(\|\partial_\rho v\|^2_{\omega_{2,\beta}}+\|v\|^2_{\omega_{0,\beta}}\big).\tag{4.12}$$

LEMMA 4.1. *For any $v\in H^1_{\omega_{2,\beta}}(\Lambda)\cap L^2_{\omega_{1,\beta}}(\Lambda)\cap A^r_{1,\beta}(\Lambda)$ and integer $r\ge1$,*

$$\|P^1_{N,\beta}v-v\|_{1,\omega_{2,\beta}}+\|P^1_{N,\beta}v-v\|_{\omega_{1,\beta}}\le c(1+\beta^{-1})(\beta N)^{\frac{1-r}{2}}|v|_{A^r_{1,\beta}}.\tag{4.13}$$

Proof. By projection theorem and the Cauchy–Schwarz inequality,

$$\begin{aligned}\|P^1_{N,\beta}v-v\|^2_{1,\omega_{2,\beta}}+\|P^1_{N,\beta}v-v\|^2_{\omega_{1,\beta}}&\le\|\phi-v\|^2_{1,\omega_{2,\beta}}+\|\phi-v\|^2_{\omega_{1,\beta}}\\&\le|\phi-v|^2_{1,\omega_{2,\beta}}+\frac32\|\phi-v\|^2_{\omega_{2,\beta}}+\frac12\|\phi-v\|^2_{\omega_{0,\beta}}\quad\forall\phi\in\mathbb{P}_N.\end{aligned}$$

Taking $\phi(\rho)=P_{N,1,\beta}v(\rho)-P_{N,1,\beta}v(\frac1\beta)+v(\frac1\beta)$, we have from (4.11), (4.12), and (3.1) that

$$\begin{aligned}\|P^1_{N,\beta}v-v\|^2_{1,\omega_{2,\beta}}+\|P^1_{N,\beta}v-v\|^2_{\omega_{1,\beta}}&\le c(1+\beta^{-2})\|\partial_\rho(\phi-v)\|^2_{\omega_{2,\beta}}\\&=c(1+\beta^{-2})|P_{N,1,\beta}v-v|^2_{A^1_{1,\beta}}\le c(1+\beta^{-2})(\beta N)^{1-r}|v|^2_{A^r_{1,\beta}}.\end{aligned}$$ □

We now go back to the convergence analysis of scheme (4.8). Let $U_N = P^1_{N,\beta}U$; then by (4.5) and (4.10),

$$a_{\mu,\beta}(U_N, \phi) = -G(\phi) + (\mathcal{I}_{G,N,2,\beta}f, \phi)_{\omega_{2,\beta}} \quad \forall \phi \in \mathbb{P}_N, \tag{4.14}$$

where

$$G(\phi) = \Big(\mu - \frac{\beta^2}{4} - 1\Big)(U - U_N, \phi)_{\omega_{2,\beta}} + (\beta - 1)(U - U_N, \phi)_{\omega_{1,\beta}} + (\mathcal{I}_{G,N,2,\beta}f - f, \phi)_{\omega_{2,\beta}}.$$

Set $\widetilde{U}_N = u_N - U_N$. Then by (4.9) and (4.14),

$$a_{\mu,\beta}(\widetilde{U}_N, \phi) = G(\phi) \quad \forall \phi \in \mathbb{P}_N. \tag{4.15}$$

Taking $\phi = \widetilde{U}_N$ in the previous formula and using (4.7) give

$$\|\widetilde{U}_N\|^2_{1,\omega_{2,\beta}} + \beta\|\widetilde{U}_N\|^2_{\omega_{1,\beta}} \le c|G(\widetilde{U}_N)|. \tag{4.16}$$

Hence, it suffices to estimate $|G(\widetilde{U}_N)|$. For simplicity, we shall use the following notation:

$$\begin{aligned}
B^{(1)}_{N,\beta,r}(v) &= c(1+\beta^2)^2(1+\beta^{-1})^2(\beta N)^{1-r}|v|^2_{A^r_{1,\beta}},\\
B^{(2)}_{N,\beta,r}(v) &= c(1+\beta)^2(1+\beta^{-1})^2(\beta N)^{1-r}|v|^2_{A^r_{1,\beta}},\\
B^{(3)}_{N,\beta,s}(v) &= c(\beta N)^{1-s}\Big(\beta^{-2}|v|^2_{A^{s-1}_{1,\beta}} + (1+\beta^{-1})\ln N|v|^2_{A^s_{2,\beta}}\Big).
\end{aligned}$$

By virtue of (4.13) and (3.21), for integers $r, s \ge 1$,

$$|G(\widetilde{U}_N)| \le B^{(1)}_{N,\beta,r}(U) + B^{(2)}_{N,\beta,r}(U) + B^{(3)}_{N,\beta,s}(f) + \frac{1}{2}\|\widetilde{U}_N\|^2_{\omega_{2,\beta}} + \frac{\beta}{2}\|\widetilde{U}_N\|^2_{\omega_{1,\beta}}.$$

Plugging the previous formula into (4.16) leads to an estimate for $\|\widetilde{U}_N\|^2_{1,\omega_{2,\beta}} + \beta\|\widetilde{U}_N\|^2_{\omega_{1,\beta}}$. Since $U - u_N = U - P^1_{N,\beta}U - \widetilde{U}_N$, we use (4.13) again to reach the following conclusion.

Theorem 4.2. *Let U and u_N be the solutions of* (4.5) *and* (4.8), *respectively, and let $\mu > \frac{1}{4}\beta^2$. If $U \in A^r_{1,\beta}(\Lambda)$ and $f \in A^s_{1,\beta}(\Lambda) \cap A^s_{2,\beta}(\Lambda)$ with integers $r, s \ge 1$, then*

$$\|U - u_N\|^2_{1,\omega_{2,\beta}} + \beta\|U - u_N\|^2_{\omega_{1,\beta}} \le c\Big(B^{(1)}_{N,\beta,r}(U) + B^{(2)}_{N,\beta,r}(U) + B^{(3)}_{N,\beta,s}(f)\Big).$$

Remark 4.1. *After solving $u_N(\rho)$ from* (4.8), *we evaluate the numerical solution of the original problem by $w_N(\rho) = e^{-\frac{\beta}{2}\rho}u_N(\rho)$. Indeed, a direct computation leads to*

$$\begin{aligned}
\|W - w_N\|_{1,\hat{\omega}_2} + \sqrt{\beta}\|W - w_N\|_{\hat{\omega}_1} &\le (1+\beta)\|U - u_N\|_{1,\omega_{2,\beta}} + \sqrt{\beta}\|U - u_N\|_{\omega_{1,\beta}}\\
&= \mathcal{O}(N^{\frac{1-r}{2}} + (\ln N)^{\frac{1}{2}}N^{\frac{1-s}{2}}),
\end{aligned}$$

where $\hat{\omega}_\alpha(\rho) = \rho^\alpha = \omega_{\alpha,0}(\rho)$. A combination of the previous formula and (3.19) *with $\alpha = 1$ yields*

$$\begin{aligned}
\sup_{\rho\in\Lambda}|\rho(W - w_N)| &= \sup_{\rho\in\Lambda}|(U - u_N)\rho e^{-\frac{\beta}{2}\rho}| \le c\|U - u_N\|_{A^1_{1,\beta}}\\
&\le c(|U - u_N|_{1,\omega_{2,\beta}} + \|U - u_N\|_{\omega_{1,\beta}}) = \mathcal{O}(N^{\frac{1-r}{2}} + (\ln N)^{\frac{1}{2}}N^{\frac{1-s}{2}}).
\end{aligned}$$

Hence, a spectral accuracy is expected from theoretical analysis.

Remark 4.2. *Given $\mu > 0$, we can always choose the adjustable factor β such that $\mu > \frac{1}{4}\beta^2$, which guarantees the well-posedness of our Galerkin formulation.*

5. Generalized Laguerre pseudospectral method for exterior problems. This section is for the generalized Laguerre pseudospectral method based on Gauss–Radau interpolation for exterior problems. As an example, we consider the following equation induced by the spherically symmetric solution of the three-dimensional problem:

$$\begin{cases} -\dfrac{1}{\rho^2}\partial_\rho(\rho^2\partial_\rho W(\rho)) + \mu W(\rho) = F(\rho), \quad \mu > 0, \quad \rho > 1, \\ \lim\limits_{\rho\to\infty} \rho^{\frac{3}{2}} W(\rho) = 0, \quad W(1) = g. \end{cases} \tag{5.1}$$

For simplicity, let $g = 0$. We first shift the interval $[1, \infty)$ to $[0, \infty)$ by using the variable transform: $\rho = x + 1$, $W(\rho) = V(x)$, $F(\rho) = G(x)$. Then (5.1) becomes

$$\begin{cases} -\dfrac{1}{(x+1)^2}\partial_x((x+1)^2\partial_x V(x)) + \mu V(x) = G(x), \quad \mu > 0, \quad x > 0, \\ \lim\limits_{x\to\infty} x^{\frac{3}{2}} V(x) = V(0) = 0. \end{cases} \tag{5.2}$$

As mentioned earlier, it is necessary to make the following transformation:

$$V(x) = e^{-\frac{\beta}{2}x}U(x), \quad G(x) = (x+1)^{-2}e^{-\frac{\beta}{2}x}f(x).$$

Then (5.2) is rewritten as

$$\begin{cases} -\partial_x^2 U(x) - \dfrac{1}{x+1}(2 - \beta(x+1))\partial_x U(x) + \dfrac{1}{x+1}\Big(\Big(\mu - \dfrac{1}{4}\beta^2\Big)(x+1) + \beta\Big)U(x) \\ \quad = \dfrac{1}{(x+1)^2} f(x), \quad \mu > 0, \quad x > 0, \\ \lim\limits_{x\to\infty} x^{\frac{3}{2}} e^{-\frac{\beta}{2}x} U(x) = U(0) = 0. \end{cases} \tag{5.3}$$

Now, let $\sigma_{\alpha,\beta}(x) = (x+1)^\alpha e^{-\beta x}$, and denote ${}_0H^1_{\sigma_{2,\beta}}(\Lambda) := \{v \in H^1_{\sigma_{2,\beta}}(\Lambda) : u(0) = 0\}$. A weak form of (5.3) is to find $U \in {}_0H^1_{\sigma_{2,\beta}}(\Lambda) \cap L^2_{\sigma_{1,\beta}}(\Lambda)$ such that

$$\widetilde{a}_{\mu,\beta}(U, v) = (f, v)_{\omega_{0,\beta}} \quad \forall v \in {}_0H^1_{\sigma_{2,\beta}}(\Lambda) \cap L^2_{\sigma_{1,\beta}}(\Lambda), \tag{5.4}$$

where the bilinear form

$$\widetilde{a}_{\mu,\beta}(u, v) = (\partial_x u, \partial_x v)_{\sigma_{2,\beta}} + \Big(\mu - \frac{1}{4}\beta^2\Big)(u, v)_{\sigma_{2,\beta}} + \beta(u, v)_{\sigma_{1,\beta}}.$$

One can verify readily that

$$|\widetilde{a}_{\mu,\beta}(u, v)| \le c\big((1+\beta)\|u\|_{1,\sigma_{2,\beta}} + \beta^{\frac{1}{2}}\|u\|_{\sigma_{1,\beta}}\big)\big((1+\beta)\|v\|_{1,\sigma_{2,\beta}} + \beta^{\frac{1}{2}}\|v\|_{\sigma_{1,\beta}}\big), \tag{5.5}$$

and for $\mu > \frac{1}{4}\beta^2$,

$$|\widetilde{a}_{\mu,\beta}(v, v)| \ge c\big(\|v\|^2_{1,\sigma_{2,\beta}} + \beta\|v\|^2_{\sigma_{1,\beta}}\big). \tag{5.6}$$

Hence, if $f \in (H^1_{\sigma_{2,\beta}}(\Lambda) \cap L^2_{\sigma_{1,\beta}}(\Lambda))'$, then (5.4) has a unique solution.

The generalized Laguerre pseudospectral scheme for (5.4) is to seek $u_N \in {}_0\mathbb{P}_N := \{u \in \mathbb{P}_N : u(0) = 0\}$ such that

$$\widetilde{a}_{\mu,\beta,N}(u_N, \phi) = (f, \phi)_{\omega_{0,\beta},R,N} \quad \forall \phi \in {}_0\mathbb{P}_N, \tag{5.7}$$

where

$$\begin{aligned}\widetilde{a}_{\mu,\beta,N}(u,v) &= (\partial_x u, \partial_x v)_{\omega_{2,\beta},R,N} + 2(\partial_x u, \partial_x v)_{\omega_{1,\beta},R,N} + (\partial_x u, \partial_x v)_{\omega_{0,\beta},R,N}\\ &+\Big(\mu - \frac{1}{4}\beta^2\Big)(u,v)_{\omega_{2,\beta},R,N} + \Big(2\mu - \frac{1}{2}\beta^2 + \beta\Big)(u,v)_{\omega_{1,\beta},R,N}\\ &+\Big(\mu - \frac{1}{4}\beta^2 + \beta\Big)(u,v)_{\omega_{0,\beta},R,N}.\end{aligned}$$

According to (2.22), (5.7) is equivalent to

$$\widetilde{a}_{\mu,\beta}(u_N, \phi) = (\mathcal{I}_{R,N,0,\beta} f, \phi)_{\omega_{0,\beta}} \quad \forall \phi \in {}_0\mathbb{P}_N. \tag{5.8}$$

5.1. A specific orthogonal projection. We next consider a specific orthogonal projection that will be used in numerical analysis of generalized Laguerre pseudospectral method for exterior problems. Let ${}_0H^1_{\omega_{\alpha,\beta}}(\Lambda) := \{v \in H^1_{\omega_{\alpha,\beta}}(\Lambda) : v(0) = 0\}$. Note that $H^1_{\omega_{2,\beta}}(\Lambda) \cap H^1_{\omega_{0,\beta}}(\Lambda) \subseteq H^1_{\sigma_{2,\beta}}(\Lambda) \cap L^2_{\sigma_{1,\beta}}(\Lambda)$. The orthogonal projection ${}_0\Pi^1_{N,\beta} : {}_0H^1_{\omega_{2,\beta}}(\Lambda) \cap H^1_{\omega_{0,\beta}}(\Lambda) \to {}_0\mathbb{P}_N$ is defined by

$$\big(\partial_x({}_0\Pi^1_{N,\beta} v - v), \partial_x \phi\big)_{\sigma_{2,\beta}} + \big({}_0\Pi^1_{N,\beta} v - v, \phi\big)_{\sigma_{2,\beta}} = 0 \quad \forall \phi \in {}_0\mathbb{P}_N. \tag{5.9}$$

In order to analyze approximation error of the previous projection, we need another auxiliary orthogonal projection. To do this, we introduce the space $H^1_{\omega_{2,\beta},\omega_{0,\beta}}(\Lambda)$, equipped with the norm $\|v\|_{1,\omega_{2,\beta},\omega_{0,\beta}} = (\|\partial_x v\|^2_{\omega_{2,\beta}} + \|v\|^2_{\omega_{0,\beta}})^{\frac{1}{2}}$.

The orthogonal projection $\widetilde{P}^1_{N,\beta} : H^1_{\omega_{2,\beta},\omega_{0,\beta}}(\Lambda) \to \mathbb{P}_N$ is defined by

$$\big(\partial_x(\widetilde{P}^1_{N,\beta} v - v), \partial_x \phi\big)_{\omega_{2,\beta}} + \big(\widetilde{P}^1_{N,\beta} v - v, \phi\big)_{\omega_{0,\beta}} = 0 \quad \forall \phi \in \mathbb{P}_N. \tag{5.10}$$

LEMMA 5.1. *If $v \in H^1_{\omega_{2,\beta},\omega_{0,\beta}}(\Lambda) \cap A^r_{1,\beta}(\Lambda)$ and an integer $r \geq 1$, then*

$$\|\widetilde{P}^1_{N,\beta} v - v\|_{1,\omega_{2,\beta},\omega_{0,\beta}} \leq c(\beta N)^{\frac{1-r}{2}} |v|_{A^r_{1,\beta}}.$$

Proof. By projection theorem,

$$\|\widetilde{P}^1_{N,\beta} v - v\|_{1,\omega_{2,\beta},\omega_{0,\beta}} \leq \|\phi - v\|_{1,\omega_{2,\beta},\omega_{0,\beta}} \quad \forall \phi \in \mathbb{P}_N.$$

We take $\phi(x) = P_{N,1,\beta} v(x) - P_{N,1,\beta} v(\frac{1}{\beta}) + v(\frac{1}{\beta})$. Then by (4.11) and (3.1),

$$\|\phi - v\|_{1,\omega_{2,\beta},\omega_{0,\beta}} \leq c\|\partial_x(\phi - v)\|_{\omega_{2,\beta}} = c|P_{N,1,\beta} v - v|_{A^1_{1,\beta}} \leq c(\beta N)^{\frac{1-r}{2}} |v|_{A^r_{1,\beta}}.$$

This completes the proof. □

We are ready to estimate $\|{}_0\Pi^1_{N,\beta} v - v\|_{1,\sigma_{2,\beta}}$. We shall use the fact that for $v \in H^1_{\omega_{\alpha,\beta}}(\Lambda)$, $v(0) = 0$ and $\alpha < 1$, we have (see Lemma 2.2 of [12])

$$\|v\|_{\omega_{\alpha,\beta}} \leq c\beta^{-1} \|\partial_x v\|_{\omega_{\alpha,\beta}}. \tag{5.11}$$

LEMMA 5.2. *For any $v \in A^r_{0,\beta}(\Lambda)$ with $v(0) = 0$ and an integer $r \geq 2$,*

$$\|{}_0\Pi^1_{N,\beta} v - v\|_{1,\sigma_{2,\beta}} \leq c(1 + \beta^{-2})(\beta N)^{1-\frac{r}{2}} |v|_{A^r_{0,\beta}}.$$

Proof. By projection theorem,

$$\|{}_0\Pi^1_{N,\beta} v - v\|_{1,\sigma_{2,\beta}} \leq \|\phi - v\|_{1,\sigma_{2,\beta}} \leq c(\|\phi - v\|_{1,\omega_{2,\beta}} + \|\phi - v\|_{1,\omega_{0,\beta}}) \quad \forall \phi \in {}_0\mathbb{P}_N.$$

Taking

$$\phi(x) = \int_0^x \widetilde{P}^1_{N-1,\beta}\partial_\xi v(\xi)\,d\xi \in {}_0\mathbb{P}_N,$$

we have from Lemma 5.1 that

$$\|\partial_x(\phi - v)\|_{\omega_{0,\beta}} = \|\widetilde{P}^1_{N-1,\beta}\partial_x v - \partial_x v\|_{\omega_{0,\beta}} \le c(\beta N)^{1-\frac{r}{2}}|v|_{A^r_{0,\beta}}, \tag{5.12}$$

which, along with (5.11) with $\alpha = 0$, gives

$$\|\phi - v\|_{\omega_{0,\beta}} \le c\beta^{-1}\|\partial_x(\phi - v)\|_{\omega_{0,\beta}} \le c\beta^{-1}(\beta N)^{1-\frac{r}{2}}|v|_{A^r_{0,\beta}}. \tag{5.13}$$

Moreover, thanks to (4.12) with $\alpha = 2$, we have from Lemma 5.1 that

$$\begin{aligned}\|\partial_x(\phi - v)\|_{\omega_{2,\beta}} &= \|\widetilde{P}^1_{N-1,\beta}\partial_x v - \partial_x v\|_{\omega_{2,\beta}} \\ &\le c\beta^{-1}(\|\partial_x(\widetilde{P}^1_{N-1,\beta}\partial_x v - \partial_x v)\|_{\omega_{2,\beta}} + \|\widetilde{P}^1_{N-1,\beta}\partial_x v - \partial_x v\|_{\omega_{0,\beta}}) \\ &\le c\beta^{-1}(\beta N)^{1-\frac{r}{2}}|v|_{A^r_{0,\beta}}.\end{aligned} \tag{5.14}$$

Furthermore, using (4.12), (5.13), and (4.13) leads to

$$\|\phi - v\|_{\omega_{2,\beta}} \le c\beta^{-1}(\|\partial_x(\phi - v)\|_{\omega_{2,\beta}} + \|\phi - v\|_{\omega_{0,\beta}}) \le c\beta^{-2}(\beta N)^{1-\frac{r}{2}}|v|_{A^r_{0,\beta}}. \tag{5.15}$$

Finally, a combination of (5.12)–(5.15) leads to the desired result. □

5.2. Convergence analysis. Let $U_N = {}_0\Pi^1_{N,\beta}U$. Then by (5.4) and (5.9),

$$\widetilde{a}_{\mu,\beta}(U_N, \phi) = -\widetilde{G}(\phi) + (\mathcal{I}_{R,N,0,\beta}f, \phi)_{\omega_{0,\beta}} \quad \forall \phi \in {}_0\mathbb{P}_N, \tag{5.16}$$

where

$$\widetilde{G}(\phi) = \Big(\mu - \frac{1}{4}\beta^2 - 1\Big)(U - U_N, \phi)_{\sigma_{2,\beta}} + \beta(U - U_N, \phi)_{\sigma_{1,\beta}} + (\mathcal{I}_{R,N,0,\beta}f - f, \phi)_{\omega_{0,\beta}}.$$

Set $\widetilde{U}_N = u_N - U_N$. Then subtracting (5.16) from (5.8) yields

$$\widetilde{a}_{\mu,\beta}(\widetilde{U}_N, \phi) = \widetilde{G}(\phi) \quad \forall \phi \in {}_0\mathbb{P}_N. \tag{5.17}$$

Taking $\phi = \widetilde{U}_N$ in the previous formula and using (5.6), we obtain

$$\|\widetilde{U}_N\|^2_{1,\sigma_{2,\beta}} + \beta\|\widetilde{U}_N\|^2_{\sigma_{1,\beta}} \le c|\widetilde{G}(\widetilde{U}_N)|. \tag{5.18}$$

Thus, it suffices to estimate $|\widetilde{G}(\widetilde{U}_N)|$. For simplicity, we will use the following notation:

$$\begin{aligned}\widetilde{B}^{(1)}_{N,\beta,r}(v) &= c(\beta^2 + 1)^2(1 + \beta^{-2})^2(\beta N)^{2-r}|v|^2_{A^r_{0,\beta}}, \\ \widetilde{B}^{(2)}_{N,\beta,r}(v) &= c(1 + \beta^{-2})^2\beta(\beta N)^{2-r}|v|^2_{A^r_{0,\beta}}, \\ \widetilde{B}^{(3)}_{N,\beta,s}(v) &= c\beta^{-1}(\beta N)^{1-s}\Big(\beta^{-2}\|\partial_x^s v\|^2_{\omega_{s-1,\beta}} + (1 + \beta^{-1})\ln N|v|^2_{A^s_{0,\beta}}\Big).\end{aligned}$$

By virtue of Theorem 3.7 and Lemma 5.2, for integers $r \ge 2$ and $s \ge 1$,

$$|\widetilde{G}(\widetilde{U}_N)| \le \widetilde{B}^{(1)}_{N,\beta,r}(U) + \widetilde{B}^{(2)}_{N,\beta,r}(U) + \widetilde{B}^{(3)}_{N,\beta,s}(f) + \frac{1}{2}\|\widetilde{U}_N\|^2_{\sigma_{2,\beta}} + \frac{\beta}{2}\|\widetilde{U}_N\|^2_{\sigma_{1,\beta}}.$$

Plugging the previous formula into (5.18) leads to an estimate for $\|\widetilde{U}_N\|^2_{1,\sigma_{2,\beta}} + \beta\|\widetilde{U}_N\|^2_{\sigma_{1,\beta}}$. Finally, we use Lemma 5.2 again to reach the following conclusion.

THEOREM 5.3. *Let U and u_N be the solutions of* (5.4) *and* (5.7)*, respectively, and $\mu > \frac{1}{4}\beta^2$. If $U \in A^r_{0,\beta}(\Lambda)$ with $U(0) = 0$, and $f \in A^s_{0,\beta}(\Lambda)$ and $\partial^s_x f \in L^2_{\omega_{s-1,\beta}}(\Lambda)$ with integers $r \geq 2$ and $s \geq 1$, then*

$$\|U - u_N\|^2_{1,\sigma_{2,\beta}} + \beta\|U - u_N\|^2_{\sigma_{1,\beta}} \leq \widetilde{B}^{(1)}_{N,\beta,r}(U) + \widetilde{B}^{(2)}_{N,\beta,r}(U) + \widetilde{B}^{(3)}_{N,\beta,s}(f).$$

6. Numerical results. We present numerical results to illustrate the efficiency of the proposed schemes.

6.1. The scheme (4.8). We first take a look at the matrix form of the system (4.8). We take the base functions $\psi_j(\rho) = \mathcal{L}^{(1,\beta)}_j(\rho)$ and let $\mathbb{P}_N = \text{span}\{\psi_0, \psi_1, \ldots, \psi_N\}$. By (2.5) and (2.3), we have

$$\psi_j(\rho) = \frac{1}{\beta}\Big(\partial_\rho \mathcal{L}^{(1,\beta)}_j(\rho) - \partial_\rho \mathcal{L}^{(1,\beta)}_{j+1}(\rho)\Big) = -\mathcal{L}^{(2,\beta)}_{j-1}(\rho) + \mathcal{L}^{(2,\beta)}_j(\rho).$$

This fact together with (2.3) and (2.6) leads to

$$a_{jk} := (\partial_\rho\psi_k, \partial_\rho\psi_j)_{\omega_{2,\beta}} = \beta^2\gamma^{(2,\beta)}_{k-1}\delta_{j,k}, \quad m_{jk} := (\psi_k, \psi_j)_{\omega_{1,\beta}} = \gamma^{(1,\beta)}_k\delta_{j,k},$$
$$s_{jk} := (\psi_k, \psi_j)_{\omega_{2,\beta}} = \begin{cases} -\gamma^{(2,\beta)}_{k-1}, & j = k-1, \\ \gamma^{(2,\beta)}_{k-1} + \gamma^{(2,\beta)}_k, & j = k, \\ -\gamma^{(2,\beta)}_k, & j = k+1, \\ 0 & \text{otherwise}. \end{cases}$$

Next, we set

$$\begin{aligned} &u_N(\rho) = \sum_{j=0}^N \hat{u}_j\psi_j(\rho), \quad \mathbf{u} = (\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_N)^T, \quad f_j = \ (f, \psi_j)_{\omega_{2,\beta},G,N}, \\ &\mathbf{f} = (f_0, f_1, \ldots, f_N)^T, \quad A = (a_{jk})_{0\leq j,k\leq N}, \quad M = (m_{jk})_{0\leq j,k\leq N}, \\ &S = (s_{jk})_{0\leq j,k\leq N}. \end{aligned} \tag{6.1}$$

Then the system (4.9) becomes

$$\Big(A + \Big(\mu - \frac{\beta^2}{4}\Big)S + \beta M\Big)\mathbf{u} = \mathbf{f}. \tag{6.2}$$

It is seen that this system is symmetric, tridiagonal, and easy to be inverted.

We now present some numerical results using the previous scheme to solve (4.1) with spherically symmetric solution $W(\rho)$. Basically, we find $u_N(\rho)$ from the system (6.2), and then evaluate the numerical solution by $w_N(\rho) = e^{-\frac{\beta}{2}\rho}u_N(\rho)$. In the following computations, let $\mu = 5$ in (4.8).

Example 1. We take the test function $W(\rho) = e^{-\gamma\rho}\sin h\rho$, with $\gamma > 0$, which decays exponentially at infinity. The corresponding solution of formula (4.4) is $U(\rho) = e^{(\beta/2-\gamma)\rho}\sin h\rho$. We measure the errors in two ways:

(i) maximum pointwise error:

$$\max_{0\leq j\leq N}\Big|\Big(W(\xi^{1,\beta}_{G,N,j}) - w_N(\xi^{1,\beta}_{G,N,j})\Big)\xi^{1,\beta}_{G,N,j}\Big| \sim \sup_{\rho\in\Lambda}|\rho(W(\rho) - w_N(\rho))|;$$

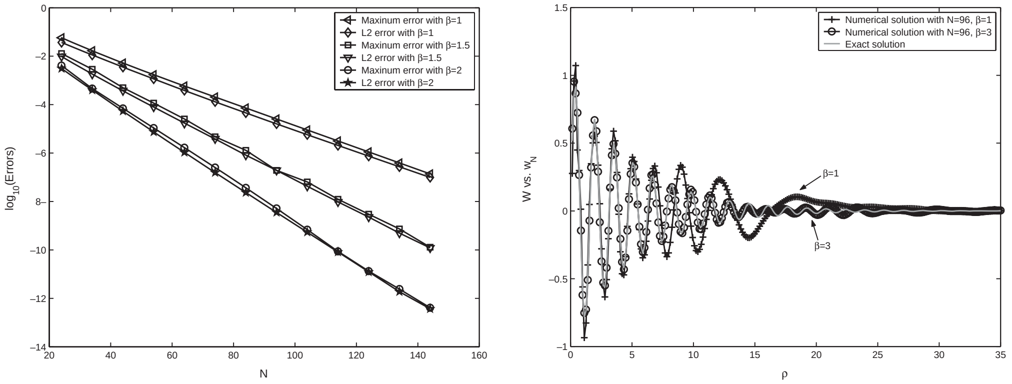


FIG. 6.1. *Convergence rate: Example* 1 *with* $\gamma = 1$ *and* $h = 3$ *on the left. Generalized Laguerre approximation: Example* 1 *with* $\gamma = 0.2$ *and* $h = 4$ *on the right.*

(ii) discrete L^2-error:

$$\|W - w_N\|_N := \|U - u_N\|_{\omega_{1,\beta},G,N} \sim \|W - w_N\|_{\hat{\omega}_1}.$$

In the left part of Figure 6.1, we plot the $\log_{10}$ of maximum error and the $\log_{10}$ of L^2-error against various N for $\gamma = 1,\ h = 3$, and different β. As predicted in Theorem 4.2 and Remark 4.1, the approximate solution will converge faster than any algebraic power, which is confirmed by the error behaviors (like $e^{-cN}, c > 0$) as shown in the figure. We also see that for fixed N, the scheme with $\beta = 2$ or $\beta = 1.5$ produces better numerical results than that with $\beta = 1$ (the usual generalized Laguerre approximation).

To see more clearly the role of β, we compare in the right part of Figure 6.1 the exact solution with $\gamma = 0.2$ and $h = 4$ with the numerical solution obtained by our pseudospectral scheme with $N = 96$ and $\beta = 1, 3$. Notice that the approximation solution with $\beta = 1$ exhibits an observable error, while the numerical solution with $\beta = 3$ is virtually indistinguishable with the exact solution. This example demonstrates that a suitable choice of the parameter β can raise the accuracy, and also enhance greatly the resolution capabilities of the generalized Laguerre approximations.

Example 2. We take $W(\rho) = \frac{\rho}{(\rho+1)^k}$ with $k > 1$, which decays algebraically at infinity. It is clear that $\rho^{\frac{3}{2}} W(\rho) \to 0$, as $\rho \to \infty$, if $k > \frac{5}{2}$.

In Figure 6.2, we plot the $\log_{10}$ of L^2-errors vs. $\sqrt{N}$ for different k and β. We see that the convergence rates are of order $O(e^{-c\sqrt{N}})$ for all cases, which are somewhat better than those predicted in Theorem 4.2 and Remark 4.1 (no more than order k). We also observe that for larger N, better numerical result can be obtained by choosing suitable $\beta < 1$ if the solution decays slowly (cf. the left part of Figure 6.2, where $W(\rho) = O(\rho^{-1.51})$), while conversely for the solution decaying very fast (cf. the right part of Figure 6.2, where $W(\rho) = O(\rho^{-4})$).

Example 3. We take $W(\rho) = \frac{\sin h\rho}{(1+\rho)^k}$ with $k > 0$, which decays algebraically with oscillation.

In Figure 6.3, we plot the $\log_{10}$ of L^2-errors vs. $\log_{10} N$. In the left part, we take $h = 3,\ k = 4$, and $\beta = 1, 2, 3, 4$, while in the right part, we fix $\beta = 2$ and $h = 3$, and test different $k = 3, 4, 5$. It is clear that in all cases, the errors decay at certain

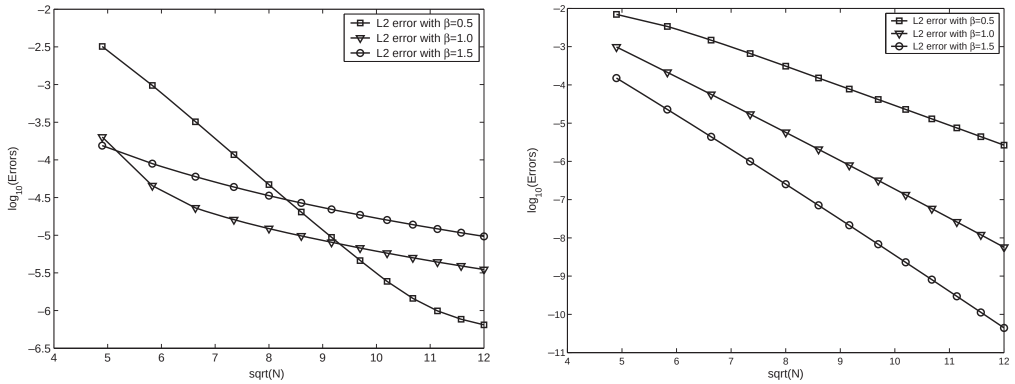


FIG. 6.2. *Convergence rate of generalized Laguerre pseudospectral method: Example* 2 *with* $k = 2.51$ *on the left; Example* 2 *with* $k = 5$ *on the right.*

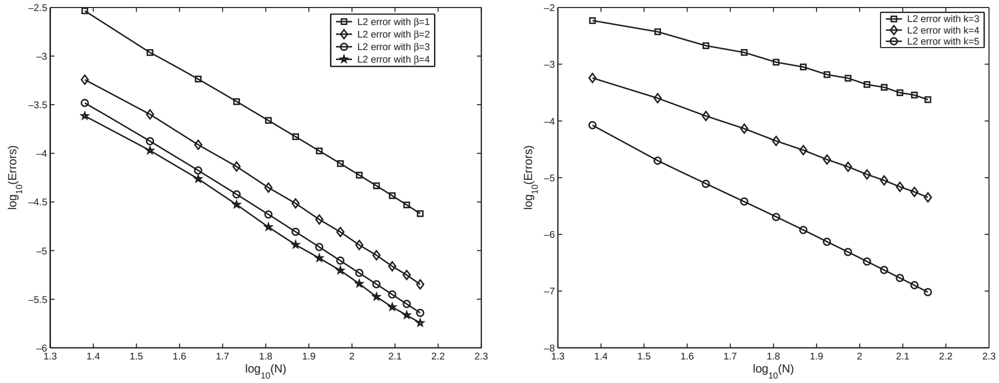


FIG. 6.3. *Convergence rate of generalized Laguerre pseudospectral method: Example* 3 *with* $h = 3$ *and* $k = 4$ *on the left; Example* 3 *with* $h = 3$ *and* $\beta = 2$ *on the right.*

algebraic rate. Once again, we see from the left part of this figure that a suitable parameter β can produce better numerical results. On the other hand, the right part shows that the faster the exact solution decays, the smaller the numerical errors would be. The previous facts coincide again well with our theoretical results.

6.2. The scheme (5.8). We next describe an efficient implementation for scheme (5.8). Set $\psi_j(x) := \mathcal{L}_j^{(0,\beta)}(x) - \mathcal{L}_{j+1}^{(0,\beta)}(x)$, $j \geq 0$, $\beta > 0$. Clearly, $\psi_j(0) = 0$. Hence, ${}_0\mathbb{P}_N = \text{span}\{\psi_0, \psi_1, \ldots, \psi_{N-1}\}$. We now study the structures of the corresponding matrices. Thanks to (2.5), we have $\partial_x \psi_j(x) = \beta \mathcal{L}_j^{(0,\beta)}(x)$, which, along with (2.3), implies that $(1+x)^2 \partial_x \psi_j(x)$ is a linear combination of $\mathcal{L}_l^{(0,\beta)}$, $j-2 \leq l \leq j+2$. This fact with (2.3), (2.5), and (2.6) leads to

$$\begin{aligned} a_{jk} &:= (\partial_x \psi_k, \partial_x \psi_j)_{\sigma_{2,\beta}} = 0 \quad &&\text{if} \quad |j-k| > 2, \\ b_{jk} &:= (\psi_k, \psi_j)_{\sigma_{2,\beta}} = 0 &&\text{if} \quad |j-k| > 3, \\ c_{jk} &:= (\psi_k, \psi_j)_{\sigma_{1,\beta}} = 0 &&\text{if} \quad |j-k| > 2. \end{aligned} \tag{6.3}$$

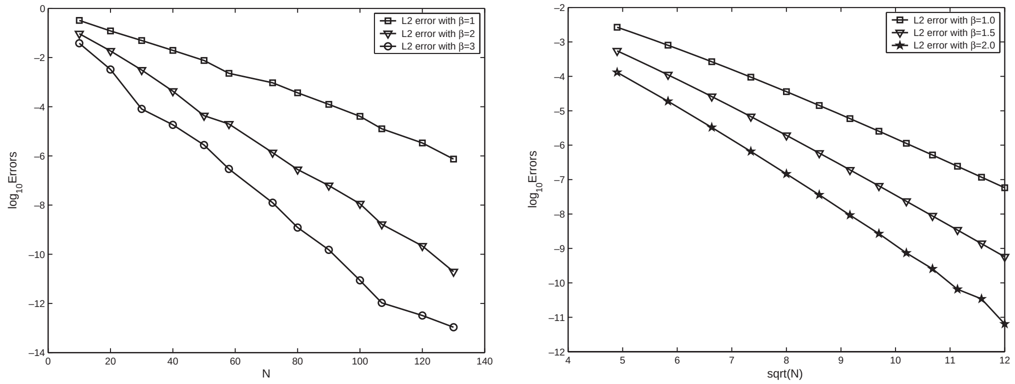


FIG. 6.4. *Convergence rate: Example* 4 *with* $\gamma = 1$ *and* $h = 3$ *on the left; Example* 5 *with* $h = 1$ *and* $k = 5$ *on the right.*

By setting

$$\begin{aligned} &u_N = \sum_{j=0}^{N-1} \hat{u}_j \psi_j(\rho), \quad \mathbf{u} = (\hat{u}_0, \hat{u}_1, \ldots, \hat{u}_{N-1})^{\mathrm{T}}, \\ &f_j = (f, \psi_j)_{\omega_{0,\beta},R,N}, \quad \mathbf{f} = (f_0, f_1, \ldots, f_{N-1})^{\mathrm{T}}, \\ &A = (a_{jk})_{0\le j,k\le N-1}, \quad B = (b_{jk})_{0\le j,k\le N-1}, \quad C = (c_{jk})_{0\le j,k\le N-1}, \end{aligned} \tag{6.4}$$

the system (5.8) becomes

$$\Big(A + \Big(\mu - \frac{\beta^2}{4}\Big)B + \beta C\Big)\mathbf{u} = \mathbf{f}. \tag{6.5}$$

The coefficient matrix is symmetric and has only several nonvanishing diagonals. Moreover, the nonzero entries can be determined explicitly by using properties of generalized Laguerre polynomials as shown in section 2.

We present below two numerical examples to show the efficiency of generalized Laguerre pseudospectral methods for exterior problems. Let $\mu = 5$ in (5.7).

Example 4. We take the test function $W(\rho) = e^{-\gamma(\rho-1)} \sin h(\rho - 1)$ with $\gamma > 0$ and $\rho \ge 1$, which decays exponentially at infinity. The corresponding solution of (5.3) is $U(\rho) = e^{(\beta/2-\gamma)\rho} \sin h\rho$. We denote the discrete L^2-error by $\|W - w_N\|_N := \|U - u_N\|_{\omega_{0,\beta},R,N}$.

In the left part of Figure 6.4, we plot the $\log_{10}$ of L^2-error against various N for $\gamma = 1,\ h = 3$, and different β. We observe a convergence rate of order $O(e^{-cN})$, as predicted in Theorem 5.3. Moreover, for fixed N, the scheme with $\beta = 3$ or $\beta = 2$ produces better numerical results than that with $\beta = 1$.

Example 5. We take $W(\rho) = \frac{\sin(h(\rho-1))}{\rho^k}$ with $k > 0$ and $\rho \ge 1$, which decays algebraically with oscillation. It is clear that $\rho^{\frac{3}{2}} W(\rho) \to 0$, as $\rho \to \infty$, if $k > \frac{3}{2}$.

In the right part of Figure 6.4, we plot the $\log_{10}$ of L^2-error vs. $\sqrt{N}$ for $h = 1,\ k = 5$, and different β. It is seen that the convergence rates are of order $O(e^{-c\sqrt{N}})$, which are somewhat better than what were predicted in Theorem 5.3 (no more than order k). Once again, the error behaviors confirm that a suitable choice of β gives better numerical results than that obtained from the usual generalized Laguerre approximation ($\beta = 1$).

7. Concluding remarks. In this paper, we established a set of results on generalized Laguerre–Gauss-type interpolation in nonuniformly weighted Sobolev spaces with the weight function $\omega_{\alpha,\beta}(x) = x^{\alpha}e^{-\beta x},\ \alpha > -1, \beta > 0$, which provided us useful tools in developing and analyzing generalized pseudospectral methods for a variety of problems in unbounded domains.

Several advantages justified our choice of working on the orthogonal system $\{\mathcal{L}_l^{(\alpha,\beta)}(x)\}$ with general parameters $\alpha > -1, \beta > 0$.

- The parameter α played an essential part in forming the pseudospectral schemes, which was chosen to agree with the degree of singular coefficients of leading terms in underlying equations. For instance, we take $\alpha = 2$ for three-dimensional problems as in (4.1) and (5.1), while we should take $\alpha = 1$ for two-dimensional problems.
- The adjustable parameter β offers great flexibility to match various asymptotic behaviors of the solutions at infinity. In fact, a suitable choice of β depends on certain coefficients which determine the asymptotic behaviors of solutions such as the parameter μ in (4.1) and (5.1).
- The parameter β somehow played a role similar to a scaling factor, which could improve the numerical resolution. But they are not exactly the same. Indeed, in the scaling method with variable transformation, one approximates the function $v(\beta x)$ by the basis $\{\mathcal{L}_l^{(\alpha)}(\beta x)\}$. However, we approximate the function $v(x)$ directly.
- As shown in sections 4 and 5, we could always choose a suitable value of β to guarantee the well-posedness of our Galerkin formulation, provided that some conditions are fulfilled.
- From theoretical point of view, our analysis included usual Laguerre ($\alpha = 0$ and $\beta = 1$) and standard generalized Laguerre ($\alpha > -1$ and $\beta = 1$) approximations as special cases. Moreover, our estimates improved the previously published results for the special case $\alpha = 0$ and $\beta = 1$ (cf. [19]). Roughly speaking, the factor $N^{\gamma}, \gamma > 0$ appearing in the upper bound of the interpolation error of [19] is now replaced by $\ln N$.

In this paper, our pseudospectral method was designed for transformed equations (cf. (4.4) and (5.3)). We may also take the generalized Laguerre functions $\widehat{\mathcal{L}}_l^{(\alpha,\beta)}(x) = e^{-\frac{\beta}{2}x}\mathcal{L}_l^{(\alpha,\beta)}(x)$ as the base functions that are mutually orthogonal with the weight $\hat{\omega}_{\alpha}(x) = x^{\alpha}$. In this case, the weights of Gauss quadrature, $\hat{\omega}_{Z,N,j}^{(\alpha,\beta)} = e^{\beta\xi_{Z,N,j}^{(\alpha,\beta)}}\omega_{Z,N,j}^{(\alpha,\beta)},\ Z = G, R,\ 0 \le j \le N$. Then, the pseudospectral scheme for (4.4) is to find $w_N \in \widehat{\mathbb{P}}_N := \{\phi : \phi = e^{-\frac{\beta}{2}x}\psi \quad \forall \psi \in \mathbb{P}_N\}$ such that

$$(\partial_x w_N, \partial_x \phi)_{\hat{\omega}_2,G,N} + \mu(w_N, \phi)_{\hat{\omega}_2,G,N} = (f, \phi)_{\hat{\omega}_2,G,N} \quad \forall \phi \in \widehat{\mathbb{P}}_N,$$

where $(\cdot,\cdot)_{\hat{\omega}_2,G,N}$ is the corresponding discrete inner product.

REFERENCES

[1] J. Bergh and J. Löfström, *Interpolation Spaces: An Introduction,* Spring-Verlag, Berlin, 1976.

[2] C. Bernardi and Y. Maday, *Spectral methods,* in Handbook of Numerical Analysis, Vol. 5, Techniques of Scientific Computing, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 209–485.

[3] J. P. Boyd, *Chebyshev and Fourier Spectral Methods,* 2nd ed., Dover, Mineola, NY, 2001.

[4] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics,* Springer-Verlag, Berlin, 1988.

[5] O. Coulaud, D. Funaro, and O. Kavian, *Laguerre spectral approximation of elliptic problems in exterior domains,* Comput. Methods Appl. Mech. Engrg., 80 (1990), pp. 451–458.

[6] D. Funaro, *Polynomial Approximations of Differential Equations,* Springer-Verlag, Berlin, 1992.

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

[8] B.-Y. Guo, *Spectral Methods and Their Applications,* World Scientific, Singapore, 1998.

[9] B.-Y. Guo, *Error estimation for Hermite spectral method for nonlinear partial differential equations,* Math. Comp., 68 (1999), pp. 1067–1078.

[10] B.-Y. Guo and J. Shen, *Laguerre–Galerkin method for nonlinear partial differential equations on a semi-infinite interval,* Numer. Math., 86 (2000), pp. 635–654.

[11] B.-Y. Guo, J. Shen, and C.-L. Xu, *Generalized Laguerre approximation and its applications to exterior problems,* J. Comput. Math., 22 (2004), pp. 113–130.

[12] B.-Y. Guo and X.-Y. Zhang, *A new generalized Laguerre spectral approximation and its applications,* J. Comput. Appl. Math., 181 (2005), pp. 342–363.

[13] V. Irazo and A. Falqués, *Some spectral approximations for differential equations in unbounded domains,* Comput. Methods Appl. Mech. Engrg., 98 (1992), pp. 105–126.

[14] Y. Maday, B. Pernaud-Thomas, and H. Vandeven, *Une réhabilitation des méthodes spéctrales de type Laguerre,* Rech. Aerospat., 6 (1985), pp. 353–379.

[15] G. Mastroanni and D. Occorsio, *Lagrange interpolation at Laguerre zeros in some weighted uniform spaces,* Acta Math. Hungar., 91 (2001), pp. 27–52.

[16] G. Mastroanni and G. Monegato, *Nyström interpolants based on zeros of Laguerre polynomials for some Weiner–Hopf equations,* IMA J. Numer. Anal., 17 (1997), pp. 621–642.

[17] J. Shen, *Stable and efficient spectral methods in unbounded domains using Laguerre functions,* SIAM J. Numer. Anal., 38 (2000), pp. 1113–1133.

[18] G. Szegö, *Orthogonal Polynomials,* AMS, Providence, RI, 1959.

[19] C.-L. Xu and B.-Y. Guo, *Laguerre pseudospectral method for nonlinear partial differential equations,* J. Comput. Math., 20 (2002), pp. 413–428.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2590–2606

IDENTIFICATION OF ASYMPTOTIC DECAY TO SELF-SIMILARITY FOR ONE-DIMENSIONAL FILTRATION EQUATIONS*

LAURENT GOSSE† AND GIUSEPPE TOSCANI‡

Abstract. The objective of this paper is the derivation and the analysis of a simple explicit numerical scheme for general one-dimensional filtration equations. It is based on an alternative formulation of the problem using the pseudoinverse of the density's repartition function. In particular, the numerical approximations can be proven to satisfy a contraction property for a Wasserstein metric. Various numerical results illustrate the ability of this numerical process to capture the time-asymptotic decay towards self-similar solutions even for fast-diffusion equations.

Key words. degenerate parabolic equation, porous medium equation, Wasserstein metric

AMS subject classifications. 76S05, 65M06, 35K65

DOI. 10.1137/040608672

1. Introduction and examples.

This paper focuses onto the numerical analysis of the following Cauchy problem:

$$\partial_t u = \partial_{xx}\Phi(u), \quad u(t=0,x) = u_0(x) \geq 0, \quad x \in \mathbb{R},\ t > 0, \tag{1.1}$$

where $u_0 \in L^1(\mathbb{R})$ and $\Phi \in C^2(\mathbb{R}^+)$. It is also customary to assume $\Phi : \mathbb{R}^+ \to \mathbb{R}^+$ to be increasing. In the special case $\Phi(u) = u^m$, $m > 1$, one speaks about the *porous media equation* which describes the flow of a gas through a porous interface according to some constitutive relation linking its velocity to the pressure like Darcy's law. Another interesting situation corresponds to $0 < m < 1$ and is referred to as the *fast-diffusion equation*. The general case of the *filtration equations* can be encountered within the theory of heat transfer assuming the thermal conductivity to be a function of the temperature. A comprehensive introduction to these topics is provided in [25].

The numerical analysis of (1.1) is delicate for at least two reasons: the appearance of singularities for solutions with compact support when $\Phi'(0) = 0$, and the so-called retention property, which means that its size keeps growing as time increases; we shall briefly recall in section 2 theoretical results which are useful on a computational level. Implicit discretizations are thus of common use after, e.g., [6, 14, 20, 21, 17, 16, 23] (other references of interest are [1, 2, 11, 12, 13, 15, 18, 19, 22]); it leads to the resolution of a strictly elliptic problem for $w = \Phi(u)$ at every time-step Δt. Unfortunately, this very stable approach is of little help when investigating the *long-time behavior* of (1.1). Indeed, because of its spreading dynamics, the equation will ask for repetitive regridding. We refer to [3, 8, 7, 9, 10, 24, 25, 26, 27] for theoretical background on the asymptotics of (1.1), mainly in the case $\Phi(u) = u^m$, $m > 1$. We stress that one of the goals of the present work is to provide a tool which allows to achieve numerical studies for the cases still unknown.

*Received by the editors May 20, 2004; accepted for publication (in revised form) April 19, 2005; published electronically January 27, 2006. This work was partially supported by the EEC project HYKE (Hyperbolic and Kinetic Equations) HPRN-CT-2002-00282.

http://www.siam.org/journals/sinum/43-6/60867.html

†Istituto per le Applicazioni del Calcolo (sezione di Bari), Via G. Amendola 122, 70126 Bari, Italy (l.gosse@area.ba.cnr.it).

‡Dipartimento di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy (giuseppe.toscani@unipv.it).

This paper is, therefore, intended to introduce a new numerical approach able to solve both issues in a one-dimensional (1D) context. Loosely speaking, it consists in considering the *repartition function* ϱ of the density u, which is a monotone function, discretizing its values and evolving in time its pseudoinverse $X(t,\varrho)$ which satisfies (3.4) for $t>0$. This Lagrangian strategy is explained in detail in section 3.1, whereas stability and convergence properties are stated in section 3.2. Interestingly, a discrete contraction principle in a Wasserstein metric is shown in section 3.3.

At last, section 4 is concerned with numerical results: we check the decay towards a Gaussian distribution for the heat equation together with two cases of fast-diffusion equations. Then we come to present a decay towards the so-called Barenblatt–Pattle similarity profile for $\Phi(u)=\frac{u^2}{2}$ and a doubly degenerate Buckley–Leverett equation. Finally, some concluding remarks are drawn together with possible extensions to, e.g., viscous pressureless gas equations [4] and radial solutions of two-dimensional (2D) filtration equations.

2. L^1 theory for general porous media equations.

We first notice that there is no restriction in assuming $\Phi(0)=0$ in (1.1). A *weak solution* is generally defined as a distribution $u\in L^2_{loc}(\mathbb{R}^+;L^2(\mathbb{R}))$ such that $\Phi(u)\in L^2_{loc}(\mathbb{R}^+;H^1(\mathbb{R}))$, thus satisfying (1.1) in a weak sense for test functions belonging to $H^1(\mathbb{R})$. Existence and uniqueness results in this framework are recalled for instance in [20, 21]; we shall not pursue in this direction here.

2.1. Existence and uniqueness with nonnegative data.

We follow [25] and are focused hereafter with the Cauchy problem for slow-diffusion equations:

$$\partial_t u=\partial_{xx}(u^m),\quad u(t=0,.)=u_0\in L^1(\mathbb{R}),\quad x\in\mathbb{R},\ t>0. \tag{2.1}$$

We assume $m>1$ and $u_0\geq 0$ in order to study nonnegative L^1 solutions, (e.g., densities) with finite mass. In this context, asymptotics for (2.1) are now well known.

Definition 2.1. *A nonnegative function $u\in C^0(\mathbb{R}^+;L^1(\mathbb{R}))$ is a strong solution of* (2.1) *if*

- $u^m,\partial_t u,\partial_{xx}(u^m)\in L^1_{loc}(\mathbb{R}^+_*;L^1(\mathbb{R}))$,
- $\partial_t u=\partial_{xx}(u^m)$ *holds almost everywhere in* $\mathbb{R}^+_*\times\mathbb{R}$,
- $u(t=0,.)=u_0$.

In contrast to strongly parabolic equations for which $\Phi'(u)\neq 0$, strong solutions are not by no means classical; it is known that they are endowed with Hölder C^α continuity in space only. In one dimension, the exponent has been found to be $\alpha=\min(1,1/(m-1))$ [26]. Relying on the regularity properties of strong solutions to (2.1), we deduce easily two important properties of strong solutions:

- $\int_{\mathbb{R}}u(t,x)\,dx=\int_{\mathbb{R}}u_0(x)\,dx$ for all $t\in\mathbb{R}^+_*$ (conservation of mass);
- $\int_{\mathbb{R}}\max(0,u_1(t,x)-u_2(t,x))\,dx\leq\int_{\mathbb{R}}\max(0,u_1(s,x)-u_2(s,x))\,dx$ for all $t\geq s\geq 0$ (L^1-contraction property).

This last property of course implies uniqueness of strong solutions in the sense of Definition 2.1. A quite general result reads as follows.

Theorem 2.2. *For all $0\leq u_0\in L^1(\mathbb{R})$, there exists a unique strong solution of* (2.1), *$u\in C^0(\mathbb{R}^+;L^1(\mathbb{R}))\cap L^\infty(\mathbb{R}^+_*\times\mathbb{R})$, which satisfies*

- *for all $1\leq p\leq+\infty$, $u_0\in L^p(\mathbb{R})\Rightarrow\|u(t,.)\|_{L^p(\mathbb{R})}\leq\|u_0\|_{L^p(\mathbb{R})}$, $t>0$;*
- *let $v=\frac{mu^{m-1}}{m-1}$; it holds that $\partial_{xx}v(t,\cdot)\geq\frac{1}{(1+m)t}$ (semisuperharmonicity).*

Refined regularity properties are now given.

PROPOSITION 2.3. *Let $u_0 \in L^1 \cap C^0(\mathbb{R})$ be strictly positive and let u be its corresponding strong solution. Then $u \in C^\infty(\mathbb{R}_*^+ \times \mathbb{R}) \cap C^0(\mathbb{R}^+ \times \mathbb{R})$ is strictly positive and realizes a classical solution of* (2.1).

For instance, $u_0(x) = \frac{1}{\pi(1+x^2)}$ generates a unique classical solution of (2.1); this class of initial data will be extensively studied numerically. More generally, one can consider $u_0(x) = \frac{C_p}{(1+x^2)^p}$, $p \geq 1$.

2.2. Asymptotic decay towards source solutions. We observe that equation (1.1) can be rewritten as

$$\partial_t u = \partial_x (D(u)\partial_x u), \qquad D(u) = \Phi'(u).$$

In the special case of (2.1), $D(u) = mu^{m-1}$ is often called the *diffusivity*. It is a well-known fact that degeneracy levels for which Φ' vanishes, (e.g., at $u = 0$ for $\Phi(u) = u^m$) induce a phenomenon called *finite speed of propagation.*

THEOREM 2.4. *Let $0 \leq u_0 \in L^1 \cap L^\infty(\mathbb{R})$ and let u be the corresponding strong solution of* (2.1). *Assume that u_0 is supported in a bounded set of $\mathbb{R}$. Then for any positive time $t > 0$, the support of $u(t,.)$ is also bounded.*

The support of $u(t,\cdot)$ is generally strictly bigger than the one of u_0; this is the *retention property.* Making use of modern analytical tools, one can be more precise [7].

THEOREM 2.5. *Let $(u_0, v_0) \in L^1 \cap L^\infty(\mathbb{R})$ be nonnegative with unit masses and u, v their corresponding strong solutions in the sense of Definition* 2.1. *We define*

$$\Omega_u(t) = \{x \in \mathbb{R} \textit{ such that } u(t,x) > 0\}$$

and the analogue for v. Then it holds true that for all $t > 0$,

- $|\inf(\Omega_u(t)) - \inf(\Omega_v(t))| \leq W_\infty^0$,
- $|\sup(\Omega_u(t)) - \sup(\Omega_v(t))| \leq W_\infty^0$,

where the constant $W_\infty^0 \in \mathbb{R}^+$ depends only on m, u_0, v_0.

Its proof is based on a careful use of a Monge–Kantorowich related metric that we shall discuss in more detail later on; see section 3.3 and [28]. Indeed, as a particular case of (2.1), one can make the following mild hypotheses on the data:

$$u_0 \in L^1 \cap L^\infty(\mathbb{R}), \quad \int_{\mathbb{R}} x.u_0(x)\,dx = 0, \quad \Omega_u(0) \subset \text{ compact of } \mathbb{R}.$$

Then, as $t \to +\infty$, the corresponding strong solution to (2.1) decays towards a *similarity* (or source-type) solution,

$$U(t,x,C) = \frac{1}{t^\mu} \max\left\{0, \left(C - k\frac{x^2}{t^{2\mu}}\right)^{\frac{1}{m-1}}\right\}, \quad \mu = \frac{1}{1+m}, \quad k = \mu\frac{m-1}{2m}, \tag{2.2}$$

the normalization constant $C > 0$ ensuring that $U(t,.,C)$ has unit mass. One can also define the so-called *similarity variable* $\alpha(t)$ solution of

$$\alpha'(t) = \frac{1}{\alpha(t)^m}, \qquad \alpha(0) = 0 \tag{2.3}$$

for which (2.2) reads

$$U(t,x,C) = \frac{1}{\alpha(t)} \max\left\{0, \left(\tilde{C} - \frac{m-1}{2m}\frac{x^2}{\alpha(t)^2}\right)^{\frac{1}{m-1}}\right\}. \tag{2.4}$$

Of course, plugging $U(t, x, \|u_0\|_{L^1})$, $t \geq \tau > 0$, in place of v inside Theorem 2.5 yields an easy bound on the support of any strong solution of (2.1).

It has recently been shown that even for the general case (1.1) for which results are more sparse than (2.1), the third moment of $u(t,x)$ can play the role of an auxiliary variable in order to investigate the long-time behavior.

THEOREM 2.6 (see Toscani [24]). *Let* $0 \leq u_0 \in L^1 \cap L^\infty(\mathbb{R})$ *be of compact support in* $\mathbb{R}$*, u the corresponding strong solution of* (2.1)*, and* $E(t)$ *its third moment:*

$$E(t) = \int_{\mathbb{R}} \frac{x^2}{2} u(t,x)\, dx.$$

Then the similarity variable $\alpha(t)$ *satisfies, as* $t \to +\infty$*,*

$$\frac{E(t)}{\alpha(t)^2} \to E_B = \int_{\mathbb{R}} \frac{x^2}{2} U(t=1, x, C)\, dx,$$

where E_B *is the third moment of the source solution* (2.4) *at time* $t = 1$.

Hence a feasible route to study numerically the long-time asymptotics of (1.1) is to consider its *scaled solutions*,

$$f(t,x) = \sqrt{E(t)}\; u\Big(t, x\sqrt{E(t)}\Big), \tag{2.5}$$

which can hopefully be expected to stabilize as $t \to +\infty$ onto an asymptotic profile $f_\infty(x)$ independent of t. Of course, in case one considers (2.1) with convenient initial data, $f(t,.)$ will converge onto the corresponding Barenblatt–Pattle similarity solution according to the decay results of, e.g., [7, 26]; $t \mapsto E(t)$ is also expected to display a powerlike behavior.

3. An explicit numerical approximation.

We consider now a slightly more general problem than (2.1); namely, (1.1) completed by $0 \leq u_0 \in L^1 \cap L^\infty(\mathbb{R})$, compactly supported with unit mass. We shall also assume for convenience that the second moment vanishes:

$$\int_{\mathbb{R}} x u_0(x)\, dx = 0.$$

This property propagates for $t > 0$, as is easily seen from the formal computation:

$$\partial_t \int_{\mathbb{R}} x u(t,x)\, dx = -\int_{\mathbb{R}} \partial_x \Phi(u)(t,x)\, dx = 0.$$

3.1. Derivation of the numerical process.

As we can already notice, the decay towards similarity solutions can be slow, and because of the retention property, a direct simulation of (1.1) (or even (2.1)) will surely require quite a big computational domain with a possibly fine mesh. This clearly constitutes a numerical difficulty we propose to overcome in an original way, as follows.

- Let us introduce the distribution function associated to the probability density u_0,

$$\varrho_0(x) = \int_{-\infty}^{x} u_0(y)\, dy \in [0,1], \qquad \varrho_0 \in W^{1,1}_{loc}(\mathbb{R}),$$

which is obviously nondecreasing in the x variable. We can thus define its (nondecreasing) pseudoinverse:

$$\begin{array}{rccl} x_0: & [0,1] & \to & \mathbb{R}, \\ & \bar{\varrho} & \mapsto & x_0(\bar{\varrho}) := \inf\{y \in \mathbb{R} \text{ such that } \varrho_0(y) = \bar{\varrho}\}. \end{array} \tag{3.1}$$

If (1.1) holds in the sense of distributions, then also

$$\partial_t \varrho = \partial_x(\Phi(\partial_x \varrho)), \qquad \varrho(t=0,.) = \varrho_0, \tag{3.2}$$

from which one gets $u(t,x) = \partial_x \varrho(t,x)$.

- For any $\bar{\varrho} \in [0,1]$, we can define the *reciprocal mapping*,

$$\begin{array}{rccl} X: & \mathbb{R}^+ & \to & \mathbb{R}, \\ & t & \mapsto & X(t,\bar{\varrho}), \end{array}$$

by means of the implicit function theorem in case $\partial_x \varrho \neq 0$ such that

$$X(t=0,\bar{\varrho}) = x_0(\bar{\varrho}), \qquad \varrho(t, X(t,\bar{\varrho})) = \bar{\varrho}. \tag{3.3}$$

From the second condition in (3.3), one deduces easily

$$\frac{d}{dt}\varrho(t, X(t,\bar{\varrho})) = (\partial_t \varrho + \partial_t X . \partial_x \varrho)(t, X(t,\bar{\varrho})) = 0.$$

This yields the time evolution of $X(.,\varrho)$ (we drop the overbar for ease of reading),

$$\partial_t X = -\frac{\partial_t \varrho}{\partial_x \varrho} = -\frac{\partial_x(\Phi(\partial_x \varrho))}{\partial_x \varrho} \Rightarrow \partial_t X + \partial_\varrho \left(\Phi\left(\frac{1}{\partial_\varrho X} \right) \right) = 0, \tag{3.4}$$

since $\partial_\varrho X = 1/\partial_x \varrho$ holds for smooth enough functions.

Therefore, our numerical approach to (1.1) with convenient (unit mass, centered) initial data stems from computing the pseudoinverse of ϱ_0, $X(t=0,\cdot)$, evolving it in time by means of an *explicit* marching scheme for (3.4) in order to deduce the values of $\varrho(t, X(t,.)) \in [0,1]$ thanks to (3.3). Working on this pseudoinverse $X(t,.)$ allows us to pass through the expanding support issue for any arbitrary large time $t > 0$ since the computational domain is now fixed, $\varrho \in [0,1]$. The retention phenomenon manifests itself through the constant increase of $|\sup_\varrho(X(t,\varrho))|$ and $|\inf_\varrho(X(t,\varrho))|$ as t grows.

We now discretize the ϱ and t axes and define

$$X_k^n \simeq X(t^n, \varrho_k), \qquad t^n = n\Delta t \text{ for } k \in \mathcal{K} \subset \mathbb{N},\ n \in \mathbb{N}. \tag{3.5}$$

A numerical scheme for (3.4) reads

$$X_k^{n+1} = X_k^n - \frac{\Delta t}{|C_k|}\left\{ \Phi\left(\frac{\varrho_{k+1} - \varrho_k}{X_{k+1}^n - X_k^n} \right) - \Phi\left(\frac{\varrho_k - \varrho_{k-1}}{X_k^n - X_{k-1}^n} \right) \right\}, \tag{3.6}$$

where $|C_k| = \varrho_{k+\frac{1}{2}} - \varrho_{k-\frac{1}{2}}$ stands for the width of the control cell centered on ϱ_k with $\varrho_{k+\frac{1}{2}} = \varrho_0(x_{k+\frac{1}{2}})$. As ϱ_0 is at least absolutely and Lipschitz continuous, a convenient choice is given by linear interpolation, $\varrho_{k+\frac{1}{2}} = \frac{1}{2}(\varrho_k + \varrho_{k+1})$, which yields

$$|C_k| = \frac{1}{2}(\varrho_{k+1} - \varrho_{k-1}).$$

Equation (3.6) should be completed with boundary conditions at the edges of the computational domain $\varrho \in [0,1]$. For convenience, we selected Neumann-type conditions: $\Phi(u) = \Phi(\partial_x \varrho) = 0$ in $\varrho = 0$ and $\varrho = 1$. This gives on the left side

$$X_0^{n+1} = X_0^n - \frac{\Delta t}{|C_0|}\Phi\left(\frac{\varrho_1 - \varrho_0}{X_1^n - X_0^n}\right) \leq X_0^n,$$

together with a similar expression on the right side. This furthermore yields,

$$\forall n \in \mathbb{N}, \quad \sum_k |C_k| X_k^n = \sum_k |C_k| X_k^0 \simeq \int_0^1 X(t=0,\varrho)\, d\varrho = \int_{\mathbb{R}} x u_0(x)\, dx.$$

We stress that the ϱ_k's stand for a cumulative mass variable and thus do *not* depend on time. In order to reconstruct $\tilde{\varrho}(t^n,.)$, an approximation of $\varrho(t,.)$ at a given time $t \simeq t^n$, one has to interpolate the family of numerical values ϱ_k, X_k^n, t^n, since

$$\tilde{\varrho}(t^n, X_k^n) = \varrho_k \simeq \varrho(t^n, X_k^n),$$

up to the numerical truncation errors on X_k^n coming from the discretization (3.6). Then one deduces $u(t^n,.)$ by centered divided differences; such a numerical differentiation process may weaken the convergence though. An obvious consequence of this discretization is that the total variation in space of $\tilde{\varrho}(t,.)$ is constant in time.

Other useful quantities for the study of the asymptotic behavior of (2.1) are the moments $m_{2n+1}(t) = \int_{\mathbb{R}} x^{2n} u(t,x)\, dx$, $n \in \mathbb{N}$, which satisfy

$$\frac{d}{dt} m_{2n+1}(t) = 2n(2n-1) m_{2n-1}(t) \text{ for } \Phi(u) = u^m.$$

In the general case of (1.1), one still has

$$\text{(3.7)} \qquad \frac{d}{dt} m_3(t) = 2\int_{\mathbb{R}} \Phi(u)(t,x)\, dx, \qquad m_3(t) = 2E(t) = \int_0^1 X(t,\varrho)^2\, d\varrho.$$

This last equality provides us with a very convenient way to compute the scaled solution $f(t^n,.)$ (2.5) relying on our marching scheme (3.6).

3.2. Stability and consistency of the scheme. To fix ideas, we introduce now a regular computational mesh determined by $\Delta x > 0$, $x_k := k\Delta x$, $k \in \mathbb{N}$. Then we compute the sequence $u_k^0 = u_0(x_k)$ and thus $X_k^0 = X(0,\varrho_k) = x_k$ with $\varrho_0(x_k) = \varrho_k$.

Of course, because of the retention property, the derivation of bounds for the X_k^n's is doomed in advance because we expect $\sup_\varrho X(t,\varrho)$ to diverge when $t \to +\infty$. However, we can prove that the scheme (3.6) is *monotonicity-preserving.*

LEMMA 3.1. *Let* $0 < u_0 \in L^1 \cap L^\infty(\mathbb{R})$ *and let* $\Phi \in C^1(\mathbb{R})$ *be an increasing function; we denote* $0 < a := \inf_{k\in\mathcal{K}} \left(X_{k+1}^0 - X_k^0\right)$. *Then, under the CFL condition,*

$$\text{(3.8)} \qquad \frac{\Delta t}{a^2} \sup_k \left\{ \frac{\varrho_{k+1} - \varrho_k}{|C_k|} \Phi'\left(\frac{\varrho_{k+1} - \varrho_k}{X_{k+1}^0 - X_k^0}\right)\right\} \leq 1,$$

the scheme (3.6) *is monotonicity-preserving. Moreover, there hold, for* $n \in \mathbb{N}$,

$$\text{(3.9)} \qquad \sum_k |\delta X_{k+\frac{1}{2}}^n| \left|\frac{\delta \varrho_{k+\frac{1}{2}}}{\delta X_{k+\frac{1}{2}}^n}\right|^p \leq \|u_0\|_{L^p(\mathbb{R})}^p, \qquad p \geq 1,$$

and the uniform Lipschitz estimate

$$\sup_k \left| \frac{\delta \varrho_{k+\frac{1}{2}}}{\delta X^n_{k+\frac{1}{2}}} \right| \leq \|u_0\|_{L^\infty(\mathbb{R})}. \tag{3.10}$$

The estimates (3.9) and (3.10) are of course the discrete analogues of the continuous ones recalled in Theorem 2.2.

Proof. We first want to prove that $X^{n+1}_{k+1} - X^{n+1}_k$ is a positive combination of its neighbors at time t^n. To this end, we proceed by induction: let us assume that $X^n_{k+1} - X^n_k \geq a > 0$; from (3.6) we get

$$\begin{aligned} X^{n+1}_{k+1} - X^{n+1}_k = X^n_{k+1} - X^n_k - \Bigg\{ \frac{\Delta t}{|C_{k+1}|} \left(\Phi\Big(\frac{\varrho_{k+2} - \varrho_{k+1}}{X^n_{k+2} - X^n_{k+1}}\Big) - \Phi\Big(\frac{\varrho_{k+1} - \varrho_k}{X^n_{k+1} - X^n_k}\Big) \right) \\ - \frac{\Delta t}{|C_k|} \left(\Phi\Big(\frac{\varrho_{k+1} - \varrho_k}{X^n_{k+1} - X^n_k}\Big) - \Phi\Big(\frac{\varrho_k - \varrho_{k-1}}{X^n_k - X^n_{k-1}}\Big) \right) \Bigg\}. \end{aligned}$$

Thanks to the hypothesis, we can apply the mean-value theorem to the function Φ in the preceding expression. We introduce some notation: $\delta X^n_{k+\frac{1}{2}} := X^n_{k+1} - X^n_k \geq 0$, $\delta \varrho_{k+\frac{1}{2}} := \varrho_{k+1} - \varrho_k \geq 0$, and so on. This boils down to

$$\begin{aligned} \delta X^{n+1}_{k+\frac{1}{2}} = \delta X^n_{k+\frac{1}{2}} - \Bigg\{ \frac{\Delta t}{|C_{k+1}|} \Phi'_{k+1} \left(\frac{\delta \varrho_{k+\frac{3}{2}} \delta X^n_{k+\frac{1}{2}} - \delta \varrho_{k+\frac{1}{2}} \delta X^n_{k+\frac{3}{2}}}{\delta X^n_{k+\frac{3}{2}} \delta X^n_{k+\frac{1}{2}}} \right) \\ - \frac{\Delta t}{|C_k|} \Phi'_k \left(\frac{\delta \varrho_{k+\frac{1}{2}} \delta X^n_{k-\frac{1}{2}} - \delta \varrho_{k-\frac{1}{2}} \delta X^n_{k+\frac{1}{2}}}{\delta X^n_{k+\frac{1}{2}} \delta X^n_{k-\frac{1}{2}}} \right) \Bigg\}, \end{aligned}$$

where Φ'_{k+1}, Φ'_k stand for some midpoint values of the derivative of Φ at time t^n. Now, taking into account the signs of all the present quantities and rearranging terms, we obtain

$$\begin{aligned} \delta X^{n+1}_{k+\frac{1}{2}} = \delta X^n_{k+\frac{1}{2}} \left\{ 1 - \frac{\Delta t \Phi'_{k+1}}{\delta X^n_{k+\frac{3}{2}} \delta X^n_{k+\frac{1}{2}}} \frac{\delta \varrho_{k+\frac{3}{2}}}{|C_{k+1}|} - \frac{\Delta t \Phi'_k}{\delta X^n_{k+\frac{1}{2}} \delta X^n_{k-\frac{1}{2}}} \frac{\delta \varrho_{k-\frac{1}{2}}}{|C_k|} \right\} \\ + \frac{\Delta t \Phi'_{k+1}}{\delta X^n_{k+\frac{3}{2}} \delta X^n_{k+\frac{1}{2}}} \frac{\delta \varrho_{k+\frac{1}{2}}}{|C_{k+1}|} \delta X^n_{k+\frac{3}{2}} + \frac{\Delta t \Phi'_k}{\delta X^n_{k+\frac{1}{2}} \delta X^n_{k-\frac{1}{2}}} \frac{\delta \varrho_{k+\frac{1}{2}}}{|C_k|} \delta X^n_{k-\frac{1}{2}}. \end{aligned}$$

From this positive combination, we infer that

$$\begin{aligned} \frac{\delta X^{n+1}_{k+\frac{1}{2}}}{\delta \varrho_{k+\frac{1}{2}}} = \frac{\delta X^n_{k+\frac{1}{2}}}{\delta \varrho_{k+\frac{1}{2}}} \left\{ 1 - \frac{\Delta t \Phi'_{k+1}}{\delta X^n_{k+\frac{3}{2}} \delta X^n_{k+\frac{1}{2}}} \frac{\delta \varrho_{k+\frac{3}{2}}}{|C_{k+1}|} - \frac{\Delta t \Phi'_k}{\delta X^n_{k+\frac{1}{2}} \delta X^n_{k-\frac{1}{2}}} \frac{\delta \varrho_{k-\frac{1}{2}}}{|C_k|} \right\} \\ + \frac{\Delta t \Phi'_{k+1}}{\delta X^n_{k+\frac{3}{2}} \delta X^n_{k+\frac{1}{2}}} \frac{\delta \varrho_{k+\frac{3}{2}}}{|C_{k+1}|} \left(\frac{\delta X^n_{k+\frac{3}{2}}}{\delta \varrho_{k+\frac{3}{2}}} \right) + \frac{\Delta t \Phi'_k}{\delta X^n_{k+\frac{1}{2}} \delta X^n_{k-\frac{1}{2}}} \frac{\delta \varrho_{k-\frac{1}{2}}}{|C_k|} \left(\frac{\delta X^n_{k-\frac{1}{2}}}{\delta \varrho_{k-\frac{1}{2}}} \right), \end{aligned}$$

which is the desired convex combination under the condition (3.8); this ensures $X^{n+1}_{k+1} - X^{n+1}_k \geq a > 0$. Since $\mathbb{R}^+_* \ni x \mapsto 1/x$ is a convex function, the estimates (3.9) and (3.10) follow by Jensen's inequality. ∎

We stress that the monotonicity property of the X^n_k's is crucial in order to define an approximation $\tilde{\varrho}(t^n, .)$ as being the graph of a monovalued function, which is the least one may expect in this context. In particular, for $p = 1$, (3.9) boils down to $\sup_k \varrho_k \leq 1$.

Of course, in practice, one could obey the restriction (3.8) according to the *real* value of $\inf_k \left(X_{k+1}^n - X_k^n\right)$ at time t^n in order to allow Δt to vary in an *adaptive* way as times increase. We took advantage of this in the numerical tests shown subsequently. In order to keep $\Delta t > 0$, one needs to assume that $X(t,.)$ is strictly increasing, which implies that $u_0 > 0$. This can be partially dropped in practice; see section 4.

The next lemma is an important step towards the convergence result.

LEMMA 3.2. *The scheme* (3.6) *is consistent with* (3.2).

Proof. Let us define the function $\varrho^{\Delta x}$ defined by C^1 interpolation of the values

$$\forall\ (k,n) \in \mathcal{K} \times \mathbb{N}, \qquad \varrho^{\Delta x}(t^n, X_k^n) = \varrho_k. \tag{3.11}$$

From the very definition of $\varrho^{\Delta x}$ (3.11) and the scheme (3.6), one derives

$$\begin{aligned}\varrho^{\Delta x}(t^n, X_k^n) &= \varrho^{\Delta x}(t^{n+1}, X_k^{n+1}) \\ &= \varrho^{\Delta x}\left(t^{n+1}, X_k^n - \frac{\Delta t}{|C_k|}\left\{\Phi\Big(\frac{\varrho_{k+1}-\varrho_k}{X_{k+1}^n - X_k^n}\Big) - \Phi\Big(\frac{\varrho_k - \varrho_{k-1}}{X_k^n - X_{k-1}^n}\Big)\right\}\right).\end{aligned}$$

Then the mean-value theorem gives for some $\zeta_k^{n+1} = \lambda X_k^n + (1-\lambda)X_k^{n+1}$, $\lambda \in [0,1]$,

$$\begin{aligned}\varrho^{\Delta x}(t^{n+1}, X_k^{n+1}) = \varrho^{\Delta x}(t^n, X_k^n)& \\ + \frac{\Delta t}{|C_k|}\partial_x \varrho^{\Delta x}(t^{n+1}, \zeta_k^{n+1})&\left\{\Phi\Big(\frac{\varrho_{k+1}-\varrho_k}{X_{k+1}^n - X_k^n}\Big) - \Phi\Big(\frac{\varrho_k - \varrho_{k-1}}{X_k^n - X_{k-1}^n}\Big)\right\}.\end{aligned}$$

We now observe that

$$\frac{\partial_x \varrho^{\Delta x}(t^{n+1}, \zeta_k^{n+1})}{|C_k|} = \frac{\partial_x \varrho^{\Delta x}(t^{n+1}, \zeta_k^{n+1})}{\varrho^{\Delta x}(t^{n+1}, X_{k+\frac12}^{n+1}) - \varrho^{\Delta x}(t^{n+1}, X_{k-\frac12}^{n+1})} = \frac{1}{X_{k+\frac12}^{n+1} - X_{k-\frac12}^{n+1}}$$

up to high-order terms. Replacing the other values ϱ_k inside (3.6) by the corresponding $\varrho^{\Delta x}(t^n,.)$ leads to the following *finite volume* discretization of (3.2):

$$\begin{aligned}\varrho^{\Delta x}(t^{n+1}, X_k^n) &= \varrho^{\Delta x}(t^n, X_k^n) + \frac{\Delta t}{X_{k+\frac12}^{n+1} - X_{k-\frac12}^{n+1}} \\ \times \left\{\Phi\Big(\frac{\varrho^{\Delta x}(t^n, X_{k+1}^n) - \varrho^{\Delta x}(t^n, X_k^n)}{X_{k+1}^n - X_k^n}\Big)\right. &- \left.\Phi\Big(\frac{\varrho^{\Delta x}(t^n, X_k^n) - \varrho^{\Delta x}(t^n, X_{k-1}^n)}{X_k^n - X_{k-1}^n}\Big)\right\}.\end{aligned} \tag{3.12}$$ □

We now derive a time-modulus of equicontinuity for the aforementioned scheme.

LEMMA 3.3. *Under the assumptions of Lemma* 3.1 *and the CFL restriction* (3.8), *the scheme* (3.12) *satisfies*

$$\sup_k |\varrho^{\Delta x}(t^{n+1}, X_k^n) - \varrho^{\Delta x}(t^n, X_k^n)| = O(\sqrt{\Delta t}). \tag{3.13}$$

Proof. This readily follows from the expression (3.12), the CFL condition, and the Lipschitz estimate (3.10). □

THEOREM 3.4. *Under the assumptions of Lemma* 3.1 *and the CFL restriction* (3.8), *the sequence of approximate solutions* $\varrho^{\Delta x}$ *is relatively compact as* $\Delta x \to 0$ *in* $L^p_{loc}(\mathbb{R}_*^+ \times \mathbb{R})$; *it converges towards the unique solution in the sense of distributions of*

$$\partial_t \varrho = \partial_x(\Phi(\partial_x \varrho)), \quad \varrho(t=0,.) = \varrho_0 \in W^{1,p}(\mathbb{R}), \quad 1 \leq p \leq +\infty.$$

Proof. The proof is a bare consequence of the preceding lemmas together with a time-modulus of equicontinuity, as we explain now. Let us start from (3.12); multiplying by a smooth function with compact support $\varphi(t^{n+1}, X_k^n)$ and summing gives

$$\frac{1}{\Delta t}\sum_{k,n} |X^{n+1}_{k+\frac12} - X^{n+1}_{k-\frac12}|\varphi(t^{n+1}, X_k^n)\Big(\varrho^{\Delta x}(t^{n+1}, X_k^n) - \varrho^{\Delta x}(t^n, X_k^n)\Big)$$
$$= \sum_{k,n} \varphi(t^{n+1}, X_k^n)\left\{\Phi\left(\frac{\delta\varrho_{k+\frac12}}{\delta X^n_{k+\frac12}}\right) - \Phi\left(\frac{\delta\varrho_{k-\frac12}}{\delta X^n_{k-\frac12}}\right)\right\}.$$

Summing by parts yields

$$\frac{1}{\Delta t}\sum_{k,n} \varrho^{\Delta x}(t^n, X_k^n)\Big(-\varphi(t^{n+1}, X_k^n)|X^{n+1}_{k+\frac12} - X^{n+1}_{k-\frac12}| + \varphi(t^n, X_k^n)|X^n_{k+\frac12} - X^n_{k-\frac12}|\Big)$$
$$= \sum_{k,n} \Phi\left(\frac{\delta\varrho_{k+\frac12}}{\delta X^n_{k+\frac12}}\right)\Big(\varphi(t^{n+1}, X_k^n) - \varphi(t^{n+1}, X^n_{k+1})\Big).$$

We deduce, using the preceding notation,

$$\frac{1}{\Delta t}\sum_{k,n} \varrho^{\Delta x}(t^n, X_k^n)\Big(-\varphi(t^{n+1}, X_k^n)\delta X_k^{n+1} + \varphi(t^n, X_k^n)\delta X_k^n\Big)$$
$$-\sum_{k,n} \Phi\left(\frac{\delta\varrho_{k+\frac12}}{\delta X^n_{k+\frac12}}\right)\Big(\varphi(t^{n+1}, X_k^n) - \varphi(t^{n+1}, X^n_{k+1})\Big) = 0.$$

We can rewrite this in integral form as follows:

$$\sum_{k,n} \int_{t^n}^{t^{n+1}} \int_{X^n_{k-\frac12}}^{X^n_{k+\frac12}} \varrho^{\Delta x}(t^n, X_k^n)\Big(-\frac{\varphi(t^{n+1}, X_k^n) - \varphi(t^n, X_k^n)}{\Delta t}\Big)$$
$$+\Phi\left(\frac{\delta\varrho_{k+\frac12}}{\delta X^n_{k+\frac12}}\right)\Big(\frac{\varphi(t^n, X^n_{k+1}) - \varphi(t^n, X_k^n)}{\delta X_k^n}\Big)\,dx\,dt$$
$$= \Delta t\sum_{k,n} \varrho(t^n, X_k^n)\varphi(t^n, X_k^n)(\delta X_k^{n+1} - \delta X_k^n)$$
$$+\Phi\left(\frac{\delta\varrho_{k+\frac12}}{\delta X^n_{k+\frac12}}\right)\Big(\varphi(t^{n+1}, X_k^n) - \varphi(t^{n+1}, X^n_{k+1}) - \varphi(t^n, X_k^n) + \varphi(t^n, X^n_{k+1})\Big).$$

At this point, we use the fact that $\Phi(\delta\varrho/\delta X)$ and $\varrho(t^n, X_k^n)$ are bounded, and φ is smooth in both variables; then we rewrite the first term of the right-hand side as

$$-\sum_{k,n}\Delta t\int_{X^n_{k-\frac12}}^{X^n_{k+\frac12}} (\varrho^{\Delta x}\varphi)(t^n, X_k^n) - (\varrho^{\Delta x}\varphi)(t^{n-1}, X_k^{n-1})\,dx$$

and the second term as

$$-\sum_{k,n}\Delta t\int_{t^n}^{t^{n+1}}\int_{X_k^n}^{X^n_{k+1}} \Phi\left(\frac{\delta\varrho_{k+\frac12}}{\delta X^n_{k+\frac12}}\right)\partial_{tx}\varphi(\tau, \xi)\,d\xi\,d\tau.$$

Now, since $0 < u_0 \in L^1 \cap L^\infty(\mathbb{R})$, ϱ_0 is a strictly increasing Lipschitz function of x; hence the family $(X_k^n)_k$ covers the whole axis $\mathbb{R}$. Then, by regularity, $|X^n_{k+1} - X^n_k| \to 0$

and $|X_k^{n+1} - X_k^n| \to 0$ for $n \in \mathbb{N}$ as $\Delta x \to 0$ since by Lemma 3.1, (3.6) is a convex combination. This is enough to derive the weak form of the equation. Uniqueness in the limit follows from the classical argument of Oleinik for weak solutions; see [25]. □

We close this section by mentioning that the assumption $u_0 > 0$ in Lemma 3.1 is essentially needed in order to ensure that $(X_k^n)_{k\in\mathbb{N}}$ permits us to cover the whole real line as $\Delta x \to 0$. We shall consider in section 4.4 initial data of compact support which are strictly positive only inside their support; in this case, only the support of $u(n\Delta t, .)$ can be expected to be recovered.

3.3. Study of the Wasserstein metric. We mainly follow [7, 10, 28] to study *contraction properties* of the scheme (3.6) within the Wasserstein metric framework. Denoting $\mathbb{P}_p(\mathbb{R})$ as the set of all probability measures on $\mathbb{R}$ with moments of order $1 \le p < +\infty$ and $\Pi(\nu_1, \nu_2)$ as any of the probability measures on $\mathbb{R}^2$ admitting $\nu_{1,2} \in \mathbb{P}_p(\mathbb{R})$ as marginal distributions, the *Wasserstein p-metric* reads

$$W_p(\nu_1,\nu_2) := \left(\inf_{\pi\in\Pi(\nu_1,\nu_2)} \int_{\mathbb{R}^2} |x-y|^p \, d\pi(x,y) \right)^{\frac{1}{p}}, \qquad 1 \le p < +\infty. \tag{3.14}$$

Any probability measure admits a distribution function, which can be chosen right-continuous, nondecreasing, and taking values inside $[0,1]$. A nondecreasing pseudoinverse can be defined as for (3.1). Hence if X_1, X_2 stand for the pseudoinverses of the repartition functions of $\nu_1, \nu_2 \in \mathbb{P}_p(\mathbb{R})$, the distance (3.14) can be rewritten as

$$W_p(\nu_1,\nu_2) := \left(\int_0^1 |X_1(\varrho) - X_2(\varrho)|^p \, d\varrho \right)^{\frac{1}{p}}, \qquad 1 \le p < +\infty. \tag{3.15}$$

According to (1.1), a formal computation leads easily to a contraction property for the metric $W_2(.,.)$. Let $X(t,\varrho), Y(t,\varrho)$ stand for two reciprocal mappings associated to nonnegative and centered initial data of (1.1) $u_0, v_0 \in L^1 \cap L^\infty(\mathbb{R})$ with unit mass

$$\begin{aligned} \frac{d}{dt}\int_0^1 |X(t,\varrho) - Y(t,\varrho)|^2 \, d\varrho &= -2\int_0^1 (X-Y)\partial_\varrho \left\{ \Phi\Big(\frac{1}{\partial_\varrho X}\Big) - \Phi\Big(\frac{1}{\partial_\varrho Y}\Big) \right\}(t,\varrho)\, d\varrho \\ &= 2\int_0^1 \partial_\varrho(X-Y) \left\{ \Phi\Big(\frac{1}{\partial_\varrho X}\Big) - \Phi\Big(\frac{1}{\partial_\varrho Y}\Big) \right\}(t,\varrho)\, d\varrho \\ &\le 0, \end{aligned}$$

because Φ is increasing. Then a similar property can be shown to hold for the outcome of the explicit scheme (3.6).

THEOREM 3.5. *Let u_0, v_0 be two nonnegative initial data in $L^1 \cap L^\infty(\mathbb{R})$ for* (1.1) *and let X, Y be their reciprocal mappings. Under the CFL restriction* (3.8)*, the scheme* (3.6) *is contractive in any Wasserstein metric W_p; more precisely, there holds,*

$$\forall n \in \mathbb{N}, \quad \sum_k |C_k||X_k^{n+1} - Y_k^{n+1}|^p \le \sum_k |C_k||X_k^n - Y_k^n|^p, \quad p \ge 1. \tag{3.16}$$

Proof. Mimicking the preceding formal computation, we aim at establishing

$$\delta W_p := \frac{1}{\Delta t}\sum_k |C_k| \left\{ |X_k^{n+1} - Y_k^{n+1}|^p - |X_k^n - Y_k^n|^p \right\} \le 0.$$

We get from (3.6) that

$$\begin{aligned} X_k^{n+1} - Y_k^{n+1} &= X_k^n - Y_k^n - \left\{ \frac{\Delta t}{|C_k|} \left(\Phi\Big(\frac{\varrho_{k+1}-\varrho_k}{X_{k+1}^n - X_k^n}\Big) - \Phi\Big(\frac{\varrho_k - \varrho_{k-1}}{X_k^n - X_{k-1}^n}\Big)\right)\right. \\ &\qquad \left. - \frac{\Delta t}{|C_k|} \left(\Phi\Big(\frac{\varrho_{k+1}-\varrho_k}{Y_{k+1}^n - Y_k^n}\Big) - \Phi\Big(\frac{\varrho_k - \varrho_{k-1}}{Y_k^n - Y_{k-1}^n}\Big)\right)\right\} \\ &= X_k^n - Y_k^n - \frac{\Delta t}{|C_k|} \left\{ \Big(\Phi_{k+\frac12}(X_{k+1}^n - X_k^n) - \Phi_{k+\frac12}(Y_{k+1}^n - Y_k^n)\Big)\right. \\ &\qquad \left. - \Big(\Phi_{k-\frac12}(X_k^n - X_{k-1}^n) - \Phi_{k-\frac12}(Y_k^n - Y_{k-1}^n)\Big)\right\}, \end{aligned}$$

where we used the notation

$$\Phi_{k+\frac12}(\delta X) := \Phi\Big(\frac{\varrho_{k+1}-\varrho_k}{\delta X}\Big).$$

Thanks to the bound given by Lemma 3.1, we know that $\delta X \geq a > 0$, so the function $\Phi_{k+\frac12}$ is smooth and the mean-value theorem can be applied. The outcome is

$$\begin{aligned} X_k^{n+1} - Y_k^{n+1} &= (X_k^n - Y_k^n)\left(1 + \frac{\Delta t}{|C_k|}\big(\Phi'_{k+\frac12} + \Phi'_{k-\frac12}\big)\right) \\ &\quad - \frac{\Delta t}{|C_k|}\Phi'_{k+\frac12}(X_{k+1}^n - Y_{k+1}^n) - \frac{\Delta t}{|C_k|}\Phi'_{k-\frac12}(X_{k-1}^n - Y_{k-1}^n), \end{aligned}$$

with $\Phi'_{k+\frac12}$ standing for some midpoint value of the derivative of $\Phi_{k+\frac12}$ with respect to δX. Hence, since

$$\Phi'_{k+\frac12}(\delta X) = -\frac{\varrho_{k+1}-\varrho_k}{\delta X^2}\Phi'\Big(\frac{\varrho_{k+1}-\varrho_k}{\delta X}\Big),$$

the CFL condition (3.8) ensures that the last expression is a convex combination. By means of Jensen's inequality, and thanks to the fact that the fluxes are null on the borders of the domain, this yields $\delta W_p \leq 0$ and we are done. □

A consequence of this is that in case one would want to use the discretization (3.6) for a problem (1.1) with a partly atomic probability measure, one can initialize the scheme with a somewhat smoother initial data relying on this contraction property. Moreover, this also ensures that the propagation speed of the free boundaries is correct relying on Theorem 2.5.

We stress that the estimate (3.16) doesn't imply a decay of the support $(X_k^n)_{k\in\mathcal{K}}$ which would somewhat contradict the retention phenomenon. Indeed, the CFL condition (3.8) cannot allow us to choose $Y_k^n \equiv 0$.

4. Numerical results.

All the following tests have been carried out relying on the explicit scheme (3.6); the initial data u_0 is sampled on a set of 257 points, which gives a space-step Δx equal to the length of the domain divided by 256. The ϱ_k's are then deduced by numerical quadrature. The time-step is chosen in an adaptive way, as explained after the proof of Lemma 3.1.

4.1. Validation: The heat equation.

In order to test the scheme on a simple and well-known case, we set up (1.1) with $\Phi(u) = \frac{u}{2}$. The initial data is chosen rather far away from the expected equilibrium state:

$$u_0(x) = \frac12\left(\frac{1}{\pi(1+(x-5)^2)} + \frac{1}{\pi(1+(x+5)^2)}\right), \qquad x \in [-20, 20]. \tag{4.1}$$

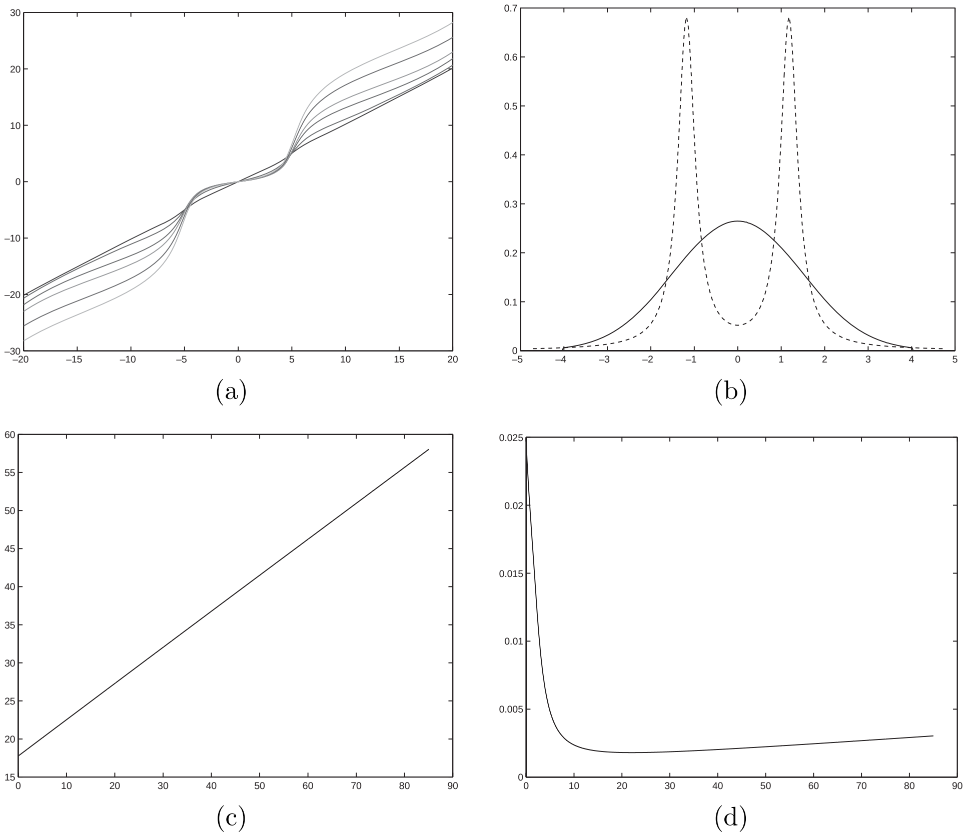


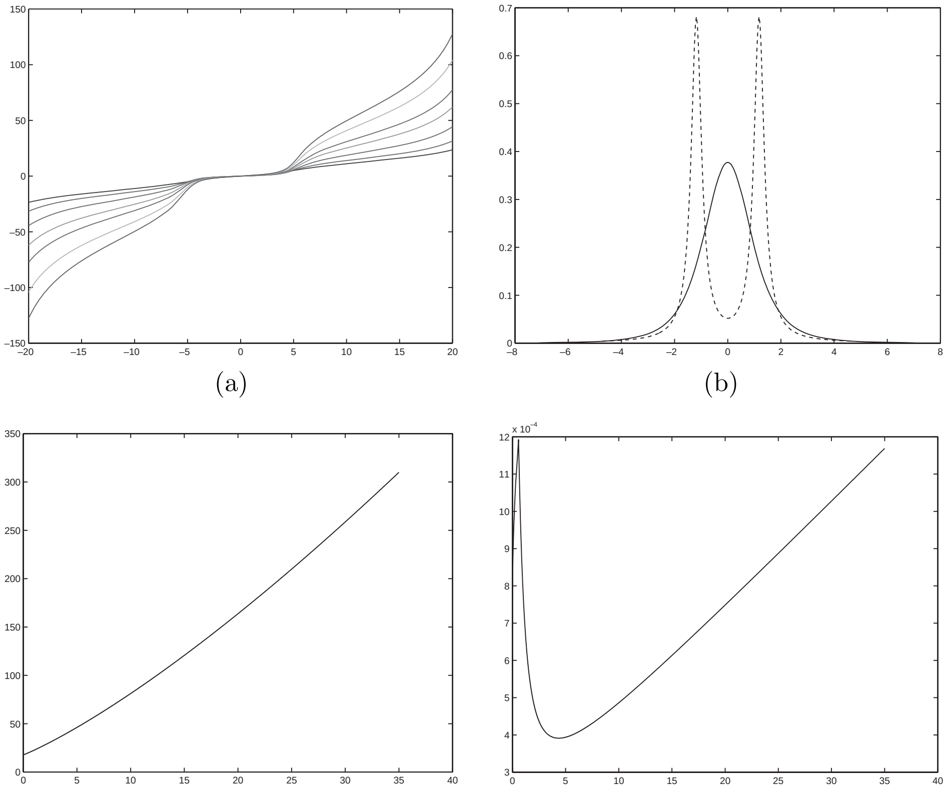
FIG. 1. (a) *Numerical values for* $X(t,x)$ *for* $t = 1, 5, 15, 25, 45, 65$*;* (b) *scaled initial data* $f(t = 0, .)$ (*dotted line*) *and stationary solution* f_∞ (*solid line*)*;* (c) *evolution of* $E(t)$ *with time* t*;* (d) *evolution of* Δt *with time* t.

We observe that even if $\int_{\mathbb{R}} x^2 u_0(x)\,dx$ isn't bounded, one can set up the scheme (3.6) for x inside a compact interval of $\mathbb{R}$. The results at time $t = 85$ are displayed in Figure 1. Along with the evolution of X_k^n as $t^n = 1, 5, 15, 25, 45, 65$, we observe a linear increase of $t \mapsto E(t)$ as shown theoretically and a correct decay onto a Gaussian distribution for the scaled solution $f(t, .)$. We show the corresponding (numerically) stationary profile. The time-step decreases a lot when the two bumps merge but increases afterwards as the solution $u(t, .)$ no longer changes its shape.

4.2. Two cases of fast-diffusion equations. We now display in Figures 2 and 3 a similar experiment with two fast-diffusion equations, respectively, $\Phi(u) = \sqrt{u}$ and $\Phi(u) = u^{\frac{1}{4}}$. The initial data and the computational domain are still given by (4.1). Several major differences show up in this case compared to the heat equation:

- the support of the solution extends much more quickly as the exponent m is decreased, as can be seen on the graphs of the X_k^n's;
- the scaled solutions $f(t, .)$ stabilize at much earlier times ($t \simeq$ 25–30);
- the asymptotic profile is more peaked with a lower value of m;
- the mapping $t \mapsto E(t)$ now looks convex.

However, as for the heat equation, the asymptotic solution has infinite support and is thus C^∞ as a consequence of Proposition 2.3; see also [26]. We stress that the tails



(a) (b)

(c) (d)

FIG. 2. (a) *Numerical values for* $X(t,x)$ *for* $t = 0.5, 2, 5, 10, 15, 25, 35$*,* $m = 0.5$*;* (b) *scaled initial data* $f(t = 0, .)$ (*dotted line*) *and stationary solution* f_∞ (*solid line*)*;* (c) *evolution of* $E(t)$ *with time* t*;* (d) *evolution of* Δt *with time* t*.*

of the initial data can be seen to be close to the ones of the asymptotic profile; see [24] for remarks in this direction.

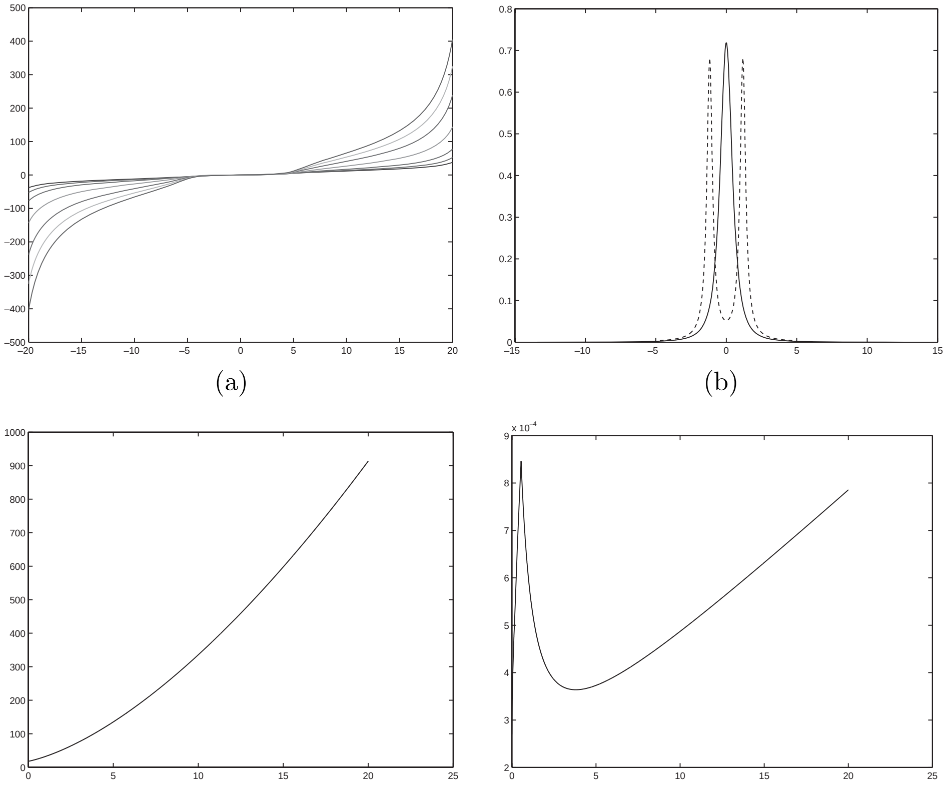
4.3. The porous media equation and Barenblatt's solution. We investigated the case of the classical porous medium equation, namely, $\Phi(u) = \frac{u^2}{2}$ with the data (4.1). Since it isn't compactly supported, we didn't observe the well-known decay towards the corresponding Barenblatt–Pattle solution, but instead, f_∞ exhibits a similar profile with two tails on each side, as shown in Figure 4. Also the mapping $t \mapsto E(t)$ looks concave and the stabilization time is much greater than in the two preceding examples ($t \simeq 200$). The variations of the time-step are moderate in comparison with the fast-diffusion equations.

4.4. Buckley–Leverett's doubly degenerate equation. Finally, we studied a more singular problem given by (1.1) with

$$\Phi(u) = \frac{u^2}{u^2 + 0.5(1-u)^2}.$$

The derivative Φ' vanishes at two points $u = 0$ and $u = 1$. We set up the following smooth initial data extended by zero outside of $[-1, 1]$:

$$u_0 = \cos(\pi x/2)^2, \qquad x \in [-1, 1].$$



(a)

(b)

(c)

(d)

FIG. 3. (a) *Numerical values for $X(t,x)$ for $t = 0.5, 1, 2, 5, 10, 15, 20$, $m = 0.25$;* (b) *scaled initial data $f(t = 0, .)$* (*dotted line*) *and stationary solution f_∞* (*solid line*)*;* (c) *evolution of $E(t)$ with time t;* (d) *evolution of Δt with time t.*

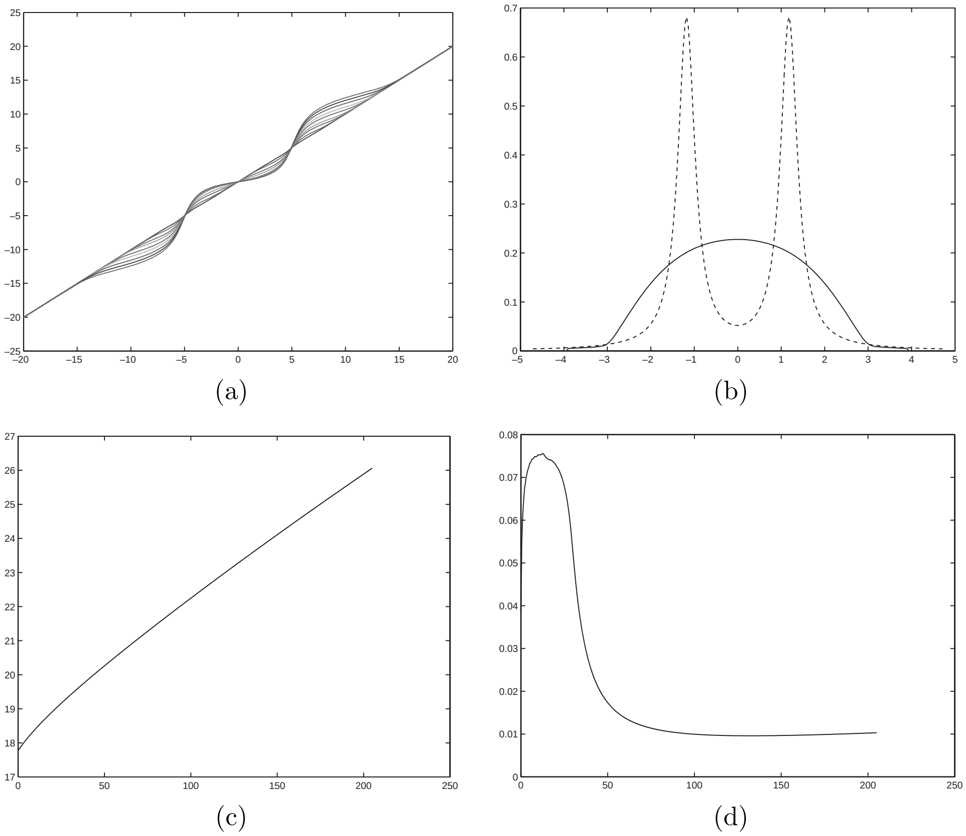
In this case, even if this hasn't been rigorously proven yet, we may expect to observe a decay of this compactly supported function towards a Barenblatt–Pattle profile asymptotically. This can be observed in Figure 5. We can also check that this problem shares other features with the slow-diffusion equation since the mapping $t \mapsto E(t)$ seems concave. However, the support of the solution grows more quickly as times increase.

5. Conclusion and outlook. We introduced and studied analytically in this paper a new numerical scheme for 1D filtration equations of the type (1.1). As a main feature, it allows us to observe the asymptotic decay of solutions towards self-similar ones without requesting important changes of the computational domain (as it would be the case for a conventional discretization; see, e.g., [14, 17, 16, 18, 20, 21, 23]). Moreover, a contraction property in Wasserstein metrics can be easily established. As a final remark, let us stress that a similar derivation can be applied to 1D nonlinear Fokker–Planck equations,

$$\partial_t u = \partial_x(\partial_x V(t,x)u) + \partial_{xx}(\Phi(u)),$$

for which the evolution of the reciprocal mapping $X(t,\varrho)$ would be given by (see [10])

$$\partial_t X + (\partial_x V)(t,X) + \partial_\varrho(\Phi(1/\partial_\varrho X)) = 0,$$



(a) (b)

(c) (d)

FIG. 4. (a) *Numerical values for $X(t, x)$ for $t = 1, 5, 15, 25, 45, 65, 95, 125, 155$;* (b) *scaled initial data $f(t = 0, .)$* (*dotted line*) *and stationary solution f_∞* (*solid line*)*;* (c) *evolution of $E(t)$ with time t;* (d) *evolution of Δt with time t.*

thus only asking for mild changes with respect to (3.4)–(3.6). Very fast diffusion equations could also be handled within the present framework, i.e., $\partial_t u + \partial_{xx}(u^{-m}) = 0$, $0 < m < 1$.

From this last computation, one can, moreover, extract information concerning radial solutions of 2D equations; let $u(t, x, y)$ solve $\partial_t u = \Delta\Phi(u)$, Δ standing for the Laplace operator in $\mathbb{R}^2$, while meeting the requirement that $u(t, x, y) = \tilde{u}(t, r)$, $r = \sqrt{x^2 + y^2}$. Then one deduces easily an equation on $\tilde{u}$, namely, $\partial_t \tilde{u} = \partial_{rr}\Phi(\tilde{u}) + \frac{1}{r}\partial_r \Phi(\tilde{u})$, for which the last term creates a difficulty with respect to the aforementioned Fokker–Planck computation. This can be circumvented as follows: one observes that

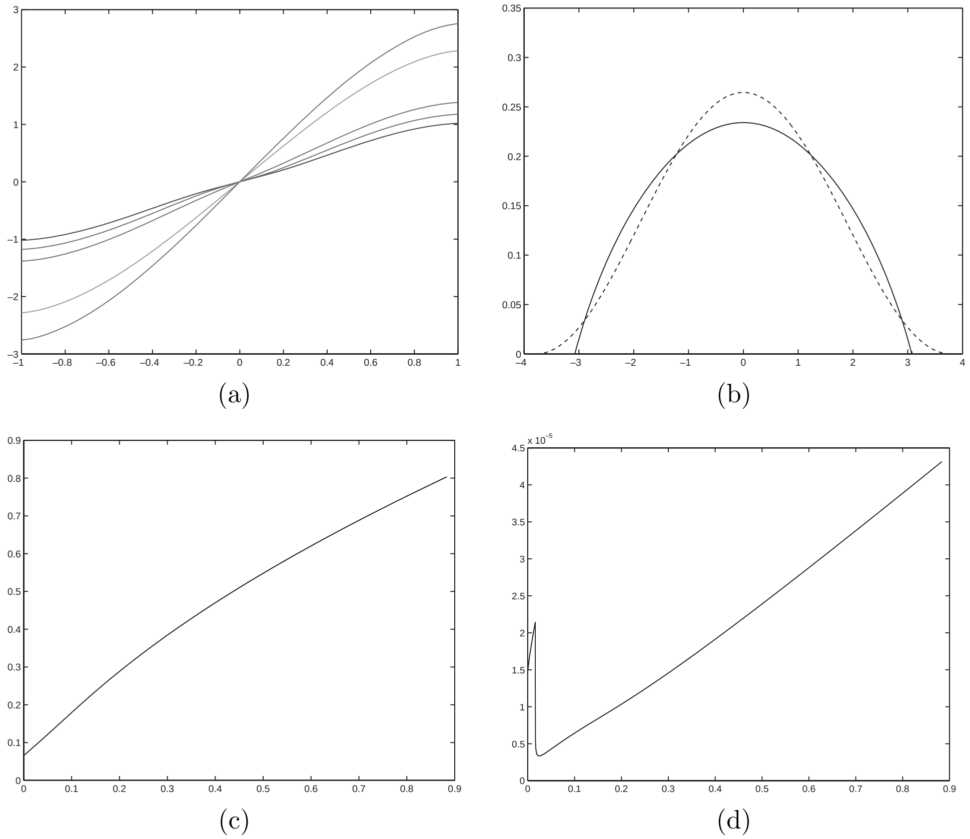
$$\varrho(t, r) = \int_0^r s\, \tilde{u}(t, s)\, ds, \qquad \partial_t \int_0^\infty s\, \tilde{u}(t, s)\, ds \equiv 0.$$

Then it is possible to define R as the reciprocal mapping such that,

$$\forall \bar{\varrho} \in [0, 1], \qquad \varrho(t, R(t, \bar{\varrho})) = \bar{\varrho},$$

and satisfying the equation

$$\partial_t R + R\partial_\varrho \Phi\left(\frac{1}{R\partial_\varrho R}\right) = 0.$$



(a) (b)

(c) (d)

Fig. 5. (a) *Numerical values for* $X(t,x)$ *for* $t = 0.05, 0.1, 0.2, 0.5, 0.88$*;* (b) *scaled initial data* $f(t=0,.)$ (*dotted line*) *and stationary solution* f_∞ (*solid line*)*;* (c) *evolution of* $E(t)$ *with time* t*;* (d) *evolution of* Δt *with time* t.

A third possible extension is suggested in [4, 5] for “viscous pressureless gas equations”; in this case, the pseudoinverse evolves according to

$$\partial_t X = v_0(\varrho) - \partial_\varrho(\Phi(1/\partial_\varrho X)),$$

where Φ can be computed explicitly from the Eulerian viscosity term. The function $\Phi(u) = u\ln(u)$ is of special interest as it corresponds to a linear perturbation in Eulerian coordinates. In this context, $v_0(\varrho)$ is implicitly defined from the initial velocity in Eulerian coordinates via $v_0(\varrho) = u_0 \circ X(t=0,\varrho)$.

Acknowledgments. The first author warmly thanks the Mathematics Department of the University of Pavia for its hospitality and C. Salani.

REFERENCES

[1] T. Arbogast, M. F. Wheeler, and N. Y. Zhang, *A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow through porous media,* SIAM J. Numer. Anal., 33 (1996), pp. 1669–1687.

[2] D. Aregba-Driollet, R. Natalini, and S. Tang, *Diffusive kinetic explicit schemes for nonlinear degenerate parabolic systems,* Math. Comp., 73 (2004), pp. 63–94.

[3] A. Arnold, P. A. Markowich, G. Toscani, and A. Unterreiter, *On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker–Planck equations,* Commun. Partial Differential Equations, 26 (2001), pp. 43–100.

[4] Y. Brenier, *Hydrodynamic structure of the augmented Born–Infeld equations,* Arch. Ration. Mech. Anal., 172 (2004), pp. 65–91.

[5] Y. Brenier, *Order Preserving Vibrating Strings and Applications to Electrodynamics and Magnetohydrodynamics,* preprint.

[6] A. E. Berger, H. Brezis, and J. C. W. Rogers, *A numerical method for solving the problem $u_t - \Delta f(u) = 0$,* RAIRO Anal. Numer., 13 (1979), pp. 297–312.

[7] J. A. Carrillo, M. P. Gualdani, and G. Toscani, *Finite speed of propagation in porous media by mass transportation methods,* C. R. Acad. Sci. Paris Ser. I, 338 (2004), pp. 815–818.

[8] J. A. Carrillo, A. Jüngel, P. A. Markowich, G. Toscani, and A. Unterreiter, *Entropy dissipation methods for degenerate parabolic problems and generalized Sobolev inequalities,* Monatsh. Math., 133 (2001), pp. 1–82.

[9] J. A. Carrillo and G. Toscani, *Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity,* Indiana Univ. Math. J., 49 (2000), pp. 113–142.

[10] J. A. Carrillo and G. Toscani, *Wasserstein Metric and Large-Time Asymptotics of Nonlinear Diffusion Equations,* preprint.

[11] C. Ebmeyer, *Error estimates for a class of degenerate parabolic equations,* SIAM J. Numer. Anal., 35 (1998), pp. 1095–1113.

[12] L. Gosse and G. Toscani, *Space localization and well-balanced schemes for discrete kinetic models in the diffusive limit,* SIAM J. Numer. Anal., 41 (2003), pp. 641–658.

[13] L. Gosse and G. Toscani, *Asymptotic-preserving and well-balanced schemes for radiative transfer and the Rosseland approximation,* Numer. Math., 198 (2004), pp. 223–250.

[14] W. Jäger and J. Kačur, *Solution of porous medium type systems by linear approximation schemes,* Numer. Math., 60 (1991), pp. 407–427.

[15] S. Jin, L. Pareschi, and G. Toscani, *Diffusive relaxation schemes for multiscale discrete velocity kinetic equations,* SIAM J. Numer. Anal., 35 (1998), pp. 2405–2439.

[16] J. Kačur, *Solution of free boundary problems by relaxation schemes,* SIAM J. Numer. Anal., 36 (1999), pp. 290–316.

[17] J. Kačur, A. Handlovičová, and M. Kačurová, *Solution of nonlinear diffusion problems by linear approximation schemes,* SIAM J. Numer. Anal., 30 (1993), pp. 1703–1722.

[18] K. Mikula, *Numerical solution of nonlinear diffusion with finite extinction phenomenon,* Acta Math. Univ. Comenian., 64 (1995), pp. 173–184.

[19] T. Nakaki and K. Tomoeda, *A finite difference scheme for some nonlinear diffusion equations in a absorbing medium: Support splitting phenomena,* SIAM J. Numer. Anal., 40 (2002), pp. 945–964.

[20] R. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration,* SIAM J. Numer. Anal., 25 (1988), pp. 784–814.

[21] R. Nochetto, A. Schmidt, and C. Verdi, *A posteriori error estimation and adaptivity for degenerate parabolic problems,* Math. Comp., 69 (1999), pp. 1–24.

[22] K. Oelschläger, *Simulation of the solution of a viscous porous medium equation by a particle method,* SIAM J. Numer. Anal., 40 (2002), pp. 1716–1762.

[23] I. S. Pop and W. A. Yong, *A numerical approach to degenerate parabolic equations,* Numer. Math., 92 (2002), pp. 357–381.

[24] G. Toscani, *A central limit theorem for solutions of the porous medium equation,* J. Evol. Equ., 5 (2005), pp. 185–203.

[25] J. L. Vazquez, *An introduction to the mathematical theory of the porous medium equation,* in Shape Optimization and Free Boundaries, Kluwer, Montreal, PQ, Canada, 1990, pp. 347–389.

[26] J. L. Vazquez, *Asymptotic behavior for the porous medium equation posed in the whole space,* J. Evol. Equ., 3 (2003), pp. 67–118.

[27] J. L. Vazquez, *Asymptotic behavior and propagation properties of the one-dimensional flow of gas in a porous medium,* Trans. Amer. Math. Soc., 277 (1983), pp. 507–527.

[28] C. Villani, *Topics in Optimal Mass Transportation,* Grad. Stud. Math. 58, AMS, Providence, RI, 2002.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2607–2623

GALERKIN METHODS BASED ON HERMITE SPLINES FOR SINGULAR PERTURBATION PROBLEMS[*]

SONG-TAO LIU[†] AND YUESHENG XU[†‡]

Abstract. We develop Galerkin methods for solving the singularly perturbed two-point boundary value problem of high-order elliptic differential equations. These methods are based on Hermite splines with knots adapted to the singular behavior of the solution of the problem. We prove an optimal order of uniform convergence for the method with respect to the perturbation parameter. Specifically, we present a sufficient condition on the mesh of grid points that ensures the corresponding approximate solution has the optimal order of uniform convergence in the energy norm. We also construct *optimal* meshes that satisfy the sufficient condition. Numerical examples are presented to illustrate the method and the corresponding theoretical estimates.

Key words. singular perturbation, Galerkin methods, Hermite splines, grid meshes, optimal order of uniform convergence

AMS subject classifications. 65L10, 65L12, 65L60

DOI. 10.1137/040607411

1. Introduction.

The numerical solution of the singularly perturbed two-point boundary value problem is a challenging task. The effect of the boundary layers makes it difficult to develop numerical methods for solving the problem with an optimal order of uniform convergence (cf. [B, LT, O, SO1, SO2, TT]). Many authors have tried to tackle this difficulty. Bakhvalov [B] introduced special grids based on mesh generating functions and used them to develop numerical methods for solving the problems. Uniformly convergent classical difference schemes on the special meshes may be found in [G, V]. The Shishkin mesh is one of the simplest piecewise equidistant meshes. The uniform convergence of the Shishkin mesh has been discussed in [MOS, RST, S]. For the Shishkin mesh, Sun and Stynes [SS1, SS2] provided almost optimal uniform convergence results for the finite element methods for the singularly perturbed high-order elliptic two-point boundary value problem based on piecewise polynomial approximations. Their results show that the traditional finite element methods using the Shishkin mesh are suitable for the high-order singularly perturbed problems. Tang and Trummer [TT] proposed a pseudospectral methods for treating the boundary layer problem for the singular perturbation problem. A recent paper of Roos and Linss [RL] studied convergence properties of the simple upwind difference scheme and a Galerkin method on generalized Shishkin grids. Sufficient conditions on the mesh-characterizing function for the convergence of the method, uniformly with respect to the perturbation parameter, are proposed. The idea was extended in [L] to the two-dimensional case.

[*]Received by the editors April 27, 2004; accepted for publication (in revised form) April 26, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/sinum/43-6/60741.html

[†]Department of Mathematics, 215 Carnegie Hall, Syracuse University, Syracuse, NY 13244-1150 (sliu07@syr.edu, yxu06@syr.edu).

[‡]Institute of Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing 100080, People's Republic of China. This author was supported in part by NSF grants DMS-9973427 and CCR-0407476, by Natural Science Foundation of China grant 10371122, and by the Chinese Academy of Sciences under the program "One Hundred Distinguished Young Chinese Scientists."

We introduce in this paper an optimal Galerkin method for solving the singularly perturbed high-order elliptic two-point boundary value problem of reaction-diffusion type using Hermite splines with knots adapted to the boundary layer behavior of the solution. The main results of this paper include a simple sufficient condition on the mesh-sizes which ensures the *optimal* order of *uniform* convergence of the Galerkin method based on the Hermite splines built on the mesh. The order is the same as the approximation order of the Hermite spline space with *uniform* knots. We also construct an optimal mesh which realizes the sufficient condition. In this construction, we combine ideas from [QS, QST, RST, SS1, SS2, V] and those from [CHX]. That is, we choose the mesh-size for each of the subintervals so that the error of the Hermite spline approximation on each subinterval is bounded in an optimal order uniformly with respect to the perturbation parameter. In this development, we use estimates of the derivatives of the exact solution to guide the design of the mesh.

This paper is organized into five sections. In section 2, we outline a setting of reaction-diffusion-type problems and describe the Galerkin methods for solving such problems. We propose in section 3 a sufficient condition on the mesh-size which ensures the optimal order of uniform convergence of the approximate solution. In section 4, we construct a specific mesh which satisfies the condition and generates a Hermite spline approximate solution having the optimal order of uniform convergence. Finally, in section 5, we present numerical experiments that demonstrate the methods and confirm the theoretical estimates.

2. The Galerkin method based on Hermite splines of arbitrary knots. We describe in this section the Galerkin method for solving reaction-diffusion problems using the Hermite splines of arbitrary knots. For this purpose, we describe the singularly perturbed high-order elliptic two-point boundary value problem of reaction-diffusion type. Let $m \geq 2$ be an integer, let $\varepsilon \in (0, 1]$ be a perturbation parameter, and let a_j, $j \in Z_{2(m-1)+1}$ with $Z_m := \{0, 1, \ldots, m-1\}$, and f be sufficiently smooth functions defined on $I := [0, 1]$. We introduce the differential operator L_ε by

$$\begin{aligned} L_\varepsilon u := {} & (-1)^m \varepsilon^2 u^{(2m)} + (-1)^{m-1}\big(a_{2(m-1)} u^{(m-1)}\big)^{(m-1)} \\ & + \sum_{k=2}^{m} (-1)^{m-k}\big(a_{2(m-k)+1} u^{(m-k+1)} + a_{2(m-k)} u^{(m-k)}\big)^{(m-k)} \end{aligned}$$

and consider the boundary value problem

$$\begin{aligned} (L_\varepsilon u)(x) &= f(x), \qquad x \in (0, 1), \\ u^{(j)}(0) = u^{(j)}(1) &= 0, \qquad j \in Z_m. \end{aligned} \tag{2.1}$$

We denote by $L_2(I)$ the space of real-valued square integrable functions on I with the associated inner product $(\cdot, \cdot)$. Let $H^k(I)$, $k = 1, 2, \ldots, m$, be the Sobolev spaces on I with the norm $\|\cdot\|_k$ and the seminorm $|\cdot|_k$ (see, for example, [C]). For convenience, we let $H^0(I) := L_2(I)$ and $\|\cdot\|_0 := \|\cdot\|_{L_2(I)}$. We denote by $\|\cdot\|_\infty$ the essential maximum norm on $L_\infty(I)$ and by $\|\cdot\|_{k,\infty}$ the maximum norm on $C^k(I)$ for $k \in Z_m$, i.e., $\|u\|_{k,\infty} := \sum_{j=0}^{k} \|u^{(j)}\|_\infty$. Let $H_0^m := H_0^m(I)$ be the closure of the set $\{v \in C^m(I) : v^{(k)}(0) = v^{(k)}(1) = 0,\ k \in Z_m\}$ in the Sobolev norm $\|\cdot\|_m$. The energy norm is defined by $\|v\|_\varepsilon := \{\varepsilon^2 |v|_m^2 + \|v\|_{m-1}^2\}^{1/2}$.

The bilinear form $A_\varepsilon(\cdot,\cdot)$ is defined by

$$A_\varepsilon(u,v) := \left(\varepsilon^2 u^{(m)}, v^{(m)}\right) + \left(a_{2(m-1)}u^{(m-1)}, v^{(m-1)}\right) \\ + \sum_{k=2}^{m} \left(a_{2(m-k)+1}u^{(m-k+1)} + a_{2(m-k)}u^{(m-k)}, v^{(m-k)}\right).$$

We now present a sufficient condition that guarantees the coercivity of the bilinear form. To this end, we introduce an index set by $I^+ := \{j : \alpha_j \geq 0,\ j \in Z_m\}$, where constants $\alpha_{m-1} \leq a_{2(m-1)}$ and $\alpha_{m-k} \leq a_{2(m-k)} - \frac{1}{2}a'_{2(m-k)+1}$, $k = 2, 3, \ldots, m$. We assume that there is a decomposition for α_j, $j \in I^- := Z_m \setminus I^+$, that is,

$$\alpha_j = \sum_{k \in I^+ \cap \{j+1, j+2, \ldots, m-1\}} \alpha_{j,k}$$

such that $\eta_k \geq 0$, for $k \in I^+ \setminus \{m-1\}$, and $\eta_{m-1} > 0$, where

$$\eta_k := \alpha_k + \sum_{j \in I^-, j<k} \alpha_{j,k} 2^{-(k-j)} \quad \text{for} \quad k \in I^+.$$

Then, by making use of the estimate

$$|v|_{s-1}^2 \leq \frac{|v|_s^2}{2}, \quad v \in H_0^m(I), \quad s \in N_m := \{1, 2, \ldots, m\},$$

we have the coercivity of the bilinear form $A_\varepsilon(\cdot,\cdot)$, i.e.,

$$A_\varepsilon(v,v) \geq \varepsilon^2 |v|_m^2 + \eta_{m-1}|v|_{m-1}^2.$$

By integration by parts, we rewrite boundary value problem (2.1) in a variational form in which we seek $u \in H_0^m$ such that

$$A_\varepsilon(u,v) = (f,v) \ \text{ for all } \ v \in H_0^m. \tag{2.2}$$

The solution of (2.2) is a weak solution of (2.1). Equation (2.2) will serve as the basic equation for numerical computation of the solution. Under the hypotheses on a_j, it can be proved that there exist positive constants c_1, c_2 such that

$$|A_\varepsilon(v,w)| \leq c_1 \|v\|_\varepsilon \|w\|_\varepsilon, \qquad v, w \in H_0^m, \tag{2.3}$$

and

$$A_\varepsilon(v,v) \geq c_2 \|v\|_\varepsilon^2, \qquad v \in H_0^m. \tag{2.4}$$

Thus, the bilinear form $A_\varepsilon(\cdot,\cdot)$ is coercive with respect to $\|\cdot\|_\varepsilon$. By the Lax–Milgram theorem, existence and uniqueness of the solution of (2.2) are guaranteed. Since the coefficient functions a_k, $k \in Z_{2(m-1)+1}$, and f are sufficiently smooth, we know that the solution is also sufficiently smooth [Gr, GT], and thus it is identical to the classical solution. For this reason, we will not distinguish the weak solution from the classical one.

We now introduce the space of the Hermite splines. Let N be a positive integer and let $0 = x_0 < x_1 < \cdots < x_N = 1$ be an arbitrary mesh of the interval I with $h_i := x_i - x_{i-1}$, $i \in \mathbb{N}_N$, and $h := \max\{h_i : i \in \mathbb{N}_N\}$. We denote by I_i the interval

$[x_{i-1}, x_i)$. For a positive integer r, we let $P_{2r}(I_i)$ denote the space of polynomials of degree $2r-1$ on I_i. The space of the Hermite splines is defined by

$$V_N := \{v(x) \in H_0^r(I) : v|_{I_i} \in P_{2r}(I_i),\ i \in \mathbb{N}_N\}. \tag{2.5}$$

It can be seen that $V_N \subset H_0^r(I)$ and by the Sobolev embedding theorem, $V_N \subset C_0^{r-1}(I)$. The dimension of V_N is $r(N-1)$. For $v \in C_0^{r-1}(I)$ we denote by Πv its Hermite spline interpolant from V_N satisfying the conditions $(\Pi v)^{(k)}(x_i) = v^{(k)}(x_i)$, $k \in Z_r$, $i \in \mathbb{N}_{N-1}$. It is well known (cf. [Sc]) that there exists a positive constant c such that for all $v \in C_0^{2r}(I_i)$,

$$|v - \Pi v|_{j,\infty,I_i} \leq c h_i^{2r-j} |v|_{2r,\infty,I_i}, \qquad j \in Z_{2r+1}, \tag{2.6}$$

where $|v|_{k,\infty,I_i}$ denotes the maximum norm of $v^{(k)}$ on the interval I_i. Here and in what follows, constant c is used to denote the generic positive constant that is independent of ε and the mesh.

In order to have a conforming Galerkin method, we require $r \geq m$ through out the rest of this paper. The Galerkin method based on the Hermite splines is to seek $u_N \in V_N$ such that

$$A_\varepsilon(u_N, v) = (f, v), \qquad v \in V_N. \tag{2.7}$$

Recalling that u satisfies (2.2), we have that $A_\varepsilon(u - u_N, v) = 0$, $v \in V_N$. By Cea's lemma [C], this equation implies that

$$\|u - u_N\|_\varepsilon \leq c \inf_{v \in V_N} \|u - v\|_\varepsilon. \tag{2.8}$$

We next consider the *discrete* Galerkin method, which takes into account the effect of numerical integration required for computing the inner products. We let a_k^N, $k \in Z_{2(m-1)+1}$, and f^N be the Hermite spline interpolation of degree $2\ell - 1$ to a_k and f, respectively, with uniform knots $\frac{j}{N}$, $j \in Z_{N+1}$. Then we have the approximation orders

$$\|a_k^N - a_k\|_\infty \leq c N^{-2\ell} \|a_k\|_{2\ell,\infty} \quad \text{and} \quad \|f^N - f\|_\infty \leq c N^{-2\ell} \|f\|_{2\ell,\infty}. \tag{2.9}$$

The *discrete* bilinear form is

$$\begin{aligned} A_\varepsilon^N(u,v) := {}& \big(\varepsilon^2 u^{(m)}, v^{(m)}\big) + \big(a_{2(m-1)}^N u^{(m-1)}, v^{(m-1)}\big) \\ &+ \sum_{k=2}^{m} \big(a_{2(m-k)+1}^N u^{(m-k+1)} + a_{2(m-k)}^N u^{(m-k)}, v^{(m-k)}\big). \end{aligned}$$

Correspondingly, the discrete Galerkin method is to seek $u_N^N \in V_N$ such that

$$A_\varepsilon^N(u_N^N, v) = (f^N, v), \qquad v \in V_N. \tag{2.10}$$

It was proved in Lemma 4.1 of [SS1] that $A_\varepsilon^N(\cdot,\cdot)$ is coercive with respect to $\|\cdot\|_\varepsilon$. This leads to the following estimate of error between u_N and u_N^N.

THEOREM 2.1. *Let u_N and u_N^N be the solution of* (2.7) *and* (2.10), *respectively. Suppose that a_k and f are approximated by a_k^N and f^N, respectively, with errors* (2.9). *Then*

$$\|u_N - u_N^N\|_\varepsilon \leq c N^{-2\ell} (\|f\|_{2\ell,\infty} + \|u\|_\varepsilon).$$

Proof. To prove this theorem, we set $B_\varepsilon(v) := A_\varepsilon^N(u_N^N, v) - A_\varepsilon(u_N, v)$ and observe from (2.7) and (2.10) that

$$B_\varepsilon(v) = (f^N - f, v), \qquad v \in V_N. \tag{2.11}$$

Using the definition of A_ε and A_ε^N with a rearrangement of terms leads to the equation

$$\begin{aligned} B_\varepsilon(v) = A_\varepsilon^N(u_N^N - u_N, v) &+ \sum_{k=1}^{m} \big((a_{2(m-k)}^N - a_{2(m-k)}) u_N^{(m-k)}, v^{(m-k)} \big) \\ &+ \sum_{k=2}^{m} \big((a_{2(m-k)+1}^N - a_{2(m-k)+1}) u_N^{(m-k+1)}, v^{(m-k)} \big). \end{aligned}$$

Letting $v = u_N^N - u_N$ in the equation above and using the coercivity of A_ε^N and the first estimate of (2.9), we conclude that

$$c\|u_N - u_N^N\|_\varepsilon^2 - cN^{-2\ell}\|u_N\|_{m-1}\|u_N - u_N^N\|_\varepsilon \le B_\varepsilon\big(u_N^N - u_N\big).$$

On the other hand, by (2.11) and the second estimate of (2.9), we have that

$$B_\varepsilon\big(u_N^N - u_N\big) = \big(f^N - f, u_N^N - u_N\big) \le cN^{-2\ell}\|f\|_{2\ell,\infty}\|u_N^N - u_N\|_\varepsilon.$$

Combining the above two estimates, we obtain that

$$\|u_N - u_N^N\|_\varepsilon \le cN^{-2\ell}(\|f\|_{2\ell,\infty} + \|u_N\|_{m-1}). \tag{2.12}$$

Noting that

$$\|u_N\|_{m-1} \le \|u_N\|_\varepsilon \le \|u - u_N\|_\varepsilon + \|u\|_\varepsilon \le \|u\|_\varepsilon + c \inf_{v \in V_N} \|u - v\|_\varepsilon \le c\|u\|_\varepsilon,$$

the desired estimate follows from (2.12). □

3. The optimal order of uniform convergence. In this section, we derive a sufficient condition on the meshes having $O(N)$ number of grid points that ensures the corresponding approximate solution u_N having optimal order of uniform convergence. For this purpose, we recall properties of the solution for the singularly perturbed boundary value problems (2.1). From the book [O], we may write the solution u of problem (2.1) in terms of sufficiently differentiable functions E, F, G for which

$$u = E + F + G,$$

and such that for all $x \in I$ and for $j \in \mathbb{N} = \{0, 1, \dots\}$,

$$\begin{aligned} |G^{(j)}(x)| \le c, \quad |E^{(j)}(x)| &\le c\varepsilon^{m-1-j} \exp(-\alpha x/\varepsilon), \\ |F^{(j)}(x)| &\le c\varepsilon^{m-1-j} \exp(-\alpha(1-x)/\varepsilon). \end{aligned} \tag{3.1}$$

Here, constant $\alpha = \alpha_{m-1}$. Clearly, functions G, E, F describe, respectively, the reduced problem solution and two boundary layers at endpoints 0, 1.

Based on estimates (3.1), we introduce a generating function defined by

$$h^0(x) := \frac{\varepsilon}{N} \exp\left(\frac{\alpha x}{2r\varepsilon}\right), \tag{3.2}$$

and choose the mesh-sizes h_i, $i \in \mathbb{N}_{\tilde{N}}$, to satisfy the condition that

$$h_i \leq \min\{h^0(x_{i-1}), h^0(1-x_i), 1/N\} \tag{3.3}$$

and

$$\tilde{N} \leq cN. \tag{3.4}$$

Meshes having this property are called *optimal meshes*. A construction of optimal meshes will be presented in the next section. The construction of these meshes is motivated by the equal distribution of errors in the subintervals. We remark that these meshes improve the convergence order of the well-known Bakhvalov meshes and Shishkin meshes. We will prove that they lead to the optimal order of convergence and linear complexity. Recall that the Bakhvalov-type meshes map equidistant grids to nonuniform meshes by using the layer functions or their modifications. The Shishkin meshes, which are piecewise uniform meshes, are particularly simple in their constructions. It is not clear in the literature from a theoretical point of view if the Bakhvalov-type meshes give the optimal order of uniform convergence. Nonetheless, it is known from [SS1, SS2] that the Shishkin meshes give the optimal uniform convergence order *up to a logarithmic factor*.

Lemma 3.1. *Let E, F, and G be functions satisfying the condition* (3.1) *and let h_i be chosen according to* (3.3). *Then*

$$\|E - \Pi E\|_j \leq cN^{-2r+j}, \quad j \in Z_r, \quad \|E - \Pi E\|_\varepsilon \leq cN^{-2r+m}, \tag{3.5}$$

$$\|F - \Pi F\|_j \leq cN^{-2r+j}, \quad j \in Z_r, \quad \|F - \Pi F\|_\varepsilon \leq cN^{-2r+m}, \tag{3.6}$$

and

$$\|G - \Pi G\|_j \leq cN^{-2r+j}, \quad j \in Z_r, \quad \|G - \Pi G\|_\varepsilon \leq cN^{-2r+m}\max\{\varepsilon, N^{-1}\}. \tag{3.7}$$

Proof. The proof for (3.7) is standard. We present only the proof for the estimate for E, since the one for F is obtained similarly. To this end, we first estimate $|E - \Pi E|_{m,\infty,I_i}$, $i \in \mathbb{N}_{\tilde{N}}$. By estimate (2.6), there exists a positive constant c such that

$$\varepsilon|E - \Pi E|_{m,\infty,I_i} \leq c\varepsilon|E|_{2r,\infty,I_i} h_i^{2r-m}. \tag{3.8}$$

Using (3.3), we obtain that

$$h_i \leq h^0(x_{i-1}) = \frac{\varepsilon}{N}\exp\left(\frac{\alpha x_{i-1}}{2r\varepsilon}\right). \tag{3.9}$$

Substituting the bound on $|E|_{2r,\infty,I_i}$ and (3.9) into the right-hand side of (3.8) yields

$$\varepsilon|E - \Pi E|_{m,\infty,I_i} \leq c\varepsilon(\varepsilon^{m-1-2r})\exp\left(-\frac{\alpha x_{i-1}}{\varepsilon}\right)\left[\frac{\varepsilon}{N}\exp\left(\frac{\alpha x_{i-1}}{2r\varepsilon}\right)\right]^{2r-m} \leq cN^{-2r+m}.$$

Next, we estimate $|E - \Pi E|_{j,\infty,I_i}$ for $j \in Z_r$. By using Lemma 2.6, the bound (3.9) on h_i and the hypothesis on the function E, we have that

$$|E - \Pi E|_{j,\infty,I_i} \leq |E|_{2r,\infty,I_i} h_i^{2r-j} \leq cN^{-2r+j}\varepsilon^{m-j}\exp\left(-\frac{j\alpha x_{i-1}}{2r\varepsilon}\right)$$

for $j \in Z_r$. Thus, we have for $j \in Z_r$ that

$$|E - \Pi E|_{j,\infty,I_i} \le cN^{-2r+j}. \tag{3.10}$$

It follows from (3.10) that

$$\|(E - \Pi E)^{(j)}\|_2 \le \left(\sum_{i=1}^{\tilde{N}} |E - \Pi E|^2_{j,\infty,I_i} |I_i| \right)^{1/2} \le cN^{-2r+j},$$

which yields the desired results. □

We now use Lemma 3.1 to estimate the error of interpolation Πu, where u is the solution of problem (2.1).

Proposition 3.2. *Let u be the solution of problem* (2.1) *and let h_i, $i \in \mathbb{Z}_{\tilde{N}+1}$, be chosen according to* (3.3). *Then there exists a positive constant c independent of ε or N such that*

$$\|u - \Pi u\|_j \le cN^{-2r+j}, \quad j \in Z_r, \quad \textit{and} \quad \|u - \Pi u\|_\varepsilon \le cN^{-2r+m}. \tag{3.11}$$

Proof. Since the solution u of problem (2.1) has the form $u = E + F + G$, where functions E, F, and G satisfy condition (3.1), using the estimates in Lemma 3.1 for the functions E, F, and G, respectively, we conclude the desired estimate in the proposition. □

Our next task is to establish an error bound on $\|\Pi u - u_N\|_\varepsilon$.

Lemma 3.3. *Let u and u_N be the solutions of problems* (2.2) *and* (2.7), *respectively. Suppose that the mesh satisfies condition* (3.3). *Then*

$$\|\Pi u - u_N\|_\varepsilon \le cN^{-2r+m}.$$

Moreover, if $r = m$, then

$$\|\Pi u - u_N\|_{m-1} \le cN^{-m-1} \quad \textit{and} \quad |\Pi u - u_N\|_\varepsilon \le cN^{-m-1}.$$

Proof. To prove the first estimate, we note that $A_\varepsilon(u - u_N, \Pi u - u_N) = 0$, and observe that

$$\|\Pi u - u_N\|^2_\varepsilon \le cA_\varepsilon(\Pi u - u_N, \Pi u - u_N) = cA_\varepsilon(\Pi u - u, \Pi u - u_N). \tag{3.12}$$

Inequality (3.12) implies that $\|\Pi u - u_N\|_\varepsilon \le c\|\Pi u - u\|_\varepsilon$. By Proposition 3.2, we conclude the first estimate.

To obtain the special results when r is chosen as m, we estimate $A_\varepsilon(\Pi u - u, \Pi u - u_N)$. To do this, we first prove that

$$\big((\Pi u - u)^{(m)}, (\Pi u - u_N)^{(m)}\big) = 0. \tag{3.13}$$

Recall that the function $\Pi u - u$ has the interpolation property that $(\Pi u - u)^{(j)}(x_i) = 0$, $j \in Z_m$, $i \in Z_{\tilde{N}+1}$. Using integration by parts with this interpolation property, we conclude for any polynomial p of degree $m - 1$ on interval I_i that

$$\int_{x_{i-1}}^{x_i} (\Pi u - u)^{(m)}(x) p(x)\, dx = \int_{x_{i-1}}^{x_i} (\Pi u - u)(x) p^{(m)}(x)\, dx = 0.$$

Since $(\Pi u - u_N)^{(m)}|_{I_i}$ is a polynomial of degree $m-1$, $i \in \mathbb{N}_{\tilde{N}}$, we see that (3.13) holds. Combining (3.13) and the formula of $A_\varepsilon(\Pi u - u, \Pi u - u_N)$, we obtain that

$$\begin{aligned} A_\varepsilon(\Pi u - u, \Pi u - u_N) = {} & \big(a_{2(m-1)}(\Pi u - u)^{(m-1)}, (\Pi u - u_N)^{(m-1)}\big) \\ & + \sum_{k=2}^{m} \big(a_{2(m-k)}(\Pi u - u)^{(m-k)} + a_{2(m-k)+1} \\ & \times (\Pi u - u)^{(m-k+1)}, (\Pi u - u_N)^{(m-k)}\big). \end{aligned}$$

It follows that

$$A_\varepsilon(\Pi u - u, \Pi u - u_N) \le c\|\Pi u - u\|_{m-1}\|\Pi u - u_N\|_{m-1}.$$

Using the first estimate of Proposition 3.2, we find that

$$A_\varepsilon(\Pi u - u, \Pi u - u_N) \le cN^{-m-1}\|\Pi u - u_N\|_\varepsilon. \tag{3.14}$$

Combining (3.12) and (3.14) proves this third estimate. The second estimate follows directly from the third. ∎

Next we estimate the errors $u - u_N$ and $u - u_N^N$.

THEOREM 3.4. *Let u, u_N, and u_N^N be the solutions of problems* (2.2), (2.7), *and* (2.10)*, respectively. Suppose that the mesh satisfies condition* (3.3). *Then*

$$\|u - u_N\|_\varepsilon \le cN^{-2r+m}, \qquad \|u - u_N^N\|_\varepsilon \le c(N^{-2r+m} + N^{-2\ell}).$$

Moreover, if $r = m$, then

$$\begin{aligned} &\|u - u_N\|_{m-1} \le cN^{-m-1}, \quad \|u - u_N^N\|_\varepsilon \le c(N^{-m} + N^{-2\ell}), \\ &\|u - u_N^N\|_{m-1} \le c(N^{-m-1} + N^{-2\ell}). \end{aligned}$$

Proof. The estimate on $\|u - u_N\|_\varepsilon$ is a direct consequence of Proposition 3.2 and Lemma 3.3, while the estimate on $\|u - u_N^N\|_\varepsilon$ is obtained by using the first estimate of this theorem and Theorem 2.1.

Next, we prove the special results when we choose $r = m$. According to Proposition 3.2, if $r = m$, we have that $\|u - \Pi u\|_{m-1} \le cN^{-m-1}$. Combining this estimate with the first estimate in Proposition 3.2, we arrive at the inequality

$$\|u - u_N\|_{m-1} \le \|u - \Pi u\|_{m-1} + \|\Pi u - u_N\|_{m-1} \le N^{-m-1}.$$

The estimates on $\|u - u_N^N\|_\varepsilon$ follow directly from the second estimate of this theorem with $r = m$. To prove the last estimate, we note that

$$\|u - u_N^N\|_{m-1} \le \|u - u_N\|_{m-1} + \|u_N - u_N^N\|_{m-1} \le \|u - u_N\|_{m-1} + \|u_N - u_N^N\|_\varepsilon.$$

The third estimate in this theorem and the result of Theorem 2.1 ensure the last estimate. ∎

Theorem 3.4 suggests that the Hermite spline approximations of an appropriate order for a_k and f must be used to ensure optimal order of approximation of u_N^N.

4. A construction of optimal meshes. We describe a specific construction of a mesh of interval I and prove that it is optimal in the sense that it satisfies both conditions (3.3) and (3.4).

We define the grid points according to the generating function h^0 by the recursive formula

$$x_0 := 0, \quad x_i := x_{i-1} + h^0(x_{i-1}), \quad i = 1, 2, \ldots, M, \tag{4.1}$$

where M is a positive integer such that

$$h^0(x_{M-1}) < 1/N \quad \text{and} \quad h^0(x_M) \geq 1/N. \tag{4.2}$$

We consider two cases. In the first case where $x_M < 1/2$, $x_0 < x_1 < \cdots < x_M$ forms a mesh for the interval $[0, x_M]$ and by symmetry, $1 - x_M < 1 - x_{M-1} < \cdots < 1 - x_0$ forms a mesh for the interval $[1 - x_M, 1]$. For the middle interval $[x_M, 1 - x_M]$, we define a uniform mesh with $N_1 := \lfloor (1 - 2x_M)N \rfloor + 1$ grid points such that the mesh-size $\tilde{h} := (1 - 2x_M)/N_1 \leq 1/N$, where $\lfloor x \rfloor$ denotes the largest integer not greater than x. Thus, points

$$x_j = x_{j-1} + \tilde{h}, \quad j = M+1, M+2, \ldots, M+N_1,$$

subdivide the middle interval. Set $x_{M+N_1+j} := 1 - x_{M-j}$, $j \in Z_{M+1}$. Hence the mesh

$$\Delta : \quad 0 = x_0 < x_1 < \cdots < x_{2M+N_1} = 1$$

is the desired mesh for I, which consists of three parts, the left and right meshes for the boundary layers and the middle uniform mesh. We now describe a construction when the second case $x_M \geq 1/2$ occurs. In this case, there is no middle uniform mesh. Note that there exists a nonnegative integer μ such that

$$x_{M-\mu-1} < \frac{1}{2} \leq x_{M-\mu} < \cdots < x_M.$$

We let $M' := M - \mu$ and observe that

$$x_{M'-1} < 1/2 \quad \text{and} \quad x_{M'} \geq 1/2. \tag{4.3}$$

We use M' to replace M in (4.1). Clearly, the integer M' satisfying condition (4.3) is less than or equal to the integer M satisfying condition (4.2). We redefine $x_{M'} := 1/2$ and $x_{M'+j} = 1 - x_{M'-j}$, $j \in \mathbb{N}_{M'}$. Thus, in this case

$$\Delta : 0 = x_0 < x_1 < \cdots < x_{2M'} = 1$$

is a desired mesh for interval I.

The main purpose of this section is to show that the Hermite spline interpolation developed based on the mesh constructed above has an optimal order of convergence and the *linear* order of computational complexity measured by the number of grid points. For this purpose, we consider a sequence x_n generated by a given generating function g. Suppose that y_0 is given and g is a real-valued function defined on the real line. We define a sequence x_n by letting

$$x_0 = y_0, \quad x_n = x_{n-1} + g(x_{n-1}), \quad n \in \mathbb{N}.$$

We call x_n the sequence generated by y_0 and g. The next lemma is concerned with this sequence.

LEMMA 4.1. *Suppose that $y_0 < y_1$ and g is a real-valued nondecreasing function on R. Let x_n be the sequence generated by y_0 and g. If $g(y_0) > 0$, then there exists a unique integer M_g such that*

$$x_{M_g} \leq y_1 \quad \textit{and} \quad x_{M_g} + g(x_{M_g}) > y_1.$$

Proof. By the definition of sequence x_n and the hypothesis on g, we find that x_n is a strictly increasing sequence. Set $g_0 := g(y_0)$ and observe by induction that $x_n \geq x_0 + ng_0$ for $n \in \mathbb{N}$. Thus, by hypothesis that $g_0 > 0$ we obtain that $\lim_{n\to\infty} x_n = \infty$. It follows that there exists a positive integer n such that $x_n > y_1$. Choose n_0 as the smallest of such integers n. We conclude that $x_{n_0} > y_1$ and $x_{n_0-1} \leq y_1$, proving the lemma. ◻

We will call (y_0, y_1, g) an admissible triple if $y_0 < y_1$, g is nondecreasing and $g(y_0) > 0$, and call M_g the index determined by the admissible triple in the last lemma. Let $\mathcal{A}$ denote the class of admissible triples. We define a mapping $\mathcal{M}$ from $\mathcal{A}$ to positive integers by assigning to each triple the index M_g and denote it by

$$M_g = \mathcal{M}(y_0, y_1, g). \tag{4.4}$$

Lemma 4.1 ensures that the integer M_g is well defined. In the next lemma, we compare two indices M_{g_1} and M_{g_2} corresponding to two generating functions g_1 and g_2.

LEMMA 4.2. *Suppose that (y_0, y_1, g_1) and (y_0, y_1, g_2) are two admissible triples. If $g_1 \leq g_2$, then*

$$\mathcal{M}(y_0, y_1, g_1) \geq \mathcal{M}(y_0, y_1, g_2).$$

Proof. Let $M_{g_i} := \mathcal{M}(y_0, y_1, g_i)$ for $i = 1, 2$. Suppose that for each $i = 1, 2$, x_n^i, $n \in Z_{M_{g_i}+1}$, there are two sequences generated by y_0 and g_i. It can be shown by induction on n that $x_n^1 \leq x_n^2$, $n \in \mathbb{N}$. Thus, the result of this lemma follows. ◻

We next consider a specific generating function g which is an affine function.

LEMMA 4.3. *Suppose that $y_0 < y_1$ and for $k > 0$ and $h_0 := ky_0 + b > 0$, and define $g(x) = kx + b$. Let $M_g := \mathcal{M}(y_0, y_1, g)$. Then*

$$\left(\frac{(1+k)^{M_g} - 1}{k}\right) h_0 \leq y_1 - y_0.$$

Proof. Suppose that x_n, $n \in Z_{M_g+1}$, is the sequence of points generated by the nondecreasing function g and set $h_n := x_{n+1} - x_n$. Hence, $h_n = g(x_n) = g(x_{n-1} + h_{n-1})$. By the definition of g, we have that

$$h_n = kx_{n-1} + kh_{n-1} + b = g(x_{n-1}) + kh_{n-1} = h_{n-1} + kh_{n-1} = (1+k)h_{n-1}.$$

Repeatedly using this equation, we obtain the formula that $h_n = (1+k)^n h_0$. Consequently, we conclude that

$$x_{M_g} - y_0 = \sum_{j=0}^{M_g-1} h_j = h_0 \sum_{j=0}^{M_g-1} (1+k)^j = \left(\frac{(1+k)^{M_g} - 1}{k}\right) h_0.$$

Noting that $x_{M_g} \leq y_1$, this completes the proof. ◻

To estimate the complexity of the Hermite spline interpolation based on the grid points constructed at the beginning of this section, we need to estimate the integer M appearing in the mesh. To this end, we consider a special admissible triple (x^*, x', h^0), where h^0 is defined by (3.2) and

$$x^* := \frac{\varepsilon}{\beta} \log\left(\frac{N}{\beta}\right) \quad \text{and} \quad x' := \frac{\varepsilon}{\beta} \log\left(\frac{1}{\varepsilon}\right) \quad \text{with} \quad \beta := \frac{\alpha}{2r}.$$

LEMMA 4.4. *Suppose that ε and N satisfy the condition*

$$\frac{e}{e^{2^{\gamma(N-1)}}} < \varepsilon < \frac{\beta}{N} \tag{4.5}$$

for some positive constant γ independent of ε and N. Then $\mathcal{M}(x^*, x', h^0) < \gamma(N-1)$.

Proof. The hypothesis $\varepsilon < \beta/N$ ensures that $x^* < x'$ and thus (x^*, x', h^0) is an admissible triple. We will use Lemmas 4.2 and 4.3 to estimate the index M_{h^0} generated by h^0.

Let $y_0 := x^*$, $y_1 := x'$. It can be verified that

$$h^0(y_0) = \frac{\varepsilon}{\beta} \quad \text{and} \quad \frac{d}{dx} h^0(y_0) = 1.$$

Choose $k = 1$, $h_0 := \frac{\varepsilon}{\beta}$, and $b := h_0 - k y_0$ and define $g(x) := kx + b$. Thus (y_0, y_1, g) is an admissible triple. Note that the straight line $y = g(x)$ passes through the point $(y_0, h^0(y_0))$ with slope $k = 1$. It is easily confirmed that $g(x) < h^0(x)$ for all $x > y_0$. By Lemma 4.2, we have that

$$\mathcal{M}(y_0, y_1, h^0) \le \mathcal{M}(y_0, y_1, g).$$

Let $M_g := \mathcal{M}(y_0, y_1, g)$. It follows from Lemma 4.3 that $h_0\left(2^{M_g} - 1\right) \le y_1 - y_0 \le y_1$. This leads to the estimate

$$M_g \le \log_2\left(\ln\left(\frac{1}{\varepsilon}\right) + 1\right). \tag{4.6}$$

By hypothesis that $\frac{e}{e^{2^{\gamma(N-1)}}} < \varepsilon$, we conclude that $\log_2(\ln(1/\varepsilon) + 1) < \gamma(N-1)$. This proves the result of this lemma. □

We are now ready to prove the main result of this section.

THEOREM 4.5. *Let u be the solution of problem* (2.1)*, let Δ be the mesh constructed as above, and let Π be the piecewise polynomial interpolation operator associated with the mesh Δ. Then Πu satisfies the error bound* (3.11)*. If, in addition, the condition* (4.5) *is satisfied, then the number of grid points is bounded by cN.*

Proof. By the construction of the mesh Δ, it is ready to verify that the mesh satisfies condition (3.3) and thus, by Proposition 3.2, Πu satisfies the error bound (3.11).

It remains to prove the bound on the number of grid points in the mesh Δ. Recalling that the number of the grid points in the mesh is $\tilde{N} = 2M + N_1$ for case one and $\tilde{N} = 2M'$ for case two, it suffices to prove that N_1, M, and M' are bonded by a constant multiple of N. Because

$$N_1 = \lfloor (1 - 2x_M)N \rfloor + 1 \le N + 1 \le 2N \quad \text{and} \quad M' < M$$

we need only to show that $M \le cN$.

We consider one case in the construction of the mesh. Recalling the definition of x^*, we divide the interval $[0, x_M]$ into two intervals $I_1 := [0, x^*]$ and $I_2 := [x^*, x_M]$ and estimate the number of grid points in these intervals separately. Let M_1 and M_2 denote the number of grid points in intervals I_1 and I_2, respectively. Then $x_{M_1} \leq x^* \leq x_{M_1+1}$ and $M = M_1 + M_2$.

We now estimate the value of M_1 with the help of the integral of function $\phi(x) := 1/h^0(x)$, $x \in I$. To this end, we introduce a piecewise linear function

$$\underline{\phi}(x) := \phi(x_{i-1}) \frac{x_i - x}{x_i - x_{i-1}}, \quad x \in [x_{i-1}, x_i), \quad i \in \mathbb{N}_{M_1}.$$

Using the definition of ϕ and noticing that ϕ is decreasing on $[0, x_{M_1}]$, we find for $x \in (x_{i-1}, x_i)$ that

$$\phi'(x) := -\frac{\beta}{\varepsilon h^0(x)} = -\frac{\beta}{\varepsilon}\phi(x) \geq -\frac{\beta}{\varepsilon}\phi(x_{i-1}).$$

Also, for $x \in (x_{i-1}, x_i)$, we have that $\underline{\phi}'(x) = -\phi^2(x_{i-1})$. Condition $x_i < x^*$ implies that $\beta/\varepsilon \leq \phi(x_{i-1})$. It follows from this inequality and the fact that ϕ is decreasing on $[x_{i-1}, x_i)$ that $\phi'(x) \geq \underline{\phi}'(x)$ for $x \in (x_{i-1}, x_i)$. This estimate with the fact that $\phi(x_{i-1}) = \underline{\phi}(x_{i-1})$ implies that

$$\underline{\phi}(x) \leq \phi(x), \qquad x \in [0, x_{M_1}). \tag{4.7}$$

By the construction of the grid points x_i and the definition of function ϕ, we obtain that

$$\int_0^{x_{M_1}} \underline{\phi}(x)\, dx = \frac{1}{2}\sum_{i=1}^{M_1} \phi(x_{i-1})(x_i - x_{i-1}) = \frac{M_1}{2}. \tag{4.8}$$

On the other hand, we have the estimate

$$\int_0^{x_{M_1}} \phi(x)\, dx \leq \int_0^1 \frac{dx}{h^0(x)} \leq \frac{N}{\beta}. \tag{4.9}$$

Combining (4.7), (4.8), and (4.9), we conclude that $M_1 \leq \frac{2N}{\beta}$.

We next prove that $M_2 < \gamma N$. Recalling the mesh generating procedure, we have that $h^0(x) \leq 1/N$, $x \in [0, x_{M-1}]$. The definition of x' implies that $h^0(x') = 1/N$. Noting that h^0 is a strictly monotone increasing function, from inequalities

$$h^0(x_{M-1}) < h^0(x') = \frac{1}{N} \leq h^0(x_M)$$

we immediately obtain that $x_{M-1} < x' \leq x_M$. By Lemma 4.4, we conclude that $M_2 < \gamma N$. Therefore, in this case we have that $M \leq (\frac{2}{\beta} + \gamma)N$, completing the proof of this theorem. □

We remark that the mesh generating function can be modified to be

$$h^0(x) = S\frac{\varepsilon}{N} \exp\left(\frac{\alpha x}{2r\varepsilon}\right),$$

where S is a constant. The purpose of introducing the constant S is to adjust the number of grid points to be distributed in the boundary layers at two ends. For example, if S is larger, then by our construction, fewer grid points will be generated for the boundary layer. Its effect is illustrated in the numerical examples given in the next section. The constant S does not affect the convergence order, nor the order of the number of grid points.

5. Numerical examples. In this section, we present numerical examples to demonstrate the method and to verify the theoretical results proved in this paper. We begin with an introduction of the space of the cubic Hermite splines which correspond to $r = 2$. For $x \in [-1, 1]$ we assume

$$\begin{aligned}&\psi_{10}(x) := (x+1)^2(1-2x), \quad \psi_{11}(x) := (1-x)^2(2x+1),\\ &\psi_{20}(x) := x(x+1)^2, \quad \psi_{21}(x) := x(x-1)^2\end{aligned}$$

and define two cubic splines ϕ_1 and ϕ_2 on $\mathbb{R}$ by

$$\phi_1(x) := \begin{cases}\psi_{10}(x), & x \in [-1,0],\\ \psi_{11}(x), & x \in (0,1],\\ 0, & x \in \mathbb{R}\setminus[-1,1],\end{cases} \quad \text{and} \quad \phi_2(x) := \begin{cases}\psi_{20}(x), & x \in [-1,0],\\ \psi_{21}(x), & x \in (0,1],\\ 0, & x \in \mathbb{R}\setminus[-1,1].\end{cases}$$

It can be verified that they satisfy the Hermite interpolation conditions

$$\phi_1(j) = \delta(j), \quad \phi_1'(j) = 0, \quad \phi_2(j) = 0, \quad \phi_1'(j) = \delta(j), \quad j \in \mathbb{Z},$$

where for $j \in \mathbb{Z}$, $\delta(j) = 0$ if $j \neq 0$ and 1 if $j = 0$. Applications of the Hermite cubic splines in the numerical solution of boundary value problems of ordinary differential equations may be found in [JL]. On a mesh $0 < x_0 < x_1 < \cdots < x_{\tilde{N}} = 1$, we may scale and shift ϕ_1, ϕ_2 to construct the basis functions for the space $V_{\tilde{N}}$ by

$$\phi_{1,i}(x) := \begin{cases}\psi_{10}(\frac{x-x_i}{x_i-x_{i-1}}), & x \in [x_{i-1}, x_i],\\ \psi_{11}(\frac{x-x_i}{x_{i+1}-x_i}), & x \in (x_i, x_{i+1}],\\ 0, & x \in I\setminus[x_{i-1}, x_{i+1}],\end{cases}$$

and

$$\phi_{2,i}(x) := \begin{cases}\psi_{20}(\frac{x-x_i}{x_i-x_{i-1}}), & x \in [x_{i-1}, x_i],\\ \frac{h_{i+1}}{h_i}\psi_{21}(\frac{x-x_i}{x_{i+1}-x_i}), & x \in (x_i, x_{i+1}],\\ 0, & x \in I\setminus[x_{i-1}, x_{i+1}],\end{cases}$$

for $i \in \mathbb{N}_{\tilde{N}-1}$. According to the interpolation properties of ϕ_1 and ϕ_2, we may verify for $i \in \mathbb{N}_{\tilde{N}-1}$ and $j \in Z_{\tilde{N}+1}$ that

$$\phi_{1,i}(x_j) = \delta(i-j), \quad \phi_{1,i}'(x_j) = 0, \quad \phi_{2,i}(x_j) = 0, \quad \phi_{2,i}'(x_j) = \frac{1}{h_i}\delta(i-j).$$

The space $V_{\tilde{N}}$ of the cubic (corresponding to $m = 2$) Hermite splines is defined by $V_{\tilde{N}} := \operatorname{span}\{\phi_{i,j} : i = 1, 2,\ j \in \mathbb{N}_{\tilde{N}-1}\}$. It is easily seen that $V_{\tilde{N}}$ is a dense subspace of $H_0^2(I)$.

Example 1. In this example, we consider the reaction-diffusion problem given by

$$\begin{aligned}\varepsilon^2 u^{(4)}(x) + ((1 + x(1-x))u')' &= f(x), \qquad x \in (0,1),\\ u(0) = u'(0) = u(1) = u'(1) &= 0,\end{aligned}$$

where f is chosen so that this problem has the exact solution

$$u = \varepsilon\left(\frac{\exp(-x/\varepsilon) + \exp(-(1-x)/\varepsilon)}{1 + \exp(-1/\varepsilon)} - 1\right) + \frac{1 - \exp(-1/\varepsilon)}{1 + \exp(-1/\varepsilon)}x(1-x) + x^2(1-x)^2.$$

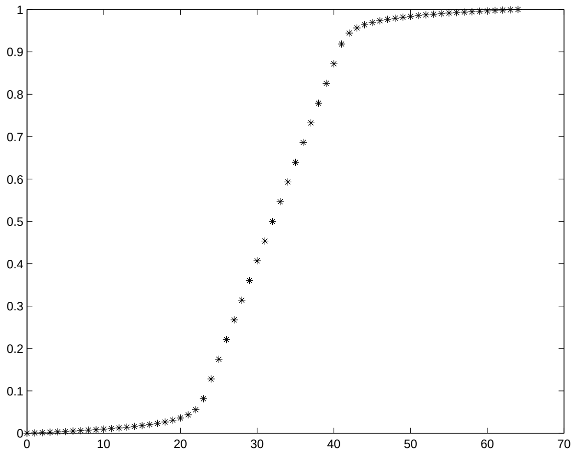


FIG. 1.

The solution exhibits two boundary layers near $x = 0$ and $x = 1$. The weak form of this problem corresponds to $m = 2$.

The generating function for this example is chosen as

$$h^0(x) = \frac{4\varepsilon}{N} \exp\left(\frac{x}{4\varepsilon}\right),$$

that is, we choose $S := \frac{2m}{\alpha} = 4$. We first present a figure that shows the distribution of the grid points for this problem. In this figure we choose $\alpha = 1$ and $\varepsilon = 3.905 \times 10^{-3}$ and $N = 21$. Correspondingly, the total number of grid points is 64. According to our method, we have 21 grid points distributed in each of the two boundary layers near $x = 0$ and $x = 1$. The graph in Figure 1 shows the distribution of 64 grid points.

To illustrate the optimal order of the uniform convergence of the proposed method, we set $e_{\tilde{N}} := u - u_{\tilde{N}}^{\tilde{N}}$. In Table 1, we compare the values of $\|e_{\tilde{N}}\|_\varepsilon$ and the convergence order O of the proposed method based on the optimal mesh with those results based on the Shishkin mesh (see p. 126 of [SS1]). The convergence order O is computed numerically by the formula

$$O = \log\left(\frac{e_{\tilde{N}_1}}{e_{\tilde{N}_2}}\right) / \log\left(\frac{\tilde{N}_2}{\tilde{N}_1}\right),$$

where $\tilde{N}_1$, $\tilde{N}_2$ are two neighboring numbers of grid points in the table. The numerical results show evidently that the proposed optimal mesh provides the optimal order $O \approx m = r = 2$ of convergence for $\|e_{\tilde{N}}\|_\varepsilon$, and it performs better than the Shishkin mesh.

To verify the estimate in norm $\|\cdot\|_{m-1}$, a result in Theorem 3.4, we list in Table 2 the values of $\|e_{\tilde{N}}\|_1$ and convergence order O. We observe that the computed convergence order is $O = 3$.

Example 2. We consider the second order reaction-diffusion problems

$$\begin{aligned} -\varepsilon^2 u''(x) + (2 + \sin(x))u(x) &= f(x), \qquad x \in (0,1), \\ u(0) = u(1) &= 0, \end{aligned}$$

Table 1

Comparison of convergence orders for numerical methods using the proposed mesh and the Shishkin mesh.

	Proposed mesh			Shishkin mesh		
ε	$\tilde{N}$	$\|e_{\tilde{N}}\|_\varepsilon$	O	$\tilde{N}$	$\|e_{\tilde{N}}\|_\varepsilon$	O
3.905-3	63	1.13-4		64	1.80-3	
	127	2.72-5	2.03	128	6.78-4	1.40
	255	6.78-6	1.99	256	2.41-4	1.49
	511	1.68-6	2.01	512	8.09-5	1.57
	1026	4.17-7	2.00	1024	2.59-5	1.64
6.104-5	63	1.79-5		64	2.24-4	
	127	3.97-6	2.14	128	8.47-5	1.40
	255	9.58-7	2.04	256	3.01-5	1.49
	513	2.34-7	2.02	512	1.01-5	1.58
	1023	5.86-8	2.01	1024	3.24-6	1.64
3.816-6	64	8.65-6		64	5.62-5	
	126	1.33-6	2.76	128	2.12-5	1.40
	254	2.61-7	2.32	256	7.53-6	1.49
	510	6.07-8	2.09	512	2.53-6	1.57
	1023	1.48-8	2.02	1024	8.11-7	1.64

Table 2

Convergence order for $\|e_{\tilde{N}}\|_1$ *for Example* 1.

ε	$\tilde{N}$	$\|e_{\tilde{N}}\|_1$	O
3.905-3	64	1.32-5	
	127	1.58-6	3.10
	254	1.99-7	2.99
	510	2.40-8	3.03
	1027	2.95-9	2.99
6.104-5	66	1.64-5	
	129	1.95-6	3.18
	256	2.35-7	3.09
	511	2.88-8	3.04
	1024	3.51-9	3.03
3.816-6	63	1.96-5	
	128	2.06-6	3.18
	257	2.39-7	3.09
	512	2.90-8	3.06
	1029	3.48-9	3.04

where f is chosen so that

$$u(x) = \exp(-x/\varepsilon) + \exp(-(1-x)/\varepsilon) + x(1-x) - (1 + \exp(-1/\varepsilon))$$

is the exact solution. In this case, $m = 1$. Since $r = 2$, according to the theoretical estimate the optimal convergence order is $\|u - u_{\tilde{N}}^{\tilde{N}}\|_\varepsilon \leq cN^{-3}$. The numerical results shown in Table 3 confirms this estimate.

TABLE 3
Convergence order for $\|e_{\tilde{N}}\|_\varepsilon$ *for Example* 2.

ε	$\tilde{N}$	$\|\|e_{\tilde{N}}\|\|_\varepsilon$	O
3.905-3	65	3.35-6	
	128	4.59-7	2.93
	255	6.27-8	2.89
	513	8.12-9	2.92
	1025	1.07-9	2.93
6.104-5	63	5.51-7	
	126	7.05-8	2.97
	256	9.01-9	2.90
	512	1.21-9	2.90
	1024	1.60-10	2.92
3.816-6	63	1.38-7	
	125	1.89-8	2.90
	257	2.25-9	2.95
	512	3.01-10	2.92
	1026	4.03-11	2.89

Acknowledgments. The authors are grateful to Professor M. Stynes for bringing references [L, RL] to their attention after the first draft of this paper was written. A special thank is given to a referee whose remark leads to a relaxation of the hypothesis on the coercivity of the bilinear form.

REFERENCES

[B] A. S. BAKHVALOV, *K optimizacii methdov resenia kraevyh zadac prinalicii pogranicnogo sloja.*, Zh. Vychisl. Mat. Mat. Fiz., 9 (1969), pp. 841–859.

[C] P. G. CIARLET, *The Finite Element Methods for Elliptic Problems*, North-Holland Publishing Company, North-Holland, Amsterdam, 1978.

[CHX] Y. Z. CAO, T. HERDMAN, AND Y. XU, *A hybrid collocation method for Volterra integral equations with weakly singular kernels*, SIAM J. Numer. Anal., 41 (2003), pp. 364–381.

[G] E. C. GARTLAND, *Graded-mesh difference schemes for singularly perturbed to-point boundary value problems*, Math. Comp., 51 (1988), pp. 631–657.

[Gr] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains*, Pitman Advanced Publishing, Boston, 1985.

[GT] D. GILBARG AND N. S. TRUDINGER, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer-Verlag, Berlin, 1983.

[JL] R. Q. JIA AND S. T. LIU, *Wavelet bases of Hermite cubic splines on the interval*, Adv. Comput. Math., to appear.

[L] T. LINSS, *Analysis of a Galerkin finite element method on a Bakhvalov-Shishkin mesh for a linear convection-diffusion problem*, IMA J. Numer. Anal., 20 (2000), pp. 621–632

[LT] W. LIU AND T. TANG, *Error analysis for a Galerkin-spectral method with coordinate transformation for solving singularly perturbed problems*, Appl. Numer. Math., 38 (2001), pp. 315–345.

[MOS] J. J. H. MILLER, E. O'RIORDAN, AND G. I. SHISHKIN, *Fitted Numerical Methods for Singularly Perturbation Problems*, World Scientific, Singapore, 1996.

[O] R. E. O'MALLEY, *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[QS] Y. QIU AND D. M. SLOAN, *Analysis of difference approximations to a singular perturbated two-point boundary value problem on an adaptively generated grid*, J. Comput. Appl. Math., 101 (1999), pp. 1–25.

[QST] Y. QIU, D. M. SLOAN, AND T. TANG, *Numerical solution of a singularly perturbated two-point boundary value problem using equidistribution: Analysis of convergence*, J. Comput. Appl. Math., 116 (2000), pp. 121–143.

[RL] H. G. ROOS AND T. LINSS, *Sufficient conditions for uniform convergence on layer-adapted* *grids*, Computing, 63 (1999), pp. 27–45.

[RST] H. G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.

[Sc] L. L. Schumaker, *Spline Functions: Basic Theory*, John Wiley & Sons, New York, 1981.

[S] G. I. Shishkin, *Grid approximation of singularly perturbed parabolic equations with internal layers*, Sov. J. Numer. Anal. Math. Model., 3 (1988), pp. 393–407.

[SO1] M. Stynes and E. O'Riordan, *A finite element method for a singularly perturbed boundary value problem*, Numer. Math., 50 (1986), pp. 1–15.

[SO2] M. Stynes and E. O'Riordan, *An analysis of a singularly perturbed two-point boundary value problem using only finite element techniques*, Math. Comp., 56 (1991), pp. 663–675.

[SS1] G. Sun and M. Stynes, *Finite element methods for singularly perturbed high-order elliptic two-point boundary problems,* I: *Reaction-diffusion-type problems*, IMA J. Numer. Anal., 15 (1995), pp. 117–139.

[SS2] G. Sun and M. Stynes, *Finite element methods for singularly perturbed high-order elliptic two-point boundary problems,* II: *Convection-diffusion-type problems*, IMA J. Numer. Anal., 15 (1995), pp. 197–219.

[TT] T. Tang and M. R. Trummer, *Boundary layer resolving pseudospectral methods for singular perturbation problems*, SIAM J. Sci. Comput., 17 (1996), pp. 430–438.

[V] R. Vulanovic, *Non-equidistant generalizations of the Gushchin-Shchennikov scheme*, Z. Angew. Math. Mec., 67 (1987), pp. 625–632.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2624–2646

MULTIADAPTIVE GALERKIN METHODS FOR ODES III: A PRIORI ERROR ESTIMATES*

ANDERS LOGG†

Abstract. The multiadaptive continuous/discontinuous Galerkin methods mcG(q) and mdG(q) for the numerical solution of initial value problems for ordinary differential equations are based on piecewise polynomial approximation of degree q on partitions in time with time steps which may vary for different components of the computed solution. In this paper, we prove general order a priori error estimates for the mcG(q) and mdG(q) methods. To prove the error estimates, we represent the error in terms of a discrete dual solution and the residual of an interpolant of the exact solution. The estimates then follow from interpolation estimates, together with stability estimates for the discrete dual solution.

Key words. multiadaptivity, individual time steps, local time steps, ODE, continuous Galerkin, discontinuous Galerkin, mcG(q), mdG(q), a priori error estimates, existence, stability, Peano kernel theorem, interpolation estimates, piecewise smooth

AMS subject classifications. 65L05, 65L07, 65L20, 65L50, 65L60, 65L70

DOI. 10.1137/040604133

1. Introduction.

This is part 3 in a sequence of papers [32, 33] on multiadaptive Galerkin methods, mcG(q) and mdG(q), for approximate (numerical) solution of ODEs of the form

$$\begin{aligned} \dot{u}(t) &= f(u(t), t), \quad t \in (0, T], \\ u(0) &= u_0, \end{aligned} \tag{1.1}$$

where $u : [0, T] \to \mathbb{R}^N$ is the solution to be computed, $u_0 \in \mathbb{R}^N$ a given initial condition, $T > 0$ a given final time, and $f : \mathbb{R}^N \times (0, T] \to \mathbb{R}^N$ a given function that is Lipschitz-continuous in u and bounded.

In the previous two parts of our series on multiadaptive Galerkin methods, we proved a posteriori error estimates, through which the time steps are adaptively determined from residual feedback and stability information, obtained by solving a dual linearized problem. In this paper, we prove a priori error estimates for mcG(q) and mdG(q). We also prove the stability estimates and interpolation estimates which are essential to the a priori error analysis.

Standard methods for the time-discretization of (1.1) require that the resolution is equal for all components $U_i(t)$ of the computed approximate solution $U(t)$ of (1.1). This includes all standard Galerkin or Runge–Kutta methods; see [9, 4, 23, 24, 41, 2]. Using the same time step sequence $k = k(t)$ for all components could become very costly if the different components of the solution exhibit multiple time scales of different magnitudes. We therefore propose a new representation of the solution in which the difference in time scales is reflected in the *componentwise* time-discretization of (1.1), that is, each component $U_i(t)$ is computed using an individual time step sequence $k_i = k_i(t)$.

*Received by the editors February 12, 2004; accepted for publication (in revised form) May 4, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/sinum/43-6/60413.html

†Toyota Technological Institute at Chicago, 1427 East 60th Street, Chicago, IL 60637 (logg@tti-c.org).

The multiadaptive Galerkin methods mcG(q) and mdG(q) first presented in [32] are formulated as extensions of the standard continuous and discontinuous Galerkin methods cG(q) and dG(q), studied earlier in detail by Hulme [28, 27], Jamet [29], Delfour, Hager, and Trochu [7], Eriksson, Johnson, and Thomée [16, 30, 11, 12, 10, 13, 14, 15, 8], and Estep et al. [17, 18, 19, 21, 20]. As such, the analysis of the mcG(q) and mdG(q) methods can be carried out within the existing framework, but the extension to multiadaptive time-stepping leads to some technical challenges, in particular, proving the appropriate interpolation estimates.

Local (multiadaptive) time-stepping has been explored before to some extent for specific applications, including specialized integrators for the n-body problem [37, 5, 1] and low-order methods for conservation laws [39, 22, 6]. Early attempts at local time-stepping include [25, 26]. Recently, a new class of related methods, known as asynchronous variational integrators (AVI) with local time steps, has been proposed [31].

1.1. Main results. The main results of this paper are a priori error estimates for the mcG(q) and mdG(q) methods, respectively, of the form

$$\|e(T)\|_{l_p} \leq CS(T) \big\| k^{2q} u^{(2q)} \big\|_{L_\infty([0,T],l_1)} \tag{1.2}$$

and

$$\|e(T)\|_{l_p} \leq CS(T) \big\| k^{2q+1} u^{(2q+1)} \big\|_{L_\infty([0,T],l_1)} \tag{1.3}$$

for $p = 2$ or $p = \infty$, where C is an interpolation constant, $S(T)$ is a (computable) stability factor, and $k^{2q}u^{(2q)}$ (or $k^{2q+1}u^{(2q+1)}$) combines local time steps $k_i = k_i(t)$ with derivatives of the exact solution u. The norm $L_\infty(I, \|\cdot\|)$ is defined by $\|v\|_{L_\infty(I,\|\cdot\|)} = \sup_{t\in I} \|v(t)\|$. These estimates state that the mcG(q) method is of order $2q$ and that the mdG(q) method is of order $2q + 1$ in the local time step. We refer to section 6.2 for the exact results. It should be noted that superconvergence is obtained only at synchronized time levels, such as the end-point $t = T$. For the general nonlinear problem, we obtain exponential estimates for the stability factor $S(T)$. In [34], we prove that for a parabolic model problem, the stability factor remains bounded and of unit size, independent of T (up to a logarithmic factor).

1.2. Notation. The following notation is used throughout this paper. Each component $U_i(t)$, $i = 1, \ldots, N$, of the approximate m(c/d)G(q) solution $U(t)$ of (1.1) is a piecewise polynomial on a partition of $(0, T]$ into M_i subintervals. Subinterval j for component i is denoted by $I_{ij} = (t_{i,j-1}, t_{ij}]$, and the length of the subinterval is given by the *local time step* $k_{ij} = t_{ij} - t_{i,j-1}$. This is illustrated in Figure 1. On each subinterval I_{ij}, $U_i|_{I_{ij}}$ is a polynomial of degree q_{ij} and we refer to $(I_{ij}, U_i|_{I_{ij}})$ as an *element*.

Furthermore, we shall assume that the interval $(0, T]$ is partitioned into blocks between certain synchronized time levels $0 = T_0 < T_1 < \cdots < T_M = T$. We refer to the set of intervals $\mathcal{T}_n$ between two synchronized time levels T_{n-1} and T_n as a *time slab*:

$$\mathcal{T}_n = \{I_{ij} : T_{n-1} \leq t_{i,j-1} < t_{ij} \leq T_n\}.$$

We denote the length of a time slab by $K_n = T_n - T_{n-1}$. We also refer to the entire collection of intervals I_{ij} as the partition $\mathcal{T}$.

Since different components use different time steps, a local interval I_{ij} may contain nodal points for other components, that is, some $t_{i'j'} \in (t_{i,j-1}, t_{ij})$. We denote the set of such internal nodes on a local interval I_{ij} by $\mathcal{N}_{ij}$.

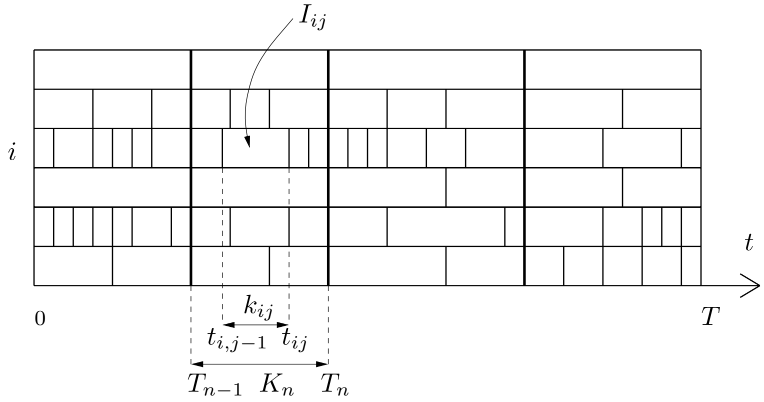


FIG. 1. *Individual partitions of the interval* $(0,T]$ *for different components. Elements between common synchronized time levels are organized in time slabs. In this example, we have* $N = 6$ *and* $M = 4$.

1.3. Outline of the paper. The outline of this paper is as follows. In section 2, we give the full definition of the multiadaptive Galerkin methods mcG(q) and mdG(q). We also introduce the dual methods mcG(q)* and mdG(q)*, which are of importance to the a priori error analysis. In sections 3 and 4, respectively, we then prove existence and stability of the discrete solutions as defined in section 2.

In section 5, we prove the interpolation estimates that we later use to prove the a priori error estimates in section 6. Proving the interpolation estimates is technically challenging, since the function to be interpolated may be discontinuous within the interval of interpolation. To measure the regularity of the interpolated function, it is then necessary to take into consideration the size of the jump in function value and derivatives at each point of discontinuity.

Finally, in section 7, we present some numerical evidence for the a priori error estimates by solving a simple model problem and showing that we obtain the predicted convergence rates, k^{2q} and k^{2q+1}, respectively, for the mcG(q) and mdG(q) methods.

2. Definition of methods. In this section, we give the definitions of the multiadaptive Galerkin methods mcG(q) and mdG(q). The multiadaptive methods are obtained as extensions of the standard (monoadaptive) Galerkin methods cG(q) and dG(q) by extending the trial and test spaces to allow individual time step sequences for different components.

As an important tool for the a priori error analysis in section 6, we also introduce the discrete dual problem and the discrete dual methods mcG(q)* and mdG(q)*.

2.1. Multiadaptive continuous Galerkin, mcG(q). To formulate the mcG(q) method, we define the *trial space* V and the *test space* $\hat{V}$ as

$$\begin{aligned} V &= \left\{ v \in [\mathcal{C}([0,T])]^N : v_i|_{I_{ij}} \in \mathcal{P}^{q_{ij}}(I_{ij}),\ j = 1,\ldots,M_i,\ i = 1,\ldots,N \right\}, \\ \hat{V} &= \left\{ v : v_i|_{I_{ij}} \in \mathcal{P}^{q_{ij}-1}(I_{ij}),\ j = 1,\ldots,M_i,\ i = 1,\ldots,N \right\}, \end{aligned} \tag{2.1}$$

where $\mathcal{P}^q(I)$ denotes the linear space of polynomials of degree q on an interval $I \subset \mathbb{R}$. In other words, V is the space of vector-valued continuous piecewise polynomials of degree $q = (q_i(t))$ with $q_i(t) \geq 1$ on the partition $\mathcal{T}$, and $\hat{V}$ is the space of vector-valued (possibly discontinuous) piecewise polynomials of degree $q - 1 = (q_i(t) - 1)$ on the same partition.

We now define the mcG(q) method for (1.1) as follows: Find $U \in V$ with $U(0) = u_0$ such that

$$\int_0^T (\dot{U}, v)\,dt = \int_0^T (f(U,\cdot), v)\,dt \quad \forall v \in \hat{V}, \tag{2.2}$$

where $(\cdot,\cdot)$ denotes the $\mathbb{R}^N$ inner product. With a suitable choice of test function v, it follows that the global problem (2.2) can be restated as a sequence of successive local problems for each component: For $i = 1, \ldots, N$, $j = 1, \ldots, M_i$, find $U_i|_{I_{ij}} \in \mathcal{P}^{q_{ij}}(I_{ij})$ with $U_i(t_{i,j-1})$ given such that

$$\int_{I_{ij}} \dot{U}_i v\,dt = \int_{I_{ij}} f_i(U,\cdot) v\,dt \quad \forall v \in \mathcal{P}^{q_{ij}-1}(I_{ij}), \tag{2.3}$$

where the initial condition is specified for $i = 1, \ldots, N$ by $U_i(0) = u_i(0)$.

We define the *residual* R of the approximate solution U by $R_i(U,t) = \dot{U}_i(t) - f_i(U(t),t)$. In terms of the residual, we can rewrite (2.3) in the form

$$\int_{I_{ij}} R_i(U,\cdot) v\,dt = 0 \quad \forall v \in \mathcal{P}^{q_{ij}-1}(I_{ij}), \quad j = 1, \ldots, M_i, \quad i = 1, \ldots, N, \tag{2.4}$$

that is, the residual is orthogonal to the test space on each local interval. We refer to (2.4) as the *Galerkin orthogonality* of the mcG(q) method.

2.2. Multiadaptive discontinuous Galerkin, mdG(q). For mdG(q), we define the trial and test spaces by

$$V = \hat{V} = \{v : v_i|_{I_{ij}} \in \mathcal{P}^{q_{ij}}(I_{ij}),\ j = 1, \ldots, M_i,\ i = 1, \ldots, N\}, \tag{2.5}$$

that is, both trial and test functions are vector-valued (possibly discontinuous) piecewise polynomials of degree $q = (q_i(t))$ with $q_i(t) \geq 0$ on the partition $\mathcal{T}$. By definition, the mdG(q) solution $U \in V$ is left-continuous.

We now define the mdG(q) method for (1.1) as follows: Find $U \in V$ with $U(0^-) = u_0$ such that

$$\sum_{i=1}^{N} \sum_{j=1}^{M_i} \left[[U_i]_{i,j-1} v_i\big(t_{i,j-1}^+\big) + \int_{I_{ij}} \dot{U}_i v_i\,dt \right] = \int_0^T (f(U,\cdot), v)\,dt \quad \forall v \in \hat{V}, \tag{2.6}$$

where $[U_i]_{i,j-1} = U_i(t_{i,j-1}^+) - U_i(t_{i,j-1}^-)$ denotes the jump in $U_i(t)$ across the node $t = t_{i,j-1}$, and where $v(t^+) = \lim_{s \to t^+} v(s)$.

The mdG(q) method in local form, corresponding to (2.3), reads as follows: For $i = 1, \ldots, N$, $j = 1, \ldots, M_i$, find $U_i|_{I_{ij}} \in \mathcal{P}^{q_{ij}}(I_{ij})$ such that

$$[U_i]_{i,j-1} v(t_{i,j-1}) + \int_{I_{ij}} \dot{U}_i v\,dt = \int_{I_{ij}} f_i(U,\cdot) v\,dt \quad \forall v \in \mathcal{P}^{q_{ij}}(I_{ij}), \tag{2.7}$$

where the initial condition is specified for $i = 1, \ldots, N$ by $U_i(0^-) = u_i(0)$.

The residual R is defined on the inner of each local interval I_{ij} by $R_i(U,t) = \dot{U}_i(t) - f_i(U(t),t)$. In terms of the residual, (2.7) can be restated in the form

$$[U_i]_{i,j-1} v\big(t_{i,j-1}^+\big) + \int_{I_{ij}} R_i(U,\cdot) v\,dt = 0 \quad \forall v \in \mathcal{P}^{q_{ij}}(I_{ij}) \tag{2.8}$$

for $j = 1, \ldots, M_i$, $i = 1, \ldots, N$. We refer to (2.8) as the Galerkin orthogonality of the mdG(q) method.

2.3. The dual problem. The dual problem is the standard tool for error analysis, a priori or a posteriori, of Galerkin finite element methods for the numerical solution of differential equations; see [8, 3]. For the a posteriori error analysis of the multiadaptive Galerkin methods mcG(q) and mdG(q) in [32], we formulate a continuous dual problem. For the a priori error analysis of this paper, we formulate instead a discrete dual problem. The discrete dual problem was first introduced for the family of discontinuous Galerkin methods dG(q) in [16]. As we shall see, the discrete dual problem can be expressed as a Galerkin method for a continuous problem.

The discrete dual solution $\Phi : [0,T] \to \mathbb{R}^N$ is a Galerkin approximation of the exact solution $\phi : [0,T] \to \mathbb{R}^N$ of the continuous dual problem

$$\begin{aligned} -\dot{\phi}(t) &= J^\top(\pi u, U, t)\phi(t) + g(t), \quad t \in [0,T), \\ \phi(T) &= \psi, \end{aligned} \tag{2.9}$$

where πu is an interpolant or a projection of the exact solution u of (1.1), $g : [0,T] \to \mathbb{R}^N$ is a given function, $\psi \in \mathbb{R}^N$ is a given initial condition, and

$$J^\top(\pi u, U, t) = \left(\int_0^1 \frac{\partial f}{\partial u}(s\pi u(t) + (1-s)U(t), t)\, ds \right)^\top, \tag{2.10}$$

that is, an appropriate mean value of the transpose of the Jacobian of the right-hand side $f(\cdot, t)$ evaluated at $\pi u(t)$ and $U(t)$. Note that by the chain rule, we have

$$J(\pi u, U, \cdot)(U - \pi u) = f(U, \cdot) - f(\pi u, \cdot). \tag{2.11}$$

The data (ψ, g) of the dual problem allow us to obtain error estimates for different functionals $L_{\psi,g}$ of the error $e = U - u$.

We define below two new Galerkin methods for the dual problem (2.9): the dual methods mcG$(q)^*$ and mdG$(q)^*$. We will later use the mcG$(q)^*$ method to express the error of the mcG(q) solution of (1.1) in terms of the mcG$(q)^*$ solution of (2.9). Similarly, we will express the error of the mdG(q) solution of (1.1) in terms of the mdG$(q)^*$ solution of (2.9).

2.4. Multiadaptive dual continuous Galerkin, mcG$(q)^*$. In the formulation of the dual method of mcG(q), we interchange the trial and test spaces of mcG(q). With the same definitions of V and $\hat{V}$ as in (2.1), we thus define the mcG$(q)^*$ method for (2.9) as follows: Find $\Phi \in \hat{V}$ with $\Phi(T^+) = \psi$ such that

$$\int_0^T (\dot{v}, \Phi)\, dt = \int_0^T (J(\pi u, U, \cdot)v, \Phi) + L_{\psi,g}(v) \tag{2.12}$$

for all $v \in V$ with $v(0) = 0$, where

$$L_{\psi,g}(v) \equiv (v(T), \psi) + \int_0^T (v, g)\, dt. \tag{2.13}$$

Notice the extra condition that the test functions should vanish at $t = 0$, which is introduced to make the dimension of the test space equal to the dimension of the trial space. Integrating by parts, (2.12) can alternatively be expressed in the form

$$\sum_{i=1}^N \sum_{j=1}^{M_i} \left[-[\Phi_i]_{ij} v_i(t_{ij}) - \int_{I_{ij}} \dot{\Phi}_i v_i\, dt \right] = \int_0^T (J^\top(\pi u, U, \cdot)\Phi + g, v)\, dt. \tag{2.14}$$

2.5. Multiadaptive dual discontinuous Galerkin, mdG$(q)^*$. With the same definitions of V and $\hat{V}$ as in (2.5), we define the mdG$(q)^*$ method for (2.9) as follows: Find $\Phi \in \hat{V}$ with $\Phi(T^+) = \psi$ such that

$$\sum_{i=1}^{N} \sum_{j=1}^{M_i} \left[[v_i]_{i,j-1} \Phi_i\big(t_{i,j-1}^+\big) + \int_{I_{ij}} \dot{v}_i \Phi_i \, dt \right] = \int_0^T (J(\pi u, U, \cdot) v, \Phi) \, dt + L_{\psi,g}(v) \tag{2.15}$$

for all $v \in V$ with $v(0^-) = 0$. Integrating by parts, (2.15) can alternatively be expressed in the form

$$\sum_{i=1}^{N} \sum_{j=1}^{M_i} \left[-[\Phi_i]_{ij} v_i\big(t_{ij}^-\big) - \int_{I_{ij}} \dot{\Phi}_i v_i \, dt \right] = \int_0^T (J^\top(\pi u, U, \cdot)\Phi + g, v) \, dt. \tag{2.16}$$

3. Existence of solutions.

To prove existence of the discrete mcG(q), mdG(q), mcG$(q)^*$, and mdG$(q)^*$ solutions defined in the previous section, we formulate fixed point iterations for the construction of solutions. Existence then follows from the Banach fixed point theorem if the time steps are sufficiently small.

Lemma 3.1 (fixed point iteration). *Let $\mathcal{T}_n$ be a time slab with synchronized time levels T_{n-1} and T_n. With time reversed for the dual methods (to simplify the notation), the* mcG(q)*,* mdG(q)*,* mcG$(q)^*$*, and* mdG$(q)^*$ *methods can all be expressed in the following form: For all $I_{ij} \in \mathcal{T}_n$, find $\{\xi_{ijn}\}$ (the degrees of freedom for U_i on I_{ij}) such that*

$$\xi_{ijn} = u_i(0) + \int_0^{t_{i,j-1}} f_i(U, \cdot) \, dt + \int_{I_{ij}} w_n^{[q_{ij}]}(\tau_{ij}(t)) f_i(U, \cdot) \, dt, \tag{3.1}$$

where $\tau_{ij}(t) = (t - t_{i,j-1})/(t_{ij} - t_{i,j-1})$ and $\{w_n^{[q_{ij}]}\}$ is a set of polynomial weight functions on $[0, 1]$.

Proof. The result follows from the definitions of the mcG(q), mdG(q), mcG$(q)^*$, and mdG$(q)^*$ methods, using an appropriate basis for the trial and test spaces. See [34] for details. □

Theorem 3.2 (existence of solutions). *Let $K = \max K_n$ be the maximum time slab length and define the Lipschitz constant $L_f > 0$ by*

$$\|f(x,t) - f(y,t)\|_{l_\infty} \le L_f \|x - y\|_{l_\infty} \quad \forall t \in [0, T] \ \forall x, y \in \mathbb{R}^N. \tag{3.2}$$

If now

$$K C L_f < 1, \tag{3.3}$$

where $C = C(q) > 0$ is a constant depending only on the order and method, the fixed point iteration (3.1) *converges to the unique solution of* (2.2), (2.6), (2.12), *and* (2.15), *respectively.*

Proof. The result follows by Lemma 3.1 and an application of the Banach fixed point theorem. See [34] for details. □

4. Stability of solutions.

Write the dual problem (2.9) for $\phi = \phi(t)$ in the form

$$\begin{aligned} -\dot{\phi}(t) + A^\top(t)\phi(t) &= g, \quad t \in [0, T), \\ \phi(T) &= \psi. \end{aligned} \tag{4.1}$$

For simplicity, we consider only the case $g = 0$. With $w(t) = \phi(T - t)$, we have $\dot{w}(t) = -\dot{\phi}(T - t) = -A^\top(T - t)w(t)$, and so (4.1) can be written as a forward problem for w in the form

$$\begin{aligned} \dot{w}(t) + B(t)w(t) = 0, \quad t \in (0, T], \\ w(0) = w_0, \end{aligned} \tag{4.2}$$

where $w_0 = \psi$ and $B(t) = A^\top(T - t)$. Below, w represents either u or $\phi(T - \cdot)$ and, correspondingly, W represents either the discrete mc/dG(q) approximation U of u or the discrete mc/dG(q)* approximation Φ of ϕ.

4.1. A general exponential estimate.

The general exponential stability estimate is based on the following version of the discrete Gronwall inequality.

Lemma 4.1 (discrete Gronwall inequality). *Assume that $z, a : \mathbb{N} \to \mathbb{R}$ are nonnegative, $a(m) \leq 1/2$ for all m, and $z(n) \leq C + \sum_{m=1}^{n} a(m)z(m)$ for all n. Then $z(n) \leq 2C \exp(\sum_{m=1}^{n-1} 2a(m))$ for $n = 1, 2, \ldots$.*

Proof. By a standard discrete Gronwall inequality [38], $z(n) \leq C \exp(\sum_{m=0}^{n-1} a(m))$ if $z(n) \leq C + \sum_{m=0}^{n-1} a(m)z(m)$ for $n \geq 1$ and $z(0) \leq C$. Here, $(1 - a(n))z(n) \leq C + \sum_{m=1}^{n-1} a(m)z(m)$, and so $z(n) \leq 2C + \sum_{m=1}^{n-1} 2a(m)z(m)$, since $1 - a(n) \geq 1/2$. The result now follows if we take $a(0) = z(0) = 0$. □

Theorem 4.2 (stability estimate). *Let W be the* mcG(q), mdG(q), mcG(q)*, *or* mdG(q)* *solution of* (4.2). *Then there is a constant $C = C(q)$, depending only on the highest order* $\max q_{ij}$, *such that if $K_n C \|B\|_{L_\infty([T_{n-1},T_n],l_p)} \leq 1$ for $n = 1, \ldots, M$, then*

$$\|W\|_{L_\infty([T_{n-1},T_n],l_p)} \leq C\|w_0\|_{l_p} \exp\left(\sum_{m=1}^{n-1} K_m C \|B\|_{L_\infty([T_{m-1},T_m],l_p)}\right) \tag{4.3}$$

for $n = 1, \ldots, M$, $1 \leq p \leq \infty$.

Proof. By Lemma 3.1, we can write the mcG(q), mdG(q), mcG(q)*, and mdG(q)* methods in the form $\xi_{ijn'} = w_i(0) + \int_0^{t_{i,j-1}} f_i(W, \cdot)\, dt + \int_{I_{ij}} w_{n'}^{[q_{ij}]}(\tau_{ij}(t)) f_i(W, \cdot)\, dt$. Applied to the linear model problem (4.2), we have $\xi_{ijn'} = w_i(0) - \int_0^{t_{i,j-1}} (BW)_i\, dt - \int_{I_{ij}} w_{n'}^{[q_{ij}]}(\tau_{ij}(t))(BW)_i\, dt$, and so

$$\begin{aligned} |\xi_{ijn'}| &\leq |w_i(0)| + \left| \int_0^{t_{i,j-1}} (BW)_i\, dt \right| + \left| \int_{I_{ij}} w_{n'}^{[q_{ij}]}(\tau_{ij}(t))(BW)_i\, dt \right| \\ &\leq |w_i(0)| + C \int_0^{t_{ij}} |(BW)_i|\, dt \leq |w_i(0)| + C \int_0^{T_n} |(BW)_i|\, dt, \end{aligned}$$

where T_n is smallest synchronized time level for which $t_{ij} \leq T_n$. It now follows that for all $t \in [T_{n-1}, T_n]$, we have $|W_i(t)| \leq C|w_i(0)| + C \int_0^{T_n} |(BW)_i|\, dt$, and so

$$\|W(t)\|_{l_p} \leq C\|w_0\|_{l_p} + C \int_0^{T_n} \|BW\|_{l_p}\, dt = C\|w_0\|_{l_p} + C \sum_{m=1}^{n} \int_{T_{m-1}}^{T_m} \|BW\|_{l_p}\, dt.$$

The result now follows by letting $\bar{W}_n = \|W\|_{L_\infty([T_{n-1},T_n],l_p)}$. □

Remark 4.1. *See* [34] *for an extension to multiadaptive time-stepping of the strong stability estimate Lemma* 6.1 *for parabolic problems in* [11].

5. Interpolation estimates. In this section, we introduce a pair of carefully chosen interpolants, $\pi_{\mathrm{cG}}^{[q]}$ and $\pi_{\mathrm{dG}}^{[q]}$, which are central to the a priori error analysis of the mcG(q) and mdG(q) methods. The interpolants are defined in section 5.1. In section 5.2, we discuss the interpolation of piecewise smooth functions, that is, the interpolation of functions which may be discontinuous within the interval of interpolation, and then present the basic general interpolation estimates for the two interpolants $\pi_{\mathrm{cG}}^{[q]}$ and $\pi_{\mathrm{dG}}^{[q]}$.

For the a priori error analysis of the mcG(q) and mdG(q) methods, we will also need a special interpolation estimate for the function $\varphi = J^\top \Phi$, where J is the Jacobian of the right-hand side f of (1.1) and Φ is the discrete dual solution as defined in section 2, including estimates for the size of the jump in function value and derivatives for the function φ at points of discontinuity. These estimates are proved in section 5.3, based on a representation formula for the mcG(q) and mdG(q) solutions of (1.1).

5.1. Interpolants. The interpolant $\pi_{\mathrm{cG}}^{[q]} : V \rightarrow \mathcal{P}^q([a,b])$ is defined by the following conditions:

$$\begin{aligned} &\pi_{\mathrm{cG}}^{[q]} v(a) = v(a) \quad \text{and} \quad \pi_{\mathrm{cG}}^{[q]} v(b) = v(b), \\ &\int_a^b \big(v - \pi_{\mathrm{cG}}^{[q]} v\big) w \, dx = 0 \quad \forall w \in \mathcal{P}^{q-2}([a,b]), \end{aligned} \tag{5.1}$$

where V denotes the set of functions that are piecewise $\mathcal{C}^{q+1}$ and bounded on $[a,b]$. In other words, $\pi_{\mathrm{cG}}^{[q]} v$ is the polynomial of degree q that interpolates v at the end-points of the interval $[a,b]$ and additionally satisfies $q-1$ projection conditions. This is illustrated in Figure 2. We also define the dual interpolant $\pi_{\mathrm{cG}^*}^{[q]}$ as the standard L_2-projection onto $\mathcal{P}^{q-1}([a,b])$.

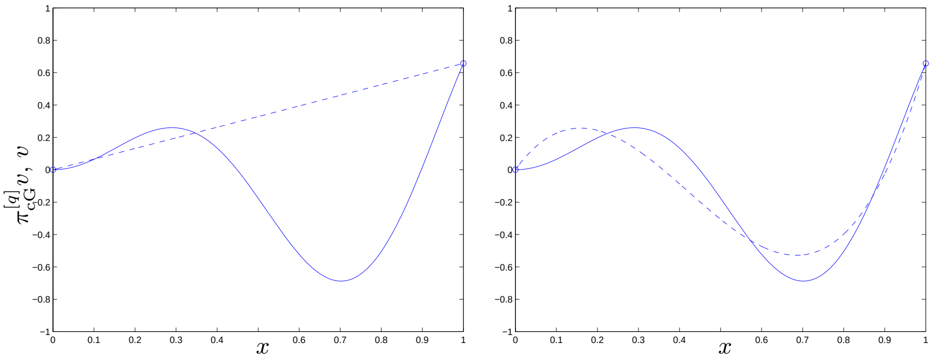


FIG. 2. *The interpolant* $\pi_{\mathrm{cG}}^{[q]} v$ *(dashed) of the function* $v(x) = x\,\sin(7x)$ *(solid) on* $[0,1]$ *for* $q = 1$ *(left) and* $q = 3$ *(right).*

The interpolant $\pi_{\mathrm{dG}}^{[q]} : V \rightarrow \mathcal{P}^q([a,b])$ is defined by the following conditions:

$$\begin{aligned} &\pi_{\mathrm{dG}}^{[q]} v(b) = v(b), \\ &\int_a^b \big(v - \pi_{\mathrm{dG}}^{[q]} v\big) w \, dx = 0 \quad \forall w \in \mathcal{P}^{q-1}([a,b]), \end{aligned} \tag{5.2}$$

that is, $\pi_{\mathrm{dG}}^{[q]} v$ is the polynomial of degree q that interpolates v at the right end-point of the interval $[a,b]$ and additionally satisfies q projection conditions. This is illustrated

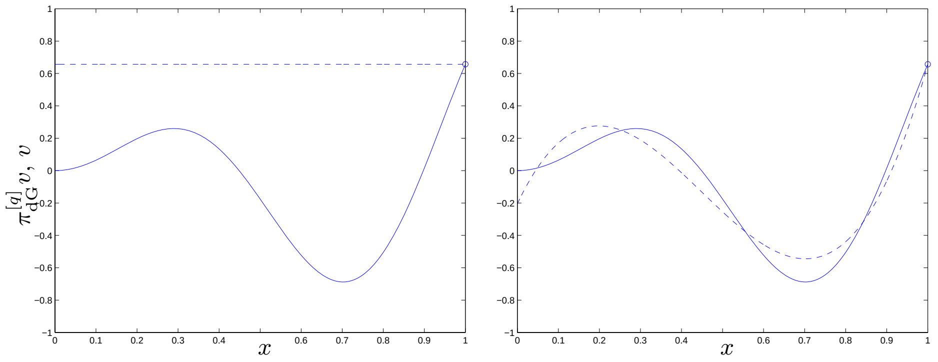


FIG. 3. *The interpolant $\pi_{\mathrm{dG}}^{[q]}v$ (dashed) of the function $v(x) = x\,\sin(7x)$ (solid) on $[0,1]$ for $q = 0$ (left) and $q = 3$ (right).*

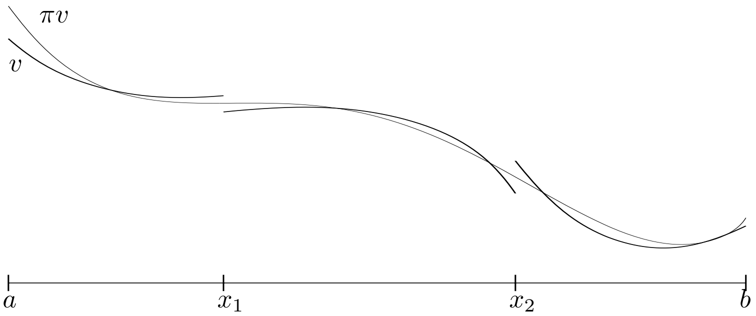


FIG. 4. *A piecewise smooth function v and its interpolant πv.*

in Figure 3. The dual interpolant $\pi_{\mathrm{dG}^*}^{[q]}$ is defined similarly, with the difference being that the left end-point $x = a$ is used for interpolation.

5.2. Basic interpolation estimates. To estimate the size of the interpolation error $\pi v - v$ for a given function v, we express the interpolation error in terms of the regularity of v and the length of the interpolation interval, $k = b - a$. Specifically, when $v \in \mathcal{C}^{q+1}([a,b]) \subset V$ for some $q \geq 0$, we obtain estimates of the form

$$\big\|(\pi v)^{(p)} - v^{(p)}\big\| \leq Ck^{q+1-p}\big\|v^{(q+1)}\big\|, \quad p = 0, \dots, q+1, \tag{5.3}$$

where $\|\cdot\| = \|\cdot\|_{L_\infty([a,b])}$ denotes the maximum norm on $[a,b]$. This estimate is a simple consequence of the Peano kernel theorem [40] if one can show that the interpolant $\pi : V \to \mathcal{P}^q([a,b]) \subset V$ is linear and bounded on V and that π is exact on $\mathcal{P}^q([a,b]) \subset V$, that is, $\pi v = v$ for all $v \in \mathcal{P}^q([a,b])$.

In the general case, where the interpolated function v is only piecewise smooth (see Figure 4), we also need to include the size of the jump $[v^{(p)}]_x$ in function value and derivatives at each point x of discontinuity within (a,b) to measure the regularity of the interpolated function v. In [34], we prove the following extensions of the basic estimate (5.3).

LEMMA 5.1. *If π is linear and bounded on V and is exact on $\mathcal{P}^q([a,b]) \subset V$, then there is a constant $C = C(q) > 0$ such that for all v piecewise $\mathcal{C}^{q+1}$ on $[a,b]$ with*

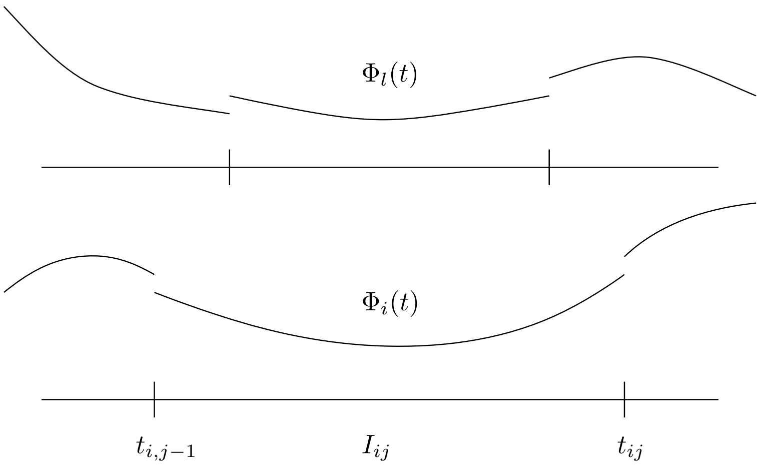


FIG. 5. *If some other component $l \neq i$ has a node within I_{ij}, then Φ_l may be discontinuous within I_{ij}, causing φ_i to be discontinuous within I_{ij}.*

discontinuities at $a < x_1 < \cdots < x_n < b$,

$$\big\|(\pi v)^{(p)} - v^{(p)}\big\| \leq Ck^{r+1-p}\big\|v^{(r+1)}\big\| + C\sum_{j=1}^{n}\sum_{m=0}^{r} k^{m-p}\big|\big[v^{(m)}\big]_{x_j}\big| \tag{5.4}$$

for $p = 0, \ldots, r+1$, $r = 0, \ldots, q$.

LEMMA 5.2. *If π is linear and bounded on V and is exact on $\mathcal{P}^q([a,b]) \subset V$, then there is a constant $C = C(q) > 0$ such that for all v piecewise $\mathcal{C}^{q+1}$ on $[a,b]$ with discontinuities at $a < x_1 < \cdots < x_n < b$,*

$$\big\|(\pi v)^{(p)}\big\| \leq C\big\|v^{(p)}\big\| + C\sum_{j=1}^{n}\sum_{m=0}^{p-1} k^{m-p}\big|\big[v^{(m)}\big]_{x_j}\big| \tag{5.5}$$

for $p = 0, \ldots, q$.

Lemmas 5.1 and 5.2 apply to both the $\pi_{\mathrm{cG}}^{[q]}$ interpolant (for $q \geq 1$) and the $\pi_{\mathrm{dG}}^{[q]}$ interpolant (for $q \geq 0$) defined in section 5.1. The linearity of both interpolants follows directly from the definition of the interpolants. The proofs that both interpolants are bounded and exact on $\mathcal{P}^q([a,b])$ are given in detail in [34] and [35].

5.3. A special interpolation estimate.

To prove a priori error estimates for mcG(q) and mdG(q) in section 6, we need to estimate the interpolation error $\pi\varphi - \varphi$ for the function φ defined by

$$\varphi_i = (J^\top(\pi u, u, \cdot)\Phi)_i = \sum_{l=1}^{N} J_{li}(\pi u, u, \cdot)\Phi_l, \quad i = 1, \ldots, N. \tag{5.6}$$

We note that φ_i may be discontinuous within I_{ij} if I_{ij} contains a node for some other component, which is generally the case. This is illustrated in Figure 5. Note that on the right-hand side f is linearized around a mean value of πu and u.

An interpolation estimate for $\pi\varphi - \varphi$ follows directly from Lemma 5.1. To use this estimate, we need to estimate the size of the jump in function value and derivatives at each internal node t_{ij} of the partition $\mathcal{T}$. To obtain this estimate, we need to make a number of additional assumptions on the right-hand side f of (1.1) and the partition $\mathcal{T}$. These assumptions are discussed in section 5.3.2. Based on the assumptions and the representation formula presented in section 5.3.1, we obtain the jump estimates in section 5.3.3 and, finally, in section 5.3.4, the interpolation estimate for φ.

5.3.1. A representation formula. The proof of jump estimates for the multiadaptive Galerkin methods mcG(q) and mdG(q) is based on expressing the solutions as certain interpolants. These representations are obtained as follows. Let U be the mcG(q) or mdG(q) solution of (1.1) and define, for $i = 1, \dots, N$,

$$\tilde{U}_i(t) = u_i(0) + \int_0^t f_i(U(s), s)\, ds. \tag{5.7}$$

Similarly, for Φ the mcG(q)* or mdG(q)* solution of (2.9), we define, for $i = 1, \dots, N$,

$$\tilde{\Phi}_i(t) = \psi_i + \int_t^T f_i^*(\Phi(s), s)\, ds, \tag{5.8}$$

where $f^*(\Phi, \cdot) = J^\top(\pi u, U, \cdot)\Phi + g$. We note that $\dot{\tilde{U}} = f(U, \cdot)$ and $-\dot{\tilde{\Phi}} = f^*(\Phi, \cdot)$.

It now turns out that U can be expressed as an interpolant of $\tilde{U}$. Similarly, Φ can be expressed as an interpolant of $\tilde{\Phi}$. We present these representations in Lemmas 5.3 and 5.4. We remind the reader about the interpolants $\pi_{\mathrm{cG}}^{[q]}$, $\pi_{\mathrm{cG}^*}^{[q]}$, $\pi_{\mathrm{dG}}^{[q]}$, and $\pi_{\mathrm{dG}^*}^{[q]}$, defined in section 5.1.

LEMMA 5.3. *The* mcG(q) *solution U of* (1.1) *can expressed in the form $U = \pi_{\mathrm{cG}}^{[q]}\tilde{U}$. Similarly, the* mcG($q$)* *solution Φ of* (2.9) *can be expressed in the form $\Phi = \pi_{\mathrm{cG}^*}^{[q]}\tilde{\Phi}$, that is, $U_i = \pi_{\mathrm{cG}}^{[q_{ij}]}\tilde{U}_i$ and $\Phi_i = \pi_{\mathrm{cG}^*}^{[q_{ij}]}\tilde{\Phi}_i$ on each local interval I_{ij}.*

Proof. The representation formulas follow by the definitions of the mcG(q) and mcG(q)* methods and the interpolants $\pi_{\mathrm{cG}}^{[q]}$ and $\pi_{\mathrm{cG}^*}^{[q]}$. See [34] for details. □

LEMMA 5.4. *The* mdG(q) *solution U of* (1.1) *can expressed in the form $U = \pi_{\mathrm{dG}}^{[q]}\tilde{U}$. Similarly, the* mdG($q$)* *solution Φ of* (2.9) *can be expressed in the form $\Phi = \pi_{\mathrm{dG}^*}^{[q]}\tilde{\Phi}$, that is, $U_i = \pi_{\mathrm{dG}}^{[q_{ij}]}\tilde{U}_i$ and $\Phi_i = \pi_{\mathrm{dG}^*}^{[q_{ij}]}\tilde{\Phi}_i$ on each local interval I_{ij}.*

Proof. The representation formulas follow by the definitions of the mdG(q) and mdG(q)* methods and the interpolants $\pi_{\mathrm{dG}}^{[q]}$ and $\pi_{\mathrm{dG}^*}^{[q]}$. See [34] for details. □

5.3.2. Assumptions. To estimate the size of the jump in function value and derivatives for the function φ defined in (5.6), we make the following assumptions. Given a time slab $\mathcal{T}$, assume that for each pair of local intervals I_{ij} and I_{mn} within the time slab, we have

$$q_{ij} = q_{mn} = \bar{q} \tag{A1}$$

and

$$k_{ij} > \alpha\, k_{mn} \tag{A2}$$

for some $\bar{q} \geq 0$ and some $\alpha \in (0, 1)$. The dependence on α in the error estimates is weak (see Remark 5.1), so assumption (A2) does not prevent multiadaptivity.

We also assume that the problem (1.1) is autonomous,

$$\partial f_i / \partial t = 0, \quad i = 1, \dots, N, \tag{A3}$$

noting that the dual problem nevertheless will be nonautonomous in general. Furthermore, we assume that

$$\text{(A4)} \qquad \|f_i\|_{D^{\bar{q}+1}(\mathcal{T})} < \infty, \quad i = 1, \dots, N,$$

where $\|\cdot\|_{D^p(\mathcal{T})}$ is defined for $v : \mathbb{R}^N \to \mathbb{R}$ and $p \geq 0$ by $\|v\|_{D^p(\mathcal{T})} = \max_{n=0,\dots,p} \|D^n v\|_{L_\infty(\mathcal{T},l_\infty)}$, with the norm $\|D^n v\|_{L_\infty(\mathcal{T},l_\infty)}$ defined by $\|D^n v\, w^1 \cdots w^n\|_{L_\infty(\mathcal{T})} \leq \|D^n v\|_{L_\infty(\mathcal{T},l_\infty)} \|w^1\|_{l_\infty} \cdots \|w^n\|_{l_\infty}$ for all $w^1, \dots, w^n \in \mathbb{R}^N$, and $D^n v$ the nth-order tensor given by

$$D^n v\, w^1 \cdots w^n = \sum_{i_1=1}^{N} \cdots \sum_{i_n=1}^{N} \frac{\partial^n v}{\partial x_{i_1} \cdots \partial x_{i_n}} w^1_{i_1} \cdots w^n_{i_n}.$$

Furthermore, we choose $C_f \geq \max_{i=1,\dots,N} \|f_i\|_{D^{\bar{q}+1}(\mathcal{T})}$ such that

$$\text{(5.9)} \qquad \|d^p/dt^p (\partial f/\partial u)^\top (x(t))\|_{l_\infty} \leq C_f C_x^p$$

for $p = 0, \dots, \bar{q}$, and

$$\text{(5.10)} \qquad \left\| [d^p/dt^p (\partial f/\partial u)^\top (x(t))]_t \right\|_{l_\infty} \leq C_f \sum_{n=0}^{p} C_x^{p-n} \left\| \left[x^{(n)} \right]_t \right\|_{l_\infty}$$

for $p = 0, \dots, \bar{q} - 1$ and any given $x : \mathbb{R} \to \mathbb{R}^N$, where $C_x > 0$ denotes a constant such that $\|x^{(n)}\|_{L_\infty(\mathcal{T},l_\infty)} \leq C_x^n$ for $n = 1, \dots, p$. Note that $C_f = C_f(t)$ defines a piecewise constant function on the partition $0 = T_0 < T_1 < \cdots < T_M = T$. Note also that assumption (A4) implies that each f_i is bounded by C_f.

We further assume that there is a constant $c_k > 0$ such that

$$\text{(A5)} \qquad k_{ij} C_f \leq c_k$$

for each local interval I_{ij}. We summarize the list of assumptions as follows:

(A1) the local orders q_{ij} are equal within each time slab;
(A2) the local time steps k_{ij} are semiuniform within each time slab;
(A3) f is autonomous;
(A4) f and its derivatives are bounded;
(A5) the local time steps k_{ij} are small.

5.3.3. Estimates of derivatives and jumps. To estimate higher-order derivatives, we face the problem of taking higher-order derivatives of $f(U(t), t)$ with respect to t. In Lemmas 5.5 and 5.6, we present basic estimates for composite functions $v \circ x$ with $v : \mathbb{R}^N \to \mathbb{R}$ and $x : \mathbb{R} \to \mathbb{R}^N$. The proofs are based on a straightforward application of the chain rule and Leibniz rule and are given in full detail in [34].

LEMMA 5.5. *Let $v : \mathbb{R}^N \to \mathbb{R}$ be $p \geq 0$ times differentiable in all its variables, let $x : \mathbb{R} \to \mathbb{R}^N$ be p times differentiable, and let $C_x > 0$ be a constant such that $\|x^{(n)}\|_{L_\infty(\mathbb{R},l_\infty)} \leq C_x^n$ for $n = 1, \dots, p$. Then there is a constant $C = C(p) > 0$ such that*

$$\text{(5.11)} \qquad \left\| \frac{d^p (v \circ x)}{dt^p} \right\|_{L_\infty(\mathbb{R})} \leq C \|v\|_{D^p(\mathbb{R})} C_x^p.$$

LEMMA 5.6. *Let $v : \mathbb{R}^N \to \mathbb{R}$ be $p + 1 \geq 1$ times differentiable in all its variables, let $x : \mathbb{R} \to \mathbb{R}^N$ be p times differentiable, except possibly at some $t \in \mathbb{R}$, and let*

$C_x > 0$ be a constant such that $\|x^{(n)}\|_{L_\infty(\mathbb{R},l_\infty)} \leq C_x^n$ for $n = 1,\dots,p$. Then there is a constant $C = C(p) > 0$ such that

$$\left|\left[\frac{d^p(v\circ x)}{dt^p}\right]_t\right| \leq C\|v\|_{D^{p+1}(\mathbb{R})}\sum_{n=0}^{p} C_x^{p-n}\big\|\left[x^{(n)}\right]_t\big\|_{l_\infty}. \tag{5.12}$$

We now prove estimates for derivatives and jumps of the mcG(q) or mdG(q) solution U of the general nonlinear problem (1.1), under the assumptions listed in section 5.3.2. Similarly, one can obtain estimates for the discrete dual solution Φ and the function φ defined in (5.6), from which the desired interpolation estimates follow.

To obtain estimates for the multiadaptive solution U, we first prove estimates for the function $\tilde{U}$ defined in section 5.3.1. The estimates for U then follow by induction.

To simplify the estimates, we introduce the following notation. For given $p > 0$, let $C_{U,p} \geq C_f$ be a constant such that

$$\big\|U^{(n)}\big\|_{L_\infty(\mathcal{T},l_\infty)} \leq C_{U,p}^n, \quad n = 1,\dots,p. \tag{5.13}$$

For $p = 0$, we define $C_{U,0} = C_f$. Temporarily, we assume that there is a constant $c_k' > 0$ such that for each p,

$$k_{ij}C_{U,p} \leq c_k'. \tag{A5$'$}$$

This assumption will be removed in Lemma 5.9. In the following lemma, we use assumptions (A1), (A3), and (A4) to derive estimates for $\tilde{U}$ in terms of $C_{U,p}$ and C_f.

LEMMA 5.7 (derivative and jump estimates for $\tilde{U}$). *Let U be the* mcG(q) *or* mdG(q) *solution of* (1.1) *and define $\tilde{U}$ as in* (5.7)*. If assumptions* (A1), (A3), *and* (A4) *hold, then there is a constant $C = C(\bar{q}) > 0$ such that*

$$\big\|\tilde{U}^{(p)}\big\|_{L_\infty(\mathcal{T},l_\infty)} \leq CC_{U,p-1}^p, \quad p = 1,\dots,\bar{q}+1, \tag{5.14}$$

and

$$\big\|\left[\tilde{U}^{(p)}\right]_{t_{i,j-1}}\big\|_{l_\infty} \leq C\sum_{n=0}^{p-1} C_{U,p-1}^{p-n}\big\|\left[U^{(n)}\right]_{t_{i,j-1}}\big\|_{l_\infty}, \quad p = 1,\dots,\bar{q}+1, \tag{5.15}$$

for each local interval I_{ij}, where $t_{i,j-1}$ is an internal node of the time slab $\mathcal{T}$.

Proof. By definition, $\tilde{U}_i^{(p)} = \frac{d^{p-1}}{dt^{p-1}}f_i(U)$, and so the results follow directly by Lemmas 5.5 and 5.6, noting that $C_f \leq C_{U,p-1}$. □

By Lemma 5.7, we now obtain the following estimate for the size of the jump in function value and derivatives for U.

LEMMA 5.8 (jump estimates for U). *Let U be the* mcG(q) *or* mdG(q) *solution of* (1.1)*. If assumptions* (A1)–(A5) *and* (A5$'$) *hold, then there is a constant $C = C(\bar{q}, c_k, c_k', \alpha) > 0$ such that*

$$\big\|\left[U^{(p)}\right]_{t_{i,j-1}}\big\|_{l_\infty} \leq Ck_{ij}^{r+1-p}C_{U,r}^{r+1}, \quad p = 0,\dots,r+1, \quad r = 0,\dots,\bar{q}, \tag{5.16}$$

for each local interval I_{ij}, where $t_{i,j-1}$ is an internal node of the time slab $\mathcal{T}$.

Proof. The proof is by induction. We first note that at $t = t_{i,j-1}$, we have

$$\begin{aligned}\left[U_i^{(p)}\right]_t &= U_i^{(p)}(t^+) - \tilde{U}_i^{(p)}(t^+) + \tilde{U}_i^{(p)}(t^+) - \tilde{U}_i^{(p)}(t^-) + \tilde{U}_i^{(p)}(t^-) - U_i^{(p)}(t^-)\\ &\equiv e_+ + e_0 + e_-.\end{aligned}$$

By Lemma 5.3 (or Lemma 5.4), U is an interpolant of $\tilde{U}$ and so, by Lemma 5.1, we have

$$|e_+| \le C k_{ij}^{r+1-p} \big\| \tilde{U}_i^{(r+1)} \big\|_{L_\infty(I_{ij})} + C \sum_{x \in \mathcal{N}_{ij}} \sum_{m=1}^{r} k_{ij}^{m-p} \big| \big[\tilde{U}_i^{(m)} \big]_x \big|$$

for $p = 0, \ldots, r+1$ and $r = 0, \ldots, \bar{q}$. Note that the second sum starts at $m = 1$ rather than at $m = 0$, since $\tilde{U}$ is continuous. Similarly, we have

$$|e_-| \le C k_{i,j-1}^{r+1-p} \big\| \tilde{U}_i^{(r+1)} \big\|_{L_\infty(I_{i,j-1})} + C \sum_{x \in \mathcal{N}_{i,j-1}} \sum_{m=1}^{r} k_{i,j-1}^{m-p} \big| \big[\tilde{U}_i^{(m)} \big]_x \big|.$$

To estimate e_0, we note that $e_0 = 0$ for $p = 0$, since $\tilde{U}$ is continuous. For $p = 1, \ldots, \bar{q}+1$, Lemma 5.7 gives $|e_0| = |[\tilde{U}_i^{(p)}]_t| \le C \sum_{n=0}^{p-1} C_{U,p-1}^{p-n} \|[U^{(n)}]_t\|_{l_\infty}$. By assumption (A2), it then follows that (5.16) holds for $r = 0$.

Assume now that (5.16) holds for $r = \bar{r} - 1 \ge 0$. Then, by Lemma 5.7 and assumption (A5′), it follows that

$$\begin{aligned} |e_+| &\le C k_{ij}^{\bar{r}+1-p} C_{U,\bar{r}}^{\bar{r}+1} + C \sum_{x \in \mathcal{N}_{ij}} \sum_{m=1}^{\bar{r}} k_{ij}^{m-p} \sum_{n=0}^{m-1} C_{U,m-1}^{m-n} \big\| [U^n]_x \big\|_{l_\infty} \\ &\le C k_{ij}^{\bar{r}+1-p} C_{U,\bar{r}}^{\bar{r}+1} + C \sum k_{ij}^{m-p} C_{U,m-1}^{m-n} k_{ij}^{(\bar{r}-1)+1-n} C_{U,\bar{r}-1}^{(\bar{r}-1)+1} \\ &\le C k_{ij}^{\bar{r}+1-p} C_{U,\bar{r}}^{\bar{r}+1} \Big(1 + \sum (k_{ij} C_{U,\bar{r}-1})^{m-1-n} \Big) \le C k_{ij}^{\bar{r}+1-p} C_{U,\bar{r}}^{\bar{r}+1}. \end{aligned}$$

Similarly, we obtain the estimate $|e_-| \le C k_{ij}^{\bar{r}+1-p} C_{U,\bar{r}}^{\bar{r}+1}$. Finally, we use Lemma 5.7 and assumption (A5′) to obtain the estimate

$$\begin{aligned} |e_0| &\le C \sum_{n=0}^{p-1} C_{U,p-1}^{p-n} \big\| [U^n]_t \big\|_{l_\infty} \le C \sum_{n=0}^{p-1} C_{U,p-1}^{p-n} k_{ij}^{(\bar{r}-1)+1-n} C_{U,\bar{r}-1}^{(\bar{r}-1)+1} \\ &= C k_{ij}^{\bar{r}+1-p} C_{U,\bar{r}}^{\bar{r}+1} \sum_{n=0}^{p-1} (k_{ij} C_{U,\bar{r}})^{p-1-n} \le C k_{ij}^{\bar{r}+1-p} C_{U,\bar{r}}^{\bar{r}+1}. \end{aligned}$$

Summing up, we thus obtain $|[U_i^{(p)}]_t| \le |e_+| + |e_0| + |e_-| \le C k_{ij}^{\bar{r}+1-p} C_{U,\bar{r}}^{\bar{r}+1}$, and so (5.16) follows by induction. □

By Lemmas 5.7 and 5.8, we now obtain the following estimate for derivatives of the solution U.

Lemma 5.9 (derivative estimates for U). *Let* U *be the* mcG(q) *or* mdG(q) *solution of* (1.1). *If assumptions* (A1)–(A5) *hold, then there is a constant* $C = C(\bar{q}, c_k, \alpha) > 0$ *such that*

$$\big\| U^{(p)} \big\|_{L_\infty(\mathcal{T}, l_\infty)} \le C C_f^p, \quad p = 1, \ldots, \bar{q}. \tag{5.17}$$

Proof. By Lemma 5.3 (or Lemma 5.4), U is an interpolant of $\tilde{U}$ and so, by Lemma 5.1, we have

$$\big\| U_i^{(p)} \big\|_{L_\infty(I_{ij})} = \big\| (\pi \tilde{U}_i)^{(p)} \big\|_{L_\infty(I_{ij})} \le C' \big\| \tilde{U}_i^{(p)} \big\|_{L_\infty(I_{ij})} + C' \sum_{x \in \mathcal{N}_{ij}} \sum_{m=1}^{p-1} k_{ij}^{m-p} \big| \big[\tilde{U}_i^{(m)} \big]_x \big|$$

for some constant $C' = C'(\bar{q})$. For $p = 1$, we thus obtain the estimate

$$\|\dot{U}_i\|_{L_\infty(I_{ij})} \leq C' \|\dot{\tilde{U}}_i\|_{L_\infty(I_{ij})} = C' \|f_i(U)\|_{L_\infty(I_{ij})} \leq C' C_f$$

by assumption (A4), and so (5.17) holds for $p = 1$.

For $p = 2, \ldots, \bar{q}$, assuming that (A5$'$) holds for $C_{U,p-1}$, we use Lemmas 5.7 and 5.8 (with $r = p - 1$) and assumption (A2) to obtain

$$\begin{aligned}\big\|U_i^{(p)}\big\|_{L_\infty(I_{ij})} &\leq C C_{U,p-1}^p + C \sum_{x \in \mathcal{N}_{ij}} \sum_{m=1}^{p-1} k_{ij}^{m-p} \sum_{n=0}^{m-1} C_{U,m-1}^{m-n} \big\| \big[U^{(n)}\big]_x \big\|_{l_\infty} \\ &\leq C C_{U,p-1}^p + C \sum k_{ij}^{m-p} C_{U,m-1}^{m-n} k_{ij}^{(p-1)+1-n} C_{U,p-1}^{(p-1)+1} \\ &\leq C C_{U,p-1}^p \Big(1 + \sum (k_{ij} C_{U,m-1})^{m-n}\Big) \leq C C_{U,p-1}^p,\end{aligned}$$

where $C = C(\bar{q}, c_k, c_k', \alpha)$. This holds for all components i and all local intervals I_{ij} within the time slab $\mathcal{T}$, and so

$$\big\|U^{(p)}\big\|_{L_\infty(\mathcal{T}, l_\infty)} \leq C C_{U,p-1}^p, \quad p = 1, \ldots, \bar{q},$$

where by definition $C_{U,p-1}$ is a constant such that $\|U^{(n)}\|_{L_\infty(\mathcal{T}, l_\infty)} \leq C_{U,p-1}^n$ for $n = 1, \ldots, p-1$. Starting at $p = 1$, we now define $C_{U,1} = C_1 C_f$ with $C_1 = C' = C'(\bar{q})$. It then follows that (A5$'$) holds for $C_{U,1}$ with $c_k' = C' c_k$, and thus

$$\big\|U^{(2)}\big\|_{L_\infty(\mathcal{T}, l_\infty)} \leq C C_{U,2-1}^2 = C C_{U,1}^2 \equiv C_2 C_f^2,$$

where $C_2 = C_2(\bar{q}, c_k, \alpha)$. We may thus define $C_{U,2} = \max(C_1 C_f, \sqrt{C_2} C_f)$. Continuing, we note that (A5$'$) holds for $C_{U,2}$, and thus

$$\big\|U^{(3)}\big\|_{L_\infty(\mathcal{T}, l_\infty)} \leq C C_{U,3-1}^3 = C C_{U,2}^3 \equiv C_3 C_f^3,$$

where $C_3 = C_3(\bar{q}, c_k, \alpha)$. In this way, we obtain a sequence of constants $C_1, \ldots, C_{\bar{q}}$, depending only on $\bar{q}$, c_k, and α, such that $\|U^{(p)}\|_{L_\infty(\mathcal{T}, l_\infty)} \leq C_p C_f^p$ for $p = 1, \ldots, \bar{q}$, and so (5.17) follows if we take $C = \max_{i=1,\ldots,\bar{q}} C_i$. □

Having now removed the additional assumption (A5$'$), we obtain the following version of Lemma 5.8.

LEMMA 5.10 (jump estimates for U). *Let U be the* mcG(q) *or* mdG(q) *solution of* (1.1). *If assumptions* (A1)–(A5) *hold, then there is a constant $C = C(\bar{q}, c_k, \alpha) > 0$ such that*

$$\big\| \big[U^{(p)}\big]_{t_{i,j-1}} \big\|_{l_\infty} \leq C k_{ij}^{\bar{q}+1-p} C_f^{\bar{q}+1}, \quad p = 0, \ldots, \bar{q}, \tag{5.18}$$

for each local interval I_{ij}, where $t_{i,j-1}$ is an internal node of the time slab $\mathcal{T}$.

Similarly, we obtain estimates for the discrete dual solution Φ and the function φ. In Lemma 5.11, we present the estimates for the function φ.

LEMMA 5.11 (estimates for φ). *Let φ be defined as in* (5.6). *If assumptions* (A1)–(A5) *hold, then there is a constant $C = C(\bar{q}, c_k, \alpha) > 0$ such that*

$$\big\|\varphi_i^{(p)}\big\|_{L_\infty(I_{ij})} \leq C C_f^{p+1} \|\Phi\|_{L_\infty(\mathcal{T}, l_\infty)}, \quad p = 0, \ldots, q_{ij}, \tag{5.19}$$

and

$$\big|\big[\varphi_i^{(p)}\big]_x\big| \leq C k_{ij}^{r_{ij}-p} C_f^{r_{ij}+1} \|\Phi\|_{L_\infty(\mathcal{T}, l_\infty)} \quad \forall x \in \mathcal{N}_{ij}, \quad p = 0, \ldots, q_{ij} - 1, \tag{5.20}$$

with $r_{ij} = q_{ij}$ for the mcG(q) *method and $r_{ij} = q_{ij} + 1$ for the* mdG(q) *method. This holds for each local interval I_{ij} within the time slab $\mathcal{T}$.*

5.3.4. Interpolation estimates. Using the basic interpolation estimate of section 5.2, we now obtain the following important interpolation estimates for the function φ.

LEMMA 5.12 (interpolation estimates for φ). *Let φ be defined as in* (5.6). *If assumptions* (A1)–(A5) *hold, then there is a constant $C = C(\bar{q}, c_k, \alpha) > 0$ such that*

$$\left\|\pi_{\mathrm{cG}}^{[q_{ij}-2]}\varphi_i - \varphi_i\right\|_{L_\infty(I_{ij})} \leq Ck_{ij}^{q_{ij}-1}C_f^{q_{ij}}\|\Phi\|_{L_\infty(\mathcal{T},l_\infty)}, \quad q_{ij} = \bar{q} \geq 2, \tag{5.21}$$

and

$$\left\|\pi_{\mathrm{dG}}^{[q_{ij}-1]}\varphi_i - \varphi_i\right\|_{L_\infty(I_{ij})} \leq Ck_{ij}^{q_{ij}}C_f^{q_{ij}+1}\|\Phi\|_{L_\infty(\mathcal{T},l_\infty)}, \quad q_{ij} = \bar{q} \geq 1, \tag{5.22}$$

for each local interval I_{ij} within the time slab $\mathcal{T}$.

Proof. To prove (5.21), we use Lemma 5.1, with $r = q_{ij} - 2$ and $p = 0$, together with Lemma 5.11, to obtain

$$\begin{aligned}\left\|\pi_{\mathrm{cG}}^{[q_{ij}-2]}\varphi_i - \varphi_i\right\|_{L_\infty(I_{ij})} &\leq Ck_{ij}^{q_{ij}-1}\left\|\varphi_i^{(q_{ij}-1)}\right\|_{L_\infty(I_{ij})} + C\sum_{x\in\mathcal{N}_{ij}}\sum_{m=0}^{q_{ij}-2}k_{ij}^m\left|\left[\varphi_i^{(m)}\right]_x\right| \\ &\leq Ck_{ij}^{q_{ij}-1}C_f^{q_{ij}}\|\Phi\|_{L_\infty(\mathcal{T},l_\infty)} + C\sum_{x\in\mathcal{N}_{ij}}\sum_{m=0}^{q_{ij}-2}k_{ij}^m k_{ij}^{q_{ij}-m}C_f^{q_{ij}+1}\|\Phi\|_{L_\infty(\mathcal{T},l_\infty)} \\ &= Ck_{ij}^{q_{ij}-1}C_f^{q_{ij}}\|\Phi\|_{L_\infty(\mathcal{T},l_\infty)} + Ck_{ij}^{q_{ij}}C_f^{q_{ij}+1}\|\Phi\|_{L_\infty(\mathcal{T},l_\infty)},\end{aligned}$$

from which the estimate follows. The estimate for $\pi_{\mathrm{dG}}^{[q_{ij}-1]}\varphi_i - \varphi_i$ is obtained similarly. ▯

REMARK 5.1. *Note that there is only a weak dependence on c_k and α, since the jump term contains an extra factor k_{ij}. If higher-order terms can be ignored, then the dependence on c_k and α can be removed.*

6. A priori error estimates.

To prove a priori error estimates for the mcG(q) and mdG(q) methods, we derive error representations in section 6.1 and then obtain the a priori error estimates in section 6.2 for the general nonlinear case. We refer to [34] for a sharp a priori error estimate in the case of a parabolic model problem.

6.1. Error representation.

For each of the two methods, mcG(q) and mdG(q), we represent the error in terms of the discrete dual solution Φ and an interpolant πu of the exact solution u of (1.1), using the special interpolants $\pi u = \pi_{\mathrm{cG}}^{[q]}u$ or $\pi u = \pi_{\mathrm{dG}}^{[q]}u$ defined in section 5.

We write the error $e = U - u$ in the form

$$e = \bar{e} + (\pi u - u), \tag{6.1}$$

where $\bar{e} \equiv U - \pi u$ is represented in terms of the discrete dual solution and the residual of the interpolant. An estimate for the second part of the error, $\pi u - u$, follows directly from an interpolation estimate.

In Lemma 6.1, we present the error representation for the mcG(q) method, and then present the corresponding representation for the mdG(q) method in Lemma 6.2. The error representations are obtained directly by choosing $\bar{e}$ as a test function for the discrete dual problems (2.12) and (2.15).

LEMMA 6.1 (error representation for mcG(q)). *Let U be the* mcG(q) *solution of* (1.1)*, let Φ be the corresponding* mcG(q)* *solution of the dual problem* (2.9)*, and let*

πu be any trial space approximation of the exact solution u of (1.1) *that interpolates u at the end-points of every local interval. Then*

$$L_{\psi,g}(\bar{e}) \equiv (\bar{e}(T), \psi) + \int_0^T (\bar{e}, g)\, dt = -\int_0^T (R(\pi u, \cdot), \Phi)\, dt,$$

where $\bar{e} \equiv U - \pi u$.

LEMMA 6.2 (error representation for mdG(q)). *Let U be the* mdG(q) *solution of* (1.1)*, let Φ be the corresponding* mdG(q)* *solution of the dual problem* (2.9)*, and let πu be any trial space approximation of the exact solution u of* (1.1) *that interpolates u at the right end-point of every local interval. Then*

$$L_{\psi,g}(\bar{e}) = -\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left[[\pi u_i]_{i,j-1}\Phi_i\big(t_{i,j-1}^{+}\big) + \int_{I_{ij}} R_i(\pi u, \cdot)\Phi_i\, dt\right],$$

where $\bar{e} \equiv U - \pi u$.

With a special choice of interpolant, $\pi u = \pi_{\mathrm{cG}}^{[q]} u$ and $\pi u = \pi_{\mathrm{dG}}^{[q]} u$, respectively, we obtain the following versions of the error representations.

COROLLARY 6.3 (error representation for mcG(q)). *Let U be the* mcG(q) *solution of* (1.1) *and let Φ be the corresponding* mcG(q)* *solution of the dual problem* (2.9). *Then*

$$L_{\psi,g}(\bar{e}) = \int_0^T \big(f\big(\pi_{\mathrm{cG}}^{[q]} u, \cdot\big) - f(u, \cdot), \Phi\big)\, dt.$$

Proof. Integrate by parts and use the definition of the interpolant $\pi_{\mathrm{cG}}^{[q]}$. □

COROLLARY 6.4 (error representation for mdG(q)). *Let U be the* mdG(q) *solution of* (1.1) *and let Φ be the corresponding* mdG(q)* *solution of the dual problem* (2.9). *Then*

$$L_{\psi,g}(\bar{e}) = \int_0^T \big(f\big(\pi_{\mathrm{dG}}^{[q]} u, \cdot\big) - f(u, \cdot), \Phi\big)\, dt.$$

Proof. Integrate by parts and use the definition of the interpolant $\pi_{\mathrm{dG}}^{[q]}$. □

6.2. A priori error estimates for the general nonlinear problem. Using the error representations of section 6.1, the stability estimates of section 4, and the interpolation estimates of section 5, we now prove our main results: a priori error estimates for general order mcG(q) and mdG(q).

THEOREM 6.5 (a priori error estimate for mcG(q)). *Let U be the* mcG(q) *solution of* (1.1) *and let Φ be the corresponding* mcG(q)* *solution of the dual problem* (2.9). *Then there is a constant $C = C(q) > 0$ such that*

$$|L_{\psi,g}(\bar{e})| \leq CS(T)\big\|k^{q+1}\bar{u}^{(q+1)}\big\|_{L_\infty([0,T],l_2)}, \tag{6.2}$$

where $(k^{q+1}\bar{u}^{(q+1)})_i(t) = k_{ij}^{q_{ij}+1}\|u_i^{(q_{ij}+1)}\|_{L_\infty(I_{ij})}$ for $t \in I_{ij}$, and where the stability factor $S(T)$ is given by $S(T) = \int_0^T \|J^\top(\pi_{\mathrm{cG}}^{[q]} u, u, \cdot)\Phi\|_{l_2}\, dt$. Furthermore, if assumptions (A1)–(A5) *hold, then there is a constant $C = C(q, c_k, \alpha) > 0$ such that*

$$|L_{\psi,g}(\bar{e})| \leq C\bar{S}(T)\big\|k^{2q}\bar{\bar{u}}^{(2q)}\big\|_{L_\infty([0,T],l_1)}, \tag{6.3}$$

where $(k^{2q}\bar{\bar{u}}^{(2q)})_i(t) = k_{ij}^{2q_{ij}} C_f^{q_{ij}-1} \|u_i^{(q_{ij}+1)}\|_{L_\infty(I_{ij})}$ *for* $t \in I_{ij}$, *and where the stability factor* $\bar{S}(T)$ *is given by*

$$\bar{S}(T) = \int_0^T C_f \|\Phi\|_{L_\infty(\mathcal{T},l_\infty)}\, dt = \sum_{n=1}^M K_n C_f \|\Phi\|_{L_\infty(\mathcal{T}_n,l_\infty)}.$$

Proof. By Corollary 6.3, we obtain

$$L_{\psi,g}(\bar{e}) = \int_0^T \big(f(\pi_{\mathrm{cG}}^{[q]}u,\cdot) - f(u,\cdot),\Phi\big)\, dt = \int_0^T \big(\pi_{\mathrm{cG}}^{[q]}u - u, J^\top\big(\pi_{\mathrm{cG}}^{[q]}u,u,\cdot\big)\Phi\big)\, dt.$$

By Lemma 5.1, it now follows that

$$|L_{\psi,g}(\bar{e})| \leq C\|k^{q+1}\bar{u}^{q+1}\|_{L_\infty([0,T],l_2)} \int_0^T \big\|J^\top\big(\pi_{\mathrm{cG}}^{[q]}u,u,\cdot\big)\Phi\big\|_{l_2}\, dt,$$

which proves (6.2). To prove (6.3), we note that by definition, $\pi_{\mathrm{cG}}^{[q_{ij}]}u_i - u_i$ is orthogonal to $\mathcal{P}^{q_{ij}-2}(I_{ij})$ for each local interval I_{ij}, and so, recalling that $\varphi = J^\top(\pi_{\mathrm{cG}}^{[q]}u,u,\cdot)\Phi$,

$$L_{\psi,g}(\bar{e}) = \sum_{i,j}\int_{I_{ij}} \big(\pi_{\mathrm{cG}}^{[q_{ij}]}u_i - u_i\big)\varphi_i\, dt = \sum_{i,j}\int_{I_{ij}} \big(\pi_{\mathrm{cG}}^{[q_{ij}]}u_i - u_i\big)\big(\varphi_i - \pi_{\mathrm{cG}}^{[q_{ij}-2]}\varphi_i\big)\, dt,$$

where we take $\pi_{\mathrm{cG}}^{[q_{ij}-2]}\varphi_i \equiv 0$ for $q_{ij} = 1$. By Lemmas 5.1 and 5.12, it now follows that

$$\begin{aligned}
|L_{\psi,g}(\bar{e})| &\leq \int_0^T \big|\big(\pi_{\mathrm{cG}}^{[q]}u - u, \varphi - \pi_{\mathrm{cG}}^{[q-2]}\varphi\big)\big|\, dt \\
&= \int_0^T \big|\big(k^{q-1}C_f^{q-1}\big(\pi_{\mathrm{cG}}^{[q]}u - u\big), k^{-(q-1)}C_f^{-(q-1)}\big(\varphi - \pi_{\mathrm{cG}}^{[q-2]}\varphi\big)\big)\big|\, dt \\
&\leq C\big\|k^{2q}\bar{\bar{u}}^{(2q)}\big\|_{L_\infty([0,T],l_1)} \int_0^T C_f\|\Phi\|_{L_\infty(\mathcal{T},l_\infty)}\, dt \\
&= C\bar{S}(T)\big\|k^{2q}\bar{\bar{u}}^{(2q)}\big\|_{L_\infty([0,T],l_1)},
\end{aligned}$$

where $\bar{S}(T) = \int_0^T C_f\|\Phi\|_{L_\infty(\mathcal{T},l_\infty)}\, dt = \sum_{n=1}^M K_n C_f \|\Phi\|_{L_\infty(\mathcal{T}_n,l_\infty)}$. □

Similarly, we obtain the following a priori error estimate for the mdG(q) method.

THEOREM 6.6 (a priori error estimate for mdG(q)). *Let* U *be the* mdG(q) *solution of* (1.1) *and let* Φ *be the corresponding* mdG(q)* *solution of the dual problem* (2.9). *Then there is a constant* $C = C(q) > 0$ *such that*

$$|L_{\psi,g}(\bar{e})| \leq CS(T)\big\|k^{q+1}\bar{u}^{(q+1)}\big\|_{L_\infty([0,T],l_2)}, \tag{6.4}$$

where $(k^{q+1}\bar{u}^{(q+1)})_i(t) = k_{ij}^{q_{ij}+1}\|u_i^{(q_{ij}+1)}\|_{L_\infty(I_{ij})}$ *for* $t \in I_{ij}$, *and where the stability factor* $S(T)$ *is given by* $S(T) = \int_0^T \|J^\top(\pi_{\mathrm{dG}}^{[q]}u,u,\cdot)\Phi\|_{l_2}\, dt$. *Furthermore, if assumptions* (A1)–(A5) *hold, then there is a constant* $C = C(q, c_k, \alpha) > 0$ *such that*

$$|L_{\psi,g}(\bar{e})| \leq C\bar{S}(T)\big\|k^{2q+1}\bar{\bar{u}}^{(2q+1)}\big\|_{L_\infty([0,T],l_1)}, \tag{6.5}$$

where $(k^{2q+1}\bar{\bar{u}}^{(2q+1)})_i(t) = k_{ij}^{2q_{ij}+1}C_f^{q_{ij}}\|u_i^{(q_{ij}+1)}\|_{L_\infty(I_{ij})}$ *for* $t \in I_{ij}$, *and where the stability factor* $\bar{S}(T)$ *is given by*

$$\bar{S}(T) = \int_0^T C_f \|\Phi\|_{L_\infty(\mathcal{T},l_\infty)}\, dt = \sum_{n=1}^M K_n C_f \|\Phi\|_{L_\infty(\mathcal{T}_n,l_\infty)}.$$

Using the stability estimate proved in section 4, we obtain the following bound for the stability factor $\bar{S}(T)$.

LEMMA 6.7. *Assume that $K_n C_q C_f \leq 1$ for all time slabs $\mathcal{T}_n$, with $C_q > 0$ the constant in Theorem* 4.2*, and take $g = 0$ in* (2.9)*. Then*

$$\bar{S}(T) \leq \|\psi\|_{l_\infty} e^{C_q \bar{C}_f T}, \tag{6.6}$$

where $\bar{C}_f = \max_{[0,T]} C_f$.

Proof. By Theorem 4.2, we obtain

$$\|\Phi\|_{L_\infty(\mathcal{T}_n, l_\infty)} \leq C_q \|\psi\|_{l_\infty} \exp\left(\sum_{m=n+1}^{M} K_m C_q C_f \right) \leq C_q \|\psi\|_{l_\infty} e^{C_q \bar{C}_f (T - T_n)},$$

and so

$$\begin{aligned} \bar{S}(T) = \sum_{n=1}^{M} K_n C_f \|\Phi\|_{L_\infty(\mathcal{T}_n, l_\infty)}\, dt &\leq \|\psi\|_{l_\infty} \sum_{n=1}^{M} K_n C_q \bar{C}_f e^{C_q \bar{C}_f (T - T_n)} \\ &\leq \|\psi\|_{l_\infty} \int_0^T C_q \bar{C}_f e^{C_q \bar{C}_f t}\, dt \leq \|\psi\|_{l_\infty} e^{C_q \bar{C}_f T}. \qquad \square \end{aligned}$$

Finally, we rewrite the estimates of Theorems 6.5 and 6.6 for special choices of data ψ and g. We first take $\psi = 0$. With $g_n = 0$ for $n \neq i$, $g_i(t) = 0$ for $t \notin I_{ij}$, and

$$g_i(t) = \operatorname{sgn}(\bar{e}_i(t))/k_{ij}, \quad t \in I_{ij},$$

we obtain $L_{\psi,g}(\bar{e}) = \frac{1}{k_{ij}} \int_{I_{ij}} |\bar{e}_i(t)|\, dt$ and so $\|\bar{e}_i\|_{L_\infty(I_{ij})} \leq C L_{\psi,g}(\bar{e})$ by an inverse estimate. By definition, it follows that $\|e_i\|_{L_\infty(I_{ij})} \leq C L_{\psi,g}(\bar{e}) + C k_{ij}^{q_{ij}+1} \|u_i^{q_{ij}+1}\|_{L_\infty(I_{ij})}$. Note that for this choice of g, we have $\|g\|_{L_1([0,T], l_2)} = \|g\|_{L_1([0,T], l_\infty)} = 1$.

We also make the choice $g = 0$. Noting that $\bar{e}(T) = e(T)$, since $\pi u(T) = u(T)$, we obtain

$$L_{\psi,g}(\bar{e}) = (e(T), \psi) = |e_i(T)|$$

for $\psi_i = \operatorname{sgn}(e_i(T))$ and $\psi_n = 0$ for $n \neq i$, and

$$L_{\psi,g}(\bar{e}) = (e(T), \psi) = \|e(T)\|_{l_2}$$

for $\psi = e(T)/\|e(T)\|_{l_2}$. Note that for both choices of ψ, we have $\|\psi\|_{l_\infty} \leq 1$.

With these choices of data, we obtain the following versions of the a priori error estimates.

COROLLARY 6.8 (a priori error estimate for mcG(q)). *Let U be the* mcG(q) *solution of* (1.1)*. Then there is a constant $C = C(q) > 0$ such that*

$$\|e\|_{L_\infty([0,T], l_\infty)} \leq C S(T) \left\| k^{q+1} \bar{u}^{(q+1)} \right\|_{L_\infty([0,T], l_2)}, \tag{6.7}$$

where the stability factor $S(T) = \int_0^T \|J^\top(\pi_{\mathrm{cG}}^{[q]} u, u, \cdot)\Phi\|_{l_2}\, dt$ is taken as the maximum over $\psi = 0$ and $\|g\|_{L_1([0,T], l_\infty)} = 1$. Furthermore, if assumptions (A1)–(A5) *and the assumptions of Lemma* 6.7 *hold, then there is a constant $C = C(q, c_k, \alpha)$ such that*

$$\|e(T)\|_{l_p} \leq C \bar{S}(T) \left\| k^{2q} \bar{u}^{(2q)} \right\|_{L_\infty([0,T], l_1)} \tag{6.8}$$

for $p = 2, \infty$, where the stability factor $\bar{S}(T)$ is given by $\bar{S}(T) = e^{C_q \bar{C}_f T}$.

Corollary 6.9 (a priori error estimate for mdG(q)). *Let U be the* mdG(q) *solution of* (1.1). *Then there is a constant $C = C(q) > 0$ such that*

$$\|e\|_{L_\infty([0,T],l_\infty)} \leq CS(T) \big\| k^{q+1} \bar{u}^{(q+1)} \big\|_{L_\infty([0,T],l_2)}, \tag{6.9}$$

where the stability factor $S(T) = \int_0^T \|J^\top(\pi_{\mathrm{dG}}^{[q]} u, u, \cdot)\Phi\|_{l_2}\, dt$ is taken as the maximum over $\psi = 0$ and $\|g\|_{L_1([0,T],l_\infty)} = 1$. Furthermore, if assumptions (A1)–(A5) *and the assumptions of Lemma* 6.7 *hold, then there is a constant $C = C(q, c_k, \alpha)$ such that*

$$\|e(T)\|_{l_p} \leq C\bar{S}(T) \big\| k^{2q+1} \bar{\bar{u}}^{(2q+1)} \big\|_{L_\infty([0,T],l_1)} \tag{6.10}$$

for $p = 2, \infty$, where the stability factor $\bar{S}(T)$ is given by $\bar{S}(T) = e^{C_q \bar{C}_f T}$.

The stability factor $S(T)$ that appears in the a priori error estimates is obtained from the discrete solution Φ of the dual problem (4.1), and can thus be computed by solving the discrete dual problem. Numerical computation of the stability factor reveals the exact nature of the problem, in particular, whether or not the problem is parabolic; if the stability factor is of unit size and does not grow, then the problem is parabolic by definition; see [36].

6.3. A note on quadrature errors.

The error representations presented in section 6.1 are based on the Galerkin orthogonalities of the mcG(q) and mdG(q) methods. In particular, for the mcG(q) method, we assume that

$$\int_0^T (R(U, \cdot), \Phi)\, dt = 0.$$

In the presence of quadrature errors, this term is nonzero. As a result, we obtain an additional term of the form

$$\int_0^T (\tilde{f}(U, \cdot) - f(U, \cdot), \Phi)\, dt,$$

where $\tilde{f}$ is the interpolant of f corresponding the quadrature rule that is used. A convenient choice of quadrature for the mcG(q) method is Lobatto quadrature with $q+1$ nodal points [32], which means that the quadrature error is of order $2(q+1)-2 = 2q$ and so (super)convergence of order $2q$ is obtained also in the presence of quadrature errors. Similarly for the mdG(q) method, we use Radau quadrature with $q+1$ nodal points, which means that the quadrature error is of order $2(q+1) - 1 = 2q+1$, and so the $2q+1$ convergence order of mdG(q) is also maintained under quadrature.

7. A numerical example.

We conclude by demonstrating the convergence of the multiadaptive methods in the case of a simple test problem.

Consider the problem

$$\begin{aligned} \dot{u}_1 &= u_2, \\ \dot{u}_2 &= -u_1, \\ \dot{u}_3 &= -u_2 + 2u_4, \\ \dot{u}_4 &= u_1 - 2u_3, \\ \dot{u}_5 &= -u_2 - 2u_4 + 4u_6, \\ \dot{u}_6 &= u_1 + 2u_3 - 4u_5 \end{aligned} \tag{7.1}$$

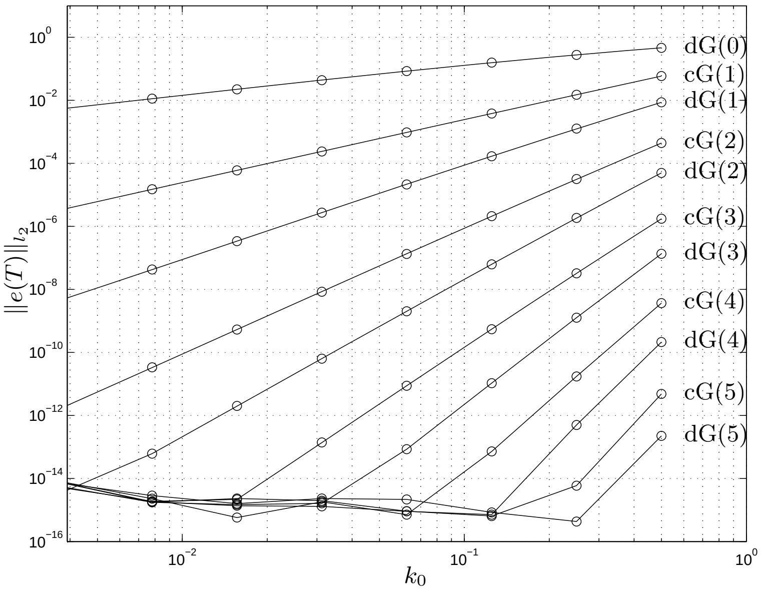


FIG. 6. *Convergence of the error at final time for the solution of the test problem* (7.1) *with* mcG(q) *and* mdG(q), $q \leq 5$.

TABLE 1
Order of convergence p for mcG(q).

mcG(q)	1	2	3	4	5
p	1.99	3.96	5.92	7.82	9.67
$2q$	2	4	6	8	10

TABLE 2
Order of convergence p for mdG(q).

mdG(q)	0	1	2	3	4	5
p	0.92	2.96	4.94	6.87	9.10	–
$2q+1$	1	3	5	7	9	11

on $[0,1]$ with initial condition $u(0) = (0,1,0,2,0,3)$. The solution is given by $u(t) = (\sin t, \cos t, \sin t + \sin 2t, \cos t + \cos 2t, \sin t + \sin 2t + \sin 4t, \cos t + \cos 2t + \cos 4t)$. For given $k_0 > 0$, we take $k_i(t) = k_0$ for $i = 1,2$, $k_i(t) = k_0/2$ for $i = 3,4$, and $k_i(t) = k_0/4$ for $i = 5,6$, and study the convergence of the error $\|e(T)\|_{l_2}$ with decreasing k_0. From the results presented in Figure 6 and Tables 1 and 2, it is clear that the predicted order of convergence is obtained.

REFERENCES

[1] S. G. ALEXANDER AND C. B. AGNOR, *n-body simulations of late stage planetary formation with a simple fragmentation model*, ICARUS, 132 (1998), pp. 113–124.

[2] U. M. ASCHER AND L. R. PETZOLD, *Computer Methods for Ordinary Differential Equations and Differential-Algebraic Equations*, SIAM, Philadelphia, 1998.

[3] R. BECKER AND R. RANNACHER, *An optimal control approach to a posteriori error estimation in finite element methods*, Acta Numer., 10 (2001), pp. 1–102.

[4] J. C. BUTCHER, *The Numerical Analysis of Ordinary Differential Equations—Runge–Kutta and General Linear Methods*, Wiley, New York, 1987.

[5] R. Davé, J. Dubinski, and L. Hernquist, *Parallel treeSPH*, New Astron., 2 (1997), pp. 277–297.

[6] C. Dawson and R. C. Kirby, *High resolution schemes for conservation laws with locally varying time steps*, SIAM J. Sci. Comput., 22 (2001), pp. 2256–2281.

[7] M. Delfour, W. Hager, and F. Trochu, *Discontinuous Galerkin methods for ordinary differential equations*, Math. Comp., 36 (1981), pp. 455–473.

[8] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numer., 4 (1995), pp. 105–158.

[9] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, London, 1996.

[10] K. Eriksson and C. Johnson, *Adaptive Finite Element Methods for Parabolic Problems* III*: Time Steps Variable in Space*, in preparation.

[11] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* I: *A linear model problem*, SIAM J. Numer. Anal., 28 (1991), pp. 43–77.

[12] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* II: *Optimal order error estimates in $l_\infty l_2$ and $l_\infty l_\infty$*, SIAM J. Numer. Anal., 32 (1995), pp. 706–740.

[13] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* IV: *Nonlinear problems*, SIAM J. Numer. Anal., 32 (1995), pp. 1729–1749.

[14] K. Eriksson and C. Johnson, *Adaptive finite element methods for parabolic problems* V: *Long-time integration*, SIAM J. Numer. Anal., 32 (1995), pp. 1750–1763.

[15] K. Eriksson, C. Johnson, and S. Larsson, *Adaptive finite element methods for parabolic problems* VI: *Analytic semigroups*, SIAM J. Numer. Anal., 35 (1998), pp. 1315–1325.

[16] K. Eriksson, C. Johnson, and V. Thomée, *Time discretization of parabolic problems by the discontinuous Galerkin method*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 611–643.

[17] D. Estep, *A posteriori error bounds and global error control for approximations of ordinary differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1–48.

[18] D. Estep and D. French, *Global error control for the continuous Galerkin finite element method for ordinary differential equations*, M2AN Math. Model. Numer. Anal., 28 (1994), pp. 815–852.

[19] D. Estep, M. Larson, and R. Williams, *Estimating the error of numerical solutions of systems of nonlinear reaction–diffusion equations*, Mem. Amer. Math. Soc., 696 (2000), pp. 1–109.

[20] D. Estep and A. Stuart, *The dynamical behavior of the discontinuous Galerkin method and related difference schemes*, Math. Comp., 71 (2002), pp. 1075–1103.

[21] D. Estep and R. Williams, *Accurate parallel integration of large sparse systems of differential equations*, Math. Models Methods Appl. Sci., 6 (1996), pp. 535–568.

[22] J. E. Flaherty, R. M. Loy, M. S. Shephard, B. K. Szymanski, J. D. Teresco, and L. H. Ziantz, *Adaptive local refinement with octree load balancing for the parallel solution of three-dimensional conservation laws*, J. Parallel Distrib. Comput., 47 (1997), pp. 139–152.

[23] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* I*—Nonstiff Problems*, Springer Ser. Comput. Math. 8, Springer, New York, 1991.

[24] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II*—Stiff and Differential-Algebraic Problems*, Springer Ser. Comput. Math. 14, Springer, New York, 1991.

[25] T. J. R. Hughes, I. Levit, and J. Winget, *Element-by-element implicit algorithms for heat-conduction*, J. Engrg. Mech.-ASCE, 109 (1983), pp. 576–585.

[26] T. J. R. Hughes, I. Levit, and J. Winget, *An element-by-element solution algorithm for problems of structural and solid mechanics*, Comput. Methods Appl. Mech. Engrg., 36 (1983), pp. 241–254.

[27] B. L. Hulme, *Discrete Galerkin and related one-step methods for ordinary differential equations*, Math. Comp., 26 (1972), pp. 881–891.

[28] B. L. Hulme, *One-step piecewise polynomial Galerkin methods for initial value problems*, Math. Comp., 26 (1972), pp. 415–426.

[29] P. Jamet, *Galerkin-type approximations which are discontinuous in time for parabolic equations in a variable domain*, SIAM J. Numer. Anal., 15 (1978), pp. 912–928.

[30] C. Johnson, *Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations*, SIAM J. Numer. Anal., 25 (1988), pp. 908–926.

[31] A. Lew, J. E. Marsden, M. Ortiz, and M. West, *Asynchronous variational integrators*, Arch. Ration. Mech. Anal., 167 (2003), pp. 85–146.

[32] A. Logg, *Multi-adaptive Galerkin methods for ODEs* I, SIAM J. Sci. Comput., 24 (2003), pp. 1879–1902.

[33] A. Logg, *Multi-adaptive Galerkin methods for ODEs* II: *Implementation and applications*, SIAM J. Sci. Comput., 25 (2003), pp. 1119–1141.

[34] A. Logg, *Automation of Computational Mathematical Modeling*, Ph.D. thesis, Chalmers University of Technology, Sweden, 2004.

[35] A. Logg, *Interpolation Estimates for Piecewise Smooth Functions in One Dimension*, Technical report 2004–02, Chalmers Finite Element Center Preprint Series, 2004.

[36] A. Logg, *Multi-adaptive time-integration*, Appl. Numer. Math., 48 (2004), pp. 339–354.

[37] J. Makino and S. Aarseth, *On a Hermite integrator with Ahmad–Cohen scheme for gravitational many-body problems*, Publ. Astron. Soc. Japan, 44 (1992), pp. 141–151.

[38] P. Niamsup and V. N. Phat, *Asymptotic stability of nonlinear control systems described by difference equations with multiple delays*, Electron. J. Differential Equations, 11 (2000), pp. 1–17.

[39] S. Osher and R. Sanders, *Numerical approximations to nonlinear conservation laws with locally varying time and space grids*, Math. Comp., 41 (1983), pp. 321–336.

[40] M. J. D. Powell, *Approximation Theory and Methods*, Cambridge University Press, Cambridge, UK, 1988.

[41] L. Shampine, *Numerical Solution of Ordinary Differential Equations*, Chapman & Hall, London, 1994.

SIAM J. NUMER. ANAL.
Vol. 43, No. 6, pp. 2647–2667

NEW A PRIORI FEM ERROR ESTIMATES FOR EIGENVALUES[*]

ANDREW V. KNYAZEV[†] AND JOHN E. OSBORN[‡]

Abstract. We analyze the Ritz–Galerkin method for symmetric eigenvalue problems and prove a priori eigenvalue error estimates. For a simple eigenvalue, we prove an error estimate that depends mainly on the approximability of the corresponding eigenfunction and provide explicit values for all constants. For a multiple eigenvalue we prove, in addition, what is apparently the first truly a priori error estimates that show the levels of the eigenvalue errors depending on approximability of eigenfunctions in the corresponding eigenspace. These estimates reflect a known phenomenon that different eigenfunctions in the corresponding eigenspace may have different approximabilities, thus resulting in different levels of errors for the approximate eigenvalues. For clustered eigenvalues, we derive eigenvalue error bounds that do not depend on the width of the cluster. Our results are readily applicable to the classical Ritz method for compact symmetric integral operators and to finite element method eigenvalue approximation for symmetric positive definite differential operators.

Key words. eigenvalue problem, operator, invariant subspace, multiple eigenvalues, clustered eigenvalues, approximation, Ritz method, Ritz value, finite element method, a priori error estimates, angles between subspaces

AMS subject classification. 65F35

DOI. 10.1137/040613044

1. Introduction.

We revisit the classical subject of a priori eigenvalue error estimates for the Ritz–Galerkin approximation of symmetric eigenvalue problems, with application to finite element method (FEM) eigenvalue approximation. A priori estimates have traditionally been used to prove the convergence of FEM eigenvalue approximation and to determine the convergence rate when the mesh is refined. These estimates are typically based on the approximability of the eigenfunctions by the FEM subspace and can be used to explain certain interesting features of eigenvalue approximation. For example, see [1, 2, 3, 4] for explanations of why the third vibration mode of an L-shaped membrane is easier to approximate than the first two, and why two Ritz values approximating a double eigenvalue may converge at different rates.

The main result of the present paper—briefly stated—is that the eigenvalue errors depend mainly on just the approximability of the corresponding invariant subspaces, whether the eigenvalues are well separated, multiple, or clustered. Our results differ from those in [2, 3, 4] in particular in that the information required by the new estimates is minimal and is covered by *explicitly given constants* that can be relatively easily obtained a posteriori from approximate eigenvalues and eigenfunctions. The question of whether our theorems give completely *computable eigenvalue bounds* thus is reduced to explicitly estimating the main factor, namely the approximability of invariant subspaces.

[*] Received by the editors August 6, 2004; accepted for publication (in revised form) July 25, 2005; published electronically January 27, 2006. Preliminary results of this paper were presented at the meeting "State of the Art in the Finite Element Method" at the City University of Hong Kong in 1998. An extended version of the paper was published as a technical report by the Center for Computational Mathematics of the University of Colorado at Denver [15].

http://www.siam.org/journals/sinum/43-6/61304.html

[†] Department of Mathematics, University of Colorado at Denver, Denver, CO 80217 (knyazev@na-net.ornl.gov, http://www-math.cudenver.edu/~aknyazev). This author was partially supported by National Science Foundation award DMS 0208773.

[‡] Department of Mathematics, University of Maryland, College Park, MD 20742 (jeo@math.umd.edu, http://www.wam.umd.edu/~jeo). This author was partially supported by National Science Foundation award DMS 0341982.

Computing the approximability is, however, difficult except in fairly trivial situations. One traditional approach is to use approximation theory results based on the smoothness of the eigenfunctions. Many results of this type are known, but usually the constants in the estimates are generic and not easily computable in practice. Assessing the smoothness of the eigenfunctions, meaning obtaining an estimate for an appropriate higher Sobolev norm of the eigenfunction in question, can be done using an appropriate regularity theory for the underlying partial differential equation. Much is known about the regularity of the eigenfunctions, but, again, the constants are typically generic and cannot be easily estimated, with rather trivial exceptions. Nevertheless, a priori eigenvalue error analysis is a classical approach that has proved to be useful.

Early examples of a priori eigenvalue error estimates can be found, e.g., in [17]. Later, it became clear that the eigenvalue error was governed by the approximability of the exact eigenfunctions by the approximation space. In [5], Birkhoff et al. showed that the error for the jth eigenvalue was bounded by a constant times the sum of the norms squared of the approximation errors of the all eigenfunctions corresponding to the first j eigenvalues. In [22], Weinberger improved this result, showing that in the estimate for the relative eigenvalue error the constant simply equals one; see Remark 2.1 for the exact formulation. Knyazev in [11] (see also [8]) further improved this result by replacing the norms of the approximation errors of individual eigenfunctions with the angle that measures the approximability of the invariant subspace spanned by these eigenfunctions. We reproduce this latter result by Knyazev in the present paper, in Theorem 2.4, and show that it is sharp.

The estimates of [5, 22, 11] suggest that the jth eigenvalue error depends on the approximability of all the eigenfunctions in the corresponding eigenspace, as well as of all the eigenfunctions corresponding to the previous eigenvalues. In reality, this is not the case. Numerical experiments for the L-shaped membrane eigenvalue problem show that the accuracy of approximation for the third eigenvalue is significantly better than for the first two. This can be explained as follows. We first note that the first two eigenfunctions of the L-shaped membrane eigenvalue problem are singular because of the re-entrant corner, but the third eigenfunction is analytic because of symmetry, and hence easily approximated, especially by the p-method (see [1]). Second, Vainikko in Krasnosel'skii et al. [16] and Chatelin [7] derive estimates of the eigenvalue error mainly in terms of just the approximability of the eigenfunctions in the corresponding eigenspace. Coupling this approximability result with this eigenvalue error estimate, we obtain the accurate eigenvalue approximation for the third eigenvalue.

Moreover, Vainikko in Krasnosel'skii et al. [16] and Chatelin [7] show that the multiplicative constant in the estimate of the relative eigenvalue error approaches 1 under the approximability assumption on the family of the approximating spaces; see section 3.2 for details. In [3], Babuška and Osborn determine that the closeness of the constant to 1 depends on the approximability of the operator of the original problem by the Ritz method; again, see section 3.2.

Our first main results—Theorems 2.7 and 3.2—clarify the estimate of [3] and improve the constant. All our constants are explicitly given, and no asymptotic assumptions are made. In the FEM context, our results are readily applicable for a fixed mesh without making the traditional assumption, cf. Strang and Fix [2], that the mesh size is small enough.

When the eigenvalue of interest is of multiplicity $q > 1$, different eigenfunctions in the corresponding eigenspace may have different approximabilities, thus resulting in different levels of error for the approximate eigenvalues; i.e., the q Ritz values, corresponding to the multiple eigenvalue, may approach the eigenvalue with different rates. It is important to have eigenvalue error estimates that capture this phenomenon.

The error bounds of Krasnosel'skii et al. [16] and Chatelin [7] effectively require approximability of all eigenfunctions in the corresponding eigenspace that provides an estimate for the largest eigenvalue error only. In [2, 3, 4], Babuška and Osborn perform analysis that differentiates levels of eigenvalue error depending on approximability of different eigenfunctions in the eigenspace, but their estimates are not truly a priori, except for the estimate for the smallest eigenvalue error, which depends mainly on the approximability of the most easily approximated eigenfunction within the eigenspace.

Our results for multiple eigenvalues—Theorems 2.11 and 3.3—clarify and improve these results of [2, 3, 4]. For example, if the eigenspace is spanned by three eigenfunctions of different approximation qualities, our results estimate the corresponding quality of each of the three Ritz values.

Error estimates for clustered eigenvalues are not well examined in the literature. The results presented in this paper are valid for clustered eigenvalues, as well as for multiple eigenvalues, and give error estimates that do not depend on the width of the cluster. Ovtchinnikov in [19], independently derives similar estimates, which he calls "cluster robust." Our estimates, compared to those of [19], are more compact and use less information.

In our proofs, we heavily use approximation error estimates for eigenspaces and invariant subspaces obtained by Knyazev in [13].

The paper intentionally contains some material that may be considered redundant, in order to improve readability. A critic once wrote about Beethoven's Symphony No. 2 in D major, op. 36 that it "would surely benefit from the abbreviation of some passages and the deletion of others." If we are allowed to use musical terms in our defense and to compare our paper to a symphony, it consists of four movements:

The first, fast, movement is subsections 2.1–2.5. Subsection 2.1 sets the stage for an abstract setting of a compact symmetric operator on a Hilbert space. We briefly introduce the angles instruments in the developmental subsection 2.2 and then, in subsection 2.3, the main theme, a priori estimates for eigenvalues. Subsection 2.4 is the most important in the first movement—it brings us the main theme in its most "ideal" form in Theorem 2.7, without a proof. Theorem 2.7 is an error estimate for a jth eigenvalue mainly in terms of the approximation error of the corresponding eigenfunctions. In subsection 2.5, the theme appears with slight variations for multiple and clustered eigenvalues. It becomes apparent that a major development is needed.

The second, slow, movement is the massive subsection 2.6. The main theme is significantly extended and generalized, with a complete vigorous proof, to carry a considerable improvement, in Theorem 2.11, for multiple and clustered eigenvalues. A number of possible variations surface at the end of the second movement.

The third, dance-like, movement is subsection 3.1, which is a brief reminiscence of the first two movements. The same theme is essentially repeated, but in a different key, for the variational Galerkin method in a context applicable to FEM eigenvalue approximation for second order symmetric positive definite differential operators. Our last main results—Theorem 3.2 and Theorem 3.3—appear in this subsection.

The fast finale, subsection 3.2, takes the material of the previous movement and contrasts it from earlier work. It opens in a relaxed manner and cites several well-known results. In closing, it reaches a climax by showing how to obtain differential levels of eigenvalue error depending on approximability of different eigenfunctions in the eigenspace.

It would, of course, be appealing to have practical numerical examples of our a priori analysis providing computable eigenvalue bounds, e.g., for the Laplacian in a polygonal domain. It is known that the eigenfunctions of the Laplacian are smooth (analytic, in fact) inside the domain, but are generally singular in the corners. In certain cases, however, the eigenfunction is smooth, e.g., the already discussed third eigenfunction of the Laplacian in the L-shaped domain is smooth. It is very interesting to try to use this kind of information to compute eigenvalue bounds. But due to the difficulties of computing the approximability of invariant subspaces, discussed in the first paragraphs of the introduction, such a project lies beyond the scope of this paper. For some computational examples we refer to [3, 4].

2. Estimates for a compact symmetric operator.

2.1. An abstract eigenvalue problem. We consider in this section a compact symmetric positive definite operator T defined on a real separable Hilbert space H, with inner product (u,v) and norm $\|u\| = \sqrt{(u,u)}$. The spectral theory of such operators is well known; see, e.g., [9]. The spectrum consists of nonzero eigenvalues of finite multiplicity, together with 0, which is in the continuous spectrum. The eigenvectors can be chosen to be orthonormal. We denote the eigenvalues and corresponding eigenvectors of T by $\mu_1 \geq \mu_2 \geq \cdots > 0$ and $u_1, u_2, \ldots,$ where $(u_i, u_j) = \delta_{ij}$. We are interested in approximating the eigenpairs (μ_i, u_i) of T by the Ritz method. Given a finite-dimensional subspace $\tilde{U}$ of H, referred to as the trial subspace, the Ritz approximation to T is the operator $\tilde{T} = (\tilde{Q}T)|_{\tilde{U}}$, where $\tilde{Q}$ is the orthogonal projector onto $\tilde{U}$. The operator $\tilde{T}$ is symmetric positive definite. The eigenpairs of $\tilde{T}$ are called the Ritz pairs of T; we regard them as approximations of the eigenpairs of T. We denote the eigenvalues and corresponding eigenvectors of $\tilde{T}$ by $\tilde{\mu}_1 \geq \tilde{\mu}_2 \geq \cdots \geq \tilde{\mu}_n > 0$, where $n = \dim \tilde{U}$, and $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_n$, where $(\tilde{u}_i, \tilde{u}_j) = \delta_{ij}$. The numbers $\tilde{\mu}_i$ are called the Ritz values and the vectors $\tilde{u}_i$ are called the Ritz vectors. In this paper we are specifically concerned with approximating the eigenvalues of T by Ritz values as $\mu_i \approx \tilde{\mu}_i$. It is an immediate consequence of the max-min characterization of eigenvalues that $\tilde{\mu}_i \leq \mu_i, i = 1, \ldots, n$.

We make the assumptions that operator T is positive definite and compact just to simplify the arguments. The majority of our results in this section can be easily modified to hold without these assumptions. The most important modification is a replacement of the ratios such as $(\mu_j - \tilde{\mu}_j)/\mu_j$ that appear in the left-hand sides of most of our estimates below with $(\mu_j - \tilde{\mu}_j)/(\mu_j - \mu_{inf})$, where μ_{inf} is the algebraically smallest point of the spectrum of T (when T is positive definite, evidently $\mu_{inf} = 0$). Such a modification makes our estimates invariant with respect to a scalar shift $T - \alpha I$ in T for any real scalar α.

2.2. Principal angles between subspaces. If M and N are nontrivial finite-dimensional subspaces of H, we will quantify the approximability of M by N using the sine of the largest principal angle from M to N, which is defined by

$$\sin \angle\{M; N\} = \sup_{u \in M, \|u\|=1} \text{dist}\,(u, N) = \sup_{u \in M, \|u\|=1} \inf_{v \in N} \|u - v\|. \tag{2.1}$$

For nonzero vectors u and v, if $M = \text{span}\{u\}$, we write $\sin \angle\{u; N\}$ for $\sin \angle\{M; N\}$; and if $M = \text{span}\{u\}$ and $N = \text{span}\{v\}$, we write $\sin \angle\{u; v\}$ for $\sin \angle\{M; N\}$.

It is immediate that $0 \leq \sin \angle\{M; N\} \leq 1$ and that $\sin \angle\{M; N\} = 0$ if and only if $M \subseteq N$. If $\dim M > \dim N$, then $\sin \angle\{M; N\} = 1$. If $\dim M = \dim N < \infty$, then $\sin \angle\{M; N\} = \sin \angle\{N; M\}$. In the remainder of this paper, we will typically have $\dim M \leq \dim N$.

We will need the following simple observations; cf. Lemma 3.4 of [6].

LEMMA 2.1. *Let the subspace M be split into an orthogonal sum of subspaces $M = M_1 \oplus M_2$. Then (see* [15]*),*

$$\sin^2 \angle\{M; N\} \leq \sin^2 \angle\{M_1; N\} + \sin^2 \angle\{M_2; N\}. \tag{2.2}$$

Applying (2.2) recursively, we immediately obtain the following.

COROLLARY 2.2. *Let vectors $\{u_i, i = 1, \ldots, \dim M\}$ form an orthogonal basis for the subspace M. Then*

$$\sin^2 \angle\{M; N\} \leq \sum_i \sin^2 \angle\{u_i; N\}. \tag{2.3}$$

We call angle $\angle\{M; N\}$ the largest since it is also well known (see, e.g., [14]), so that smaller angles between subspaces can be defined as follows. Using P and Q, the orthogonal projectors onto M and N, respectively, the sine of the largest angle equals the largest singular value of the operator $(I - Q)P$. Introducing the notation $s_1((I-Q)P) \geq s_2((I-Q)P) \geq \cdots \geq s_{\dim M}((I-Q)P)$ for the $\dim M$ *largest* singular values of the operator $(I-Q)P$, we define the ith angle from subspace M to subspace N using its sine: $\sin \angle_i\{M; N\} = s_{\dim M-i+1}((I - Q)P)$, $i = 1, \ldots, \dim M$, assuming that all angles lie on the closed interval $[0, \pi/2]$. The complete set of $\dim M$ angles from subspace M to subspace N gives detailed information on approximability of M by N; e.g., if the smallest angle vanishes, the subspaces M and N have a nontrivial intersection.

Later in the paper we use the following property of angles (see [14]):

$$\angle_j\{M; N\} = \inf_{L \subseteq M,\, \dim L = j} \angle\{L; N\},\ j = 1, \ldots, \dim M. \tag{2.4}$$

Finally, we will also need the following generalization of Corollary 2.2.

LEMMA 2.3. *Let vectors $\{u_i, i = 1, \ldots, \dim M\}$ form an orthogonal basis for the subspace M and be arranged in such a way that*

$$\angle\{u_1; N\} \leq \cdots \leq \angle\{u_{\dim M}; N\}.$$

Then

$$\sin^2 \angle_j\{M; N\} \leq \sum_{i=1}^{j} \sin^2 \angle\{u_i; N\}, \qquad j = 1, \ldots, \dim M. \tag{2.5}$$

Proof. We deduce from (2.4) that

$$\sin^2 \angle_j\{M; N\} \leq \sin^2 \angle\{\operatorname{span}\{u_1, \ldots, u_j\}; N\}.$$

Now, the statement of the lemma, (2.5), immediately follows from (2.3) applied to $M = \operatorname{span}\{u_1, \ldots, u_j\}$. □

2.3. Estimates based on the approximability of all previous eigenvectors. Sharp eigenvalue error estimates are usually derived under the assumption that the eigenvector corresponding to the eigenvalue being estimated is well approximated by the trial subspace.

We derive an estimate for the error in approximating μ_j, the jth eigenvalue of T, by $\tilde{\mu}_j$, the jth Ritz value of T, i.e., the jth eigenvalue of $\tilde{T}$. Let $U_{1,\ldots,j}$ denote the span of the eigenvectors $u_1, \dots, u_j$, and let $P_{1,\dots,j}$ be the orthogonal projector onto $U_{1,\dots,j}$. For $u \neq 0$, let $\mu(u) = (Tu, u)/(u, u) = (u, u)_T/(u, u)$ be the Rayleigh quotient associated with T. Here $(\cdot, \cdot)_T$ is a second inner product on H. We will refer to orthogonality in $(\cdot, \cdot)_T$ as T-orthogonality. Note that $\mu(u) > 0$ since T is positive definite.

Our first theorem is known; it was proved in [11] and reproduced in [8]. For the particular case $j = \dim \tilde{U}$, a different proof was then suggested in [10, 12].

THEOREM 2.4. *For $j = 1, 2, \dots, n = \dim \tilde{U}$ we have*

$$0 \leq \frac{\mu_j - \tilde{\mu}_j}{\mu_j} \leq \sin^2 \angle\{U_{1,\dots,j}; \tilde{U}\} = \|(I - \tilde{Q})P_{1,\dots,j}\|^2. \tag{2.6}$$

The estimate (2.6) is sharp; see [15].

Remark 2.1. By Corollary 2.2 we have

$$\sin^2 \angle\{U_{1,\dots,j}; \tilde{U}\} \leq \sum_{i=1}^{j} \sin^2 \angle\{u_i; \tilde{U}\} = \sum_{i=1}^{j} \|(I - \tilde{Q})u_i\|^2;$$

therefore, the estimate

$$\frac{\mu_j - \tilde{\mu}_j}{\mu_j} \leq \sum_{i=1}^{j} \|(I - \tilde{Q})u_i\|^2 \tag{2.7}$$

follows directly from Theorem 2.4. Estimate (2.7) is well known (see, e.g., [20, 22]); on the right-hand side we have the sum of the squares of the approximation errors for the eigenvectors $u_1, \dots, u_j$. If $j = 1$, the estimates (2.6) and (2.7) are identical.

2.4. Estimates based mainly on the approximability of the target eigenvector. Theorem 2.4 has a major weakness; namely, the right-hand side of estimate (2.6) for the target eigenvalue μ_j involves the approximability of all functions in $U_{1,\dots,j}$. The result thus suggests that the eigenvalue error $(\mu_j - \tilde{\mu}_j)/\mu_j$ depends on the approximation errors for all eigenfunctions $u_1, \dots, u_j$. We now mention two results suggesting that this is not the case; that, in fact, the ratio $(\mu_j - \tilde{\mu}_j)/\mu_j$ depends mainly on just the approximation error for u_j, the target eigenfunction. First, consider the following

LEMMA 2.5. *For $j = 1, 2, \dots, n = \dim \tilde{U}$, the estimate*

$$\begin{aligned} \frac{\mu_j - \tilde{\mu}_j}{\mu_j} &= \|(I - \tilde{P}_j)u_j\|^2 - \frac{1}{\mu_j}((I - P_j)\tilde{u}_j, T(I - P_j)\tilde{u}_j) \\ &\leq \sin^2 \angle\{u_j, \tilde{u}_j\} \end{aligned} \tag{2.8}$$

holds, where $\tilde{P}_j$ is the orthogonal projector onto $\operatorname{span}\{\tilde{u}_j\}$.

The first line of (2.8) follows from the chain of identities in the proof of Lemma 3.5 of [6].

Next consider the following lemma.

LEMMA 2.6. *If $(\tilde{u}_j, u_j) \neq 0$, the estimate*

$$\begin{aligned} \frac{\mu_j - \tilde{\mu}_j}{\mu_j} &= \|(I - \tilde{Q})u_j\|^2 + \frac{1}{\mu_j}\left(T(I - \tilde{Q})u_j, \frac{(I - P_j)\tilde{u}_j}{\|P_j\tilde{u}_j\|}\right) \\ &\leq \left(1 + \frac{\|(I - \tilde{Q})T\|}{\mu_j} \frac{\tan \angle\{u_j, \tilde{u}_j\}}{\sin \angle\{u_j, \tilde{U}\}}\right) \sin^2 \angle\{u_j, \tilde{U}\} \end{aligned} \tag{2.9}$$

holds, where P_j is the orthogonal projector onto $\operatorname{span}\{u_j\}$.

The identity in the first line of (2.9) is based on an argument from the proof of Theorem 4.1 in [3] (see also [18]). For a complete proof, see [15].

It is informative to compare (2.8) with (2.9). The first term on the right-hand side of the first line of (2.8) is larger than that of (2.9). However, the second term in the first line of (2.8) is negative, and thus is dropped in the second line of (2.8). The second term on the right-hand side in the first line of (2.9), while generally not negative, in typical applications (when $\|(I - \tilde{Q})T\|$ is small) is significantly smaller compared to the first term; in other words, the term added to 1 in the second line of (2.9) in such applications is small because of the multiplier $\|(I - \tilde{Q})T\|$. We conclude that both (2.8) and (2.9) suggest that $(\mu_j - \tilde{\mu}_j)/\mu_j$ depends mainly on the approximation error for u_j.

Both estimates (2.8) and (2.9), in addition to being dependent on the eigenfunction u_j, depend explicitly on the approximate eigenfunction $\tilde{u}_j$: (2.8) in the main term and (2.9) in the constant. Our next theorem is based on a novel alternative technique, where the approximate eigenfunction $\tilde{u}_j$ is not used in the proof and does not appear in the theorem statement.

Theorem 2.7. *For a fixed index j such that $1 \le j \le n = \dim \tilde{U}$, suppose that*

$$\min_{i=1,\dots,j-1} |\tilde{\mu}_i - \mu_j| \neq 0. \tag{2.10}$$

Then

$$\begin{aligned} 0 \le \frac{\mu_j - \tilde{\mu}_j}{\mu_j} &\le \|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-1})u_j\|^2 \\ &\le \left(1 + \frac{\|(I - \tilde{Q})T\tilde{P}_{1,\dots,j-1}\|^2}{\min_{i=1,\dots,j-1} |\tilde{\mu}_i - \mu_j|^2}\right) \sin^2 \angle\{u_j; \tilde{U}\}, \end{aligned} \tag{2.11}$$

where $\tilde{P}_{1,\dots,j-1}$ is the orthogonal projector onto $\tilde{U}_{1,\dots,j-1} = \operatorname{span}\{\tilde{u}_1, \dots, \tilde{u}_{j-1}\}$ (if $j = 1$, we define $\tilde{P}_{1,\dots,j-1} = 0$ and do not use (2.10)*).*

For brevity, we do not prove the theorem here, but instead refer to [15], and to our proof of Theorem 2.11 later in the paper, which is a generalization of Theorem 2.7.

Since $\|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-1})u_j\| \le \|(I - \tilde{P}_j)u_j\|$, our new estimate (2.11) clearly improves (2.8). A direct comparison of the constants in (2.9) and (2.11) in a general case does not appear to be simple because of the unresolved dependence of (2.9) on $\tilde{u}_j$. However, we have

$$\begin{aligned} \frac{\|(I - \tilde{Q})T\tilde{P}_{1,\dots,j-1}\|^2}{\min_{i=1,\dots,j-1} |\tilde{\mu}_i - \mu_j|^2} &\le \frac{\|(I - \tilde{Q})T\|^2}{\min_{i=1,\dots,j-1} |\tilde{\mu}_i - \mu_j|^2} \\ &\le \frac{\|(I - \tilde{Q})T\|}{\mu_j}, \end{aligned}$$

assuming

$$\|(I - \tilde{Q})T\| \le \frac{\min_{i=1,\dots,j-1} |\tilde{\mu}_i - \mu_j|^2}{\mu_j}. \tag{2.12}$$

Since $\tan \angle\{u_j, \tilde{u}_j\} \ge \sin \angle\{u_j, \tilde{U}\}$, we can conclude that our estimate (2.11) is sharper than (2.9) under the assumption (2.12). We note that in the FEM context assumption (2.12) is realistic as, for typical problems, $\|(I - \tilde{Q})T\|$ vanishes when the mesh parameter tends to zero.

Let us finally comment that the ratio

$$\frac{\|(I-\tilde{Q})T\tilde{P}_{1,\dots,j-1}\|^2}{\min_{i=1,\dots,j-1}|\tilde{\mu}_i-\mu_j|^2}=\frac{\|(I-\tilde{Q})(T/\mu_j)\tilde{P}_{1,\dots,j-1}\|^2}{\min_{i=1,\dots,j-1}|\tilde{\mu}_i/\mu_j-1|^2}$$

in (2.11) is "dimensionless," i.e., invariant with respect to scaling of T. Here, the quantity in the denominator, $\min_{i=1,\dots,j-1}(\tilde{\mu}_i/\mu_j-1)$, in the limit, where all $\tilde{\mu}_i\to\mu_i$, turns into μ_{j-1}/μ_j-1, the one-sided relative gap in the spectrum at μ_j.

2.5. Corollaries of Theorems 2.4 and 2.7 for multiple eigenvalues. Here we address in details the case when the eigenvalue μ_j is a multiple of multiplicity $q>1$. Our Theorems 2.4 and 2.7 hold for multiple eigenvalues since we never assumed the eigenvalues were simple. However, the case of multiple eigenvalues has special features, which we want to highlight. Let us start with the simplest case, where we are interested only in estimates for the largest eigenvalue μ_1. From Theorem 2.4 we easily derive the following.

Corollary 2.8. *Let*

$$\mu_1=\mu_2=\dots=\mu_q>\mu_{q+1}$$

and $q\le n=\dim\tilde{U}$. *For* $j=1,2,\dots,q$ *we have*

$$\begin{aligned}0\le\frac{\mu_1-\tilde{\mu}_j}{\mu_1}&\le\inf_{\substack{U_{1,\dots,j}\subset U_{1,\dots,q}\\ \dim U_{1,\dots,j}=j}}\sin^2\angle\{U_{1,\dots,j};\tilde{U}\}\\ &=\sin^2\angle_j\{U_{1,\dots,q};\tilde{U}\}.\end{aligned}\tag{2.13}$$

Estimate (2.13) has two important properties. First, it controls the error for *every* Ritz values corresponding to the first eigenvalue μ_1. Second, it shows that different Ritz values may have different approximation qualities, depending on approximability of the eigenspace $U_{1,\dots,q}$ by the trial subspace $\tilde{U}$ of the Ritz method, where the approximability is measured by the angles from $U_{1,\dots,q}$ to $\tilde{U}$ and, thus, can be estimated a priori.

In general, the multiple eigenvalue of interest may not be the largest:

$$\mu_{p-1}>\mu_p=\mu_{p+1}=\dots=\mu_j=\dots=\mu_{p+q-1}>\mu_{p+q}.\tag{2.14}$$

Applying Theorem 2.4, we obtain the following.

Corollary 2.9. *Suppose* (2.14) *is satisfied and* $p+q-1\le n$. *For any index* $j=p,p+1,\dots,p+q-1$ *we have*

$$0\le\frac{\mu_p-\tilde{\mu}_j}{\mu_p}\le\inf_{\substack{U_{1,\dots,p-1}\subset U_{1,\dots,j}\subset U_{1,\dots,p+q-1}\\ \dim U_{1,\dots,j}=j}}\sin^2\angle\{U_{1,\dots,j};\tilde{U}\}.$$

Proof. The subspace $U_{1,\dots,j}$ has a fixed part $U_{1,\dots,p-1}\subset U_{1,\dots,j}$, but the rest of it we can choose within $U_{p,\dots,\min\{p+q-1,n\}}$ as we like. □

Corollary 2.9 preserves the desired properties of Corollary 2.8; i.e., it provides a different estimate for each Ritz value of interest, but it requires approximability of all previous eigenvectors.

Let us now turn our attention to Theorem 2.7. The only relevant assumption in Theorem 2.7 is that (2.10) is satisfied so that the denominator in the constant in

Theorem 2.7 is not zero. Let us analyze the likely behavior of this constant for the particular case $q = 2$ so that

$$\mu_{p-1} > \mu_p = \mu_{p+1} > \mu_{p+2}. \tag{2.15}$$

There are two relevant possibilities for j in Theorem 2.7: $j = p$ or $j = p+1$. Assuming that all Ritz values $\tilde{\mu}_i$ approximate the corresponding eigenvalues μ_i, which is typical for FEM applications (see section 3.2 for details), we observe that in (2.10)

$$\min_{i=1,\ldots,j-1} |\tilde{\mu}_i - \mu_j| \approx \mu_{j-1} - \mu_j.$$

Thus, if $j = p$, the denominator is asymptotically positive; specifically, it is asymptotically equal to $\mu_{p-1} - \mu_p$, and the estimate of Theorem 2.7 is asymptotically valid; while if $j = p + 1$, the denominator in the constant in Theorem 2.7 asymptotically vanishes. This discussion demonstrates that Theorem 2.7 provides an asymptotically valid estimate only for one out of the $q = 2$ Ritz values. On the positive side, however, we can freely choose the eigenvector u_j within the eigenspace corresponding to μ_p to minimize the right-hand side of (2.11). Let us reformulate Theorem 2.7 to reflect these observations.

Corollary 2.10. *Suppose that the eigenvalue μ_p, where $p > 1$, has multiplicity $q > 1$ so that* (2.14) *holds, and that $p + q - 1 \leq n$, and denote the corresponding eigenspace by $U_{p,\ldots,p+q-1}$. As in Theorem* 2.7, *suppose that*

$$\min_{i=1,\ldots,p-1} |\tilde{\mu}_i - \mu_p| \neq 0.$$

Then

$$\begin{aligned} 0 \leq \frac{\mu_p - \tilde{\mu}_p}{\mu_p} &\leq \min_{u \in U_{p,\ldots,p+q-1},\, \|u\|=1} \|(I - \tilde{Q} + \tilde{P}_{1,\ldots,p-1})u\|^2 \\ &\leq \left(1 + \frac{\|(I-\tilde{Q})T\tilde{P}_{1,\ldots,p-1}\|^2}{\min_{i=1,\ldots,p-1} |\tilde{\mu}_i - \mu_p|^2}\right) \min_{u \in U_{p,\ldots,p+q-1},\, \|u\|=1} \sin^2 \angle\{u; \tilde{U}\} \\ &= \left(1 + \frac{\|(I-\tilde{Q})T\tilde{P}_{1,\ldots,p-1}\|^2}{\min_{i=1,\ldots,p-1} |\tilde{\mu}_i - \mu_p|^2}\right) \sin^2 \angle_1\{U_{p,\ldots,p+q-1}; \tilde{U}\}. \end{aligned} \tag{2.16}$$

Proof. We take $j = p$ in Theorem 2.7 and notice that we can choose u_j to be any normalized vector in the eigenspace $U_{p,\ldots,p+q-1}$ and finally use (2.4). □

It is useful to compare Corollary 2.9 with Corollary 2.10. Corollary 2.9 gives different estimates for every Ritz value out of the q Ritz values corresponding to the multiple eigenvalue μ_p, but requires approximability of all previous eigenvectors. In Corollary 2.10, the approximability of previous eigenvectors appears only in the constant, but it gives an estimate only for the largest Ritz value out of the q.

We want to obtain a result that combines the advantages of Corollaries 2.9 and 2.10 and removes their weaknesses. E.g., if $q = 3$ and the eigenspace corresponding to the triple eigenvalue μ_p is spanned by eigenfunctions of different approximation quality, we want to have three error estimates for μ_p reflecting it and not depending strongly on approximability of previous eigenfunctions.

2.6. A new estimate that covers multiple and clustered eigenvalues.

Our new result is a generalization of Theorem 2.7 that gives us the desired estimates for a multiple eigenvalue corresponding to an eigenspace spanned by eigenfunctions of different approximation quality. In addition, the new estimate also covers the case of clustered eigenvalues, i.e., the constant in the new estimate does not depend on the width of the eigenvalue cluster.

THEOREM 2.11.[1] *For fixed indexes j and m satisfying $1 \le j \le n$ and $1 \le m \le j$, let $U_{j-m+1,\dots,j}$ be the m-dimensional invariant subspace corresponding to eigenvalues $\mu_{j-m+1} \ge \cdots \ge \mu_j$ and $P_{j-m+1,\dots,j}$ be the orthogonal projector on $U_{j-m+1,\dots,j}$. If*

$$\min_{i=1,\dots,j-m} |\tilde{\mu}_i - \mu_j| \ne 0, \tag{2.17}$$

then

$$\begin{aligned} 0 \le \frac{\mu_j - \tilde{\mu}_j}{\mu_j} &\le \|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m})P_{j-m+1,\dots,j}\|^2 \\ &\le \left(1 + \frac{\|(I-\tilde{Q})T\tilde{P}_{1,\dots,j-m}\|^2}{\min_{i=1,\dots,j-m}|\tilde{\mu}_i - \mu_j|^2}\right)\|(I-\tilde{Q})P_{j-m+1,\dots,j}\|^2, \end{aligned} \tag{2.18}$$

where $\tilde{P}_{1,\dots,j-m}$ is the orthogonal projector onto $\tilde{U}_{1,\dots,j-m} = \operatorname{span}\{\tilde{u}_1, \dots, \tilde{u}_{j-m}\}$ (if $j = m$, we set $\tilde{P}_{1,\dots,j-m} = 0$ and do not use (2.17)*). If $m = j$, the present theorem turns into Theorem* 2.4*; if $m = 1$, it turns into Theorem* 2.7.

Proof. The operators $I - \tilde{Q} + \tilde{P}_{1,\dots,j-m}$ and $P_{j-m+1,\dots,j}$ are orthogonal projectors; thus, $\|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m})P_{j-m+1,\dots,j}\| \le 1$. If $\|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m})P_{j-m+1,\dots,j}\| = 1$, the first estimate in (2.18) is trivially true. Now we suppose

$$\|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m})P_{j-m+1,\dots,j}\| < 1. \tag{2.19}$$

Then, since $\dim U_{j-m+1,\dots,j} = m$, the subspace $(\tilde{Q} - \tilde{P}_{1,\dots,j-m})U_{j-m+1,\dots,j}$ is also m-dimensional by Theorem 6.34 in Chapter I in [9].

We choose a normalized vector $\bar{u}$ such that

$$\bar{u} \in (\tilde{Q} - \tilde{P}_{1,\dots,j-m})U_{j-m+1,\dots,j}, \qquad \mu(\bar{u}) = \min_{w \in (\tilde{Q} - \tilde{P}_{1,\dots,j-m})U_{j-m+1,\dots,j} \setminus \{0\}} \mu(w),$$

and introduce the orthogonal and T-orthogonal decomposition

$$\bar{u} = u + v, \quad u \in U_{1,\dots,j}, \quad v \in U_{1,\dots,j}^{\perp}.$$

Since $\bar{u} \in (\tilde{Q} - \tilde{P}_{1,\dots,j-m})U_{j-m+1,\dots,j}$, $\|\bar{u}\| = 1$, and $u = \bar{u} - v$ is the orthogonal projection of $\bar{u}$ onto $U_{1,\dots,j}$, we see, using again Theorem 6.34 in Chapter I in [9], that

$$\begin{aligned} \|v\| &= \sin \angle\{\bar{u}; U_{1,\dots,j}\} \\ &\le \sin \angle\{\bar{u}; U_{j-m+1,\dots,j}\} \\ &\le \sin \angle\{(\tilde{Q} - \tilde{P}_{1,\dots,j-m})U_{j-m+1,\dots,j}; U_{j-m+1,\dots,j}\} \\ &= \|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m})P_{j-m+1,\dots,j}\|. \end{aligned} \tag{2.20}$$

It now follows from (2.19) and (2.20) that $\|v\| < 1$; thus $u \ne 0$, and $\mu(u)$ is defined.

We next prove the following chain of inequalities:

$$\mu(\bar{u}) \le \tilde{\mu}_j \le \mu_j \le \mu(u). \tag{2.21}$$

[1] It came to our attention that a similar result was independently obtained in the revised version of [19] to appear in *Linear Algebra and Its Applications.*

Indeed, the first inequality,

$$\begin{aligned}\mu(\bar{u}) &= \min_{w\in(\tilde{Q}-\tilde{P}_{1,\dots,j-m})U_{j-m+1,\dots,j}\setminus\{0\}} \mu(w)\\ &\le \max_{\substack{W\subseteq \operatorname{Im}(\tilde{Q}-\tilde{P}_{1,\dots,j-m})\\ \dim W=m}} \min_{w\in W\setminus\{0\}} \mu(w)=\tilde{\mu}_j,\end{aligned}$$

follows from the min-max principle for Ritz values, since the dimension of the subspace $(\tilde{Q}-\tilde{P}_{1,\dots,j-m})U_{j-m+1,\dots,j}$ is m. The second inequality, $\tilde{\mu}_j \le \mu_j$, is an immediate consequence of the max-min principle. The third inequality, $\mu_j \le \mu(u)$, follows from the fact that $u \in U_{1,\dots,j}$.

The identity

$$\mu(\bar{u}) = \frac{(Tu,u)+(Tv,v)}{(u,u)+(v,v)}$$

can be rewritten as

$$\mu(u)-\mu(\bar{u}) = \begin{cases} [\mu(\bar{u})-\mu(v)]\dfrac{(v,v)}{(u,u)},\ v\neq 0,\\ 0,\ v=0.\end{cases} \tag{2.22}$$

For $v \neq 0$, it follows directly from (2.21) and (2.22) that

$$\begin{aligned}0\le \mu_j-\tilde{\mu}_j &\le \mu(u)-\mu(\bar{u})\\ &= [\mu(\bar{u})-\mu(v)]\frac{(v,v)}{(u,u)}\\ &\le \tilde{\mu}_j\frac{\|v\|^2}{\|u\|^2} \qquad\qquad (\text{since } \mu(v)>0);\end{aligned}$$

hence, since $\|v\|^2+\|u\|^2=1$ and $(\mu_j-\tilde{\mu}_j)(1-\|v\|^2)\le\tilde{\mu}_j\|v\|^2$, we get

$$0\le\frac{\mu_j-\tilde{\mu}_j}{\mu_j}\le\|v\|^2. \tag{2.23}$$

If $v=0$, then from (2.22) we see that $\mu(u)=\mu(\bar{u})$, which, together with (2.21), shows that $\tilde{\mu}_j=\mu_j$. Thus, estimate (2.23) is also valid for $v=0$.

Combining estimates (2.20) and (2.23), we obtain the first estimate in (2.18).

Finally, by Lemma 2.1,

$$\|(I-(\tilde{Q}-\tilde{P}_{1,\dots,j-m}))P_{j-m+1,\dots,j}\|^2\le\|(I-\tilde{Q})P_{j-m+1,\dots,j}\|^2+\|\tilde{P}_{1,\dots,j-m}P_{j-m+1,\dots,j}\|^2.$$

The second term can be estimated using Theorem 3.2 of [13]:

$$\|\tilde{P}_{1,\dots,j-m}P_{j-m+1,\dots,j}\|\le\frac{\|(I-\tilde{Q})T\tilde{P}_{1,\dots,j-m}\|}{\min_{i=1,\dots,j-m}|\tilde{\mu}_i-\mu_j|}\|(I-\tilde{Q})P_{j-m+1,\dots,j}\|.$$

Combining the first estimate in (2.18) with the last two inequalities completes the proof. □

Alternatively, Lemma 2.1 can be used to estimate $\|\tilde{P}_{1,\dots,j-m}P_{j-m+1,\dots,j}\|$, which results in

$$\begin{aligned}\|\tilde{P}_{1,\dots,j-1}P_{j-m+1,\dots,j}\|^2 &= \left\|\sum_{i=1}^{j-1}\tilde{P}_iP_{j-m+1,\dots,j}\right\|^2\\ &\le\sum_{i=1}^{j-1}\|\tilde{P}_iP_{j-m+1,\dots,j}\|^2.\end{aligned}$$

Every term $\|\tilde{P}_i P_{j-m+1,\dots,j}\|^2$ in the sum above can be estimated using results of [13]. For simplicity, let $m = 1$; then by Theorem 2.1 in [13],

$$\|\tilde{P}_i P_j\| \le \frac{\|T\tilde{u}_i - \tilde{\mu}_i \tilde{u}_i\|}{|\tilde{\mu}_i - \mu_j|} \sin \angle\{u_j; \tilde{U}\},$$

where $\|\tilde{u}_i\| = \|u_j\| = 1$, so we get

$$0 \le \frac{\mu_j - \tilde{\mu}_j}{\mu_j} \le \left(1 + \sum_{i=1}^{j-1} \frac{\|T\tilde{u}_i - \tilde{\mu}_i \tilde{u}_i\|^2}{|\tilde{\mu}_i - \mu_j|^2}\right) \sin^2 \angle\{u_j; \tilde{U}\},$$

which in some cases may provide a smaller constant compared to that of (2.18) with $m = 1$.

Remark 2.2. A careful examination of the proof of the first estimate in (2.18) of Theorem 2.11 shows that we can replace the orthoprojector $P_{j-m+1,\dots,j}$ with an orthoprojector P_L to any m-dimensional subspace L of $U_{1,\dots,j}$: the argument still holds and the first estimate in (2.18) can be improved:

$$0 \le \frac{\mu_j - \tilde{\mu}_j}{\mu_j} \le \inf_{L \subseteq U_{1,\dots,j},\, \dim L = m} \|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m}) P_L\|^2. \tag{2.24}$$

The right-hand side of (2.24) allows a nice geometric interpretation, using definition (2.4) of the angles between subspaces:

$$\inf_{L \subseteq U_{1,\dots,j},\, \dim L = m} \|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m}) P_L\|^2 = \sin^2 \angle_m\{U_{1,\dots,j}; \tilde{U} \cap (\tilde{U}_{1,\dots,j-m})^\perp\}.$$

This may lead to a potential improvement of the second estimate (2.18)—provided one can estimate the right-hand side of (2.24) using terms similar to those of the second estimate in (2.18).

We can derive a simple estimate of the right-hand side of (2.24), using the fact, which follows from dimensionality arguments, that

$$\dim (\tilde{U}_{1,\dots,j-m})^\perp \cap U_{1,\dots,j} \ge m.$$

Since

$$\|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m})u\| = \|(I - \tilde{Q})u\|,\ u \in (\tilde{U}_{1,\dots,j-m})^\perp \cap U_{1,\dots,j},$$

restricting the choice of L to the intersection above in (2.24) we derive that

$$0 \le \frac{\mu_j - \tilde{\mu}_j}{\mu_j} \le \inf_{L \subseteq (\tilde{U}_{1,\dots,j-m})^\perp \cap U_{1,\dots,j},\, \dim L = m} \|(I - \tilde{Q}) P_L\|^2. \tag{2.25}$$

In FEM applications typically (because of the approximability assumption) we have $\dim((\tilde{U}_{1,\dots,j-m})^\perp \cap U_{1,\dots,j}) = m$ so the inf in (2.25) is then redundant.

Estimate (2.25) improves (2.6). We note that m is a free parameter in (2.25) and can be chosen arbitrarily, $1 \le m \le j$. We finally note that (2.25) is not truly an a priori estimate since the right-hand side of it depends on the Ritz vectors $\tilde{u}_1, \dots, \tilde{u}_{j-m}$ that are not known a priori.

Let us now reformulate Theorem 2.11 in the context of the multiple eigenvalue in order to obtain a generalization of Corollary 2.10. Theorem 2.11 gives us enough flexibility to establish a different error estimate for each of the q Ritz values corresponding to the multiple eigenvalue of multiplicity q.

Corollary 2.12. *Suppose that the eigenvalue μ_p, where $p > 1$, has multiplicity $q > 1$ so that* (2.14) *holds and that $p + q - 1 \leq n$. Suppose that*

$$\min_{i=1,\dots,p-1} |\tilde{\mu}_i - \mu_p| \neq 0.$$

Then, for $j = p, \dots, p+q-1$, we have

$$\begin{aligned} 0 \leq \frac{\mu_p - \tilde{\mu}_j}{\mu_p} &\leq \|(I - \tilde{Q} + \tilde{P}_{1,\dots,p-1})P_{p,\dots,j}\|^2 \\ &\leq \left(1 + \frac{\|(I-\tilde{Q})T\tilde{P}_{1,\dots,p-1}\|^2}{\min_{i=1,\dots,p-1}|\tilde{\mu}_i - \mu_p|^2}\right)\|(I-\tilde{Q})P_{p,\dots,j}\|^2, \end{aligned} \tag{2.26}$$

where $\tilde{P}_{1,\dots,p-1}$ is the orthogonal projector onto $\tilde{U}_{1,\dots,p-1} = \operatorname{span}\{\tilde{u}_1, \dots, \tilde{u}_{p-1}\}$ and $P_{p,\dots,j}$ is the orthogonal projector onto any $(j-p+1)$-dimensional subspace of the eigenspace $U_{p,\dots,p+q-1}$ corresponding to the eigenvalue μ_p. The optimal choice of the projector $P_{p,\dots,j}$ allows us to replace the term $\|(I-\tilde{Q})P_{p,\dots,j}\|^2$ in estimate (2.26) *with $\sin^2 \angle_{j-p+1}\{U_{p,\dots,p+q-1}, \tilde{U}\}$.*

Proof. We simply take $m = j - p + 1$ in Theorem 2.11. □

To see the improvement of Corollary 2.12 over Theorem 2.7, consider the following situation. Suppose μ_2 has multiplicity 2, so $p = q = 2$. Then

$$\min_{i=1,\dots,p-1} |\tilde{\mu}_i - \mu_p| \approx \mu_1 - \mu_2 > 0,$$

provided $\tilde{\mu}_1$ is close enough to μ_1. Taking $j = 2$ in Corollary 2.12 yields

$$\frac{\mu_2 - \tilde{\mu}_2}{\mu_2} \lesssim \left(1 + \frac{\|(I-\tilde{Q})T\tilde{P}_1\|^2}{(\mu_1 - \mu_2)^2}\right)\|(I-\tilde{Q})P_2\|^2, \tag{2.27}$$

while taking $j = 3$ yields

$$\frac{\mu_3 - \tilde{\mu}_3}{\mu_3} = \frac{\mu_2 - \tilde{\mu}_3}{\mu_2} \lesssim \left(1 + \frac{\|(I-\tilde{Q})T\tilde{P}_1\|^2}{(\mu_1 - \mu_2)^2}\right)\|(I-\tilde{Q})P_{2,3}\|^2. \tag{2.28}$$

In (2.27), the eigenvalue error is bounded by a constant that is slightly larger than 1 times the square of the best approximation error for u_2; while in (2.28), we have the square of the best approximation error for $\operatorname{span}\{u_2, u_3\}$ = the eigenspace for $\mu_2 = \mu_3$. Note that estimating $(\mu_3 - \tilde{\mu}_3)/\mu_3$ with Theorem 2.7 yields no asymptotically valid estimate (cf. the discussion preceding Corollary 2.10).

Results giving different estimates for $(\mu_p - \tilde{\mu}_j)/\mu_p$, $j = p, \dots, p+q-1$ (cf. Corollaries 2.9 and 2.12) were first proved in [2]; see also [3, 4]. Our presentation simplifies and clarifies the analysis in [2, 3] and provides explicit constants. In section 3.2 we compare these results in detail. For an example of a multiple eigenvalue with eigenfunctions of differing approximabilities, see [2, 4].

Let us finally highlight the opportunities that Theorem 2.11 provides for error estimates of clustered eigenvalues in the following situation. Let

$$\mu_1 > \mu_2 \approx \mu_3 > \mu_4,$$

and suppose we are interested in error estimates for μ_2 and μ_3, assuming that $\tilde{\mu}_1 \approx \mu_1$ and $\tilde{\mu}_2 \approx \mu_2$. We do not even need Theorem 2.11 to estimate the error for μ_2: Theorem

2.7 with $j = 2$ already gives us an asymptotically valid estimate (2.27), and the fact that μ_2 is clustered (or multiple as above) is irrelevant. Theorem 2.7 with $j = 3$ does not provide an asymptotically valid estimate for the error in μ_3 since the term $|\mu_3 - \tilde{\mu}_2| \approx 0$ appears in the denominator.

Applying Theorem 2.11 with $j = 3$ we have the option of choosing the free parameter $m = 1$, 2, or 3. Taking $m = 1$ reduces Theorem 2.11 to Theorem 2.7, which does not work well in this situation as we just discussed. Taking $m = 2$ yields a good estimate

$$\frac{\mu_3 - \tilde{\mu}_3}{\mu_3} \lesssim \left(1 + \frac{\|(I - \tilde{Q})T\tilde{P}_1\|^2}{(\mu_1 - \mu_3)^2}\right) \|(I - \tilde{Q})P_{2,3}\|^2. \tag{2.29}$$

Taking $m = 3$ reduces Theorem 2.11 to Theorem 2.4,

$$\frac{\mu_3 - \tilde{\mu}_3}{\mu_3} \leq \|(I - \tilde{Q})P_{1,2,3}\|^2. \tag{2.30}$$

Comparing the right-hand sides of (2.29) and (2.30), we see that (2.29) provides a sharper estimate than (2.30) if $\mu_1 - \mu_3$ is large enough and u_1 cannot be well approximated by the trial subspace. To summarize, choosing different m in Theorem 2.11 allows us to reduce the constants in estimating errors for clustered eigenvalues at the cost of enlarging the invariant subspace that needs to be well approximated by the trial subspace. Note that in neither (2.29) nor (2.30) does the constant depend on the width of the eigenvalue cluster $\mu_2 \approx \mu_3$. Ovtchinnikov in [19] calls such estimates "cluster robust."

3. Application to the variational Galerkin method and comparisons.

We now consider the previous abstract results in two important contexts. First, suppose we have an eigenvalue problem for a symmetric positive compact integral operator T defined on $H = L_2$ and apply the classical Ritz method for integral operators. All our results apply immediately and provide relative eigenvalue error estimates for the largest eigenvalues in terms of L_2 approximability of the corresponding eigenfunctions.

Our second application is to the variational Galerkin method for symmetric positive definite differential operators. Here, we essentially need to reformulate our results for the inverse of T, but the operator T cannot be just simply replaced with its inverse since this would change the Ritz values. A proper inversion involves a simultaneous change of the scalar product as it is implicitly done in the next subsection. Our estimates of the previous section have to be somewhat rewritten in this context, since they are not invariant with respect to such an inversion.

3.1. The variational Galerkin method.

Suppose, as above, that H is a real separable Hilbert space with inner product (u, v) and norm $\|u\| = \sqrt{(u, u)}$, and suppose we are given two symmetric bilinear forms $B(u.v)$ and $D(u, v)$ on $H \times H$. The bilinear form $B(u, v)$ is assumed to satisfy

$$|B(u, v)| \leq C_1\|u\|\|v\| \quad \text{for all } u, v \in H \tag{3.1}$$

and

$$C_0\|u\|^2 \leq B(u, u) \quad \text{for all } u \in H \quad \text{with } C_0 > 0. \tag{3.2}$$

It follows from (3.1) and (3.2) that $\|u\|_B = \sqrt{B(u, u)}$ and $\|u\|$ are equivalent norms on H. For the remainder of this section we use $B(u, v)$ and $\|u\|_B$ as the inner product and norm, respectively, on H, and denote the resulting space by H_B. We also measure all angles in H_B, i.e., with respect to $B(u,v)$. Regarding $D(u,v)$ we assume that $0 < D(u,u)$ for all nonzero vectors $u \in H$ and that the unit ball of the norm $||\cdot||_D$ is compact in H.

We consider the following variationally formulated symmetric eigenvalue problem:

$$\left\{ \begin{array}{l} \text{Seek } \lambda \in R \text{ and } 0 \neq u \in H_B \text{ satisfying} \\ B(u,v) = \lambda D(u,v) \text{ for all } v \in H_B. \end{array} \right. \tag{3.3}$$

Under our assumptions, problem (3.3) has eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \cdots \nearrow +\infty$ and corresponding eigenvectors $u_1, u_2, \ldots$, which satisfy $B(u_i, u_j) = \lambda_i D(u_i, u_j) = \delta_{ij}$.

We are interested in approximating the eigenpairs of (3.3) by the variational Ritz method. Toward this end, we suppose we are given a finite-dimensional subspace $\tilde{U}$ of H_B, and consider the following finite-dimensional, variationally formulated eigenvalue problem:

$$\left\{ \begin{array}{l} \text{Seek } \tilde{\lambda} \in R \text{ and } 0 \neq \tilde{u} \in \tilde{U} \text{ satisfying} \\ B(\tilde{u},v) = \lambda D(\tilde{u},v), \text{ for all } v \in \tilde{U}. \end{array} \right. \tag{3.4}$$

Problem (3.4), being a finite-dimensional eigenvalue problem, has $n = \dim \tilde{U}$ positive eigenvalues $\tilde{\lambda}_1 \leq \tilde{\lambda}_2 \leq \cdots \leq \tilde{\lambda}_n$, and corresponding eigenvectors $\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_n$, which satisfy $B(\tilde{u}_i, \tilde{u}_j) = \tilde{\lambda}_i D(\tilde{u}_i, \tilde{u}_j) = \delta_{ij}, i,j = 1,\ldots,n$. The Poincaré inequalities $\lambda_i \leq \tilde{\lambda}_i$, $i = 1,\ldots,n$, and the min-max characterization of eigenvalues of problems (3.3) and (3.4) hold under our assumptions. We then view $\tilde{\lambda}_i$ as an approximation to λ_i, i.e., $\lambda_i \approx \tilde{\lambda}_i, i = 1,\ldots,n$.

Next we introduce the operator $T : H_B \to H_B$ defined by

$$\left\{ \begin{array}{l} Tf \in H_B, \\ B(Tf,v) = D(f,v) \text{ for all } v \in H_B \end{array} \right. \tag{3.5}$$

and the operator $\tilde{T} : \tilde{U} \to \tilde{U}$ defined by

$$\left\{ \begin{array}{l} \tilde{T}f \in \tilde{U},\ f \in \tilde{U}, \\ B(\tilde{T}f,v) = D(f,v) \text{ for all } v \in \tilde{U}. \end{array} \right. \tag{3.6}$$

The operator T is the solution operator for the "boundary value problem" corresponding to the eigenvalue problem (3.3). By our assumption, the unit ball of $\|\cdot\|_D$ is compact in H and, therefore, in H_B, thus, the operator T is compact in H_B. Of course, $\tilde{T}$, being an operator on a finite-dimensional space, is also compact. It follows directly from definition (3.5) that T is symmetric and positive definite on H_B and from definition (3.6) that $\tilde{T}$ is symmetric and positive definite on $\tilde{U}$ (with respect to $B(u,v)$). It is easily seen that, if, as above, $\tilde{Q}$ is the orthogonal projector of H_B onto $\tilde{U}$, then $\tilde{T} = (\tilde{Q}T)|_{\tilde{U}}$.

The eigenvalues of problem (3.3) and of the operator T are reciprocals: $\lambda_i = 1/\mu_i$, $i = 1,2,\ldots$; problem (3.3) and the operator T have the same eigenvectors u_i. Likewise, the eigenvalues of problem (3.4) and of the operator $\tilde{T}$ are reciprocals: $\tilde{\lambda}_i = 1/\tilde{\mu}_i$, $i = 1,2,\ldots,n$; problem (3.4) and the operator $\tilde{T}$ have the same eigenvectors $\tilde{u}_i$. As in the previous section, we choose $\{u_i\}$ and $\{\tilde{u}_i\}$ to be orthonormal systems, in the context of the present section, that is, in H_B.

The FEM approximation of eigenvalue problems for symmetric differential operators can be viewed as a variational Ritz method, and the FEM eigenvalue errors can be estimated using the theorems of the previous section.

We can utilize Theorems 2.4 and 2.7, applied to T and $\tilde{T}$ on H_B, to estimate the eigenvalue error $(\tilde{\lambda}_i - \lambda_i)/\tilde{\lambda}_i$. Here $U_{1,\dots,j}$ denotes the span of the eigenvectors $u_1, \dots, u_j$, and $P_{1,\dots,j}$ is the H_B orthogonal projector onto $U_{1,\dots,j}$.

THEOREM 3.1. *For $j = 1, \dots, n = \dim \tilde{U}$ we have*

$$0 \le \frac{\tilde{\lambda}_j - \lambda_j}{\tilde{\lambda}_j} \le \sin^2 \angle_B\{U_{1,\dots,j}; \bar{U}\} = \|(I - \tilde{Q})P_{1,\dots,j}\|_B^2. \tag{3.7}$$

Remark 3.1. By analogy with Remark 2.1, from Theorem 3.1 we get the following estimate, mathematically equivalent to estimate (2.7):

$$0 \le \frac{\tilde{\lambda}_j - \lambda_j}{\tilde{\lambda}_j} \le \sum_{i=1}^{j} \|(I - \tilde{Q})u_i\|_B^2,$$

which can be rewritten as

$$0 \le \frac{\tilde{\lambda}_i - \lambda_i}{\lambda_i} \le \frac{\sum_{i=1}^{j} \|(I - \tilde{Q})u_i\|_B^2}{1 - \sum_{i=1}^{j} \|(I - \tilde{Q})u_i\|_B^2}, \tag{3.8}$$

assuming that the denominator in the latter expression is positive. Estimate (3.8) is well known (see, e.g., Theorem 2.1 in Chapter 4 of [22]); a similar estimate is proved in [5].

To formulate the next theorem—an analogue of Theorem 2.7—we recall that $\tilde{P}_{1,\dots,j-1}$ is the orthogonal projector of H_B onto $\tilde{U}_{1,\dots,j-1} = \operatorname{span}\{\tilde{u}_1, \dots, \tilde{u}_{j-1}\}$, where $\tilde{u}_i$ are eigenvectors of (3.4).

THEOREM 3.2. *For a fixed index j such that $1 \le j \le n = \dim \tilde{U}$, suppose*

$$\min_{1,\dots,j-1} |\tilde{\lambda}_i - \lambda_j| \neq 0. \tag{3.9}$$

Then

$$\begin{aligned} 0 \le \frac{\tilde{\lambda}_j - \lambda_j}{\tilde{\lambda}_j} &\le \|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-1})u_j\|_B^2 \\ &\le \left(1 + \max_{i=1,\dots,j-1} \frac{\tilde{\lambda}_i^2 \lambda_j^2}{|\tilde{\lambda}_i - \lambda_j|^2} \|(I - \tilde{Q})T\tilde{P}_{1,\dots,j-1}\|_B^2\right) \sin^2 \angle_B\{u_j; \tilde{U}\}. \end{aligned} \tag{3.10}$$

Similarly, we can apply Theorem 2.11 to obtain the following.

THEOREM 3.3. *For fixed indexes j and m satisfying $1 \le j \le n$ and $1 \le m \le j$, let $U_{j-m+1,\dots,j}$ be the m-dimensional invariant subspace corresponding to eigenvalues $\lambda_j \ge \cdots \ge \lambda_{j-m+1}$, and let $P_{j-m+1,\dots,j}$ be the H_B orthogonal projector on $U_{j-m+1,\dots,j}$. If*

$$\min_{i=1,\dots,j-m} |\tilde{\lambda}_i - \lambda_j| \neq 0, \tag{3.11}$$

then

$$\begin{aligned} 0 \le \frac{\tilde{\lambda}_j - \lambda_j}{\tilde{\lambda}_j} &\le \|(I - \tilde{Q} + \tilde{P}_{1,\dots,j-m})P_{j-m+1,\dots,j}\|_B^2 \\ &\le \left(1 + \max_{i=1,\dots,j-m} \frac{\tilde{\lambda}_i^2 \lambda_j^2}{|\tilde{\lambda}_i - \lambda_j|^2} \|(I - \tilde{Q})T\tilde{P}_{1,\dots,j-m}\|_B^2\right) \|(I - \tilde{Q})P_{j-m+1,\dots,j}\|_B^2, \end{aligned} \tag{3.12}$$

where $\tilde{P}_{1,\dots,j-m}$ *is the* H_B *orthogonal projector onto* $\tilde{U}_{1,\dots,j-m} = \operatorname{span}\{\tilde{u}_1, \dots, \tilde{u}_{j-m}\}$ *(if* $j = m$*, we set* $\tilde{P}_{1,\dots,j-m} = 0$ *and do not use* (3.11)*). If* $m = j$*, the present theorem turns into Theorem* 3.1*; if* $m = 1$*, it turns into Theorem* 3.2*.*

Let us finally reformulate Theorem 3.3 in the context of the multiple eigenvalue by analogy with Corollary 2.12.

Corollary 3.4. *Suppose that the eigenvalue* λ_p*, where* $p > 1$*, has multiplicity* $q > 1$*, so that*

$$\lambda_{p-1} < \lambda_p = \lambda_{p+1} = \cdots = \lambda_{p+q-1} < \lambda_{p+q} \tag{3.13}$$

holds, and that $p + q - 1 \le n$*. Suppose that*

$$\min_{i=1,\dots,p-1} |\tilde{\lambda}_i - \lambda_p| \neq 0.$$

Then, for $j = p, \dots, p + q - 1$*, we have*

$$0 \le \frac{\tilde{\lambda}_j - \lambda_p}{\tilde{\lambda}_j} \le \|(I - \tilde{Q} + \tilde{P}_{1,\dots,p-1})P_{p,\dots,j}\|_B^2$$

$$\le \left(1 + \max_{i=1,\dots,p-1} \frac{\tilde{\lambda}_i^2 \lambda_p^2}{|\tilde{\lambda}_i - \lambda_p|^2} \|(I - \tilde{Q})T\tilde{P}_{1,\dots,p-1}\|_B^2\right) \|(I - \tilde{Q})P_{p,\dots,j}\|_B^2, \tag{3.14}$$

where $\tilde{P}_{1,\dots,p-1}$ *is the* H_B *orthogonal projector onto* $\tilde{U}_{1,\dots,p-1} = \operatorname{span}\{\tilde{u}_1, \dots, \tilde{u}_{p-1}\}$ *and* $P_{p,\dots,j}$ *is the* H_B *orthogonal projector onto any* $(j - p + 1)$*-dimensional subspace of the eigenspace* $U_{p,\dots,p+q-1}$ *corresponding to the eigenvalue* λ_p*. The main term, the multiplier* $\|(I-\tilde{Q})P_{p,\dots,j}\|_B^2$*, in* (3.14) *can be replaced with* $\sin^2 \angle_{j-p+1}\{U_{p,\dots,p+q-1}, \tilde{U}\}$ *by choosing the projector* $P_{p,\dots,j}$ *in the optimal way, where the angle is measured in* H_B*.*

3.2. Comparison with known asymptotic estimates for eigenvalues. Estimate (3.10) should be compared with estimates of Vainikko in Krasnosel'skii et al. [16], Chatelin [7], and Babuška and Osborn [3], all of which address a slightly different context that we now describe.

In addition to all assumptions of the previous subsection, let $\{U^h\}$ be a family of finite-dimensional subspaces of H_B depending on a parameter $h > 0$ called the mesh parameter. For a fixed h, we use $U^h = \tilde{U}$ as the trial subspace for the variational Ritz method. Let $Q^h = \tilde{Q}$ be the H_B orthogonal projector on U^h. We make the following approximability assumption on the family $\{U^h\}$:

$$\|(I - Q^h)u\|_{H_B} = \inf_{v^h \in U^h} \|u - v^h\|_{H_B} \to 0 \quad \text{as } h \to 0 \quad \text{for each } u \in H_B. \tag{3.15}$$

To be consistent with our new h-based notation, we denote the approximate eigenvalues by $\lambda_j^h = \tilde{\lambda}_j$ and the corresponding eigenvectors by $u_j^h = \tilde{u}_j$. It is well known that under assumption (3.15) we have $\lambda_j^h \to \lambda_j$ as $h \to 0$ for each fixed j.

We compare our results to estimates of [3, 7, 16] that are asymptotic, $h \to 0$, upper (and lower) bounds for the ratio $(\lambda_j^h - \lambda_j)/\lambda_j^h$ in [7, 16] and for the ratio $(\lambda_j^h - \lambda_j)/\lambda_j$ (notice a slightly different denominator) in [3]. Since

$$\frac{\lambda_j^h - \lambda_j}{\lambda_j} = \frac{\lambda_j^h - \lambda_j}{\lambda_j^h} + \frac{(\lambda_j^h - \lambda_j)^2}{\lambda_j \lambda_j^h},$$

where the second term in the sum on the right can be asymptotically ignored, the results of [7, 16] asymptotically estimate the same eigenvalue errors as those of [3]. Results of [3] provide upper bounds for $(\lambda_j^h - \lambda_j)/\lambda_j$ that trivially serve also as upper bounds for $(\lambda_j^h - \lambda_j)/\lambda_j^h$. Moreover, it is possible to show that the lower bounds for $(\lambda_j^h - \lambda_j)/\lambda_j$ in [3] also hold for $(\lambda_j^h - \lambda_j)/\lambda_j^h$ without any changes. Here, we will formulate all the results (except for (3.23)) in terms of $(\lambda_j^h - \lambda_j)/\lambda_j^h$ to be consistent with our estimates.

We start our discussion with the case of a simple eigenvalue λ_j and later turn our attention to the case of multiple eigenvalues. The convergence rate for a simple eigenvalue is determined by the following well-known result: let real r_j^h be defined by

$$0 \le \frac{\lambda_j^h - \lambda_j}{\lambda_j^h} = \left(1 + r_j^h\right) \|(I - Q^h)u_j\|_B^2; \tag{3.16}$$

then $r_j^h \to 0$ as $h \to 0$; see subsection 18.6 (pp. 285–286) of [16] and subsection 6.2 (pp. 315–317) of [7]. Babuška and Osborn [3] showed that

$$|r_j^h| \le d_j \sup_{\|g\|_D = 1} \|(I - Q^h)Tg\|_B^2 \to 0 \tag{3.17}$$

and that (cf. (2.9))

$$|r_j^h| \le d_j \sup_{\|g\|_B = 1} \|(I - Q^h)Tg\|_B \to 0, \tag{3.18}$$

where $d_j > 0$ are unknown generic constants.

Our present estimate (3.10) using the h notation takes the form (3.16) with

$$r_j^h \le \max_{i=1,\dots,j-1} \frac{(\lambda_i^h)^2 \lambda_j^2}{|\lambda_i^h - \lambda_j|^2} \|(I - Q^h)TP_{1,\dots,j-1}^h\|_B^2. \tag{3.19}$$

The first multiplier in the right-hand side of (3.19) is asymptotically (as $h \to 0$) a constant,

$$\frac{(\lambda_i^h)^2 \lambda_j^2}{|\lambda_i^h - \lambda_j|^2} \to \frac{\lambda_{j-1}^2 \lambda_j^2}{|\lambda_{j-1} - \lambda_j|^2},$$

provided that the eigenvalue λ_j is simple. The second multiplier is bounded by

$$\begin{aligned} \|(I - Q^h)TP_{1,\dots,j-1}^h\|_B^2 &\le \|(I - Q^h)T\|_B^2 \\ &= \sup_{\|g\|_B = 1} \|(I - Q^h)Tg\|_B^2 \\ &\le \frac{1}{\lambda_1} \sup_{\|g\|_D = 1} \|(I - Q^h)Tg\|_B^2; \end{aligned}$$

thus, our estimate (3.19) is an improvement of both estimates (3.17) (our constant is explicitly written) and (3.18) (we have the small multiplier squared) of [3]. However, our estimate (3.19) provides only an upper bound for r_j^h, while (3.17) and (3.18) also give the lower bounds because they estimate the absolute value $|r_j^h|$. Let us note that the denominator $|\lambda_{j-1} - \lambda_j|^2$ may be small, but the term in the numerator is bounded from above by a constant times $\sup_{\|g\|_D = 1} \|(I - Q^h)Tg\|_B^2 \to 0$ as $h \to 0$.

Now suppose eigenvalue λ_p has multiplicity q so that (3.13) holds, and let $P_{p,\dots,p+q-1}$ be the H_B orthogonal projector on the q-dimensional eigenspace, corresponding to $\lambda_p = \lambda_{p+1} = \cdots = \lambda_{p+q-1}$ as in Corollary 3.4 in subsection 18.6 (pp. 285–286) of Krasnosel'skii et al. [16] and Chatelin in subsection 6.2 (pp. 315–317) of [7] prove that

$$0 \le \frac{\lambda_j^h - \lambda_p}{\lambda_j^h} = \left(1 + r_j^h\right) \frac{\|(I - Q^h) P_{p,\dots,p+q-1} u_j^h\|_B^2}{\|P_{p,\dots,p+q-1} u_j^h\|_B^2}, \qquad j = p, \dots, p+q-1, \tag{3.20}$$

where $r_j^h \to 0$ as $h \to 0$. An evident difficulty in using estimate (3.20) for a priori error analysis is that the approximate eigenfunctions u_{j+i-1}^h are not known a priori. If we consider the worst case, it leads to the estimate, which bounds the error for all q Ritz values, using

$$\|(I - Q^h) P_{p,\dots,p+q-1}\|_B^2 = \sin^2 \angle\{U_{p,\dots,p+q-1}; U_h\}. \tag{3.21}$$

Let us remind the reader that an angle without an index denotes the largest angle, according to our agreement in subsection 2.2, and that in this and the previous subsections all angles are measured in H_B.

In some cases (see [2, 4] for examples), the eigenspace may be spanned by eigenfunctions of different approximation qualities, and it is interesting to analyze how this affects the error for different Ritz values. As mentioned in the introduction, such results were first proved by Babuška and Osborn in [2]. In [3], they completed such an analysis for the smallest of the q Ritz values, using

$$\inf_{u \in U_{p,\dots,p+q-1},\, \|u\|_B = 1} \|(I - Q^h) u\|_B^2 = \sin^2 \angle_1\{U_{p,\dots,p+q-1}; U_h\}, \tag{3.22}$$

which depends on the approximability of the most easily approximated eigenfunction in the eigenspace. Thus, estimates based on (3.21) and (3.22) represent two extremes: (3.21) uses the largest angle and serves to estimate the largest error (thus effectively all q errors at once), while (3.22) uses the smallest angle and can estimate only one, the smallest, eigenvalue error.

For the intermediate multiple eigenvalue error, Babuška and Osborn in [3] established the following result: for $j = p, \dots, p+q-1$ let the quantities r_j^h be redefined by

$$0 \le \frac{\lambda_j^h - \lambda_p}{\lambda_p} = \left(1 + r_p^h\right) \inf_{\substack{u \in U_{p,\dots,p+q-1}, \\ u \in (U^h_{p,\dots,j-1})^{\perp_B}, \\ \|u\|_B = 1}} \|(I - Q^h) u\|_B^2; \tag{3.23}$$

then

$$|r_j^h| \le d_j \sup_{\|g\|_B = 1} \|(I - Q^h) T g\|_B, \tag{3.24}$$

with generic constants $d_j > 0$, and where the orthogonal complement $\left(U^h_{p,\dots,j-1}\right)^{\perp_B}$ is taken in H_B. In [4], it is shown that r_j^h in (3.23) are bounded, but more detailed estimates (3.24) appear only in [3].

We note that the constraints on u in (3.23) are similar to those in (2.25) except that (2.25) involves orthogonalization to all previous Ritz vectors, while (3.23) needs only orthogonalization to previous Ritz vectors corresponding to the multiple eigenvalue under consideration. Both (2.25) and (3.23) are not truly a priori estimates since their right-hand sides depends on Ritz vectors that are not known a priori.

In contrast, our estimate (3.14) can be formulated as follows: let for $j = p, \dots, p + q - 1$ the quantities r_j^h be yet again redefined by

$$0 \le \frac{\lambda_j^h - \lambda_p}{\lambda_j^h} = \left(1 + r_j^h\right) \sin^2 \angle_{j-p+1}\{U_{p,\dots,p+q-1}, U^h\}; \tag{3.25}$$

then

$$r_j^h \le \max_{i=1,\dots,p-1} \frac{(\lambda_i^h)^2 \lambda_p^2}{|\lambda_i^h - \lambda_p|^2} \|(I - Q^h) T P_{1,\dots,p-1}^h\|_B^2.$$

We have already shown that our upper bound for r_p^h is better than that given by estimate (3.24): the constant is explicitly written and the h-dependent part is smaller. Let us turn our attention to the main term of the right-hand side of (3.25), namely, the $\sin^2 \angle_{j-p+1}\{U_{p,\dots,p+q-1}, U^h\}$ multiplier, and demonstrate that it is smaller than the main term of the right-hand side of (3.23) and that it can be easily estimated from above using (2.5).

We first highlight again that this multiplier can be estimated a priori since it does not depend on Ritz vectors, contrary to the main term of estimate (3.23). Second, we can directly compare the main terms in (3.23) and (3.25). Indeed, by (2.4), and since $\dim\{\left(U_{p,\dots,j-1}^h\right)^\perp \cap U_{p,\dots,p+q-1}\} \ge j - p + 1$, we have for $j = p, \dots, p + q - 1$:

$$\begin{aligned}
\sin^2 \angle_{j-p+1}\{U_{p,\dots,p+q-1}, U^h\} &= \inf_{L \subseteq U_{p,\dots,p+q-1},\, \dim L = j-p+1} \sin^2 \angle\{L; U^h\} \\
&\le \sin^2 \angle\{\left(U_{p,\dots,j-1}^h\right)^\perp \cap U_{p,\dots,p+q-1}; U^h\} \\
&= \inf_{\substack{u \in U_{p,\dots,p+q-1}, \\ u \in \left(U_{p,\dots,j-1}^h\right)^\perp, \\ \|u\|_B = 1}} \|(I - Q^h) u\|_B^2,
\end{aligned}$$

so our estimate (3.25) is sharper than (3.23).

Using the term $\sin^2 \angle_{j-p+1}\{U_{p,\dots,p+q-1}, U^h\}$ has yet another advantage: namely, it permits the application of (2.5). Suppose the vectors $\{u_i, i = p, \dots, p + q - 1\}$ form an orthogonal basis for the subspace $U_{p,\dots,p+q-1}$ and are arranged in such a way that

$$\angle\{u_p; U^h\} \le \cdots \le \angle\{u_{p+q-1}; U^h\}.$$

Then, by (2.5),

$$\sin^2 \angle_{j-p+1}\{U_{p,\cdots,p+q-1}; U^h\} \le \sum_{i=p}^{j} \sin^2 \angle\{u_i; U^h\}, \qquad j = p, \dots, p + q - 1.$$

In other words, if the eigenspace $U_{p,\dots,p+q-1}$ is spanned by eigenfunctions of different approximation qualities, our result assesses the quality of each of the Ritz values corresponding to the multiple eigenvalue.

Conclusions. We derive eigenvalue error bounds for the Ritz method that have several novel features:

- For a simple eigenvalue, our estimates improve those previously known and provide explicit values for all constants.

- For a multiple eigenvalue we prove, in addition, what is apparently the first truly a priori error estimates that show the levels of the eigenvalue errors depending on approximability of eigenfunctions in the corresponding eigenspace.
- For clustered eigenvalues, our results provide elegant eigenvalue error bounds that do not depend on the width of the cluster.

In the FEM eigenvalue approximation context, our results improve earlier known results and are readily applicable for a fixed mesh without making the traditional assumption about the mesh size being small enough.

REFERENCES

[1] I. Babuška, B. Q. Guo, and J. E. Osborn, *Regularity and numerical solution of eigenvalue problems with piecewise analytic data*, SIAM J. Numer. Anal., 26 (1989), pp. 1534–1560.

[2] I. Babuška and J. E. Osborn, *Estimates for the errors in eigenvalue and eigenvector approximation by Galerkin methods, with particular attention to the case of multiple eigenvalues*, SIAM J. Numer. Anal., 24 (1987), pp. 1249–1276.

[3] I. Babuška and J. E. Osborn, *Finite element-Galerkin approximation of the eigenvalues and eigenvectors of selfadjoint problems*, Math. Comp., 52 (1989), pp. 275–297.

[4] I. Babuška and J. E. Osborn, *Eigenvalue problems*, in Handbook of Numerical Analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 641–787.

[5] G. Birkhoff, C. de Boor, B. Swartz, and B. Wendroff, *Rayleigh-Ritz approximation by piecewise cubic polynomials*, SIAM J. Numer. Anal., 3 (1966), pp. 188–203.

[6] J. H. Bramble, J. E. Pasciak, and A. V. Knyazev, *A subspace preconditioning algorithm for eigenvector/eigenvalue computation*, Adv. Comput. Math., 6 (1996), pp. 159–189.

[7] F. Chatelin, *Spectral Approximations of Linear Operators*, Academic Press, New York, 1983.

[8] E. G. D'yakonov, *Optimization in Solving Elliptic Problems*, CRC Press, Boca Raton, FL, 1996.

[9] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1976.

[10] A. V. Knyazev, *Computation of Eigenvalues and Eigenvectors for Mesh Problems: Algorithms and Error Estimates*, Department of Numerical Mathematics, USSR Academy of Sciences, Moscow, 1986 (in Russian).

[11] A. V. Knyazev, *Sharp a priori error estimates of the Rayleigh-Ritz method without assumptions of fixed sign or compactness*, Math. Notes, 38 (1986), pp. 998–1002.

[12] A. V. Knyazev, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.

[13] A. V. Knyazev, *New estimates for Ritz vectors*, Math. Comp., 66 (1997), pp. 985–995.

[14] A. V. Knyazev and M. E. Argentati, *Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates*, SIAM J. Sci. Comput., 23 (2002), pp. 2008–2040.

[15] A. V. Knyazev and J. Osborn, *New A Priori FEM Error Estimates for Eigenvalues*, Tech. report, UCD-CCM 215, Center for Computational Mathematics, University of Colorado at Denver, Denver, CO, 2004. Available online at http://math.cudenver.edu/ccm/reports/

[16] M. A. Krasnosel'skii, G. M. Vainikko, P. P. Zabreiko, Ya. B. Rutitskii, and Y. Ya. Stetsenko, *Approximate Solution of Operator Equations*, Wolters-Noordhoff, Groningen, 1972.

[17] M. N. Krylov, *Les méthodes de solution approachée des problèmes de la physique mathématique*, Mém. Sci Math. 49, (1931), pp. 1–69.

[18] J. E. Osborn, *Spectral approximation for compact operators*, Math. Comput., 29 (1975), pp. 712–725.

[19] E. Ovtchinnikov, *Cluster Robust Error Estimates for the Rayleigh–Ritz Approximation* II: *Estimates for Eigenvalues*, Tech. report 210, Center for Computational Mathematics, University of Colorado at Denver, Denver, CO, 2004. Available online at http://math.cudenver.edu/ccm/reports/rep210.pdf.gz Linear Algebra Appl., to appear.

[20] B. N. Parlett, *The Symmetric Eigenvalue Problem*, SIAM, Philadelphia, PA, 1998.

[21] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[22] H. F. Weinberger, *Variational Methods for Eigenvalue Approximation*, SIAM, Philadelphia, 1974.

SIAM J. Numer. Anal.
Vol. 43, No. 6, pp. 2668–2689

SHARP CONVERGENCE ESTIMATES FOR THE PRECONDITIONED STEEPEST DESCENT METHOD FOR HERMITIAN EIGENVALUE PROBLEMS*

E. E. OVTCHINNIKOV†

Abstract. The paper is concerned with convergence estimates for the preconditioned steepest descent method for the computation of the smallest eigenvalue of a Hermitian operator. Available estimates are reviewed and new estimates are introduced that improve on the known ones in certain respects. In addition to the estimates for the error reduction after one iteration, we consider estimates for the so-called asymptotic convergence factor defined as the upper limit of the average error reduction per iteration. The paper focuses on sharp estimates, i.e., those that cannot be improved without using additional information.

Key words. eigenvalue computation, preconditioning, steepest descent, convergence estimates

AMS subject classifications. 65F15, 65N12, 65N22, 65N30

DOI. 10.1137/040620643

1. Introduction.

The steepest descent method is a classical method for finding minima of smooth real-valued functionals that is based on the following simple and natural idea. Let $\psi(v)$ be a real-valued smooth functional in a Euclidean space $\mathcal{E}$, which we for the moment assume for simplicity to be real. Given an approximation u^i to the minimum point u of $\psi(v)$, we look for the next approximation u^{i+1} in the direction in which $\psi(v)$ decreases faster than in any other direction. Since $\psi(v)$ is assumed to be smooth, this steepest descent direction is opposite to the direction of the gradient of $\psi(v)$ at $v = u^i$, which we denote $\nabla\psi(u^i)$. Hence, we arrive at the iterative scheme

$$u^{i+1} = u^i - \omega_i \nabla\psi(u^i), \tag{1.1}$$

where ω_i are suitably chosen parameters. A natural choice for ω_i is the minimum argument of $\psi_i(\omega) \equiv \psi(u^i - \omega\nabla\psi(u^i))$, and it is this particular choice that has become associated with the term "steepest descent method."[1]

The steepest descent method can be used to solve the linear system $Lu = f$ with a Hermitian positive definite operator L owing to the fact that the solution u of this system minimizes the functional $\phi(v) = (Lv, v) - 2(v, f)$. It can also be used to find the smallest eigenvalue λ_1 of a Hermitian operator L, which minimizes the Rayleigh quotient functional $\lambda(v) = (Lv, v)/(v, v)$.

The gradient of a functional ψ at a point v is orthogonal to the level set of ψ to which v belongs, i.e., the set of w such that $\psi(w) = \psi(v)$. Hence, one way to understand the convergence behavior of the steepest descent method is to analyze the shape of the level sets. In the case of the linear system $Lu = f$ with a Hermitian positive definite L, we have $\psi(v) = \phi(v) = \phi(u) + (L(v-u), (v-u))$, and hence each

*Received by the editors December 3, 2004; accepted for publication (in revised form) July 25, 2005; published electronically January 27, 2006.
http://www.siam.org/journals/sinum/43-6/62064.html

†University of Westminster, London HA1 3TP, UK (eeo@wmin.ac.uk).

[1] A more proper term would be "locally optimal steepest descent method" because "steepest" actually refers to the direction of the descent rather than to any particular point in that direction. In this paper the term "steepest descent" is used in the general sense in order to cover the results in [1] and [7, 8], which are given for a different choice of ω_i.

level set is an ellipsoid with the center at u and the axes collinear to the eigenvectors u_j of L. The respective lengths of the semiaxes are $1/\sqrt{\lambda_j}$, where λ_j are the respective eigenvalues. If the condition number of L; i.e., the ratio λ_{-1}/λ_1 of the largest eigenvalue[2] λ_{-1} to λ_1 is large then the level sets of $\phi(u)$ in the plane $\operatorname{span}\{u_1, u_{-1}\}$ look like those of a very narrow ravine, and the convergence can be slow. Indeed, if we take $u^0 = u + (1/\lambda_1)u_1 + (1/\lambda_{-1})u_{-1}$ as the initial guess and use locally optimal ω_i, then the iterates u^i approach u zigzagging across the bottom of the "ravine" (the axis collinear to u_1), and the convergence is slow.

Introducing a new scalar product in $\mathcal{E}$ changes the geometry of this space and hence the shape of the level sets of functionals. Indeed, if we introduce the scalar product $(\cdot,\cdot)_L = (L\cdot,\cdot)$ and the corresponding norm $\|\cdot\|_L = (\cdot,\cdot)_L^{1/2}$, then $\phi(v) = \phi(u) + \|v-u\|_L^2$, i.e., the level sets are spherical, and hence the steepest descent direction at any $u^0 \in \mathcal{E}$ points to the solution u. In the general case, if we use the scalar product $(\cdot,\cdot)_N$, where N is a Hermitian positive definite operator, then the shape of the level sets depends on the condition number of $N^{-1}L$: the smaller the condition number, the closer the level sets to spherical, and hence the faster the convergence. Denoting $K = N^{-1}$, we have in the above scalar product $\nabla\phi(u) = 2K(Lu - f)$, and the iterative scheme (1.1) becomes what is known as the preconditioned steepest descent (PSD):

$$u^{i+1} = u^i - \tau_i K r^i, \quad r^i = Lu^i - f. \tag{1.2}$$

In a similar way, in the case of the functional $\lambda(u)$ we arrive at the iterative scheme

$$u_1^{i+1} = u_1^i - \tau_i K r_1^i, \quad r_1^i = (L - \lambda_1^i)u_1^i, \tag{1.3}$$

where u_1^i approximate u_1, $\lambda_1^i = \lambda(u_1^i)$, and τ_i minimizes $\lambda^i(\tau) \equiv \lambda(u_1^i - \tau K r_1^i)$.

Since $Kr^i = KLu^i - Kf$, (1.2) can be viewed as an iterative scheme for solving the system $KLu = Kf$. This a more familiar interpretation of the preconditioning for a linear system that also explains the terminology: K (or, sometimes $N = K^{-1}$) is called the preconditioner because the multiplication of the system by K changes the condition number of this system. In the case of an eigenvalue problem, however, such an interpretation is not very helpful because the multiplication of $Lu = \lambda u$ by K results in $KLu = \lambda K u$, and it is not so obvious why the solution of the latter generalized eigenvalue problem is any easier than that of the former standard eigenvalue problem. Turning to the above geometrical interpretation and assuming λ_1 to be simple, we observe first that the level sets of $\lambda(v)$ in the vicinity of u_1 are coaxial cones with the axis collinear to u_1 and ellipsoidal cross-sections. Considerations similar to the above suggest that the speed of convergence of the steepest descent iterations is determined by the "roundness" of the level sets, which is reflected by the "roundness" of their orthogonal cross-sections. It is not difficult to see that if v lies in an orthogonal cross-section, i.e., $v - u_1$ is orthogonal to an eigenvector u_1 corresponding to λ_1, then

$$\begin{aligned}\lambda(v) &= \lambda_1 + ((L-\lambda_1)(v-u_1),(v-u_1)) + \mathcal{O}\left(\|v-u_1\|^4\right)\\ &= \lambda_1 + (K(L-\lambda_1)(v-u_1),(v-u_1))_N + \mathcal{O}\left(\|v-u_1\|^4\right).\end{aligned}$$

Hence, one is led to the conclusion that asymptotically the convergence of (1.3) is determined by the condition number of the operator $K(L-\lambda_1)$ restricted to the

[2] In using this negative index for the largest eigenvalue we follow the notation of [13].

subspace orthogonal to u_1, and indeed, the asymptotic estimate from [14] shows that this is the case.

The nonasymptotic quantitative convergence analysis of the preconditioned steepest descent method for eigenvalue problems proved to be much more difficult than the above simple considerations and several decades had passed before a nonasymptotic convergence result equivalent to that in [14] was obtained. A comprehensive review of the convergence results that were obtained in those decades can be found in [1]—in this paper we only mention some of them, focusing on their accuracy. Special attention is paid to the convergence estimates that are *sharp*, i.e., cannot be improved without using additional information. Such estimates have obvious theoretical importance, being pieces of ultimate knowledge about the convergence properties of the PSD method. Recent research in which the author takes part suggests that the accuracy of the convergence estimates for PSD may be of indirect practical importance, as these estimates can be employed in the *a posteriori* error estimation. It is observed in [4] that $\tilde{\lambda}_1 - \lambda_1 \le (\tilde{\lambda}_1 - \lambda_{PSD})/(1 - q_{PSD})$, where $\tilde{\lambda}_1$ is an approximation to λ_1, λ_{PSD} is the new approximation to λ_1 computed by the PSD method and q_{PSD} is the upper bound for the eigenvalue error reduction after one PSD iteration. Hence, the accuracy of the bound q_{PSD} affects the accuracy of the above eigenvalue error estimate.

In this paper we show that the convergence estimate in [14] is asymptotically sharp, i.e., the main term in the error bound it provides cannot be made smaller without using additional information. We show that the same holds for some other available convergence results. Further, we present sharp estimates for the so-called *asymptotic convergence factor* (cf. section 5), which, in a sense, represents the average error reduction per iteration, and, in this particular respect, gives one a better idea about the long-term convergence behavior than the worst error reduction after one iteration. Finally, we present new sharp nonasymptotic convergence estimates and compare them with the available ones.

It should be noted that the results of this paper can also be applied to the PSD for the generalized eigenvalue problem $\hat{L}\hat{u} = \lambda \hat{M}\hat{u}$ with Hermitian $\hat{L}$ and Hermitian positive definite $\hat{M}$. Indeed, the gradient of the Rayleigh quotient $\hat{\lambda}(\hat{v}) = (\hat{L}\hat{v}, \hat{v})/(\hat{M}\hat{v}, \hat{v})$ in the scalar product $(\hat{K}^{-1}\cdot, \cdot)$ is proportional to $\hat{K}(\hat{L} - \hat{\lambda}(\hat{v})\hat{M})\hat{v}$ and hence the iterations (1.1) become

$$\hat{u}_1^{i+1} = \hat{u}_1^i - \tau_i \hat{K}\hat{r}_1^i, \quad \hat{r}_1^i = (\hat{L} - \hat{\lambda}(\hat{u}_1^i)\hat{M})\hat{u}_1^i. \tag{1.4}$$

Denoting $u_1^i = \hat{M}^{1/2}\hat{u}_1^i$, $L = \hat{M}^{-1/2}\hat{L}M^{-1/2}$, and $K = \hat{M}^{1/2}\hat{K}\hat{M}^{1/2}$ we have $\hat{\lambda}(\hat{u}_1^i) = \lambda(u_1^i)$ and $\hat{K}\hat{r}_1^i = \hat{M}^{-1/2}Kr_1^i$ and hence (1.4) transforms into (1.3).

The outline of the paper is as follows. In section 2 we discuss the very first convergence result for the iterations (1.3), obtained in asymptotic form by Samokish [14], and its remarkable features, in particular, its relation to the superlinearly convergent (exact) Jacobi–Davidson method [15]. In section 3 we review and compare some available nonasymptotic convergence results. In section 4 we analyze the asymptotic accuracy of the estimate by Samokish and two nonasymptotic estimates: one by Knyazev in [5] and the other by the author in [11]. We show that all three are asymptotically sharp under their respective assumptions by providing for each a simple example whereby the ratio of the two consecutive eigenvalue errors asymptotically coincide with its theoretical upper limit given in the respective estimate. In section 5 we strengthen the above asymptotical sharpness results by showing that in each case the aforementioned ratio remains asymptotically close to its upper limit at all iterations. In section 6 we introduce new sharp convergence estimates using what appears to be a novel approach to the convergence analysis of the PSD iterations (1.3). We introduce a Hermitian positive definite operator that maps the residual vector onto the error vector $u_1^i - \omega u_1$, where ω is a nonzero scalar, and we estimate the reduction in the eigenvalue error after one iteration (1.3) via the closeness of the preconditioner to the above "ideal" preconditioner. We note that in the case where L is positive definite, the closeness of K to the "ideal" preconditioner can be easily estimated via the condition number of KL, thus producing a convergence result given in more familiar terms. Finally, in section 7 we consider some block versions of the PSD method and discuss available convergence results. Using the asymptotical results of sections 4 and 5, we show that the convergence estimate for a particular block version of PSD obtained by the author in [12] is asymptotically sharp.

In the paper we enumerate the eigenvalues λ_i in increasing order and, apart form section 7, without taking multiplicity into account, i.e, λ_2 is the second *distinct* eigenvalue etc. The eigenvectors are assumed to be normalized. As already specified above, the operator L is assumed to be Hermitian and, for some results, positive definite. The preconditioner K is always assumed to be Hermitian positive definite. Below, $\mathcal{E}$ is no longer assumed to be real.

2. The estimate by Samokish.

The first convergence result for the iterations (1.3) in a Hilbert space was obtained in 1958 by B. A. Samokish[3] [14]. For the reader's convenience, this result is reproduced below in the nonasymptotic form and in the finite-dimensional (Euclidean) case.

THEOREM 2.1. *Let κ_1 and κ_{-1} be, respectively, the smallest positive and largest eigenvalue of $K(L-\lambda_1)$. Denote $\epsilon_i = \sqrt{\|K\|(\lambda_1^i - \lambda_1)}$, where $\lambda_1^i = \lambda(u_1^i)$ and the sequence u_1^i is generated by* (1.3) *with locally optimal τ_i. If*

$$\tau(\sqrt{\kappa_{-1}} + \epsilon_i)\epsilon_i < 1, \tag{2.1}$$

where $\tau = 2/(\kappa_1 + \kappa_{-1})$, then

$$\lambda_1^{i+1} - \lambda_1 \le \left(\frac{\gamma_1 + \tau\sqrt{\kappa_{-1}}\epsilon_i}{1 - \tau(\sqrt{\kappa_{-1}} + \epsilon_i)\epsilon_i} \right)^2 (\lambda_1^i - \lambda_1), \tag{2.2}$$

where

$$\gamma_1 = \frac{1-\xi_1}{1+\xi_1}, \quad \xi_1 = \frac{\kappa_1}{\kappa_{-1}}. \tag{2.3}$$

Proof. Denote $L_1 = L - \lambda_1$, $u = u_1^i$, $\lambda = \lambda_1^i$, $r = (L-\lambda)u$, $u' = u_1^{i+1}$, $\lambda' = \lambda_1^{i+1}$ and $\epsilon = \epsilon_i$, and assume $\|u\| = 1$. We have

$$u' = u - \tau K r = (1 - \tau K L_1)u + \tau(\lambda - \lambda_1)Ku. \tag{2.4}$$

It is easy to see that

$$\|(1 - \tau K L_1)v\|_{L_1}^2 \le \gamma_1 \|v\|_{L_1}^2 \tag{2.5}$$

[3] Essentially the same estimate was later reproduced independently in [10] in the context of the generalized Davidson method and in the finite-dimensional case (cf. [11]).

and that $\|v\|_{L_1}^2 = (\lambda(v) - \lambda_1)\|v\|^2$. Hence,

$$\begin{aligned}\|u'\|_{L_1} = \sqrt{\lambda' - \lambda_1}\|u'\| &\leq \gamma_1\|u\|_{L_1} + \tau(\lambda - \lambda_1)\|Ku\|_{L_1} \\ &\leq \gamma_1\sqrt{\lambda - \lambda_1} + \tau(\lambda - \lambda_1)\sqrt{\kappa_{-1}}\|u\|_K \\ &\leq \gamma_1\sqrt{\lambda - \lambda_1} + \tau(\lambda - \lambda_1)\sqrt{\kappa_{-1}\|K\|} \\ &= (\gamma_1 + \tau\sqrt{\kappa_{-1}}\epsilon)\sqrt{\lambda - \lambda_1}.\end{aligned}$$

Since $\|u'\| \geq 1 - \tau\|Kr\|$, it remains to estimate $\|Kr\|$. We have $r = (L - \lambda)u = L_1 u - (\lambda - \lambda_1)u$, and hence

$$\begin{aligned}\|Kr\| \leq \|KL_1u\| + (\lambda - \lambda_1)\|Ku\| &\leq \sqrt{\|K\|}\|L_1u\|_K + (\lambda - \lambda_1)\|K\| \\ &\leq \sqrt{\kappa_{-1}\|K\|}\|u\|_{L_1} + (\lambda - \lambda_1)\|K\| = \sqrt{\kappa_{-1}}\epsilon + \epsilon^2,\end{aligned}$$

and we arrive at (2.2). □

The above simple estimate, obtained by Samokish in the asymptotic form, is a very remarkable one. First of all, it is the first ever convergence estimate for an iterative method for eigenvalue problems using what is essentially the preconditioning technique in all but name.[4] At the same time, as we will see in the next section, until quite recently it remained asymptotically more accurate than any of the estimates for the PSD method that have been obtained ever since. Another remarkable feature of (2.2) is its similarity to the convergence estimate for the PSD method for linear systems. Indeed, in the case of the degenerate system $(L - \lambda_1)u = f$ we have

$$\|u^{i+1} - u\|_{L-\lambda_1} \leq \gamma_1\|u^i - u\|_{L-\lambda_1}, \tag{2.6}$$

whereas in the case of the eigenvalue problem $Lu = \lambda u$ we have (cf. the proof of Theorem 2.1):

$$\|u_1^{i+1} - u_1\|_{L-\lambda_1} \leq \gamma_1\|u_1^i - u_1\|_{L-\lambda_1} + \mathcal{O}\left(\|u_1^i - u_1\|_{L-\lambda_1}^2\right).$$

Yet another remarkable fact about (2.2) is that until quite recently it remained the only estimate for the preconditioned steepest descent that predicted superlinear convergence for certain preconditioners. Indeed, consider the preconditioner K given by

$$K = (\alpha P_1 + (L - \lambda_1))^{-1}, \tag{2.7}$$

where $\alpha > 0$ and P_1 is the orthogonal projector onto the invariant subspace corresponding to λ_1. We have $\xi_1 = 1$ and hence $\gamma_1 = 0$, and by Theorem 2.1 the convergence of (1.3) is superlinear. We note that the PSD with the preconditioner given by (2.7) (which might be called quasi-optimal since it delivers superlinear convergence) is closely related to the Jacobi–Davidson method [15], whereby $g = Kr_1^i$ is computed by approximately solving the system

$$L_i g \equiv (\alpha P_{1,i} + (1 - P_{1,i})(L - \lambda_1^i)(1 - P_{1,i}))g = r_1^i,$$

where $P_{1,i}$ is the orthogonal projector onto $\operatorname{span}\{u_1^i\}$ and $\alpha > 0$ is an arbitrary parameter.[5] It is easy to see that L_i^{-1} converges to K given by (2.7) as $i \to \infty$,

[4] Strictly speaking the term "preconditioning" should not be used in the context of [14], where the steepest descent is applied to an unbounded operator in a Hilbert space.

[5] It is not difficult to see that g does not depend on α and α may actually be excluded from the computations; this parameter is introduced here merely to make the operator L_i positive definite.

which, in particular, implies that if this system is solved exactly, then the Jacobi–Davidson method converges superlinearly (for the case of inexact solution, see, e.g., [9, 11]; see also [10] where the above is used as a motivation for rediscovering the Samokish estimate).

3. Some other estimates. We have seen above that obtaining an asymptotical convergence estimate for PSD is a relatively simple task. As mentioned in the introduction, obtaining nonasymptotic convergence estimates proved to be much more difficult. Until quite recently one of the best available nonasymptotic convergence results for this method has been the following estimate by D'yakonov and Orekhov [3]: assuming that L is positive definite and that $\lambda_1^i < \lambda_2$, we have

$$\lambda_1^{i+1} - \lambda_1 \le q(\lambda_1^i)(\lambda_1^i - \lambda_1), \tag{3.1}$$

where

$$q(\lambda) = \max\left\{1 - \xi\frac{\lambda_2 - \lambda_1}{\lambda_2}, \frac{1 - \xi\frac{\lambda_2 - \lambda_1^i}{\lambda_2}}{1 + \xi\frac{\lambda_2 - \lambda_1^i}{\lambda_2}\frac{\lambda_1^i - \lambda_1}{\lambda_1}}\right\}, \quad \xi = cond(KL)^{-1}.$$

Compared to (2.2), the above estimate is asymptotically less accurate. Indeed, in the subspace orthogonal to the invariant subspace corresponding to λ_1 we have

$$\frac{\lambda_2 - \lambda_1}{\lambda_2}L \le L - \lambda_1 \le \frac{\lambda_{-1} - \lambda_1}{\lambda_{-1}}L.$$

Hence, $\xi_1 \ge \tilde{\xi}_1$, where

$$\tilde{\xi}_1 = \xi\frac{\lambda_2 - \lambda_1}{\lambda_2}\frac{\lambda_{-1}}{\lambda_{-1} - \lambda_1} \ge \xi\frac{\lambda_2 - \lambda_1}{\lambda_2}, \tag{3.2}$$

and we see that the error reduction factor in (2.2) is asymptotically less than a square of that in (3.1).

Recent papers [7, 8, 6] consider a version of PSD, referred to as PINVIT (for Preconditioned INVerse ITerations), whereby $\tau_i = 1$ and K is scaled so that the smallest and largest eigenvalues of KL are, respectively, $1 - \gamma$ and $1 + \gamma$ for some $\gamma < 1$. A sharp convergence estimate for PINVIT is given in [8]; the explicit form of this estimate is fairly cumbersome, but the two-dimensional analysis in section 5 of [7] implies that it can be formulated as follows:

$$\Delta_k(\lambda^{i+1}) \le q(\gamma, \lambda_k, \lambda_{k+1}, \lambda^i)^2\Delta_k(\lambda^i), \quad \Delta_k(\lambda) = \frac{\lambda - \lambda_k}{\lambda_{k+1} - \lambda}, \tag{3.3}$$

where k is such that $\lambda_k \le \lambda^i < \lambda_{k+1}$ and $q(u, v, w, t)$ is the solution of the equation

$$q = \frac{v}{w} + u\left(1 - \frac{v}{w}\right)\sqrt{\cos^2\alpha + q^2\sin^2\alpha}, \quad \sin^2\alpha = \frac{w}{t}\frac{t - v}{w - v}.$$

If $k = 1$, then, according to (3.3), the error in λ_1 is reduced by a factor that approaches $(\gamma + (1-\gamma)\lambda_1/\lambda_2)^2$ as λ^i approaches λ_1. We observe that the above asymptotic factor is noticeably smaller than that by (3.1), but is still greater than that in (2.2).

The next estimate of this section is a particular case of the estimate of Theorem 3.3 in [5]. Assuming that $\lambda_1^i < \lambda_2$ we have

$$\Delta_1(\lambda_1^{i+1}) \le \left(1 - (1 - \tilde{\gamma}_1^2)\max\{\xi, g_i\}\right)\Delta_1(\lambda_1^i), \quad \tilde{\gamma}_1 = \frac{1 - \tilde{\xi}_1}{1 + \tilde{\xi}_1}, \tag{3.4}$$

where $\Delta_1(\lambda)$ is the same as in (3.3), $\tilde{\xi}_1$ is given by (3.2), and

$$g_i = (1 + \eta\tilde{\gamma}_1^2\epsilon_i)^{-1}\left(1 + \frac{\eta^2\xi\epsilon_i}{4}\right)^{-1}\left(1 + \eta\sqrt{\frac{\tilde{\gamma}_1^2\epsilon_i}{\xi + (\tilde{\gamma}_1^2 - \xi)\epsilon_i}}\right)^{-1},$$

where $\eta = \xi^{-1} - 1$ and $\epsilon_i = (\lambda_2/\lambda_1^i)(\lambda_1^i - \lambda_1)/(\lambda_2 - \lambda_1)$. By the above estimate, the error in λ_1 is reduced after each iteration by a factor that approaches $\tilde{\gamma}_1^2$ as $i \to \infty$. This asymptotic factor is smaller than those in the previous two estimates. Still, the estimate (3.4) is less accurate than (2.2) in the sense that $\tilde{\gamma}_1 \geq \gamma_1$ and, furthermore, $\tilde{\gamma}_1$ is always positive, whereas $\gamma_1 = 0$; e.g., for K given by (2.7).

The last convergence result of this section approaches the estimate by Samokish much closer than any of the above: assuming $\lambda_1^i < \lambda_2$, and denoting by $\kappa_{-1,\alpha}$ the largest eigenvalue of $K(\alpha P_1 + (L - \lambda_1))$, we have [11]

$$\lambda_1^{i+1} - \lambda_1 \leq \gamma_{1,i,\alpha}^2(\lambda_1^i - \lambda_1), \tag{3.5}$$

where

$$\gamma_{1,i,\alpha} = \frac{1 - \xi_{1,\alpha}(1 - \epsilon_i)}{1 + \xi_{1,\alpha}(1 - \epsilon_i)}, \quad \xi_{1,\alpha} = \frac{\kappa_1}{\kappa_{-1,\alpha}},$$

$$\epsilon_i = \frac{(\sigma + 4)\delta_i}{\sigma\delta_i + (1 + \delta_i)^2}, \quad \delta_i = \frac{\lambda_1^i - \lambda_1}{\lambda_2 - \lambda_1}, \quad \sigma = \frac{\lambda_2 - \lambda_1}{\alpha}.$$

We observe that

$$\gamma_{1,i,\alpha} = \gamma_{1,\alpha} + \mathcal{O}\left(\lambda_1^i - \lambda_1\right), \quad \gamma_{1,\alpha} = \frac{1 - \xi_{1,\alpha}}{1 + \xi_{1,\alpha}},$$

and hence if K is the pseudoinverse of $L - \lambda_1$, in which case $\kappa_1 = \kappa_{-1,\alpha} = 1$ and $\gamma_{1,\alpha} = 0$, then $\lambda_1^{i+1} = \mathcal{O}\left((\lambda_1^i - \lambda_1)^3\right)$, i.e., the convergence is cubic (in this respect (3.5) is even more accurate than (2.2), which predicts quadratic convergence). Furthermore, substituting $\sqrt{\lambda_1^i - \lambda_1}$ for α, we obtain $\gamma_{1,i,\alpha} = \gamma_1 + \mathcal{O}\left(\sqrt{\lambda_1^i - \lambda_1}\right)$, i.e., the estimate (3.5) is asymptotically equivalent to (2.2).

4. Asymptotical sharpness. The estimates (2.2), (3.4), and (3.5) are rather cumbersome and difficult to grasp—however, their asymptotics is fairly simple. Indeed, it is not difficult to see that (2.2), (3.4), and (3.5) imply, respectively, the following three estimates:

$$\lambda_1^{i+1} - \lambda_1 \leq \left(\gamma_1 + \mathcal{O}\left(\sqrt{\lambda_1^i - \lambda_1}\right)\right)^2(\lambda_1^i - \lambda_1), \tag{4.1}$$

$$\lambda_1^{i+1} - \lambda_1 \leq \left(\tilde{\gamma}_1 + \mathcal{O}\left(\sqrt{\lambda_1^i - \lambda_1}\right)\right)^2(\lambda_1^i - \lambda_1), \tag{4.2}$$

$$\lambda_1^{i+1} - \lambda_1 \leq \left(\gamma_{1,\alpha} + \mathcal{O}\left(\lambda_1^i - \lambda_1\right)\right)^2(\lambda_1^i - \lambda_1). \tag{4.3}$$

We remind the reader that

$$\gamma_1 = \frac{1 - \xi_1}{1 + \xi_1}, \quad \tilde{\gamma}_1 = \frac{1 - \tilde{\xi}_1}{1 + \tilde{\xi}_1}, \quad \gamma_{1,\alpha} = \frac{1 - \xi_{1,\alpha}}{1 + \xi_{1,\alpha}}, \tag{4.4}$$

where ξ_1 is the ratio of the smallest positive to the largest eigenvalue of $K(L-\lambda_1)$, $\tilde{\xi}_1$ is given by (3.2) with $\xi = cond(KL)^{-1}$, and $\xi_{1,\alpha}$ is the ratio of the smallest positive eigenvalue of $K(L-\lambda_1)$ to the largest eigenvalue of $K(\alpha P_1 + (L-\lambda_1))$. In this section we show that the previous three estimates are asymptotically sharp, i.e., the main terms on the right-hand side cannot be made smaller without using additional information. We start with (4.1) and (4.3).

THEOREM 4.1. *For any* $0 \le \gamma < 1$ *there exist* u_1^i *and* K *such that* $\gamma_1 = \gamma$ *and*

$$\lambda_1^{i+1} - \lambda_1 = \left(\gamma_1 + \mathcal{O}\left(\lambda_1^i - \lambda_1\right)\right)^2 (\lambda_1^i - \lambda_1).$$

Proof. Let $K = (\omega P_1 + (L-\lambda_1)P_2 + \kappa^{-1}(L-\lambda_1)P)^{-1}$, where $\omega > 0$, $P_i = (\cdot, u_i)u_i$, $i = 1,2$, $P = 1 - P_1 - P_2$, and $\kappa \ge 1$. In the notation of Theorem 2.1, we obviously have $\kappa_1 = 1$ and $\kappa_{-1} = \kappa$, and hence $\xi_1 = \kappa^{-1}$ and $\gamma_1 = (\kappa-1)/(\kappa+1)$. To simplify the notation, let us denote $u = u_1^i$, $\lambda = \lambda_1^i$, $\tilde{u} = u_1^{i+1}$, and $\tilde{\lambda} = \lambda_1^{i+1}$. Let $u = u_1 + x_2 u_2 + x_3 u_3$, where x_2 and x_3 are real numbers. Elementary calculations show that

$$\tau = \frac{(\lambda_2-\lambda_1)x_2^2 + \kappa(\lambda_3-\lambda_1)x_3^2}{(\lambda_2-\lambda_1)x_2^2 + \kappa^2(\lambda_3-\lambda_1)x_3^2} + \delta_\tau \equiv \tau_* + \delta_\tau,$$

where $|\delta_\tau| \le c_\tau \epsilon^2$, $\epsilon = \sqrt{x_2^2 + x_3^2}$, and c_τ does not depend on x_2 and x_3; we observe that $\kappa^{-1} \le \tau_* \le 1$. Thus, $\tilde{u} = \tilde{x}_1 u_1 + \tilde{x}_2 u_2 + \tilde{x}_3 u_3$, where $\tilde{x}_1 = 1 - \tau(\lambda_1 - \lambda)/\omega$ and

$$\begin{aligned}
\tilde{x}_2 &= \left(1 - \tau\frac{\lambda_2-\lambda}{\lambda_2-\lambda_1}\right)x_2 = (1 - \tau_* + \delta_2)x_2 \\
&= \left(\frac{\kappa(\kappa-1)(\lambda_3-\lambda_1)x_3^2}{(\lambda_2-\lambda_1)x_2^2 + \kappa^2(\lambda_3-\lambda_1)x_3^2} + \delta_2\right)x_2, \\
\tilde{x}_3 &= \left(1 - \tau\kappa\frac{\lambda_3-\lambda}{\lambda_3-\lambda_1}\right)x_3 = (1 - \kappa\tau_* + \delta_3)x_3 \\
&= \left(\frac{(1-\kappa)(\lambda_2-\lambda_1)x_2^2}{(\lambda_2-\lambda_1)x_2^2 + \kappa^2(\lambda_3-\lambda_1)x_3^2} + \delta_3\right)x_3,
\end{aligned}$$

where $|\delta_i| \le c_i\epsilon^2$ and c_i does not depend on x_2 and x_3, $i = 2,3$. Now, if $(\lambda_2-\lambda_1)x_2^2 = \kappa(\lambda_3-\lambda_1)x_3^2$, then $1 - \tau_* = \kappa\tau_* - 1 = (\kappa-1)/(\kappa+1)$ and we have

$$\tilde{x}_2 = \left(\frac{\kappa-1}{\kappa+1} + \delta_2\right)x_2, \quad \tilde{x}_3 = -\left(\frac{\kappa-1}{\kappa+1} - \delta_3\right)x_3,$$

and hence

$$\begin{aligned}
\tilde{\lambda} = \lambda(\tilde{u}) = \frac{\lambda_1 + \lambda_2\tilde{x}_2^2 + \lambda_3\tilde{x}_3^2}{\tilde{x}_1^2 + \tilde{x}_2^2 + \tilde{x}_3^2} &= \lambda_1 + \frac{(\lambda_2-\lambda_1)\tilde{x}_2^2 + (\lambda_3-\lambda_1)\tilde{x}_3^2}{1+\epsilon^2} \\
&= \lambda_1 + \left(\frac{\kappa-1}{\kappa+1} + \mathcal{O}\left(\epsilon^2\right)\right)^2 \frac{(\lambda_2-\lambda_1)x_2^2 + (\lambda_3-\lambda_1)x_3^2}{1+\epsilon^2}
\end{aligned}$$

$$= \lambda_1 + \left(\frac{\kappa-1}{\kappa+1} + \mathcal{O}\left(\epsilon^2\right)\right)^2 (\lambda - \lambda_1) = \lambda_1 + \left(\gamma_1 + \mathcal{O}\left(\lambda-\lambda_1\right)\right)^2(\lambda-\lambda_1). \qquad \square$$

THEOREM 4.2. *For any* $0 \le \gamma < 1$ *there exist* u_1^i *and* K *such that* $\gamma_{1,\alpha} = \gamma$ *and*

$$\lambda_1^{i+1} - \lambda_1 = \left(\gamma_{1,\alpha} + \mathcal{O}\left(\lambda_1^i - \lambda_1\right)\right)^2 (\lambda_1^i - \lambda_1).$$

Proof. We just use the proof of Theorem 4.1 with $\omega = \alpha$, observing that $\kappa_1 = 1$ and $\kappa_{-1,\alpha} = \kappa$, and hence $\xi_{1,\alpha} = \kappa^{-1}$ and $\gamma_{1,\alpha} = (\kappa - 1)/(\kappa + 1)$. □

THEOREM 4.3. *Let L be positive definite. For any $0 < \xi \leq 1$ there exist u_1^i and K such that $cond(KL) = \xi^{-1}$ and*

$$\lambda_1^{i+1} - \lambda_1 = \left(\tilde{\gamma}_1 + \mathcal{O}\left(\lambda_1^i - \lambda_1\right)\right)^2 \left(\lambda_1^i - \lambda_1\right). \tag{4.5}$$

Proof. Using the notation of the proof of Theorem 4.1, let us take $K = L^{-1}(1 - P) + \kappa L^{-1}P$, where $\kappa \geq 1$. The smallest eigenvalue of KL is obviously 1 and the largest is κ; hence $\xi = \kappa^{-1}$. Let $u = u_1 + x_2u_2 + x_{-1}u_{-1}$, where u_{-1} is an eigenvector of L corresponding to its largest eigenvalue λ_{-1}, and x_2 and x_{-1} are real numbers. Elementary calculations show that the locally optimal value of τ is

$$\tau = \frac{\frac{(\lambda_2-\lambda_1)^2}{\lambda_2}x_2^2 + \kappa\frac{(\lambda_{-1}-\lambda_1)^2}{\lambda_{-1}}x_{-1}^2}{\frac{(\lambda_2-\lambda_1)^3}{\lambda_2^2}x_2^2 + \kappa^2\frac{(\lambda_{-1}-\lambda_1)^3}{\lambda_{-1}^2}x_{-1}^2} + \mathcal{O}\left(\epsilon^2\right),$$

where $\epsilon = \sqrt{x_2^2 + x_{-1}^2}$. Thus, $\tilde{u} = \tilde{x}_1u_1 + \tilde{x}_2u_2 + \tilde{x}_{-1}u_{-1}$, where $\tilde{x}_1 = 1 - \tau(\lambda_1 - \lambda)/\lambda_1$ and

$$\tilde{x}_2 = \left(1 - \tau\frac{\lambda_2 - \lambda}{\lambda_2}\right)x_2 = \left(\frac{\left(\kappa\frac{\lambda_{-1}-\lambda_1}{\lambda_{-1}} - \frac{\lambda_2-\lambda_1}{\lambda_2}\right)\kappa\frac{(\lambda_{-1}-\lambda_1)^2}{\lambda_{-1}}x_{-1}^2}{\frac{(\lambda_2-\lambda_1)^3}{\lambda_2^2}x_2^2 + \kappa^2\frac{(\lambda_{-1}-\lambda_1)^3}{\lambda_{-1}^2}x_{-1}^2} + \mathcal{O}\left(\epsilon^2\right)\right)x_2$$

$$\tilde{x}_{-1} = \left(1 - \tau\kappa\frac{\lambda_{-1} - \lambda}{\lambda_{-1}}\right)x_{-1} = \left(\frac{\left(\frac{\lambda_2-\lambda_1}{\lambda_2} - \kappa\frac{\lambda_{-1}-\lambda_1}{\lambda_{-1}}\right)\frac{(\lambda_2-\lambda_1)^2}{\lambda_2}x_2^2+}{\frac{(\lambda_2-\lambda_1)^3}{\lambda_2^2}x_2^2 + \kappa^2\frac{(\lambda_{-1}-\lambda_1)^3}{\lambda_{-1}^2}x_{-1}^2} + \mathcal{O}\left(\epsilon^2\right)\right)x_{-1}.$$

If x_2 and x_{-1} are such that

$$\frac{(\lambda_2 - \lambda)^2}{\lambda_2}x_2^2 = \kappa\frac{(\lambda_{-1} - \lambda)^2}{\lambda_{-1}}x_{-1}^2,$$

then

$$\tilde{x}_2 = \left(\frac{\kappa\frac{\lambda_{-1}-\lambda_1}{\lambda_{-1}} - \frac{\lambda_2-\lambda_1}{\lambda_2}}{\frac{\lambda_2-\lambda_1}{\lambda_2} + \kappa\frac{\lambda_{-1}-\lambda_1}{\lambda_{-1}}} + \mathcal{O}\left(\epsilon^2\right)\right)x_2 = \left(\tilde{\gamma}_1 + \mathcal{O}\left(\epsilon^2\right)\right)x_2,$$

$$\tilde{x}_{-1} = \left(\frac{\frac{\lambda_2-\lambda_1}{\lambda_2} - \kappa\frac{\lambda_{-1}-\lambda_1}{\lambda_{-1}}}{\frac{\lambda_2-\lambda_1}{\lambda_2} + \kappa\frac{\lambda_{-1}-\lambda_1}{\lambda_{-1}}} + \mathcal{O}\left(\epsilon^2\right)\right)x_{-1} = -\left(\tilde{\gamma}_1 + \mathcal{O}\left(\epsilon^2\right)\right)x_{-1},$$

and we arrive at (4.5). □

Remark 1. The fact that (4.2) is sharp does not contradict the fact that it is less accurate than (4.1) because (4.2) uses less accurate information about K and L; instead of ξ_1, the ratio of the smallest positive to the largest eigenvalue of $K(L - \lambda_1)$ that is used in (4.1), the estimate (4.2) uses its lower bound $\tilde{\xi}_1$.

5. Sharp upper bounds for the asymptotic convergence factor. For any nonzero u_1^0 that is not an eigenvector of L corresponding to λ_1 the following quantity can be introduced:

$$q_1(u_1^0) = \varlimsup_{i\to\infty} q_{1,i}(u_1^0), \quad q_{1,i}(u_1^0) \equiv \left(\frac{\lambda_1^i - \lambda_1}{\lambda_1^0 - \lambda_1}\right)^{\frac{1}{i}},$$

where u_1^i is the sequence generated by (1.3) starting with the initial guess u_1^0, and $\lambda_1^i = \lambda(u_1^i)$. In this paper we refer to $q_1(u_1^0)$ as the *asymptotic convergence factor* (a.c.f.) for the sequence $\{u_1^i\}_{i=0}^{\infty}$. We observe that

$$q_{1,i}(u_1^0) = \left(\frac{\lambda_1^i - \lambda_1}{\lambda_1^{i-1} - \lambda_1} \frac{\lambda_1^{i-1} - \lambda_1}{\lambda_1^{i-2} - \lambda_1} \cdots \frac{\lambda_1^1 - \lambda_1}{\lambda_1^0 - \lambda_1} \right)^{\frac{1}{i}};$$

i.e., $q_{1,i}(u_1^0)$ is the geometrical average of the error reductions on the first i iterations (1.3). The estimates (4.1), (4.2), and (4.3) imply that γ_1^2, $\tilde{\gamma}_1^2$, and $\gamma_{1,\alpha}^2$ are upper bounds for the a.c.f. $q_1(u_1^0)$. In this section we show that these bounds are sharp in the sense that by a proper choice of K each of them can be made equal to the supremum of $q_1(u_1^0)$ taken over all initial guesses u_1^0 for which u_1^i converge to an eigenvector corresponding to λ_1. Below we denote the latter quantity by q_1, i.e.,

$$q_1 \equiv \sup_{u_1^0 \in \mathcal{E} : \lambda(u_1^0) > \lambda_1, q_1(u_1^0) < 1} q_1(u_1^0). \tag{5.1}$$

THEOREM 5.1. *For any $\gamma \in [0, 1)$ there exists K such that $q_1 = \gamma_1^2 = \gamma^2$.*

Proof. First, let us show that $q_1 \le \gamma_1^2$. In view of the condition $q_1(u_1^0) < 1$ we only need to consider sequences u_1^i that converge to an eigenvector corresponding to λ_1. Denote by $\gamma_{1,i}^2$ the coefficient in front of $\lambda_1^i - \lambda_1$ in (2.2). For any sequence of u_1^i in focus there exists i_0 such that $\lambda_1^{i_0} < \lambda_2$ and $\gamma_{1,i_0} < 1$. Hence, using (2.2) for $i \ge i_0$ and the fact that $\lambda_1^{i+1} \le \lambda_1^i$, we have

$$\begin{aligned} \ln \frac{\lambda_1^i - \lambda_1}{\lambda_1^0 - \lambda_1} \le \ln \frac{\lambda_1^i - \lambda_1}{\lambda_1^{i_0} - \lambda_1} &\le \ln \prod_{j=i_0}^{i-1} \left(\frac{\gamma_1 + \tau\sqrt{\kappa_{-1}}\epsilon_j}{1 - \tau(\sqrt{\kappa_{-1}} + \epsilon_j)\epsilon_j} \right)^2 \\ &\le 2((i - i_0)\ln \gamma_1 + \delta_i) \end{aligned} \tag{5.2}$$

where

$$\delta_i = \sum_{j=i_0}^{i-1} (\ln(1 + \tau\sqrt{\kappa_{-1}}\epsilon_j/\gamma_1) - \ln(1 - \tau(\sqrt{\kappa_{-1}} + \epsilon_j)\epsilon_j)).$$

Denoting $a = \tau\sqrt{\kappa_{-1}}/\gamma_1$ and $b = \tau(\sqrt{\kappa_{-1}} + \epsilon_{i_0})$ and using elementary estimates $\ln(1 + x) \le x$ and $-\ln(1 - x) \le x(1 - \ln(1 - x))$ we obtain for δ_i the upper bound that does not depend on i:

$$\delta_i \le \sum_{j=i_0}^{i-1} (a\epsilon_j + b\epsilon_j(1 - \ln(1 - b\epsilon_j))) \le \frac{a}{1 - \gamma_{1,i_0}}\epsilon_{i_0} + \frac{b(1 - \ln(1 - b\epsilon_{i_0}))}{1 - \gamma_{1,i_0}}\epsilon_{i_0}.$$

Hence, dividing (5.2) by i and passing to the limit $i \to \infty$, we obtain $q_1(u_1^0) \le \gamma_1^2$.

Now, let us take the same K as in the proof of Theorem 4.1, in which case $\gamma_1 = \gamma$. If $\gamma_1 = \gamma = 0$, then $q_1(u_1^0) = 0$ for any sequence u_1^0 that converges to an eigenvector in $\mathcal{I}_1$ (cf. above), and hence $q_1 = 0$. Below we consider the case $\gamma_1 > 0$, i.e., $\kappa > 1$. Let $u_1^0 = u_1 + x_{2,0}u_2 + x_{3,0}u_3$, where $x_{2,0}$ and $x_{3,0}$ are real numbers such that

$$(\lambda_2 - \lambda_1)x_{2,0}^2 = \kappa(\lambda_3 - \lambda_1)x_{3,0}^2.$$

From the proof of Theorem 4.1 we see that $u_1^i = x_{1,i}u_1 + x_{2,i}u_2 + x_{3,i}u_3$ for some real $x_{1,i}, x_{2,i}$ and $x_{3,i}$, and that the ratio of $x_{2,i}^2 + x_{3,i}^2$ to $x_{1,i}^2$ decreases on each iteration. Let us rescale u_1^i so that $x_{1,i} = 1$. Denoting $\epsilon_i = \sqrt{x_{2,i}^2 + x_{3,i}^2}$, we have $x_{2,i} = \epsilon_i \cos\psi_i$ and $x_{3,i} = \epsilon_i \sin\psi_i$. Consulting the proof of Theorem 4.1, we observe that $u_1^{i+1} = x_{1,i+1}u_1 + x_{2,i+1}u_2 + x_{3,i+1}u_3$, where

$$\begin{aligned} x_{1,i+1} &= 1 + \delta_{1,i}, \\ x_{2,i+1} &= \left(\frac{\kappa(\kappa-1)(\lambda_3-\lambda_1)x_{3,i}^2}{(\lambda_2-\lambda_1)x_{2,i}^2 + \kappa^2(\lambda_3-\lambda_1)x_{3,i}^2} + \delta_{2,i} \right) x_{2,i}, \\ x_{3,i+1} &= \left(\frac{(1-\kappa)(\lambda_2-\lambda_1)x_{2,i}^2}{(\lambda_2-\lambda_1)x_{2,i}^2 + \kappa^2(\lambda_3-\lambda_1)x_{3,i}^2} + \delta_{3,i} \right) x_{3,i}, \end{aligned}$$

$|\delta_{j,i}| \le c\epsilon_i^2$, and c does not depend on ψ_i and ϵ_i. From the above relationships we obtain

$$\tan\psi_{i+1} = \frac{(1-\kappa)\tan^2\psi_0 + \delta_{3,i}(\tan^2\psi_0 + \kappa\tan^2\psi_i)}{(\kappa-1)\tan^2\psi_i + \delta_{2,i}(\tan^2\psi_0 + \kappa\tan^2\psi_i)} \tan\psi_i. \tag{5.3}$$

Hence, denoting $t_i = |\tan\psi_i/\tan\psi_0|$, we have

$$t_{i+1} = \left| \frac{1 - \frac{\delta_{3,i}}{\kappa-1}(1+\kappa t_i^2)}{1 + \frac{\delta_{2,i}}{\kappa-1}(\kappa + t_i^{-2})} \right| t_i^{-1}.$$

Recalling that $|\delta_{j,i}| \le c\epsilon_i^2$, and denoting $\varepsilon_i = c\epsilon_i^2/(\kappa-1)$, we observe that if $\varepsilon_i(\kappa + t_i^{-2}) < 1$, then

$$t_{i+1} \le \frac{1 + \varepsilon_i(1+\kappa t_i^2)}{1 - \varepsilon_i(\kappa + t_i^{-2})} t_i^{-1},$$

and if $\varepsilon_i(1+\kappa t_i^2) < 1$, then

$$t_{i+1}^{-1} \le \frac{1 + \varepsilon_i(\kappa + t_i^{-2})}{1 - \varepsilon_i(1+\kappa t_i^2)} t_i.$$

Assume that λ_1^0 is close enough to λ_1 so that $\gamma_{1,0} < 1$. Since $\gamma_{1,i} \le \gamma_{1,j}$ for $i > j$, we have $\lambda_1^i - \lambda_1 \le \gamma_{1,0}^{2i}(\lambda_1^0 - \lambda_1)$. Further,

$$\epsilon_i^2 = \tan^2(u_1^i, u_1) \le (\lambda_1^i - \lambda_1)/(\lambda_2 - \lambda_1^i)$$

(cf., e.g., [5]), and thus

$$\varepsilon = \sum_{i=0}^{\infty} \varepsilon_i = \frac{c}{\kappa-1}\sum_{i=0}^{\infty} \epsilon_i^2 = \sum_{i=0}^{\infty} \mathcal{O}\left(\lambda_1^i - \lambda_1\right) = \mathcal{O}\left(\lambda_1^0 - \lambda_1\right).$$

Hence, applying Lemma A.1, we have $\tan^2\psi_i = \tan^2\psi_0\,(1 + \mathcal{O}(\varepsilon))$ and

$$\begin{aligned} \tilde{x}_{2,i+1} &= \left(\frac{(\kappa-1)\tan^2\psi_i}{\tan^2\psi_0 + \kappa\tan^2\psi_i} + \delta_{2,i} \right) x_{2,i} = \left(\frac{\kappa-1}{\kappa+1} + \mathcal{O}(\varepsilon) \right) x_{2,i}, \\ \tilde{x}_{3,i+1} &= \left(\frac{(1-\kappa)\tan^2\psi_0}{\tan^2\psi_0 + \kappa\tan^2\psi_i} + \delta_{3,i} \right) x_{3,i} = \left(\frac{1-\kappa}{\kappa+1} + \mathcal{O}(\varepsilon) \right) x_{3,i}, \end{aligned}$$

and hence

$$\begin{aligned}\lambda_1^{i+1} - \lambda_1 &= \frac{(\lambda_2 - \lambda_1)\tilde{x}_{2,i+1}^2 + (\lambda_3 - \lambda_1)\tilde{x}_{3,i+1}^2}{\tilde{x}_{1,i+1}^2 + \tilde{x}_{2,i+1}^2 + \tilde{x}_{3,i+1}^2} \\ &= \left(\frac{\kappa - 1}{\kappa + 1} + \mathcal{O}(\varepsilon)\right)^2 \frac{(\lambda_2 - \lambda_1)x_{2,i}^2 + (\lambda_3 - \lambda_1)x_{3,i}^2}{1 + \mathcal{O}(\epsilon_i^2)} \\ &= (\gamma_1 + \mathcal{O}(\varepsilon))^2 \left(1 + \mathcal{O}\left(\epsilon_i^2\right)\right)(\lambda_1^i - \lambda_1) \\ &= (\gamma_1 + \mathcal{O}(\varepsilon))^2 (\lambda_1^i - \lambda_1) = (\gamma_1 + \mathcal{O}(\varepsilon))^{2(i+1)} (\lambda_1^0 - \lambda_1).\end{aligned}$$

Thus, $q_1 \geq (\gamma_1 - \delta)^2$, where $\delta = \mathcal{O}(\varepsilon) = \mathcal{O}\left(\lambda(u_1^0) - \lambda_1\right)$, provided that $\lambda(u_1^0) - \lambda_1$ is small enough and, by taking supremum over all such u_1^0, we obtain $q_1 = \gamma_1^2 = \gamma^2$. □

By introducing certain modifications into the above calculations, we obtain the respective results in terms of $\gamma_{1,\alpha}$ and $\tilde{\gamma}_1$.

THEOREM 5.2. *For any $\gamma \in [0, 1)$ there exists K such that $q_1 = \gamma_{1,\alpha}^2 = \gamma^2$.*

Proof. We just need to use the proof of Theorem 5.1 with $\omega = \alpha$ in the definition of K. □

THEOREM 5.3. *For any $\xi \in [0, 1)$ there exists K such that $q_1 = \tilde{\gamma}_1^2$, where $\tilde{\gamma}_1$ is given by* (4.4) *with $\tilde{\xi}_1 = \xi$.*

Proof. Let K be the same as in the proof of Theorem 4.3. Comparing the formulas for $\tilde{x}_2$ and $\tilde{x}_3$ in the proof of Theorem 4.1 with those for $\tilde{x}_2$ and $\tilde{x}_{-1}$ in the proof of Theorem 4.3, we observe that the latter formulas can be obtained by changing κ in the former as follows:

$$\kappa \to \kappa \frac{\lambda_2}{\lambda_2 - \lambda_1} \frac{\lambda_{-1} - \lambda_1}{\lambda_{-1}}.$$

It is not difficult to verify that in order to prove Theorem 5.3 we just need to change κ in the above way in the proof of Theorem 5.1. □

6. New sharp estimates. We have already seen in section 2 that with some preconditioners the convergence of PSD can be superlinear. Actually, there even exists a preconditioner that delivers one-step convergence: assuming that $\lambda_1^0 = \lambda(u_1^0) < \lambda_2$, let us consider

$$K = (\alpha P_1 + (L - \lambda_1^0)(1 - P_1))^{-1},$$

where $\alpha > 0$. Taking $\tau_0 = 1$, we have

$$\begin{aligned}u_1^1 &= u_1^0 - K(L - \lambda_1^0)u_1^0 = u_1^0 - K((L - \lambda_1^0)P_1 + (L - \lambda_1^0)(1 - P_1)) \\ &= u_1^0 - (\alpha P_1 + (L - \lambda_1^0)(1 - P_1))^{-1}((\lambda_1 - \lambda_1^0)P_1 + (L - \lambda_1^0)(1 - P_1)) \\ &= u_1^0 - \left(\frac{\lambda_1 - \lambda_1^0}{\alpha} P_1 + 1 - P_1\right) u_1^0 = \left(1 + \frac{\lambda_1^0 - \lambda_1}{\alpha}\right) P_1 u_1^0\end{aligned}$$

(note that the above assumption on λ_1^0 implies $P_1 u_1^0 \neq 0$), and hence $\lambda_1^1 = \lambda_1$. The result below shows that in the case of a general Hermitian positive K the reduction in the eigenvalue error on ith iteration can be estimated via the closeness of K to (a multiple of) $(\alpha P_1 + (L - \lambda_1^i)(1 - P_1))^{-1}$.

THEOREM 6.1. *Let P_1 be the orthogonal projector onto the invariant subspace of L corresponding to λ_1, and denote*

$$L_{\alpha,\lambda} = \alpha P_1 + (L - \lambda)(1 - P_1).$$

If $\lambda_1^i < \lambda_2$, then the following estimate is valid for iterations (1.3) *with the locally optimal choice of τ_i:*

$$\Delta(\lambda_1^{i+1}) \le \frac{\gamma_{\alpha,\lambda_1^i}^2}{1+\left(1-\gamma_{\alpha,\lambda_1^i}^2\right)\frac{\lambda_1^i-\lambda_1}{\alpha}}\Delta(\lambda_1^i), \tag{6.1}$$

where $\alpha > 0$,

$$\Delta(\lambda) = \frac{\lambda-\lambda_1}{\lambda_2-\lambda}, \quad \gamma_{\alpha,\lambda} = \frac{1-\xi_{\alpha,\lambda}}{1+\xi_{\alpha,\lambda}}, \quad \xi_{\alpha,\lambda} = \frac{\kappa_{1,\alpha,\lambda}}{\kappa_{-1,\alpha,\lambda}},$$

and $\kappa_{1,\alpha,\lambda}$ and $\kappa_{-1,\alpha,\lambda}$ are, respectively, the smallest and the largest eigenvalue of $KL_{\alpha,\lambda}$. The estimate (6.1) *is sharp in the sense that for any L, $\lambda \in [\lambda_1, \lambda_2)$, $\alpha > 0$, and $\gamma \in [0,1]$ there exist K and u_1^i such that $\gamma_{\alpha,\lambda_1^i} = \gamma$, $\lambda(u_1^i) = \lambda$, and both sides of* (6.1) *coincide.*

Proof. See section A.1. □

Note that if L is positive definite, then

$$\alpha_0 L \le \alpha P_1 + (L-\lambda)(1-P_1) \le \alpha^0 L,$$

where

$$\alpha_0 = \min\left\{\frac{\alpha}{\lambda_1}, \frac{\lambda_2-\lambda}{\lambda_2}\right\}, \quad \alpha^0 = \max\left\{\frac{\alpha}{\lambda_1}, \frac{\lambda_{-1}-\lambda}{\lambda_{-1}}\right\}.$$

The above relationship shows that $\xi_{\alpha,\lambda} \ge cond(KL)^{-1}\alpha_0/\alpha^0$.

The arbitrariness of the parameter α reflects the fact that one-step convergence discussed at the beginning of this section takes place for any nonzero value of α. Below we use two particular choices of α to obtain estimates that are sharp and at the same time rather simple in appearance.

Theorem 6.2. *Assuming that $\lambda_1^i \equiv \lambda(u_1^i) < \lambda_2$, the following estimate is valid for iterations* (1.3) *with the locally optimal choice of τ_i:*

$$\lambda_1^{i+1} - \lambda_1 \le \gamma_{\lambda,i}^2(\lambda_1^i - \lambda_1), \quad \gamma_{\lambda,i} = \frac{1-\xi_{\lambda,i}}{1+\xi_{\lambda,i}}, \tag{6.2}$$

where $\xi_{\lambda,i}$ is the inverse of the condition number of the operator

$$K((\lambda_2-\lambda_1^i)P_1 + (L-\lambda_1^i)(1-P_1)),$$

and P_1 is the orthogonal projector onto the invariant subspace of L corresponding to λ_1. The estimate (6.2) *is sharp in the sense that for any L, $\lambda \in [\lambda_1, \lambda_2)$ and $\gamma \in [0,1]$ there exist K and u_1^i such that $\gamma_{\lambda,i} = \gamma$, $\lambda_1^i \equiv \lambda(u_1^i) = \lambda$, and both sides of* (6.2) *coincide.*

Proof. For $\alpha = \lambda_2 - \lambda_1^i$ we have $(\lambda_1^i - \lambda_1)/\alpha = \Delta(\lambda_1^i)$. Hence, denoting $\tan^2\phi_i = \Delta(\lambda_1^i)$ and $\tan^2\phi_{i+1} = \Delta(\lambda_1^{i+1})$, we obtain

$$\tan^2\phi_{i+1} \le \frac{\gamma_{\lambda,i}^2\tan^2\phi_i}{1+\tan^2\phi_i - \gamma_{1,i}^2\tan^2\phi_i} = \frac{\gamma_{\lambda,i}^2\sin^2\phi_i}{1-\gamma_{\lambda,i}^2\sin^2\phi_i},$$

which implies $\sin^2\phi_{i+1} \le \gamma_{\lambda,i}^2\sin^2\phi_i$, i.e.,

$$\frac{\lambda_1^{i+1}-\lambda_1}{\lambda_2-\lambda_1} \le \gamma_{\lambda,i}^2\frac{\lambda_1^i-\lambda_1}{\lambda_2-\lambda_1}.$$

The sharpness of (6.2) follows trivially from the sharpness of (6.1). □

A different choice of α leads to the following estimate in terms of the inverses $\mu_j = 1/\lambda_j$ of the eigenvalues of L.

THEOREM 6.3. *In the notation and under the assumptions of Theorem* 6.2, *with the additional assumption that L is positive definite, the following estimate is valid for the inverses $\mu_j^i = 1/\lambda_j^i$ of the approximate eigenvalues generated by* (1.3)*:*

$$\mu_1 - \mu_1^{i+1} \leq \gamma_{\mu,i}^2(\mu_1 - \mu_1^i), \quad \gamma_{\mu,i} = \frac{1-\xi_{\mu,i}}{1+\xi_{\mu,i}}, \tag{6.3}$$

where $\xi_{\mu,i}$ is the inverse of the condition number of the operator

$$K\left(\frac{\mu_1^i - \mu_2}{\mu_1}P_1 + (\mu_1^i L - I)(1-P_1)\right),$$

and $\mu_j = 1/\lambda_j$. The estimate (6.3) *is sharp in the sense that for any L, $\mu \in [\mu_2, \mu_1)$, and $\gamma \in [0,1]$ there exist K and u_1^i such that $\gamma_{\mu,i} = \gamma$, $\mu_1^i \equiv \mu(u_1^i) \equiv 1/\lambda(u_1^i) = \mu$, and both sides of* (6.3) *coincide.*

Proof. For $\alpha = (\mu_1^i - \mu_2)/(\mu_1\mu_1^i)$ we have

$$\frac{\lambda_1^i - \lambda_1}{\alpha} = \frac{\mu_1 - \mu_1^i}{\mu_1^i - \mu_2}.$$

Hence, substituting the above α into (6.1) we arrive at (6.3) in the same way as in the proof of Theorem 6.2. Again, the sharpness of (6.3) follows from the sharpness of (6.1). □

Since

$$(\mu_1^i - \mu_2)L \leq \frac{\mu_1^i - \mu_2}{\mu_1}P_1 + (\mu_1^i L - I)(1-P_1) \leq (\mu_1^i - \mu_{-1})L,$$

the above result has the following corollary.

COROLLARY 6.4. *In the notation and under the assumptions of Theorem* 6.3, *the following estimate is valid:*

$$\mu_1 - \mu_1^{i+1} \leq \tilde{\gamma}_{\mu,i}^2(\mu_1 - \mu_1^i), \tag{6.4}$$

where

$$\tilde{\gamma}_{\mu,i} = \frac{1-\tilde{\xi}_{\mu,i}}{1+\tilde{\xi}_{\mu,i}}, \quad \tilde{\xi}_{\mu,i} = \xi\frac{\mu_1^i - \mu_2}{\mu_1^i - \mu_{-1}},$$

and ξ is the inverse of the condition number of KL.

Let us compare the new estimates with those discussed previously in this paper.

Compared with the original asymptotic estimate by Samokish, the new estimates have the obvious advantage of being nonasymptotic, i.e., not containing unknown asymptotically insignificant terms. Compared with the estimate (2.2) reproducing the estimate by Samokish in nonasymptotic form, the new estimates have the advantage of being valid under a weaker assumption $\lambda_1^i < \lambda_2$ as compared to (2.1). Yet another advantage of (6.2) and (6.3) is the fact that these estimates are sharp, i.e., cannot be further improved without using additional information. At the same time, it should be admitted that the main asymptotic term in (2.2) is generally smaller than those in the new estimates.

Turning to the estimates (3.3) and (3.4), we observe first that they should be compared with the estimate (6.4), which is given in the same terms. It is not difficult to verify that the main asymptotic term in (6.4) coincides with that in (3.4) and is smaller than that in (3.3). At the same time, (6.4) is the simplest of the three estimates at hand.

7. Estimates for block versions of PSD. The block PSD (BPSD) methods combine the idea of the preconditioned steepest descent with the Rayleigh–Ritz method in the following manner: the new subspace $\mathcal{H}^{i+1}$ spans the first n Ritz vectors (enumerated in the ascending order of the corresponding Ritz values) in a trial subspace that contains the subspace

$$\mathcal{H}^{i+\frac{1}{2}} = \operatorname{span}\{u_j^i - \tau_{ij} K r_j^i\}_{j=1}^n, \tag{7.1}$$

where $u_1^i, \ldots, u_n^i$ are the Ritz vectors of L in $\mathcal{H}^i$ corresponding to the Ritz values λ_j^i, and $r_j^i = r(u_j^i) = Lu_j^i - \lambda_j^i u_j^i$. We remind the reader that the Ritz values and vectors in a subspace $\mathcal{V}$, denoted below by $\lambda_j(\mathcal{V})$ and $u_j(\mathcal{V})$, are the eigenpairs of the projection of L onto $\mathcal{V}$, i.e., of the problem

$$(Lu_j(\mathcal{V}), v) = \lambda_j(\mathcal{V})(u_j(\mathcal{V}), v) \quad \forall\, v \in \mathcal{V}. \tag{7.2}$$

In [1, 6] a version of the BPSD method is studied where $\tau_{ij} = 1$ and $\mathcal{H}^{i+1} = \mathcal{H}^{i+\frac{1}{2}}$; below we refer to this version as "the simple BPSD." In [2], $\mathcal{H}^{i+1}$ is defined in the same way but with each τ_{ij} being locally optimal, i.e., minimizing $\lambda(u_j^i - \tau_{ij} K r_j^i)$. In yet another BPSD method, $\mathcal{H}^{i+1}$ is defined as follows:

$$\mathcal{H}^{i+1} = \operatorname{span}\{u_j(\mathcal{H}_{lo}^{i+\frac{1}{2}})\}_{j=1}^n, \quad \mathcal{H}_{lo}^{i+\frac{1}{2}} = \mathcal{H}^i + \operatorname{span}\{K r_j^i\}_{j=1}^n. \tag{7.3}$$

We note that $\mathcal{H}_{lo}^{i+\frac{1}{2}} \supset \mathcal{H}^{i+\frac{1}{2}}$ for any τ_{ij}, and hence, by the minimax principle, the Ritz values in $\mathcal{H}_{lo}^{i+\frac{1}{2}}$ are not greater, and hence not further away from the exact ones, than those in any $\mathcal{H}^{i+\frac{1}{2}}$ of the form (7.1). Hence, (7.3) might be called locally optimal BPSD.

In [1] one can find the following asymptotic convergence estimate[6] for the simple BPSD:

$$\lambda_k^i - \lambda_k \le c_\varepsilon \left(q(\gamma, \lambda_k, \lambda_{n+1}, \lambda_k) + \varepsilon\right)^{2i}, \tag{7.4}$$

where $q(u, v, w, t)$ is the same as in (3.3), and ε is an arbitrary small positive real number. By considering the initial subspace

$$\mathcal{H}^0 = \operatorname{span}\{u_1, \ldots, u_{j-1}, u_j + \tau u_{n+1}, u_{j+1}, \ldots, u_n\} \tag{7.5}$$

it is not difficult to verify that (7.4) is asymptotically (as $\sin(\mathcal{H}^0, \mathcal{I}_n) \to 0$, where $\mathcal{I}_n = \operatorname{span}\{u_1, \ldots, u_n\}$) sharp in the sense that $q(\gamma, \lambda_k, \lambda_{n+1}, \lambda_k)$ cannot generally be replaced with any smaller value. In [6] one can find the following much more pessimistic estimate that essentially coincides with (3.3):

$$\Delta_k(\lambda_j^{i+1}) \le q(\gamma, \lambda_k, \lambda_{k+1}, \lambda_j^i)^2 \Delta_k(\lambda^i), \tag{7.6}$$

[6] The cited paper also has a nonasymptotic estimate, which is more cumbersome and is not reproduced here for simplicity of presentation.

where k is such that $\lambda_k \leq \lambda_j^i < \lambda_{k+1}$ and $q(u, v, w, t)$ and $\Delta_k(\lambda)$ are the same as in (3.3). Surprisingly, the above estimate is sharp, which appears to be in contradiction with the previous estimate. The explanation of this "contradiction" is actually quite simple: (7.4) assumes that $\mathcal{H}^0$ is sufficiently close to $\mathcal{I}_n$, whereas (7.6) makes no assumptions about the subspace $\mathcal{H}^i$ (cf. [1]), and it is not difficult to verify that in the case

$$\mathcal{H}^i = \operatorname{span}\{u_1, \ldots, u_{j-1}, u_k + \tau u_{k+1}, u_{k+2}, \ldots, u_{n+1}\}, \tag{7.7}$$

the factor $q(\gamma, \lambda_k, \lambda_{k+1}, \lambda_j^i)^2$ cannot be replaced by any smaller value. Note also that if we define the a.c.f. for λ_k as

$$q_k(\mathcal{H}^0) = \overline{\lim_{i\to\infty}} \left(\frac{\lambda_k^i - \lambda_k}{\lambda_k^0 - \lambda_k} \right)^{\frac{1}{i}}, \tag{7.8}$$

then $q(\gamma, \lambda_k, \lambda_{k+1}, \lambda_k)^2$ is an upper bound for $q_k(\mathcal{H}^0)$ taken over all $\mathcal{H}^0$ for which kth Ritz value λ_k^i converges to λ_k, as can be seen from the estimate (7.6) and the example of the initial subspace given by (7.7). However, if we reduce the set of initial subspaces $\mathcal{H}^0$ to those for which $\mathcal{H}^i$ converges to $\mathcal{I}_n$, then we have a smaller upper bound $q(\gamma, \lambda_k, \lambda_{n+1}, \lambda_k)^2$.

Apart from the assumptions on the subspace $\mathcal{H}^i$, the above two estimates differ in the following: (7.6) is recursive, i.e., it estimates the reduction in the error after one iteration, whereas (7.4) is not. Recursive estimates are more convenient than nonrecursive in certain respects, but, as has been pointed out in [12], it is not generally possible to obtain for BPSD iterations a convergence estimate of the form $\lambda_j^{i+1} - \lambda_j \leq q(\lambda_j^i - \lambda_j)$ even if we make assumptions that would guarantee that λ_j^i converges to λ_j. The latter observation has led to the idea to look for estimates for groups of eigenpairs rather than for individual ones. This novel approach to the convergence analysis of subspace iterations resulted in the following estimate[7] for the locally optimal BPSD (7.3): assuming that L is positive definite and $\mu_k^i > \mu_{k+1}$ and $\mu_m^i > \mu_{m+1}$ (where μ's are the inverses of λ's) for some $k \leq m$, one has

$$\sum_{j=1}^{k} (\mu_j - \mu_j^{i+1}) \leq \frac{\gamma_{k,m}^2 + \varepsilon_{k,m,i}}{1 + \varepsilon_{k,m,i}} \sum_{j=1}^{k} (\mu_j - \mu_j^i), \tag{7.9}$$

where

$$\gamma_{k,m} = \frac{1 - \xi_{k,m}}{1 + \xi_{k,m}}, \quad \xi_{k,m} = \xi \frac{\mu_k - \mu_{m+1}}{\mu_k - \mu_{-1}} \tag{7.10}$$

and

$$\varepsilon_{k,m,i} = \mathcal{O}\left(\tan^2(\mathcal{I}_k^i, \mathcal{I}_k)_L\right) + \mathcal{O}\left(\sin^2(\mathcal{I}_m^i, \mathcal{I}_m)_L\right) + \mathcal{O}\left(\sum_{j=1}^{k} \|r_j^i\|_{L^{-1}}^2 \right),$$

where $\mathcal{I}_l^i = \operatorname{span}\{u_1^i, \ldots, u_l^i\}$ and $\mathcal{I}_l = \operatorname{span}\{u_1, \ldots, u_l\}$. The above estimate is cluster robust in the sense that $\varepsilon_{k,m,i}$ does not depend on the distances between $\mu_1, \ldots, \mu_k$; moreover, for $k < m$ any distances between consecutive eigenvalues (precisely, those

[7] For simplicity, here we present this result in asymptotic form.

between μ_k and μ_{k+1} and between μ_m and μ_{m+1}) appear only in the asymptotically insignificant term $\varepsilon_{k,m,i}$.

From (7.9) it is not difficult to derive the following estimate for the a.c.f. defined by (7.8):

$$(7.11)\quad q_k(\mathcal{H}^0) \le \gamma_{k,m}^2 \ \forall\ \mathcal{H}^0 \subset \mathcal{E} : \lambda_m(\mathcal{H}^0) < \lambda_{m+1},\ 1 \le k \le m \le n = \dim \mathcal{H}^0.$$

The assumption on $\lambda_m(\mathcal{H}^0)$ in the above estimate is essential and reflects the positive effect of the convergence of λ_m^i to λ_m on the convergence of λ_k^i to λ_k for $k < m$. Indeed, if λ_m^i converges to λ_m for some $m > k$, then for some i_0 we have $\lambda_k^{i_0} < \lambda_k$ and $\lambda_m^{i_0} < \lambda_m$ (cf. [12]) and the a.c.f. is bounded from above by $\gamma_{k,m}^2$, whereas in the case of $\mathcal{H}^0$ given by (7.7) we can only take $m = k$, which yields a larger upper bound $\gamma_{k,k}^2$, similar to that following from (7.6).

Using the results of the previous section, it is easy to show that the estimate (7.11) is sharp in the following sense.

THEOREM 7.1. *For any $\xi \in [0, 1)$ there exists K such that*

$$(7.12)\qquad q_k = \sup_{\mathcal{H}^0 : \mathcal{I}_m^i \to \mathcal{I}_m} q_k(\mathcal{H}^0) = \gamma_{k,m}^2,$$

where $\gamma_{k,m}$ is given by (7.10) *with $\xi_{k,m} = \xi$.*

Proof. From (7.11) it follows that $q_k \le \gamma_{k,m}^2$. To prove that $q_k = \gamma_{k,m}^2$, consider the case of $\mathcal{H}^0$ given by

$$\mathcal{H}^0 = \operatorname{span}\{u_1, \ldots, u_{k-1}, u_k + \tau u_{m+1}, u_{k+1}, \ldots, u_m, u_{m+2}, \ldots, u_{n+1}\}.$$

Let P be the orthogonal projector onto the invariant subspace corresponding to λ_{m+2} and let $K = L^{-1}(1 - P) + \xi L^{-1} P$. It is easy to see that with such K and $\mathcal{H}^0$ we have $u_j^i = u_j$ for $j = 1, \ldots, k-1, k+1, \ldots, m$, whereas u_k^i is computed recursively by (1.3). Let L' be the restriction of L to the subspace orthogonal to $u_1, \ldots, u_{k-1}, u_{k+1}, \ldots, u_m$. The eigenvalues of L' listed in increasing order are $\lambda_k, \lambda_{m+1}, \lambda_{m+2}, \ldots$, and PSD iterations for computing the minimal eigenvalue of L' started from u_k^0 produce the same sequence u_k^i as (7.3). It remains to be noted that the value of $\tilde{\gamma}_1$ for L' coincides with $\gamma_{k,m}$. Hence, by applying Theorem 5.3 to these iterations, we see that $q_k = \gamma_{k,m}^2$. □

Appendix A. Auxiliary results.

A.1. The proof of Theorem 6.1. To simplify the notation, let us denote $u = u_1^i$, $\lambda = \lambda_1^i$, $u' = u_1^{i+1}$, $\lambda' = \lambda_1^{i+1}$, $\kappa_1 = \kappa_{1,\alpha,\lambda_1^i}$, $\kappa_{-1} = \kappa_{-1,\alpha,\lambda_1^i}$, $\gamma = \gamma_{\alpha,\lambda_1^i}$, and $L_\lambda = L_{\alpha,\lambda_1^i}$ (we note that L_λ is positive definite in view of the assumption that $\lambda < \lambda_2$).

Denoting $v = L_\lambda^{1/2} u$ and $v' = L_\lambda^{1/2} u'$, and taking $\tau_i = \tau_* \equiv 2/(\kappa_1 + \kappa_{-1})$, we have

$$v' = v - \tau L_\lambda^{1/2} K L_\lambda^{1/2} L_\lambda^{-1/2}(L - \lambda) L_\lambda^{-1/2} v = v - \hat{K}(-\epsilon P_1 + P_1^\perp) v,$$

where $\epsilon = (\lambda - \lambda_1)/\alpha$, $P_1^\perp = 1 - P_1$ and $\hat{K} = \tau_* L_\lambda^{1/2} K L_\lambda^{1/2}$; we note that $1 - \gamma \le \hat{K} \le 1 + \gamma$. Hence,

$$v' = v - (-\epsilon P_1 + P_1^\perp) v - (\hat{K} - 1)(-\epsilon P_1 + P_1^\perp) v = (1 + \epsilon) P_1 v - w,$$

where $w = (\hat{K} - 1)(-\epsilon P_1 + P_1^\perp) v$. From the above relationship we have

$$\|P_1 v'\| \ge (1 + \epsilon)\|P_1 v\| - \|P_1 w\|, \quad \|P_1^\perp v'\| = \|P_1^\perp w\|,$$

and for w we have

$$\|w\| \le \gamma\|(-\epsilon P_1 + P_1^\perp)v\| = \gamma\sqrt{\epsilon^2\|P_1 v\|^2 + \|P_1^\perp v\|^2} = \gamma\|P_1 v\|\sqrt{\epsilon^2 + t^2},$$

where $t = \|P_1^\perp v\|/\|P_1 v\|$. Since

$$\|P_1 v\|^2 = \|P_1 L_\lambda^{1/2} u\|^2 = \|L_\lambda^{1/2} P_1 u\|^2 = \|P_1 u\|_{L_\lambda}^2 = \alpha\|P_1 u\|^2$$

and

$$\|P_1^\perp v\|^2 = \|P_1^\perp u\|_{L_\lambda}^2 = ((L-\lambda)P_1^\perp u, P_1^\perp u) = (\lambda - \lambda_1)\|P_1 u\|^2,$$

we have $t^2 = \epsilon$, and hence

$$\|w\| \le \gamma\|P_1 v\|\sqrt{\epsilon(1+\epsilon)}.$$

Now, denoting $\cos\omega = \|P_1 w\|/\|w\|$, $\omega \ge 0$, we have

$$\|P_1 v'\| \ge (1+\epsilon)\|P_1 v\| - \|w\|\cos\omega, \quad \|P_1^\perp v'\| = \|w\|\sin\omega,$$

and hence

$$\frac{\|P_1^\perp v'\|^2}{\|P_1 v'\|^2} \le \frac{\|w\|^2\sin^2\omega}{((1+\epsilon)\|P_1 v\| - \|w\|\cos\omega)^2} \le \frac{\gamma^2\epsilon(1+\epsilon)\sin^2\omega}{\left(1+\epsilon-\gamma\sqrt{\epsilon(1+\epsilon)}\cos\omega\right)^2}$$

$$= \frac{\epsilon}{1+\epsilon}\gamma^2\frac{\sin^2\omega}{\left(1-\sqrt{\frac{\epsilon}{1+\epsilon}}\gamma\cos\omega\right)^2} = a^2\frac{1-x^2}{(1-ax)^2} \equiv f(x), \tag{A.1}$$

where $a = \gamma\sqrt{\epsilon/(1+\epsilon)} \le 1$ and $x = \cos\omega$. Elementary calculations show that the maximum of $f(x)$ on $[0,1]$ is achieved at $x = a$, hence the right-hand side of (A.1) is not greater than

$$f(a) = \frac{a^2}{1-a^2} = \frac{\epsilon\gamma^2}{1+(1-\gamma^2)\epsilon} = \frac{\gamma^2}{1+(1-\gamma^2)\frac{\lambda-\lambda_1}{\alpha}}\frac{\lambda-\lambda_1}{\alpha}.$$

It remains to estimate the left-hand side of (A.1) from below. We have (cf. the above calculations for $\|P_1 v\|$ and $\|P_1^\perp v\|$):

$$\|P_1 v'\|^2 = \|P_1 u'\|_{L_\lambda}^2 = \alpha\|P_1 u'\|^2$$

and

$$\begin{aligned}\|P_1^\perp v'\|^2 &= ((L-\lambda)P_1^\perp u', P_1^\perp u') = ((L-\lambda_1)P_1^\perp u', P_1^\perp u') - (\lambda-\lambda_1)\|P_1^\perp u'\|^2\\ &= (\lambda'-\lambda_1)\|u'\|^2 - (\lambda-\lambda_1)\|P_1^\perp u'\|^2 = (\lambda'-\lambda_1)\|P_1 u'\|^2 - (\lambda-\lambda')\|P_1^\perp u'\|^2.\end{aligned}$$

Let us now verify that $\lambda' \le \lambda$. Assuming that the opposite is true, we would have $\|P_1^\perp v'\|^2 \ge (\lambda'-\lambda_1)\|P_1 u'\|^2$ and (A.1) would imply

$$\frac{\lambda'-\lambda_1}{\alpha} \le \frac{\|P_1^\perp v'\|^2}{\|P_1 v'\|^2} \le \frac{\gamma^2}{1+(1-\gamma^2)\frac{\lambda-\lambda_1}{\alpha}}\frac{\lambda-\lambda_1}{\alpha} \le \frac{\lambda-\lambda_1}{\alpha},$$

which is a contradiction. Hence, $\lambda' \le \lambda$ and, using the well-known estimate (see, e.g., [5])

$$\frac{\|P_1^\perp u'\|^2}{\|P_1 u'\|^2} = \tan^2(u', \mathcal{I}_1) \le \frac{\lambda' - \lambda_1}{\lambda_2 - \lambda'},$$

where $\mathcal{I}_1$ is the invariant subspace corresponding to λ_1, we obtain

$$\frac{\|P_1^\perp v'\|^2}{\|P_1 v'\|^2} = \frac{1}{\alpha}\left(\lambda' - \lambda_1 - (\lambda - \lambda')\frac{\|P_1^\perp u'\|^2}{\|P_1 u'\|^2}\right)$$
$$\ge \frac{1}{\alpha}\left(\lambda' - \lambda_1 - (\lambda - \lambda')\frac{\lambda' - \lambda_1}{\lambda_2 - \lambda'}\right) = \frac{\lambda' - \lambda_1}{\lambda_2 - \lambda'}\frac{\lambda_2 - \lambda}{\alpha},$$

which, together with the above estimate for the right-hand side, leads to (6.1).

To prove that (6.1) is sharp, let us denote $\phi = \arctan\sqrt{\epsilon} > 0$, where $\epsilon = (\lambda - \lambda_1)/\alpha$, and consider the case $v = x_1 u_1 + x_2 u_2 + x_3 u_3$, where u_1, u_2, and u_3 are eigenvectors corresponding to $\lambda_1 < \lambda_2 < \lambda_3$ (any of these eigenvalues may be multiple), and

$$x_1 = \cos\phi, \quad x_2 = -\frac{\gamma\sin\phi}{\sqrt{1 + (1-\gamma^2)\tan^2\phi}}, \quad x_3 = \frac{\sqrt{1-\gamma^2}\tan\phi}{\sqrt{1 + (1-\gamma^2)\tan^2\phi}}$$

(note that $x_2^2 + x_3^2 = \sin^2\phi$). Let $\cos\omega = \gamma\sqrt{\epsilon/(1+\epsilon)} = \gamma\sin\phi$, and denote $\hat{w} = \cos\omega \cdot u_1 + \sin\omega \cdot u_2$, $v^\perp = -\epsilon P_1 v + P_1^\perp v$ (we have $(v^\perp, v) = -\epsilon\|P_1 v\|^2 + \|P_1^\perp v\|^2 = -\tan^2\phi\cos^2\phi + \sin^2\phi = 0$, hence the notation) and $\hat{v}^\perp = v^\perp/\|v^\perp\|$. Let $K = L_\lambda^{-1/2}\hat{K}L_\lambda^{-1/2}$, where $\hat{K}$ is defined by

$$\hat{K}u = u + \gamma(u - 2(u,z)z), \quad z = \|\hat{v}^\perp - \hat{w}\|^{-1}(\hat{v}^\perp - \hat{w}).$$

It is easy to verify that $\kappa_1 = 1 - \gamma$ and $\kappa_{-1} = 1 + \gamma$ (hence $\tau_* = 1$) and that $(\hat{K} - 1)\hat{v}^\perp = \gamma\hat{w}$. Since

$$\|v^\perp\|^2 = \epsilon^2\|P_1 v\|^2 + \|P_1^\perp v\|^2 = \tan^4\phi\cos^2\phi + \sin^2\phi = \tan^2\phi$$

we have

$$\begin{aligned} w = (\hat{K} - 1)v^\perp = \|v^\perp\|(\hat{K} - 1)\hat{v}^\perp &= \gamma\tan\phi\left(\gamma\sin\phi \cdot u_1 + \sqrt{1 - \gamma\sin^2\phi} \cdot u_2\right) \\ &= \gamma^2\tan\phi\sin\phi \cdot u_1 + \gamma\tan\phi\sqrt{1 - \gamma\sin^2\phi} \cdot u_2 \\ &= \gamma^2\tan^2\phi\cos\phi \cdot u_1 + \gamma\sin\sqrt{1 + (1-\gamma^2)\tan^2\phi} \cdot u_2 \end{aligned}$$

and hence

$$\begin{aligned} v' = (1+\epsilon)P_1 v - w &= (1 + (1-\gamma^2)\tan^2\phi)\cos\phi \cdot u_1 - \gamma\sin\sqrt{1 + (1-\gamma^2)\tan^2\phi} \cdot u_2 \\ &= (1 + (1-\gamma^2)\tan^2\phi)(x_1 u_1 + x_2 u_2) \end{aligned}$$

Since $v' \in \operatorname{span}\{u_1, u_2\}$, and therefore $u' \in \operatorname{span}\{u_1, u_2\}$, we have

$$\frac{\|P_1^\perp v'\|^2}{\|P_1 v'\|^2} = \frac{x_2^2}{x_1^2} = \frac{\gamma^2\tan^2\phi}{1 + (1-\gamma^2)\tan^2\phi}, \quad \frac{\|P_1^\perp u'\|^2}{\|P_1 u'\|^2} = \frac{\lambda' - \lambda_1}{\lambda_2 - \lambda'}.$$

Further,

$$\frac{\|P_1^\perp v'\|^2}{\|P_1 v'\|^2} = \frac{((L-\lambda)P_1^\perp u', P_1^\perp u')}{\alpha\|P_1 u'\|^2} = \frac{\lambda_2-\lambda}{\alpha}\frac{\|P_1^\perp u'\|^2}{\|P_1 u'\|^2} = \frac{\lambda_2-\lambda}{\alpha}\frac{\lambda'-\lambda_1}{\lambda_2-\lambda'}.$$

Recalling that $\tan^2\phi = \epsilon^2 = (\lambda-\lambda_1)/\alpha$, and observing that λ' is a function of λ, γ, and α, we arrive at the identity

$$\Delta(\lambda') = \frac{\gamma^2}{1+(1-\gamma^2)\frac{\lambda-\lambda_1}{\alpha}}\Delta(\lambda), \quad \lambda\in[\lambda_1,\lambda_2), \quad \gamma\in[0,1], \quad \alpha>0.$$

It remains to be shown that $\lambda' \le \lambda(u-\tau K(L-\lambda)u)$ for any τ. Let us denote $v'_\tau = v-\tau\hat{K}v^\perp$ and $u'_\tau = L_\lambda^{-1/2}v'_\tau$. Using the fact that $v' = (1+(1-\gamma^2)\tan^2\phi)(x_1u_1+x_2u_2)$ (cf. above), we have

$$\begin{aligned} v'_\tau &= v-\tau(v^\perp+w) = v'-(\tau-1)(v^\perp+w) = v'-(\tau-1)(v-v') \\ &= v'-(\tau-1)((1+(1-\gamma^2)\tan^2\phi)^{-1}v'+x_3u_3-v') \equiv \beta v'-(\tau-1)x_3u_3, \end{aligned}$$

and hence $u'_\tau = \beta u' - (\lambda_3-\lambda)^{-1/2}(\tau-1)x_3u_3$. Since $u'\in\operatorname{span}\{u_1,u_2\}$, this implies $\lambda(u'_\tau)\ge\lambda(u')=\lambda'$.

A.2. Auxiliary result used in Theorem 5.1.

LEMMA A.1. *Let ϵ be a sequence of positive real numbers such that*

$$\epsilon = \sum_{i=0}^{\infty}\epsilon_i < \infty.$$

Let κ be a positive real number and let t_i be a sequence of positive real numbers starting with $t_0=1$ and satisfying the following conditions: if $\epsilon_i(\kappa+t_i^{-2})<1$, then

$$t_{i+1} \le \frac{1+\epsilon_i(\kappa t_i^2+1)}{1-\epsilon_i(\kappa+t_i^{-2})}t_i^{-1} \tag{A.2}$$

and if $\epsilon_i(\kappa t_i^2+1)<1$, then

$$t_{i+1}^{-1} \le \frac{1+\epsilon_i(\kappa+t_i^{-2})}{1-\epsilon_i(\kappa t_i^2+1)}t_i. \tag{A.3}$$

If ϵ is small enough, then

$$t^{-1}\le t_i\le t = 1+c\,\epsilon, \tag{A.4}$$

where $c>0$ does not depend on i and ϵ.

Proof. Consider the following function:

$$\Phi_\epsilon(t) = F(\epsilon(\kappa t^2+1)) + F(\epsilon(\kappa+t^2)), \quad F(x) = x(1-\ln(1-x)). \tag{A.5}$$

For $\epsilon<\epsilon_\kappa\equiv\min\{(4\kappa+1)^{-1},(\kappa+4)^{-1}\}$, the definition domain of $\Phi_\epsilon(t)$ (as a real function of the real argument t) includes the interval $[1,2]$. Obviously, $\Phi_\epsilon(t)\to0$ as $\epsilon\to0$: let $\epsilon<\epsilon_\kappa$ be small enough so that $\Phi_\epsilon(2)<\ln 2$. Since $\Phi_\epsilon(1)>\ln 1$, the equation $\Phi_\epsilon(t)=\ln t$ has a solution $t=t(\epsilon)\in[1,2]$ that has the following asymptotics in ϵ: $t=\exp(F(\epsilon(\kappa t^2+1))+F(\epsilon(\kappa+t^2)))=1+\mathcal{O}(\epsilon)$.

Let us now denote $a = \kappa t^2 + 1$ and $b = \kappa + t^2$, and consider the sequence u_n defined as follows:

$$u_0 = 1, \quad u_n = \prod_{k=0}^{n-1} \frac{1+\epsilon_{2k+1}a}{1-\epsilon_{2k+1}b}\frac{1+\epsilon_{2k}b}{1-\epsilon_{2k}a}, \ n > 0$$

(note that $\epsilon_i a \le \epsilon(4\kappa+1) < 1$ and $\epsilon_i b \le \epsilon(\kappa+4) < 1$). Using elementary inequalities $\ln(1+x) \le x$ and $-\ln(1-x) \le x(1-\ln(1-x))$, we have

$$\begin{aligned}\ln u_n &= \sum_{k=0}^{n-1}(\ln(1+\epsilon_{2k+1}a) + \ln(1+\epsilon_{2k}b) - \ln(1-\epsilon_{2k+1}b) - \ln(1-\epsilon_{2k}a)) \\ &\le \sum_{k=0}^{n-1}(\epsilon_{2k+1}a + \epsilon_{2k}b + \epsilon_{2k+1}b + \epsilon_{2k}a - \epsilon_{2k+1}b\ln(1-\epsilon_{2k+1}b) - \epsilon_{2k}a\ln(1-\epsilon_{2k}a))\end{aligned}$$

and, since $\ln(1-x) \le 0$ for $0 < x \le 1$ and $\epsilon_i \le \epsilon$,

$$\begin{aligned}\ln u_n &\le (a(1-\ln(1-\epsilon a)) + b(1-\ln(1-\epsilon b)))\sum_{k=0}^{n}(\epsilon_{2k}+\epsilon_{2k+1}) \\ &\le (a(1-\ln(1-\epsilon a)) + b(1-\ln(1-\epsilon b)))\epsilon = \Phi_\epsilon(t) = \ln t;\end{aligned}$$

i.e., $u_n \le t$. By similar arguments, for the sequence

$$v_0 = 1, \quad v_n = \prod_{k=0}^{n-1} \frac{1+\epsilon_{2k+1}b}{1-\epsilon_{2k+1}a}\frac{1+\epsilon_{2k}a}{1-\epsilon_{2k}b}, \ n > 0$$

we have $v_n \le t$.

Let us now turn to the sequence t_i. Since $t_0 = 1 \le t$ and $\epsilon_0(\kappa+t_0^{-2}) < \epsilon(\kappa+4) < 1$, we have

$$t_1 \le \frac{1+\epsilon_0(\kappa t_0^2+1)}{1-\epsilon_0(\kappa+t_0^{-2})}t_0^{-1} \le \frac{1+\epsilon_0(\kappa t^2+1)}{1-\epsilon_0(\kappa+t^2)} \le \frac{1+\epsilon_1(\kappa+t^2)}{1-\epsilon_1(\kappa t^2+1)}\frac{1+\epsilon_0(\kappa t^2+1)}{1-\epsilon_0(\kappa+t^2)} = v_1 \le t$$

and, since $\epsilon_0(\kappa t_0^2+1) < \epsilon(4\kappa+1) < 1$, we have

$$t_1^{-1} \le \frac{1+\epsilon_0(\kappa+t_0^{-2})}{1-\epsilon_0(\kappa t_0^2+1)}t_0 \le \frac{1+\epsilon_0(\kappa+t^2)}{1-\epsilon_0(\kappa t^2+1)} \le u_1 \le t.$$

Using the above two estimates we, in turn, obtain

$$\begin{aligned}t_2 &\le \frac{1+\epsilon_1(\kappa t_1^2+1)}{1-\epsilon_1(\kappa+t_1^{-2})}t_1^{-1} \le \frac{1+\epsilon_1(\kappa t_1^2+1)}{1-\epsilon_1(\kappa+t_1^{-2})}\frac{1+\epsilon_0(\kappa+t_0^{-2})}{1-\epsilon_0(\kappa t_0^2+1)}t_0 \\ &\le \frac{1+\epsilon_1(\kappa t^2+1)}{1-\epsilon_1(\kappa+t^2)}\frac{1+\epsilon_0(\kappa+t^2)}{1-\epsilon_0(\kappa t^2+1)} = u_1 \le t \\ t_2^{-1} &\le \frac{1+\epsilon_1(\kappa+t_1^{-2})}{1-\epsilon_1(\kappa t_1^2+1)}t_1 \le \frac{1+\epsilon_1(\kappa+t_1^{-2})}{1-\epsilon_1(\kappa t_1^2+1)}\frac{1+\epsilon_0(\kappa t_0^2+1)}{1-\epsilon_0(\kappa+t_0^{-2})}t_0^{-1} \\ &\le \frac{1+\epsilon_1(\kappa+t^2)}{1-\epsilon_1(\kappa t^2+1)}\frac{1+\epsilon_0(\kappa t^2+1)}{1-\epsilon_0(\kappa+t^2)} = v_1 \le t\end{aligned}$$

etc. □

REFERENCES

[1] J. H. Bramble, J. E. Pasciak, and A. V. Knyazev, *A subspace preconditioning algorithm for eigenvector/eigenvalue computation*, Adv. Comput. Math., 6 (1996), pp. 159–189.

[2] V. E. Bulgakov, M. V. Belyi, and K. M. Mathisen, *Multilevel aggregation method for solving large-scale generalized eigenvalue problems in structural dynamics*, Internat. J. Numer. Methods Engrg., 40 (1997), pp. 453–471.

[3] E. G. D'yakonov and M. Yu. Orehov, *Minimization of the computational labor in determining the first eigenvalues of differential operators*, Math. Notes, 27 (1980), pp. 795–812.

[4] U. Hetmaniuk, A. Knyazev, R. Lehoucq, and E. Ovtchinnikov, *The use of the residual norm for the eigenvalue error estimation*, Technical report, Sandia National Laboratory, Albuquerque, NM, 2005.

[5] A. V. Knyazev, *Convergence rate estimates for iterative methods for a mesh symmetric eigenvalue problem*, Soviet J. Numer. Anal. Math. Modelling, 2 (1987), pp. 371–396.

[6] A. V. Knyazev and K. Neymeyr, *A geometric theory for preconditioned inverse iteration.* III. *A short and sharp convergence estimate for generalized eigenvalue problems*, Linear Algebra Appl., 358 (2003), pp. 95–114.

[7] K. Neymeyr, *A geometric theory for preconditioned inverse iteration.* I. *Extrema of the Rayleigh quotient*, Linear Algebra Appl., 322 (2001) pp. 61–85.

[8] K. Neymeyr, *A geometric theory for preconditioned inverse iteration.* II. *Convergence estimates*, Linear Algebra Appl., 322 (2001) pp. 87–104.

[9] Y. Notay, *Convergence analysis of inexact Rayleigh quotient iteration*, SIAM J. Matrix. Anal. Appl., 24 (2003) pp. 627–644.

[10] S. Oliveira, *On the convergence rate of a preconditioned subspace eigensolver*, Computing, 63 (1999) pp. 219–231.

[11] E. Ovtchinnikov, *Convergence estimates for the generalized Davidson method for symmetric eigenvalue problems.* I. *The preconditioning aspect*, SIAM J. Numer. Anal., 41 (2003) pp. 258–271.

[12] E. Ovtchinnikov, *Cluster robustness of preconditioned gradient subspace iteration eigensolvers*, Linear Algebra Appl., accepted, 2004.

[13] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, NJ, 1980.

[14] B. Samokish, *The steepest descent method for an eigenvalue problem with semi-bounded operators*, Izv. Vyssh. Uchebn. Zaved. Mat., 5 (1958) pp. 105–114, in Russian.

[15] G. L. G. Sleijpen and H. A. van der Vorst, *A Jacobi–Davidson iteration method for linear eigenvalue problems*, SIAM J. Matrix Anal. Appl., 17 (1996), pp. 401–425.